Научная статья

УДК 004.624
DOI 10.25205/1818-7900-2024-22-4-5-16

О повышении качества выходных данных в вопросно-ответной системе, обрабатывающей климатическую информацию

Ольга Юрьевна Гавенко[1]
Наталья Александровна Шашок[2]

[1,2]Федеральный исследовательский центр информационных и вычислительных технологий, Новосибирск, Россия

[1]Новосибирский государственный университет
Новосибирск, Россия

[1]olga.yu.gavenko@mail.ru, https://orcid.org/0000-0003-3619-1120
[2]n.shashok@alumni.nsu.ru, https://orcid.org/0009-0007-3658-6110

Аннотация

Разработка вопросно-ответной системы (QA), обрабатывающей климатическую информацию, опирается на использование разнородных климатических данных в различных форматах (текстовые, числовые, графические, видео, аудио, географические и данные мониторинга). Обязательным элементом вопросно-ответной системы должен являться инструмент, позволяющий обрабатывать и анализировать подобные данные.

Процессы поиска и извлечения данных выступают центральной частью рассматриваемой системы, поскольку от них во многом зависит качество сгенерированного ответа. Точный способ извлечения данных имеет решающее значение для выходных данных системы QA, а также для проблем принятия решений, так как существуют ситуации, в которых LLM генерирует ответы, соответствующие контексту, но фактически являющиеся неверными и не соответствующими входным данным. Использование правильных метрик и алгоритмов для некоторых типов данных и неправильных для других может привести к превышению допустимого порога нерелевантных данных, что, в свою очередь, может снизить качество ответов. Дополненная поисковая генерация (Retrieval-augmented Generation, RAG) также может использоваться для оптимизации входных данных для этой задачи.

В работе рассматриваются различные алгоритмы извлечения данных и ранжирования документов, а также возможность использования ансамблей агентов LLM при разработке вопросно-ответной системы, обрабатывающей климатическую информацию.

Ключевые слова

оптимизация входных данных, вопросно-ответные системы, разработка RAG-системы, обработка мультимодальных данных

Для цитирования

Гавенко О. Ю., Шашок Н. А. О повышении качества выходных данных в вопросно-ответной системе, обрабатывающей климатическую информацию // Вестник НГУ. Серия: Информационные технологии. 2024. Т. 22, № 4. С. 5–16. DOI 10.25205/1818-7900-2024-22-4-5-16

On Increasing the Quality of the Climate Observations Question-Answering System's Output Data

Olga Yu. Gavenko[1], Natalia A. Shashok[2]

Federal Research Center for Information and Computational Technologies
Novosibirsk, Russian Federation

[1]Novosibirsk State University,
Novosibirsk, Russian Federation

[1]olga.yu.gavenko@mail.ru, https://orcid.org/0000-0003-3619-1120
[2]n.shashok@alumni.nsu.ru, https://orcid.org/0009-0007-3658-6110

Abstract

The development of the climate observations question-answer (QA) information system relies on heterogeneous climate data in various formats (text, numerical, graphic, video, audio, geographic and monitoring data). A mandatory element of such a system is a tool that allows processing and analyzing such data.

Searching and retrieving data is a central part of the system in question, since the quality of the generated answer heavily depends on it. The exact way the data is retrieved is critical to the output of a QA system as well as to decision-making problems, since there are situations in which the LLM generates a contextually appropriate but factually incorrect answers that do not match the input. Using correct metrics and algorithms for some data types and incorrect ones for others can cause the permissible threshold of irrelevant data to be exceeded, which in turn can cause the quality of the answers to decrease. Retrieval-augmented generation (RAG) systems can also be used to optimize input data for that task.

This work discusses various algorithms for data extraction and document ranking, as well as the possibility of using ensembles of LLM agents in development of the QA system that works with climate data.

Keywords

Input data optimization, question answering systems, RAG system development, multimodal data processing.

For citation

Gavenko O., Shashok N. On increasing the quality of the climate observations question-answering system's output data. *Vestnik NSU. Series: Information Technologies*, 2024, vol. 22, no. 4, pp. 5–16 (in Russ.) DOI 10.25205/1818-7900-2024-22-4-5-16

Введение

Разработка информационной системы типа «вопрос-ответ», обрабатывающей и анализирующей климатические данные, соответствует целевому направлению Климатической доктрины Российской Федерации от 26 октября 2023 г.[1], определяющей климатическую политику Российской Федерации.

Климатические данные могут быть получены как из внешних источников (систем мониторинга и глобальной сети), так и из доступных внутренних хранилищ, и могут существовать в различных форматах (текстовые, числовые, графические, видео-, аудио-, географические и данные мониторинга), при этом для обработки должны быть доступны не только имеющиеся в хранилищах данные, но и данные, поступающие в систему в непрерывном режиме. Обязательным элементом подобной информационной системы должен быть инструмент, позволяющий обрабатывать и анализировать разнородные динамические и статические данные с целью их использования в алгоритмах построения и генерации ответов для решения широкого круга задач, связанных с поддержкой принятия решений.

Очевидно, что вопросы, на которые вопросно-ответная система, обрабатывающая климатические данные, должна быть способна ответить, могут быть различной сложности. Так, вопрос, направленный на определение автора какой-либо конкретной научной работы, требует обработки только этой самой работы, если она существует; для ответа на вопрос о том, находится ли озеро Байкал в Уральских горах, достаточен будет один документ из достовер-

[1] Russia's new Climate Doctrine approved, http://www.en.kremlin.ru/acts/news/72598

ного источника, описывающий местоположение озера Байкал. Однако существуют вопросы, требующие изучения большого количества информации, в частности, вопросы, связанные с проведением сравнительного анализа каких-либо данных в течение некоторого периода времени. В качестве примера можно рассмотреть следующую задачу: определение совокупного количества CO_2, образующегося на конкретной территории в Российской Федерации в связи с работой промышленных предприятий региона. Для ответа на подобный вопрос может потребоваться изучение нескольких научных статей, в которых есть информация по каждому предприятию, но за отсутствием необходимых статей могут использоваться данные мониторинга и спутниковой съемки, выявляющей изменения содержания CO_2 в атмосфере за некоторый период.

Система может не предусматривать получения ответов на риторические вопросы и вопросы по более широкой тематике; представляется допустимым, что при попытке ответить на вопрос, не связанный с целевой тематикой системы, система может либо сгенерировать слабо связанный с вопросом ответ, либо полностью отказаться от его генерации. Подобное поведение вполне соответствует самому определению задачи автоматического нахождения ответа на вопросы [1]: это задача, которая направлена на генерацию правильного ответа на вопросы, специфичные для некоторой предметной области, на основе заданного контекста или базы знаний.

В эпоху развития глубокого обучения и больших языковых моделей задача генерации ответов может быть решена более сложными методами по сравнению с такими классическими методами, как прямое сопоставление, поиск ключевых слов, разметка частей речи и разбор фрагментов, использовавшимися ранее, начиная с 1970-х годов [2; 3]. Тем не менее основной критерий решения задачи не меняется: это построение правильного и корректного ответа на вопросы в предметной области и, возможно, вне предметной области, с помощью предварительной обработки входных данных и последующего использования результата для поиска в базе данных предварительно обработанных документов. Основное различие с более ранними разработками проявляется в использовании современных подходов к вычислению и предварительной обработке данных. Такие подходы включают преобразование или связывание частей документов различных типов с векторами и поиск частей документов или данных с нужным контекстом в общем векторном пространстве, а затем их использование для предоставления необходимого контекста для использования в некоторой большой языковой модели (LLM), такой как GPT[2], Gemini[3], Gemma[4], Falcon[5] и других.

Подход к решению задачи генерации ответа на вопрос, рассматриваемый в представленной работе – Retrieval Augmented Generation (RAG) [4] – заключается в следующем. На первом этапе используется система поиска векторной информации (IR) для получения документов, релевантных запросу пользователя, после чего применяется модуль извлечения информации (IE) для фильтрации данных, а именно отсечения нерелевантных запросу данных и осуществление выборки необходимых для предоставления контекста частей документов. На заключительном этапе полученная информация объединяется с запросом пользователя для формирования полного контекста, необходимого для корректной генерации ответа. При подобном подходе система IR обычно предоставляет пользователю возможность генерировать ответы самостоятельно на основе найденных документов. Модуль IE, а также LLM, генерирующая ответ, в свою очередь, могут автоматизировать задачу извлечения информации.

Модель LLM может использовать API поиска векторного хранилища для генерации более точных ответов, которые пользователь получает после генерации более краткого вопроса для LLM. Общий вид такого потока данных представлен на рис. 1.

[2] GPT-4 | OpenAI. https://openai.com/index/gpt-4/

[3] Gemini, https://gemini.google.com/

[4] Google AI Gemma open models - Gemini API, https://ai.google.dev/gemma

[5] Falcon LLM, https://falconllm.tii.ae

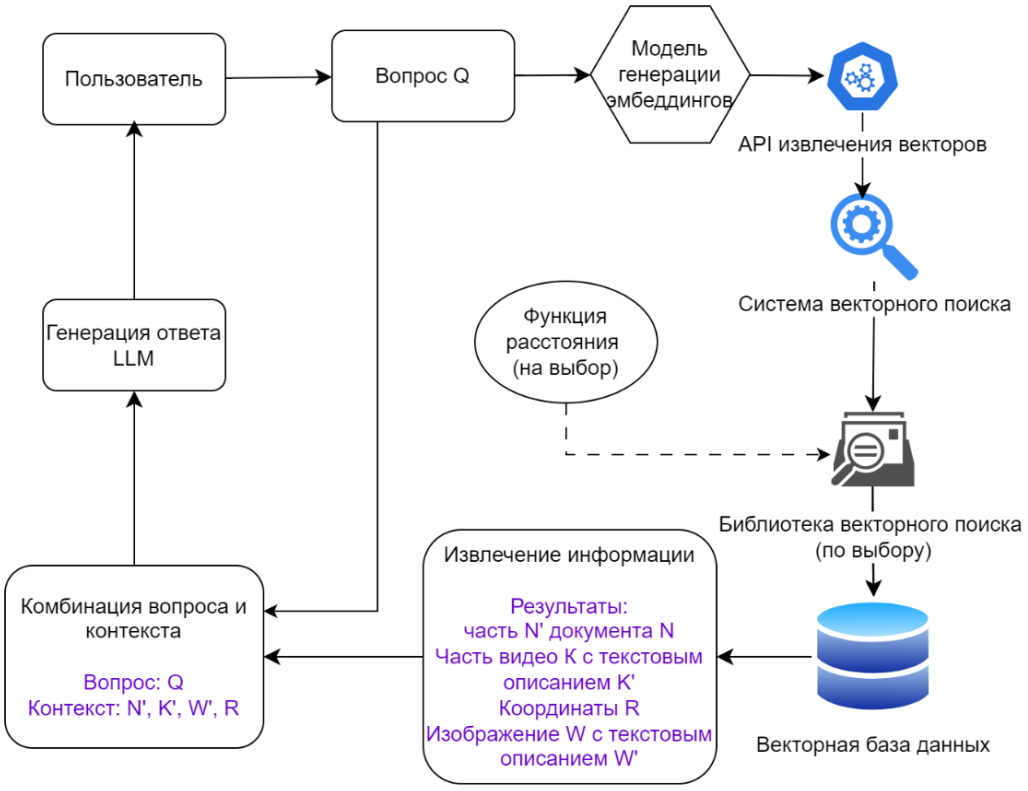

Рис. 1. Общий вид потока данных вопросно-ответной системы, обрабатывающей климатическую информацию
Fig. 1. A general view into the data flow within the retrieval augmented generation architecture for the climate observation question answering system

С учетом того, что объем данных, используемых разрабатываемой системой, может быть довольно большим, хранение данных в разных хранилищах и объединение их с помощью нескольких модулей с одним и тем же API-интерфейсом, зависящим от типа входных данных, на основе которых определялись эмбеддинги, представляется целесообразным, что изменяет поток данных, как показано на рис. 2.

Можно отметить несколько причин, по которым разделение хранилищ по типу или происхождению данных с последующим их объединением под одним API представляется обоснованным. Во-первых, общий размер набора данных не должен сильно ухудшать параметр скорости ответа системы на ввод запроса пользователем. Системы, подобные разрабатываемой системе обработки климатических данных, на данный момент используют весьма разный объем документов: от 35–50 тысяч отдельных утверждений[6] до миллионов документов, новостных каналов и сообщений из сети Интернет (например, MediSearch[7]). Разнообразные наборы данных, используемые для разработки подобных вопросно-ответных систем, такие как Cohere's Wikipedia Embeddings[8], CORD-19 [5], NewsQA[9], SQuAD [6] и другие, отличаются большим объемом данных, и со временем этот объем растет, поскольку эти наборы данных расширяются их разработчиками. Во-вторых, предоставление системе возможности выбирать, какие именно данные ей нужны, и указывать точный источник данных, с большой долей веро-

[6] SberQuAD (Sberbank Question Answering Dataset), https://paperswithcode.com/dataset/sberquad

[7] MediSearch, https://medisearch.io

[8] Wikipedia (en) embedded with cohere.ai multilingual-22-12 encoder, https://huggingface.co/datasets/Cohere/wikipedia-22-12-en-embeddings

[9] Dataset Card for NewsQA, https://huggingface.co/datasets/Maluuba/newsqa

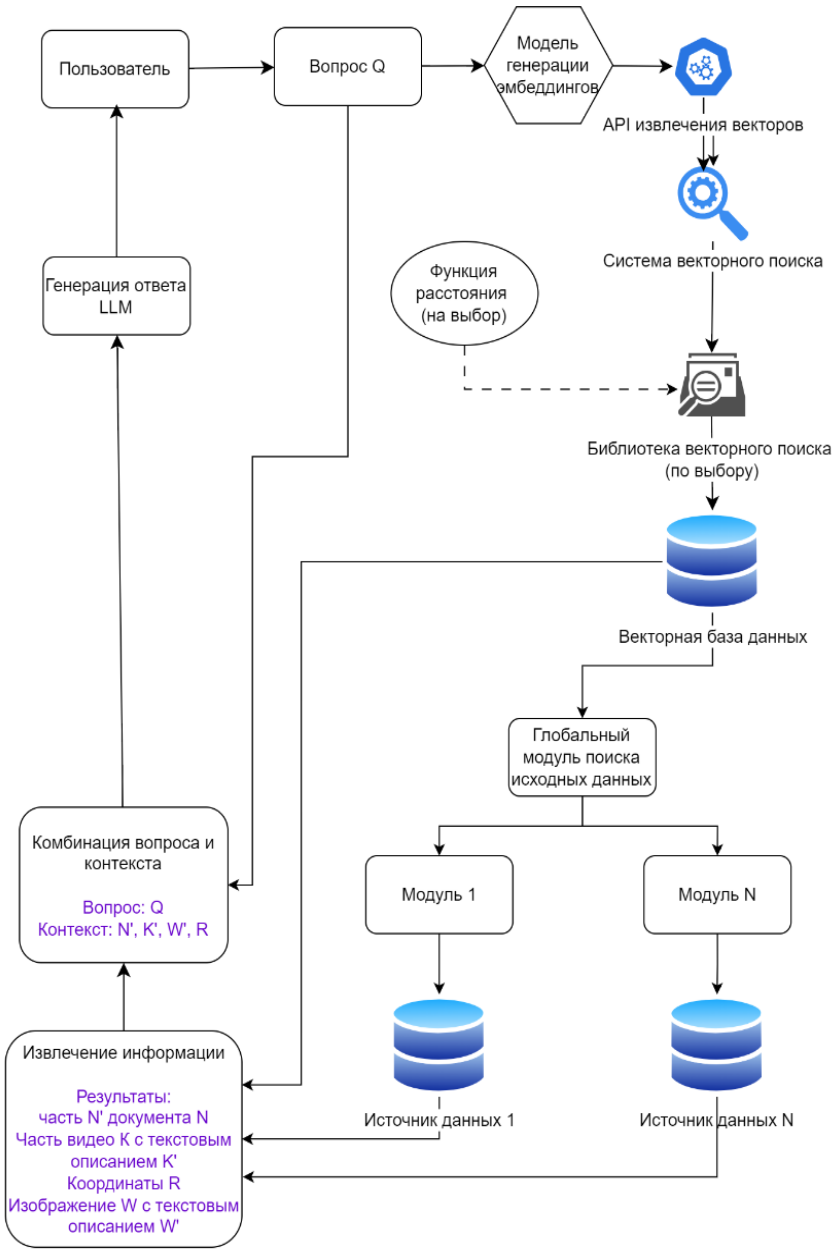

Рис. 2. Уточненный вид потока данных вопросно-ответной системы, обрабатывающей климатическую информацию
Fig. 2. A more detailed view into the data flow within the retrieval augmented generation architecture for the climate observation question answering system

ятности снизит нагрузку на сами источники данных, поскольку, при необходимости получить данные только из одного источника, запросы к остальным источникам осуществляться не будут. В-третьих, при административной поддержке системы также может быть полезно знать, какой именно источник данных содержит определенные данные.

Следует подчеркнуть, что описанный выше подход в его базовой постановке может быть недостаточным. Известна проблема так называемых «галлюцинаций», когда модель использует по большей части только «лучший», по своему «мнению», полученный результат данных без дополнительного анализа или верификации, и при этом генерирует фактически не-

верные ответы [7]. Примером таких «галлюцинаций» может быть текущее состояние ИИ LLM Google, который был обучен использовать лучшие результаты из всех доступных веб-документов без последующего анализа на достоверность источника[10], а также на релевантность и корректность предоставленных данных. Это приводит к тому, что LLM Google рассматривает веб-документы, содержащие недостоверную информацию, в том числе, написанную пользователями сети Интернет для умышленного введения других людей в заблуждение, как заслуживающую использования при генерации ответов. В идеальном случае алгоритм PageRank [8] должен гарантировать, что подобных ситуаций происходить не будет, однако существуют примеры обратного[11].

Учитывая вышеизложенное, планируется первоначально протестировать описываемую RAG-систему на базе знаний, состоящей, как минимум, из 40 тысяч документов, разделенных между двумя узлами источников данных. На данный момент такое количество документов нельзя назвать достаточным для конечной системы, однако использование небольшого количества документов представляется разумным в целях тестирования как системы, так и проверки документов на достоверность. Впоследствии предполагается расширение базы знаний и такой разработки архитектуры системы и ее реализации, чтобы она могла потенциально обрабатывать как минимум такое же количество документов, которое содержится в наборе данных CORD-19.

Подходы к решению проблемы

Следует подробнее рассмотреть некоторые вопросы, возникающие при описанных выше проблемах вопросно-ответной системы, обрабатывающей климатическую информацию: в частности, проблему повышения качества генерации ответов на вопросы и задачи, возникающие при разработке подхода к извлечению информации.

На данный момент не существует единого общепризнанного проверенного подхода повышения качества генерации ответов с помощью больших языковых моделей. Тем не менее представляется возможным выстроить систему таким образом, чтобы можно было эмпирически найти необходимые параметры, при которых система показала бы хорошие результаты, поскольку существуют некоторые отдельные решения повышения качества генерируемых ответов, и комбинация этих подходов могла бы дать более весомый результат, чем использование только одного.

Рассмотрим эти подходы детально.

Правильное извлечение информации

Перед генерацией ответа необходимо правильным образом извлекать информацию, подаваемую в качестве контекста LLM. Настраиваться может в том числе и сам процесс информационного поиска, например, через выбор библиотек и мер близости, которые используются для поиска документов в векторном пространстве.

Помимо этого, необходимо учитывать, что база знаний разрабатываемой вопросно-ответной системы хранения и обработки климатической информации является принципиально мультимодальной, поскольку она может содержать не только текстовые данные, но и карты, данные климатических наблюдений, изображения, видео и звуки. Примерами таких данных являются, помимо прочего, экологические словари, такие как [9], данные, полученные от Федеральной службы по гидрометеорологии и мониторингу окружающей среды (карты, новости,

[10] https://support.google.com/websearch/answer/14901683

[11] https://www.nytimes.com/2024/05/24/technology/google-ai-overview-search.html

таблицы)[12], данные повторного анализа системы климатического прогнозирования[13], научные статьи открытого доступа из таких проектов, как CyberLeninka[14] и др.

Определение «мультимодальные данные», которое дается данным разных форматов в ряде ИТ-задач, отражает совокупность разнородных документов, обрабатываемых и анализируемых современными информационными системами. Количество задач, требующих разработки подходов к обработке специфичных мультимодальных данных, увеличивается из года в год, и это связано, прежде всего, с преимуществом систем, способных анализировать данные разных форматов, в сравнении с системами, обрабатывающими только один тип данных, поскольку подобное характерно для самых разных областей современной науки и техники. В ряде задач схожие данные определяются как гетерогенные данные, что также отражает суть проблемы: в задачах, в которых должны использоваться различные данные, имеющие разную структуру, источники и методы обработки, принципиально важно определить подход, учитывающий их аутентичность, и выбрать соответствующие инструменты, позволяющие одновременно обрабатывать и анализировать данные разных форматов, а затем использовать полученные результаты в вопросно-ответных системах для поддержки принятия решений.

Для задач, связанных с обработкой климатической информации вопросно-ответными системами, способность системы обрабатывать мультимодальные данные является центральной, учитывая, что климатическая информация может поступать из различных источников, которые возможно сгруппировать по ряду признаков. В общем случае климатические данные могут быть следующих форматов: текстовые, числовые, графические, видео-, аудио-, картографические данные и данные мониторинга. Помимо этого, всю совокупность климатических данных можно разделить на две части. Во-первых, это статические данные, загружаемые из источников, находящихся в фиксированных и редко обновляемых базах данных, содержащих словари, тезаурусы, справочную литературу и т. д., т. е. источники, информация в которых слабо изменяется по своему содержанию, например, словарные статьи, содержащие текстовые и графические данные. Во-вторых, динамические данные, которые изменяются во временных интервалах, накапливаются или удаляются как неактуальные, – это данные мониторинга, а также картографические данные. Очевидно, что графические данные могут быть как частью статической информации, если они входят в словарную статью, так и изменяемыми в хронологическом аспекте, если они получены в результате обработки данных мониторинга. Таким образом, для корректной обработки информации необходимо дополнительно учитывать источники данных, их тип и характер.

Однако такая неоднородность данных, как по типу, так и по характеру, усложняет и выбор инструментов, и разработку архитектуры вопросно-ответной системы, работающей с мультимодальными данными, а также привносит другие проблемы, поскольку использование корректных метрик и алгоритмов для одних типов данных и некорректных для других может привести к превышению допустимого порога нерелевантных данных, что в свою очередь потенциально приводит к снижению качества ответа. Очевидно, что при проектировании способ генерации эмбеддингов для привязки к документам должен быть определен на основе используемого подхода к извлечению данных.

Один из таких подходов состоит в следующем. Системы обработки мультимодальных данных могут использовать различные алгоритмы извлечения и ранжирования для каждого из используемых типов данных, с учетом того, что различные типы данных в различных хранилищах могут храниться с использованием неидентичных векторных пространств. Противоположным подходом к извлечению данных могло бы быть объединение данных из различных

[12] ЕИП Росгидромета, https://eip.meteo.ru/opendata

[13] Climate Data Guide: Climate Forecast System Reanalysis (CFSR), https://climatedataguide.ucar.edu/climate-data/climate-forecast-system-reanalysis-cfsr

[14] Научная электронная библиотека «КиберЛенинка», https://cyberleninka.ru

модальностей в некоторую однородную форму – либо с использованием общего векторного пространства, либо путем координирования различных пространств, и урезания либо расширения мерности. Последний подход представляется многообещающим при работе с неоднородными данными сильно отличных между собой типов.

Определение уровня шума в выдаче

Документ, возвращаемый системой извлечения данных, может быть релевантным запросу, но при этом устаревшим, связанным с обозначенной тематикой, но не содержащим ответ на вопрос, либо вообще случайным. Последние рассматриваются как шум, от которого необходимо очистить данные для повышения качества сгенерированного ответа.

Однако некоторые исследования показывают, что это может быть не так [10], и некоторые нерелевантные документы, рассматриваемые как шум, имеют возможность повысить качество генерации, если они правильным образом размещены в контекст, подаваемый на вход LLM, в то время как семантически связанные документы, не содержащие ответа, значительно снижают качество. Важно отметить, что добавление шума ухудшает некоторые метрики оценки поиска, такие как fall-out rate; из этого можно сделать вывод, что в разрабатываемую систему необходимо добавить механизм, позволяющий при генерации ответа выбрать уровень «шума» в выдаче контекста в целях поиска границы нерелевантности, после которой качество ответов начинает падать.

Использование ансамблей

Еще одним подходом к улучшению результатов поиска является одновременное использование нескольких различных методов поиска и объединение результатов с помощью голосования или других методов комбинирования, этот подход называется ансамблем [11]. При таком подходе можно упорядочить части контекста при подаче на вход LLM таким образом, который может быть недостижимым при использовании одного метода, что, в свою очередь, может повлиять на качество сгенерированных ответов. Для проверки этого утверждения представляется целесообразным добавить возможность настройки включения пользователем использования ансамблей и выбора алгоритмов голосования при составлении контекста для генерации ответа.

Методы генерации эмбеддингов

Учитывая, что некоторые документы могут содержать большой объем данных, необходимо определить ответы на следующие вопросы: какую часть документа следует преобразовать в эмбеддинг; следует ли каждому документу составлять некоторое краткое описание или необходимо разбивать документы на фрагменты; насколько большим должно быть описание или часть документа; можно ли использовать два этих подхода совместно, несмотря на потерю данных при составлении обобщения; следует ли ранжировать или классифицировать фрагменты перед генерацией по ним эмбеддингов.

Выбор правильного для обработки размера фрагмента представляется целесообразным по следующим причинам. Небольшой размер фрагмента может обеспечить более высокую детализацию, точно указав необходимый контекст в документах. С другой стороны, это может означать, что некоторый контекст может быть упущен. Использование же фрагментов крупных размеров может определять большие затраты по времени при генерации ответа, поскольку больше информации передается LLM для обработки. В целях тестирования полезно иметь несколько векторных индексов в векторной базе данных или непересекающиеся векторные пространства, где каждый индекс либо векторное пространство соответствует разному раз-

меру фрагмента, и позволить пользователю выбирать, с каким размером фрагмента он хочет работать; однако это потенциально может привести к большой нагрузке на базу данных, поскольку размер используемых данных растет при увеличении количества индексов и векторных пространств.

Описанные выше подходы по фрагментированию данных перед их переводом в эмбеддинги могут быть актуальны не только для текстовых данных, но и для карт, данных мониторинга, изображений и видео; например, часть карты, находящейся в процессе преобразования во встраивания, может быть связана конкретно с Алтайским краем, другая часть может быть связана с Москвой, и эти части могут быть расположены в разных частях векторного пространства.

Плановый процесс загрузки данных и создания эмбеддингов в системе представлен на рис. 3.

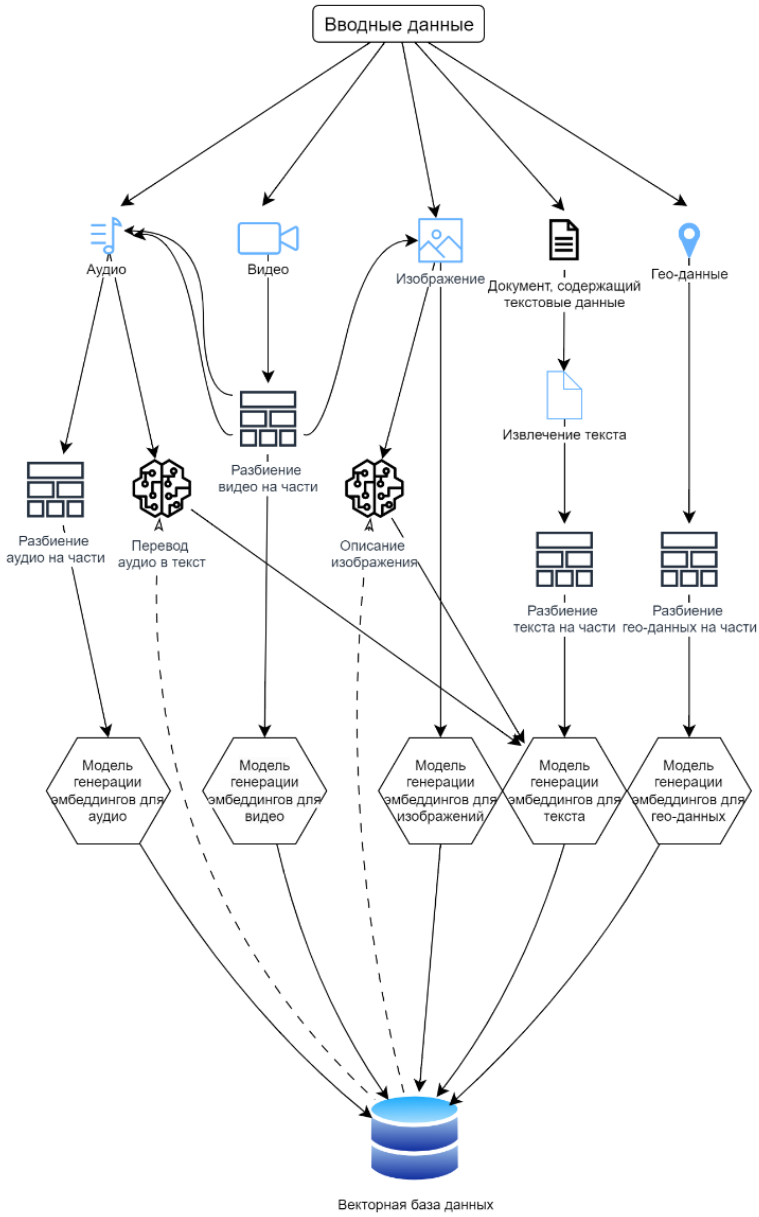

Рис. 3. Целевой вариант процесса загрузки данных и генерации эмбеддингов для дальнейшего извлечения и поиска информации

Fig. 3. The proposed process of data uploading and embeddings generation for further extraction and information retrieval

Заключение

Таким образом, учитывая вышеизложенное, можно сделать вывод, что при проектировании механизма поиска информации для вопросно-ответной системы, обрабатывающей климатические данные, необходимо предоставить конечному пользователю возможность устанавливать уровень шума, определять, как именно извлекать данные, использовать ли ансамблирование, какой тип данных использовать, в том числе выбирать векторное расстояние. Это даст возможность всесторонне протестировать модуль генерации ответов и определить, какие параметры извлечения данных обеспечивают наибольшую корректность сгенерированных ответов.

Список литературы

1. **Hirschman L., Gaizauskas R.** Natural language question answering: the view from here // Natural Language Engineering Journal. 2001. Vol. 7, no. 4. P. 275–300. DOI: 10.1017/S1351324901002807
2. **Keen P. G. W, Michael S. S. M.** Decision support systems: an organizational perspective. Michigan, Addison-Wesley, 1978.
3. **Woods W. A.** Progress in natural language understanding: an application to lunar geology // Proceedings of the national computer conference and exposition (AFIPS ‘73), 1974, Association for Computing Machinery, New York, NY, USA, p. 441–450. DOI: https://doi.org/10.1145/1499586.1499695
4. **Lewis P., Perez E., et al.** Retrieval-augmented generation for knowledge-intensive NLP tasks // Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS ‘20), 2020, Curran Associates Inc., Red Hook, NY, USA, Article 793, p. 9459–9474. DOI: 10.48550/arXiv.2005.11401
5. **Wang L., Lo K. et al.** CORD-19: The COVID-19 Open Research Dataset. ArXiv, abs/2004.10706, 2020. DOI: 10.48550/arXiv.2004.10706
6. **Rajpurkar P., Zhang J., Lopyrev K., Liang P.** Squad: 100,000+ questions for machine comprehension of text // Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, Association for Computational Linguistics, Austin, Texas, USA, p. 2383–2392. DOI: 10.18653/v1/D16-1264
7. **Magesh V., Surani F., Dahl M., Suzgun M. et al.** Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. ArXiv, abs/2405.20362, 2024. DOI: 10.48550/arXiv.2405.20362
8. **Page L., Brin S., Motwani R., Winograd T.** The PageRank Citation Ranking: Bringing Order to the Web. Technical Report, Stanford InfoLab, 1999.
9. **Фадеев С. В.** Экологический словарь. СПб., 2011.
10. **Florin C., Giovanni T. et al:** The Power of Noise: Redefining Retrieval for RAG Systems // Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2024, Association for Computing Machinery, New York, NY, USA, p. 719–729. DOI: 10.1145/3626772.3657834
11. **Cormack G. V., Clarke C. L., Büttcher S.** Reciprocal rank fusion outperforms condorcet and individual rank learning methods // Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, 2009, Association for Computing Machinery, New York, NY, USA, p. 758–759. DOI: 10.1145/1571941.1572114

References

1. **Hirschman L., Gaizauskas R.** Natural language question answering: the view from here. Natural Language Engineering Journal, 2001, vol. 7, no. 4, pp. 275–300. DOI: 10.1017/S1351324901002807
2. **Keen P. G. W, Michael S. S. M.** Decision support systems: an organizational perspective. Michigan, Addison-Wesley, 1978.
3. **Woods W. A.** Progress in natural language understanding: an application to lunar geology. *Proceedings of the national computer conference and exposition (AFIPS '73)*, 1974, Association for Computing Machinery, New York, NY, USA, pp. 441–450. DOI: https://doi.org/10.1145/1499586.1499695
4. **Lewis P., Perez E., et al.** Retrieval-augmented generation for knowledge-intensive NLP tasks. *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20)*, 2020, Curran Associates Inc., Red Hook, NY, USA, Article 793, pp. 9459–9474. DOI: 10.48550/arXiv.2005.11401
5. **Wang L., Lo K. et al.** CORD-19: The Covid-19 Open Research Dataset. ArXiv, abs/2004.10706, 2020. DOI: 10.48550/arXiv.2004.10706
6. **Rajpurkar P., Zhang J., Lopyrev K., Liang P.** Squad: 100,000+ questions for machine comprehension of text. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, Association for Computational Linguistics, Austin, Texas, USA, pp. 2383–2392. doi: 10.18653/v1/D16-1264
7. **Magesh V., Surani F., Dahl M., Suzgun M. et al.** Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. ArXiv, abs/2405.20362, 2024. DOI: 10.48550/arXiv.2405.20362
8. **Page L., Brin S., Motwani R., Winograd T.** The PageRank Citation Ranking: Bringing Order to the Web. Technical Report, Stanford InfoLab, 1999.
9. **Fadeev S. V.** Ekologicheskij slovar'. Saint Petersburg, 2011 (in Russ.)
10. **Florin C., Giovanni T., et al:** The Power of Noise: Redefining Retrieval for RAG Systems. *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024, Association for Computing Machinery, New York, NY, USA pp. 719-729. DOI: 10.1145/3626772.3657834
11. **Cormack G. V., Clarke C. ., Büttcher S.** Reciprocal rank fusion outperforms condorcet and individual rank learning methods. *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, 2009, Association for Computing Machinery, New York, NY, USA, pp. 758–759. DOI: 10.1145/1571941.1572114

Сведения об авторах

Гавенко Ольга Юрьевна, доктор технических наук, кандидат филологических наук, ведущий научный сотрудник Федерального исследовательского центра информационных и вычислительных технологий; старший преподаватель кафедры математического моделирования Новосибирского государственного университета

Шашок Наталья Александровна, аспирант Федерального исследовательского центра информационных и вычислительных технологий

Information about the Authors

Olga Yu. Gavenko, Doctor of Sciences (Technical Sciences), Candidate of Sciences (Philology), Leading Researcher, Federal Research Center for Information and Computational Technologies. Senior lecturer of the Department of Mathematical Modeling, Novosibirsk State University

Natalia A. Shashok, Ph. D Student. Federal Research Center for Information and Computational Technologies

Статья поступила в редакцию 07.12.2024;
одобрена после рецензирования 26.12.2024; принята к публикации 26.12.2024

The article was submitted 07.12.2024;
approved after reviewing 26.12.2024; accepted for publication 26.12.2024.