УДК 007:681.512.2

ФОРМИРОВАНИЕ LLM-ОРИЕНТИРОВАННЫХ РЕСУРСОВ ЗНАНИЙ НА ОСНОВЕ ГЕНЕРАЦИИ ДОПОЛНЕННОЙ ИНФОРМАЦИИ

И. Ю. Каширин, д.т.н., профессор кафедры ВПМ РГРТУ Рязань, Россия;
orcid.org/0000-0003-1694-7410, e-mail: igor-kashirin@mail.ru

Рассматривается новая технология проектирования вопросно-ответных систем, построенных на основе больших генеративных языковых моделей LLM (Large Language Models). Исследуются недостатки LLM, главным из которых является отсутствие в этих моделях актуальных сведений, появившихся в информационных сервисах за сравнительно недавнее время.

Новая технология опирается на современный подход к развитию больших моделей ресурсами новых актуальных или специфических знаний. Такие системы получили название RAG-систем дополненной генерации (Retrieval-Augmented Generation, RAG). Для улучшения качества диалога в них используются дополнительные базы данных. Автором статьи предлагается использовать для генерации новых ресурсов знаний метод расширения семантического пространства при векторизации естественно-языковых текстов. Основой метода является система операций на множестве иерархических чисел, генерируемых в качестве семантических индексов словарных понятий и словарных определений событий. Это дает возможность более точно вычислять семантическую близость словарных конструкций. Новый подход, в частности, может использоваться для специализированных предметных областей.

Программная реализация предложенной технологии получила воплощение в RAG-системе IYuRAG v.1.0. При проектировании был задействован разработанный автором ранее модуль сбора тематических корпусов CorpusMining v.2.1, основанный на инструментарии Googlesearch и BeautifulSoup4 в среде Python v.3.10, Anaconda v.2.1. Кроме того, был применен инструментарий LLM RoBERTa-transformers 4.7.

RAG-система IYuRAG v.1.0 дает возможность генерировать ресурсы знаний в предметной области «Политические новости/Вооруженные конфликты». Вопросно-ответный модуль RAG-системы расширяет возможности существующих LLMs.

***Целью статьи** является презентация нового метода проектирования RAG-систем на основе применения иерархических чисел для расширения семантического пространства в больших нейросетевых генеративных моделях.*

***Ключевые слова**: генерация дополненной информации, эмбеддинги иерархических чисел, нейросетевые трансформеры, анализ естественного языка, онтологические таксономии, семантическое пространство.*

DOI: 10.21667/1995-4565-2025-94-85-97

Введение

Большие нейросетевые языковые модели (Large Language Models, LLM) [1] находятся сейчас в непрерывном и весьма интенсивном развитии. Их функционал ежегодно расширяется. Так, например, эти модели могут почти полностью заменить программистов начинающего уровня (junior) и существенным образом разгружают работу программистов продвинутого уровня (middle) GPT 4,5 [2]. На текущем этапе развития LLM, такие как Gemini 2.5 Pro, дают возможность анализировать и воспроизводить даже динамическое видео [3, 4]. LLM встают в один ряд с поисковиками уровня Google и Yandex, превосходя их по точности ответов на фактографические запросы и возможностям решения генеративных задач.

В отличие от поисковиков большие языковые модели используют эмбеддинги (embeddings), представляющие собой многомерные векторы чисел, в которые преобразуются такие словарные конструкции (токены), как части слов, слова и словосочетания. Эти векторы

имеют в LLM тысячи измерений, формируемых для отражения семантического сходства словарных конструкций [5].

В то же время современные большие языковые вопросно-ответные системы (QA Systems) имеют существенные недостатки:

− обладая поистине гигантскими объемами полезной информации, они совершенно ничего «не знают» о достижениях или просто событиях, возникших относительно недавно;

− они абсолютно бесполезны в узкоспециализированных областях знаний, тем более не могут владеть индивидной или корпоративной информацией;

− при недостатке знаний эти модели генерируют так называемые «галлюцинации», т. е. пытаются нафантазировать что-то очень похожее на правильный квалифицированный ответ;

− эти системы не могут указать источник полученной ими экономической или даже академической информации;

− попытки дообучения таких систем (fine-tuning) приводят не только к большим затратам, но и к информационным сбоям за счет изменения (рассогласования) полученных в результате предварительного обучения числовых эмбеддингов.

Компенсировать перечисленные сложности призваны недавно появившиеся технологии проектирования интеллектуальных RAG-систем дополненной генерации (Retrieval-Augmented Generation, RAG) [6], которые используют для генерации ответов на вопросы дополнительные базы данных, содержащие новую и специализированную информацию. Вопросам проектирования таких систем с применением оригинальной модификации эмбеддингов в целях получения большего эффекта посвящена настоящая статья.

Основные принципы технологии RAG

Существенную и весьма эффективную модификацию LLM получили в новых архитектурах QA систем, использующих в дополнение к уже существующему инструментарию программ-ретриверов [7]. Ретриверы (Retrievers) являются средствами автоматизированного или автоматического поиска фрагментов текста (документов, абзацев, предложений) из большой коллекции данных, релевантных (соответствующих по содержанию) заданному пользователем вопросу.

Архитектурно QA системы состоят из индексатора (Indexer), ретривера (Retriever) и генератора ответа (Answer Extractor). Примерная простейшая архитектурная схема представлена на рисунке 1. Индексатор создает индекс (например, с использованием TF-IDF, BM25 или векторных представлений) для всей коллекции документов. Это позволяет отыскивать релевантные документы. Ретривер получает вопрос пользователя и использует индекс для поиска наиболее подходящих документов или фрагментов текста. Генератор ответа (часто просто возвращающий нужный фрагмент текста) извлекает предполагаемый ответ из найденного текста.

Простейшие RAG-системы принято называть *наивными RAG-системами*. В них при индексации обрабатываются разнотипные документы, такие как doc, HTML, pdf, Excel файлы. При этом используются парсеры и фильтры, выделяющие и очищающие информацию файлов, превращая их в простейший текст, который разбивается на локальные фрагменты строго ограниченного размера (например, 512 байт). Текст векторизуется, преобразуясь в эмбеддинги языковыми LM-моделями, а затем сохраняется в векторной базе данных (ВБД) для дальнейшего информационного поиска по критериям семантической близости [5]. Запрос пользователя также векторизуется, вычисляется его близость к документам (фрагментам текстов) ВБД. Для дальнейшей обработки выделяются лучшие N фрагментов, которые в первоначальном виде не могут служить ответом пользователю. Ответ формируется LLM-генератором, т. е. какой-либо готовой LLM, на вход которой пересылаются найденные фрагменты (chunks) и промпты (подсказки, что именно нужно сгенерировать в качестве ответа). LLM генерирует ответ на основе собственных знаний и найденных ретривером фрагментов, а также промптов.

К недостаткам наивных RAG-систем относятся:

− возможные потери информативных фрагментов;

− частое нарушение связности смысловой организации текста;

− несогласованность или повтор пересекающихся фрагментов ответа;

− включение в результат недостоверных или токсичных данных;

− генерация текстов только из готовых фрагментов, исключая формирование новых ответов;

− возможное отсутствие ответов при отклонении вопроса от известной системе формы.

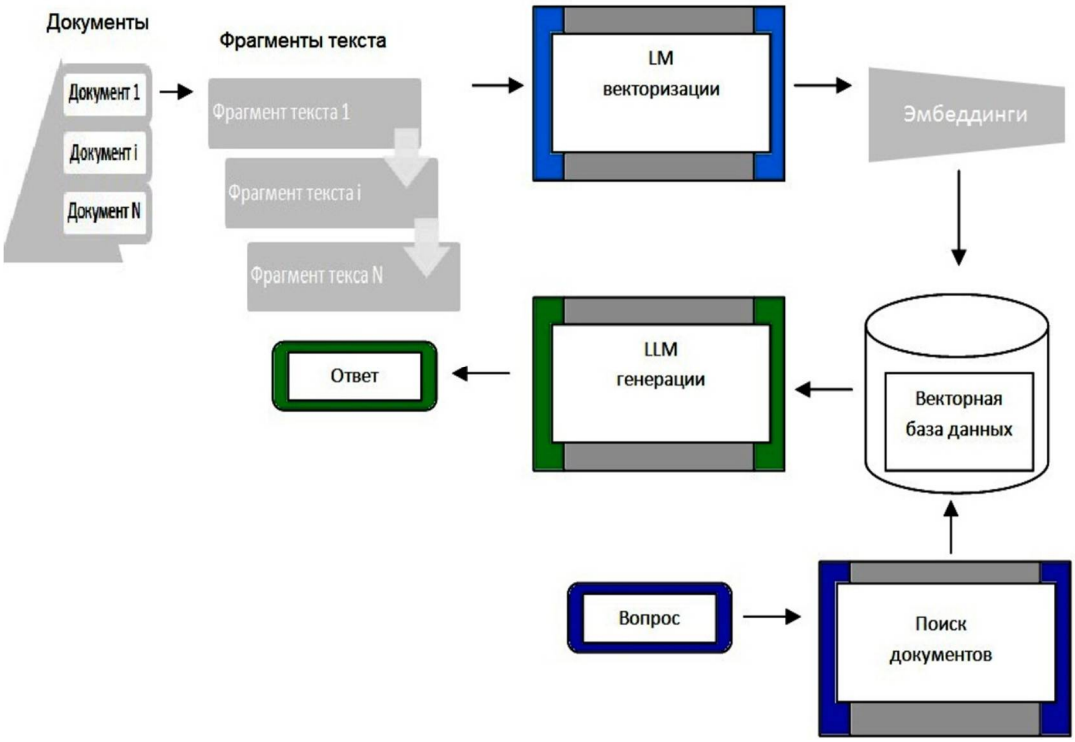

Рисунок 1 – Примерная архитектурная схема простых вопросно-ответных систем
Figure 1 – Approximate architectural diagram of simple question-and-answer systems

Перечисленные недостатки во многом удается преодолеть в более продвинутых интеллектуальных системах, называемых *развитыми RAG-системами* (рисунок 2).

Развитые RAG-системы превосходят наивные системы, используя комбинированный поиск в несколько «проходов». Реализуется предварительный и уточненный поиск. При поиске в них используются следующие технологии [8, 9]:

− применение метаданных (дополнительная информация о данных);

− метод «скользящего окна» (sliding windows), в котором непомещающийся в тензор ретривера текст разбивается на перекрывающиеся фрагменты с определенным малым шагом (stride);

− дробная сегментация (fine-grained segmentation), заключающаяся в точном выделении нужных фрагментов текста на детальном уровне;

− оптимизация всего процесса поиска;

− использование априорных графов знаний, отражающих основные понятия предметной области вопросов/ответов.

Агентные RAG-системы превосходят ранее представленные системы, поскольку используют специальные модули-агенты, позволяющие не только обрабатывать вопросы и генерировать ответы в статических информационных ресурсах, но и учитывать текущую динамику развития событий. Для этого агенты обращаются к внешним сервисам, например Интернет-поисковикам, API-сервисам и сторонним ретриверам. Упрощенная архитектура агентных RAG-систем приведена на рисунке 3.

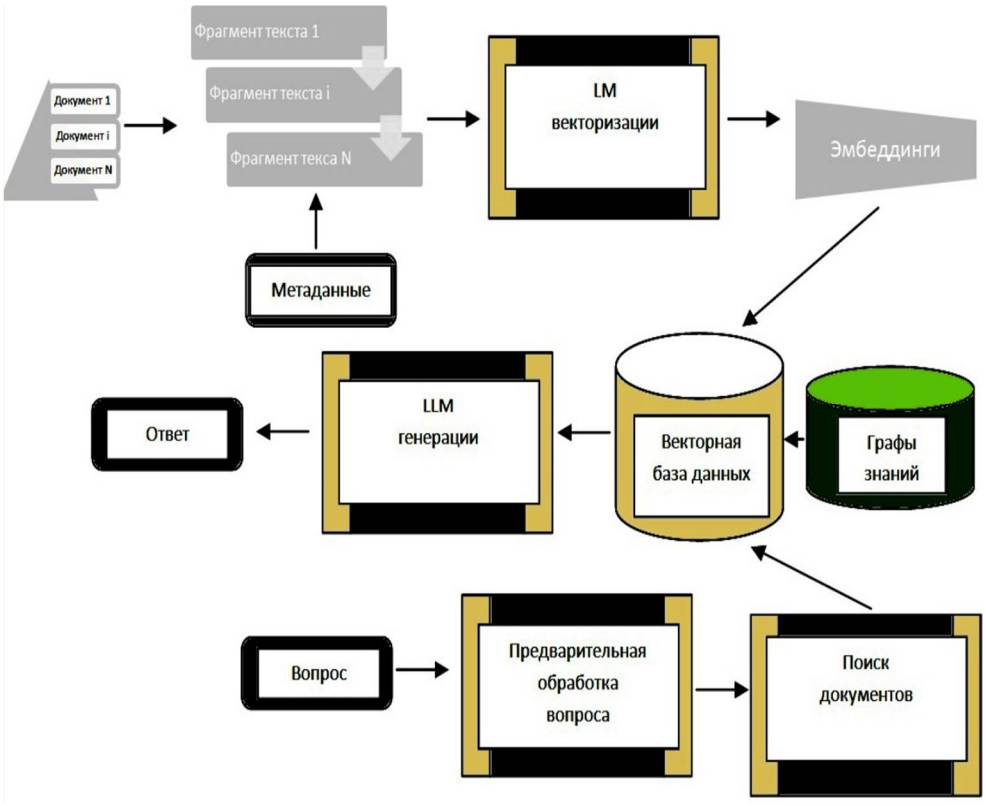

Рисунок 2 – Архитектурная схема развитых RAG-систем
Figure 2 – Architectural diagram of advanced RAG systems

Сравнение основных характеристик наивных и развитых RAG-систем дано в таблице 1.

Таблица 1 – Сравнительные характеристики наивных и развитых RAG-систем
Table 1 – Comparative characteristics of naive and advanced RAG systems

	Наивные (простые) RAG-системы	**Развитые RAG-системы**
Задача системы	Поиск релевантных фрагментов	Генерация нового связного ответа
Главный модуль	Ретривер	Ретривер с генеративной моделью
Результат	Выбранный фрагмент текста	Связные предложения
Применение LLM	Минимальное	Максимальное
Актуальность ответа	Определяется возрастом исходных документов	Используются актуальные данные
Сложность композиции ответа	На уровне нескольких релевантных фрагментов	Возможен сложный связный текст

К отличительным особенностям таких систем можно отнести реализацию следующих функциональных задач:

− итерационное уточнение результатов поиска ответа с повышением релевантности;

− формирование сложных многокомпонентных ответов с разбиением общей задачи поиска на подзадачи;

− рефлексивная самооценка генеративных результатов с исправлением возможных ошибок или неточностей;

− мультиагентный поиск, допускающий общение разных агентов между собой;

− доступ к внешним постоянно меняющимся со временем информационным сервисам.

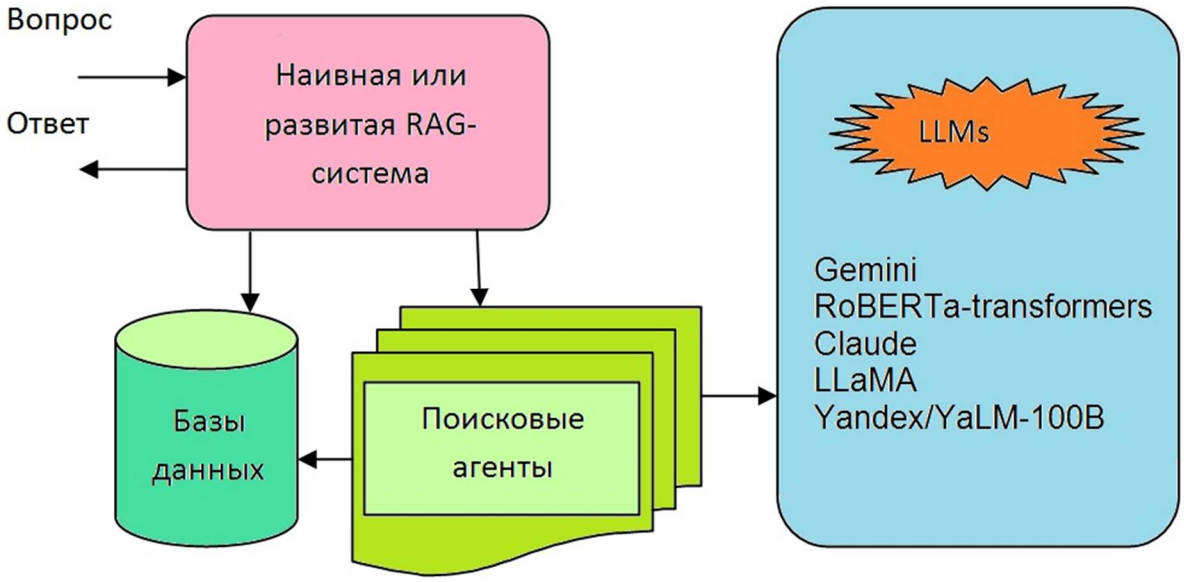

Рисунок 3 – Архитектурная схема агентных RAG-систем
Figure 3 – Architectural diagram of agent-based RAG systems

RAG-системы с дополненными LLM-ориентированными ресурсами знаний

RAG-системы с LLM-ориентированными ресурсами знаний являются еще одной разновидностью современных развитых RAG-систем. Все RAG-системы используют векторное представление документов, отражающее семантическую зависимость между отдельными словами, словосочетаниями и более крупными фрагментами текста. Размеченные документы с выделенными, возможно пересекающимися, фрагментами словарных конструкций становятся интеллектуальными кластерами, хранящими фрагменты ответов на всевозможные вопросы пользователей. Кластеры помещаются в базу данных, используемую в дальнейшем в поиске документов, которые могут служить подходящим материалом для ответа на вопрос. Таким способом производится выбор документов, релевантных не только по ключевым словосочетаниям, но и по семантическому сходству. Накопленный набор документов (база данных) позиционируется как *ресурс знаний*, составляющий содержательную основу RAG-системы.

Кроме больших языковых моделей, представленных на рисунке 3, на сегодняшний день лидерами LLM в открытом доступе, применяемыми в RAG-системах, ориентированных на ресурсы знаний, являются: MaziyarPanahi/calme-3.1-instruct-78b, dfurman/CalmeRys-78B-Orpo-v0.1, huihui-ai/Qwen2.5-72B-Instruct-abliterated, Qwen/Qwen2.5-72B-Instruct [10].

Для формирования ресурса знаний из внешних источников информации обычно используется сложный шаблон запроса, формируемый из вопроса пользователя. Простейшие примеры таких шаблонов приведены в [11]. Отбор релевантных документов, как правило, является сложным многопроходным процессом, состоящим из нескольких этапов, объединенных в единый конвейер. На каждом из этапов выбор уточняется решением таких задач, как:

- согласование и приведение к единому формату различной по типу информации;
- устранение избыточных и пересекающихся фрагментов материала;
- приведение к компактному виду единообразных перечислений, таблиц, статистических сводок и т.п.;
- применение промптов при наличии сложного изначального вопроса пользователя;
- устранение синтаксической и лексической омонимии в формируемом ответе;
- адаптация ответа под индивидуальные особенности пользователя (при использовании моделей пользователей).

Еще одним новым способом повышения релевантности документов для формирования ресурсов знаний, предлагаемым автором настоящей статьи, является дополнение RAG-систем расширенными эмбеддингами словарных конструкций. При этом также возможно применение собственной логики обработки знаний. Такое расширение заключается во введении в семантическое пространство языковой модели новых смысловых признаков, образующих новые размерности для включаемых в него векторов.

Это особенно актуально для предметных областей дополненной (retrieval-augmented) информации, имеющей смысловую специфику связанную, например, с выявлением целевой направленности новостных статей влиятельных международных изданий [11].

При этом не только появляется возможность регулярного обновления информационных ресурсов вопросно-ответных систем и указания ссылок на источники сведений, но и обеспечивается оперативность получения новой информации за счет экономичной адаптации RAG-систем. Действительно, такая технология не требует дорогостоящего дообучения LLM (файнтюнинга). Кроме того, у администраторов информационного ресурса и инженеров по знаниям появляется возможность эффективного выявления фейк новостей и целе-направленной токсичности, введенных авторами материалов при формулировании новых сведений [12].

В предлагаемой технологии расширение эмбеддинга реализуется за счет применения онтологической схемы специализированной предметной области, понятия (словарные конструкции) которой проиндексированы иерархическими числами, впервые введенными в [13]. Для упрощения проектного решения онтологические схемы предлагается ограничить только таксономическими элементами онтологий, в частности, родовидовой и причинно-следственной таксономиями. Подробный пример родовидовой таксономии приведен в [7]. При существовании возможности сюда можно добавить также меронимическую таксономию, базирующуюся на отношении «часть-целое».

Как и в более ранних работах автора статьи [12, 13], примером может быть фрагмент из предметной области «новостные публикации международной политики». Индексирование родовидовой таксономии было рассмотрено ранее в [14].

Фрагмент причинно-следственной таксономии «вооруженный конфликт» представлен рисунком 4. На рисунке элементами «-1» обозначены разделители для композиции иерархических чисел, соответствующих одному событию, а элементами «-2» – разделитель, указывающий на новый уровень иерархии в таксономии. Каждая вершина иерархии отражает какую-либо словарную конструкцию, соответствующую политическому событию. Ребра графа задают причинно-следственное отношение между событиями. К вершинам прикреплены иерархические индексы, дающие возможность с помощью метрического алгоритма рассчитать семантическую близость понятий, соответствующих вершинам. Сами понятия могут быть выражены словосочетанием (сложной словарной конструкцией). Примеры: «Политическая эскалация», «Начало вооруженных столкновений».

Суть метрического алгоритма вычисления семантической близости вершин заключается в определении схожести соответствующих иерархических чисел на основе сравнения их фрагментов.

К любому таксономическому дереву, не имеющему единой корневой вершины, виртуально добавляется предельно абстрактное событие (понятие) «начало сотворения мира».

Схожесть понятий-вершин тем выше, чем ближе в графе таксономии эти вершины расположены по вертикали и горизонтали. Схожесть понятий-вершин понижается тем больше, чем дальше друг от друга эти вершины расположены по горизонтальной и вертикальной составляющим графа таксономии.

Для двух событий вычисляется их ближайший общий предок. События, не имеющие общих предков, считаются независимыми, и величина их семантического сходства приравнивается к нулю.

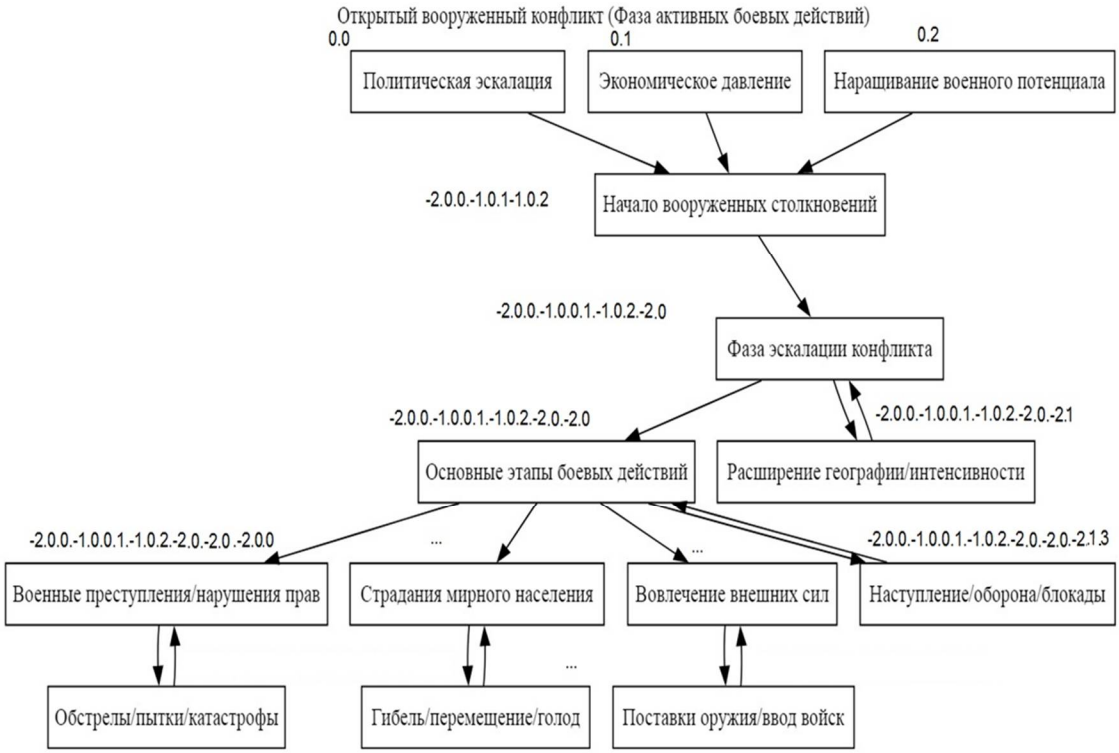

Рисунок 4 – Причинно-следственная таксономия понятий предметной области «Вооруженный конфликт»
Figure 4 – Causal taxonomy of concepts in «Armed Conflict» subject area

Графический смысл величин, предварительно вычисляемых для определения семантического сходства двух событий, представлен на рисунке 5.

Вычисление семантического сходства S для понятий, определяющих события, реализуется следующим алгоритмом, основанным на причинно-следственных таксономиях с иерархическими числами, являющихся индексами вершин таксономии.

1. Вычислить количество G общих вершин-предков для двух сравниваемых понятий (соответствующих вершин).
2. Если G = 0, величина S = 0. Завершить алгоритм. Иначе продолжить решение.
3. Вычислить количество предков до более глубокой общей вершины для сравниваемых вершин, левой и правой (L, R) соответственно.
4. Вычислить разность этих величин (насколько одно событие раньше в таксономии, чем другое) d = |L-R |.
5. Вычислить иерархический индекс вершины, являющейся предком более глубокой вершины из двух сравниваемых, причем для того предка, который расположен на одинаковой глубине от общего предка для менее глубокой вершины.
6. Вычислить количество горизонтальных вершин H, лежащих между вершиной с вычисленным индексом и менее глубокой из изначально сравниваемых вершин. Если это количество равно нулю, то более глубокая вершина является прямым потомком менее глубокой. Тогда менее глубока вершина – общий предок двух вершин.
7. Вычислить числовое значение семантического сходства по формуле:

$$\Delta = (1/(d+1) + 1/(H+1))/2,$$

$$S(L,R,G) = \begin{cases} 1, \Delta = 1, \\ G * \Delta / Max, \Delta \neq 1. \end{cases}$$

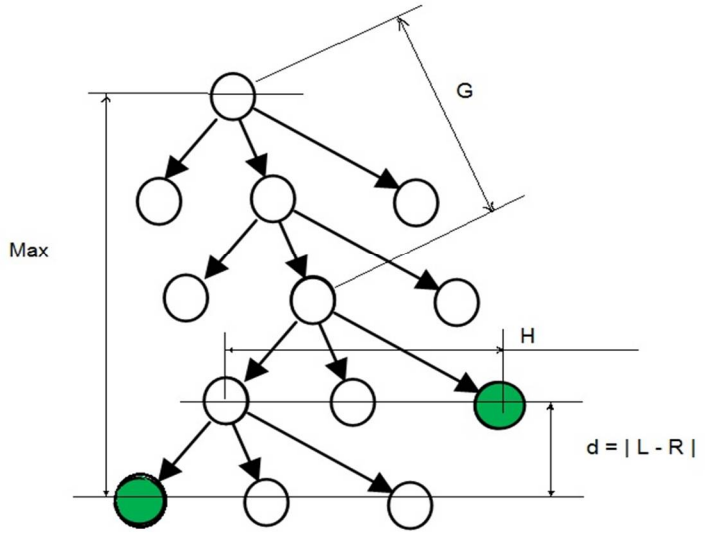

Рисунок 5 – Графический смысл величин, вычисляемых для определения семантического сходства
Figure 5 – Graphical meaning of the values calculated to determine semantic similarity

Бо́льшая часть приведенного алгоритма может быть рассчитана с помощью алгебры иерархических чисел [15].

В качестве примера можно рассмотреть вычисление семантического сходства для двух вершин таксономии, представленной рисунком 4: «Военные преступления/Нарушения прав населения» (иерархический индекс -2.0.0-1.0.0.1.-1.0.2.-2.0.-2.0.0) и «Расширение географии/интенсивности боевых действий» (-2.0.0-1.0.0.1.-1.0.2.-2.0.-2.1).

Обозначим левую вершину как x, а правую как y.

1. Согласно семантике алгебраической операции «º» можно выделить ближайшего общего предка этих вершин:

x º y = -2.0.0-1.0.0.1.-1.0.2.-2.0.-2.0.0 º -2.0.0-1.0.0.1.-1.0.2.-2.0.-2.1 = -2.0.0-1.0.0.1.-1.0.2.-2.0.-2.

Функция «||» вычисления количества вершин от текущей вершины до корневой даст в результате значение величины G: |-2.0.0-1.0.0.1.-1.0.2.-2.0.-2| = 3.

2. Поскольку G ≠ 0, вершины x и y имеют ненулевое семантическое сходство.
3. Используя функцию «||», вычисляем L и R:

L =|-2.0.0-1.0.0.1.-1.0.2.-2.0.-2.0.0| = 5, R = |-2.0.0-1.0.0.1.-1.0.2.-2.0.-2.1| = 6.

4. d = abs(L – R) = 1, где «abs» ─ абсолютное значение.
5. Для вычисления индекса предка вершины x, расположенного на одном уровне с вершиной y, необходимо от x подняться в таксономии на d вершин (алгебраическая операция «--»): -2.0.0-1.0.0.1.-1.0.2.-2.0.-2.0.0-- = -2.0.0-1.0.0.1.-1.0.2.-2.0.-2.0.
6. Функцией «Hor» можно вычислить число горизонтальных вершин между x и y:

Hor (x, y) = Hor(-2.0.0-1.0.0.1.-1.0.2.-2.0.-2.0.0, -2.0.0-1.0.0.1.-1.0.2.-2.0.-2.1) = 1.

Заключительные вычисления производятся в обычной десятичной арифметике:

$$\Delta = (1/(1+1) + 1/(1+1))/2 = 0.5,\ S(5, 6, 3) = 3 * 0.5 / 4 = 0.125,$$

поскольку величина Max = 4 вычисляется как длина пути от корневой вершины до любой из максимально глубоких вершин (например, «Обстрелы/пытки/катастрофы»):

| -2.0.0-1.0.0.1.-1.0.2.-2.0.-2.0.0.-2.0| = 4.

Таким образом, величина семантического сходства между событиями x и y равна 0.125.

Технологическая схема использования такого расширения эмбеддингов в RAG-системе на примере вопроса о работах автора приведена на рисунке 6.

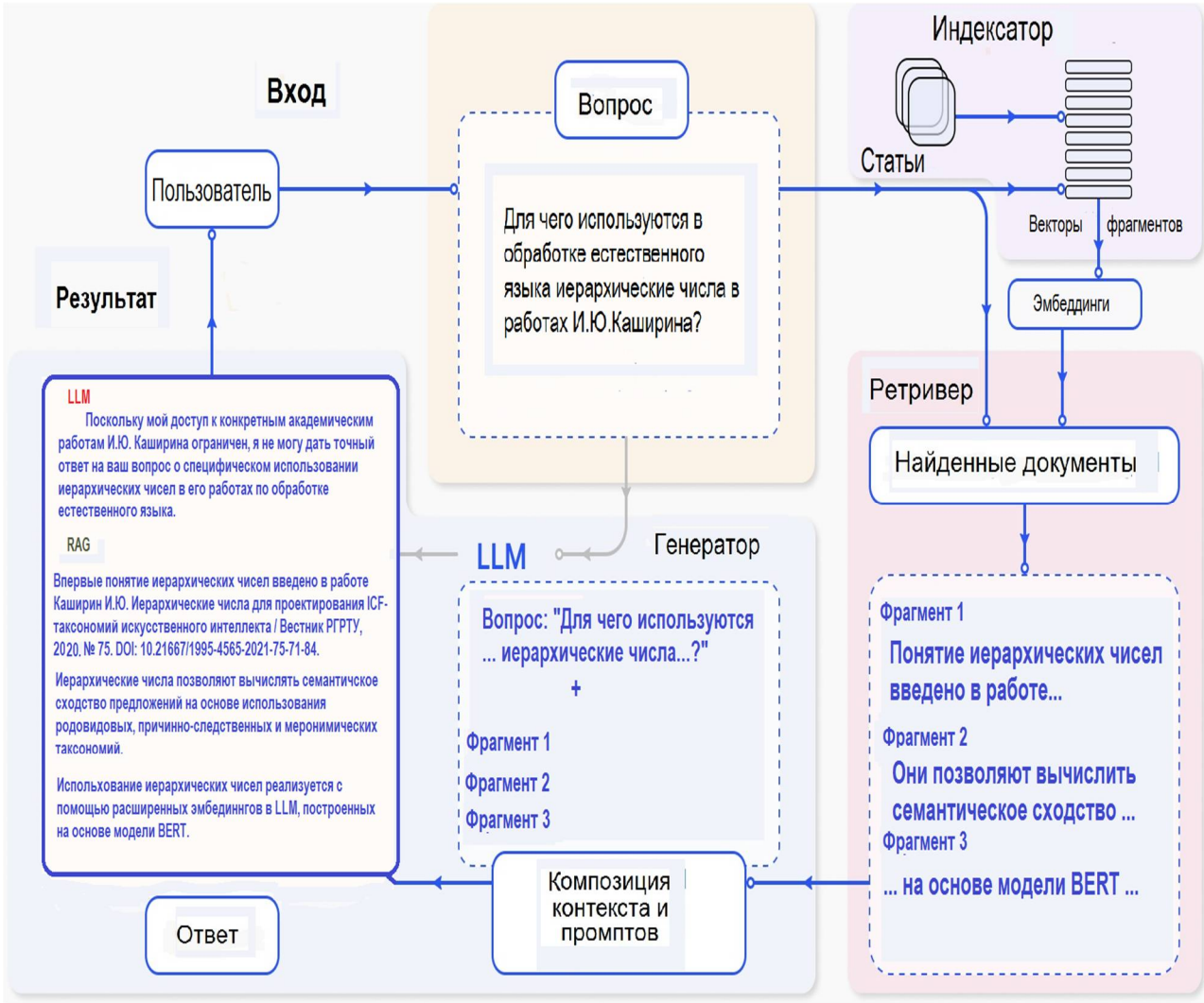

Рисунок 6 – Схема работы дообученной RAG
Figure 6 – Work scheme of the pre-trained RAG

Экспериментальное исследование

Эксперименты проводились на базе обучающего корпуса, состоящего из более 2000 электронных политических публикаций следующих информационных англоязычных сервисов. Были исследованы следующие электронные ресурсы.

Западные журналы: aljazeera.com, bloomberg.com, cnn.com, indianexpress.com, msnbc.com, nytimes.com, politico.com, theguardian.com.

Восточные журналы: en.kremlin.ru, en.globalaffairs.ru, government.ru/en/news/, interfax.com, meduza.io/en, rbth.com, rossiyasegodnya.com, RT.com, sputnikglobe.com, tass.com.

Для сбора обучающего корпуса и последующего ***создания ресурса знаний*** в целях дополнения моделей автором статьи был разработан и использован модуль сбора тематических корпусов CorpusMining v. 2.1, использующий инструментарий Googlesearch и BeautifulSoup4 в среде Python v.3.10, Anaconda v. 2.1. Кроме того, была разработана собственная RAG-подсистема IYuRAG v.1.0, генерирующая ресурсы знаний в предметной области «Политические новости/Вооруженные конфликты». Для вопросно-ответного модуля RAG-системы IYuRAG 1.0 использовалась LLM RoBERTa-transformers 4.7 .

Результаты сравнительного тестирования приведены далее.

Вопрос, заданный 28.09.2025:

«Что нового произошло за неделю в Палестино-Израильском конфликте?»

Сокращенные ответы приведены на рисунке 7.

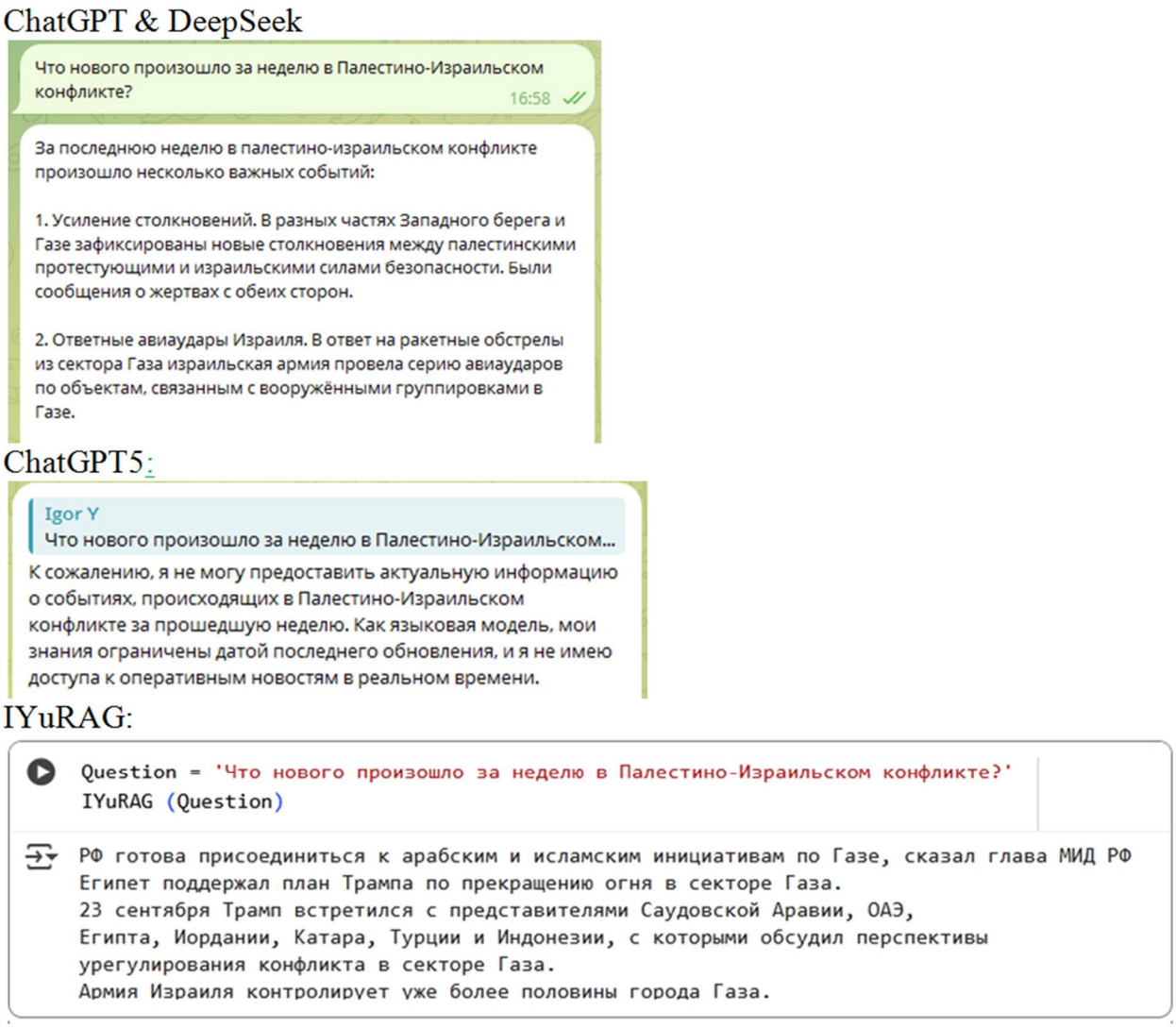

Рисунок 7 – Краткие результаты работы LLM и RAG
Figure 7 – Summary of LLM and RAG results

Как следует из приведенных результатов, ChatGPT & DeepSeek попыталась синтезировать ответ, основанный на общих исторических знаниях о Палестино-Израильском конфликте, поскольку эти последние новости не входили в ее обучающий корпус на момент предобучения. Ответ получился весьма похожим на правду, но далеким от освещения реальных событий на дату посылки запроса. ChatGPT5 честно сообщила о наличии у нее какой-либо новостной информации только на дату последнего обновления ее версии. Следовательно, обе перечисленные большие модели не пользуются online-доступом к текущим внешним информационным ресурсам.

Ответ RAG-системы IYuRAG 1.0 на базе LLM RoBERTa-transformers 4.7 использовал RAG-схему доступа к текущим новостным сайтам, в результате чего был получен приемлемый конкретный результат.

Заключение

Произведенные эксперименты позволяют сделать вывод о высокой эффективности технологии расширения семантического пространства на основе использования таксономий с иерархическими числами. Технология позволяет накапливать актуальные ресурсы знаний для дополнения ими RAG-систем. Результаты апробации RAG-системы IYuRAG 1.0 показали проигрыш большим языковым моделям GPT5 и GPT DeepSeek по времени обработки запросов примерно в два раза. Однако общее абсолютное время ответа не

превышало 1-1,5 минуты, что позволяет считать тестируемую систему применимой на практике в ограниченной предметной области политических новостей.

В то же время при оперативной предварительной настройке IYuRAG 1.0 предоставляет возможность получения более новой информации (например, за прошлую неделю) из новостных электронных ресурсов.

Библиографический список

1. **Minaee1 S., Mikolov T., Nikzad N., Chenaghlu M., Socher R., Amatriain X., Gao J.** Large Language Models: A Survey. arXiv:2402.06196v3 [cs.CL] 23 Mar 2025.

2. **Thompson A.D.** GPT-4.5 [Electronic resource]. Update date: January 2024, February 2025. URL: https://lifearchitect.ai/gpt-4-5/ (date of application: 18.08.2025).

3. **Baddepudi A. Y., Lučić M**. Advancing the frontier of video understanding with Gemini 2.5, 2025. URL https://developers.googleblog.com/en/gemini-2-5-video-understan ding/.

4. **Fu C., Dai Y., Luo Y., Li L., Ren S., Zhang R., Wang Z., Zhou C., Shen Y., Zhang M.,** et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 24108–24118, 2025. URL https://openaccess.thecvf.com/content/CVPR2024/html/Fu_Video-MME_The _First-Ever_Comprehensive_ Evaluation_Benchmark_of_Multi-Modal_LLMs_in_C VPR_2024_paper.html.

5. **Miller G.A., Charles W.G.** Contextual Correlates of Semantic Similarity. Language and Cognitive Processes. 1991. Vol. 6(1). Pp. 1-28.

6. **Bang A., Zhang S., Dredze M.** RAG LLMs are Not Safer: A Safety Analysis of Retrieval-Augmented Generation for Large Language Models. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies 2025. Vol. 1: Long Papers. Pp. 5444-5474, Albuquerque, New Mexico. Association for Computational Linguistics.

7. **Каширин И.Ю**. Токенизация политических текстов в BERT-моделях с использованием ICF+-онтологий // Информационные технологии. 2024. Т. 30. № 12. С. 622-632. DOI: 10.17587/ it.30.622-632

8. **Xu R., Yu Y., Ho J., Yang C.** «Weakly-supervised scientific document classification via retrieval-augmented multi-stage training». In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval: 2023. Pp. 2501-2505.

9. **Hu M., Zhao X., Wei J., Wu J., Sun X., Li Z., Zhang Y.** «rT5: A Retrieval-Augmented Pre-trained Model for Ancient Chinese Entity Description Generation» In International Conference on NLP and Chinese Computing, Cham: Springer (2023). Pp. 736-748.

10. **Huggingface.** Open LLM Leaderboard Archived. [Electronic resource] Update date: January 2024, February 2025. URL: https://huggingface.co/spaces/ open-llm-leaderboard/open_llm_leaderboard#/ (date of application: 12.09.2025).

11. **Каширин И.Ю.** Извлечение фактов из естественно-языковых текстов методом унификации семантических паттернов. Вестник Рязанского государственного радиотехнического университета. 2025. № 91. С. 36-49. DOI: 10.21667/1995-4565-2025-91-36-49.

12. **Каширин И.Ю.** Нейросети нового многополярного мира: классификация электронных новостей. Вестник Рязанского государственного радиотехнического университета. 2024. № 87. С. 29-40. DOI: 10.21667/1995-4565-2024-87-29-40.

13. **Каширин И.Ю.** Иерархические числа для проектирования ICF-таксономий искусственного интеллекта. Вестник Рязанского государственного радиотехнического университета. 2020. № 71. С. 71-82. DOI: 10.21667/1995-4565-2020-71-71-82.

14. **Каширин И.Ю.** Векторизация текста на основе ICF+ онтологии в ансамблях моделей машинного обучения для классификации электронных ресурсов. Вестник Рязанского государственного радиотехнического университета. 2024. № 90. С. 41-53. DOI: 10.21667/1995-4565-2024-90-41-53.

15. **Kashirin I.Yu.** Theory of hierarchical numbers in calculation problems semantic similarity of natural language constructions. [Electronic resource]. Update date: January 2024, February 2025.URL: https://kashirin.net/Theory_of_hierarchical_numbers/ (date of application: 29.09.2025).

UDC 007:681.512.2

FORMATION OF LLM-ORIENTED KNOWLEDGE RESOURCES BASED ON GENERATION OF AUGMENTED INFORMATION

I. Yu. Kashirin, Dr. in technical sciences, full professor, RSREU, Ryazan, Russia;
orcid.org/0000-0003-1694-7410, e-mail: igor-kashirin@mail.ru

A new technology for designing question-answering systems based on large generative language models (LLMs) is being considered. The disadvantages of LLMs, the main being the lack of up-to-date information that has appeared in information services in relatively recent time, are investigated.

The new technology is based on a modern approach to the development of large models with the resources of new up-to-date or specific knowledge. Such systems are called RAG systems of augmented generation (Retrieval-Augmented Generation, RAG). Additional databases are used to improve the quality of dialogue in these systems. The author of the article suggests using the method of expanding the semantic space for vectorization of natural language texts to generate new knowledge resources. The method is based on a system of operations on a set of hierarchical numbers generated as semantic indices of dictionary concepts and dictionary definitions of events. This makes it possible to more accurately calculate the semantic proximity of dictionary constructions. The new approach can be used for specialized subject areas.

Software implementation of the proposed technology has been implemented in IYuRAG v.1.0 RAG system. The design involved the previously developed CorpusMining v.2.1 module for collecting thematic corpora, which is based on Googlesearch and BeautifulSoup4 tools in Python v.3.10 and Anaconda v.2.1. In addition, LLM RoBERTa-transformers toolkit was used.

IYuRAG v.1.0 RAG system provides an opportunity to generate knowledge resources in «Political news/ Armed conflicts» domain. RAG system's question-and-answer module enhances the capabilities of existing LLMs.

The aim of this article is to present a new method for designing RAG systems based on the use of hierarchical numbers to expand the semantic space in large neural network generative models.

Keywords*: augmented information generation, hierarchical number embeddings, neural network transformers, natural language analysis, ontological taxonomies, and semantic space.*

DOI: 10.21667/1995-4565-2025-94-85-97

REFERENCES

1. **Minaee1 S., Mikolov T., Nikzad N., Chenaghlu M., Socher R., Amatriain X., Gao J.** Large Language Models: A Survey. *arXiv:2402.06196v3 [cs.CL] 23 Mar 2025*.

2. **Thompson A.D.** GPT-4.5 [*Electronic resource*]. Update date: January 2024, February 2025. URL: https://lifearchitect.ai/gpt-4-5/ (date of application: 18.08.2025).

3. **Baddepudi A.Y., Lučić M.** Advancing the frontier of video understanding with Gemini 2.5, 2025. *URL https://developers.googleblog.com/en/gemini-2-5-video-understan ding/*.

4. **Fu C., Dai Y., Luo Y., Li L., Ren S., Zhang R., Wang Z., Zhou C., Shen Y., Zhang M.**, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *In Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025, pp. 24108-24118, URL https: // openaccess.thecvf.com/content/CVPR2024/html/Fu_Video-MME_The_First-Ever_Comprehensive_Evaluation_Benchmark_of_Multi-Modal_LLMs_in_C VPR_2024_paper.html.

5. **Miller G.A., Charles W.G.** Contextual Correlates of Semantic Similarity. *Language and Cognitive Processes*. 1991, vol. 6(1). pp. 1-28.

6. **Bang A., Zhang S., Dredze M.** RAG LLMs are Not Safer: A Safety Analysis of Retrieval-Augmented Generation for Large Language Models. *In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies.* 2025, vol. 1, Long Papers, pp 5444-5474, Albuquerque, New Mexico. Association for Computational Linguistics.

7. **Kashirin I.Yu.** Tokenizaciya politicheskikh tekstov v BERT-modelyakh s ispol'zovaniem ICF+-ontologij. *Informacionnye tekhnologii*. 2024, vol. 30, no. 12, pp. 622-632. DOI: 10.17587/it.30.622-632

8. **Xu R., Yu Y., Ho J., Yang C.** «Weakly-supervised scientific document classification via retrieval-augmented multi-stage training». *In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval:* 2023, pp. 2501-2505.

9. **Hu M., Zhao X., Wei J., Wu J., Sun X., Li Z., Zhang Y.** «rT5: A Retrieval-Augmented Pre-trained Model for Ancient Chinese Entity Description Generation» *In International Conference on NLP and Chinese Computing, Cham: Springer*. 2023, pp. 736-748.

10. **Huggingface**. Open LLM Leaderboard Archived. [*Electronic resource*] Update date: January 2024, February 2025.*URL: https://huggingface.co/spaces/ open-llm-leaderboard/open_llm_leaderboard#/ (date of application: 12.09.2025)*.

11. **Kashirin I.Yu.** Izvlechenie faktov iz estestvenno-yazykovykh tekstov metodom unifikacii semanticheskikh patternov. *Vestnik Ryazanskogo gosudarstvennogo radiotekhnicheskogo universiteta*. 2025, no. 91, pp. 36-49. DOI: 10.21667/1995-4565-2025-91-36-49. (in Russia).

12. **Kashirin I.Yu**. Nejroseti novogo mnogopolyarnogo mira: klassifikaciya ehlektronnykh novostej. *Vestnik Ryazanskogo gosudarstvennogo radiotekhnicheskogo universiteta*. 2024, no. 87, pp. 29-40. DOI: 10.21667/1995-4565-2024-87-29-40. (in Russia).

13. **Kashirin I.Yu.** Ierarkhicheskie chisla dlya proektirovaniya ICF-taksonomij iskusstvennogo intellekta. *Vestnik Ryazanskogo gosudarstvennogo radiotekhnicheskogo universiteta*. 2020, no. 71, pp. 71-82. DOI: 10.21667/1995-4565-2020-71-71-82. (in Russia).

14. **Kashirin I.Yu.** Vektorizaciya teksta na osnove ICF+ ontologii v ansamblyakh modelej mashinnogo obucheniya dlya klassifikacii ehlektronnykh resursov. *Vestnik Ryazanskogo gosudarstvennogo radiotekhnicheskogo universiteta.* 2024, no. 90, pp. 41-53. DOI: 10.21667/1995-4565-2024-90-41-53. (in Russia).

15. **Kashirin I.Yu.** Theory of hierarchical numbers in calculation problems semantic similarity of natural language constructions. [*Electronic resource*]. Update date: January 2024, February 2025.URL: https://kashirin.net/Theory_of_hierarchical_numbers/ (date of application: 29.09.2025).