

УДК 528.9:004.891.2

DOI 10.33764/2411-1759-2025-30-5-33-40

Формирование картографических материалов по текстам тревелогов с использованием механизмов генерации с дополненной выборкой

А. А. Колесников¹✉, С. С. Жданов¹

¹ Новосибирский государственный технический университет, языковой центр «Лингва»,
г. Новосибирск, Российская Федерация

e-mail: alexeykw@mail.ru

Аннотация. Предложено решение проблемы ограничений в ответах больших языковых моделей относительно актуальных и специализированных данных посредством использования методов генерации с дополненной выборкой. Рассмотрены существующие подходы и технические решения данного варианта построения запросов к базам данных, а также примеры использования в задачах картографирования маршрутов из тревелогов, их анализа и проверки на основе существующих картографических и геоинформационных материалов. Описан механизм преобразования имеющихся картографических материалов в текстовое описание посредством использования типовых шаблонов для разных видов объектов, где отдельные параметры заполняются на основе конкретных текстовых и числовых значений, получаемых из геометрических характеристик и атрибутивных значений объектов векторной карты. Выполнена апробация предлагаемого подхода на основе геоинформационной системы QGIS, векторной базы данных FIASS, векторизатора и большой языковой модели YandexGPT.

Ключевые слова: тематическая картография, языковая модель, тревелог, генерация с дополненной выборкой, машинное обучение

Для цитирования:

Колесников А. А., Жданов С. С. Формирование картографических материалов по текстам тревелогов с использованием механизмов генерации с дополненной выборкой // Вестник СГУГиТ. – 2025. – Т. 30, № 5. – С. 33–40. – DOI 10.33764/2411-1759-2025-30-5-33-40

Введение

Несмотря на то что основная часть картографических материалов представлена в графическом виде, современные технологии обработки текста могут найти свое применение в области картографии и геоинформатики. Среди примеров такого использования можно привести построение запросов на естественном языке для получения координат объектов, расстояний и площадей, построения маршрутов,

географического описания территорий [1, 2], поиска инструментов для выполнения определенных действий в ГИС [3], построения SQL запросов и программного кода [4], структуризации и формализации процессов обработки пространственных данных [5]. Технически данный подход реализуется сейчас в большинстве случаев посредством применения больших языковых моделей, что позволяет выполнять эти операции максимально быстро, составляя запросы на естественном языке, при этом за-

частую нет необходимости быть специалистом в области геоинформационных технологий. Основная проблема этого подхода заключается в том, что большие языковые модели обучаются на определенных корпусах текстов, которые ограничены по объему и временному интервалу сбора [6]. То есть в результатах ответа на запрос не получить самую актуальную информацию (здесь также влияет достаточно длительное время самого процесса обучения) и ту, которая не попала в изначальный набор документов, например, текущие значения статистических параметров (численность населения определенного города за последний месяц). Для решения данной проблемы используются методы генерации с дополненной выборкой (англ. Retrieval-Augmented Generation, RAG) [7]. Они позволяют объединить поиск информации из обычных баз данных с генеративными языковыми моделями для предоставления точных и контекстно обоснованных ответов, в том числе с использованием распределенных систем хранения. В данной статье рассматриваются существующие подходы и технические решения данных процессов, а также примеры их использования применительно к задачам картографирования маршрутов из травелогов, последующего анализа и проверки с использованием существующих картографических материалов, геоинформационных баз данных. Интеграция технологий обработки естественного языка в виде больших языковых моделей в процессы картографирования маршрутов по текстам травелогов представляет собой инновационные технические решения в обоих направлениях исследований. Этот синтез позволяет улучшить интерпретацию и использование данных о путешествиях, обеспечивая лучшие способы графического представления маршрутов и более глубокий анализ повествования, а также, с другой стороны, получить возможность представления картографических и геоинформационных материалов в виде текстовых описаний.

Методы и материалы

При автоматизации анализа текстов травелогов большие языковые модели могут улуч-

шить распознавание элементов маршрутов путем использования процесса извлечения именованных сущностей. Их применение необходимо для получения значимых данных из путевых заметок, таких как местоположения, виды деятельности, даты, транспорт, впечатления и т. д. Последние исследования показывают, что традиционные методы распознавания именованных сущностей часто недостаточно качественны при разборе сложных ситуаций, связей между понятиями в пределах абзаца. Naraki и др. [8] предлагают гибридный подход к аннотированию, который сочетает разметку специалистами с возможностями больших языковых моделей. Их исследование демонстрирует, что предлагаемый метод значительно повышает качество извлечения именованных сущностей, одновременно решая такие проблемы, как шумовые аннотации и дисбаланс классов. Используя большие языковые модели, этот гибридный метод не только улучшает качество анализа наборов данных травелогов, но и представляет собой менее ресурсозатратное решение проблем, присущих ручному аннотированию. Wang и др. [9] представляют GPT-NER, который адаптирует большую языковую модель для последовательной обработки задач поиска именованных сущностей. Эта модель преобразует данную задачу в процесс генерации, с которой большие языковые модели могут справиться более компетентно. Их выводы показывают, что GPT-NER превосходит традиционные методы, особенно в условиях ограниченных обучающих выборок, что достаточно актуально для текстов травелогов. Извлечение информации является важным компонентом в преобразовании неструктурированных данных описаний путешествий в формализованные структуры, которые можно легко анализировать и использовать. Zhichao и др. [10] исследуют производительность больших языковых моделей для китайского языка с открытым исходным кодом в задачах анализа текстов, уделяя особое внимание их возможностям в условиях нулевого и малого количества попыток. Их сравнительный анализ подчеркивает сильные стороны и ограничения моделей с открытым исходным кодом по сравнению с коммерческими вариантами, такими как ChatGPT,

YandexGPT, GigaChat, предоставляя информацию о том, как эти технологии могут использоваться в реальных приложениях для путешествий. Результаты этого исследования показывают, что в то время как основные большие языковые модели адекватно работают на английском языке, проблемы, возникающие в других языках, таких как китайский, могут повлиять на эффективность извлечения данных тревелогов. Это ставит важные вопросы об адаптации LLM к многоязычным данным о путешествиях и о том, как их можно оптимизировать для улучшения общего качества обработки за счет эффективного извлечения информации, в том числе и на основе использования механизмов RAG.

Структура RAG состоит из элементов Retriever, отвечающих за извлечение соответствующих документов или знаний из большого внешнего корпуса документов, и Generator, представляющего собой языковую модель, синтезирующую контекстно зависимый ответ из извлеченной информации и входного запроса [11, 12].

Retriever обычно основан на методах векторного или токенового поиска. Его роль заключается в поиске фрагментов информации (представленного как частями текстов, так и полными документами), которые релевантны заданному запросу. Векторный поиск базируется на числовых последовательностях (эмбедингах), которые сгенерированы из запросов и документов, сходство которых измеряется с помощью метрики расстояния между символами (например, косинусное сходство) [13]. С точки зрения пространственных данных вектора могут формироваться как из текстовых атрибутов, так и из преобразованных в строки представлений координат, пространственных связей, числовых атрибутов. Токенный поиск базируется на сравнении отдельных слов, например, используя методы TF-IDF или BM25. В результате поиска выводится набор (с заданным ранее определенным максимальным числом) документов или отрывков текста, которые затем передаются генератору. Generator обычно представляет собой большую языковую модель, которая генерирует связный ответ на основе как входного запроса, так и извлеченных документов.

Выделяют три основных типа реализации RAG: базовый, расширенный и модульный. Для реализации базового использования RAG требуется сформировать наборы данных, к которым будут выполняться запросы, преобразовать в вектора (создать эмбединги) и загрузить их в векторную базу данных, задействовать механизм извлечения релевантных элементов на основе разбора текстовых запросов и встроить их в ответы, генерируемые большой языковой моделью, оценить качество полученного результата. Расширенный вариант отличается от базовой схемы наличием этапов предварительной и постобработки извлеченных данных. Предварительная обработка подразумевает улучшение степени релевантности полученного из векторной базы данных набора блоков текста за счет удаления нерелевантной информации, устранения двусмысленности и неточностей, сохранения контекста, обновления устаревших документов и т. д. Кроме этого могут возникать ситуации, когда исходный запрос пользователя может быть не лучшим вариантом для подачи в большую языковую модель, поэтому используются методы переписывания запросов на основе их характеристик для улучшения качества итоговой генерации. Этап постобработки улучшает качество объединения извлеченных данных с запросом пользователя с точки зрения устранения возможного превышения ограничения окна контекста, внесения шума и ухудшения качества ответа. Для решения этих проблем используются повторное ранжирование результатов и сжатие контекста. Повторное ранжирование переупорядочивает извлеченные блоки текста на основе контекстного сходства между содержимым и запросом пользователя. Метод сжатия контекста позволяет уменьшать шум путем сокращения нерелевантной информации и общей длины строк, выделения наиболее важных отрывков. Модульный подход объединяет первые два и делает каждый из этапов независимым от конкретной программной реализации, позволяя использовать разные элементы конвейера обработки от разных разработчиков, дополнительно расширяя функциональность за счет рекурсивного извлечения, использования подзапросов, генерации синтетических ответов [14–17].

Такие подходы являются универсальными, основным отличием между ними будет способ векторизации исходных данных. В статье предлагается подход обработки картографических материалов в механизмах RAG на примере обработки данных травелогов.

Результаты

Основываясь на рассмотренных выше методах RAG, предлагается три возможных варианта использования при обработке данных травелогов:

- 1) формирование базы данных, позволяющей реализовать извлечение нужных сведений из текстовых данных совместно с запросами в геоинформационной системе;
- 2) выполнение сравнения характеристик и текстового описания определенных объектов (например, идентифицируемых по наз-

ванию) для поиска несоответствий между двумя базами исходных текстов и картографической информацией, преобразованной в текст;

3) поиск наиболее вероятного местоположения на основе сравнения текстового описания и картографической информации, преобразованной в текст.

Для второго и третьего вариантов требуется преобразовать имеющиеся картографические материалы в текстовое описание. Данный процесс выполняется посредством использования типовых шаблонов для разных видов объектов, где отдельные параметры заполняются на основе конкретных текстовых и числовых значений, получаемых из геометрических характеристик и атрибутивных значений объектов векторной карты в ГИС. Итоговая схема предлагаемого подхода представлена на рисунке.

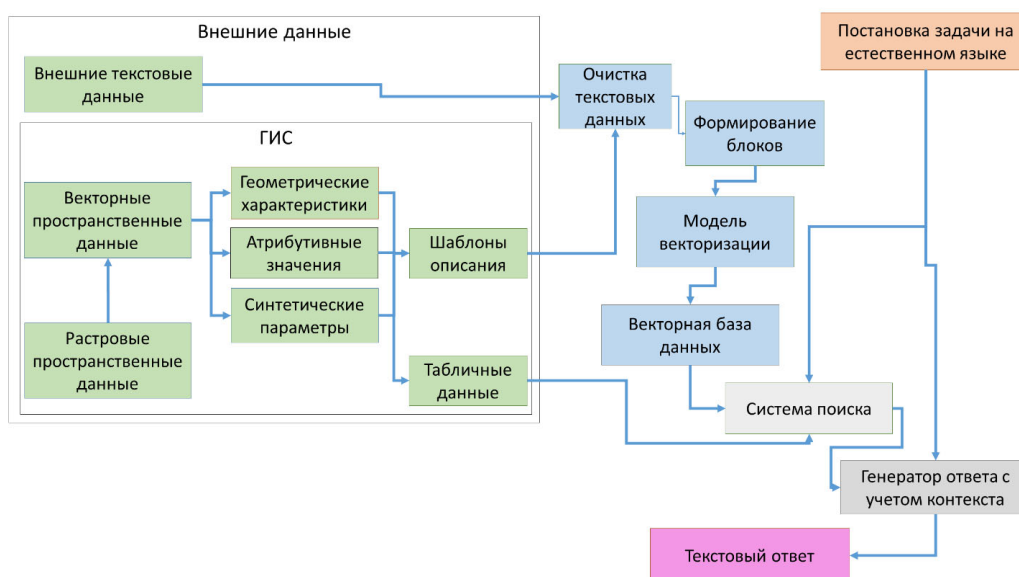


Схема использования RAG при совместной обработке текстовых и пространственных данных

На основе анализа текстов предлагаются следующие наборы элементов описания объектов карты:

- населенные пункты: название, тип, дата основания, численность, преобладающий тип строений, преобладающая этажность, основные значимые объекты;
- предприятия и объекты сервиса: название, адрес, тип продукции;

– реки: название, протяженность, основные объекты на берегах, судоходность, перепад глубин, характер относительно основного рельефа;

– озера: название, площадь, периметр, основные объекты на берегах, судоходность, перепад глубин, соленость;

– леса: название, площадь, преобладающий тип растительности, тип геометрии;

– крупные объекты рельефа: название, тип объекта, площадь, длина, перепад высот/глубин, классификация на основе ТРІ;

– дорожная сеть: название, тип объекта, типы транспорта, протяженность, основные объекты, граничащие с элементом дорожной сети.

Примерами таковых шаблонов могут быть фразы: ориентирами будет <объект>, находящийся на севере на расстоянии <расстояние> метров, и <объект>, находящийся на западе на расстоянии <расстояние> метров; река <название> имеет <тип рельефа> характер, протяженность <длина объекта> километров (где значения параметров, обозначенные в треугольных скобках, вычисляются посредством геоинформационной системы; предлагаемые шаблоны размещены в репозитории по адресу <https://github.com/AlexeyKW/travelogs>).

Для апробации предлагаемого подхода реализовано программное обеспечение на основе геоинформационной системы QGIS, векторной базы данных FIASS, векторизатора и большой языковой модели YandexGPT. Предлагаемое решение позволило реализовать функциональность запросов на естественном языке применительно к имеющейся базе данных травелогов и векторным картографическим материалам, выполнять сравнительный анализ (по метрикам BLEU, ROUGE, SemScore) текстового описания отдельных объектов, идентифицируемых по их названию, координировать фрагменты текстового описания на основе сравнения с существующими текстовыми описаниями картографических объектов.

Заключение

На основе опыта использования RAG-технологий для картографирования текстов травелогов можно сделать общий вывод о том, что они позволяют автоматизировать (на величину порядка 40 %) ряд процессов

сравнительного анализа текстовых и картографических данных. Этот вариант позволяет сочетать преимущества больших языковых моделей, а также точные и актуальные значения из геоинформационных баз данных.

Предложенные три варианта использования RAG в виде построения запросов на естественном языке, выполнения сравнительного анализа и координирования текстового описания объекта позволили автоматизировать процессы верификации и анализа текстов травелогов. Разработанные шаблоны генерации описания местности и отдельных объектов позволяют обрабатывать картографические данные большими языковыми моделями.

Результаты проделанной работы свидетельствуют о том, что необходимо провести дальнейшие исследования в области тонкой настройки (fine-tuning) языковых моделей и их использования для мультивременных геоинформационных баз данных с целью отслеживания изменений топонимов с течением времени, формирования теоретического (интерполяционного) местоположения объектов местности по их контекстному описанию в тексте.

Статья подготовлена в рамках выполнения гранта РНФ № 24-28-01431 «Репрезентация пространства Украины в русской культуре конца XVIII – XIX веков (на материале отечественных травелогов): дискурсы, нарративы, топосы» (<https://rscf.ru/project/24-28-01431/>), реализуемого по направлению Стратегии научно-технологического развития Российской Федерации (утверждена Указом Президента Российской Федерации от 1 декабря 2016 г. № 642 «О Стратегии научно-технологического развития Российской Федерации») «Н5 – Противодействие техногенным, биогенным, социокультурным угрозам, терроризму и идеологическому экстремизму, а также киберугрозам и иным источникам опасности для общества, экономики и государства».

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Mai G., Huang W., Sun J., Song S., Mishra D., Liu N., Gao S., Liu T., Cong G., Hu Y., Cundy C., Li Z., Zhu R., Lao N. On the Opportunities and Challenges of Foundation Models for GeoAI (Vision Paper) // ACM Trans. Spatial Algorithms Syst, 2024. – 10(2)-11. – P. 46. – DOI 10.1145/3653070.

2. Jonathan R., Luddecke T., Das S., Han K., Albanie S. GPT4GEO: How a Language Model Sees the World's Geography [Electronic resource] // ArXiv abs/2306.00020, 2023. – URL: <https://arxiv.org/abs/2306.00020> (дата обращения: 05.11.2024).
3. Li Z., Ning H. Autonomous GIS: the next-generation AI-powered GIS // International Journal of Digital Earth, 2023. – 16 (2). – P. 4668–4686. – DOI 10.1080/17538947.2023.2278895.
4. Ran T., Xu J. Mapping with ChatGPT // ISPRS International Journal of Geo-Information. 2023. – Vol. 12. – No. 7. – P. 284. – DOI 10.3390/ijgi12070284.
5. Sambandam P., Yuvaraj D., Padmakumari P., Swaminathan S. Deep attention based optimized Bi-LSTM for improving geospatial data ontology // Data & Knowledge Engineering. – 2023. – Vol. 144. – P. 102123. – DOI 10.1016/j.datak.2022.102123.
6. Скрынникова И. В. Интерпретационные проблемы LLM: загадки образного языка // Когнитивные исследования языка. – 2024. – № 2-2 (58). – С. 638–642.
7. Lewis P., Piktus A., Petroni F. Retrieval-augmented generation for knowledge-intensive NLP tasks // Advances in Neural Information Processing Systems (Virtual, Online, 06–12 декабря 2020 г.). – P. 34.
8. Naraki Y., Yamaki R., Ikeda Y., Horie T., Naganuma H. Augmenting NER Datasets with LLMs: Towards Automated and Refined Annotation // ArXiv abs:2404.01334v1 [cs.CL], 2024. – URL: <https://arxiv.org/abs/2404.01334> (дата обращения: 05.11.2024).
9. Wang S., Sun X., Li X., Ouyang R., Wu F., Zhang T., Li J., Wang G. GPT-NER: Named Entity Recognition via Large Language Models // ArXiv abs: 2304.10428, 2023. – URL: <https://arxiv.org/abs/2304.10428> (дата обращения: 05.11.2024).
10. Zhichao Z., Zhao Q., Li J., Ge Y., Ding X., Gu T., Zou J., Lv S., Wang S., Yang J. Comparative Analysis of Large Language Models in Chinese Medical Named Entity Recognition // Bioengineering, 2024. – Vol. 11. – No. 10. – P. 982. – DOI 10.3390/bioengineering11100982.
11. Ja P. Automating Threat Intelligence Analysis with Retrieval Augmented Generation (RAG) for Enhanced Cybersecurity Posture // International Journal of Science and Research, 2024. – Vol. 13, No. 5. – P. 251–255. – DOI 10.21275/sr24502103758.
12. Mozharovskii E. Evaluating Retrieval-Augmented Generation (RAG) techniques in enhancing LMS for coding tasks // Universum: технические науки, 2024. – No. 6-6 (123). – P. 22–26.
13. Томилов Н. А., Туров В. П., Бабаянц А. А., Платонов А. В. Метод хранения векторных представлений в сжатом виде с применением кластеризации // Научно-технический вестник информационных технологий, механики и оптики, 2024. – Т. 24, № 1. – С. 112–117. – DOI 10.17586/2226-1494-2024-24-1-112-117.
14. Unlu O., Shin Ch. J., Mailly J. Retrieval-Augmented Generation-Enabled GPT-4 for Clinical Trial Screening // NEJM AI, 2024. – Vol. 1, No. 7. – DOI 10.1056/aioa2400181.
15. Martin A., Witschel H. F., Mandl M., Stockhecke M. Semantic Verification in Large Language Model-based Retrieval Augmented Generation // Proceedings of the AAAI Symposium Series, 2024. – Vol. 3, No. 1. – P. 188–192. – DOI 10.1609/aaais.v3i1.31199.
16. Федоров В. О., Поляков Р. А. Большие языковые модели с поисковой расширенной генерацией: обзор и перспективы // Оригинальные исследования. – 2023. – Т. 13, № 12. – С. 43–47.
17. Kang S. H., Kim S. J. M-RAG: Enhancing Open-domain Question Answering with Metadata Retrieval-Augmented Generation // Journal of the Korea Institute of Information and Communication Engineering, 2023. – Vol. 27, No. 12. – P. 1489–1500. – DOI 10.6109/jkiice.2023.27.12.1489.

Об авторах

Алексей Александрович Колесников – кандидат технических наук, доцент, научный сотрудник.

Сергей Сергеевич Жданов – доктор филологических наук, доцент, ведущий научный сотрудник.

Получено 12.11.2024

© А. А. Колесников, С. С. Жданов, 2025

From travelogues to maps: generating cartographic materials using retrieval-augmented generation

A. A. Kolesnikov¹✉, S. S. Zhdanov¹

¹ Novosibirsk State Technical University, Lingua Language Center,
Novosibirsk, Russian Federation

e-mail: alexeykw@mail.ru

Abstract. The authors propose an approach to overcoming the limitations of large language models in generating responses involving current and domain-specific data by applying retrieval-augmented generation (RAG) methods. The study reviews existing approaches and technical solutions for database querying, with a focus on their application to such tasks as mapping and analyzing travelogue routes and validating results against cartographic and geoinformation resources. A mechanism for transforming cartographic materials into textual descriptions through standardized templates is introduced, where parameters are automatically filled in from the geometric and attribute characteristics of vector map objects. The approach was validated using the QGIS, FIASS vector database, vectorization tool, and the YandexGPT large language model.

Keywords: thematic mapping, language model, travelogues, retrieval-augmented generation, machine learning

REFERENCES

1. Mai, G., Huang, W., Sun, J., Song, S., Mishra, D., Liu, N., Gao, S., Liu, T., Cong, G., Hu, Y., Cundy, C., Li, Z., Zhu, R., Lao, N. (2024) On the Opportunities and Challenges of Foundation Models for GeoAI (Vision Paper). *ACM Trans. Spatial Algorithms Syst*, 10(2)-11, 46. DOI 10.1145/3653070.
2. Jonathan, R., Luddecke, T., Das, S., Han, K., Albanie, S. (2023) GPT4GEO: How a Language Model Sees the World's Geography *ArXiv abs/2306.00020*. Retrieved from <https://arxiv.org/abs/2306.00020> (date of the application: 05.11.2024).
3. Li, Z., Ning, H. (2023) Autonomous GIS: the next-generation AI-powered GIS. *International Journal of Digital Earth*, 16(2), 4668–4686, DOI 10.1080/17538947.2023.2278895.
4. Ran, T., Xu, J. Mapping with ChatGPT. (2023) *ISPRS International Journal of Geo-Information*, 12, 7, 284, DOI:10.3390/ijgi12070284.
5. Sambandam, P., Yuvaraj, D., Padmakumari, P., Swaminathan, S. (2023) Deep attention based optimized Bi-LSTM for improving geospatial data ontology. *Data & Knowledge Engineering*, 144, 102123, DOI 10.1016/j.datak.2022.102123.
6. Skrynnikova, I. (2024) Interpretation problems of LLM: mysteries of figurative language. *Kognitivnyye issledovaniya yazyka [Cognitive studies of language]*, 2-2(58), 638-642 [in Russian].
7. Lewis, P., Piktus, A., Petroni, F. (2020) Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems* (Virtual, Online, 06–12 december 2020), 34.
8. Naraki, Y., Yamaki, R., Ikeda, Y., Horie, T., Naganuma, H. (2024) Augmenting NER Datasets with LLMs: Towards Automated and Refined Annotation. *ArXiv abs:2404.01334v1 [cs.CL]*. Retrieved from <https://arxiv.org/abs/2404.01334> (date of the application: 05.11.2024).
9. Wang, S., Sun, X., Li, X., Ouyang, R., Wu, F., Zhang, T., Li, J., Wang, G. (2023) GPT-NER: Named Entity Recognition via Large Language Models. *ArXiv abs: 2304.10428*. Retrieved from <https://arxiv.org/abs/2304.10428> (date of the application: 05.11.2024).
10. Zhichao, Z., Zhao, Q., Li, J., Ge, Y., Ding, X., Gu, T., Zou, J., Lv, S., Wang, S., Yang, J. (2024) Comparative Analysis of Large Language Models in Chinese Medical Named Entity Recognition. *Bioengineering*, 11, 10, 982, DOI 10.3390/bioengineering11100982.

11. Ja, P. (2024) Automating Threat Intelligence Analysis with Retrieval Augmented Generation (RAG) for Enhanced Cybersecurity Posture. *International Journal of Science and Research*, 13, 5, 251-255, DOI 10.21275/sr24502103758.
12. Mozharovskii, E. (2024) Evaluating Retrieval-Augmented Generation (RAG) techniques in enhancing LMS for coding tasks. *Universum: tekhnicheskiye nauki [Universum: technical sciences]*, 6-6(123), 22-26 [in Russian].
13. Tomilov, N. A., Turov, V. P., Babayants, A. A., Platonov, A. V. (2024) Method of storing vector representations in compressed form using clustering. *Nauchno-tekhnicheskiy vestnik informatsionnykh tekhnologiy, mekhaniki i optiki [Scientific and Technical Bulletin of Information Technologies, Mechanics and Optics]*, 24, 1, 112-117, DOI 10.17586/2226-1494-2024-24-1-112-117 [in Russian].
14. Unlu, O., Shin, Ch. J., Mailly, J. (2024) Retrieval-Augmented Generation–Enabled GPT-4 for Clinical Trial Screening. *NEJM AI*, 1, 7, DOI 10.1056/aioa2400181.
15. Martin, A., Witschel, H. F., Mandl, M., Stockhecke, M. (2024) Semantic Verification in Large Language Model-based Retrieval Augmented Generation. *Proceedings of the AAAI Symposium Series*. 3, 1, 188-192, DOI 10.1609/aaaiss.v3i1.31199.
16. Fedorov V. O., Polyakov R. A. (2023) Large language models with search extended generation: review and prospects. *Original'nyye issledovaniya [Original research]*, 13, 12, 43–47 [in Russian].
17. Kang, S. H., Kim, S. J. (2023) M-RAG: Enhancing Open-domain Question Answering with Metadata Retrieval-Augmented Generation. *Journal of the Korea Institute of Information and Communication Engineering*, 27, 12, 1489-1500, DOI 10.6109/jkiice.2023.27.12.1489.

Author details

Alexey A. Kolesnikov – Ph. D., Associate Professor, Researcher.

Sergey S. Zhdanov – D.Sc., Associate Professor, Leading Researcher.

Received 12.11.2024

© A. A. Kolesnikov, S. S. Zhdanov, 2025