

DOI: <https://doi.org/10.60797/jbg.2026.31.3>

EDN: HMYZDQ

ИСПОЛЬЗОВАНИЕ МЕТОДОВ RAG И ПАРАМЕТРОВ МЕДИЦИНСКИХ СНИМКОВ ДЛЯ ИНТЕРПРЕТАЦИИ ПОСТАНОВКИ ЦИТОЛОГИЧЕСКИХ ДИАГНОЗОВ ЩИТОВИДНОЙ ЖЕЛЕЗЫ

Научная статья

Основин С.С.^{1,*}, Зайцев К.С.², Дюльдин Е.В.³, Боброва Е.В.⁴, Кулик С.Д.⁵⁵ORCID : 0000-0002-9578-9010;^{1, 2, 3, 4, 5}Национальный исследовательский ядерный университет МИФИ, Москва, Российская Федерация

* Корреспондирующий автор (1300stas1300[at]gmail.com)

Аннотация

В настоящий момент заметно быстрое развитие больших языковых моделей и проникновение их во все сферы жизни, поэтому особенно актуальной становится задача понимания принимаемых системами на основе ИИ-решений, и того насколько можно доверять их ответам [1]. Это особенно важно для медицинских областей, где риск допустимой ошибки становится крайне высок. В настоящей исследовательской работе предлагается использовать специальную систему, отвечающую на вопрос «Почему именно такой результат?», который возникает при постановке диагноза автоматическими методами (например, методами компьютерного зрения) при анализе цитологических образцов щитовидной железы. Предлагаемый подход основан на архитектуре Retrieval Augmented Generation (RAG), которая позволяет извлекать различные данные из специализированной базы знаний, а также использовать количественные и числовые признаки, извлечённые из сегментированных изображений клеток для конкретного рассматриваемого случая [2], [3]. Такими признаками являются площадь ядра, соотношение размеров ядро/цитоплазма, плотность клеток, и другие характеристики. В рамках исследования в качестве базы знаний выступают различные медицинские доменные пособия по цитологии, но главным источником является международная система классификации Bethesda или TBSRTC (The Bethesda System for Reporting Thyroid Cytopathology) [4], [5]. Эта система классификации, используется для оценки результатов цитологического исследования щитовидной железы.

Цель исследования — попытаться смоделировать диагностическое рассуждение медицинского эксперта, который рассматривая различные числовые показатели, извлеченные из изображений, основывается на данных базы знаний для вынесения решения. В этом подходе специалист формирует не только прогноз (одна из 6 категорий TBSRTC), но и соответствующее обоснование причин поставленного диагноза, ссылаясь на конкретные визуальные признаки и положения медицинского руководства. Для генерации ответов были использованы различные текстовые языковые модели (Large Language Model, LLM), показывающие разные результаты; контекст извлекается из базы знаний с TBSRTC и признаковых векторов. Для получения численных параметров проведена интеграция с внешней системой визуальной сегментации содержащую параметры анализов.

Основной результат — демонстрация работоспособности RAG, создавать последовательные объяснения и отвечать на вопросы о причинно-следственных связях, соответствующих клинической логике при принятии системной того или иного решения.

Ключевые слова: RAG, система Bethesda (TBSRTC), диагностика щитовидной железы, биомедицина, большие языковые модели, искусственный интеллект в медицине, интеграции численных визуальных данных.

THE USE OF RAG METHODS AND MEDICAL IMAGING PARAMETERS TO INTERPRET CYTOLOGICAL DIAGNOSES OF THE THYROID GLAND

Research article

Osnovin S.S.^{1,*}, Zaitsev K.S.², Dyuldin Y.V.³, Bobrova Y.V.⁴, Kulik S.D.⁵⁵ORCID : 0000-0002-9578-9010;^{1, 2, 3, 4, 5}Moscow Engineering Physics Institute, Moscow, Russian Federation

* Corresponding author (1300stas1300[at]gmail.com)

Abstract

At present, there is a marked acceleration in the development of large language models and their spread into all areas of life; consequently, the task of understanding the decisions made by AI-based systems — and the extent to which their responses can be trusted — is becoming particularly relevant [1] This is especially important in medical fields, where the risk of error is extremely high. This research paper proposes the use of a specialised system designed to answer the question "Why this particular result?", which is raised when a diagnosis is made using automated methods (such as computer vision techniques) during the analysis of thyroid cytology samples. The suggested approach is based on the Retrieval Augmented Generation (RAG) architecture, which allows various data to be retrieved from a specialised knowledge base, as well as the use of quantitative and numerical traits extracted from segmented cell images for the specific case under consideration [2], [3]. Such features include nuclear area, the nucleus-to-cytoplasm ratio, cell density, and other characteristics. Within the scope of the study, various medical domain-specific cytology guidelines serve as the knowledge base, but the main source is the international Bethesda classification system or TBSRTC (The Bethesda System for Reporting Thyroid Cytopathology) [4], [5]. This classification system is used to evaluate the results of thyroid cytological examinations.



The aim of the study is to attempt to model the diagnostic reasoning of a medical expert who, by examining various numerical indicators extracted from an image, relies on data from a knowledge base to reach a decision. In this approach, the specialist not only formulates a prognosis (one of the 6 TBSRTC categories), but also provides a corresponding rationale for the diagnosis, referring to specific visual signs and the provisions of medical guidelines. Various large language models (LLMs) were used to generate responses, yielding different results; context is extracted from a knowledge base containing TBSRTC and feature vectors. To obtain numerical parameters, integration was carried out with an external visual segmentation system containing analysis parameters.

The main result is demonstrating RAG's ability to generate coherent explanations and answer questions about cause-and-effect relationships that align with clinical logic when making a particular systemic decision.

Keywords: RAG, the Bethesda system (TBSRTC), thyroid diagnostics, biomedicine, large language models, artificial intelligence in medicine, integration of numerical visual data.

Введение

Современные методы работы с большими языковыми моделями, демонстрируют не достижимые ранее результаты в задачах автоматической интерпретации биомедицинских заключений и результатов. В том числе при распознавании медицинских изображений. Однако часто для медицинских работников, такая система представляется как «чёрный ящик», поскольку объяснить выбор того или иного решения модель ИИ напрямую не может. Подобная особенность ограничивает внедрение в клиническую практику многих методов искусственного интеллекта, поскольку врач не может доверять решению, если не понимает его оснований и объяснений. Еще более актуальным данный вопрос становится, когда диагностические категории напрямую определяют тактику лечения, например, оперативное вмешательство, наблюдение или дообследование, которые могут обернуться ухудшением состояния пациентов в случае неправильного диагноза.

Международная система классификации Bethesda для цитологических исследований щитовидной железы предоставляет строгую, клинически валидированную схему интерпретации цитологических препаратов, связывая различные признаки с риском злокачественности. Эта система представляет собой надежный «внешний источник знаний», который крайне удобен для построения систем на основе искусственного интеллекта. Уже сейчас, генеративные языковые модели, способны формулировать сложные рассуждения на естественном языке, однако главным недостатком являются галлюцинации и недостаток специфичности языковых моделей с небольшим числом параметров. Метод Retrieval Augmented Generation позволяет объединить точность извлечения из структурированного знания с гибкостью генерации [6]. Цель нашего исследования — разработать и протестировать вычислительную систему, способную имитировать диагностическое рассуждение цитолога при интерпретации результатов анализа щитовидной железы, генерируя не только диагноз, но и его клинически валидное обоснование на основе количественных изображений и экспертного руководства.

Методы и принципы исследования

Система использует два независимых источника входных данных:

1. Знания экспертов: текстовые и структурированные данные связанные с TBSRTC, включающие описания диагностических категорий (I–VI), типичные морфологические признаки (архитектурные паттерны, особенности ядер), соответствующие риски злокачественности и рекомендации по лечению.

2. Количественные числовые данные, полученные с помощью отдельного инструмента для анализа изображений (разработанного внешней командой). Эти данные включают количество клеток на каждом изображении, их типы и структуры, которые они образуют. Пример характеристик, клеток и клеточных скоплений:

«Количество бесформенных структур с упорядоченным расположением клеток — 2. Количество бесформенных структур с неупорядоченным расположением клеток — 35. Количество микрофолликулярных структур — 0. Количество трабекулярных структур — 0. Количество папиллярных структур — 17. Средняя площадь клеточных скоплений — 67351,65. Среднее количество тиреоидных клеток в клеточных скоплениях — 29,63. Количество тиреоидных клеток Гюртле — 223. Количество цепочечных клеток с несколькими ядрами — 0. Количество тиреоидных клеток с псевдовключениями — 11. Средняя площадь тиреоидных клеток в железе — 6494».

Для того чтобы интегрировать эти источники данных, была разработана следующая архитектура, реализующая основные принципы RAG систем, схема которой представлена на рисунке 1.

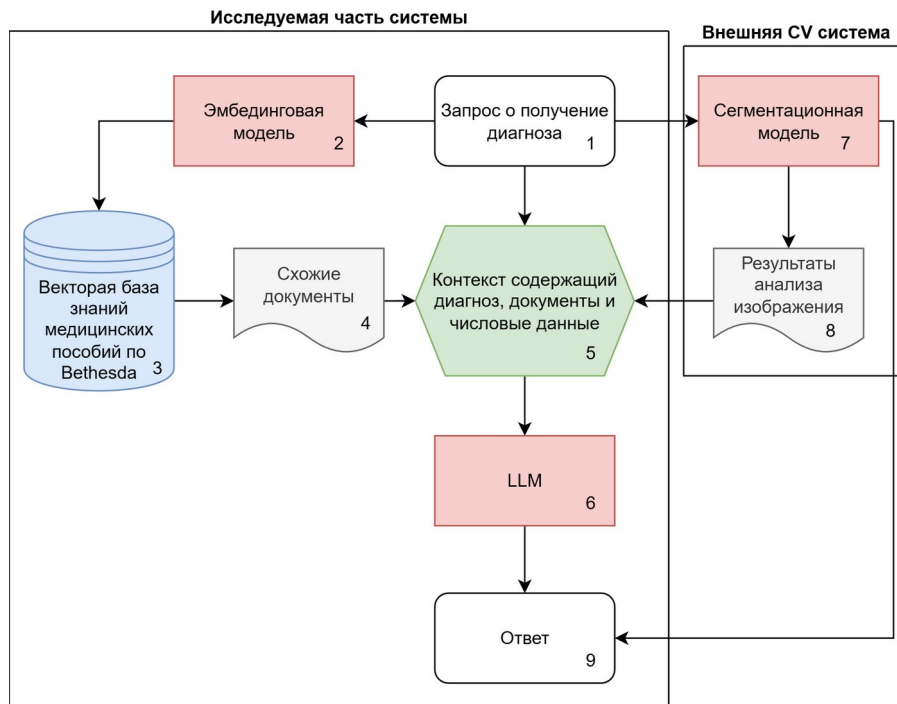


Рисунок 1 - Архитектура созданной системы
DOI: <https://doi.org/10.60797/jbg.2026.31.3.1>

Из схемы видно, что система реализована в виде модулей и представляет собой конвейер получаемых ответов на запросы о поставленном диагнозе другой моделью. При получении результата анализа цитологического изображения с помощью компьютерного зрения формируется вектор запроса с помощью эмбединговой модели (блок 2). После этого с помощью векторной базы из корпуса знаний TBSRTC (блок 3) извлекаются наиболее релевантные фрагменты, например, вся информация, связанная с ее V-ой категорией (блок 4). После этого идёт обращение во внешнюю CV систему, разработанную другой командой исследователей (блок 7). С помощью нее мы извлекаем из анализируемого изображения численные признаки, о которых упоминалось ранее (блок 8). В конечном итоге извлеченные данные формируются в промпт, который полностью подаётся LLM (блок 5 и 6). В данном случае для формирования на англоязычных результатах лучше всего себя показала модель microsoft/phi-4, на русскоязычной базе лучше работает Qwen/Qwen3-8B. На выходе системы получаем — естественно-языковое объяснение с указанием категории TBSRTC и обоснованием (блок 9).

Основные результаты

В рамках исследования была реализована и протестирована на нескольких десятках примеров вычислительная архитектура, демонстрирующая принципиальную возможность применения подходов RAG для формирования логического пояснения постановки диагноза в узкоспециализированной медицинской предметной области — цитологической диагностике по системе TBSRTC. На этом этапе исследования работа фокусируется не на оценке диагностической эффективности, а на моделировании процесса интеграции количественных данных с анализа изображения и структурированного экспертного знания с целью генерации клинически осмысленных обоснований.

Была сконструирована и протестирована полноценная фаза проектирования системы, включающая:

1. Нормализацию и семантическое представление количественных признаков, полученных от модуля визуальной сегментации
2. Формализация текстового корпуса, сформированного на основе официального руководства TBSRTC, с выделением областей полезных для постановки диагноза, а также выявление пороговых значений и рекомендаций по вынесению решения для каждой из шести категорий.
3. Реализацию механизма динамического извлечения релевантных фрагментов контекста через семантический поиск (с использованием модели эмбедингов thenlper/gte-small, индексированных в FAISS).
4. Проектирование и валидацию шаблонов промптов, обеспечивающих строгую привязку генерируемого ответа к извлечённому контексту и входным признакам без отклонения в галлюцинации.
5. Интеграцию с открытыми LLM (для изучения результатов использовались такие как: microsoft/phi-4, Llama-3-8B, Mistral-7b, Qwen/Qwen3-8B).

На тестовых сценариях (подготовленных на основе конкретных обезличенных клинических случаев, предоставленных центром эндокринологии) продемонстрирована способность системы генерировать синтаксически корректные и семантически согласованные объяснения, в которых:

1. Явно указывается диагностическая категория Bethesda.
2. Перечисляются доминирующие морфологические признаки, зафиксированные в количественном виде.
3. Приводится их рассуждения на основании руководств.
4. Формулируется логический вывод, соответствующий клиническим терминам.



Ниже приведен пример сформулированного предлагаемой системой текста:

«Пациенту была присвоена категория VI по классификации Bethesda на основании нескольких ключевых признаков, соответствующих описаниям злокачественных новообразований в этой категории:

1. Признаки папиллярного рака щитовидной железы:

– Наличие 17 папиллярных структур.

– Обнаружение 11 тиреоидных клеток с псевдовключениями.

Эти признаки соответствуют описанию папиллярного рака щитовидной железы в категории VI, где эпителиальные клетки образуют папиллярные структуры и содержат внутриядерные псевдовключения.

2. Признаки неуточненного рака щитовидной железы:

– Большое количество бесформенных структур с неупорядоченным расположением (35), что свидетельствует о дезорганизованном росте клеток.

– Большая средняя площадь клеточных скоплений (67 351,65) и значительное среднее количество тиреоидных клеток в скоплении (29,63).

Эти характеристики соответствуют описанию неуточненного рака щитовидной железы, характеризующегося крупными полиморфными клетками в скученных скоплениях.

3. Показатели медулярного рака щитовидной железы:

– Хотя результаты не содержат подробной информации, общая картина предполагает возможное поражение медулярного слоя, учитывая наличие крупных клеток и потенциальное совпадение с другими показателями злокачественности.

Учитывая эти наблюдения, сочетание морфологических признаков убедительно подтверждает диагноз злокачественной опухоли щитовидной железы, соответствующей критериям категории VI по Bethesda».

Важным достижением является установление корректного баланса между гибкостью генерации и строгой опорой на базу знаний, созданную из медицинского домена. При отключении механизма поиска дополнений, то есть при работе LLM без каких-либо подобранных документов, доля некорректных или избыточных ответов резко возрастает. Модели с небольшим количеством параметров, такие как 7B и меньше, зачастую очень слабо могут отвечать на вопросы в столь узкой области без дополнительного дообучения, из-за этого практически сразу возникают галлюцинации, в ходе которых создаются вымышленные данные, которыми модель пытается искусственно подкрепить свой ответ.

По итогу данного исследования, нами была подтверждена гипотеза о том, что использование RAG позволяет существенно повысить фактологическую точность и доменную согласованность генерируемых логических объяснений в условиях узкой медицинской специализации, даже при использовании больших языковых моделей в их базовой конфигурации.

Обсуждение

В работе F. Petroni «Language Models as Knowledge Bases?» [7] убедительно показано, что даже без дообучения такие модели, как BERT, способны извлекать релевантные фактические связи из собственных параметров, демонстрируя неожиданную эффективность в задачах типа закрытого-заполнения и даже в открытых вопросно-ответных системах.

В другом исследовании G. Izacard «Atlas: few-shot learning with retrieval augmented language models» [8] модель продемонстрировала, что тщательно спроектированная и предварительно обученная RAG-система способна осваивать сложные задачи, чувствительные к знаниям, всего по нескольким десяткам примеров. Поразительно, но 64 примера на Natural Questions позволяют авторам превзойти 540-миллиардную параметрическую модель, причём с разрывом в 50 раз в размере параметров.

Подтверждением эффективности работы RAG можно увидеть в статье B. Gu «Probabilistic medical predictions of large language models» [9], где авторы показали, что даже при использовании продвинутых свободно доступных LLM, включая модели с 70 млрд параметров (без обучения на медицинских данных), вероятностные оценки, генерируемые через промптинг, демонстрируют низкую предсказательную, рассуждающую способность и нестабильны на несбалансированных медицинских данных.

Переход от абстрактных архитектур к реальным научным инструментам с использованием RAG иллюстрирует проект «Liu W. DrBioRight 2.0: an LLM-powered bioinformatics chatbot for large-scale cancer functional proteomics analysis» [10] — чат-бот для анализа функциональной протеомики рака, построенный на LLM и интегрированный с уникальной базой данных, включающей почти 8000 образцов пациентов и 500 высококачественных антител.

Мы так же сразу перешли к практическим решениям задач, но несколько в другой области медицины. Полученные нами результаты позиционируют архитектуру RAG как весьма привлекательную методологическую основу для создания ИИ-помощников в медицинской сфере с интерпретируемыми возможностями в условиях, когда:

1. Существует хорошо формализованная предметная область, как в методологии TBSRTC.

2. Диагностические решения, основанные на соответствии количественных показателей с пороговыми критериями.

3. Существует необходимость поддержания логического рассуждения в соответствии с клиническими протоколами.

Таким образом, это свидетельствует в пользу практического использования такой системы благодаря её возможностям к адаптации в других областях. Модуль извлечения может быть заменён при изменении рекомендаций, а генеративный компонент адаптирован к различным медицинским стандартам и русскоязычным терминам, например, через дообучение на специализированных данных, связанных с цитологией (включая полноценные врачебные заключения). Следует отметить, что предлагаемый подход не является самостоятельной диагностической системой,



работающей без участия человека, а выступает в качестве надстройки для поддержки принятия решений, ускоряющий принятие решения и повышающей прозрачность ИИ-помощника для медицинского работника. Очевидно, что в дальнейшем предполагается расширение возможностей использования этого подхода на большем количестве узких областей. Конкретно же для тех данных, которые рассмотрены в статье, в НМИЦ эндокринологии им. академика И.И. Дедова Минздрава России проводится проверка клинической полезности предлагаемой системы RAG. Исследование оценок влияния выводов системы на экспертные решения клиницистов и интеграция ее с реальными диагностическими системами, используемыми в НМИЦ, составляют предмет дальнейших исследований.

Заключение

Итоги этой работы продемонстрировали принципиальную возможность использования и применения методов RAG для моделирования диагностического рассуждения в ограниченной медицинской предметной области, такой как, цитология щитовидной железы. Проведенное исследование показало, что интеграция количественных показателей данных совместно со сформированной и обработанной профессиональной базой знаний, работающей через механизм динамического контекстного извлечения и дополнения информации, позволяет генерировать логические объяснения и интерпретации диагнозу, поставленному моделями глубокого обучения. При этом формулировки, генерируемые LLM на основании сформированного контекста, соответствуют структуре и логике клинических руководств, позволяя уменьшить возможные галлюцинации и другие проблемы, свойственные языковым моделям с небольшим количеством параметров и без дополнительного дообучения на специализированных данных.

Созданный исследовательский прототип медицинской системы подтвердил, что RAG может быть крайне эффективным инструментом для повышения прозрачности и достоверности генеративных моделей в различных узкоспециализированных биомедицинских задачах, где крайне критична точность и понимание принимаемого решения. Предложенное решение доказывает возможность сочетания преимуществ глубокого обучения и структурированного экспертного знания при формировании дополнительного контекста к LLM. Полученные результаты позволяют расширить дальнейшую разработку и масштабировать выборку предметной области для исследования в системах поддержки принятия решений, а также проводить исследования качества и влияния совместно со специалистами медицинской области.

Конфликт интересов

Не указан.

Conflict of Interest

None declared.

Рецензия

Все статьи проходят рецензирование. Но рецензент или автор статьи предпочли не публиковать рецензию к этой статье в открытом доступе. Рецензия может быть предоставлена компетентным органам по запросу.

Review

All articles are peer-reviewed. But the reviewer or the author of the article chose not to publish a review of this article in the public domain. The review can be provided to the competent authorities upon request.

Список литературы / References

1. Lee J. A comprehensive evaluation of quantized instruction-tuned large language models: experimental analysis up to 405B / J. Lee, S. Park, J. Kwon [et al.] // arXiv preprint. — 2024. — URL: <https://arxiv.org/html/2409.11055v1> (accessed: 09.12.2025).
2. Gupta S. A comprehensive survey of retrieval-augmented generation (RAG): evolution, current landscape and future directions / S. Gupta, R. Ranjan, S.N. Singh // arXiv preprint. — 2024. — URL: <https://arxiv.org/abs/2410.12837> (accessed: 07.12.2025).
3. Lewis P. Retrieval-augmented generation for knowledge-intensive NLP tasks / P. Lewis, E. Perez, A. Piktus [et al.] // arXiv preprint. — 2021. — URL: <https://arxiv.org/abs/2005.11401> (accessed: 07.12.2025).
4. Cibas E.S. The Bethesda System for Reporting Thyroid Cytopathology / E.S. Cibas, S.Z. Ali // *Thyroid*. — 2018. — Vol. 19. — № 11. — P. 1159–1165.
5. Ali S.Z. The 2023 Bethesda System for reporting thyroid cytopathology / S.Z. Ali, Z.W. Baloch, B. Cochand-Priollet [et al.] // *Thyroid*. — 2023. — Vol. 12. — № 5. — P. 319–325.
6. Лозовая К.В. Программная реализация методов аугментации данных для повышения качества работы нейронных сетей / К.В. Лозовая, С.А. Парыгина // *Современные информационные технологии. Теория и практика : материалы VI Всероссийской научно-практической конференции*. — Череповец : ЧГУ, 2024. — С. 23–29.
7. Petroni F. Language models as knowledge bases? / F. Petroni, T. Rocktäschel, S. Riedel et al. // *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. — Hong Kong, 2019. — P. 2463–2473.
8. Izacard G. Atlas : few-shot learning with retrieval augmented language models/ G. Izacard, P. Lewis, M. Lomeli [et al.] // arXiv preprint. — 2022. — URL: <https://arxiv.org/abs/2208.03299> (accessed: 07.12.2025).
9. Gu B. Probabilistic medical predictions of large language models / B. Gu, R.J. Desai, K.J. Lin [et al.] // *npj Digital Medicine*. — 2024. — Vol. 7. — P. 1–9. — DOI: 10.1038/s41746-024-01366-4.
10. Liu W. DrBioRight 2.0: an LLM-powered bioinformatics chatbot for large-scale cancer functional proteomics analysis / W. Liu, J. Li, Y. Tang [et al.] // *Nature Communications*. — 2025. — Vol. 16. — № 1. — P. 1–6. — DOI: 10.1038/s41467-025-57430-4.

**Список литературы на английском языке / References in English**

1. Lee J. A comprehensive evaluation of quantized instruction-tuned large language models: experimental analysis up to 405B / J. Lee, S. Park, J. Kwon [et al.] // arXiv preprint. — 2024. — URL: <https://arxiv.org/html/2409.11055v1> (accessed: 09.12.2025).
2. Gupta S. A comprehensive survey of retrieval-augmented generation (RAG): evolution, current landscape and future directions / S. Gupta, R. Ranjan, S.N. Singh // arXiv preprint. — 2024. — URL: <https://arxiv.org/abs/2410.12837> (accessed: 07.12.2025).
3. Lewis P. Retrieval-augmented generation for knowledge-intensive NLP tasks / P. Lewis, E. Perez, A. Piktus [et al.] // arXiv preprint. — 2021. — URL: <https://arxiv.org/abs/2005.11401> (accessed: 07.12.2025).
4. Cibas E.S. The Bethesda System for Reporting Thyroid Cytopathology / E.S. Cibas, S.Z. Ali // *Thyroid*. — 2018. — Vol. 19. — № 11. — P. 1159–1165.
5. Ali S.Z. The 2023 Bethesda System for reporting thyroid cytopathology / S.Z. Ali, Z.W Baloch, B. Cochand-Priollet [et al.] // *Thyroid*. — 2023. — Vol. 12. — № 5. — P. 319–325.
6. Lozovaya K.V Programmaya realizatsiya metodov augmentatsii dannikh dlya povsheniya kachestva raboti neironnikh setei [Software implementation of data augmentation methods to improve the performance of neural networks] / K.V Lozovaya, S.A. Parigina // *Sovremennye informacionnye tekhnologii. Teoriya i praktika [Modern information technologies. Theory and practice] : proceedings of the VI All-Russian Scientific and Practical Conference*. — Cherepovets : ChGU, 2024. — P. 23–29.[in Russian]
7. Petroni F. Language models as knowledge bases? / F. Petroni, T. Rocktäschel, S. Riedel et al. // *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. — Hong Kong, 2019. — P. 2463–2473.
8. Izacard G. Atlas : few-shot learning with retrieval augmented language models/ G. Izacard, P. Lewis, M. Lomeli [et al.] // arXiv preprint. — 2022. — URL: <https://arxiv.org/abs/2208.03299> (accessed: 07.12.2025).
9. Gu B. Probabilistic medical predictions of large language models / B. Gu, R.J. Desai, K.J. Lin [et al.] // *npj Digital Medicine*. — 2024. — Vol. 7. — P. 1–9. — DOI: 10.1038/s41746-024-01366-4.
10. Liu W. DrBioRight 2.0: an LLM-powered bioinformatics chatbot for large-scale cancer functional proteomics analysis / W. Liu, J. Li, Y. Tang [et al.] // *Nature Communications*. — 2025. — Vol. 16. — № 1. — P. 1–6. — DOI: 10.1038/s41467-025-57430-4.