Виртуальный собеседник для генерации шаблонов специализированных документов на основе ключевых фраз

А. П. Степанов

Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)

E-mail stepetal94@yandex.ru

***Аннотация*. Последние несколько лет характеризуются широким внедрением больших языковых моделей (Large Language Models, LLMs) в качестве ассистента для выполнения задач по генерации и редактированию текстовых данных. В онлайн-доступе имеются как платные, так и бесплатные виртуальные собеседники (chatbot), однако их использование сопряжено с угрозой конфиденциальности информации. Для того, чтобы адресовать эту проблему, выполняют разворачивание больших языковых моделей с открытым исходным кодом внутри защищенного контура. Однако, ввиду значительной стоимости вычислительных мощностей, возникает необходимость решения поставленных задач с использованием LLM с как можно меньшим числом параметров. Такие модели не способны генерировать качественный текст узкой предметной области. Решить подобный вопрос позволяют как техника генерации с дополненной выборкой (Retrieval-Augmented Generation, RAG), так и дообучение на корпусе данных. В случае нехватки обучающей выборки единственной возможностью остается RAG. В работе описывается архитектура виртуального собеседника, а также метод генерации, расширяющий возможности RAG применительно к созданию шаблонов специализированных документов. Тестирование осуществлялось на наборе деперсонализированных медицинских заключений магнитной резонансной томографии (МРТ).**

Ключевые слова: виртуальный собеседник; большая языковая модель; шаблон специализированного документа; магнитно-резонансная томография; генерация с дополненной выборкой

I. Введение

В настоящее время большие языковые модели (Large Language Models, LLMs) активно внедряются во все ключевые области деятельности человека. Например, LLMs используются как один из инструментов по нахождению оптимальных параметров при решении задачи дизайна, проведения автоматизации вычислений, определения признаков структурной целостности объектов строительства. Мультимодальные большие языковые модели, обладающие способностью логического вывода, могут применяться при проектировании зданий [1], для анализа результатов обследований, исследования повреждений и дефектов, возникающих в строительных конструкциях [2, 3]. Решение данных задач необходимо для того, чтобы предотвратить прогрессирующие обрушения и снизить риск аварий, возникающих при эксплуатации зданий и сооружений. Важно подчеркнуть, что для эффективного использования LLMs необходим интерфейс взаимодействия с ними. Виртуальные собеседники (chatbots) как раз предоставляют такую возможность [4, 5]. На текущий момент имеется большое разнообразие виртуальных собеседников, предоставляемых как сервис. Первым из них принято считать ChatGPT [6], появление которого ознаменовало новую эпоху в исследованиях в области искусственного интеллекта. Основным достоинством подобных ассистентов является то, что для их использования не нужно обладать большими вычислительными мощностями. Однако их применение сопряжено с риском конфиденциальности информации, что во многих сферах, например в медицине, является критичным [7, 8]. Поэтому до сих пор актуальной остается задача создания виртуального собеседника, обладающего минимальными требованиями к аппаратному обеспечению.

Еще одним критерием является направленность на решение конкретной задачи: чем более специализированную задачу необходимо решить, тем меньше вероятность найти большую языковую модель, которая бы с ней справилась. Одной из таких задач является генерация текстов узкой предметной области вообще и шаблонов специализированных документов в частности. Применительно к области медицины до сих пор до конца не решена проблема полной генерации шаблона медицинского заключения магнитной резонансной томографии (МРТ) [9–12] в условиях ограниченных ресурсов и с учетом авторского стиля. Один из путей решения заключается в дообучении большой языковой модели на корпусе текстов, однако подобный подход требует, чтобы у каждого автора был датасет достаточного размера [13].

В докладе описывается метод, который опирается на технику генерации с дополненной выборкой (Retrieval Augmented Generation, RAG) и позволяет создать шаблон специализированного документа на основе ключевых фраз, учитывая стиль автора. Данный метод положен в основу виртуального собеседника, описание которого также приведено в работе. Тестирование метода осуществлялось на массиве деперсонализированных медицинских заключений МРТ.

II. Материалы и методы

A. Описание набора данных

Исходный датасет включал в себя 908 медицинских заключений МРТ, принадлежащие 7 специалистам. На этапе предобработки все тексты были деперсонализированы и разбиты на фрагменты:

описательную часть и обобщение. Каждая из описательных частей дополнительно была декомпозирована на абзацы, абзацы на предложения, а предложения на блоки текста длиной в несколько слов. Полученная информация была записана в векторную базу данных (БД) вместе с сопутствующими метаданными, необходимыми для последующей фильтрации выборки.

B. Генерация с дополненной выборкой

Генерация с дополненной выборкой предназначена для расширения базы знаний большой языковой модели на основании информации, полученной из внешних источников [14]. Техника RAG используется, чтобы адресовать две основные проблемы: недостаточный размер обучающей выборки и устаревание информации в наборе данных. RAG опирается на векторную базу данных, в которую записаны векторизованные представления текстов (embeddings). В общем случае RAG включает 3 основных этапа: 1) Выборка наиболее подходящих запросу пользователя текстовых фрагментов из векторной БД и формирование контекста. 2) Обогащение инструкции (промпта) для большой языковой модели полученным контекстом. 3) Передача промпта большой языковой модели для генерации. Базовая техника RAG имеет ряд недостатков [15], связанных с полнотой и точностью выборки, галлюцинациями большой языковой модели и ограничением размера контекстного окна, что приводит к необходимости разработки её модификации.

C. Предлагаемая модификация метода RAG

Разработанный метод развивает базовую технику RAG и позволяет сгенерировать шаблон специализированного документа на основе ключевых фраз.

Обозначения:

- *K(N)* – множество *N* элементов, каждый из которых *K(i) (i = 1 … N)* представляет собой ключевую фразу.
- *P(G)* – множество *G* элементов, каждый из которых *P(g) (g = 1 … G)*, представляет из себя абзац, соответствующий ключевой фразе *K(i)*.
- *D(Q)* – множество *Q* элементов, каждый из которых *D(q) (q = 1 … Q)* есть документ, наиболее близкий по метрике косинусной близости к параграфам из множества *P(G)*.
- *C(M)* – множество *M* элементов, каждый из которых *C(m) (m = 1 … M)*, есть набор перекрывающихся текстовых блоков размером несколько слов (chunks).
- *R* – результат генерации большой языковой моделью описательной части согласно инструкции и контекста.
- *Pr* – инструкция для большой языковой модели.
- *LLM* – большая языковая модель.
- *DB* – векторная база данных.

Предусловия:

В векторную базу данных заносится следующая информация:

- Набор документов, векторное представление каждого из них и метаданные (наименование документа).
- Набор абзацев каждого из документов, векторных представлений абзацев и метаданные (идентификатор документа, которому принадлежит абзац).
- Набор текстовых фрагментов для каждого абзаца, векторных представлений фрагментов, метаданные (идентификатор абзаца, из которого были получены текстовые фрагменты)

Исходные данные:

K(N) – множество ключевых фраз.

Описание метода:

- Для каждого ключевого слова *K(i)* в векторной базе данных *DB* найти наиболее близкий текстовый фрагмент *C(m)*. Для текстового фрагмента *C(m)* в векторной базе данных *DB* найти наиболее близкий абзац *P(g)*. В итоге имеем сопоставление *K(i)→P(g)*.
- Извлечь из векторной базы данных *DB* набор описательных частей *D(Q)*, наиболее близких абзацам *P(G)*.
- Интегрировать *D(Q)* в результирующую инструкцию *Pr* для большой языковой модели, передать инструкцию *LLM* и получить результирующую описательную часть *R*.

Схема метода изображена на рис. 1. Инструкция к большой языковой модели приведена в табл. I.

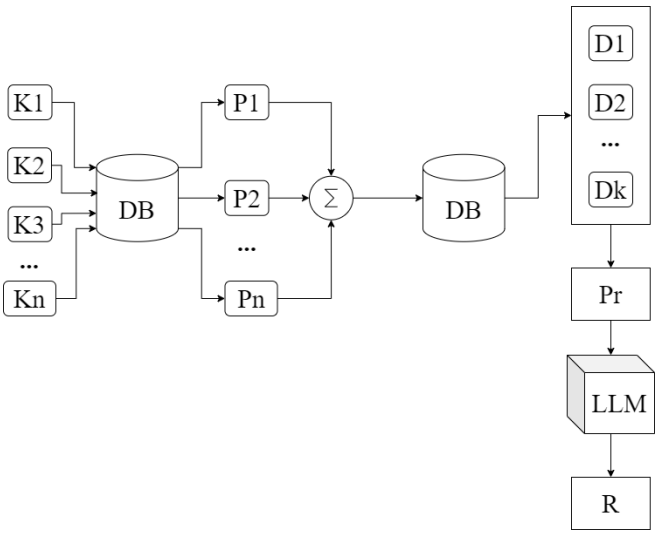

Рис. 1. Схема разработанного метода

ТАБЛИЦА I. ИНСТРУКЦИИ ДЛЯ БОЛЬШОЙ ЯЗЫКОВОЙ МОДЕЛИ

Системная инструкция	**Пользовательская инструкция**
Ты - профессиональный радиолог. Твоя основная задача - создание медицинского заключения. Внимательно изучи запрос пользователя и предоставь профессионально составленное заключение	Дан набор шаблонов медицинских заключений: {*context*} Необходимо сделать следующее: 1. Определить, какой из шаблонов наиболее близок по смыслу к следующим абзацам: {*paragraphs*} 2. Изменить этот шаблон таким образом, чтобы он включал в себя указанные выше абзацы. 3. Устранить в полученном результате дублирующие предложения, ОБЯЗАТЕЛЬНО оставив указанные выше абзацы. 4. Выведи ТОЛЬКО полученный результат. Не рассуждай. Не используй вступительные слова.

Инструкция к большой языковой модели из табл. I включает в себя переменную *context*, содержащую массив описательных частей и переменную *paragraphs*, которая содержит набор абзацев, близких по метрике косинусной близости к исходным ключевым фразам. Системная компонента инструкции позволяет настроить большую языковую модель на генерацию текста по заданной тематике.

D. Мультиагентное приложение

При построении архитектуры виртуального собеседника использовался мультиагентный подход, основанный на графовой структуре данных. Каждый узел – это программный агент, реализующий одну из трёх атомарных операций: извлечение данных из векторной БД, генерацию описательной части медицинского заключения с помощью разработанной модификации метода RAG, генерацию обобщения описательной части. Фреймворком для реализации графовой структуры и обеспечения поддержки нескольких агентов выступил LangChain. В качестве большой языковой модели применялась LLama3.1 8b, которая была развернута на сервере ollama. Это позволило значительно снизить требования к объему памяти видеокарты. Векторной базой данных предстала ChromaDB. Векторизация русскоязычного текста была осуществлена с помощью нейросетевой модели FRIDA. Транскрибирование аудиопотока было выполнено нейросетевой моделью Whisper. Качество созданного текста оценивалось при помощи набора метрик ROUGE [16], BERTScore [17], BLEU [18].

III. РЕЗУЛЬТАТЫ

Результаты оценки качества генерации при помощи разработанного метода в зависимости от специалиста приведены на рис. 2 и рис. 3.

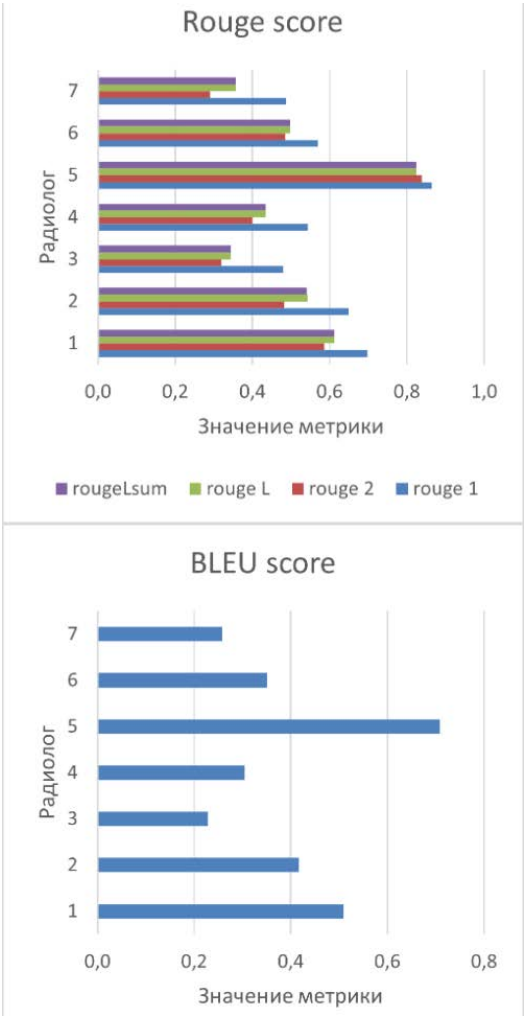

Рис. 2. Оценка качества генерации с помощью метрик ROUGE и BLEU

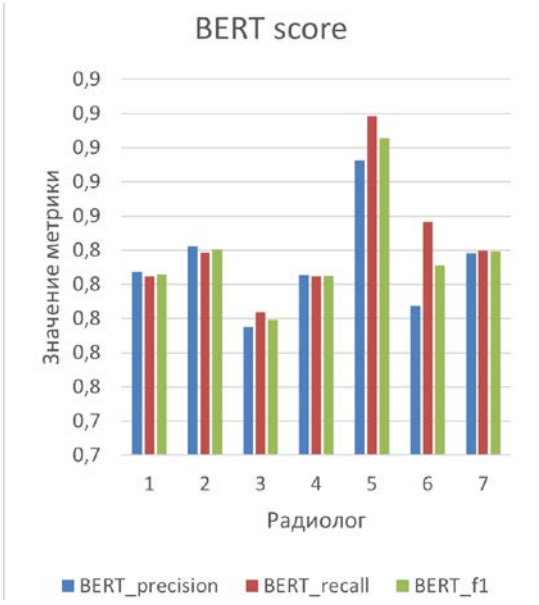

Рис. 3. Оценка качества генерации с помощью метрики BERTScore

Значения метрики ROUGE демонстрируют, что большая языковая модель определяет как короткие, так и длинные последовательности корректно, а это в свою очередь свидетельствует о связности сгенерированного текста. Значения метрики BELU показывают, что, несмотря на сходство сгенерированного и опорных текстов, изменение порядка слов и употребление синонимичных конструкций влияют на итоговые результаты. Поскольку ROUGE и BLEU метрики не позволяют оценить семантическую похожесть текстов, была произведена оценка с помощью метрики BERTScore, значения которой свидетельствуют о близости сгенерированного и опорного текстов по смыслу.

Интерфейс виртуального собеседника и результат генерации представлен на рис. 4. Изначально пользователю системы предоставляется возможность ввести ключевые фразы, на основании которых будет построен шаблон медицинского заключения МРТ. Этот шаблон можно редактировать с помощью отправки дополнительных команд большой языковой модели. Стоит отметить, что для сокращения времени предусмотрен голосовой ввод инструкций.

IV. Заключение

Таким образом, был разработан метод, позволяющий генерировать шаблон специализированного документа на основе ключевых фраз с учетом стиля автора. Метод был положен в основу виртуального собеседника, архитектурное исполнение которого предполагает использование в условиях ограниченных ресурсов. С помощью метрик ROUGE, BLUE и BERTScore была проведена оценка качества генерации шаблона медицинского заключения МРТ. Дальнейшие исследования будут направленны на оптимизацию разработанного метода, улучшение интерфейса виртуального собеседника, а также на создание опросника для качественной оценки результатов.

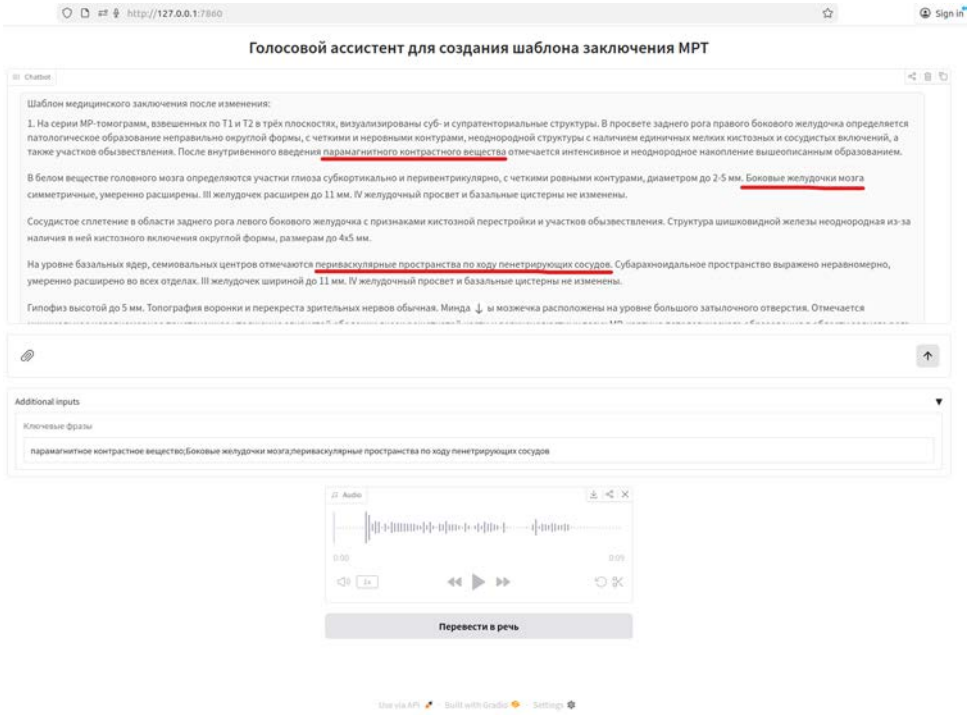

Рис. 4. Интерфейс виртуального собеседника

Список литературы

[1] Elsisi A. Large Language Models Application in Civil and Structural Engineering: Review [Электронный ресурс] // SSRN. 2025. Режим доступа: https://ssrn.com/abstract=5803822 (Дата обращения: 06.02.2026).

[2] Белый Г.И. Влияние усталостных трещин в стенке на прочность подкрановых балок / Г.И. Белый, А.Е. Кубасевич // Вестник МГСУ. 2023. Т. 18, № 11. С. 1780–1790. DOI: https://doi.org/10.22227/1997-0935.2023.11.1780-1790.

[3] Несущая способность усиленных узлов стальных ферм из гнутосварных профилей на продавливание [Электронный ресурс] / Ш. М. Мамедов [и др.] // Инженерный вестник Дона. 2023. № 6. – Режим доступа: http://www.ivdon.ru/ru/magazine/archive/n6y2023/8486. – (Дата обращения: 06.02.2026).

[4] Adamopoulou E., Moussiades L. An overview of chatbot technology //IFIP international conference on artificial intelligence applications and innovations. Cham : Springer International Publishing, 2020. P. 373-383.

[5] Cahn J. CHATBOT: Architecture, design, & development //University of Pennsylvania School of Engineering and Applied Science Department of Computer and Information Science. 2017. Vol. 46.

[6] Training language models to follow instructions with human feedback / L. Ouyang [и др.] // Advances in neural information processing systems. 2022. V. 35. P. 27730-27744.

[7] Large language models in medicine: the potentials and pitfalls: a narrative review / Omiye J. A. [и др.] //Annals of internal medicine. 2024. V. 177, No. 2. P. 210-220.

[8] Костров С. А., Потапов М. П. Большие языковые модели в медицине: актуальные этические вызовы // Медицинская этика. 2025. №2. С. 23-34. – DOI: https://doi.org/10.24075/medet.2025.008

[9] Constructing a Large Language Model to Generate Impressions from Findings in Radiology Reports / L. Zhang [и др.] // Radiology. 2024. V. 312, No. 3. P. e240885. DOI: https://doi.org/10.1148/radiol.240885.

[10] GPT-Driven Radiology Report Generation with Fine-Tuned Llama 3 / Ş. Voinea [и др.] // Bioengineering (Basel). 2024. V. 11, No. 10. P. 1043. – DOI: https://doi.org/10.3390/bioengineering11101043.

[11] Генерация врачебных заключений и классификация по Bethesda с использованием глубокого обучения / Е.В. Боброва [и др.] // Международный журнал открытых информационных технологий. 2023. Т. 11, №10. С. 119−129.

[12] An open-source fine-tuned large language model for radiological impression generation: a multi-reader performance study / A. Serapio [и др.] // BMC Med Imaging. 2024. V. 24, No. 1. P. 254. DOI: https://doi.org/10.1186/s12880-024-01435-w.

[13] Stepanov A.P., Shichkina Y.A. Fine-Tuning LLM's for Domain-Specific Text Generation in Environments with Limited Resource Capabilities // 2025 VI International Conference on Neural Networks and Neurotechnologies (NeuroNT). Saint Petersburg: IEEE, 2025. P. 56−58. – DOI: https://doi.org/10.1109/NeuroNT66873.2025.11049975.

[14] Retrieval-augmented generation for knowledge-intensive NLP tasks / P. Lewis [и др.] // Advances in neural information processing systems. 2020. V. 33, P. 9459-9474.

[15] Retrieval-augmented generation for large language models: A survey / Y. Gao [и др.] // arXiv preprint. – 2023. Режим доступа: https://arxiv.org/abs/2312.10997 (Дата обращения: 06.02.2026).

[16] Lin C. Y. Rouge: A package for automatic evaluation of summaries // *In Text summarization branches out*. 2004. P. 74−81.

[17] Bertscore: Evaluating text generation with bert / Zhang, T [и др.] // arXiv preprint. – 2019. Режим доступа: https://arxiv.org/abs/1904.09675 (Дата обращения: 06.02.2026).

[18] Bleu: a method for automatic evaluation of machine translation / K. Papineni [и др.] // In Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002. P. 311-318.