КАШИРИНА И.Л.,
д.т.н., профессор, МИРЭА – Российский технологический университет, Москва, Россия;
e-mail: kashirina@mirea.ru; ORCID: 0000-0002-8664-9817

СТАРИЧКОВА Ю.В.,
к.т.н., МИРЭА – Российский технологический университет, Москва, Россия;
e-mail: starichkova@mirea.ru; ORCID: 0000-0003-1804-0761

ЛЕ Ч.К.,
МИРЭА – Российский технологический университет, Москва, Россия; e-mail: letrungkienlk4@gmail.com

ТОНКАЯ НАСТРОЙКА ЯЗЫКОВОЙ МОДЕЛИ RUBERT ДЛЯ ПОВЫШЕНИЯ ТОЧНОСТИ АНАЛИЗА МЕДИЦИНСКИХ ЗАПРОСОВ

DOI: 10.25881/18110193_2026_1_64

Аннотация. *Цель исследования состояла в повышении точности семантического поиска медицинской информации на русском языке путем тонкой настройки языковой модели RuBERT на специализированном датасете RuMedDaNet с применением метода обучения Matryoshka Representation Learning для создания компактных и эффективных векторных представлений текста.*

Материалы и методы. В исследовании использовался датасет RuMedDaNet, содержащий русскоязычные медицинские тексты. Для оптимизации производительности поиска применялись различные техники обучения эмбеддингов (векторных представлений текста), включая подход «матрёшка», позволяющий уменьшить размерность векторных представлений без существенной потери качества.

Результаты. Эксперименты показали значительное улучшение ключевых метрик поиска (NDCG, MRR) по сравнению с базовой моделью RuBERT. Обученная в исследовании языковая модель загружена на платформу Hugging Face, где теперь она доступна для открытого использования заинтересованными специалистами.

Заключение. Предложенный метод тонкой настройки RuBERT эффективен для задач поиска в медицинских RAG (Retrieval Augmented Generation)-системах. В статье обсуждаются текущие ограничения предлагаемого подхода и направления дальнейших исследований.

Ключевые слова: *RuBERT, тонкая настройка, RuMedDaNet, медицинские тексты, векторный поиск, Matryoshka Representation Learning.*

Для цитирования: *Каширина И.Л., Старичкова Ю.В., Ле Ч.К. Тонкая настройка языковой модели RuBERT для повышения точности анализа медицинских запросов. Врач и информационные технологии. 2026; 1: 64-73. DOI: 10.25881/18110193_2026_1_64.*

KASHIRINA I.L.,
DSc., Professor, MIREA – Russian Technological University, Moscow, Russia;
e-mail: kash.irina@mail.ru; ORCID: 0000-0002-8664-9817

STARICHKOVA YU.V.,
PhD., MIREA – Russian Technological University, Moscow, Russia;
e-mail: starichkova@mirea.ru; ORCID: 0000-0003-1804-0761

LE T.K.,
MIREA – Russian Technological University, Moscow, Russia; e-mail: letrungkienlk4@gmail.com

FINE-TUNING THE RUBERT LANGUAGE MODEL TO IMPROVE THE ACCURACY OF MEDICAL QUERY ANALYSIS

DOI: 10.25881/18110193_2026_1_64

Abstract. *The aim of the study was to improve the accuracy of semantic search of medical information in Russian by fine-tuning the RuBERT language model on the specialized RuMedDaNet dataset using the Matryoshka Representation Learning method to create compact and efficient vector representations of text.*

Materials and Methods. The study utilized the RuMedDaNet dataset, which contains Russian-language medical texts. Various embedding training techniques were applied to optimize performance, including the "matryoshka" approach, which enables reducing the dimensionality of vector representations without loss of quality.

Results. Experiments demonstrated a significant improvement in key search metrics (NDCG, MRR) compared to the baseline RuBERT model. The language model trained in the study has been uploaded to the Hugging Face platform, where it is now available for open use.

Conclusion. The proposed RuBERT fine-tuning method was effective for search tasks in medical RAG systems. The current limitations of the approach and directions for further research are discussed.

Keywords: *RuBERT, fine-tuning, RuMedDaNet, medical texts, information extraction, Matryoshka Representation Learning.*

For citation: *Kashirina I.L., Starichkova Yu.V., Le T.K. Fine-tuning the RuBERT language model to improve the accuracy of medical query analysis. Medical doctor and information technology. 2026; 1: 64-73. DOI: 10.25881/18110193_2026_1_64.*

ВВЕДЕНИЕ

Современное развитие медицины характеризуется стремительной цифровизацией и внедрением технологий искусственного интеллекта в процессы диагностики и консультирования. Однако эффективность существующих систем поддержки принятия врачебных решений существенно ограничена недостаточной точностью семантического анализа медицинских текстов, особенно в условиях высокой вариативности клинической терминологии и необходимости прецизионной интерпретации профессиональной информации. Качественный семантический поиск и извлечение релевантных данных становятся критическими факторами для создания надежных медицинских информационных систем, способных не только накапливать знания, но и контекстуально интерпретировать сложные клинические кейсы с учетом тонких смысловых нюансов профессиональной коммуникации.

Метод Retrieval Augmented Generation (RAG) [1] значительно расширяет возможности современных больших языковых моделей (LLM) за счёт добавления в запрос модели дополнительной релевантной информации из предметной области на этапе генерации ответов. Ключевым фактором, определяющим эффективность RAG-систем, является качество используемых моделей эмбеддингов, то есть моделей, формирующих числовое векторное представление текстовых данных. Универсальные модели, несмотря на их широкую языковую компетенцию, зачастую демонстрируют недостаточную эффективность в узкоспециализированных областях, таких как медицина, где терминология и контекст существенно отличаются от общеупотребительного языка.

Особую сложность представляет обработка медицинских текстов на русском языке из-за его богатой морфологии и ограниченного количества специализированных лингвистических ресурсов. Широко используемая на практике модель RuBERT (Russian BERT) [2] подтвердила свою высокую эффективность в задачах общего понимания русского языка, однако её производительность в медицинской сфере остаётся недостаточной. Например, в задаче ответов на вопросы из набора данных RuMedDaNet [3, 4], точность RuBERT составила лишь 67,19%, что значительно ниже человеческого уровня точности ответов врачей-специалистов (93,36%).

В данной работе предлагается подход к улучшению качества эмбеддингов для медицинских RAG-систем на русском языке. Использованная в исследовании методика включает тонкую настройку RuBERT на специализированном наборе данных RuMedDaNet, а также применение техники обучения представлений "матрёшка" (Matryoshka Representation Learning) [5] для оптимизации эффективности поиска. Исследование вносит вклад в развитие методов адаптации языковых моделей к узкоспециализированным доменам, уделяя особое внимание медицинскому NLP на русском языке — области, которая остается недостаточно изученной в сравнении с англоязычными аналогами.

ОБЗОР СВЯЗАННЫХ РАБОТ

Эффективность поиска (Retrieval) в RAG-системах во многом определяется качеством векторных представлений текста, генерируемых моделями эмбеддингов [1]. Эти модели преобразуют текстовые данные в плотные (то есть не разреженные) числовые векторы, что позволяет осуществлять семантический поиск на основе векторного сходства. Однако большинство современных моделей эмбеддингов обучаются на общедоступных корпусах текстов, что снижает их эффективность при работе со специализированными доменами.

Экспериментальные данные свидетельствуют о том, что адаптация предобученных моделей к конкретной предметной области способна существенно улучшить качество поиска. В частности, в работе [6] продемонстрировано, что тонкая настройка эмбеддингов на доменно-специфичных данных может повысить эффективность поиска на 7–22% в зависимости от размерности векторных представлений.

Модель RuBERT, разработанная DeepPavlov [2], является одной из наиболее значимых моделей для обработки русского естественного языка. Сохраняя оригинальную архитектуру BERT, эта модель была специально адаптирована для русского языка за счёт обучения на обширном корпусе текстов, включающем русскоязычную Википедию и новостные источники. Несмотря на успешное применение в различных NLP-задачах, таких как анализ тональности и языковое моделирование [6], эффективность RuBERT в узкоспециализированных областях, включая медицину, остаётся низкой. Это связано со специфичностью

медицинской терминологии и особенностями профессионального языка, которые слабо представлены в общедоступных обучающих данных.

Оценка возможностей языковых моделей в медицинской сфере осуществляется с помощью специализированных бенчмарков, в частности RuMedBench [2]. Результаты тестирования на данных этого бенчмарка демонстрируют существенный разрыв между человеческой и машинной производительностью. Наиболее показательным примером является задача ответа на вопросы из набора RuMedDaNet, где разница в точности человека и RuBERT достигает 20 процентов [2], что подчёркивает необходимость дальнейшей адаптации моделей для медицинских приложений.

Современные методы дообучения языковых моделей на специализированных данных демонстрируют высокую эффективность в профессиональных областях. Последние исследования сосредоточены на оптимизации обучения через специализированные форматы обучающих данных, такие как положительные пары (семантически близкие текстовые фрагменты), триплеты (наборы из исходного текста, его семантического аналога и противоречащего по смыслу текста) и ранжированные пары (тексты с оценкой степени их смысловой близости). Эти подходы генерируют дифференцированные обучающие сигналы, что позволяет точнее адаптировать векторные представления текста (эмбеддинги) к конкретным задачам.

Особый интерес представляет метод Matryoshka Representation Learning (MRL) [5], организующий эмбеддинги по принципу матрёшки. Ключевая идея заключается в том, чтобы распределить семантическую информацию по приоритету: наиболее значимые признаки кодируются в начальных компонентах вектора, а второстепенные — в последующих. Это позволяет в процессе обучения сокращать размерность представлений в несколько раз, сохраняя до 99,9% исходной точности. Благодаря компактности и высокой эффективности, MRL-эмбеддинги становятся оптимальным решением для систем, работающих в условиях ограниченных вычислительных ресурсов, что важно для развертывания модели в медицинских организациях.

Стоит отметить, что в настоящий момент существуют специализированные модели, такие как RuBioBERT и RuBioRoBERTa [7], которые были целенаправленно предобучены на крупных корпусах русскоязычных биомедицинских текстов (научные публикации, клинические рекомендации, выписки) и валидированы на наборах данных бенчмарка RuMedBench. Их появление является важным шагом в развитии медицинского NLP для русского языка. Как показано в [7], RuBioRoBERTa превосходит базовую модель RuBERT на 3–10% по ключевым метрикам, а в задаче RuMedNER даже превышает человеческую точность. Успех этих моделей подтверждает ценность доменно-специфичного предобучения для медицинского NLP. Однако эти модели предназначены для общего понимания медицинских текстов и используют фиксированную скрытую размерность эмбеддингов (1024). Исследование, предлагаемое в данной статье, имеет другую цель, оно направлено на получение компактных низкоразмерных векторных представлений для эффективного поиска в медицинских RAG-системах.

МАТЕРИАЛЫ И МЕТОДЫ

Данное исследование основывается на модели RuBERT, а именно на варианте DeepPavlov/rubert-base-cased. Базовая архитектура этой модели включает 12 слоёв трансформера с размером эмбеддингов в 768 измерений, что обеспечивает баланс между вычислительной эффективностью и размером векторных представлений [2]. Модель была предварительно обучена на большом русскоязычном наборе данных, что обеспечивает ей широкое понимание русского языка. Эта основа делает модель подходящей для дальнейшей адаптации к специализированным областям через тонкую настройку.

Для тонкой настройки использовался набор данных RuMedDaNet из бенчмарка RuMedBench [3]. Датасет RuMedDaNet предназначен для решения задачи ответов на вопросы типа "да/нет", охватывающую различные медицинские области, включая фармакологию, анатомию и терапевтическую медицину. Каждый пример в наборе данных состоит из: медицинского контекста, предоставляющего основную информацию; вопроса, связанного с этим контекстом; ответа "да" или "нет".

Пример:

Контекст: "Эпилепсия — это хроническое полиэтиологическое заболевание головного мозга, доминирующим проявлением которого являются повторяющиеся эпилептические приступы, возникающие вследствие

повышенного гиперсинхронного разряда нейронов головного мозга."

Вопрос: "Эпилепсия является заболеванием головного мозга человека?"

Ответ: "да" [3]

Выбор набора данных RuMedDaNet для тонкой настройки модели оптимизации эмбеддингов был обусловлен рядом ключевых факторов. Во-первых, RuMedDaNet содержит структурированные тройки "медицинский контекст — вопрос — ответ (типа "да/нет")", что идеально соответствует задаче оптимизации поиска релевантных контекстов для RAG-систем и отвечает цели улучшения точности анализа медицинских запросов через повышение качества эмбеддингов. Во-вторых, RuMedDaNet является частью общепризнанного бенчмарка RuMedBench, используемого для оценки русскоязычных медицинских моделей, что обеспечивает простую воспроизводимость результатов. В-третьих, четкие связи "вопрос-контекст" позволяют генерировать позитивные (релевантные) и негативные (нерелевантные) пары для обучения эффективных эмбеддингов.

В рамках данного исследования набор данных RuMedDaNet был адаптирован для обучения моделей построения векторных представлений текста. На основе этого набора данных были сформированы семантически связанные пары вопросов и соответствующих им контекстов, которые стали основой для дообучения модели. Для оптимизации процесса использовалась функция потерь Multiple Negatives Ranking Loss [8], которая учит модель отличать релевантные данные от случайных, используя отрицательные примеры (то есть примеры с ответом “нет”) из текущего набора данных.

Для повышения эффективности обучения был реализован подход, который можно описать как структурированное контекстно-контрастное представление данных. Его ключевая идея заключается в особом расположении семантически связанных примеров в обучающей выборке, когда тексты со схожей тематикой или близкими терминами в процессе обучения подаются модели последовательно. Это создаёт эффект усиленного контрастного восприятия — модель начинает тоньше различать нюансы между похожими медицинскими понятиями (такими как "гипертензия" и "гипертония"), одновременно сохраняя способность к широкому обобщению на уровне крупных тематических разделов (кардиология, онкология и др.). Такой подход представляется важным именно для медицинских текстов, где необходима способность модели различать терминологические нюансы. Медицинская лексика часто содержит ситуативные синонимы ("инфаркт" и "ОИМ"), контекстно-зависимые трактовки терминов, а также тонкие различия между клинически близкими состояниями. Традиционные методы обучения могут "размывать" эти различия, тогда как предлагаемая стратегия может повысить способность модели улавливать подобные детали.

Процесс тонкой настройки был реализован с использованием фреймворка SentenceTransformers [8], который предоставляет эффективные инструменты для обучения и оптимизации моделей эмбеддингов. Был применен многоэтапный подход к обучению.

Сначала модель была настроена с использованием функции потерь Multiple Negatives Ranking Loss (1) для формирования прочной основы медицинских знаний.

$$\text{Loss} = \sum_{i=1}^{P} \sum_{j=1}^{N} \max(0, f(q, p_i) - f(q, n_i) + \text{margin}) \quad (1)$$

где: P — количество положительных примеров, N — количество отрицательных примеров, q — входной запрос, p_i — i-й положительный пример, n_j — j-й отрицательный пример, f — функция, измеряющая схожесть между векторами запроса и примера, margin — гиперпараметр, определяющий желаемое разделение между положительными и отрицательными примерами.

На следующем этапе была использована функция потерь Matryoshka Loss для оптимизации модели по производительности на нескольких размерах эмбеддингов (64, 128, 256, 512 и 768).

В завершение был проведен процесс доработки с комбинацией перечисленных функций потерь для обеспечения согласованности между различными целями оптимизации.

Формула функции потерь Matryoshka Loss (2) выглядит следующим образом:

$$\text{Loss}_{\text{MRL}} = \sum_{d_i \in M} w_{d_i} \cdot \text{Loss}_{d_i} \quad (2)$$

Таблица 1 — Гиперпараметры модели

Размер батча	16
Скорость обучения	$2e^{-5}$
Количество эпох	4
С линейным разогревом и спадом	да
Весовое затухание (weight decay)	0,01
Ограничение градиента (gradient clipping)	1,0
Максимальная длина последовательности	512

где: w_{d_i} — вес, назначенный размерности d_i, $Loss_{d_i}$ — функция потерь для задачи (например, кросс-энтропия или ранжирование), применяемая к эмбеддингам размером d_i.

Гиперпараметры обучения, представленные в таблице 1, были тщательно подобраны на основе предварительных экспериментов.

Для повышения скорости обучения и эффективной работы с длинными последовательностями в данном исследовании использовался метод Flash Attention 2 [9], реализованный через механизм SDPA (Scaled Dot-Product Attention) [10]. Этот подход оптимизирует вычисления за счёт перераспределения операций внимания [11], что позволило значительно ускорить обработку сложных текстовых данных. Применение Flash Attention 2 не только сократило время обучения, но и улучшило точность анализа за счёт более глубокого учёта контекстных связей в длинных медицинских описаниях. Это особенно важно для задач, требующих одновременного анализа множества взаимосвязанных терминов, таких как диагностические критерии или фармакологические взаимодействия.

Для проверки качества поиска с использованием предложенной модели эмбеддингов использовалась тестовая часть набора данных RuMedDaNet, на основе которой был сформирован тестовый корпус медицинских контекстов (исходный датасет [3], составленный медицинскими экспертами, содержит 1564 примера для обучения и валидации и 512 примеров для тестирования). Целью эксперимента было определить, насколько точно модели извлекают релевантные фрагменты текста по заданному вопросу.

Производительность модели оценивалась с помощью двух ключевых поисковых метрик: NDCG (Normalized Discounted Cumulative Gain), которая учитывает не только наличие релевантных документов в ответе на запрос, но и их позицию в результатах поиска и MRR (Mean Reciprocal Rank), отражающую среднюю позицию первого корректного ответа.

Формула расчета NDCG имеет вид [12]:

$$NDCG = \frac{DCG}{IDCG} \quad (3)$$

где:
$$DCG = \sum_{i=1}^{k} \frac{(2^{r_i} - 1)}{log_2(i + 1)},$$

$$IDCG = \sum_{i=1}^{k} \frac{(2^{r_i^*} - 1)}{log_2(i + 1)}$$

где: k — количество рассматриваемых ответов на запрос, r_i — релевантность результата на позиции i в ранжированном списке ответов на запрос, r_i^* — релевантность результата на позиции i в идеально отсортированном списке.

Формула расчета MRR выглядит следующим образом [13]:

$$MRR = \frac{1}{N} \sum_{i=1}^{k} \frac{1}{rank_i}, \quad (4)$$

где: N — количество запросов, $rank_i$ — позиция первого релевантного результата для i-го запроса.

Эксперименты проводились для эмбеддингов разной размерности от 64 до 768 измерений. Такой подход позволил проанализировать эффективность метода Matryoshka Representation Learning, который оптимизирует распределение информации в векторах, сохраняя высокую точность даже при сокращении их размера.

Исследование выполнялось в среде Python версии 3.9.1. с использованием графического процессора NVIDIA RTX 3090 (24 ГБ VRAM). Для загрузки и предварительной обработки датасета использовалась библиотека datasets версии 3.3.1. Для построения, обучения и оценки моделей эмбеддингов применялись библиотеки sentence_transformers версии 3.3.1 и torch версии 2.5.1. Для публикации и управления моделями в репозитории Hugging Face Hub использовалась библиотека huggingface_hub версии 0.29.0.

РЕЗУЛЬТАТЫ

Первоначальная оценка эффективности проводилась на оригинальной версии модели

RuBERT без дополнительной доменной адаптации. Результаты, представленные в Таблице 2, отражают значения метрики NDCG при различных размерностях векторных представлений.

Анализ данных выявил устойчивую зависимость между размерностью эмбеддингов и качеством поиска: с уменьшением количества измерений наблюдается постепенное снижение показателей эффективности. Такая закономерность объясняется фундаментальным ограничением базовых моделей — недостаточной емкостью низкоразмерных векторных представлений для кодирования всей необходимой семантической информации. Особенно критичным это становится при работе со специализированными медицинскими терминами и понятиями, требующими точного различения смысловых нюансов. Полученные результаты подтверждают необходимость адаптации моделей для работы с узкоспециализированными доменами. После тонкой настройки на наборе данных RuMedDaNet наблюдаются существенные улучшения в производительности поиска на всех размерах эмбеддингов, что также отражено в таблице 2.

Улучшения присутствуют и по другой метрике оценки. Таблица 3 показывает значения метрики MRR для обеих моделей (базовой и дообученной). Результатом является значительное улучшение показателей во всех размерностях, особенно заметное на эмбеддингах с меньшей размерностью. Такая закономерность свидетельствует о том, что тонкая настройка позволяет модели эффективнее выделять ключевую информацию в каждом измерении, улучшая общую эффективность без снижения производительности.

Проведенные эксперименты подтвердили эффективность подхода Matryoshka Representation Learning. Анализ сохранения относительной производительности модели на разных размерностях эмбеддингов в сравнении с полными 768-мерными представлениями представлен в Таблице 4.

Результаты показывают, что модель, настроенная с использованием предлагаемого подхода, сохраняет более 94% производительности при использовании всего 128 измерений, что соответствует шестикратному уменьшению исходного размера. Даже при сокращении размерности эмбеддингов до 64 (уменьшение исходного размера в 12 раз) модель демонстрирует

Таблица 2 — Сравнение значений метрики NDCG между базовой и настроенной моделями

Размерность	Базовая модель	Настроенная модель	Улучшение
768	0,5128	0,7201	40,43%
512	0,5064	0,7018	38,59%
256	0,4891	0,7001	43,14%
128	0,4523	0,6778	49,86%
64	0,3986	0,6196	55,44%

Таблица 3 — Сравнение значений метрики MRR между базовой и настроенной моделями

Размерность	Базовая модель	Настроенная модель	Улучшение
768	0,4832	0,6720	39,07%
512	0,3773	0,6606	75,09%
256	0,3612	0,6581	82,20%
128	0,3324	0,638	91,93%
64	0,2743	0,5799	111,41%

Таблица 4 — Сохранение производительности эмбеддингов с уменьшенной размерностью относительно 768-мерных эмбеддингов

Размерность	% полной производительности (NDCG)	% полной производительности (MRR)
512	97,46%	98,03%
256	97,22%	97,93%
128	94,12%	94,94%
64	86,04%	86,29%

более 86% от исходной эффективности. Такой баланс между уменьшением размерности и сохранением точности делает данный подход особенно полезным для систем с ограниченными вычислительными ресурсами.

На практике уменьшение размерности эмбеддингов в 6 раз позволит пропорционально снизить размеры векторных баз данных в медицинских RAG-системах и повысить скорость обработки запросов к большим языковым моделям, что сегодня является крайне актуальной задачей.

Предлагаемая в данном исследовании модель получила название Med-Bert-Matryoshka-v1 и была загружена на платформу Hugging Face, где

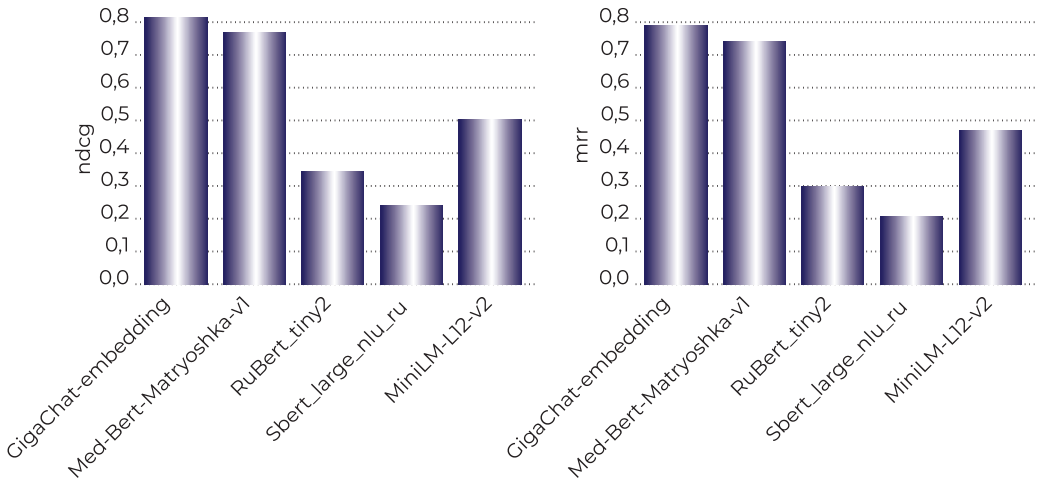

Рисунок 1 — График сравнения метрик NDCG и MRR.

теперь она доступна для открытого использования и дальнейшего развития сообществом специалистов в области искусственного интеллекта [14].

Для оценки качества работы модели в задачах информационного поиска было проведено сравнение с несколькими ведущими моделями построения эмбеддингов. В качестве метрик использовались MRR и NDCG с мерой косинусного сходства.

Среди сравниваемых моделей эмбеддингов рассматривались:

1) GigaChat-embedding [15] — проприетарная модель для построения эмбеддингов, отличающаяся высокой производительностью;
2) Med-Bert-Matryoshka-v1 [14] — модель, предлагаемая в данном исследовании для повышения эффективности поиска в медицинских векторных базах данных;
3) MiniLM-L12-v2 [16] — компактная модель построения эмбеддингов, разработанная для высокой вычислительной эффективности;
4) RuBert_tiny2 [17] — облегчённая версия RuBERT, адаптированная для работы с русскоязычными данными;
5) Sbert_large_nlu_ru [18] — крупноразмерная модель, ориентированная на обработку естественного русского языка.

Результаты исследования (рисунок 1) показали, что GigaChat-embedding, будучи коммерческой моделью, демонстрирует наивысшую эффективность по обеим метрикам (MRR и NDCG). Это объясняется значительными вычислительными ресурсами, задействованными при её обучении и оптимизации.

Тем не менее, обученная в данном исследовании open source модель Med-Bert-Matryoshka-v1 также показала высокую конкурентоспособность, заняв второе место с показателями метрик MRR = 0,71 и NDCG = 0,75. Это свидетельствует о её способности обеспечивать точное и релевантное извлечение информации из текстовых данных при формировании моделей эмбеддингов, что особенно важно в RAG — системах.

Остальные модели продемонстрировали более скромные результаты. MiniLM-L12-v2 достигла значений около 0,5 по обеим метрикам, что указывает на её ограниченную эффективность в задачах поиска, несмотря на низкую вычислительную сложность. RuBert_tiny2 и Sbert_large_nlu_ru показали ещё более низкие результаты — в диапазоне 0.3–0.35 и 0.2–0.25 соответственно, что свидетельствует о неэффективности их применения в задачах информационного поиска медицинских данных по сравнению с исследуемой моделью.

Таким образом, проведённое сравнение подтверждает, что модель Med-Bert-Matryoshka-v1, представляющая собой RuBERT, дообученную на русскоязычных медицинских данных, является эффективной моделью получения эмбеддингов для решения задач поиска информации, уступив в вычислительном эксперименте только коммерческому решению GigaChat-embedding.

ОБСУЖДЕНИЕ

Результаты данного исследования, показывающие значительный прирост качества после тонкой настройки на медицинских данных

(Таблицы 2, 3), полностью согласуются с выводом авторов RuMedBench [3, 4] о том, что универсальные языковые модели демонстрируют существенный разрыв в эффективности при работе со специализированной медицинской терминологией по сравнению с человеческим уровнем. Данное исследование предлагает один из возможных способов для преодоления этого разрыва.

Хотя базовая модель RuBERT показывает высокую эффективность в обработке общеупотребительной лексики, её применение в медицинских контекстах является ограниченным из-за недостатка специализированных знаний. Проведённая донастройка на корпусе RuMedDaNet позволяет устранить этот пробел, значительно повышая точность распознавания и кодирования медицинских данных.

Применение метода Matryoshka Representation Learning (MRL) позволило не только повысить точность, но и добиться существенного сокращения вычислительных затрат. Это согласуется с результатами работы [5] о том, что ключевая семантическая информация может быть эффективно упакована в начальных измерениях вектора. Проведенный в данном исследовании эксперимент показал, что даже при шестикратном сжатии (с 768 до 128 измерений) модель Med-Bert-Matryoshka-v1 сохраняет более 94% эффективности. Это открывает возможности для существенного снижения вычислительных затрат на генерацию эмбеддингов, их индексирование и поиск — крайне важный фактор для ресурсоёмких сред, таких как мобильные медицинские приложения или системы с ограниченными аппаратными возможностями.

Хотя тонкая настройка существенно улучшила качество работы модели, анализ ее ошибок позволил выявить ряд устойчивых проблем. В частности, модель демонстрирует снижение эффективности в следующих случаях.

Сложные вопросы, требующие многошаговых рассуждений. Модель не всегда корректно обрабатывает запросы, предполагающие анализ взаимосвязей между несколькими медицинскими концепциями, что является общей проблемой для современных языковых моделей в медицине [19]. Это указывает на то, что, несмотря на успешное усвоение терминологии, система не всегда способна распознавать сложные логические цепочки.

Редкие термины и узкоспециализированные процедуры. Вопросы, содержащие малоупотребительную лексику или редко встречающиеся в обучающих данных медицинские процедуры, обрабатываются менее точно. Данное ограничение подтверждает необходимость использования более репрезентативных и разнообразных наборов данных при дообучении.

Неоднозначные формулировки и зависимость от контекста. Запросы, допускающие множественную интерпретацию, нередко приводят к ошибкам.

Подобные случаи иллюстрируют сложность работы с медицинским языком, где точность формулировок и понимание контекста играют существенную роль. Выявленные ограничения определяют перспективные направления для дальнейших исследований. Среди возможных улучшений — расширение обучающей выборки за счёт более разнородных медицинских текстов, включая данные, ориентированные на развитие навыков клинического мышления [19], а также разработка специализированных методов обучения, лучше учитывающих семантические нюансы медицинской терминологии, интеграция формализованных медицинских знаний и онтологий в процесс обучения, как это предлагается, например, в работе [20].

Как уже отмечалось, существуют и другие успешные подходы к адаптации языковых моделей для медицины, такие как RuBioRoBERTa [7], которая показала превосходство над RuBERT в задачах общего понимания медицинских текстов. Однако данное исследование фокусируется на иной, но не менее важной задаче — оптимизации компактных векторных представлений именно для эффективного поиска в RAG-системах, а не на решении широкого круга NLP-задач. Это объясняет и выбранные в исследовании метрики оценки.

ЗАКЛЮЧЕНИЕ

Проведённое исследование демонстрирует существенное улучшение качества семантических эмбеддингов, используемых для информационного поиска в рамках архитектуры RAG. Оптимизированная модель Med-Bert-Matryoshka-v1 обеспечивает более точную релевантность извлекаемых данных, что непосредственно влияет на достоверность генерируемых ответов на медицинские запросы и важно для обеспечения клинической надежности ИИ-решений.

Перспективными направлениями для дальнейших исследований могут стать: расширение спектра используемых русскоязычных медицинских текстов для тонкой настройки, интеграция формализованных медицинских знаний в процесс обучения, а также разработка специализированных функций потерь, учитывающих особенности медицинской терминологии.

Конфликт интересов. Авторы декларируют отсутствие явных и потенциальных конфликтов интересов, связанных с публикацией настоящей статьи.

Финансирование. Авторы заявляют, что не получали финансовой поддержки при проведении данного исследования, написании и/или публикации данной статьи.

ЛИТЕРАТУРА/REFERENCES

1. Lewis P, Perez J, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. Adv Neural Inf Process Syst. 2020; 33: 9459-9474. doi: 10.48550/arXiv.2005.11401.
2. Kuratov Y, Arkhipov M. Adaptation of deep bidirectional multilingual transformers for Russian language. arXiv. Preprint posted online May 17, 2019. doi: 10.48550/arXiv.1905.07213.
3. MedBench [Internet]. Открытый набор задач в области здравоохранения [cited 2025 May 3]. Available from: https://medbench.ru/
4. Blinov P, Chertok A, Drozdov A, et al. RuMedBench: a Russian medical language understanding benchmark. Artif Intell Med. 2022; 383-392. doi: 10.1007/978-3-031-09342-5_38.
5. Kusupati A, Ordonez V, Parikh D, et al. Matryoshka representation learning. Adv Neural Inf Process Syst. 2022; 35: 30233-30249. doi: 10.48550/arXiv.2205.13147.
6. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. Proc NAACL-HLT. 2019; 1: 4171-4186. doi: 10.18653/v1/N19-1423.
7. Yalunin A, Nesterov A, Umerenkov D. RuBioRoBERTa: a pre-trained biomedical language model for Russian language biomedical text mining. arXiv. Preprint posted online April 8, 2022. doi: 10.48550/arXiv.2204.03951.
8. Reimers N, Gurevych I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. Proc EMNLP-IJCNLP. 2019; 3982-3992. doi: 10.18653/v1/D19-1410.
9. Dao T, Fu T, Ermon S, et al. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. Adv Neural Inf Process Syst. 2022; 35: 16344-16359. doi: 10.48550/arXiv.2205.14135.
10. Dao T. Flashattention-2: Faster attention with better parallelism and work partitioning. arXiv. Preprint posted online July 17, 2023. arXiv: 2307.08691. doi: 10.48550/arXiv.2307.08691.
11. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Adv Neural Inf Process Syst. 2017; 30. doi: 10.48550/arXiv.1706.03762.
12. Wang Y, et al. A theoretical analysis of NDCG type ranking measures. J Mach Learn Res. 2013; 25-54. doi: 10.48550/arXiv.1304.6480.
13. Craswell N. Mean Reciprocal Rank. In: Liu L, Özsu MT, eds. Encyclopedia of Database Systems. Springer; 2009. doi: 10.1007/978-0-387-39940-9_488.
14. TrungKiencding. Med-Bert-Matryoshka-v1 [Internet]. Hugging Face [cited 2025 May 3]. Available from: https://huggingface.co/TrungKiencding/Med-Bert-Matryoshka-v1.
15. GigaChatEmbeddings [Internet]. [cited 2025 May 3]. Available from: https://deepwiki.com/ai-forever/gigachain/3-gigachatembeddings.
16. Sentence-transformers/all-MiniLM-L12-v2 [Internet]. Hugging Face [cited 2025 May 3]. Available from: https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2.
17. Cointegrated/rubert-tiny2 [Internet]. Hugging Face [cited 2025 May 3]. Available from: https://huggingface.co/cointegrated/rubert-tiny2.
18. Ai-forever/sbert_large_nlu_ru [Internet]. Hugging Face [cited 2025 May 3]. Available from: https://huggingface.co/ai-forever/sbert_large_nlu_ru.
19. Jin, D., Pan, E., Oufattole, N., Weng, W. H., Fang, H., & Szolovits, P. (2021). What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams. Applied Sciences, 11(14), 6421. doi: 10.3390/app11146421.
20. Радюш Д.В. Методы интеграции знаний для разработки вопросно-ответных систем. Russian Technological Journal. — 2025. — №13(3). — С.21-43. [Radyush DV. Knowledge injection methods in question answering. Russian Technological Journal. 2025; 13(3): 21-43. (In Russ).] doi: 10.32362/2500-316X-2025-13-3-21-4300.