ПРИМЕНЕНИЕ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ В ЗДРАВООХРАНЕНИИ ОБЩИННЫХ МЕДИЦИНСКИХ УЧРЕЖДЕНИЙ

Пойкалайнен А.М., Кочкаров Р.А.
Финансовый университет при Правительстве РФ, Москва, Россия
ampojkalajnen@ fa.ru, rkochkarov@fa.ru

Аннотация. Для решения проблемы нехватки медицинских работников к 2030 году предполагается внедрение решений на основе больших языковых моделей и автоматизированного машинного перевода речи. Такие решения созданы преодолевать языковые барьеры, культурные различия и ограниченный доступ к данным, повышая качество и доступность медицинских услуг в странах с низким и средним уровнем дохода населения.

Ключевые слова: большие языковые модели, машинный перевод, качество медицинских услуг, общинные медицинские работники.

Введение

Проблемы, связанные с недостатком медицинской инфраструктуры, особенно в сельских регионах, и нехватка квалифицированных медицинских кадров определяют необходимость в гибких и доступных медицинских услугах [1]. Особенно это заметно в странах Африки и Азии с развивающейся экономикой, где нехватка медицинских работников, включая врачей и медсестер, является серьезной проблемой. В этих условиях работники общинных медицинских учреждений играют важную роль в сокращении разрыва в доступности здравоохранения, несмотря на проблемы, связанные с отсутствием формального обучения и поддержки. Важное значение имеет потенциал больших языковых моделей (Large Language Models, LLM) в повышении квалификации медицинских работников, что может способствовать улучшению качества медицинской помощи. Однако внедрение LLM в практику работы медицинских работников в странах с низким и средним доходом населения сталкивается с трудностями, связанными с необходимостью адаптации технологий к специфическим условиям и культурным особенностям.

В настоящей работе предлагается основные этапы разработки и внедрения LLM с целью преодоления выявленных препятствий и совершенствования LLM в здравоохранении в условиях ограниченных ресурсов. Основываясь на адаптации LLM к лингвистическим, культурным и контекстуальным особенностям, стратегия включает в себя разработку механизмов постоянной обратной связи с медицинскими работниками и применение передовых технологий машинного перевода для расширения их возможностей. Такие LLM, как ChatGPT и MedPaLM 2, достигли успеха в решении образовательных и операционных задач медицинскими работники, способствуя повышению эффективности системы здравоохранения в неблагополучных районах. Инновационные подходы к обучению и принятию медицинских решений через LLM могут значительно повлиять на качество медицинской помощи посредством доступа к современным медицинским знаниям и информации. Тем не менее, требуется внимание к адаптации инноваций к реальным условиям стран с низким уровнем дохода населения и преодолению препятствий к качественному здравоохранению.

1. Модель

1.1. Цели разработки на основе LLM

Целью разработки моделей LLM является обеспечить возможность медицинскому персоналу работать более эффективно и безопасно, уделяя внимание решению критических проблем здравоохранения, определенных задачами устойчивого развития ВОЗ для стран с низким уровнем дохода населения. Система должна обеспечивать:

- упрощенные коммуникации: получение работниками здравоохранения краткой, легко усваиваемой диагностической и образовательной информации, избегая сложного медицинского жаргона и аббревиатур для лучшего понимания;
- безопасность и точность контента: наличие механизмов проверки входных и выходных данных на предмет токсичности, дезинформации или ошибочных рекомендаций, обеспечивая обмен надежной и достоверной информацией;
- модульность и масштабируемость: способность адаптации к различным географическим и культурным особенностям для повышения универсальности применения инструментария.

- культурную чувствительность и локализацию: адаптация контента к местным языкам, обычаям и системам верований.

1.2. Разработка модели

Процесс начинается со сбора специализированного обучающего корпуса, в который входят наборы данных медицинских диалогов и рекомендации региональных органов здравоохранения, чтобы обеспечить актуальность и применимость модели в сфере коммуникаций в здравоохранении.

Далее проводится оценка различных адаптированных к данной области LLM их многоязыковые возможности и актуальность для медицинских работников. На основе набора определенных критериев выбираются модели для дальнейшей доработки. Модели дорабатываются с использованием собранного набора диалогов и рекомендаций для повышения их производительности и соответствия целевому состоянию.

Кроме того, для расширения охвата системы в различных языковых регионах рассматривается включение моделей машинного перевода, что позволяет решить проблему масштабируемости и регионального языкового контекста. На этом этапе оценивались различные модели перевода на предмет их точности и эффективности в медицинских диалогах.

Далее проводится всесторонняя оценка моделей, используется множество методик для тщательного изучения эффективности моделей с целью обеспечения их надежного и безопасного применения в медицинских учреждениях. Такие методики включают как количественный, так и качественный анализ, направленный на подтверждение эффективности, надежности и соответствия моделей стандартам медицинской коммуникации.

Также предлагаются способы интеграции полученных результатов в систему здравоохранения в странах со средним и низким уровнем дохода населения и перспективы развития.

Было проведено исследование [2] по применению БЯМ и моделей машинного перевода в области здравоохранении в регионах с ограниченными ресурсами (районы, где могут возникать проблемы с доступом к медицинским услугам, оборудованием и другими необходимыми вещами). Авторы исследования [2] уделили внимание оценке способности этих моделей предоставлять надежные медицинские рекомендации на ресурсоограниченных языках (языки, для которых ограничены лингвистические данные и ресурсы, доступные для задач NLP): телугу, хинди, суахили и арабском. На примере этого исследования строится дальнейший анализ применения перспектив развития БЯМ в здравоохранении общинных медицинских учреждений.

Обучение и данные

В наборы медицинских данных входят клинические рекомендации и руководства, предоставленные региональными органами здравоохранения, а также медицинские диалоги, состоящие из разговоров между медицинскими работниками и пациентами, и данные о неблагоприятных событиях, полученные из системы сообщений о неблагоприятных событиях (Food and Drug Administration Adverse Event Reporting System - FAERS).

Основное внимание уделялось трем основным причинам, приводящим к снижению продолжительности жизни с поправкой на инвалидность (DALYs - Disability Adjusted Life Years) [3]: ишемическая болезнь сердца (ИБС), инфекции нижних дыхательных путей (ИВДП) и неонатальный уход. DALYs представляет собой комплексный показатель, являющимся важным критерием оценки медицинского обслуживания. Акцент на DALYs позволяет фокусировать обучение на областях, оказывающих существенное влияние на здоровье, гарантируя соответствие глобальным приоритетам здравоохранения и эффективный вклад в решение наиболее актуальных проблем здравоохранения.

Наборы данных, за исключением двух диалоговых наборов данных на китайском языке, были на английском языке, поэтому возникла необходимость перевода этих наборов на несколько языков: телугу, хинди, арабский и суахили, в качестве переводчика использовался Azure AI Translator. Во избежание проблемы отсутствия разговорных выражений из-за буквального перевода терминов была проведена замена буквально переведенных терминов на общеупотребительные и культурно-чувствительные фразы и слова, тем самым повышая доступность и актуальность переведенного контента для медицинских работников.

Оценка базовых моделей

Создание и обучение базовых моделей искусственного интеллекта, таких как GPT-4[4] и Llama 2 [5], использующих обширный корпус медицинских данных из таких ресурсов, как PubMed [https://pubmed.ncbi.nlm.nih.gov/], показало обнадеживающие результаты. Они способны отвечать на

запросы на разных языках, что делает их широко применимыми. Однако, способность предоставлять медицинские рекомендации на ресурсоограниченных языках (языки, для которых ограничены лингвистические данные и ресурсы, доступные для задач NLP) требует изучения. Исследование [2] посвящено применению предварительно обученных LLM в медицинских учреждениях с целью повышения эффективности оказания медицинской помощи путем преодоления языковых барьеров, особенно в регионах с ограниченными ресурсами. В этом исследовании используются две основные стратегии, направленные на максимальное использование потенциала передовых инструментов ИИ.

Во-первых, было рассмотрено прямое применение моделей ИИ в интерпретации медицинских диалогов на нескольких языках, что позволило оценить их пригодность для предоставления надежных рекомендаций без каких-либо языковых или специфических для данной области корректировок. Во-вторых, учитывая, что существующие модели не могут полностью понять и ответить на сложные медицинские запросы на ресурсоограниченных языках, была изучена возможность интеграции эффективных инструментов машинного перевода. Предполагалось адаптировать подсказки на ресурсоограниченные языки, а затем использовать специфические для конкретной области обученные модели на английском языке. Что позволяет не только преодолеть проблемы, связанные с языковыми барьерами, но и улучшить ответы моделей, обеспечив их тонкое понимание, необходимое для получения медицинских рекомендаций.

Первоначальный подход предполагал оценку эффективности моделей на наборах данных медицинских диалогов, созданных на четырех языках - телугу, хинди, суахили и арабском, - путем определения семантического сходства между ответами, созданными ИИ, и их эквивалентами, отредактированными человеком. Длинные текстовые ответы, генерируемые ИИ, не подходят для обычных метрик, таких как метрика Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [6], потому что метод ROUGE не позволяет эффективно улавливать контекстно-зависимую информацию, являющейся важнейшим компонентом оценки семантической согласованности. Поэтому ответы оценивались с помощью многоязычного BERT (mBERT) [7]. Эта модель использует контекстуальные вкрапления для лучшего понимания нюансов смысла, заложенного в тексте, обеспечивая более сложный анализ семантического соответствия между ответами, сгенерированными моделью ИИ, и эталонными ответами.

В рамках пересмотренного подхода основное внимание уделяется оценке комплексной модели, использующей возможности двух различных моделей. Подход с двумя моделями необходим для разделения и специализации двух функций: машинного перевода и генерации ответов. Операция тонкой настройки fine-tuning [8] направлен на определение преимуществ комплексной модели (ансамбля моделей) по сравнению отдельных LLM в обеспечении точного перевода и контекстуально точных ответов.

Операция Supervised Fine-tuning

Fine-tuning реализуется на двух уровнях: оптимизация языковой модели, адаптированной к медицинской области в части расширения возможностей диагностики и оказания медицинской помощи; улучшение машинного перевода за счет интеграции контекстных данных для упрощения контекстной локализации.

Модель Biomistral [9], адаптированная к медицинскому контексту, показала лучший результат на четырех языках по сравнению с Meditron 70B [10], GPT-4 [3] и Llama 2 [4]. А в качестве модели машинного перевода использовалась Seamless M4T v2 Large [11] с открытым исходным кодом с возможностью настройки параметров, присущих общению в сфере здравоохранения с учетом культурных особенностей.

Для уменьшения ошибок в комплексной модели использовались межъязыковые эмбединги (векторизация слов) для обеспечения семантической согласованности в английском, телугу, хинди, суахили и арабском языках.

2. Оценка моделей Biomistral и Seamless M4T v2

Процесс оценки проводится по двум отдельным, но взаимосвязанным параметрам, каждый из которых специально сформирован для проверки работоспособности модели для общинных медицинских работников:

- производительность, дообученной на корпусе медицинских данных, модели Biomistral: показатель оценивает способность модели точно понимать и обрабатывать специализированную медицинскую терминологию; оценивается работа модели в диагностических ситуациях и ее способность формировать корректное предложение медицинской помощи.

- точность модели перевода: исследуется точность перевода – насколько хорошо модель сохраняет содержательную составляющую медицинской информации на разных языках; является важным аспект для безошибочной коммуникации в различных языковых средах.

2.1. Производительность модели Biomistral

Традиционно для оценки адаптированных к домену моделей в здравоохранении в основном использовались эталонные наборы данных, основанные на базах медицинских вопросов-ответов (QA), таких как PubMedQA [https://pubmedqa.github.io/] и MedQA [https://paperswithcode.com/dataset/medqa-usmle]. Хотя эталоны являются ценными инструментами для сравнения, они не всегда учитывают в ответах специфические потребности и нюансы, связанные с оказанием медицинской помощи в сельской местности.

Учитывая это ограничение, была предложена новая система для эффективной оценки применимости и полезности генерируемых ИИ ответов в условиях сельского здравоохранения [2]. Система оценки характеризуется вниманием к двум пользовательским наборам данных: (1) набор медицинских запросов, относящихся к работе медработников в сельской местности и (2) группа медицинских рекомендаций, отражающих практические и ситуационные потребности, возникающие в этих условиях. Для покрытия широкого спектра информационных потребностей наборы данных были сформированы по диагностическим запросам и рекомендательным запросам.

Процесс оценки включает в себя отправку запросов из наборов данных в настроенную медицинскую модель и проверку ответов на предмет их точности и полезности. Чтобы обеспечить всестороннюю и надежную оценку, сгенерированные ответы подвергаются проверке с помощью двух усовершенствованных языковых моделей: GPT-4 [3] и API Claude Opus [12]. Ответ считается верифицированным только в том случае, если он получил подтверждение (как семантическое, так и контекстуальное сходство) от обеих моделей, что свидетельствует о высокой степени надежности и релевантности. И наоборот, если ни один из API не подтверждает сгенерированный ответ, он классифицируется как неприемлемый.

Рис. 1 демонстрирует сравнительную эффективность итоговой модели Gen model по сравнению с моделью Biomistral с точки зрения создания релевантных ответов и применения понятной терминологии. Гистограммы показывают явное преимущество итоговой модели в том, что сложная медицинская информация становится более доступной для медработников. Улучшение читабельности и доступности ответов имеет решающее значение для эффективной коммуникации в медицинских учреждениях.

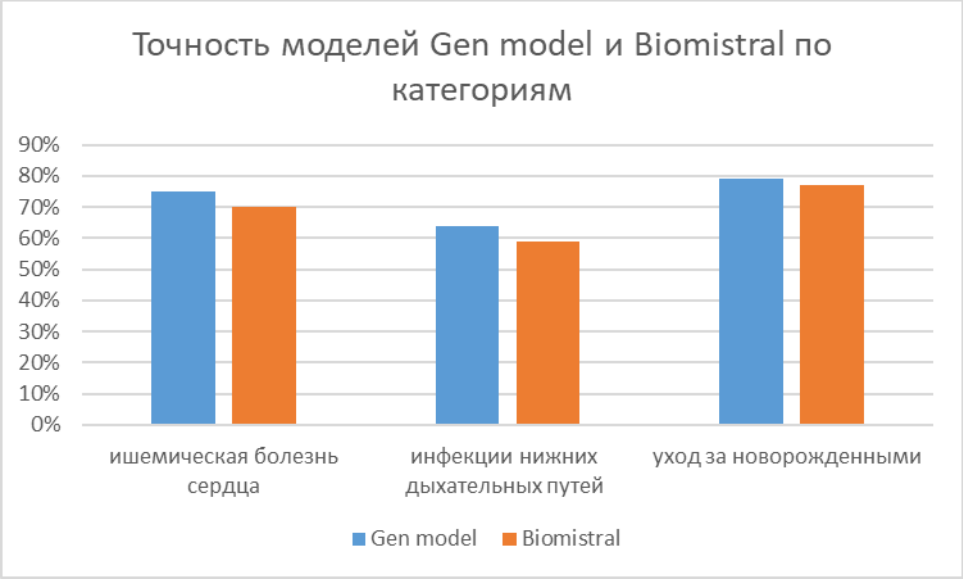

Рис. 1. Точность моделей Gen Model и Biomistral по трем основным категориям DALY

2.2. Оценка точности модели перевода

Также была проведена тщательная оценка точности модели машинного перевода (Seamless M4T v2 Large) [11], особенно способность работать с медицинской терминологией и повседневными выражениями. Для оценки эффективности машинного перевода использовалась метрика BLEU (Bilingual Evaluation Understudy) [13]. Как видно из табл. 1 [2], производительность модели перевода,

прошедшей Fine-tuning, значительно улучшилась для индийских языков (хинди, телугу) и арабского языка, однако при работе с суахили оказались небольшими. Отмечается, что на процесс перевода сильно влияют контекстуальные факторы, такие как пол и возраст, что иногда приводит к ошибкам. Например, «двухлетняя девочка» может быть неточно переведена как «женщина». Кроме того, при переводе некоторых медицинских терминов, таких как "ХОБЛ" (хроническая обструктивная болезнь легких), на языки телугу или хинди, возникали значительные трудности. Что указывает на трудности точной передачи специализированного медицинского языка в различных языковых контекстах. Все это указывает на исключительную важность понимания контекста и точности при переводе медицинской.

Таблица 1. Производительность Seamless M4T v2 Large до и после Fine-tuning

Языки	**Seamless M4T v2 Large**	
	До Fine-tuning	**После Fine-tuning**
Телугу	75.6	82.4
Хинди	73.4	83.1
Суахили	45.8	48.1
Арабский	68.5	80.5

3. Комплексная модель Gen model (general model)

Комплексная модель Gen model [2] объединяет все настроенные компоненты (параметры) и модели для создания надежного ассистента медицинских работников. Как показано на рис. 2, медицинские работники предоставляют информацию на своем местном языке, которая переводится на английский с помощью высокоточного переводчика. Переведенный текст проходит проверку на надежность и безопасность. После проверки текст обрабатывается итоговой моделью для создания ответа. Ответ снова проверяется защитными ограждениями на качество и безопасность, а затем переводится на местный язык, на котором говорит медицинский работник. Для обеспечения актуальности, предотвращения взлома, уменьшения галлюцинаций (шумов) и обеспечения безопасности ответов использовались NeMo Guardrails [14] – библиотека с открытым исходным кодом от NVidia. NeMo Guardrails предназначена для поддержания целостности пользовательских запросов и не изменяет вводимые пользователем данные. Если запрос распознается как неправильно сформированный или неясный, он возвращается пользователю для уточнения, что обеспечивает точность процесса оказания медицинской помощи.

Рис. 2. Процесс обработки пользовательских запросов

3.1. Производительность комплексной модели

В табл. 2 приведена оценка прироста производительности комплексной модели после Fine-tuning. Цифры являются приблизительными, так как возможное распространение ошибок при последовательной интеграции моделей может повлиять на реальную производительность.

Таблица 2. Прирост производительности комплексной системы Gen model

Языки	**Комплексная модель Gen model**	
	До Fine-tuning	**После Fine-tuning**
Телугу	49.2	67.2
Хинди	49.2	67.1
Суахили	28.1	41.0
Арабский	44.5	65.6

3.2. Валидация комплексной модели

Обеспечение точности и уменьшение количества дезинформации очень важны для применения языковых моделей, чтобы избежать фабрикации сценариев. Для этой оценки использовался набор

данных MedHALT [15], включающий тесты на галлюцинации рассуждений (RHT - Reasoning Hallucination Tests) и галлюцинации памяти (MHT - Memory Hallucination Tests). Наборы данных были переведены на языки хинди, телугу, арабский и суахили с помощью API Azure Cognitive Translation. Некоторые переведенные наборы данных были доработаны и контекстуально обогащены, чтобы сохранить их актуальность и применимость на всех целевых языках.

Цель тестов – оценить способность языковой модели проводить логические рассуждения с медицинскими данными и выдавать результаты, которые являются не только последовательными, но и фактологически корректными. Кроме того, они гарантируют, что модель не изобретает информацию. Тесты охватывают три различных аспекта.

- Тестирование на ложную уверенность (FCT - False Confidence Testing). Этот тест измеряет склонность модели давать ответы с излишней уверенностью, особенно когда ей не хватает информации.
- Тест «Ничего из вышеперечисленного» (NOTA - None of the Above). Этот тест оценивает способность модели распознавать и игнорировать нерелевантную или неверную информацию.
- Тест на фальшивые вопросы (FQT - Fake Questions Test). В этом тесте модели задаются фиктивные или абсурдные медицинские вопросы, чтобы оценить ее способность точно распознавать и обрабатывать такие данные.

Результаты представлены в табл. 3. Комплексная модель обошла GPT-3.5, показав более высокую эффективность. Однако Llama-2 70B оказалась лучше в тестах на ложную уверенность и фальшивые вопросы, на это повлияли неточности перевода и расхождения в наборах данных.

Таблица 3. Результаты оценки LLMs и Gen model в тестах на галлюцинации

Модели	**FCT**		**FQT**		**NOTA**	
	Accuracy	**Score**	**Accuracy**	**Score**	**Accuracy**	**Score**
GPT-3.5	34.2	33.4	71.6	12.0	27.6	18.0
Llama-2 70B	42.2	52.4	97.3	17.9	77.5	188.7
Gen model – телугу	35.3	34.3	86.3	15.9	78.7	191.1
Gen model – хинди	38.2	43.1	89.8	18.3	81.9	193.8
Gen model – арабский	34.4	33.5	88.2	17.3	78.4	189.9

3.3. Сценарии применения комплексной модели

Использование интерактивных виртуальных симуляторов на основе LLM

Виртуальные симуляторы медицинских сценариев представляют собой мощный инструмент для обучения и повышения квалификации медицинских работников. Использование LLM для создания реалистичных диалогов между виртуальным пациентом и медицинским работником позволяет добиться высокого уровня интерактивности и погружения.

Ключевыми преимуществами применения виртуальных симуляторов на основе LLM являются:

- безопасная среда для отработки практических навыков диагностики и оказания помощи без риска для пациентов;
- возможность моделирования широкого спектра медицинских сценариев, включая редкие и сложные случаи;
- сбор данных о действиях и решениях медицинских работников для анализа и обратной связи.

Для реализации виртуальных симуляторов помимо описанной комплексной модели могут быть использованы следующие технологии и модели:

- модели компьютерного зрения, например DALL-E 2 или Stable Diffusion, для генерации визуальных компонентов симуляторов;
- интеграция с игровыми движками, такими как Unity или Unreal Engine, для разработки интерактивных 3D-симуляций;
- использование технологий отслеживания взгляда и жестов для обеспечения естественного взаимодействия пользователя.

Использование LLM для автоматизации рутинных задач

Одним из ключевых преимуществ применения LLM в здравоохранении является их способность автоматизировать широкий спектр рутинных административных и документационных задач. Это позволяет медицинским работникам высвободить время и ресурсы для непосредственного оказания медицинской помощи пациентам.

Задачи, которые может решать языковая модель:

- составление медицинской документации и отчетности;
- ведение электронных медицинских карт;
- поиск и обобщение медицинской информации;
- подготовка образовательных материалов для медицинских работников.

4. Перспективы развития языковых моделей в здравоохранении

Для обеспечения долгосрочной эффективности и актуальности применения LLM в здравоохранении необходимы механизмы постоянного мониторинга и адаптации к меняющимся потребностям медицинских работников на местах.

Ключевые элементы таких механизмов обратной связи:

- сбор и анализ отзывов медицинских работников о качестве, полезности и применимости рекомендаций, предоставляемых комплексной системой;
- интеграция каналов для оперативного сообщения о проблемах, ошибках или недостатках в работе LLM-решений;
- регулярное обновление обучающих данных и Fine-tuning моделей на основе полученной обратной связи.

Для реализации механизмов обратной связи могут быть использованы и интегрированы в систему следующие технологии:

- модели анализа тональности и эмоций, как RoBERTa [16] или XLNET [17], для оценки отзывов и настроений медицинских работников;
- системы управления жалобами и обращениями, интегрированные с LLM-моделями для быстрого реагирования и анализа проблем;
- технологии машинного обучения для кластеризации и выявления закономерностей в отзывах с целью определения приоритетных направлений улучшения.

Ключевым аспектом реализации является обеспечение безопасности, надежности и интегрируемости данных решений. Для этого могут использоваться:

- методы проверки на безопасность и отсутствие вредоносного контента в генерируемых LLM-моделями решений;
- технологии Federated Learning (федеративное обучение) [18] и Differential Privacy [19] для защиты конфиденциальности данных при обучении моделей.
- стандарты и протоколы обмена данными (FHIR - Fast Healthcare Interoperability Resources [https://www.hl7.org/fhir/overview.html], HL7 – Health Level 7 [https://docs.cntd.ru/document/1200135009]) для интеграции с медицинскими информационными системами.

Комплексное применение этих технологий и моделей позволит реализовать описанные инновационные решения и обеспечить их эффективное и безопасное использование в условиях ограниченных ресурсов здравоохранения.

5. Заключение

В работе рассмотрены основные этапы разработки и внедрения LLM для решения проблемы ожидаемого глобального дефицита медицинских работников, особенно в странах с низким и средним уровнем дохода населения. Анализ, проведенный на основе исследования [2], показал, что в результате объединения модели изучения языка (LLM) со сложными технологиями машинного перевода получается эффективная многоязычная языковая модель, предназначенная для поддержки работников здравоохранения общинных медицинских учреждений. Этот подход позволяет преодолевать культурные и языковые барьеры, значительно повышая доступность и качество медицинских услуг за счет предоставления медицинским работникам контекстуально релевантных медицинских рекомендаций и диагностических инструментов.

Основополагающей особенностью подхода является ее модульная конструкция, которая позволяет быстро адаптироваться к различным культурным и языковым средам, а также значительно сократить эксплуатационные расходы за счет использования компонентов с открытым исходным кодом. Композитность решения играет ключевую роль в его эффективном и масштабируемом развертывании во многих странах, а также в его уникальной способности способствовать независимому улучшению характеристик местных языков. Этот фактор крайне важен для того, чтобы медицинская помощь,

предлагаемая комплексной моделью, была точной и культурно релевантной, тем самым сокращая неравенство в обеспечении здравоохранения населения.

Кроме этого, в работе были предложены различные сценарии применения комплексной модели, включая виртуальные симуляторы и систему автоматизации рутинных задач, были описаны их преимущества и варианты доработки. А также был проведен обзор перспектив развития языковых моделей в области здравоохранения.

Настоящая работа подчеркивает потенциал ИИ в восполнении нехватки медицинских кадров в странах с низким уровнем дохода и в улучшении глобальных показателей здоровья. Проиллюстрирована универсальная и масштабируемая модель ИИ для решения проблем здравоохранения с учетом специфики региональных языков.

Литература

1. *Kwaku Agyeman-Manu, Tedros Adhanom Ghebreyesus, Mohamed Maait, Alexandru Rafila, Lino Tom, Nísia Trindade Lima, et al.* Prioritising the health and care workforce shortage: Protect, invest, together. // The Lancet Global Health – 2023. – Vol. 11(5): e686–e687.
2. *Agasthya Gangavarapu*. Introducing L2M3, A Multilingual Medical Large Language Model to Advance Health Equity in Low-Resource Regions, 2024. URL: https://arxiv.org/html/2404.08705v1#bib.bib8.
3. World Health Organization.Indicator metadata registry details. https://www.who.int/data/gho/indicator-metadata-registry/imr-details/158 , 2023.Accessed: 2024-04-03.
4. OpenAI. Gpt-4 system card. Technical report, OpenAI, San Francisco, CA, 2023.URL: https://cdn.openai.com/papers/gpt-4-system-card.pdf.
5. *Hugo Touvron, Louis Martin, Kevin Stone and et al.* Llama 2: Open foundation and fine-tuned chat models, 2023. URL: https://arxiv.org/abs/2307.09288.
6. *Chin-Yew Lin.* ROUGE: A Package for Automatic Evaluation of Summaries. // Association for Computational Linguistics. – Barcelona, 2004. – Vol. In Text Summarization Branches Out. – P.74–81. URL: https://aclanthology.org/W04-1013
7. *Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. // Proceedings of NAACL-HLT 2019, pp 4171 -4186.
8. *Cheonsu Jeong*. Fine-tuning and Utilization Methods of Domain-specific LLMs, 2024. URL: https://arxiv.org/abs/2401.02981.
9. *Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, Richard Dufour*. BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains, 2024. URL: https://arxiv.org/abs/2402.10373.
10. *Zeming Chen, Alejandro Hernández-Cano, Angelika Romanou, and et al.* Meditron-70b: Scaling medical pretraining for large language models, 2023. URL: https://arxiv.org/abs/2311.16079.
11. *Emilia David.* Meta's new ai model translates speech directly into other languages. https://www.theverge.com/2023/8/22/23840571/meta-multilingual-speech-translation-model-ai, August 2023. Accessed: 2024-04-03.
12. Anthropic. Model card for claude 3. Technical report, Anthropic, San Francisco, CA, 2023.URL: https://www-dn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.
13. *Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu.* BLEU: a Method for Automatic Evaluation of Machine Translation. // Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). – Philadelphia, July 2002. – P.311-318.
14. *Traian Rebedea, Razvan Dinu, Makesh Narsimhan Sreedhar, Christopher Parisien, and Jonathan Cohen.* NeMo guardrails: A toolkit for controllable and safe LLM applications with programmable rails. Proc. of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. – Singapore, 2023. – P. 431–445.
15. *Logesh Kumar Umapathi, Ankit Pal, and Malaikannan Sankarasubbu.* Med-halt: Medical domain hallucination test for large language models, 2023. URL: https://arxiv.org/abs/2307.15343.
16. *Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov*. RoBERTa: A Robustly Optimized BERT Pretraining Approach. 2019. URL: https://arxiv.org/abs/1907.11692.
17. *Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, Quoc V. Le.* XLNet: Generalized Autoregressive Pretraining for Language Understanding. 2020. URL: https://arxiv.org/abs/1906.08237.
18. *Subrato Bharati, M. Rubaiyat Hossain Mondal, Prajoy Podder, V. B. Surya Prasath*. Federated learning: Applications, challenges and future directions. 2022. URL: https://arxiv.org/abs/2205.09513.
19. *Martín Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, Li Zhang.* Deep Learning with Differential Privacy. 2016. URL: https://arxiv.org/abs/1607.00133.