Содержание

Анализ AI-возможностей

В последние годы наблюдается рост интереса к большим лингвистическим моделям и способам их применения в различных областях человеческого знания. В частности, данный анализ фокусируется на использовании больших лингвистических моделей в области медицины. На графике изображена динамика роста числа работ по данной теме.

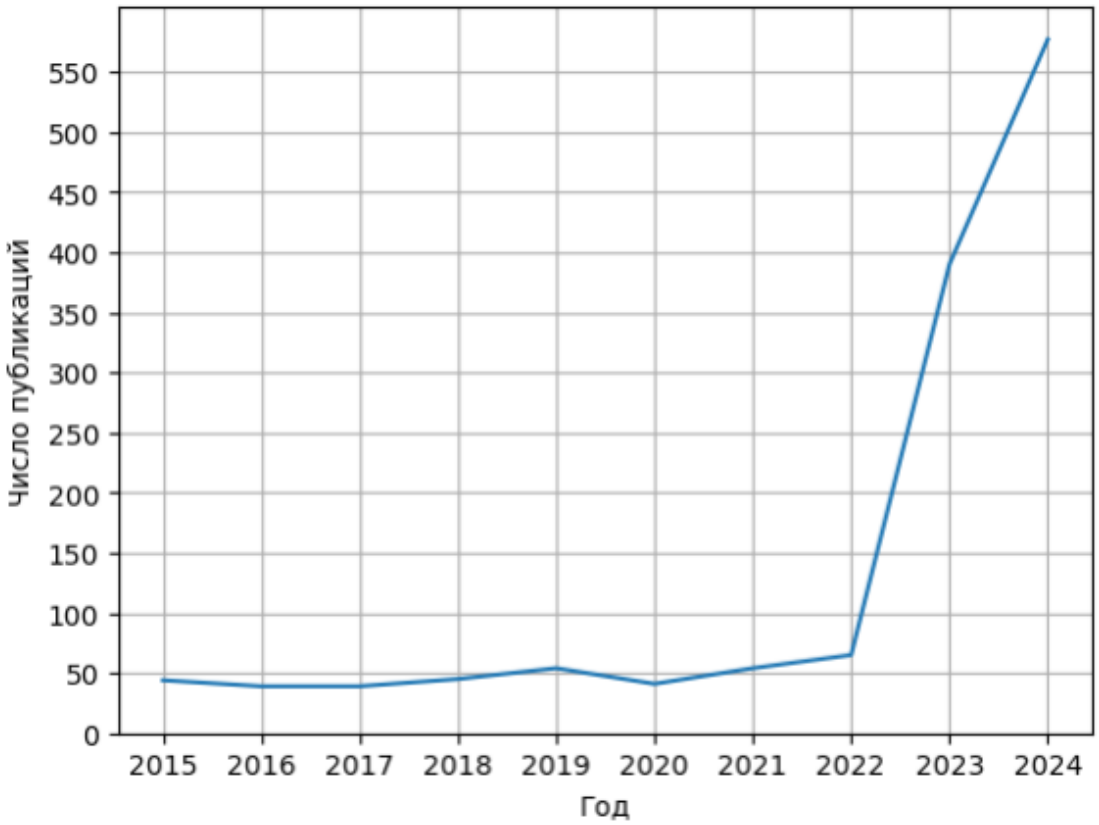

Рисунок 1 – Рост числа публикаций по теме «Большие лингвистические модели в медицине». Поиск производился с помощью агрегатора Semantic Scholar

Значительное число работ направлено на создание специализированных медицинских больших лингвистических моделей [1–6]. Медицинские большие лингвистические модели могут стать одним из наиболее значительных достижений в области медицины за последние годы, предлагая такие возможности, как быстрый анализ большого объёма медицинской литературы, персонализированное общение с пациентами и оптимизация административных процессов. Такие модели могут улучшить процесс принятия решений, повысить точность диагностики и расширить доступ к медицинским консультациям, особен-

но в условиях ограниченных ресурсов. Например, модели на основе PaLM и GPT-4, MedPaLM-2 и MedPrompt, соответственно, достигли конкурентной точности 86,5 и 90,2 по сравнению с человеческими экспертами (87,0) на экзамене на получение медицинской лицензии США (USMLE) [7].

Однако внедрение больших лингвистических моделей сопряжено со значительными трудностями. LLM подвержены неточностям, на данный момент не имеют клинической валидации и напрямую зависят от обучающих данных. Более того, большие лингвистические модели потенциально уязвимы к special token injection атакам, что может привести к утечки данных о пациентах, а их ответы часто не имеют тонкого понимания и ответственности, присущих человеческим суждениям. Несмотря на то, что медицинские LLM обладают огромным потенциалом, к их интеграции следует подходить с осторожностью, обеспечивая надежные гарантии безопасности пациентов, сохранности данных и соблюдения этических норм .

Среди основных опасностей внедрения больших лингвистических моделей авторы выделяют:

1) Ошибочные ответы. LLM могут давать убедительные, но предвзятые или неверные ответы, потенциально способные повлиять на решение пациентов без адекватного объяснения рисков и преимуществ.

2) Несбалансированные датасеты. LLM, обученные на несбалансированных наборах данных, могут давать неверные результаты, отвечая на вопросы, связанные со слабо представленными в датасете темами.

3) Отсутствие четкой ответственности. В случаях, когда ответы, сгенерированные LLM, причиняют вред, неясно, кто должен нести ответственность – разработчики, медицинские работники или учреждения. Это может привести к моральному риску, когда поставщики услуг чрезмерно полагаются на систему, пренебрегая своим профессиональным суждением.

4) Риск нарушения конфиденциальности данных. Присутствие конфиденциальной информации в датасете может привести к тому, что данные о пациенте могут быть использованы без его согласия.

5) Чрезмерная уверенность в моделях. Пациенты или врачи могут принять результаты LLM без достаточной проверки, что приведет к пагубным последствиям.

Большие лингвистические модели произвели революцию в возможностях чат-ботов искусственного интеллекта и в области обработки естественного языка в целом, продемонстрировав производительность на уровне человека в профессиональных тестах в областях медицины и радиологии. К недостаткам LLM, которые в настоящее время ограничивают их применение в радиологии, относятся галлюцинации, ограничение знаний, плохое комплексное мышление, тенденция к сохранению предвзятости. LLM имеют огромный потенциал для применения в радиологии. Использование больших лингвистических моделей позволит повысить эффективность, рентабельность и качество лечения.

Таким образом, можно сказать, что большие лингвистические модели, в особенности мультимодальные, могут быть внедрены в клиническую практику в области радиологии, но их необходимо оптимизировать и протестировать, прежде чем применять в контролируемых условиях.

Исследование решений

В литературе к настоящему времени имеется информация об экспериментах по разработке мультимодальных больших лингвистических моделей с поддержкой возможности обработки визуальной информации, в том числе и в области медицины [8—11].

В статье [8] авторы дообучили LLaVA-Med – мультимодальную большую лингвистическую модель, созданную для помощи в медицинских целях. В качестве результатов проделанной работы можно выделить:

1) Конвейер для создания аннотаций к датасету PMC-15M. Написан скрипт для загрузки изображений датасета и автоматического создания аннотаций с помощью GPT-4.

2) Разделение дообучения на 2 этапа: Concept Alignment – дообучение на большом объёме (600 тысяч) изображений с аннотациями для внедрения медицинских терминов и адаптации модели к новой задаче и Instruction-Tuning – дообучение на маленьком объёме (60 тысяч) датасета для приведения ответов модели к определенному формату.

3) Дообученная модель LLaVA1.5-7B для работы с медицинскими изображениями.

4) Проведена оценка работы модели на бенчмарках VQA-RAD, SLAKE, PathVQA.

Авторы статьи также приводят возможные направления исследований в будущем:

1) Повышение точности работы модели. Авторы отмечают, что обученная модель, как и большинство схожих моделей, подвержена ошибкам и имеет проблемы с решением задач, требующих логики.

2) Более глубокое тестирование. Авторы предлагают тестировать модели на большем количестве бенчмарков и метрик. Кроме того, в процессе тестирования модель должна быть оценена на проблемах из клинической практики.

3) Поддержка дополнительных языков. Несмотря на то, что в статье приводится пример использования обученной модели для ответа на вопросы на других языках, авторы отмечают, что никаких усилий для поддержки дополнительных языков приложено не было.

4) Разработка инструментов для инференса. Авторы отмечают, что в будущих исследованиях необходимо разработать удобные системы для взаимодействия рядового пользователя с моделью, например, веб-приложение или чат-бот.

5) Использовать другой encoder. Авторы предлагают использовать BioMedCLIP для кодирования изображений, так как он позволяет добиться более высокой точности в данной области.

В статье [9] авторы делятся опытом и результатами дообучения большой лингвистической мультимодальной нейросети Med-Flamingo-9B, основанной на

Open-Flamingo-9B. Архитектурные отличия от LlaVA-Med позволяют данной модели обрабатывать сразу несколько изображений в одном запросе. Авторами были достигнуты следующие результаты:

1) Была адаптирована и дообучена большая лингвистическая мультимодальная модель Open-Flamingo-9B.

2) Дообучение производилось на двух датасетах: MTB (Medical Textbooks Dataset), состоящий из 4271-го медицинского учебника, содержащих 0.8 миллионов изображений, и PMC-OA Dataset, состоящий из 1.6 миллионов изображений с подписями. Авторами была проведена серьезная работа по выбору качественных источников и удалению дубликатов.

3) Обработка нескольких изображений в одном запросе. Авторы также акцентируют внимание на важности обработки сразу нескольких изображений, поскольку одно изображение часто содержит недостаточное количество информации для правильного решения многих задач.

4) Авторы отмечают низкое качество существующих метрик и бенчмарков для оценки мультимодальных моделей в области медицины. Для решения этой проблемы был создан новый датасет, имитирующий экзамен для получения медицинской лицензии в США.

5) Впервые оценка правильности работы модели производилась, в том числе, с участием экспертов из области медицины. Для обеспечения удобного взаимодействия экспертов с моделью авторами статьи было разработано специальное приложение.

Недостатки и возможные пути развития:

1) Низкая точность. Авторы акцентируют внимание на том, что обученная модель не подходит для клинического использования в связи с низкой точностью и большим количеством ошибок. Возможным решением может стать увеличение размера модели или дополнительное обучение.

2) Поддержка других форматов данных. Авторы предлагают рассмотреть возможность поддержки данных других форматов, например, видеоматериалов или 3D-файлов.

В статье [10] авторы рассказывают о MiniGPT-Med, мультимодальной большой лингвистической модели, созданной для обработки медицинских изображений и текстовой информации. Обучение модели производилось на датасетах радиологических изображений, включающих в себя рентгеновские изображения, КТ-сканы и снимки МРТ. Авторы описывают следующие результаты работы:

1) Дообучена большая языковая модель LLaMA-2-chat с энкодером изображений EVA.

2) Дообучение производилось на датасетах MIMIC-CXR, SLAKE и RadQVA (124 тыс. изображений). Особое внимание уделялось таким задачам, как создание медицинских отчетов, выявление заболеваний и ответы на вопросы об изображениях.

3) Представлены идентификаторы заданий, а также использованы текстовые представления разных областей изображения для более эффективного извлечения информации из изображения.

4) Произведена оценка модели на метриках BERT-Sim и CheXbert-Sim. Авторам удалось добиться прироста в точности ответов на 19% и 5.2% соответственно относительно лучших моделей.

Среди наиболее существенных недостатков модели авторы отмечают галлюцинации, зависимость от базовой модели и отсутствие тестирования модели на задачах из реального мира.

В качестве возможных направлений исследования авторы выделяют:

1) Расширение датасета. Общедоступные датасеты могут не полностью отражать разнообразие реальных медицинских случаев. Использование более богатых наборов данных, включая данные о пациентах с различными демографическими характеристиками и редкими заболеваниями, повысит надежность модели.

2) Поддержка других форматов данных.

3) Тестирование модели на примерах из реальной медицинской практики.

4) Повышение точности. Прежде чем интегрировать медицинские модели в клиническую практику необходимо минимизировать галлюцинации.

В статье [11] представлена мультимодальная лингвистическая модель для обработки радиологических данных, названная RadFM. Авторы выделяют следующие результаты работы:

1) Собран датасет MedMD, состоящий из 16 миллионов изображений и 3D-сканов с подробными описаниями. В частности, часть датасета для обучения модели на радиологических снимках (RadMD) содержит около 3 миллионов изображений.

2) На собранном датасете дообучена модель LLaMA-13B. В качестве энкодера изображений и пространственной информации использовался 12-слойный 3D ViT.

3) Представлен новый бенчмарк – RadBench, предназначенный для оценки производительности RadFM в пяти задачах: распознавание модальностей, диагностика заболеваний, ответы на вопросы медицинской тематики, создание отчетов и обоснование диагноза.

Среди главных недостатков авторы отмечают присутствие в датасете ошибочных данных, отсутствие клинических испытаний модели, а также несоответствие бенчмарков, на которых производился анализ модели реальным задачам.

Авторы предлагают следующие решения для обозначенных проблем:

1) Проведение клинических испытаний. Проведение более обширных испытаний в реальных клинических условиях может способствовать дальнейшему подтверждению надежности и эффективности модели.

2) Использование 4D-данных. Добавление динамических 4D-данных (сканирование с временной последовательностью) может улучшить способность модели обнаруживать изменения с течением времени, что очень важно для верной постановки некоторых диагнозов.

3) Аугментация данных. Для решения проблемы дисбаланса между редкими и распространенными заболеваниями можно применить более продвинутые методы аугментации данных.

4) Обучение с учителем. Сотрудничество с экспертами в области радиологии в процессе обучения может улучшить прогнозы модели и повысить ее клиническую значимость.

В таблице 1 приведены данные об использованных авторами статей метрик и полученных результатах.

	RadVQA (BERT-Sim)	MIMIC-CXR (BERT-Sim)
MedFlamingo	0,48	0,1
LLaVA-Med	0,61	0,06
RadFM	0,62	0,45
MiniGPT-Med	0,58	0,72

Потенциальные метрики

Perplexity, PPL — классическая метрика качества языковых моделей, отражающая их способность предсказывать последовательность слов. Она определяется как экспонента от средней отрицательной логарифмической вероятности слов в тексте. Чем ниже значение метрики, тем модель лучше предсказывает текст [12].

BERTScore использует современные языковые модели (например, BERT) для вычисления семантической близости между сгенерированным текстом и эталоном. Для этого берутся эмбеддинги слов и вычисляется косинусное

сходство. Итоговое значение формируется на основе F1-метрики. BERTScore лучше коррелирует с человеческими оценками, чем классические метрики на n-граммах [13].

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) — метрика, чаще всего используемая для оценки качества автоматического суммирования текста. Она ориентирована на полноту (recall), измеряя долю n-грамм или подпоследовательностей из эталонного текста, которые присутствуют в сгенерированном.

Техническая осуществимость

Техническая осуществимость проекта определяется доступностью технологий, ресурсов и инфраструктуры, необходимых для реализации предложенных решений.

Во-первых, в настоящее время существует широкий набор инструментов и библиотек для работы с искусственным интеллектом: PyTorch, TensorFlow, HuggingFace Transformers обеспечивают доступ к современным моделям обработки естественного языка и позволяют проводить эксперименты без необходимости создания модели «с нуля».

Во-вторых, для проведения исследовательской части достаточно персонального компьютера. Для обучения модели будут использоваться ресурсы кластера ВолГТУ. Это снижает порог вхождения и делает проект реализуемым даже при ограниченных локальных ресурсах.

В-третьих, интеграция модели в прикладные системы возможна с использованием стандартных технологий: REST API, фреймворков для веб-разработки или мобильных сервисов и библиотек для работы с данными. Это позволяет создать прототип приложения с базовой функциональностью, не требуя уникальной инфраструктуры.

Основными потенциальными ограничениями являются:

1) необходимость значительных вычислительных мощностей при работе с крупными моделями;
2) требования к качеству и объёму данных для обучения и валидации;
3) вопросы безопасности и фильтрации нежелательного контента.

В целом, проект можно признать технически осуществимым, так как современные технологии и вычислительные ресурсы обеспечивают возможность реализации как исследовательской, так и практической части работы.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Capabilities of Gemini Models in Medicine / K. Saab [и др.] // ArXiv. — 2024. — T. abs/2404.18416. — URL: https://api.semanticscholar.org/CorpusID:269449780.
2. Large language models encode clinical knowledge / K. Singhal [и др.] // Nature. — 2022. — T. 620. — C. 172—180. — URL: https://api.semanticscholar.org/CorpusID:255124952.
3. CancerLLM: A Large Language Model in Cancer Domain / M. Li [и др.] // ArXiv. — 2024. — T. abs/2406.10459. — URL: https://api.semanticscholar.org/CorpusID:270559865.
4. MEDITRON-70B: Scaling Medical Pretraining for Large Language Models / Z. Chen [и др.] // ArXiv. — 2023. — T. abs/2311.16079. — URL: https://api.semanticscholar.org/CorpusID:265456229.
5. BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains / Y. Labrak [и др.] // Annual Meeting of the Association for Computational Linguistics. — 2024. — URL: https://api.semanticscholar.org/CorpusID:267740180.
6. Clinical Camel: An Open Expert-Level Medical Language Model with Dialogue-Based Knowledge Encoding / A. Toma [и др.] //. — 2023. — URL: https://api.semanticscholar.org/CorpusID:261030221.
7. What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams / D. Jin [и др.] // ArXiv. — 2020. — T. abs/2009.13081. — URL: https://api.semanticscholar.org/CorpusID:221970190.
8. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day / C. Li [и др.] // ArXiv. — 2023. — T. abs/2306.00890. — URL: https://api.semanticscholar.org/CorpusID:258999820.

9. Med-Flamingo: a Multimodal Medical Few-shot Learner / M. Moor [и др.] // ArXiv. — 2023. — Т. abs/2307.15189. — URL: https ://api.semanticscholar.org/CorpusID:260316059.
10. MiniGPT-Med: Large Language Model as a General Interface for Radiology Diagnosis / A. Alkhaldi [и др.] // ArXiv. — 2024. — Т. abs/2407.04106. — URL: https://api.semanticscholar.org/CorpusID:271039722.
11. Towards Generalist Foundation Model for Radiology / C. Wu [и др.] // ArXiv. — 2023. — Т. abs/2308.02463. — URL: https : //api.semanticscholar.org/CorpusID:260611504.
12. Jurafsky D., Martin J. H. Speech and language processing. 3rd edn. draft //Online: https://web. stanford. edu/jurafsky/slp3. – 2020.
13. Zhang T. et al. Bertscore: Evaluating text generation with bert //arXiv preprint arXiv:1904.09675. – 2019.