УДК 004.832.2

# НУЖЕН ЛИ МЕДИКАМ GPT-4: АНАЛИЗ АКТУАЛЬНОГО МИРОВОГО ОПЫТА

*Адылова Ф.Т.*[1], *Давронов Р.Р.*[1]

[1] Институт математики им. В.И. Романовского АН РУз, Ташкент, Узбекистан

fatadilova@mathinst.uz , rifqat.davronov@mathinst.uz

**Аннотация**. *Большие языковые модели (LLM) продемонстрировали замечательные возможности в понимании и генерации естественного языка в различных областях, включая медицину. В статье представлена оценка GPT-4 на основе двух точек зрения на проблему применения этой языковой модели: разработчиков из OpenAI, Microsoft и пользователей-медиков из двух европейских проектов. За последние несколько лет LLM, обученные на массивных междисциплинарных корпусах, стали мощными строительными блоками при создании систем, ориентированных на решение конкретных задач. В статье рассматривается три задачи: медицинское образование, работоспособность ChatGPT-4 в клинике (консультации, записи стенограмм беседы врача и пациента), и конкретные уровни точности диагностики (разные области медицины). Ответ на поставленный вопрос о необходимости медицинского GPT есть в мире, -он положительный.*

**Ключевые слова**: *LLM, OpenAI, GPT-4, ChatGPT-4.*

## I. ВВЕДЕНИЕ

Сочетание нехватки врачей и возросшей сложности медицины уже представляет собой серьезную проблему для своевременной, точной постановки диагнозов и лечения. Учитывая демографические изменения и старение населения, ожидается, что в ближайшие годы эта проблема с рабочей нагрузкой врачей еще больше возрастет, что подчеркивает необходимость новых технологических разработок. Непрерывное развитие искусственного интеллекта (ИИ), включая большую языковую модель (LLM), известную как генеративный предварительно обученный трансформер (GPT), позволило проводить новые исследования в разных областях медицины. Например, стали доступны приложения для записи с голоса медицинских ежедневных обходов, синтез интервью с пациентами и результатами лабораторных анализов позволяет делать эти записи напрямую, без вмешательства врача. ИИ используется для интерпретации изображений - рентгенограмм, гистологии и исследования глазного дна, для анализа и интерпретации больших исследовательских баз данных, содержащих информацию, варьирующуюся от лабораторных результатов до клинических данных.

Все эти инструменты обладают потенциалом повышения эффективности и, возможно, могут дать представление о том, чего трудно достичь с помощью традиционных методов анализа данных. Однако новые методы искусственного интеллекта не являются панацеей; они могут быть хрупкими, могут работать только в узкой области и иметь встроенные предубеждения («галлюцинации»). Базовая технология ИИ быстро меняется, и во многих случаях разрабатывается компаниями и исследователями, имеющими финансовые интересы в своих продуктах. Для растущего класса крупномасштабных моделей ИИ, кампании, обладающие необходимыми ресурсами, могут быть единственными, кто способен расширить границы систем искусственного интеллекта. Поэтому многие такие модели еще не получили широкого распространения, из-за чего практический опыт и детальное понимание рабочих характеристик модели часто доступны лишь небольшой группе разработчиков моделей.

Несмотря на потенциальные финансовые стимулы, которые могут привести к конфликту интересов, глубокое понимание искусственного интеллекта и машинного обучения и их использования в медицине требует участия людей, вовлеченных в их разработку, с одной стороны, и пользователей (медиков, пациентов), - с другой [1]. Именно в этом и состоит мотивация данной статьи, - мы сопоставим на основе многих актуальных публикаций точки зрения и тех, и других.

В 1990-х и начале 2000-х годов, даже при медленных компьютерах и ограниченной памяти, решалась проблема успешного выполнения машинами определенных медицинских задач. Благодаря значительным денежным вложениям и интеллектуальным усилиям компьютерное считывание электрокардиограмм (ЭКГ) и дифференциального количества лейкоцитов,

анализ фотографий сетчатки и кожных поражений, а также другие задачи обработки изображений стали реальностью. Многие из этих задач, решаемых с помощью машинного обучения, были в значительной степени приняты и внедрены в повседневную медицинскую практику [2].

## II. МЕТОДОЛОГИЯ ПРИМЕНЕНИЯ GPT-4: ТОЧКА ЗРЕНИЯ ИНЖЕНЕРОВ-РАЗРАБОТЧИКОВ

***GPT-4 в медицинском образовании***. Прорыв в технологии ИИ реализовали большие языковые модели (LLMs) продемонстрировав замечательную способность интерпретировать и генерировать последовательности в областях, таких как естественный язык, компьютерный код и последовательности белков. Многочисленные мощные модели основаны на архитектуре трансформер [3], адаптированы к языку и обучаются самоконтролем [4,5]. Оценки по множеству тестов, как правило, улучшались с увеличением масштаба, включая увеличение размера модели, размера набора данных и объем обучающих вычислений [6,7]. Эмпирические результаты перекликаются с теоретическим анализом, который показывает необходимость масштаба для надежности выводов из больших нейронных моделей [8].

За последние несколько лет LLM, обученные на массивных междисциплинарных корпусах, стали мощными строительными блоками при создании систем, ориентированных на решение конкретных задач [9]. Методы доработки моделей для конкретной предметной области включают тонкую настройку с использованием специализированных наборов данных, взятых из целевых приложений, и общих методы управления поведением моделей, таких как обучение с подкреплением (обратная связь с человеком), которое направляет систему к лучшему пониманию запросов конечных пользователей [10].

Исследования вычислительных методов для оказания помощи врачам включали вероятностные методы и методы теории принятия решений, продукционные экспертные системы, семантические графы, контролируемое изучение баз медицинской информации и модели глубоких нейронных сетей [11].

Такие модели могут быть обучены на конкретных медицинских корпусах или базовых моделях, предварительно обученных на огромных объемах общей языковой и/или визуальной информации, а затем адаптированы к медицинским данным посредством специальной тонкой настройки. Анализ возможностей GPT-4 в решении сложных медицинских задач с целью сравнения GPT-4 с GPT-3.5 и Flan-PaLM540B проведен разработчиками из Microsoft, Open AI [12]. Это исследование состоит из всесторонней оценки эффективности GPT-4 на 1-3 этапах экзамена по медицинскому лицензированию в США (USMLE) https://synapse-med.ru/blog/chto-takoe-usmle-ili-kak-stat-vrachom-v-ssha.

Чтобы оценить GPT-4, были рассмотрены шесть наборов данных, которые охватывают различные аспекты медицинских знаний и рассуждений. Два из этих наборов данных, - выборочный экзамен USMLE и самооценка USMLE, были получены из Национального совета медицинских экспертов (NBME), остальные четыре набора составляют значительную часть недавно введенного бенчмарка “MultiMedQA” [13].

Результаты показали, что GPT-4 демонстрирует значительное прогресс по сравнению со своими предшественниками в вопросах официального экзамена USMLE, улучшившись более чем на 30 процентных пунктов на обоих экзаменах по сравнению с GPT-3.5. GPT-4 показывает столь же резкое улучшение по сравнению с недавними независимыми показателями производительности ChatGPT, -популярного варианта GPT-3.5, оптимизированного для взаимодействия в чате [14]. В таблице 1 и 2 приведены соответствующие оценки.

**Таблица 1.** Сравнение эффективности моделей по самооценке USMLE. GPT-4 значительно превосходит GPT-3.5

| USMLE | GPT-4 (5 ответов) | GPT-4 (нет ответов) | GPT-3.5 (5 ответов) | GPT-3.5 (нет ответов) |
|---|---|---|---|---|
| **Этап 1** | **85.21** | 83.46 | 54.22 | 49.62 |
| **Этап 2** | **89.50** | 84.75 | 52.75 | 48.12 |
| **Этап 3** | **83.52** | 81.25 | 53.41 | 50.00 |
| **Среднее*** | **86.65** | 83.76 | 53.61 | 49.10 |

**Таблица 2:** Сравнение производительности моделей на выборочном экзамене USMLE. Этот набор данных рассмотрен в [14].

| USMLE | GPT-4 (5 ответов) | GPT-4 (нет ответов) | GPT-3.5 (5 ответов) | GPT-3.5 (нет ответов) | ChatGPT (нет ответов) |
|---|---|---|---|---|---|
| **Этап 1** | **85.71** | 80.67 | 52.10 | 51.26 | 55.1 |
| **Этап 2** | **83.33** | 81.67 | 58.33 | 60.83 | 59.1 |
| **Этап 3** | **90.71** | 89.78 | 64.96 | 58.39 | 60.9 |
| **Среднее *** | **86.70** | 84.31 | 58.78 | 56.91 | - |

*Рассчитывается как соотношение правильных вопросов ко всем вопросам на всех трех этапах. Размер выборки на каждом этапе немного отличается

Авторы из [12] выполнили поверку GPT-4,-т.е. вычислили меру соответствия между прогнозируемыми вероятностями правильности каждого ответа и истинными результатами. Поверка вероятности правильности ответов или любых утверждений, сгенерированных LLM, имеет решающее значение для приложений в области медицины. Хорошая поверка - не то же самое, что высокая точность прогнозирования, поскольку прогнозирующие модели могут быть точными, но плохо проверенными [15].

Методом поверок измерений является график, который объединяет предсказания по их предполагаемым вероятностям и измеряет, насколько близка средняя вероятность в каждой ячейке к истинному показателю. На рисунке 1 представлены результаты поверки моделей GPT-4 и GPT-3.5 в обоих официальных наборах данных USMLE.

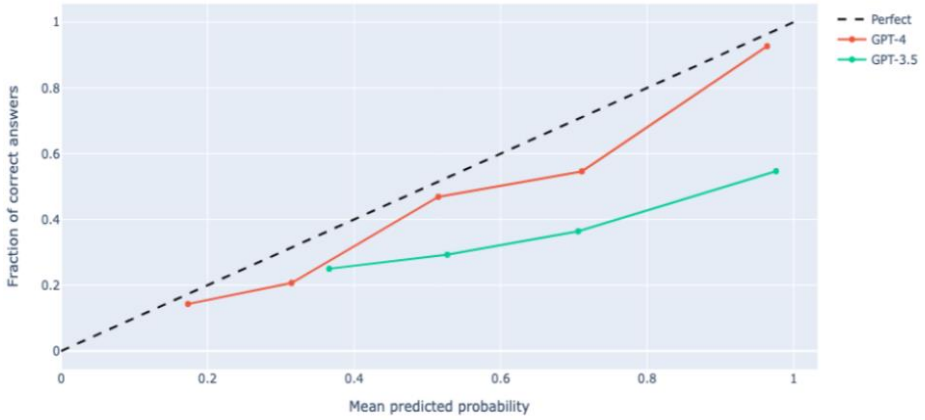


**Рис 1.** Сравнение поверок GPT-4 и GPT-3.5 по результатам экзамена USMLE.

***GPT-4 как Chatbot.*** Рассмотрим ещё один аспект GPT-4: преимущества, ограничения и риски GPT-4 в качестве чат-бота, представленный также разработчиками из Microsoft Research и MIT [16]. GPT-4 был разработан для развития общих когнитивных навыков с целью помочь пользователям выполнять множество различных задач. Приглашение к чату может быть в форме вопроса, но оно также может быть указанием на выполнение конкретной задачи, например: “Пожалуйста, прочтите и обобщите это медицинское исследование в статье.” Кроме того, подсказки не ограничиваются предложениями на английском языке; они могут быть написаны на множестве различных языков и могут содержать электронные таблицы, технические спецификации, исследовательские работы и математические уравнения. Microsoft Research совместно с OpenAI в течение последних 6 месяцев изучали возможное использование GPT-4 в здравоохранении и медицинских приложениях, чтобы лучше понять его фундаментальные возможности, ограничения и риски для здоровья человека. Речь идёт о применении в медицинской документации, взаимодействии данных, диагностике, исследованиях и образовании. Несколько других известных чат-ботов с искусственным интеллектом также были изучены для применения в медицине, - в том числе LaMDA (Google) [17].

Поскольку медицине учат на примерах, в статье приведены примеры GPT-4 Первый при-

мер демонстрирует способность GPT-4 составлять медицинские записи на основе стенограммы встречи врача и пациента [18]. В этом примере GPT-4 принимает взаимодействие голоса врача и пациента, а затем создает “медицинскую заметку” для медицинской карты пациента. GPT-4 может быть предложено ответить на вопросы о встрече, написать резюме после посещения и представить критическую информацию по обратной связи с врачом и пациентом. Хотя такое приложение, безусловно, полезно, но GPT-4 — это интеллектуальная система, которая, подобно человеку, подвержена ошибкам. Второй пример, -консультации врачей. Если задаются типичные вопросы о первоначальном осмотре пациента или краткое изложение результатов лабораторных анализов, GPT-4, как правило, даёт полезные ответы, которые могут помочь врачу решить вызывающую беспокойство проблему.

Медицинские знания делают GPT-4 потенциально полезным не только в клинических условиях, но и в научных исследованиях. GPT-4 может читать материалы медицинских исследований и участвовать в их обсуждении, например, кратко резюмируя содержание, или выявлять соответствующую предыдущую работу, оценивать выводы и задавать возможные вопросы для последующего исследования.

GPT-4 чрезвычайно мощный, у него также есть важные ограничения: система может допускать, но также и улавливать ошибки, допущенные как ИИ, так и людьми. Применения ИИ на моделях, настроенных для конкретных клинических задач, выиграли от точно определенной операционной среды. Но как оценивать общую интеллектуальность такого инструмента, как GPT-4? В какой степени пользователь может “доверять” GPT-4 или ему нужно тратить время на проверку правдивости того, что он пишет? Насколько больше требуется проверки фактов, чем корректуры, и в какой степени GPT-4 может помочь в выполнении этой задачи? GPT-4, как и недавно заявленная GPT-5 https://chat-gpt-5.ai/ не является самоцелью, — это открытие двери к новым возможностям, но и к новым рискам.

***Выводы.*** Не останавливаясь на интересных деталях этого исследованиях, процитируем главный вывод разработчиков GPT-4: исключительные характеристики GPT-4 по контрольным показателям служат достоверной оценкой его потенциала для использования в медицинском образовании и в поддержке многих аспектов оказания медицинской помощи.

Таким образом, мнение крупнейших разработчиков GPT: это очередной шаг в медицинском ИИ, но будущая работа потребует ответов на уже существующие и потенциальные критические замечания.

## III. МЕТОДОЛОГИЯ ПРИМЕНЕНИЯ GPT-4: ТОЧКА ЗРЕНИЯ КЛИНИЦИСТОВ

Точка зрения практикующих врачей нами проанализирована по двум публикациям врачей [19,20].

***Эффективность GPT-4 в диагностике.*** Врачи из Дании [19] оценили эффективность GPT-4 в диагностике сложных медицинских случаев и сравнили правильность его ответов с показателями читателей медицинских журналов. Мотивация работы: вопрос о том, насколько хорошо GPT-4 работает в реальных клинических случаях до сих пор изучен недостаточно. Например, остается неясным, в какой степени GPT-4 может помочь в клинических случаях, содержащих длинные, сложные и разнообразные описания пациентов, и как он работает в этих сложных реальных случаях по сравнению с людьми. В исследовании использовались доступные сложные задачи по клиническим случаям с исчерпывающей полнотекстовой информацией, опубликованной онлайн в период с января 2017 года по январь 2023 года. В каждом случае представлена история болезни и опрос с шестью вариантами наиболее вероятного диагноза. Для решения задач мы представили GPT-4 с подсказкой и клиническим случаем. Подсказка предписывала GPT-4 решить проблему, ответив на вопрос с множественным выбором, за которым следовал полный неотредактированный текст из отчета о клиническом случае. Лабораторная информация, содержащаяся в таблицах, была преобразована в обычный текст и включена в кейс. Использовались две версии GPT-4 от марта и сентября 2023 г. Для читателей медицинских журналов собрали количество и распределение голосов по каждому случаю. Используя эти наблюдения, смоделировали 10 000 наборов ответов на все случаи, в результате чего получилась псевдопопуляция из 10 000 обычных участников-людей. Ответы были смоделированы как независимые переменные с распределением по Бернулли (правильный/неправильный ответ) с предельными распределениями, наблюдаемыми среди читателей медицинских журналов.

Были выявлены 38 проблемных клинических случаев, и, в общей сложности, 248 614 ответов от читателей онлайн-медицинских

журналов. Наиболее распространенные диагнозы были в области инфекционных заболеваний (39,5%), в эндокринологии (13,1%) и в ревматологии (10,5%). Возраст пациентов, представленных в клинических наблюдениях, варьировался от новорожденного до 89 лет. В мартовском выпуске GPT-4 за 2023 год был поставлен правильный диагноз в среднем в 57%, тогда как читатели медицинских журналов в среднем правильно поставили диагноз в 36% случаев. Основываясь на моделировании, обнаружили, что GPT-4 показал лучшие результаты, чем 99,98% псевдопопуляции (рис. 2). Выпуск GPT-4 от сентября 2023 года правильно диагностировал 54% случаев.

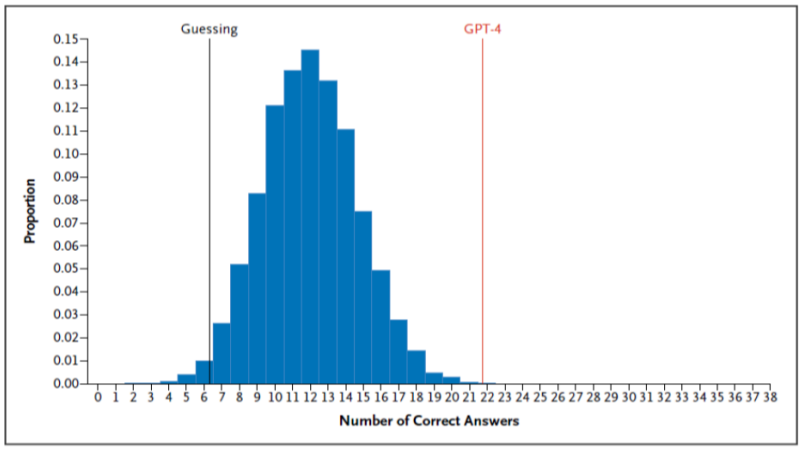


**Рис 2.** Количество правильных ответов в GPT-4 по сравнению с угадыванием и моделированием популяции читателей медицинских журналов. Количество правильных ответов GPT-4 (красная линия) на 38 реальных клинических задач с множественным выбором по сравнению с тем, что можно было бы ожидать при простом угадывании с одинаковой вероятностью для всех вариантов ответа (черная линия), и с долей правильных ответов в смоделированной популяции из 10 000 читателей медицинских журналов (синяя гистограмма).

***ChatGPT-4 в консультациях пациентов.*** Врачи из Голландии изучили эффективность ChatGPT как поддержку принятия решений с помощью ИИ, для чего провели экспериментальное исследование для интерпретации симптомов и лечения распространенных сердечно-сосудистых заболеваний, - проект AMSTELHEART-2 [20]. В исследовании протестировали способность ChatGPT правильно отвечать на простые вопросы о сердечно-сосудистых заболеваниях и интерпретировать симптомы или давать соответствующие рекомендации по лечению на основе кратких описаний случаев заболеваний, основанных на первичной медицинской помощи. Простые вопросы, а также краткие описания были введены онлайн на веб-платформе ChatGPT только на английском языке. Использовали быстрые пять тестов из Medscape по десять вопросов в каждом по темам: острый коронарный синдром, легочная и венозная тромботическая эмболия, фибрилляция предсердий, сердечная недостаточность и управление сердечно-сосудистыми рисками https://reference.medscape.com/index/section_10360_0.

Были выбраны 10 случаев консультаций пациента с врачом и 10 случаев консультаций врача общей практики с кардиологом/экспертом.

Результаты эксперимента показали, что ChatGPT правильно ответил на 74% простых вопросов, с небольшими различиями в точности в областях ишемической болезни сердца (80%), тромбоэмболии легких и вен (80%), фибрилляции предсердий (70%), сердечной недостаточности (80%) и управления сердечно-сосудистыми рисками (60%). В случае с краткими описаниями клинических случаев ответ ChatGPT в 90% случаев соответствовал фактически данному совету. В более сложных случаях, когда врачи (врачи общей практики) обращались к другим врачам (кардиологам) за помощью или поддержкой принятия решений, ChatGPT был правильным в 50% случаев и часто давал неполные или несоответствующие рекомендации по сравнению с консультацией эксперта. На рисунке 3 дана иллюстрация полученных результатов.

***Выводы.*** Таким образом, сравнение диагностической точности GPT-4 в сложных случаях с точностью читателей журнала, которые отвечали на те же вопросы в Интернете показало, что GPT-4 хорошо справлялся со сложными случаями и даже лучше, чем читатели медицинских журналов. В настоящее время GPT-4

специально не предназначен для медицинских задач, однако прогресс в разработке моделей ИИ продолжает ускоряться, что приведет к более быстрой диагностике и лучшим результатам. Результаты данного исследования, вместе с недавними находками других медиков, указывают на то, что текущая модель GPT-4 может быть перспективной в клинической практике уже сегодня.

Сильной стороной исследования эффективности ChatGPT является то, что использовали несколько подходов (простые вопросы и случаи различной сложности) для оценки точности ChatGPT в решении медицинских вопросов. Именно, использование кратких описаний случаев обеспечивает хорошую имитацию того, как ChatGPT будет работать в реальной жизни. Ограничением исследования является относительно небольшой размер выборок, но, несмотря на это, удалось выявить закономерности в показателях ChatGPT.

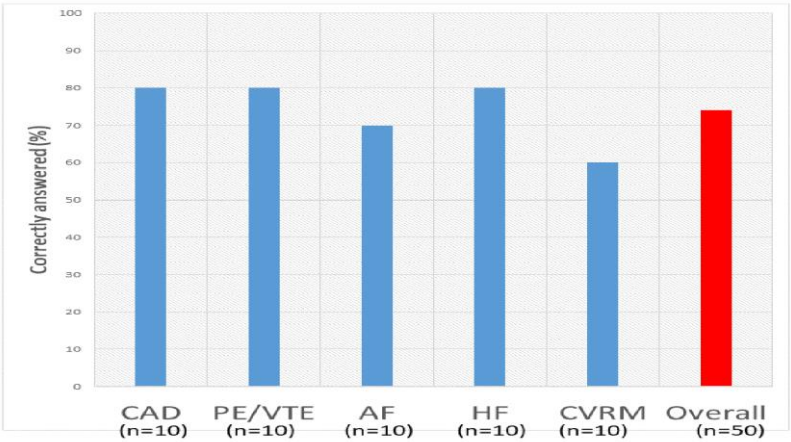


**Рис 3.** Результаты работы ChatGPT, неправильно ответившей на медицинские тесты по ключевым аспектам и основам практики распространенных сердечно-сосудистых заболеваний (n=50). CAD = ишемическая болезнь сердца (острый коронарный синдром и лечение атеросклероза коронарных артерий); PE/VTE = тромбоэмболия легочной артерии и венозная тромбоэмболия; AF= фибрилляция предсердий; HF= сердечная недостаточность; CVRM = факторы сердечно-сосудистого риска и лечение.

## IV. ЗАКЛЮЧЕНИЕ

Не так давно врачи первичной медико-санитарной помощи считали потенциал искусственного интеллекта ограниченным. Однако это мнение, вероятно, кардинально изменится в этом десятилетии, учитывая огромный интерес общества к этой новой технологии. Что же могли бы сделать эти модели искусственного интеллекта для будущего развития медицины? Во-первых, они могут быстро и точно обрабатывать большие объемы данных и могут выявлять закономерности и взаимосвязи в данных, которые людям было бы трудно обнаружить самостоятельно. ChatGPT, могли бы помочь врачам и пациентам различными способами: быстро и точно идентифицировать возможные заболевания, и планы лечения, основываясь на симптомах пациента и медицинских записях, тем более, если пациенты получат прямой доступ к ChatGPT,то они своевременно обратятся к врачу. С точки зрения врача, ChatGPT можно было бы использовать для консультирования благодаря его способности анализировать большие объемы медицинских данных для выявления закономерностей, влияющих на решения о диагностике или лечении. Чтобы взаимодействие технологий и медиков было грамотным и конструктивным, важно иметь компетентный медперсонал, чему может помочь американский опыт лицензирования врачей на основе GPT-4, а сегодня, возможно, и GPT-5.

С точки зрения исследований, ChatGPT также может оказать существенную помощь в разработке и тестировании новых лекарств, методов лечения и медицинских технологий. Анализируя данные прошлых испытаний и экспериментов, ChatGPT мог бы оптимизировать процесс и принимать более быстрые и обоснованные решения о разработке новых лекарств и методах лечения. Это могло бы привести к повышению скорости, эффективности и рентабельности медицинских разработок.

ChatGPT продемонстрировал и ограничения, сходные с теми, которые были обнаружены в более ранних языковых моделях, т.е. он не застрахован от предубеждений. Модели

обучаются на больших наборах данных, полученных с различных веб-сайтов, книг и других источников, некоторые из которых могут содержать неявные или явные предубеждения. И здесь очень важным становится вопрос корректности материала, на котором обучается медицинский GPT. В этом отношении есть хороший доказательный пример разработки корпуса для обучения модели [21,22].

Обучение LLM требуют огромного количества вычислительных ресурсов и энергии, что приводит к высоким финансовым затратам, которые представляют собой барьер для небольших организаций или исследователей. Важно стремиться к большей прозрачности в обучении и точной настройке моделей, таких как ChatGPT, чтобы обеспечить ответственную и заслуживающую доверия разработку и использование искусственного интеллекта в медицине.

## ЛИТЕРАТУРА

[1] *Andrew L. Beam, Ph.D., Jeffrey M. Drazen, M.D., Isaac S. Kohane, M.D., Ph.D., Tze-Yun Leong, Ph.D., Arjun K. Manrai, Ph.D., and Eric J. Rubin, M.D., Ph.D.* Artificial Intelligence in Medicine N Engl J Med 2023; 388:1233-1239 DOI: 10.1056/NEJMsr2214184

[2] *Charlotte J. Haug, M.D., Ph.D., and Jeffrey M. Drazen, M.D.* Artificial Intelligence and Machine Learning in Clinical Medicine, 2023N Engl J Med 2023;388:1201-8.DOI: 10.1056/NEJMra2302038

[3] *Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin.* Attention is all you need. Advances in Neural information processing systems, 30, 2017

[4] *Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al.* Improving language understanding by generative pre-training. arXiv preprint arXiv, 2018.

[5] *Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert*: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[6] *Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jerey Wu, and Dario Amodei*. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361, 2020.

[7] *Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, et al.* Holistic evaluation of language models, 2022.

[8] *Sebastien Bubeck and Mark Sellke.* A universal law of robustness via isoperimetry. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, Advances in Neural Information Processing Systems, volume 34, pages 28811{28822. Curran Associates, Inc., 2021.

[9] *Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, et al.* On the opportunities and risks of foundation models, 2021.

[10] *Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei*. Deep reinforcement learning from human preferences. Advances in neural information processing systems, 30, 2017.

[11] *Yuntao Bai, Andy Jones, Kamal Ndousse, et al.* Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback, April 2022.

[12] *Harsha Nori1, Nicholas King, Scott Mayer McKinney, Dean Carignan1, Eric Horvitz* Capabilities of GPT-4 on Medical Challenge Problems arXiv: 2303.13375 [cs.CL] https://doi.org/10.48550/arXiv.2303.13375

[13] *Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al.* Large language models encode clinical knowledge. arXiv preprint arXiv:2212.13138, 2022

[14] *Tiany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepa~no, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al.* Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. PLOS Digital Health, 2(2):e0000198, 2023

[15] *Alexandru Niculescu-Mizil and Rich Caruana*. Predicting good probabilities with supervised learning. In Proceedings of the 22nd international conference on Machine learning, pages 625{632, 2005

[16] *Peter Lee, Ph.D., Sebastien Bubeck, Ph.D., and Joseph Petro*. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine N Engl J Med

2023;388:1233-1239 DOI: 10.1056/NEJMsr2214184
[17] *Singhal K, Azizi S, Tu T, et al.* Large language models encode clinical knowledge. arXiv, December 26, 2022 https://arxiv.org/abs/2212.13138
[18] Automatically document care with the Dragon Ambient eXperience https://www.nuance.com/healthcare/ambient-clinical-intelligence.html
[19] *Alexander V. Eriksen, Scoren Mcoller, Jesper Ryg,* Use of GPT-4 to Diagnose Complex Clinical Cases NEJM AI 2023; 1(1) https://doi.org/10.1056/AIp2300031
[20] *P.Ralf, E.Harskamp, Lukas De Clercq* Performance of ChatGPT as an AI-assisted decision support tool in medicine: a proof-of-concept study for interpreting symptoms and management of common cardiac conditions (AMSTELHEART-2) medRxiv preprint doi: https://doi.org/10.1101/2023.03.25.23285475
[21] *Zeming Chen, Alejandro Hernandez Cano, Angelika Romanou, Antoine Bonnet,Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan,Andreas Kopf, Amirkeivan Mohtashami, Alexandre Sallinen et.all* MEDITRON-70B: Scaling Medical Pretraining for Large Language Models ArXiv:2311.16079v1 [cs.CL] 27 Nov 2023 https://doi.org/10.48550/arXiv.2311.16079
[22] *Musaev, M., Mussakhojayeva, S., Khujayorov, I., Khassanov, Y., Ochilov, M., Atakan Varol, H.* (2021). USC: An Open-Source Uzbek Speech Corpus and Initial Speech Recognition Experiments. In: Karpov, A., Potapova, R. (eds) Speech and Computer. SPECOM 2021. Lecture Notes in Computer Science(), vol 12997. Springer, Cham. https://doi.org/10.1007/978-3-030-87802-3_40

*Поступила в редакцию 5.01.2024*

**Цитирование:** *Адылова Ф.Т., Давронов Р.Р.* (2024). Нужен ли медикам gpt-4: анализ актуального мирового опыта. *Международный Журнал Теоретических и Прикладных Вопросов Цифровых Технологий*, 7(1), –С. 16-23. https://doi.org/10.62132/ijdt.v7i1.156

# DO MEDICINES NEED GPT-4: ANALYSIS OF CURRENT WORLD EXPERIENCE.

*Adilova F.T.*[1], *Davronov R.R.*[1]
[1] V.I.Romanovskiy Institute of Mathematics of the Academy of Sciences of the Republic of Uzbekistan, Tashkent, Uzbekistan
fatadilova@mathinst.uz , rifqat.davronov@mathinst.uz

**Abstract**. Large Language Models (LLM) have demonstrated remarkable capabilities in understanding and generating natural language in various fields, including medicine. The article presents an assessment of GMT-4 based on two points of view on the problem of applying this language model: developers from Open AI, Microsoft and medical users from two European projects. Over the past few years, LLMs trained on massive interdisciplinary corpuses have become powerful blocks in building task-oriented systems. The article considers three tasks: medical education, the efficiency of ChatGPT-4 in the clinic (consultations, transcripts of the conversation between the doctor and the patient ), and specific levels of diagnostic accuracy (different fields of medicine). The answer to the question about the need for medical GPT is in the world - it is positive.

**Keywords:** *LLM, Open AI, GPT-4, ChatGPT-4.*