УДК 004.032.26
ББК 32.813.5

DOI 10.47475/2618-9852-2023-8-2-48-54

РОЛЬ CHATGPT В НАУКЕ О ДАННЫХ

Д. О. Сулейманова, Т. Р. Магомаев

Грозненский государственный нефтяной технический университет имени академика М. Д. Миллионщикова, Грозный, Россия

ChatGPT – диалоговый интерфейс искусственного интеллекта, использующий обработку естественного языка и алгоритмы машинного обучения, – находит очень широкое применение в настоящее время. С учетом вероятного влияния этой модели на науку о данных в данной статье будет представлен обзор потенциальных возможностей и проблем, связанных с использованием ChatGPT в науке о данных.

Ключевые слова: *ChatGPT, наука о данных, синтетические данные, обработка на естественном языке, анализ данных.*

Продолжающийся рост больших данных привел к быстрому росту значимости науки о данных, когда данные сами по себе превращаются в самый ценный актив для любой организации [1; 2]. Наука о данных, которая включает в себя сбор, анализ и интерпретацию данных для извлечения информации и улучшения процессов принятия решений, определяется как «научное исследование создания, валидации и преобразования данных для создания смысла» [3]. Область науки о данных быстро развивается, чему способствуют достижения в области технологий и растущий спрос на аналитику, основанную на данных. Специалисты по обработке данных сталкиваются с многочисленными проблемами, включая необходимость повышения прозрачности и объяснимости моделей машинного обучения, проблемы конфиденциальности данных, а также трудности интеграции разрозненных источников данных. Кроме того, отрасль сталкивается с нехваткой квалифицированных специалистов.

Одной из многообещающих технологий является предварительно обученный трансформатор OpenAI ChatGenerative [8] – разговорный интерфейс искусственного интеллекта, который использует обработку естественного языка для понимания запросов человека и реагирования на них. ChatGPT основан на алгоритмах глубокого обучения, позволяющих ему генерировать высококачественные ответы на широкий спектр запросов. Это достигается за счет использования алгоритмов обработки естественного языка и машинного обучения для генерации текста, который является одновременно беглым и соответствующим контексту.

Модель состоит из нескольких уровней самоконтроля и нейронных сетей прямой связи, дающих ей возможность эффективно улавливать зависимости и взаимосвязи между словами в предложении. Механизм самоконтроля позволяет модели фокусироваться на различных частях входной последовательности при генерации выходных данных, что особенно полезно для задач обработки естественного языка.

Одним из ключевых преимуществ ChatGPT является его способность генерировать последовательные и контекстуально соответствующие ответы на вводимый текст, даже для открытых запросов, таких как беседы с чат-ботом. Это достигается за счет генерации распределения вероятностей по следующему слову в последовательности и выборки из этого распределения для генерации выходных данных. Многократно генерируя следующее слово на основе предыдущих, модель может создавать беглый и связный текст.

При этом ChatGPT имеет ряд проблем и ограничений. Некоторые проблемы связаны с неточностями, конфиденциальностью, предвзятостью и плагиатом [10]. Одной из самых больших проблем является потенциальная предвзятость [4], поскольку ChatGPT может воспроизводить и усиливать существующие предвзятости в данных, на которых он обучается, что приводит к неправильным и ошибочным прогнозам. Кроме того, существует риск плагиата, если пользователи просто копируют и вставляют информацию, сгенерированную ChatGPT, без надлежащих ссылок или подтверждения ее использования. Примечательно, что некоторые из отмеченных здесь ограничений устраняются с помощью его последней версии – GPT-4. Тем не менее это модель черного ящика, которая не может объяснить, как были сгенерированы ее выходные данные [5].

Несмотря на эти проблемы, ChatGPT полезен сообществу специалистов по обработке данных [2]. Его можно использовать для создания кода при автоматизации процессов сбора, форматирования или очистки данных; определения структур данных; указания того, какую инфографику следует создавать и какую информацию она должна содержать; создания учебных материалов; определения источников данных, необходимых для конкретных задач; создания синтетических данных; предоставления рекомендаций по соблюдению требований, регулированию и практических шагов для обеспечения законности, беспристрастности и этичности операций с данными; помощи в определении аналитических процессов, ведущих к лучшим практикам [9]. Поскольку технология продолжает развиваться, вполне вероятно, что она станет достаточно важным инструментом в области науки о данных.

Генерация синтетических данных с помощью ChatGPT включает в себя обучение языковой модели на большом массиве текстовых данных, а затем ее использование для генерации новых синтетических данных на основе шаблонов и структур, извлеченных из обучающих данных. Это можно сделать, предоставив модели приглашение или начальный текст, который затем используется для генерации нового текста, похожего по стилю и содержанию на исходные данные. Одним из главных преимуществ использования синтетических данных является то, что они могут помочь решить распространенную во многих приложениях машинного обучения проблему нехватки данных. Генерируя новые данные, похожие на исходные, модели машинного обучения можно обучать на более крупном и разнообразном наборе данных, что может привести к повышению производительности и обобщению. Кроме того, синтетические данные также могут быть использованы для расширения существующих наборов данных путем добавления новых примеров или вариаций существующих.

Однако при использовании синтетических данных в машинном обучении также есть некоторые ограничения и проблемы. Одной из главных является обеспечение того, чтобы синтетические данные были репрезентативными для реальных данных и отражали те же шаблоны и структуры, которые присутствуют в исходных данных. Это может быть особенно сложно в приложениях, где данные сложны или многогранны и где лежащие в их основе закономерности недостаточно понятны.

Другая проблема заключается в обеспечении того, чтобы синтетические данные не вносили никаких искажений или артефактов, которые могли бы повлиять на производительность или справедливость модели машинного обучения. Например, если синтетические данные генерируются на основе предвзятого или неполного набора данных, это может привести к созданию предвзятой или неточной в определенных контекстах модели. Точность и надежность ответов ChatGPT зависят от нескольких факторов, в частности, качество и разнообразие обучающих данных, сложность и двусмысленность вводимого текста, а также конкретная задача или задаваемый вопрос.

ChatGPT может испытывать трудности с предоставлением точных ответов в определенных ситуациях, таких как:

- двусмысленные или неясные вопросы; если вводимый текст двусмыслен или не содержит достаточного контекста, чтобы ChatGPT мог понять предполагаемый смысл, это может привести к неточным или нерелевантным ответам;
- вопросы, не относящиеся к предметной области; ChatGPT обучается на большом объеме текста из различных областей и тем, но может не обладать достаточными знаниями или экспертизой в определенных областях, что приводит к неточным ответам;
- предвзятые или неточные данные; данные для обучения ChatGPT получены из Интернета и могут содержать предвзятую или неточную информацию, что может повлиять на точность и надежность его ответов;
- сложный или технический язык; ChatGPT может испытывать трудности с пониманием и генерированием ответов на сложный или технический язык, такой как научная или юридическая терминология, обычно не используемая в повседневном языке;
- сложные интегралы; ChatGPT может испытывать трудности со сложными интегралами, которые требуют глубокого понимания

исчисления и других передовых математических концепций; например, если вас попросят решить интеграл, требующий использования таких методов, как интегрирование по частям, подстановка или частичные дроби, ChatGPT может выдать неправильный ответ;

- необычные системы счисления; если вас попросят выполнить вычисления в необычной системе счисления, такой как базовая 3 или базовая 16, ChatGPT может быть не в состоянии предоставить точный ответ; это связано с тем, что модель в основном обучена десятичной системе счисления и может быть недостаточно знакома с другими системами;
- многомерное исчисление; в то время как ChatGPT может обрабатывать некоторые базовые вопросы многомерного исчисления, с более сложными вопросами, включающими частные производные, градиенты и множественные интегралы, он может испытывать трудности;
- абстрактная алгебра; ChatGPT, возможно, не сможет генерировать точные ответы на вопросы, связанные с абстрактной алгеброй, такие как теория групп, теория колец и теория поля; эти темы требуют глубокого понимания передовых математических концепций и могут выходить за рамки учебных данных ChatGPT.

Несмотря на эти проблемы, использование синтетических данных становится все более популярным во многих приложениях машинного обучения, особенно в областях компьютерного зрения и обработки естественного языка. Комбинируя синтетические данные с данными реального мира, модели машинного обучения можно обучать на более крупных и разнообразных наборах данных, что может помочь повысить их точность и надежность в реальных приложениях.

Сравнение ChatGPT с другими подобными приложениями (*табл.*) также является важным шагом в оценке его сильных и слабых сторон. Существует несколько других популярных языковых моделей, которые обычно используются для обработки естественного языка, – GPT-2 и GPT-3 от OpenAI, BERT от Google и RoBERTa от Facebook (проект Meta Platforms Inc., деятельность которой в России запрещена).

Одним из ключевых преимуществ ChatGPT является его гибкость и простота использования.

Сравнительная характеристика ChatGPT

Модель	Год	Объем обучающего набора данных	Сильные и слабые стороны		Требуется тонкая настройка (да/нет)
			Сильные стороны	Слабые стороны	
ChatGPT	2020	1.5 B	Гибкость, простота использования, конкурентоспособная производительность	Ограниченная производительность при работе с небольшими наборами данных, возможность получения бессмысленных ответов	Да
GPT-2	2019	1.5 B	Высококачественная генерация языка, высокая производительность при выполнении заданий по заполнению текста	Ограниченная производительность при выполнении других задач НЛП	Да
GPT-3	2020	До 175 B	Высококачественная генерация языка, высокая производительность при решении широкого спектра задач НЛП	Ограниченная интерпретируемость, возможность предвзятости и этические проблемы	Да
BERT	2018	340 M – 1.1 B	Высокая производительность при классификации текстов и ответах на вопросы	Ограниченная производительность при выполнении задач по созданию языка	Да
RoBERTa	2019	125 M – 2.6 B	Высококачественное представление языка, высокая производительность в ряде задач НЛП	Ограниченная интерпретируемость, возможность предвзятости и этические проблемы	Да

ChatGPT может быть точно настроен для широкого спектра задач обработки естественного языка – от генерации языка до классификации текста и ответов на вопросы, с относительно небольшими изменениями архитектуры базовой модели. Это делает его универсальным инструментом, который может быть адаптирован к различным случаям использования.

Было показано, что с точки зрения производительности ChatGPT может конкурировать с другими современными языковыми моделями, особенно для генерации ответов на естественном языке. Например, недавнее исследование Microsoft Research показало, что ChatGPT превосходит другие языковые модели, включая GPT-3 и BERT, по целому ряду задач генерации языка, таких как обобщение и перефразирование.

Однако одним из потенциальных ограничений ChatGPT является его зависимость от больших объемов обучающих данных. Хотя модель может быть точно настроена на небольших наборах данных, ее производительность может быть ограничена по сравнению с другими моделями, разработанными специально для работы с ограниченными ресурсами (например, ALBERT от Google). Другое потенциальное ограничение ChatGPT – его способность генерировать последовательные и соответствующие контексту ответы. Несмотря на то, что модель продемонстрировала впечатляющую производительность при генерации ответов на естественном языке, все еще существуют случаи, когда сгенерированные ответы могут быть бессмысленными или неуместными, особенно если речь идет о сложном или с нюансами языке.

Итак, ChatGPT выделяется своей гибкостью, простотой использования и конкурентоспособной производительностью, но отмечается риск получения бессмысленных ответов на небольших наборах данных. GPT-2 свойственна высококачественная генерация языка, в то время как GPT-3 отличается высокой производительностью в широком спектре задач НЛП, но ограниченной интерпретируемостью и потенциалом предвзятости и этических проблем. BERT известен высокой производительностью в задачах классификации текстов и ответов на вопросы, но имеет ограниченную производительность в задачах генерации языка. RoBERTa выделяется своим высококачественным языковым представлением и высокой эффективностью в различных задачах НЛП, но также обладает ограниченной интерпретируемостью и потенциалом предвзятости и этических проблем. В целом гибкость, простота использования и конкурентоспособная производительность ChatGPT делают его ценным инструментом для целого ряда приложений обработки естественного языка. Однако, как и у любой языковой модели, у нее есть свои ограничения, и ее следует тщательно оценивать в контексте конкретных вариантов использования и требований к производительности.

Кроме вышеперечисленных, существуют также и иные конкуренты ChatGPT, а именно: Microsoft Azure Cognitive Services, Amazon Comprehend, IBM Watson Natural Language Understanding.

При этом в будущем стоит ожидать пополнение списка альтернатив и отечественными продуктами. Так, компания Яндекс разрабатывает свою версию чат-бота ChatGPT в рамках развития языковой модели из семейства YaLM (Yet another Language Model), модель которой аналогична решениям GPT-3. Отечественная версия нейросети станет частью «Поиска», «Алисы», «Почты», а также других сервисов. Голосовой помощник «Алиса» будет намного умнее, а сервис вроде «Почты» сможет как сам создавать письма в заданном стиле, так и излагать краткие смысловые выжимки из полученных сообщений.

Возвращаясь к главной теме статьи, отметим, что применение ChatGPT в науке о данных имеет несколько ключевых моментов. Во-первых, используя возможности ChatGPT, специалисты по обработке данных могут автоматизировать различные аспекты своего рабочего процесса, такие как очистка и предварительная обработка данных, обучение модели и интерпретация результатов [6]. Во-вторых, ChatGPT демонстрирует свой потенциал для анализа неструктурированных данных (например, отзывы клиентов, данные из социальных сетей и онлайн-обзоры), чтобы выявить новые идеи и улучшить процессы принятия решений. В-третьих, возможности автоматизации, предоставляемые ChatGPT в отношении процесса анализа данных и обработки естественного языка, могут позволить специалистам по обработке данных сосредоточиться на более сложных задачах, таких как разработка более точных прогнозных моделей и улучшение визуализации данных [5]. В-четвертых, его способность генерировать синтетические данные также делает его ценным ресурсом для специалистов по обработке данных, работающих с ограниченными или неполными наборами данных. В-пятых, нам больше не нужны обширные знания в области математики, информатики или искусственного интеллекта, чтобы сгенерировать код или программу, которые можно использовать для решения реальных проблем [7].

Ниже приведены несколько примеров, демонстрирующих статистические возможности ChatGPT:

- Языковое моделирование. ChatGPT обучается на огромных объемах текстовых данных,

что позволяет ему генерировать высокоточный и связный текст, имитирующий человеческий разговор [6].

- Классификация текста. ChatGPT может быть точно настроен для задач классификации текста, таких как анализ настроений, обнаружение спама и классификация тем. Модель может научиться классифицировать текст на основе шаблонов и ассоциаций в обучающих данных.
- Распознавание именованных сущностей. ChatGPT может распознавать и извлекать информацию об именованных сущностях в тексте – людях, местах и организациях [3]. Это полезно для таких задач, как извлечение информации и обобщение текста.
- Машинный перевод. ChatGPT также может быть точно настроен для задач машинного перевода, позволяя ему переводить текст с одного языка на другой [9]. Хотя это может быть не так точно, как специализированные модели машинного перевода, ChatGPT способен обеспечить полезную основу для задач машинного перевода.
- Ответы на вопросы. ChatGPT можно точно настроить для задач, связанных с ответами на вопросы, что позволяет ему отвечать на вопросы в зависимости от заданного контекста. Это полезно при обслуживании клиентов и взаимодействии с чат-ботом.
- Генерация текста. ChatGPT может генерировать текст, который является связным, контекстуально релевантным и имитирует человеческий язык. Эта возможность может быть использована для таких задач, как заполнение текста, создание истории и творческое написание.

В целом статистические возможности ChatGPT делают его универсальным и мощным инструментом для решения задач НЛП, способным изменить способ взаимодействия с машинами и обработки естественного языка.

ChatGPT может использоваться хозяйствующими субъектами в различных областях в условиях вводимых в России правовых и иных ограничений. Некоторые из возможных областей применения ChatGPT включают:

- Обслуживание клиентов. ChatGPT может использоваться для создания виртуальных помощников и чат-ботов, которые будут обслуживать клиентов и отвечать на их вопросы. Это может помочь компаниям снизить затраты на персонал и обеспечить 24/7 обслуживание клиентов.
- Маркетинг и реклама. ChatGPT может использоваться для создания контента для социальных сетей и рекламных кампаний. Он может генерировать заголовки, описания и тексты для постов в социальных сетях, а также ответы на вопросы пользователей.
- Финансы и инвестиции. ChatGPT может использоваться для анализа данных и прогнозирования цен на финансовых рынках. Он может также помочь создать инвестиционные стратегии и прогнозировать результаты.
- Юридические услуги. ChatGPT может использоваться для обработки большого объема юридических данных, например, для анализа и классификации договоров, ответов на юридические вопросы и т. д.
- Медицина и здравоохранение. ChatGPT может использоваться для анализа медицинских данных и создания индивидуальных рекомендаций по лечению. Он может также помочь автоматизировать процессы в медицинских учреждениях, например, обработку страховых претензий и др.
- Образование. ChatGPT может использоваться для создания интерактивных обучающих программ, например, для обучения иностранным языкам или навыкам программирования. Он может также помочь создать автоматизированные системы тестирования и оценки знаний.

Однако при использовании ChatGPT в России необходимо учитывать правовые и иные ограничения, связанные с обработкой персональных данных, защитой интеллектуальной собственности, использованием алгоритмов и технологий и др. Также важно соблюдать этические и профессиональные стандарты, связанные с использованием и размещением получаемой информации.

В заключение подчеркнем, что ChatGPT беспрецедентен и имеет потенциал произвести революцию в области науки о данных, и по мере дальнейшего развития технологии использование ChatGPT, вероятно, станет еще более распространенным, помогая специалистам по обработке данных совершенствовать свои рабочие процессы и достигать лучших результатов.

В дополнение к обсуждению многочисленных преимуществ ChatGPT для науки о данных необходимо напомнить о проблемах и ограничениях, связанных с этим нововведением, поскольку крайне важно, чтобы будущие поколения специалистов по обработке данных не только использовали эту технологию для повышения собственных навыков, но и научились делать это этично, добросовестно и с полным осознанием ее преимуществ и затрат.

Нельзя не упомянуть и юридическую составляющую вопроса. OpenAI, создатель ChatGPT,

не накладывает страновых ограничений на использование модели, если только это не запрещено законами стран, где модель будут применять. Однако необходимо учитывать, что некоторые страны могут иметь свои правила и законы в отношении обработки данных, включая обработку естественного языка. В таких случаях нужно обратиться к местным законодательным актам и рекомендациям по использованию данных технологий. Некоторые страны также могут блокировать доступ к веб-сайтам, где размещена модель ChatGPT, например, из-за цензуры Интернета или других соображений. В таком случае использование модели в этих странах может быть ограничено.

Кроме того, следует учитывать, что использование ChatGPT может быть ограничено авторскими правами и другими правовыми ограничениями на содержание, которые могут быть включены в тексты, используемые для обучения модели. В таких случаях необходимо соблюдать соответствующие правила и ограничения, чтобы избежать нарушения авторских прав и других законодательных актов. В целом страновые ограничения на использование модели ChatGPT зависят от законодательства каждой конкретной страны, однако OpenAI делает все возможное, чтобы обеспечить максимально широкое и безопасное использование своей технологии.

СПИСОК ИСТОЧНИКОВ

1. Avan A. A. Latest news about OpenAI, Google AI and what it means for data science // Date camp. 2022. № 15 (2). Pp. 34–38.

2. Adadi A. Review of algorithms effective for data processing in the era of big data // Big Data. 2021. № 32 (6). Pp. 7–13.

3. Ayer A. ChatGPT – prospects and challenges at the present stage of society development // Analytics. 2023. № 19 (4). Pp. 28–33.

4. Chou A. R., Perrigo B. The arms race in the field of artificial intelligence changes everything // Time. 2023. № 1 (3). Pp. 51–58.

5. Lund B. D., Wang T. Conversation about ChatGPT: How can artificial intelligence and GPT affect academia and libraries? // Bibliotech news of high technologies. 2023. № 3 (1). Pp. 38–44.

6. Luchini F. The real problem with synthetic data // Massachusetts Institute of Technology. 2021. № 15 (6). Pp. 1–4.

7. Mollik E. ChatGPT – a turning point for artificial intelligence // Harvard. 2023. № 5 (1). Pp. 12–17. URL: https://hbr.org/2022/12/chatgpt-is-a-tipping-point-for-ai (дата обращения: 08.04.2023).

8. Ruby M. How ChatGPT works: the model underlying the bot. On the way to data science // Tovarsdata. 2022. № 11 (7). Pp. 25–31.

9. Scialom T., Chakrabarti T., Muresan S. Debugged language models // Empirical methods of natural language processing. Abu Dhabi: DNEC. 2022. № 14 (3). Pp. 30–40.

10. Wiles J. Beyond ChatGPT: The future of generative artificial intelligence for enterprises // Gartner. 2023. № 3 (1). Pp. 13–19.

СВЕДЕНИЯ ОБ АВТОРАХ

Сулейманова Данна Олхазеровна – магистрант кафедры «Информационные системы в экономике» Грозненского государственного нефтяного технического университета имени академика М. Д. Миллионщикова, Грозный, Россия. dana.s.o.00s3@gmail.com

Магомаев Тамирлан Рамзанович – старший преподаватель кафедры «Информационные системы в экономике» Грозненского государственного нефтяного технического университета имени академика М. Д. Миллионщикова, Грозный, Россия. prikl-inf@mail.ru

THE ROLE OF CHATGPT IN DATA SCIENCE

D. O. Suleymanova, T. R. Magomaev

Grozny State Oil Technical University, Grozny, Russia

ChatGPT, an artificial intelligence dialog interface using natural language processing and machine learning algorithms, has a very wide range of applications. Given the likely impact of this model on data science, this paper will provide an overview of the potential opportunities and challenges of using ChatGPT in data science.

Keywords: *ChatGPT, data science, synthetic data, natural language processing, data analysis.*

REFERENCES

1. Avan A. A. (2022). Latest news about OpenAI, Google AI and what it means for data science. Date camp. No. 15 (2). Pp. 34–38. (In Engl.).

2. Adadi A. (2021). Review of algorithms effective for data processing in the era of big data. Big Data. No. 32 (6). Pp. 7–13. (In Engl.).

3. Ayer A. (2023). ChatGPT – prospects and challenges at the present stage of society development. Analytics. No. 19 (4). Pp. 28–33. (In Engl.).

4. Chou A. R. & Perrigo B. (2023). The arms race in the field of artificial intelligence changes everything. Time. No. 1 (3). Pp. 51–58. (In Engl.).

5. Lund B. D. & Wang T. (2023). Conversation about ChatGPT: How can artificial intelligence and GPT affect academia and libraries? Bibliotech news of high technologies. No. 3 (1). Pp. 38–44. (In Engl.).

6. Luchini F. (2021). The real problem with synthetic data. Massachusetts Institute of Technology. No. 15 (6). Pp. 1–4. (In Engl.).

7. Mollik E. (2023). ChatGPT – a turning point for artificial intelligence. Harvard. No. 5 (1). Pp. 12–17. Available at: https://hbr.org/2022/12/chatgpt-is-a-tipping-point-for-ai, accessed: 08.04.2023. (In Engl.).

8. Ruby M. (2022). How ChatGPT works: the model underlying the bot. On the way to data science. Tovarsdata. No. 11 (7). Pp. 25–31. (In Engl.).

9. Scialom T., Chakrabarti T. & Muresan S. (2022). Debugged language models. Empirical methods of natural language processing. Abu Dhabi: DNEC. No. 14 (3). Pp. 30–40. (In Engl.).

10. Wiles J. (2023). Beyond ChatGPT: The future of generative artificial intelligence for enterprises. Gartner. No. 3 (1). Pp. 13–19. (In Engl.).