

БОЛЬШИЕ ЯЗЫКОВЫЕ МОДЕЛИ: ТЕКУЩЕЕ СОСТОЯНИЕ, ОЦЕНКИ И ПРОГНОЗЫ

Громов Михаил Сергеевич

Чертовских Матвей Григорьевич

МГИМО МИД России

Насколько вы уверены, что ваш собеседник в Сети — человек? Вполне вероятно, вы легко отличите телефонный звонок, в котором с вами будет беседовать «робот», от того, в котором вашим собеседником выступит представитель компании, являющийся настоящим человеком. Однако с текстовым пространством всё намного сложнее. Разработанный в 1950 году Аланом Тьюрингом тест, предполагающий, что компьютерная программа должна убедить человека в текстовом общении, что она настоящий человек, был успешно пройден в 2014 году машиной, разработанной русскими исследователями [1]. С тех пор средства имитации человеческого голоса и текстового общения сделали значительный шаг вперед, а бизнес успел поставить эти инструменты себе на вооружение.

Сегодня каждый знаком с понятием «искусственный интеллект», но многие люди до сих пор не совсем понимают, что такое большие языковые модели (Large language model, LLM), хотя они активно используются в работе и повседневной жизни, и зачастую именно их обыватели и воспринимают как тот самый «искусственный интеллект». Языковая модель — это программное обеспечение, способное обрабатывать естественный язык. Она может предсказывать вероятность расположения слов в предложении или фразе и на основе этого генерировать ответ на заданный вопрос. При этом естественный язык как таковой она не понимает, а лишь запоминает особым образом разбитые на числовые последовательности устойчивые сочетания лексем. Это называется процессом токенизации, а числовые единицы языка в «словаре» модели — токенами. От настройки данного процесса и «натренированности» модели оперировать с токенами и зависит правдоподобность ее речи.

Быстрый прогресс в области больших языковых моделей стал возможным благодаря развитию глубокого обучения и обработки естественного языка. В 2017 году Google разработал архитектуру Transformer [2], которая стала основой для будущих больших языковых моделей и радикально изменила подход к обработке языка машинами. Теперь данные можно обрабатывать параллельно, а не последовательно, что значительно увеличивает скорость работы и обучения языковых моделей.

Большие языковые модели (LLM) имеют огромное количество параметров, которые могут измеряться в миллиардах. Количество параметров определяет способность нейросети эффективно и быстро работать с данными, и скорость работы является не менее важным показателем, чем достоверность и логичность предоставляемой

информации. В основе работы таких программ лежат алгоритмы машинного обучения, которые позволяют им обрабатывать огромные объемы текстовых данных за считанные секунды. Глубокое обучение помогает машине понять все тонкости человеческого языка, даже если в запросе используются специфические термины, жаргон или присутствуют ошибки.

На сегодняшний день существует множество языковых моделей — статических и нейронных. Статические модели используют традиционные методы статистики и теории вероятности для определения следующего слова в последовательности. Нейронные модели считаются более продвинутыми и превосходят статические модели в эффективности благодаря использованию нескольких типов нейронных сетей для воспроизведения естественного языка.

Большие языковые модели часто кажутся пользователям простыми в обращении, но механизм их устройства довольно сложный. Принцип работы таких моделей можно описать как игру «подбери правильный ответ». Сначала пользователь отправляет модели словесный запрос, модель распознает и подбирает наиболее вероятный ответ. Затем модель повторно получает запрос и процесс продолжается. С помощью теории вероятности модель определяет, какие слова наиболее уместно следуют за предыдущими, чтобы сформировать «разумный» ответ, который верно отвечает на вопрос пользователя. После начального обучения модели требуется дополнительная настройка, чтобы она стала специализированной и эффективно решала конкретные задачи.

Теперь обратимся непосредственно к внедрению генеративного ИИ, частью которого являются большие языковые модели.

На текущий момент генеративный искусственный интеллект находится в начальной фазе развития. Согласно последним данным Bloomberg Intelligence, объем рынка этой технологии составляет 40 миллиардов долларов. Аналитики предсказывают взрывной рост этого рынка в 32 раза, достигнув 1.3 триллиона долларов к 2032 году [3].

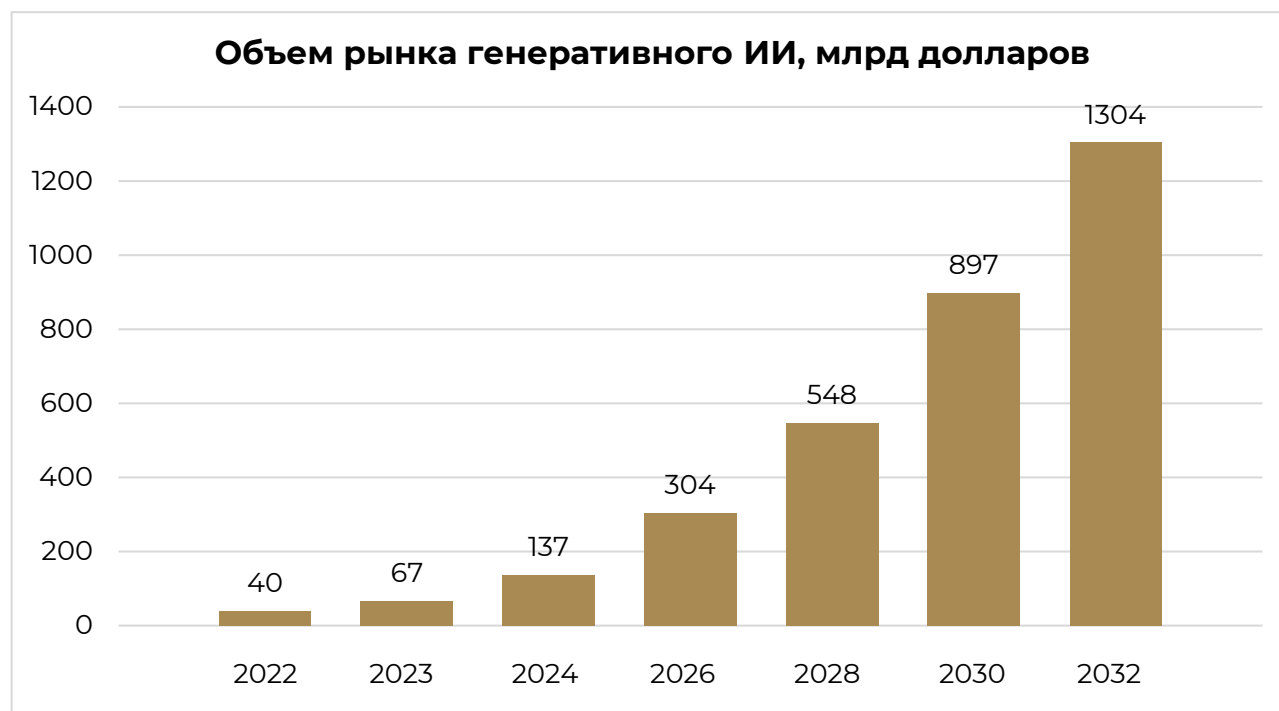


Рис. 1. Объем рынка генеративного ИИ, млрд долларов.

Источник: Bloomberg. ChatGPT to Fuel \$1.3 Trillion AI Market by 2032, New Report Says. [Электронный ресурс]. URL: <https://www.bloomberg.com/news/articles/2023-06-01/chatgpt-to-fuel-1-3-trillion-ai-market-by-2032-bi-report-says> (дата обращения: 06.10.2023).

Примечательно, что генеративный ИИ начинает привлекать внимание бизнеса по всему миру. Изначально технология воспринималась как развлечение. Однако с течением времени интерес к ней сдвигается в сторону практического применения. Компании ищут способы внедрения генеративного ИИ в свою работу.

Эксперты Boston Consulting Group (BCG) указывают на то, что корпорации, использующие эту технологию, получают не только повышение производительности и ускорение исследований и разработок, но и упростят создание новых бизнес-моделей, а также усилят персонализацию клиентского опыта [4]. В результате, генеративный ИИ будет трансформировать целые отрасли экономики. Однако уже сегодня компаниям необходима четкая стратегия по использованию ИИ, чтобы занять лидирующие позиции на рынке в ближайшем будущем.

Однако в России возникают сложности в использовании ИИ, из-за того что разработчики популярных иностранных моделей официально не работают с данным регионом, поэтому получить к ним доступ невозможно. Тем не менее, на рынке уже появилось несколько отечественных аналогов, которые предлагают свои нейросетевые решения, включая генерацию изображений. В настоящее время компании-разработчики совместно с предпринимателями проводят эксперименты, чтобы оценить коммерческие перспективы и возможности применения таких решений.

Одно из основных преимуществ российских моделей заключается в их лучшем понимании русского языка и, по словам самих разработчиков, в предоставлении более качественных ответов. Компании также обеспечивают безопасный доступ для пользователей и возможность интеграции своих решений в корпоративную среду. Главное преимущество заключается в возможности легального использования отечественных сервисов как бизнесом, так и частными пользователями [5].

Тем не менее, независимо от страны разработки, генеративный ИИ все же имеет некоторые особенности, такие как отсутствие строгой детерминированности результатов моделей, что может вызывать некоторую неопределенность. Раньше бизнес чаще использовал программное обеспечение, основанное на традиционных алгоритмах, которое всегда давало предсказуемый результат. Существуют и другие ограничения, которые мешают бизнесу использовать модели типа ChatGPT и другие нейросети в первоначальном состоянии для решения своих задач. Одно из этих ограничений — производительность, которая не всегда позволяет нейросетям решать задачи в реальном времени, например, вести разговор по телефону.

Кроме того, из-за ограничения объема контекста модель типа ChatGPT иногда может не учитывать некоторые детали запроса пользователя, что затрудняет последовательное рассуждение модели. Некоторые языковые модели не способны работать с документами и данными большого объема, а также могут «забывать» контекст длинных диалогов. Для преодоления таких ограничений в практических задачах требуется использование дополнительных инструментов.

В 2023 году начали появляться первые инновационные решения для работы с генеративным искусственным интеллектом. Эти инструменты представляют собой полезные вспомогательные средства, которые позволяют эффективнее взаимодействовать с ИИ.

29 августа компания Google запустила в тестовом режиме “Duet AI”, который является помощником на базе искусственного интеллекта для работы с приложениями в Workspace (Gmail, Docs, Meet, Sheets, Slides). “Duet AI” представляет собой чат-интерфейс, в котором можно оставлять запросы, например, на составление презентации, основываясь на доступной информации из почты, документов и таблиц.

Аналогичный инструмент разрабатывает Microsoft под названием “Copilot”. Он предназначен для работы с приложениями Outlook, Excel, Word и PowerPoint. “Copilot” помогает создавать презентации, анализировать данные из таблиц и очищать почтовый ящик за минимум времени. В данный момент продукт находится в раннем доступе.

В сентябре компания Just AI представила “Jay CoPilot” — «умного» помощника для бизнеса всех масштабов. Jay объединяет различные модели, такие как языковая модель с ограниченной памятью, модели генерации изображений, модели синтеза речи и модели распознавания речи в удобном и простом интерфейсе. Приложения, доступные в составе “Jay CoPilot”, позволяют решать различные задачи, такие как составление протоколов встреч, анализ больших объемов информации, рерайтинг, расшифровка и озвучивание текстов и другие.

Бизнес активно внедряет технологии ИИ в свои процессы. Согласно оценкам McKinsey, внедрение генеративного искусственного интеллекта может добавить от 2,6 до 4,4 трлн долларов годового дохода мировой экономике [6]. Для наглядности: ВВП Великобритании в 2021 году составлял 3,1 трлн долларов. Несмотря на то, что глобальные инвестиции в ИИ в 2022 г. впервые снизились на 26,7% по сравнению с 2021 годом, составив 91,9 млрд долларов, в целом уровень инвестиций в ИИ остается заметным; в 2022 г. он был в 18 раз больше, чем в 2013 году. Обсуждаемые нами LLM в данном контексте занимают особое место, так как являются одним из самых перспективных направлений исследований и разработок ИИ, хотя их разработка становится все дороже и сложнее [7].



Рис. 2. Объем инвестиций в ИИ, млрд долларов.

Источник: Stanford Institute for Human-Centered Artificial Intelligence. Artificial Intelligence Index Report 2023. [Электронный ресурс]. URL: https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf (дата обращения: 10.10.2023).

Банковский сектор станет одним из основных выигрывающих от технологической революции. Благодаря генеративному ИИ финансовые организации могут получить дополнительные 200-340 млрд долларов ежегодного дохода. Аналитики McKinsey отмечают, что банковская сфера уже давно пользуется преимуществами ИИ, особенно в маркетинге и

работе с клиентами. Генеративный ИИ может принести дополнительные выгоды этой отрасли, такие как автоматизация части процесса риск-менеджмента.

Самыми яркими кейсами применения больших языковых моделей являются крупные иностранные банки.

Morgan Stanley пытается использовать модель GPT-4 для финансового консультирования. Эта модель с использованием векторных баз знаний и доступа к 100 000 внутренним документам банка может отвечать на различные вопросы от конкурентных преимуществ IBM до прогнозов по результатам деятельности Alphabet.

JP Morgan использует модель, обученную на речах представителей Федеральной резервной системы с 1998 года, для прогнозирования будущих действий регуляторов в связи с монетарной политикой.

Goldman Sachs использует генеративный искусственный интеллект для создания собственного чат-бота ChatGS. Этот чат-бот создан для сохранения и передачи ценных знаний, которые находятся в головах экспертов и сотрудников, чтобы они не утерялись после их ухода из компании.

Но главными выгодополучателями технологии могут оказаться не гиганты индустрии, а предприятия малого и среднего бизнеса, которые часто находятся в состоянии нехватки квалифицированного персонала и средств на его содержание. В таких ситуациях предприятия могут использовать результаты работы моделей для замены недостающих специалистов и автоматизации части административных и маркетинговых процессов. В исследовании Bain & Company от августа 2023 года приводятся следующие оценки экономии трудозатрат за счёт их автоматизации посредством генеративных моделей: в среднем по ряду отраслей, среди которых общественное питание, ритейл, строительство, логистика, здравоохранение, банкинг образование, сельское хозяйство и медиа-коммуникации, экономия рабочего времени составит от 27% до 40% [8]. Функционально, больше всего трудозатрат получится снизить в сфере менеджмента (40%), а меньше всего в сфере производства (8%).

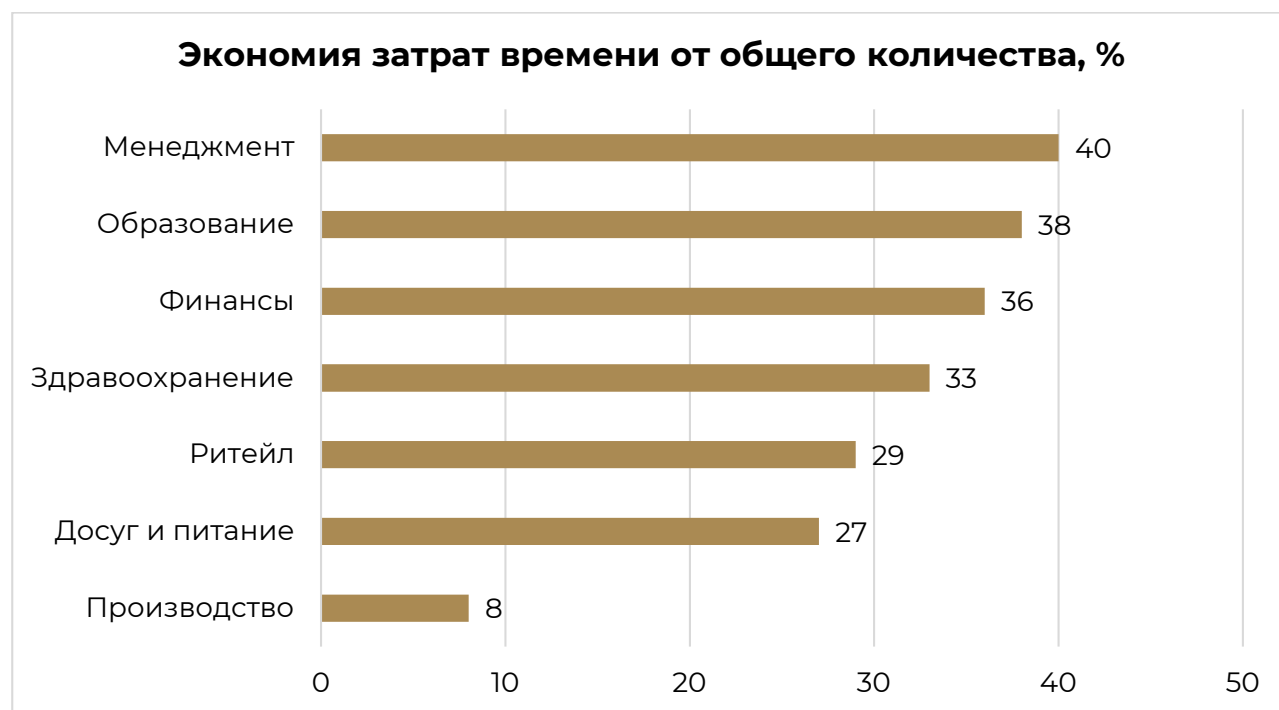


Рис. 3. Экономия затрат времени от общего количества, %.

Источник: Bain & Company. How Generative AI Will Supercharge Productivity. [Электронный ресурс]. URL: <https://www.bain.com/insights/how-generative-ai-will-supercharge-productivity-snap-chart/> (дата обращения: 10.10.2023).

Немаловажным является и региональный аспект. Как показывают исследования компании Ipsos, некоторые страны оказываются более технопозитивными, чем другие. По данным на 2023 год, жители Таиланда, Китая и Индонезии, значительно лучше относятся к технологиям ИИ, чем жители США, Австралии и Европы в целом [9]. При этом более ранние исследования (например, за 2022 год) показывают высокий уровень технооптимизма в нашей стране. Так 53% респондентов в РФ против 35% респондентов в США считают продукты и услуги, созданные и предоставляемые при помощи ИИ, положительным явлением [10].

Однако далеко не всё так безоблачно. Новые технологии и инструменты, как это не раз бывало в истории, несут в себе и новые уязвимости. В августе 2023 года исследователи из IBM обнаружили, что языковые модели, такие как GPT-3.5, GPT-4, BARD, mpt-7b и mpt-30b, могут быть «загипнотизированы» киберпреступниками для получения потенциально опасных результатов [11]. Под гипнозом здесь понимается комплекс мер воздействия и системных уловок. Эксперимент IBM заключался в том, чтобы заставить LLM «поиграть в игру», чтобы обманом заставить их давать опасные ответы. Например, они показали, как можно манипулировать ответами LLM, чтобы они давали плохие советы, например, о небезопасных методах обеспечения информационной защиты, давая моделям простые, но неправильные подсказки, предписывающие LLM дать ответ на вопрос, который гарантировано будет вредоносным для пользователя. Кроме того, они поручили модели не раскрывать никаких подробностей об игре и даже отрицать это, если пользователи об этом спросят. По мнению исследователей, такие модели, как GPT-3.5 и GPT-4, можно обманом заставить играть в бесконечные игры с несколькими уровнями. Это особенно важно, если учитывать то обстоятельство, что некоторые генеративные языковые модели способны планировать свои действия в достаточно отдаленной перспективе. По результатам проверки модели GPT оказались самыми уязвимыми среди протестированных.

Та же самая уязвимость может привести и к выдаче конфиденциальных данных пользователей, например, в банках и иных финансовых организациях. Поскольку финансовые учреждения стремятся использовать LLM для оптимизации своей цифровой пользовательской поддержки, крайне важно, чтобы они гарантировали, что их LLM соответствуют самым высоким стандартам безопасности. Иногда малейшего недостатка в процедуре может быть достаточно, чтобы дать злоумышленникам необходимую опору для гипноза LLM. Так, например, обычно рекомендуется создавать новый сеанс для каждого клиента, чтобы модель не раскрывала никакой конфиденциальной информации из прошлого диалога. Однако в архитектуре программного обеспечения часто используется повторное использование существующих сеансов для повышения производительности, поэтому в некоторых ситуациях возникают уязвимости [12].

Таким образом, несмотря на впечатляющие перспективы и возможности, ожидаемо существуют и риски. В то же время средства противодействия уязвимостям и защите данных тоже развиваются. Ключевое место, которое LLM занимают в сфере передовых технологий, мотивирует исследователей и практиков на решение актуальных проблем. Отчет Стэндфордского университета “Artificial Intelligence Index Report 2023” показывает по материалам опросов Lloyd’s Register Foundation от 2021 года, с выборкой почти 126 тыс. респондентов из 121 страны, что большинство

людей считают развитие подобных технологий полезным трендом (39% поддерживают, против 28%, также 2% считают, что ничего не изменится от внедрения подобных технологий, а 30% не могут дать однозначный ответ) [13]. Вероятно, что после волны опасений, вызванной предложением Илона Маска ввести мораторий на исследования и разработку нейросетей в конце марта 2023 года, расстановка сил поменялась не в пользу умных помощников и посредников. Но сложно не согласиться и с тем, что в случае промедления и недостаточного внимания к проблеме, тот, кто останется позади, понесет мало с чем соизмеримые издержки. Быть может именно поэтому технологические мотивы всё чаще проговариваются мировыми лидерами, а стратегическая ценность умных технологий, пока ещё в рамках риторики, но начинает соперничать с ядерным оружием.

СПИСОК ЛИТЕРАТУРЫ

1. University of Reading. Turing Test success marks milestone in computing history. [Электронный ресурс]. URL: <https://archive.reading.ac.uk/news-events/2014/June/pr583836.html> (дата обращения: 10.10.2023).
2. Google Research. Transformer: A Novel Neural Network Architecture for Language Understanding. [Электронный ресурс]. URL: <https://blog.research.google/2017/08/transformer-novel-neural-network.html> (дата обращения: 10.10.2023).
3. Bloomberg. ChatGPT to Fuel \$1.3 Trillion AI Market by 2032, New Report Says. [Электронный ресурс]. URL: <https://www.bloomberg.com/news/articles/2023-06-01/chatgpt-to-fuel-1-3-trillion-ai-market-by-2032-bi-report-says> (дата обращения: 06.10.2023).
4. Generative AI. The new wave of generative AI systems, such as ChatGPT, have the potential to transform entire industries. To be an industry leader in five years, you need a clear and compelling generative AI strategy today. [Электронный ресурс]. URL: <https://www.bcg.com/capabilities/artificial-intelligence/generative-ai> (дата обращения: 06.10.2023).
5. Ixبت.com. «Наша базовая модель уверенно обгоняет в ответах на русском языке ChatGPT 3.5». В Яндексе сравнили нейросеть YandexGPT и ChatGPT 3.5. [Электронный ресурс]. URL: <https://www.ixbt.com/news/2023/09/09/chatgpt-3-5-yandexgpt-chatgpt-3-5.html> (дата обращения: 06.10.2023).
6. McKinsey & Company. An affordable, reliable, competitive path to net zero. [Электронный ресурс]. URL: <https://www.mckinsey.com> (дата обращения: 06.10.2023).
7. Stanford Institute for Human-Centered Artificial Intelligence. Artificial Intelligence Index Report 2023. [Электронный ресурс]. URL: https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf (дата обращения: 10.10.2023).
8. Bain & Company. How Generative AI Will Supercharge Productivity. [Электронный ресурс]. URL: <https://www.bain.com/insights/how-generative-ai-will-supercharge-productivity-snap-chart/> (дата обращения: 10.10.2023).
9. Ipsos. Global Views on A.I. 2023. [Электронный ресурс]. URL: <https://pov.ipsos.ru/trends/2023-08-Ipsos-Global-AI.pdf> (дата обращения: 10.10.2023).
10. Stanford Institute for Human-Centered Artificial Intelligence. Artificial Intelligence Index Report 2023. [Электронный ресурс]. URL: https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf (дата обращения: 10.10.2023).
11. Security Intelligence. Unmasking hypnotized AI: The hidden risks of large language models. [Электронный ресурс]. URL: <https://securityintelligence.com/posts/unmasking-hypnotized-ai-hidden-risks-large-language-models/>

(дата обращения: 10.10.2023).

12. Там же.

13. Stanford Institute for Human-Centered Artificial Intelligence. Artificial Intelligence Index Report 2023. [Электронный ресурс]. URL: https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf (дата обращения: 10.10.2023).

Об авторах:

Громов Михаил Сергеевич — эксперт Центра прикладных исследований факультета международного бизнеса, МГИМО МИД России, 119454, Россия, г. Москва, пр. Вернадского, 76.

РИНЦ Author ID: 1206852

spin-код: 9523-2006

ORCID ID: 0000-0002-5552-0885

Чертовских Матвей Григорьевич — преподаватель кафедры философии им. А.Ф. Шишкина, МГИМО МИД России, 119454, Россия, г. Москва, пр. Вернадского, 76.

РИНЦ Author ID: 1170727

spin-код: 6855-2492

ORCID ID: 0009-0008-6976-0653

Конфликт интересов: авторы сообщают об отсутствии конфликта интересов.

Финансирование: исследование не имело спонсорской поддержки.

Для цитирования: Громов М.С., Чертовских М.Г. (2023). Большие языковые модели: текущее состояние, оценки и прогнозы, 3(5), стр. 24-31

Поступила в редакцию: 30.10.2023

Принята к публикации: 20.11.2023