УДК 004.8 EDN ZOXKXD

Философско-методологический анализ бенчмаркинга как средства оценки больших языковых моделей

© Р.Е. Батин

МГТУ им. Н.Э. Баумана, Москва, 105005, Россия

Представлен анализ методологии бенчмаркинга и проблематики его применения для оценки эффективности больших языковых моделей (БЯМ). Этот метод получил широкое распространение в различных научных областях — как гуманитарных, так и технических. Отмечено, что в сфере машинного обучения бенчмаркинг применяется давно и считается основным способом определения качества моделей и оценки их способностей решать разноплановые задачи, однако отсутствует строгая методология создания бенчмарков, организации процесса тестирования и интерпретации полученных результатов. Показано, что бенчмаркинг представляет собой многоаспектный и комплексный процесс, подверженный влиянию социокультурной, экономической и политической среды. Изучение данной проблематики имеет высокую актуальность как для разработки БЯМ, так и для всей области искусственного интеллекта, поскольку корректная методология оценки позволит минимизировать риски интеграции моделей в различные сферы человеческой деятельности. Рассмотрены отдельные этапы становления и развития бенчмаркинга. Особое внимание уделено критическому анализу современных методов оценки, их ограничениям и потенциальным искажениям при определении реальных возможностей интеллектуальных систем. Сформулированы концептуальные аспекты для философского осмысления бенчмаркинга и намечены направления дальнейших исследований, что составляет научную новизну данной работы.

Ключевые слова: *бенчмаркинг, большие языковые модели, оценка искусственного интеллекта, методология тестирования, эпистемология ИИ, социокультурные аспекты ИИ, валидность тестирования, операционализация знаний*

Успех больших языковых моделей (БЯМ) в задачах обработки естественного языка, а также развитие в сторону мультимодальности, т. е. возможности работы с различными видами данных, привели к интенсивному внедрению их в различные процессы, в том числе в бизнесе и государстве. Инструменты ИИ внедряют в сферы, которые прежде могли считаться не подлежащими автоматизации, например, в науку [1]. Помимо выполнения конкретных задач, данные системы могут играть роль личного ассистента, предлагая различную помощь. Так, к современным БЯМ, которые не имеют узкой специализации, любой человек может обратиться, например, с описанием симптомов своего заболевания с целью получить определенные рекомендации.

Опустим оценку рациональности такого поведения, однако уже сейчас можно наблюдать тенденцию, что БЯМ начинают играть важную роль в повседневной жизни людей.

Работа таких систем не безупречна. Следует выделить несколько категорий проблем, с которыми можно столкнуться при взаимодействии с ними. Во-первых, актуальной является проблема галлюцинирования ИИ: «...это явление, при котором нейросеть генерирует контент, который не имеет отношения к реальности или не соответствует исходным данным» [2]. Во-вторых, можно отметить достаточно большой спектр вопросов, касающихся безопасности самих моделей, а также работы с ними, которые раскрываются в понятии «согласованность» (alignment) [3]. Системы ИИ не должны рекомендовать пользователю выполнять противоправные действия, а также те, в результате которых может пострадать сам человек или быть нанесен вред обществу. В-третьих, следует выделить вопросы, связанные с наличием у таких моделей способности к рассуждению, логической, фактологической корректности. Отметим, что данными вопросами не ограничиваются все возможные проблемы, которые могут возникать при использовании БЯМ.

Сложность решения проблем отчасти заложена в самом принципе построения таких систем. Языковые модели являются стохастическими, т. е. получаемые от БЯМ данные носят вероятностный характер, а генерируются они относительно миллиардов параметров, значения которых были получены случайным образом в процессе обучения модели. Отдельные параметры или их группы (слои) не содержат сами по себе определенной семантики, что приводит к проблеме их интерпретируемости [4]. БЯМ предстают в виде «черного ящика» — невозможно однозначно определить, почему при определенных входных данных были получены те или иные ответы от системы.

Соответственно, задача обеспечения корректной работы системы, т. е. возможности получить ожидаемый, практически полезный результат, становится актуальной. Ее решение лежит в плоскости тестирования, а также интерпретации результатов этого тестирования. Частично обойти указанную ранее проблему «черного ящика» позволяют специальные подходы. Основным инструментом в данной области является бенчмаркинг. Понятие «бенчмарк», а также актуальность этой проблематики закрепляются на юридическом уровне в странах ЕС [5] и в американском «Президентском отчете по AI-науке» [6]. Это свидетельствует о тенденции к систематизации и развитию методологии оценки ИИ, в частности БЯМ.

Исследование способностей искусственного интеллекта имеет большую историческую перспективу. Точкой отсчета в данном вопросе можно считать работу А. Тьюринга [7], в ней он представил

известный мысленный эксперимент, с помощью которого предлагал исследовать способность технических средств к мышлению. Однако при практическом применении его к тестированию моделей ИИ могут возникать проблемы. Например, в тесте Тьюринга и его расширениях не предлагаются конкретные критерии, при выполнении которых, выражаясь языком статистики, должна приниматься или опровергаться гипотеза о наличии мышления. Роль «судьи» здесь отводится человеку, который должен интуитивно понять, взаимодействует ли он с человеком или с машиной, из чего делается вывод о наличии процессов мышления у исследуемой машины. Это приводит к проблеме отсутствия четких в практическом смысле критериев. Здесь стоит также вспомнить о так называемом эффекте Элизы [8, с. 456], свидетельствующем об обманчивости человеческого восприятия. При этом для рассмотрения прикладной пользы моделей требуются более устойчивые, численные показатели.

Сегодня практическим способом провести некоторое измерение над моделью машинного обучения, в частности над БЯМ, является бенчмаркинг. Данный метод может быть истолкован как сравнение объекта исследования по определенным признакам с некоторым эталонным объектом или значением [9]. Также под этим методом можно понимать сравнение нескольких объектов между собой по одной или нескольким характеристикам в рамках их тестирования на определенной задаче или в определенной среде. Примером здесь может служить исследование производительности компьютерных процессоров на тяжелых вычислительных задачах [10]. После проведения подобного тестирования составляется сравнительная таблица, а всем объектам исследования присваивается рейтинг в зависимости от результатов выполнения задачи.

Феномен «бенчмарка» в области ИИ зародился вследствие развития способов сбора и подготовки данных параллельно с модификацией самих методов обучения моделей. Приведем определение, данное в работе *AI and the Everything in the Whole Wide World Benchmark*: «В этой статье мы описываем бенчмарк как определенную комбинацию набора или наборов данных… и метрики… представляющей одну или несколько конкретных задач или наборов способностей, выбранных сообществом исследователей в качестве общей структуры для сравнения методов» [11, с. 2]. Эволюция понятия из «датасет» в «бенчмарк» происходит под воздействием социальной среды ученых или исследователей, что наглядно отражено в представленном определении. Необходимо детально рассмотреть отдельные вехи данного развития, чтобы лучше понять природу возникновения рассматриваемого явления.

Наборы данных в машинном обучении не приобретают статус бенчмарка с момента их создания. Как правило, они изначально представляют собой набор данных, формализующий определенную техническую задачу или проблему в виде совокупности пар «входные данные — выходные данные», которые впоследствии могут использоваться для оценки эффективности конкретной модели машинного обучения. Датасет формируется для решения определенной задачи, например, для распознавания символов с изображения, определения тональности текста. Существуют различные принципы построения подобных наборов данных с определенной спецификой предметной области. Можно выделить здесь основное правило: чем лучше датасет подходит под решаемую конкретную задачу, тем лучше будет итоговый результат модели. Тем не менее задача создания набора данных, который соответствовал бы всем необходимым характеристикам, требует больших усилий.

Период с конца 1950-х по начало 1960-х годов знаменуется активным развитием и становлением области машинного обучения, в частности, началом развития того направления, которое сегодня называют компьютерным зрением. Именно в рамках решения связанной с этим задачи исследователи Bell Labs, У. Хайлимен и Л. Каменцкий создали свой датасет рукописных чисел [12]. Они набрали 50 участников, чтобы те вручную написали 26 букв английского алфавита и десять цифр, в результате чего был получен набор данных из 1800 буквенно-цифровых символов, отображенных на перфокартах. Важно отметить, что три года спустя Хайлимен смог проверить работу различных моделей распознавания образов других исследователей на своих данных. Затем этот датасет получил распространение в научном сообществе, что послужило активному развитию дискурса вокруг него, а в результате были, например, разработаны новые математические методы оценки точности модели.

Сами разработчики этого набора данных отмечали в нем определенные слабые стороны. Тем не менее он стал на тот момент времени де-факто стандартом для разработки моделей распознавания, несмотря на ограничения, которые данный метод мог бы накладывать. Очевидно концептуальное изменение, переход из простого набора данных в статус бенчмарка, в рамках которого сообщество сыграло определяющую роль.

Важной вехой в развитии бенчмаркинга стало появление соревнований по машинному обучению и ИИ. Особенно важно выделить соревнования Netflix и ImageNet Large Scale Visual Recognition Challenge. Первые соревнования были организованы в 2009 г. одноименной компанией. Суть конкурса заключалась в том, чтобы по данным, предоставленным организатором, построить модель, которая способна

предлагать пользователю фильмы на основе выставленных им оценок. Соревнование смогло собрать более 50 тысяч участников. По его итогам была создана масштабная таблица лидеров, которая рассчитывалась на основе точности предсказания моделей на скрытой тестовой части данных. Однако ни один из победивших подходов не был имплементирован командой Netflix в силу излишней сложности и запутанности решения.

Вторые соревнования появились на основе существовавшего большого и детально проработанного датасета, состоявшего из набора аннотированных изображений. Целью их автора было создание «золотого стандарта» обучения моделей для классификации и распознавания объектов, а также продвижение идеи о том, что большие данные позволяют создавать лучшие алгоритмы. Соревнование просуществовало семь лет, что дало существенный толчок в развитии направления по обработке изображений.

Отдельно стоит отметить: Э. Дентон [13] в своем детальном анализе происхождения ImageNet показывает, что переход из датасета в бенчмарк часто представляет собой стихийный процесс. Этот процесс можно описать следующим образом: некоторая исследовательская группа создает на основе определенного датасета успешную модель, которая тиражируется посредством цитирования исходной работы. Затем создаются другие модели теми же авторами, которые пытаются обогнать исходную. Запускается цепная реакция, когда большая часть всего сообщества использует этот датасет.

Необходимо рассмотреть общие черты, которые прослеживаются в двух упомянутых соревнованиях. Во-первых, следует отметить мобилизацию интеллектуального ресурса вокруг одной тематики, повлекшей значительное развитие подходов и техник. Во-вторых, заметно смещение фокуса с построения инструмента для решения практической задачи к гонке за увеличение численных показателей. С одной стороны, эта гонка может способствовать поиску более разнообразных решений, с другой — переводит фокус с качественной оценки решения на количественную. Численные показатели на одном конкретном бенчмарке не отражают в полной мере прогресса технологий, скорее они характеризуют меру приспособленности технологии к этому бенчмарку. В-третьих, отметим зарождающуюся тенденцию на уменьшение роли академического сообщества и превалирование крупных финансовых игроков, которые могут задавать тренды на развитие определенных технологий и использовать это в качестве собственной рекламы.

Необходимо подвести итоги анализа исторической перспективы перехода от датасета к бенчмарку. Данный переход осуществляется при участии сообщества исследователей, а также при его росте. Он возникает посредством тиражирования в академической среде через

цитирование либо за счет привлечения крупных финансовых средств и их распределения через проведение соревнований. В результате этого один набор данных занимает центральную позицию в большинстве исследований по конкретной тематике и становится основным показателем для измерения качества работы модели.

Необходимо сделать акцент на некоторых последствиях подобного перехода. Так, центральное положение бенчмарка закрепляет недостатки и ошибки, которые могли быть допущены при его дизайне, что ведет фактически к игнорированию их во всех исследованиях, опирающихся на него. Игнорирование здесь можно трактовать как скрытое принятие возможных ограничений и слабых сторон всеми участниками процесса, выведение их в область нормы. В результате модели могут, например, иметь схожие ошибки генерализации на прикладных задачах, что уменьшает применимость получившегося инструмента. Отдельно стоит отметить смещение критерия оценки этой применимости. При возведении бенчмарка в рамки «золотого стандарта» происходит подмена качественных характеристик количественными, возникает явление гонки за превосходство численных показателей. При этом подобная оценка не дает прямого подтверждения практической ценности получившейся модели, как в примере с Netflix. Указанные выше тенденции и проблемы отчетливо находят отражение в современных бенчмарках.

Сегодня количество различных бенчмарков, которые используются сообществом для оценки БЯМ, достаточно большое. Перед тем как перейти непосредственно к анализу связанной с ними проблематики, необходимо привести их классификацию. Это позволит более отчетливо понять структуру рассматриваемой области. Авторы популярной платформы для оценки БЯМ Chatbot Arena [14] в своей работе предлагают такую классификацию:

1) по аспекту изучения:

- оценка истинности данных (ground-truth), которые воспроизводит модель,
- оценка человеческих предпочтений;

2) по изменчивости бенчмарка:

- динамические,
- статические.

Приведенные типы оснований по аспекту изучения не охватывают полную картину, которая складывается сегодня. Появляется большое количество бенчмарков, которые направлены:

1) на исследование безопасности;

2) проверку согласованности (alignment), т. е. того, насколько поведение модели согласовано с этическими, моральными и юридическими нормами;

3) проверку модели на галлюцинации — порождение фактически недостоверных данных;

4) исследование интеллектуальных способностей модели, т. е. проверка ее на соответствие критериям AGI.

Отметим, что в данную классификацию не были включены бенчмарки, которые проверяют способность модели к написанию программного кода, решению математических задач и доказательству теорем, а также многие другие узкоспециализированные бенчмарки, направленные на определенную конкретную предметную область. Задача классификации требует дальнейшего исследования и имеет большую актуальность.

Перейдем к анализу проблематики современных бенчмарков БЯМ. Авторы работы [15] выделяют девять основных позиций для критики.

1. «Проблема сбора данных, аннотации и документации». Ее можно разделить на три составляющие: «как, когда и кем» был создан бенчмарк, насколько данные соответствуют юридическим нормам и насколько сами данные в нем соответствуют поставленной цели измерения.

2. «Слабая валидность конструкции и эпистемологические претензии». Под валидностью конструкции авторы понимают соответствие целей измерения и то, что в реальности измеряется исследуемым бенчмарка. Они отмечают частые проблемы с концептуализацией и операционализацией понятий, которые в конечном счете используются в вольной манере, что напрямую отрицательно влияет на ценность таких измерений.

3. «Социокультурный контекст». Данная проблема выражается в том, что бенчмарк существует в социальной, культурной, экономической и политической среде, соответственно, подвержен их влиянию.

4. «Слабое разнообразие в бенчмарках и сферах применения». Здесь авторы подчеркивают то, что современные методы бенчмаркинга ИИ не используют несколько модальностей (аудио, видео, изображения), фокусируясь в основном на текстовых данных. Более того, сами методы сфокусированы на статической, одноразовой логике тестирования. Они выделяют важную проблему — по результатам бенчмаркинга можно сделать вывод о том, насколько успешна была модель, при этом по одному лишь численному показателю данного успеха невозможно определить, в чем именно она потерпела неудачу.

5. «Экономика, соревнования и коммерческие корни». Экономическую проблему авторы видят в доминации больших корпораций. Они приводят тезис о том, что Google, OpenAI и др. используют бенчмаркинг как средство маркетинга. Авторы также отмечают

большую зависимость академической среды от ресурсов, предоставляемых этими корпорациями.

6. «Фальсификация, мошенничество и измерение становятся целью». Здесь отмечается нарастающая тенденция к использованию различных ухищрений для лучшей оценки. Например, часто при взаимодействии с БЯМ исследователи выявляют, что в их обучающую выборку попали данные бенчмарков, с помощью которых измерялся их результат. Происходит так называемое загрязнение данных. При этом некоторые создатели бенчмарков вводят специальные правила и способы запрета и выявления подобных загрязнений. Конечно, подобный факт не перечеркивает полностью бенчмаркинг как способ оценки данных, но ставит новые вопросы и задачи для исследования.

7. «Сомнительная проверка сообщества и эффект колеи». Авторы показывают, что недостаток проверки сообществом бенчмарков, способов тестирования, данных для тестирования негативно влияет на всю индустрию. Также здесь подсвечивается проблема следования тенденциям приоритезации одного показателя для измерения над другими, что в конечном счете не позволяет в полной мере дать оценку прогресса в области.

8. «Быстрое развитие ИИ и насыщение бенчмарка». Скорость развития моделей ИИ превосходит скорость инструментария оценки, что ведет к отставанию в этой области.

9. «Сложность ИИ и “неизвестные неизвестные”». В данном пункте обсуждается эмерджентность современных БЯМ, которая может быть несоизмерима с возможностями тестирования. Авторы приводят такую цитату: «Бенчмарки БЯМ ограничены пределами создателей бенчмарков» [15].

Следует отметить, что указанные проблемы не являются особенностью современных исследований в этой области. Анализ, с одной стороны, исторических перспектив и, с другой — современной проблематики бенчмаркинга БЯМ дает основания для попытки сформулировать главные магистральные направления для их исследования.

Первое направление заключается в том, чтобы рассматривать социокультурные, экономические, политические, аксиологические аспекты создания, использования бенчмарка. Как можно заметить, именно через призму этих понятий достаточно четко раскрываются конкретные проблемы, которые могут стоять за тестированием БЯМ. Основой здесь является то, что работа над данными и моделями исторически была уделом коллективов и сообществ в силу сложности и экономической затратности как подготовки данных, так и обучения самих моделей. Соответственно, каждый автор или коллектив авторов может вкладывать определенные идеи и ценности. Это оказывает влияние на итоговый объект разработки — бенчмарк. Например, следует

выделить критику культурного смещения в работе БЯМ [16, с. 11], в том числе в сторону западного мира [17, с. 5]. Это во многом происходит из-за того, что само исследовательское сообщество имеет неравномерное распределение среди его участников. Не менее актуальными данные вопросы становятся при подключении больших корпораций, имеющих прямое финансовое влияние на всю сферу ИИ.

Можно утверждать, что бенчмарк напрямую оказывается во влиянии автора, сообщества исследований и непосредственно итоговых результатов работы получившихся моделей и сам влияет на них. Выделим векторы для дальнейшего изучения:

- исследование истории, предпосылок создания бенчмарка, общий контекст;
- автор бенчмарка: какие цели создания он преследовал;
- анализ взаимодействия «сообщество — бенчмарк». То, как бенчмаркинг оказывает влияние на сообщество, и то, как обратная связь сообщества может влиять на развитие самого бенчмарка;
- влияние корпораций, финансовых вложений и политической конъюнктуры на процесс оценивания БЯМ, а также влияние эффекта успешного или неуспешного результата на эти корпорации;
- анализ явления гонки за численными показателями среди исследовательского сообщества;
- анализ тенденций бенчмаркинга в исследовательских сообществах;
- вопросы правового регулирования сбора данных как для создания бенчмарка, так и для обучения БЯМ;
- влияние результатов бенчмаркинга на социум в целом, в частности при рассмотрении взаимодействия «человек — ИИ». Именно правильно выстроенная методология оценивания позволит гарантировать безопасность и конструктивность данной интеракции.

Второе направление состоит в исследовании эпистемологических и методологических вопросов бенчмаркинга. Оно заключается в изучении того, что именно оценивается конкретными бенчмарками, какие методы для этого используются. Здесь важно рассмотреть вопросы о том, как операционализируются эпистемологические понятия о знании и мышлении. Например, достаточно известный бенчмарк MMLU определяет понятие «понимание» через призму корректного прохождения теста по STEM дисциплинам, т. е. та модель, которая безошибочно отвечает на эти вопросы, обладает пониманием данных вопросов и самой предметной области. Такая операционализация является дискуссионной и не лишена определенных недостатков, поэтому даже авторы бенчмарка изменили свой подход, и в MMLU Pro добавили вопросы, которые требуют от модели некоторого рассуждения для ответа на вопрос.

Тем не менее важным и незатронутым остается вопрос о «мере незнания», оценки которой не происходит. Например, при прохождении бенчмарка, состоящего из вопросов в тестовой форме по физике, с результатом 90 % нельзя утверждать, что модель будет всегда справляться с 90 % физических задач и вопросов. Как уже было показано, причина этого кроется в особенностях всего процесса сбора и подготовки данных для теста. Более того, 10 % неуспешно решенных вопросов не дают возможности уверенно утверждать, насколько такая модель может быть полезна при решении той или иной задачи без проведения качественного анализа с разбором ошибок модели. Соответственно, точная формализация, операционализация и предельное понимание того, что именно будет оцениваться, являются критически важными аспектами всей области ИИ.

Следует отметить необходимость в последовательном и научном подходе к созданию самих бенчмарков: то, как именно собираются и подготавливаются данные, на основании каких критериев, как методологически должно проходить тестирование. Систематизация в данном направлении может положить начало «обратной инженерии знаний», суть которой будет заключаться в подготовке таких полных, связанных, наукоемких тестов для БЯМ и других моделей ИИ, которые позволяли бы однозначно и непротиворечиво отметить качество знаний модели.

Из этого могут проистекать следующие направления дальнейших изысканий:

- что именно исследует каждый отдельный бенчмарк, как операционализируются понятия, вокруг которых построен процесс тестирования;
- должны ли подниматься вопросы о том, какие аспекты ИИ уже изучены, насколько подробно, что еще следует изучить;
- интерпретация количественных оценок, перевод их в качественное русло. Определение мер «знания» и «незнания» модели;
- способы мошенничества и борьба с этим при прохождении бенчмарков;
- методология сбора и подготовки данных бенчмарков.

Необходимо отдельно упомянуть тип бенчмарков, который еще не рассматривался в настоящей работе. Суть их заключается в том, чтобы предоставить возможность пользователям взаимодействовать с двумя моделями, названия которых для них скрыты. В результате интеракции с моделями пользователю предлагают решить, какой из результатов для него предпочтительнее. Здесь можно видеть определенного рода преемственность идеям Тьюринга, что, в свою очередь, повлечет дальнейшую дискуссию, в том числе с обсуждением так называемого эффекта Элизы [8].

В рамках исследования был проведен анализ исторических предпосылок возникновения бенчмаркинга как средства оценки БЯМ. Был рассмотрен процесс перехода от датасета к бенчмарку, показано, как именно этот переход оказывал влияние на исследовательское сообщество, в частности рассмотрено явление гонки за численными показателями. Через эту историческую призму была исследована критика современных бенчмарков БЯМ, что позволило выделить два основных направления, в рамках которых открывается возможность к изучению бенчмаркинга. Первое направление заключается в рассмотрении бенчмарка как объекта в социокультурных, экономических, политических взаимодействиях. Немаловажен при рассмотрении аксиологический аспект его создания. Было показано, как бенчмарк может влиять на исследовательское сообщество и как само сообщество, а также иной контекст могут влиять на процесс оценки БЯМ, а также на интерпретацию этих оценок. Вторым направлением было выделено эпистемологическое и методологическое исследование бенчмаркинга, в рамках которого должны подниматься вопросы о том, что именно оценивает тот или иной бенчмарк, есть ли расхождения с его целевыми установками и фактическими способами их реализации. Здесь также необходимо рассмотреть аспекты построения датасета для тестирования, что может потребовать уже разработанную методологию инженерии знаний.

ЛИТЕРАТУРА

[1] Каримов К.С. Основные проблемы искусственного интеллекта в науке. *Постсоветский материк*, 2022, № 4, с. 59–65.

[2] Карпенко И.И., Меринов В.Ю. Галлюцинирование генеративного искусственного интеллекта: опасности и их предотвращение. *Донецкие чтения 2024: образование, наука, инновации, культура и вызовы современности: Материалы IX Международной научной конференции (Донецк, 15–17 октября 2024 г.). Т. 4: Филологические науки. Ч. 1.* Донецк, Издательство ДонГУ, 2024, с. 348–350.

[3] Zhou Z. Yu H., Zhang X., Xu R., Huang F., Li Y. How Alignment and Jailbreak Work: Explain LLM Safety through Intermediate Hidden States. *arXiv*, 2024, p. 27. URL: https://arxiv.org/abs/2406.05644 (дата обращения 01.03.2025).

[4] Singh C., Inala J., Galley M., Caruana R., Gao J. Rethinking interpretability in the era of large language models. *arXiv*, 2024, p. 7. URL: https://arxiv.org/abs/2402.01761 (дата обращения 01.03.2025).

[5] *Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act)*. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng (дата обращения 01.03.2025).

[6] *AAAI 2025 Presidential Panel on the Future of AI Research*. URL: https://aaai.org/about-aaai/presidential-panel-on-the-future-of-ai-research/ (дата обращения 01.03.2025).

[7] Turing A.M. Computing machinery and intelligence. *Mind*, 1950, vol. LIX, no. 236, pp. 433–460.

[8] Михайлов М.А., Кокодей Т.А. Риски злонамеренного использования искусственного интеллекта и возможности их минимизации. *Всероссийский криминологический журнал*, 2023, № 5, с. 452–461.
[9] Брежнева В.В. Бенчмаркинг в интернет-среде. *Труды СПбГИК*, 2006, с. 50–58.
[10] Ракитский А.А., Рябко Б.Я., Фионов А.Н. Аналитический метод сравнения и оценки производительности компьютеров и вычислительных систем. *ЖВТ*, 2014, т. 19, № 4, с. 84–98.
[11] Raji I.D., Bender E., Paullada A., Denton E., Hanna A. AI and the Everything in the Whole Wide World Benchmark. *arXiv*, 2021, p. 17. URL: https://arxiv.org/abs/2111.15366 (дата обращения 01.03.2025).
[12] Orr W., Kang E.B. AI as a Sport: On the Competitive Epistemologies of Benchmarking. *The 2024 ACM Conference on Fairness, Accountability, and Transparency. Rio de Janeiro Brazil: ACM*, 2024, pp. 1875–1884.
[13] Denton E., Hanna A., Amironesei R., Smart A., Nicole H. On the genealogy of machine learning datasets: A critical history of ImageNet. *Big Data & Society*, 2021, vol. 8, no. 2, p. 14.
[14] Chiang W.-L., Zheng L., Sheng Y., Angelopoulos A., Li T., Li D., Zhu B., Zhang H., Jordan M., Gonzalez J., Stoica I. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. *arXiv*, 2024, p. 29. URL: https://arxiv.org/abs/2403.04132 (дата обращения 01.03.2025).
[15] Eriksson M., Purificato E., Noroozian A., Vinagre J., Chaslot G., Gomez E., Fernandez-Llorca D. Can We Trust AI Benchmarks? An Interdisciplinary Review of Current Issues in AI Evaluation. *arXiv*, 2025, p. 22. URL: https://arxiv.org/abs/2502.06559 (дата обращения 01.03.2025).
[16] Liu Z. Cultural Bias in Large Language Models: A Comprehensive Analysis and Mitigation Strategies. *Journal of Transcultural Communication*, 2024, p. 21. URL: https://doi.org/10.1515/jtc-2023-0019 (дата обращения 01.03.2025).
[17] Yan T., Viberg O., Baker R.S., Kizilcec R.F. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 2024, vol. 3, no. 9, p. 9. URL: https://doi.org/10.1093/pnasnexus/pgae346 (дата обращения 01.03.2025).

Статья поступила в редакцию 11.06.2025

Ссылку на эту статью просим оформлять следующим образом:
Батин Р.Е. Философско-методологический анализ бенчмаркинга как средства оценки больших языковых моделей. *Гуманитарный вестник*, 2025, вып. 3.
EDN ZOXKXD

Батин Роман Евгеньевич — аспирант кафедры «Философия» МГТУ им. Н.Э. Баумана. e-mail: batinre@student.bmstu.ru