УДК 001

С.Р. Юников

Оценка интеллектуального поведения машины: от теста Тьюринга к benchmark-парадигме и большим языковым моделям

Аннотация:

Показана эволюция подходов к оценке интеллектуального поведения машины – от теста Тьюринга (1950) до современной benchmark-парадигмы в эпоху больших языковых моделей. Прослеживается логика смены исследовательских парадигм: от бинарного критерия через специализированные тесты (Winograd Schema Challenge, тест Лавлейс, CAPTCHA) к многомерным бенчмаркам (GLUE, MMLU, ARC-AGI) и новейшим методам эпохи LLM – мультимодальной оценке, LLM-as-a-Judge, краудсорсинговым «аренам» и динамическим бенчмаркам. Особое внимание уделяется проблеме контаминации данных.

Ключевые слова: тест Тьюринга, искусственный интеллект, бенчмарк, большие языковые модели, оценка ИИ, MMLU, ARC-AGI, LLM-as-a-Judge, контаминация данных.

Об авторе: Юников Семен Романович, МГТУ им. Н.Э. Баумана, аспирант факультета социальных и гуманитарных наук; эл. почта: unikormailgate.ru@mail.ru

Проблема оценки интеллектуального поведения машины занимает особое место в истории информатики и когнитивных наук. С одной стороны, это сугубо практическая задача: исследователям необходим надежный инструмент для измерения прогресса своих систем. С другой – за ней стоит один из глубочайших философских вопросов: что такое разум и где пролегает граница между подлинным пониманием и виртуозной имитацией? Семьдесят пять лет, прошедших с момента публикации статьи Алана Тьюринга «Computing Machinery and Intelligence» в журнале Mind, обнажили всю сложность этого вопроса [23]. За это время сменились несколько технологических укладов в области ИИ, и каждый из них ставил перед сообществом новые требования к методам оценки.

Показательна в этом отношении позиция Мелани Митчелл: наши представления об интеллекте сдвигаются по мере того, как машины осваивают области, ранее считавшиеся

прерогативой человека [16]. Когда машина «побеждает» человека, многие ожидают, что это докажет силу машинного интеллекта. Однако реакция сообщества нередко оказывается обратной: раз машина справилась с задачей, то она не была «по-настоящему» интеллектуальной. Этот «парадокс исчезновения» воспроизводился с каждым новым достижением ИИ – и с полным основанием применим к современным LLM, демонстрирующим впечатляющие баллы на стандартных бенчмарках.

Особую остроту ситуации придают результаты эмпирических исследований последних лет. В 2024 г. К. Джонс и Б. Берген провели строго контролируемый эксперимент с протоколом, максимально приближенным к оригинальному замыслу Тьюринга: GPT-4 был распознан как человек в 54% случаев – предсказание Тьюринга материализовалось с двадцатилетним опозданием [12]. Ку Мэй и коллеги в исследовании, опубликованном в PNAS, показали, что ответы GPT-4 на репрезентативных социальных опросах статистически неотличимы от человеческих по большинству параметров [15]. Эти результаты поставили под вопрос не только практическую ценность классического теста, но и базовую корректность подхода «имитация = интеллект». В этой связи целью статьи выступает систематический анализ эволюции подходов к оценке интеллектуального поведения машины – от концептуального замысла Тьюринга через промежуточные альтернативы к benchmark-парадигме и новейшим методам эпохи LLM.

В октябре 1950 г. в философском журнале Mind вышла статья «Computing Machinery and Intelligence», открывавшаяся вопросом «Могут ли машины мыслить?» [23]. Признав невозможность ответить на него напрямую, Тьюринг предложил операциональный заменитель – «игру в имитацию»: человек-судья ведет текстовый диалог с двумя невидимыми собеседниками (человеком и машиной) и пытается определить, кто из них кто. Тьюринг предсказал, что к 2000 г. средний судья при пятиминутном диалоге будет ошибаться более чем в 30% случаев. Это предсказание реализовалось с двадцатилетним опозданием – и именно это опоздание красноречиво характеризует темп прогресса ИИ.

Важно понимать исторический горизонт появления статьи. В 1950 г. электронные вычислительные машины были громоздкими установками, едва справлявшимися с арифметикой. На таком фоне тезис о потенциально мыслящей машине выглядел дерзким философским манифестом. Тьюринг опирался на собственную теорию универсальных вычислительных машин (1936): любой алгоритмически определяемый процесс может быть воспроизведен на достаточно мощном компьютере – а значит, если мышление – алгоритм,

то машина в принципе способна мыслить. Статья включала систематический разбор девяти «стандартных возражений» против машинного мышления, среди которых Тьюринг опроверг и знаменитое возражение Ады Лавлейс о неспособности машины создавать нечто по-настоящему новое [23]. Как отмечает Б. Гонсалвеш, тест следует рассматривать прежде всего как «прекрасный мысленный эксперимент», а не как практический протокол верификации [9].

Глубинный философский смысл предложения Тьюринга состоит в переходе от онтологического вопроса «что такое мышление?» к функциональному «как ведет себя мыслящее существо?». Тьюринг предложил бихевиористский критерий интеллекта: если поведение машины внешне неотличимо от человеческого, у нас нет оснований отказывать ей в статусе разумной. Это решение было прагматичным: оно уходило от неразрешимых метафизических споров о природе сознания и предлагало конкретный верифицируемый критерий. Тьюринговский функционализм нашел отзвук в широком философском движении, утверждавшем, что ментальные состояния определяются своими функциональными ролями, а не физическим субстратом, – открывая возможность для машинного разума без решения вопроса о его физической природе.

Наиболее известным контраргументом стал мысленный эксперимент Джона Серла «Китайская комната» (1980) [19]. Человек в комнате манипулирует китайскими иероглифами по инструкции и выдает правильные ответы, не понимая ни слова по-китайски. Следовательно, функциональная адекватность не равна подлинному пониманию – и тест Тьюринга не способен их различить. Практическое воплощение этой проблемы предоставила программа ELIZA (1966): простые текстовые подстановки Вейценбаума порождали у пользователей иллюзию реального собеседника. В строгом эксперименте Джонса и Бергена (2024) ELIZA получила лишь 22% – против 54% у GPT-4, – наглядно показав, что между имитацией и современными LLM пролегает качественная, а не только количественная граница [12].

Помимо философских возражений, тест страдает методологическими изъянами: он оценивает лишь текстовую коммуникацию, зависит от субъективности конкретного судьи и остается бинарным – без возможности измерить степень приближения к человеческому уровню. Именно поэтому Джеймс Стилгое предложил в Science (2023) «тест Вейценбаума»: вместо «мыслит ли машина?» спрашивать «полезна ли она и безопасна?» [21]. Как убедительно показывает Гонсалвеш, тест Тьюринга ценен как концептуальная рамка, но

неприменим как стандарт измерения – именно это противоречие и породило поиск альтернатив [9].

Первым развернутым ответом на ограниченность теста Тьюринга стал Winograd Schema Challenge (WSC), предложенный Хектором Левеском, Эрнестом Дэвисом и Леорой Моргенштерн в 2012 г. [14]. Суть теста – разрешение референциальной неоднозначности в предложениях с местоимениями, для корректной интерпретации которых требуется здравый смысл. Классический пример: «Городской совет отказал демонстрантам в разрешении, потому что они боялись насилия» – «они» относится к членам совета; стоит заменить «боялись» на «выступали за», и референт меняется. Для человека это тривиально; для машины, опирающейся на статистические паттерны, – крайне сложно. WSC обладал принципиальными преимуществами перед тестом Тьюринга: объективность, единственный правильный ответ, минимум статистических артефактов.

История WSC стала поучительной иллюстрацией цикла «создание – насыщение – замена». На протяжении 2012-2017 гг. точность систем оставалась близкой к случайному угадыванию (50–58%). Появление трансформерных моделей BERT и GPT изменило картину радикально: к 2019 г. точность превысила 90%. Коциян и коллеги в обзоре «The Defeat of the Winograd Schema Challenge» (2023) доказали, что эта «победа» достигнута не за счет понимания смысла, а за счет масштаба обучающих данных [13]. В ответ Сакагучи и коллеги создали WinoGrande – 44 тысячи примеров, прошедших процедуру «дебиасинга» [18]. WinoGrande оказался значительно сложнее, восстановив дискриминирующую силу теста – но ненадолго.

Тест Лавлейс, предложенный Брингсйордом, Белло и Ферруччи в 2001 г., ставит вопрос радикально иначе: способна ли машина к подлинному творчеству [5]? Согласно тесту, система должна произвести творческий артефакт, удовлетворяющий заданным ограничениям и не объяснимый судьей как детерминированное следствие заложенного алгоритма. Марк Ридл в 2014 г. предложил Lovelace 2.0 с более формализованными творческими заданиями – конкретным жанром, заданными словами и эмоциональной тональностью [17]. По сей день ни одна система не прошла тест Лавлейс в строгом смысле: вопрос о том, можно ли продукцию генеративных LLM считать подлинным творчеством или искусной рекомбинацией, остается открытым.

Особое место занимает CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) – «обратный тест Тьюринга» в промышленном масштабе [1].

Эволюция CAPTCHA представляет собой хронограф прогресса компьютерного зрения: текстовая CAPTCHA пала к 2014 г., reCAPTCHA v1 – около 2016 г., v2 с выбором изображений – к 2019 г. Современная reCAPTCHA v3 перешла к невидимому поведенческому анализу. Каждое «падение» версии CAPTCHA фиксировало конкретный рубеж когнитивных способностей машин – что делает CAPTCHA своеобразным барометром прогресса ИИ в области восприятия. Все три подхода – WSC, тест Лавлейс, CAPTCHA – объединяет общий изъян одномерности: каждый тестирует лишь один аспект интеллекта, оставляя остальные за кадром. Именно это и стало импульсом для создания многозадачных benchmark-наборов.

К середине 2010-х гг. в NLP царил хаос разнородных метрик: каждая исследовательская группа использовала собственные наборы данных, что делало объективное сравнение систем практически невозможным. В 2018 г. группа исследователей предложили GLUE (General Language Understanding Evaluation) – набор из девяти задач, охватывающих грамматическую приемлемость, анализ тональности, семантическое сходство и логический вывод [24]. Публичная таблица лидеров и единая совокупная метрика сделали GLUE не просто набором тестов, но инфраструктурой сравнения, мгновенно принятой сообществом. Шаврина и Малых (2021) показали, что арифметическое среднее как агрегирующая метрика имеет существенные изъяны – гармоническое и геометрическое среднее дает более справедливую картину для задач разной сложности [2].

Прогресс моделей на GLUE оказался настолько стремительным, что уже к 2019 г. трансформерные архитектуры на базе BERT вплотную приблизились к человеческому уровню – что потребовало создания SuperGLUE (2019) с более сложными задачами на многоходовое рассуждение и здравый смысл. Цикл «создание – насыщение – замена» GLUE/SuperGLUE обнажил фундаментальное противоречие benchmark-парадигмы: чем более стандартизирован тест, тем быстрее он становится целью прямой оптимизации. Это явление – частный случай «закона Гудхарта» («когда метрика становится целью, она перестает быть хорошей метрикой») – стало лейтмотивом всей последующей дискуссии об оценке ИИ.

С появлением GPT-3 стало очевидно, что оценка языковых навыков недостаточна для характеристики систем, претендующих на роль универсальных интеллектуальных агентов. В 2020 г. Дэн Хендрикс и коллеги из Калифорнийского университета в Беркли представили MMLU (Massive Multitask Language Understanding) – 57 предметных областей

от арифметики до клинической медицины на основе реальных академических и профессиональных экзаменов [11]. По данным авторов, GPT-3 при появлении показывал точность лишь на 20 процентных пунктов выше случайного угадывания – особенно слабыми оказались юриспруденция и медицина. MMLU быстро стал индустриальным стандартом; к 2023-2024 гг. топовые модели достигли 85-90%, что потребовало создания MMLU-Pro (2024) с десятью вариантами ответа и цепочечным рассуждением – переход снизил точность на 16-33 пункта. Гема и коллеги (2024) выявили, что значительная часть вопросов ряда разделов MMLU содержит ошибки – в подразделе по вирусологии их оказалось 57%, подчеркнув: качество самого бенчмарка не менее важно, чем качество оцениваемых моделей [8].

Параллельно в 2022-2023 гг. появился BIG-bench (Beyond the Imitation Game Benchmark) – беспрецедентный коллективный проект: 450 авторов из 132 организаций, 204 разнородные задачи [20]. Среди ключевых открытий – «прорывное» поведение (emergent capabilities): ряд задач, на которых малые модели дают случайные результаты, вдруг «решается» при достижении определенного масштаба – причем переход происходит резко. Это наблюдение радикально изменило представления о законах масштабирования и природе когнитивных скачков в LLM.

На фоне нараставшего скептицизма относительно знаниеориентированных бенчмарков Франсуа Шолле в 2019 г. предложил принципиально иной взгляд [7]. В работе «On the Measure of Intelligence» он критиковал господствующие бенчмарки за то, что они измеряют объем накопленных знаний, а не способность их приобретать. Шолле предложил новое определение: интеллект – это «efficiency of skill-acquisition over a space of tasks», и создал ARC-AGI (Abstraction and Reasoning Corpus) – набор визуальных головоломок, где каждая задача представляет несколько пар «входная сетка → выходная сетка», а система должна вывести правило и применить его к новому входу. Задачи тривиальны для людей (точность выше 85%), но крайне сложны для LLM: они требуют индуктивного вывода из малого числа примеров, а не паттерн-матчинга по обучающим данным.

Динамика результатов на ARC-AGI красноречива. В декабре 2024 г. модель o3 от OpenAI достигла 75,7% при стандартном уровне вычислений и 87,5% при высоком на ARC-AGI-1 [3] – первый случай превышения человеческого порога, зафиксировавший появление нового класса «больших моделей рассуждения» [4]. Обзор 82 подходов 2026 г. показал: при переходе от ARC-AGI-1 к ARC-AGI-2 производительность систем падает в 2-3 раза для всех

без исключения парадигм – программного синтеза, нейросимволических гибридов и чисто нейросетевых архитектур [22]. Люди при этом сохраняют почти идеальную точность на всех версиях. ARC Prize – соревнование с призовым фондом 1 млн долларов, учрежденное в 2024 г. [3], – наглядно демонстрирует, как академический инструмент оценки способен стать движущей силой исследовательского прогресса.

Быстрое развитие мультимодальных систем ИИ поставило новые требования к методам оценки. Задачи классических NLP-бенчмарков принципиально текстоцентричны и не позволяют оценить интеграцию информации из разных модальностей. По мере того как LLM стали применяться для открытых задач (написания текстов, кода, стратегических рекомендаций), стандартные бенчмарки с закрытыми вопросами перестали охватывать качество вывода. Возник подход LLM-as-a-Judge: использование мощной языковой модели для оценки ответов других моделей. Комплексный обзор Дж. Гу и коллег (2024), охватывающий более 150 публикаций, систематизирует таксономию методов и типичные предвзятости [10]: позиционная предвзятость (предпочтение ответам, стоящим первыми), склонность к многословию, самопредпочтение модели. Оценка ИИ должна учиться у психометрики: адаптивное тестирование (CAT), теория ответов на задания (IRT) и многомерные профили способностей позволят преодолеть ограничения статической benchmark-парадигмы.

В мае 2023 г. группа LMSYS запустила Chatbot Arena (ныне LMArena): пользователь одновременно взаимодействует с двумя анонимными моделями и выбирает лучший ответ; результаты агрегируются в рейтинг по системе Эло. К 2024 г. платформа насчитывала свыше миллиона парных сравнений более 100 моделей. Принципиальное достоинство «арена»-подхода – экологическая валидность: модели оцениваются в реальных условиях, по задачам, которые выбирают сами пользователи. Ограничения тоже очевидны: краудсорсинговая аудитория – не эксперты, а «инструктируемость» и стиль ответа могут завышать рейтинг моделей, оптимизированных на пользовательское одобрение, по сравнению с системами, нацеленными на точность. Одной из наиболее острых проблем benchmark-парадигмы выступает контаминация данных (benchmark data contamination, BDC) – проникновение тестовых примеров в обучающую выборку модели. Большинство крупных LLM обучаются на веб-данных, практически неизбежно содержащих производные от публичных бенчмарков материалы.

Наиболее перспективным ответом стали динамические бенчмарки. LiveBench (White и коллеги, 2024) автоматически генерирует новые задачи из свежих источников (последние публикации arXiv, актуальные спортивные и финансовые данные), используя временные отсечки: задача не может попасть в обучающую выборку, если создана после даты окончания обучения модели. Это делает контаминацию структурно невозможной. Комплексный обзор Чанга и коллег (2024), охватывающий более 800 публикаций [6], констатирует: только системная комбинация подходов – закрытые и открытые задания, статические и динамические бенчмарки, автоматическая и человеческая оценка – позволяет получить полноценную картину способностей системы.

Анализ эволюции подходов к оценке машинного интеллекта выявляет ряд системных противоречий, разрешение которых определит облик оценочной инфраструктуры ближайших десятилетий. Первое и главное из них – парадокс публичности бенчмарка. Чтобы быть научно значимым, тест должен быть открытым: это обеспечивает независимую верификацию и воспроизводимость. Но именно публичность делает его уязвимым для контаминации и прямой оптимизации под метрику. Закрытые тесты решают эту проблему, но порождают проблему доверия. Динамические бенчмарки – ближайшее к разрешению парадокса решение, однако они не охватывают все типы задач и требуют значительных ресурсов для поддержания актуальности.

Второй вызов – катастрофическое сокращение жизненного цикла бенчмарка. GLUE «насытился» за полтора года, SuperGLUE – за два с небольшим, MMLU – примерно за три. Скорость насыщения коррелирует с темпом прогресса моделей, определяемым доступностью вычислительных ресурсов. В этих условиях традиционная модель «создать бенчмарк – использовать несколько лет – заменить» уже не работает. Это требует либо динамической архитектуры самих бенчмарков (LiveBench), либо смещения акцента с абсолютных показателей на дифференциальный анализ задач, где отставание от человека сохраняется.

Третий вызов носит философский характер: что именно мы измеряем? Высокие баллы на академических бенчмарках не гарантируют надежной работы в реальных условиях – об этом свидетельствуют многочисленные случаи «галлюцинаций» систем, лидирующих в таблицах MMLU. Chatbot Arena приближает оценку к реальности, но оценивает предпочтение, а не объективное качество. ARC-AGI измеряет абстрактное обобщение, но его связь с практически значимыми задачами неочевидна. Ни один

инструмент не дает ответа на вопрос, «понимает» ли система или лишь имитирует понимание – вопрос, восходящий к тезису Серла о Китайской комнате [19].

На взгляд автора, наиболее перспективным направлением является синтез психометрических методов с динамической оценкой. Психометрика за столетие своего развития выработала эффективные инструменты: теорию ответов на задания (IRT), адаптивное тестирование (CAT), многомерные профили способностей. Перенос этого арсенала в область оценки ИИ позволит перейти от грубых агрегатных метрик к тонким, многомерным характеристикам: на каких типах задач и при каком уровне сложности производительность системы деградирует, по каким когнитивным осям она превосходит человека, а по каким принципиально отстает.

ARC-AGI-3, согласно планам разработчиков, должен включать задачи с активным исследованием среды – оценку не «что знает система», а «как действует в условиях неопределенности», – что принципиально сближает AI evaluation с задачами агентного ИИ. Наконец, по мере проникновения ИИ в здравоохранение, правосудие и образование оценка машинного интеллекта приобретает этическое и регуляторное измерение: как справедливо указывает Стилгое, обществу нужен не тест «мыслит ли машина», а система, удостоверяющая, что машина действует безопасно и в интересах тех, кому она служит [21].

Проведенный анализ позволяет сформулировать несколько содержательных выводов об эволюции подходов к оценке интеллектуального поведения машины. Тест Тьюринга заложил концептуальный фундамент всей дисциплины: оперируя понятием «имитация» как суррогатом «понимания», он на десятилетия определил язык и рамки дискуссии. При всей его исторической значимости тест обнаружил принципиальные ограничения – субъективность, одномерность, бинарность, уязвимость к ELIZA-эффекту, – сделавшие неизбежным поиск альтернатив. Показательно, что «прохождение» теста GPT-4 в эксперименте Джонса и Бергена не стало финальной точкой дискуссии, но лишь подтвердило: бинарный критерий имитации неадекватен сложности вопроса [12].

Альтернативные тесты первого поколения – Winograd Schema Challenge, тест Лавлейс, CAPTCHA – решили конкретные методологические проблемы, но воспроизвели общую проблему одномерности под другим углом. История WSC показательна вдвойне: она иллюстрирует и само явление насыщения, и механизм ответной реакции (создание WinoGrande), – воспроизводя рекуррентную динамику «тест – победа моделей – усложнение теста», пронизывающую всю историю оценки ИИ [14; 15; 16; 17; 18].

Benchmark-парадигма, сложившаяся с появлением GLUE [24] и достигшая зрелости в MMLU [11], BIG-bench [20] и ARC-AGI [7], обеспечила стандартизацию и измеримость прогресса в масштабах всего научного сообщества. Однако это достижение сопряжено с новыми рисками: гонкой за метриками, контаминацией данных и иллюзией понимания, создаваемой высокими баллами. Новейшие подходы – LLM-as-a-Judge, LiveBench, психометрически инспирированные протоколы – отвечают на реальные ограничения статических тестов, но каждый привносит собственные смещения. Из комплексного обзора Чанга и коллег следует: только системная комбинация методов позволяет получить сколько-нибудь полное представление о способностях системы [10; 6].

Главный вывод состоит в следующем: эволюция методов оценки машинного интеллекта отражает эволюцию самого понятия «интеллект» в научном дискурсе. От бинарного «мыслит или нет» мы прошли путь к многомерной, градуированной, предметно-специфичной оценке, в рамках которой система характеризуется не единственным баллом, а развернутым профилем способностей и ограничений. Переход – от вопроса «может ли машина имитировать человека?» к вопросу «что она действительно умеет и чего не умеет?» – является принципиальным. Будущее оценки ИИ принадлежит системам, сочетающим адаптивность психометрического тестирования, защиту от контаминации динамических бенчмарков, экологическую валидность «арен» и концептуальную глубину задач на обобщение, подобных ARC-AGI.

Библиографический список:

1. Кодов А. Тест Тьюринга и его альтернативы: эволюция оценки ИИ-систем // Sky.pro Wiki. URL: https://sky.pro/wiki/profession/alternativy-testu-tyuringa/ (дата обращения: 18.04.2026).

2. Шаврина Т. How not to Lie with a Benchmark: Rearranging NLP Leaderboards / Т. Шаврина, В. Малых // arXiv. URL: https://arxiv.org/abs/2112.01342 (дата обращения: 17.05.2026).

3. ARC Prize. ARC-AGI-1: Challenges Deep Learning // arcprize.org. URL: https://arcprize.org/arc-agi/1 (дата обращения: 18.04.2026).

4. ARC Prize 2025: Technical Report // arXiv. URL: https://arxiv.org/html/2601.10904 (дата обращения: 17.05.2026).

5. Bringsjord S. Creativity, the Turing Test, and the (Better) Lovelace Test / S. Bringsjord, P. Bello, D. Ferrucci // Minds and Machines. 2001. Vol. 11. Pp. 3-27.

6. Chang Y. A Survey on Evaluation of Large Language Models / Y. Chang, X. Wang, J. Wang // arXiv. URL: https://arxiv.org/pdf/2307.03109 (дата обращения: 17.05.2026).

7. Chollet F. On the Measure of Intelligence // arXiv. URL: https://arxiv.org/abs/1911.01547 (дата обращения: 17.05.2026).

8. Gema A. Are We Done with MMLU? / A. Gema et al. // arXiv. URL: https://arxiv.org/html/2406.04127v2 (дата обращения: 17.05.2026).

9. Gonçalves B. Turing's Test, a Beautiful Thought Experiment // arXiv. URL: https://arxiv.org/abs/2401.00009 (дата обращения: 17.05.2026).

10. Gu J. A Survey on LLM-as-a-Judge / J. Gu, X. Jiang, Z. Shi // arXiv. URL: https://arxiv.org/pdf/2411.15594 (дата обращения: 17.05.2026).

11. Hendrycks D. Measuring Massive Multitask Language Understanding / D. Hendrycks, C. Burns, S. Basart // arXiv. URL: https://arxiv.org/abs/2009.03300 (дата обращения: 17.05.2026).

12. Jones C. People cannot distinguish GPT-4 from a human in a Turing test / C. Jones, B. Bergen // Proceedings of NAACL. 2024. Pp. 5183-5210.

13. Kocijan V. The Defeat of the Winograd Schema Challenge / V. Kocijan, E. Davis, T. Lukasiewicz // Artificial Intelligence. 2023. Vol. 323. URL: https://arxiv.org/pdf/2201.02387 (дата обращения: 17.05.2026).

14. Levesque H. The Winograd Schema Challenge / H. Levesque, E. Davis, L. Morgenstern // Proceedings of the 13th International Conference on Principles of Knowledge Representation and Reasoning. 2012. Pp. 552-561.

15. Mei Q. A Turing test of whether AI chatbots are behaviorally similar to humans? / Q. Mei, Y. Xie, W. Yuan, M. Jackson // Proceedings of the National Academy of Sciences. 2024. Vol. 121, No. 9. URL: https://arxiv.org/pdf/2312.00798 (дата обращения: 17.05.2026).

16. Mitchell, M. The Turing Test and our shifting conceptions of intelligence // Science. 2024. DOI: 10.1126/science.adq9356.

17. Riedl M. The Lovelace 2.0 Test of Artificial Creativity and Intelligence // Proceedings of the AAAI Workshop. 2014. URL: https://arxiv.org/pdf/1410.6142 (дата обращения: 17.05.2026).

18. Sakaguchi K. WinoGrande: An Adversarial Winograd Schema Challenge at Scale / K. Sakaguchi, R. Le Bras, C. Bhagavatula, Y. Choi // arXiv. URL: https://arxiv.org/pdf/1907.10641 (дата обращения: 17.05.2026).

19. Searle J. Minds, Brains, and Programs // Behavioral and Brain Sciences. 1980. Vol. 3, No. 3. P. 417-457.

20. Srivastava A. Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models // Transactions on Machine Learning Research. 2023. 97 p.

21. Stilgoe J. We need a Weizenbaum test for AI // Science. 2023. DOI: 10.1126/science.adk0176.

22. The ARC of Progress towards AGI: A Living Survey of Abstraction and Reasoning // arXiv. URL: https://arxiv.org/pdf/2603.13372 (дата обращения: 17.05.2026).

23. Turing A. Computing Machinery and Intelligence // Mind. 1950. Vol. 59, No. 236. Pp. 433-460.

24. Wang A. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding / A. Wang, A. Singh, J. Michael // arXiv. URL: https://arxiv.org/abs/1804.07461 (дата обращения: 17.05.2026).

Iunikov S.R. **Assessment of Machine Intelligent Behavior: From the Turing Test to the Benchmark Paradigm and Large Language Models**

The paper analyzes the evolution of approaches to evaluating machine intelligence from the Turing test (1950) to the modern benchmark paradigm in the era of large language models. The logic of paradigm shifts is traced: from the binary criterion through specialized tests (Winograd Schema Challenge, Lovelace Test, CAPTCHA) to multidimensional benchmarks (GLUE, MMLU, ARC-AGI) and the newest LLM-era methods including multimodal evaluation, LLM-as-a-Judge, crowdsourced arenas, and dynamic benchmarks. Special attention is paid to data contamination as a systemic challenge.

Keywords: Turing test, artificial intelligence, benchmark, large language models, AI evaluation, MMLU, ARC-AGI, LLM-as-a-Judge, data contamination.