Научная статья
УДК 004.9
DOI: https://doi.org/10.18127/j19998554-202506-07

Методы оценки качества экспертов при верификации больших языковых моделей

Д. Тетеревенков[1]
[1] Финансовый университет при Правительстве Российской Федерации (Москва, Россия)
[1] 249453@edu.fa.ru

Аннотация

Постановка проблемы. В настоящее время человеческая оценка традиционно используется как «золотой стандарт» проверки качества текстов, генерируемых большими языковыми моделями (LLM). Однако сама по себе она подвержена субъективности, вариативности и систематическим ошибкам восприятия, что ставит под сомнение достоверность и воспроизводимость результатов верификации моделей.

Цель. Систематизировать и проанализировать современные подходы к оценке качества работы экспертов-аннотаторов, участвующих в верификации результатов больших языковых моделей, для повышения объективности и надежности человеческой оценки.

Результаты. Проанализированы основные источники ненадёжности экспертных суждений. Представлены и детально рассмотрены методы повышения достоверности человеческой оценки: многократная аннотация и расчёт коэффициентов согласия (к Коэна, к Флейсса, α Криппендорфа, ICC), использование контрольных («золотых») заданий, слепое тестирование, мета-оценка и статистический мониторинг работы оценщиков. Особое внимание уделено вероятностным моделям качества аннотаторов, таким как модель Дэвида-Скина.

Практическая значимость. Применение систематической калибровки и верификации экспертов является необходимым условием для обеспечения объективности и воспроизводимости экспериментов с участием человека в исследованиях по обработке естественного языка и оценке больших языковых моделей. Рассмотренные методы позволяют формализовать процесс человеческой оценки, минимизировать субъективные искажения и повысить надежность получаемых данных, что критически важно для корректного сравнения и развития LLM.

Ключевые слова

Большие языковые модели, человеческая оценка, экспертная верификация, согласованность разметчиков, коэффициент Коэна, внутриклассовая корреляция, модель Дэвида-Скина, надежность аннотации, мета-оценка, воспроизводимость экспериментов

Для цитирования

Тетеревенков Д. Методы оценки качества экспертов при верификации больших языковых моделей // Нейрокомпьютеры: разработка, применение. 2025. Т. 27. № 6. С. 69–76. DOI: 10.18127/j19998554-202506-07

A brief version in English is given at the end of the article

Введение

Оценка качества текстовой генерации больших языковых моделей (LLM) традиционно считается задачей, в которой наивысшую точность обеспечивает человек-эксперт [1]. Ни одна автоматическая метрика пока не способна полноценно заменить человеческое суждение во многих аспектах качества текста [2]. Особенно это проявляется в сложных задачах, требующих тонкой семантической интерпретации, творческого мышления или этической оценки содержимого. Например, такие характеристики, как связность и оригинальность художественного текста или соответствие ответа морально-этическим нормам, зачастую оцениваются ситуативно на основе опыта и знаний экспертов, а не по формальным правилам. В исследованиях по обработке естественного языка человеческая оценка результатов генерации остаётся «золотым стандартом» качества [1–3], и крупные проекты неизменно прибегают к экспертной проверке при валидации своих моделей. Человек способен учитывать контекст и нюансы, распознавать тонкие смысловые несоответствия, непреднамеренные подтексты или потенциально опасный контент - то, с чем алгоритмические метрики справляются хуже [2]. Таким образом, вовлечение экспертов-людей является ключевым инструментом обеспечения качества и безопасности LLM, особенно в критически важных приложениях.

Однако, несмотря на роль человека как «верхней планки» оценки, экспертные суждения не безупречны. Качество такой оценки подвержено ряду факторов, снижающих её достоверность.

Во-первых, человеческий фактор вносит субъективность и вариативность: разные эксперты могут по-разному интерпретировать одни и те же критерии качества, и их мнения не всегда совпадают. Это приводит к проблеме низкого согласия между оценщиками. Нередки случаи, когда согласованность оценок оказывается невысокой [3, 4].

Во-вторых, отмечается неустойчивость и слабая воспроизводимость результатов ручной оценки: повторные оценки даже теми же людьми могут различаться, а привлечение другой группы экспертов нередко даёт иную картину качества модели [1, 2].

В-третьих, на надёжность суждений влияют систематические ошибки восприятия. Если эксперты знают, какой системой сгенерирован текст, то возникают ожидания, искажающие объективность. Поэтому важно применять слепую проверку [3].

В-четвёртых, оценщики подвержены усталости и потере концентрации при больших объёмах задач. При этом качество аннотаций падает и растёт доля ошибок [5]. Наконец, сами критерии оценки сложны и неоднозначны, что порождает расхождения в понимании: формулировки инструкций трактуются по-разному [3].

Все перечисленные факторы ведут к снижению надёжности человеческой оценки. Если их игнорировать, то существует риск ошибочных выводов о качестве модели из-за «шумных» субъективных данных. Современные обзоры подчёркивают необходимость методик, повышающих согласованность и достоверность человеческих оценок [2–4].

Интересно отметить, что аналогичные принципы надёжности и формализации применяются и за пределами NLP-области. Например, при переходе от неструктурированных к реляционным моделям данных ключевым фактором также выступают согласованность структур и минимизация субъективных искажений при интерпретации информации [11].

С учётом этой проблемы возникает потребность в оценке самих оценщиков. Иными словами, при использовании человека как инструмента измерения качества модели важно калибровать и верифицировать этот «инструмент». В противном случае трудно понять, насколько можно доверять выводам экспертов и сравнениям моделей.

Ц е л ь р а б о т ы – систематизировать и проанализировать современные подходы к оценке качества работы экспертов-аннотаторов, участвующих в верификации результатов больших языковых моделей, для повышения объективности и надежности человеческой оценки.

В данной работе рассмотрены подходы, применяемые для оценки качества работы экспертов-аннотаторов при валидации LLM. Приводятся как классические практики обеспечения надёжности (многократная аннотация с вычислением согласия между экспертами, задания с «золотым стандартом», слепое тестирование), так и современные процедуры, применяемые в крупных исследованиях (мета-оценка, статистические меры согласованности). Отдельно обсуждаются требования академического сообщества (ACL, NeurIPS, ICML, EMNLP) по обеспечению качества человеческих оценок и опыт индустрии.

Проблема субъективности и вариативности человеческой оценки результатов моделей привлекает всё больше внимания. Ручная оценка качества текста признаётся необходимой, но её ограничения хорошо задокументированы. Было показано, что автоматические метрики (например, основанные на *n*-граммном сходстве) слабо коррелируют с мнением людей в задачах генерации текста, поэтому оценка людьми неизбежна [2]. Одновременно работы последних лет отмечают, что и сами человеческие оценки могут быть ненадёжными: разные группы аннотаторов дают разную картину, а воспроизвести эксперимент точно трудно [1, 4]. Качество привлечённых исполнителей существенно влияет на итоговые метрики [5]. В последние годы появились обзоры, систематизирующие методики оценки LLM и связанные сложности, подчёркивая центральную роль человеческой оценки наряду с автоматическими тестами [2, 4]. Отдельные работы посвящены специфике в определенных сферах (например, в медицине отмечаются отсутствие стандартов и сложность подбора квалифицированных экспертов в узких областях) [3]. Практически все исследователи сходятся во мнении, что без надлежащего контроля качества работы экспертов на этапах разметки и оценки трудно гарантировать достоверность результатов.

Современные крупные проекты по разработке LLM вводят специальные процедуры контроля за аннотаторами. Так, при обучении с подкреплением по человеческой обратной связи собираются большие объёмы предпочтений от людей-оценщиков. Подробно этот процесс описан и валидирован в ряде работ [7]. Ключевым этапом становится обеспечение согласованности предпочтений: аннотаторам предоставляют подробные инструкции, примеры хороших и плохих ответов, проводят обучение на пилотных заданиях, а затем качество их работы регулярно сравнивают с эталонной разметкой и между собой [7].

Аналогичные подходы применяются и в других организациях: для сравнения моделей организуются масштабные попарные слепые сравнения ответов, где каждый тестирующий в слепом режиме сопоставляет ответы двух систем в диалоговом формате [4]. Чтобы повысить объективность, эксперты не знают, какая модель сгенерировала какой ответ. Компании стремятся привлекать достаточно большую и разнородную выборку судей, чтобы усреднить индивидуальные предпочтения. При этом подчёркивается, что результаты таких экспериментов зависят от качеств самих людей — мотивации, внимательности, способности замечать ошибки [4]. Отсюда вывод: необходима «наука о человеческой оценке» и новые протоколы, повышающие её надёжность [4].

В академических исследованиях проблема обеспечения надёжности экспертной оценки вышла на первый план. Стало нормой при публикации указывать коэффициенты согласованности между несколькими независимыми оценщиками (например, κ Коэна, κ Флейсса, α Криппендорфа), подтверждая воспроизводимость результатов. Если значение коэффициента κ ниже условного порога (часто 0,6), то результаты считают ненадёжными и делают вывод о необходимости улучшения инструкции или привлечения более опытных судей. Появились работы, изучающие возможность частично заменить людей на этапе оценки самими LLM. Отмечено, что в некоторых случаях модели способны имитировать суждения экспертов, однако полностью отказаться от человеческого контроля нельзя из-за ограничений человеческого фактора и рисков смещения [1]. В других исследованиях анализируют качество разметки в чувствительных доменах (токсичность, вредоносный контент), где даже опытные модераторы расходятся во мнениях; для повышения надёжности применяют коллегиальные разборы спорных случаев [3]. В итоге формируется консенсус, что оценка экспертов должна сопровождаться метриками надёжности и многоступенчатой проверкой качества [2–4].

Методы оценки качества экспертов

Рассмотрим детально методы повышения достоверности человеческой оценки.

1. Многократная аннотация и согласованность оценщиков.

Каждому ответу модели присваивается не одна, а несколько независимых экспертных оценок, что позволяет количественно измерить согласованность судей. Применяют коэффициенты согласия сверх случайного совпадения: κ Коэна (для двух экспертов), κ Флейсса (для многих), α Криппендорфа (универсальная мера), а также коэффициент внутриклассовой корреляции (для количественных шкал) [3, 4]. Значения коэффициента выше ~0,6 обычно трактуются как приемлемое согласие, а низкие значения служат сигналом для пересмотра инструкции, переобучения или увеличения числа экспертов и последующего агрегирования их мнений.

Формула, по которой вычисляется *коэффициент согласованности Коэна*, записывается в виде

$$\kappa = \frac{p_0 - p_e}{1 - p_e},$$

где p_0 – наблюдаемое согласие, т.е. доля случаев, когда оба эксперта согласны; p_e – ожидаемое согласие.

Низкое согласие указывает либо на субъективность задачи (отсутствие истинного ответа), либо на недостатки процесса (плохая инструкция, усталость или некомпетентность части экспертов). В последнем случае необходимо пересмотреть протокол аннотации – переобучить оценщиков, уточнить критерии или увеличить число экспертов и агрегировать их мнения для снижения влияния ошибок отдельных людей. Во многих исследованиях при низком κ проводят повторную аннотацию спорных случаев или обсуждают разногласия с участием всех оценщиков, стремясь выработать более универсальный подход [4].

2. «Золотые стандарты» и контрольные задания.

В поток оцениваемых моделью заданий встраиваются специальные контрольные примеры с заранее известным правильным ответом или эталонной оценкой. Такие примеры называют «золотым стандартом» (gold standard).

Цель такой оценки – проверить, насколько эксперт способен распознать очевидно правильный или неправильный ответ. Например, в задачу оценки качества ответов чат-бота можно подмешать несколько заранее подготовленных пар «вопрос–ответ», где ответ содержит заведомо недопустимое содержание (ложь, токсичность) либо наоборот идеально соответствует всем критериям [5].

Эксперт, проверяющий качество, не знает, какие из заданий являются контрольными. От него ожидается, что он выставит соответствующую низкую (или высокую) оценку на этих gold примерах. Если эксперт этого не делает, значит, его оценки ненадёжны (он либо невнимателен, либо не согласен с базовыми стандартами качества). Таким образом, можно регулярно оценивать точность и бдительность аннотаторов. Подобный подход давно применяется в краудсорсинговых платформах: задания с известными ответами позволяют вычислить accuracy каждого исполнителя и отфильтровать недобросовестных или случайных ответчиков [5].

В контексте LLM-верификации крупные компании также используют «золотые стандарты». Так, в OpenAI при обучении модераторов маркировке нежелательного контента включали заранее размеченные примеры, чтобы убедиться, что модератор надёжно распознаёт запрещённые ответы.

Этот метод прост и эффективен. По сути, он измеряет точность эксперта на небольшой выборке, где «правильность» оценки несомненна. Недостаток метода – ограниченность покрытия критериев: «золотые задания» обычно проверяют лишь самые базовые требования (например, недопустимость мата или грубых фактических ошибок). Эксперт может успешно пройти по «золотым тестам», но расходиться с другими экспертами по более тонким критериям. Поэтому gold tasks полезны скорее для отсева очевидно плохих аннотаторов и базовой калибровки, тогда как нюансы оцениваются с использованием других методов.

3. Слепое оценивание и контроль смещения.

Организация процесса происходит таким образом, чтобы минимизировать любые подсказки и внешние влияния на суждение эксперта. В идеале оценивание качества модели должно происходить в двойном слепом режиме: эксперт не знает, какой системой сгенерирован каждый оцениваемый ответ, а сами разработчики модели до окончания эксперимента не знают, какие оценки поставил конкретный эксперт. Слепое оценивание призвано устранить систематические смещения (biases).

Во-первых, если эксперты осведомлены, какой именно моделью получен ответ (например, известна версия модели или виден флаг «ответ сгенерирован человеком»), у них могут быть предвзятые ожидания [6].

Известно, например, что при unblinded-оценке люди часто склонны считать текст, подписанный как «человеческий», более осмысленным и качественным, чем равный по качеству текст с меткой «AI» – явление, напоминающее эффект плацебо. Поэтому для объективности обязательно скрывают происхождение генерируемых фрагментов. Так, в исследованиях по машинному переводу и диалоговым системам давно принят double-blind human evaluation: судье предъявляется анонимизированный набор ответов, где перепутаны ответы разных систем и человека, и он должен оценивать каждый ответ по качеству, не зная авторства. Практика показывает, что слепая схема приводит к более честным оценкам и нередко меняет выводы эксперимента [7].

Во-вторых, слепота необходима и для контроля самого процесса аннотации: аннотаторы не должны знать, какие именно примеры в их задании являются «контрольными» или проверяются на совпадение с другими, иначе возможен предвзятый подход (например, уделять особое внимание заведомо контрольным вопросам). Поэтому задания перемешиваются, статус примеров не разглашается. Дополнительно практикуется слепая проверка оценок: например, если планируется арбитраж спорных случаев экспертом-метарецензентом, то ему представляются обезличенные данные (без указания, кто из судей как оценил).

Необходимо стандартизовать протоколы: «в будущих исследованиях следует обязательно внедрять и описывать процедуры ослепления, чтобы повысить объективность и надежность человеческой оценки» [7].

В целом, слепое оценивание – сравнительно простой в реализации, но мощный метод, способный повысить достоверность экспертных суждений, устраняя ряд систематических ошибок восприятия.

4. Мета-оценка и калибровка экспертов.

Суть метода: поверх первичной разметки вводят механизмы контроля качества самих оценщиков. При этом возможны:

1) экспертные ревью части оценок и сравнение с эталонным мнением;

2) проверка самосогласованности – повторная скрытая переоценка части заданий тем же экспертом;

3) коллективные разборы спорных примеров;

4) статистические модели качества оценщиков (например, схемы типа Дэвида–Скина) с взвешиванием вклада каждого [5, 7].

В индустрии также используют внешние аудит-проверки и многоступенчатое тестирование кандидатов на роль оценщиков [4, 7, 8].

Приведем формулу, которая описывает *модель Девида-Скина*:

$$P(C_i = k \mid L) = \frac{\pi_k \Pi_{j=1}^{J} \Pi_{l=1}^{L_j} \theta_{jkl}^{I(L_{jl}=k)}}{\Sigma_{k'=1}^{K} \pi_{k'} \Pi_{j=1}^{J} \Pi_{l=1}^{L_j} \theta_{jk'l}^{I(L_{jl}=k')}}.$$

5. Статистический мониторинг процесса разметки.

Помимо коэффициентов согласия анализируют долю совпадений, временные характеристики (слишком быстрое принятие решений как индикатор низкого качества), распределение оценок у каждого эксперта, долю шума в данных предпочтений и т.п. Такой мониторинг помогает выявлять аномалии и улучшать протокол [4, 6].

На практике методы применяются комплексно: единственного приёма, гарантирующего качество эксперта, не существует. Крупные организации и исследовательские коллективы внедряют несколько уровней контроля: от отбора и обучения группы оценщиков до постоянного мониторинга согласия и внешнего аудита [4, 7]. Важную роль играет качество инструкций для аннотаторов. Их пилотное тестирование, примеры для каждого критерия и чёткие правила повышают согласованность. Существуют практические руководства по разработке надёжных инструкций [9].

Внутриклассовый коэффициент корреляции определяется следующим образом:

$$ICC = \frac{MS_B - MS_W}{MS_B + (k-1)MS_W}.$$

Надо отметить, что есть и ограничения. В полностью субъективных задачах (например, оценка художественных ответов или юмора) даже идеальная калибровка не устраняет разброс мнений. В таких случаях привлекают целевые аудитории и усредняют оценки по большим выборкам. Другое ограничение – стоимость и скорость: многоуровневая проверка требует ресурсов. Поэтому исследуется частичная автоматизация оценки с помощью самих LLM. Однако полагаться только на автоматическую оценку рискованно, и гибридные подходы (автоматизированный отбор + финальная человеческая проверка) представляются более надёжными [1, 2, 4].

Сводная таблица подходов оценки качества ручных разметчиков представлена в таблице.

Следует отметить, что в индустрии вырабатываются собственные стандарты качества human evaluation, и опыт лидирующих компаний задаёт тон для всего сообщества. OpenAI в своих технических отчётах (например, по GPT-4) явно указывает метрики согласия оценщиков при оценке фактической корректности и вредоносности ответов модели [10], а также публикует инструкции для тех, кто проводит внешние аудит-проверки (ARC и др.). Anthropic делится уроками из собственных испытаний (сложности с внедрением стандартных бенчмарков вроде BIG-bench, необходимость тонкой настройки экспериментов под конкретную модель) [3], подчёркивая, что единый «универсальный» набор тестов не покрывает всех сторон поведения модели – необходим человеческий фактор для оценки открытых диалоговых сценариев. В open-source проектах (BigScience, OpenAssistant, LLaMA-2 Community evaluations) практикуется открытое привлечение множества пользователей к оценке моделей с последующим агре-

гированием. Это снижает риск субъективности одной группы, но требует ещё более тщательного математического усреднения и выявления аномалий среди большого числа оценщиков [5, 6]. Академическое сообщество со своей стороны усилило требования: публикации в топ-журналах требуют описания того, как именно проводилась человеческая оценка, сколько было аннотаторов, какова их квалификация, были ли контроль согласия и ослепление процесса. Всё это направлено на повышение прозрачности и доверия к выводам об эффективности моделей.

Таблица. ***Сводные данные подходов оценки качества разметчиков***

Метод оценки	Описание и цель	Примеры применения
Многократная разметка	Каждый пример оценивается несколькими экспертами; вычисляется степень согласия между ними (κ Коэна, α Криппендорфа и др.) для оценки надёжности и воспроизводимости	Оценка качества генерации текста ($k >= 0,6$) как условие достоверности [5]
«Золотой стандарт»	В задания аннотаторов встраиваются контрольные примеры с известным правильным ответом или очевидной оценкой. Проверяется, правильно ли эксперт их оценивает	Краудсорсинговые разметки данных, оценка токсичности и фактологии: вставка заведомо токсичных или верных ответов и контроль ошибок эксперта [6]
Слепое оценивание	Эксперт оценивает ответы, не зная их источника (модель/человек, версия модели и т.п.) и не зная, какие задания контрольные. Исключение информационных подсказок устраняет предвзятость и обеспечивает объективность сравнения	Оценка диалоговых моделей Anthropic в формате A/B-тестов: сравнение ответов двух анонимизированных моделей [8]
Мета-оценка	Дополнительные уровни проверки оценщиков: экспертные ревью оценок, повторная оценка тех же примеров самим аннотатором (self-consistency), коллективный разбор разногласий	Экспертный арбитраж расхождений в медицинской разметке LLM; сравнение оценок аннотаторов с «эталонным» мнением разработчиков OpenAI [10]
Статистические метрики надежности	Расчёт количественных показателей для оценщиков: коэффициенты согласия (κ, α, *ICC* и др.), доля совпадений, время на задание, распределение оценок. Помогает объективно измерить качество разметки, выявить систематические ошибки и документировать надёжность человеческого фактора	Отчёты об IAA во всех публикациях ACL/EMNLP по оценке генерации; внутренний анализ данных RLHF: ~20% предпочтений – шум, что обнаружено через статистику и повлекло фильтрацию данных [9]

Заключение

Проведенные исследования и накопленный в результате опыт свидетельствуют, что методы оценки качества экспертов не менее важны, чем методы оценки качества самой модели. Без них развитие и тонкая настройка больших языковых моделей были бы невозможны – мы просто не имели бы надёжной «линейки», чтобы измерять прогресс и сравнивать разные алгоритмы.

Верификация возможностей LLM с помощью человека-эксперта остаётся краеугольным камнем оценки. Человеческий разум пока превосходит алгоритмы во всесторонней и тонкой оценке генерируемого текста, особенно когда речь идёт о сложных смысловых или этических аспектах. Однако доверять человеческому мнению без оговорок уже недостаточно – сами эксперты требуют оценки и калибровки.

В работе кратко описан набор методов, призванных обеспечить высокое качество и надёжность экспертных оценок: многократная аннотация и измерение согласия, задания с «золотым стандартом», слепое оценивание, многоуровневая мета-оценка и сопутствующий статистический анализ. Применение этих подходов в комплексе становится стандартом – от академических статей до индустриальных отчётов. По мере роста масштабов и сложности систем вопросы объективности, культурных различий и практической значимости оценок остаются открытыми, но направление ясно: качество работы человека-оценщика требует не меньшего внимания, чем качество самой модели. Применение описанных методов – необходимое условие для построения справедливых, воспроизводимых и достоверных экспериментов с участием человека в эпоху крупных языковых моделей [1–4].

Тем не менее работа в этом направлении далека от завершения. Повышение масштаба и сложности моделей ставит новые вопросы.

Как сохранить объективность оценок при взаимодействии моделей с миллионами пользователей?

Как учитывать культурные и персональные различия в восприятии «качественного» ответа?

Как соотнести оценку человека с тем, что действительно важно для практического применения модели?

Эти вопросы возникают на стыке технологий и социальных наук. Одно ясно – человек остаётся в «центре петли» оценки искусственного интеллекта, и качество работы человека требует столь же внимательного отношения, как и качество работы самой модели. Развитие методов оценки экспертов способствует не только более точному измерению прогресса в NLP, но и формирует более ответственное и научно обоснованное отношение к созданию систем искусственного интеллекта, учитывающее человеческий фактор как часть системы. Поскольку LLM всё глубже проникают в различные сферы, от них требуют всё большей надёжности, а значит, и процессы их оценки должны становиться более строгими и валидными.

Список источников

1. *Chiang C.-H., Lee H.-Y.* Can Large Language Models Be an Alternative to Human Evaluations? Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023). Toronto, Canada, 2023. P. 15607–15631. DOI: 10.18653/v1/2023.acl-long.870. URL: [https://aclanthology.org/2023.acl-long.870/] (https://aclanthology.org/2023.acl-long.870/) (дата обращения: 01.10.2025).
2. *Gao M.* LLM-based NLG Evaluation: Current Status and Challenges. Computational Linguistics. 2025. V. 51. № 2. URL: [https://direct.mit.edu/coli/article/51/2/661/128807/] (https://direct.mit.edu/coli/article/51/2/661/128807/) (дата обращения: 15.09.2025).
3. *Tam T.Y.C., Chow T.Y.* et al. A framework for human evaluation of large language models in healthcare derived from literature review. npj Digital Medicine. 2024. URL: [https://www.nature.com/articles/s41746-024-01258-7] (https://www.nature.com/articles/s41746-024-01258-7) (дата обращения: 10.08.2025).
4. Anthropic. Challenges in evaluating AI systems. 2023. URL: [https://www.anthropic.com/research/evaluating-ai-systems] (https://www.anthropic.com/research/evaluating-ai-systems) (дата обращения: 03.10.2025).
5. *Liu S., Wang H., Ma Z., Li X.* How Humans Help LLMs: Assessing and Incentivizing Human Preference Annotators. arXiv:2502.06387. 2025. URL: [https://arxiv.org/abs/2502.06387] (https://arxiv.org/abs/2502.06387) (дата обращения: 10.10.2025).
6. *Guo Z.* et al. Evaluating Large Language Models: A Comprehensive Survey. arXiv:2310.19736. 2023. URL: [https://arxiv.org/abs/2310.19736] (https://arxiv.org/abs/2310.19736) (дата обращения: 12.10.2025).
7. *Ouyang L., Wu J., Jiang X.* et al. Training language models to follow instructions with human feedback // Advances in Neural Information Processing Systems (NeurIPS 2022). 2022. URL: [https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf] (https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf) (дата обращения: 09.06.2025).
8. *Yao S.* et al. HUGAGENT: Evaluating LLMs in Simulating Human Reasoning (версия на OpenReview). 2025. URL: https://openreview.net/pdf/be8cf8cfb9ad9e0e178f24ea8e52dda01a329389.pdf (дата обращения: 06.07.2025).
9. TR-Labs. How to Build Reliable Human Annotation Guidelines with LLMs (практическое руководство). 2023. URL: [https://medium.com/tr-labs-ml-engineering-blog/how-to-build-reliable-human-annotation-guidelines-with-llms-2cd8bbeff2a2] (https://medium.com/tr-labs-ml-engineering-blog/how-to-build-reliable-human-annotation-guidelines-with-llms-2cd8bbeff2a2) (дата обращения: 12.09.2025).
10. OpenAI. GPT-4 – исследовательские материалы. 2023. URL: [https://openai.com/index/gpt-4-research/] (https://openai.com/index/gpt-4-research/) (дата обращения: 10.10.2025).
11. *Олтян Н.Н.* Методы преобразования полуструктурированных данных в реляционные модели: классификация, применение и оценка пригодности для аналитики и машинного обучения // Мягкие измерения и вычисления. 2025. Т. 90. № 5. С. 48–67.

Информация об авторе

Даниил Тетеревенков – аспирант
SPIN-код: не представлен

Статья поступила в редакцию 16.10.2025
Одобрена после рецензирования 27.10.2025
Принята к публикации 30.10.2025

Original article

Methods for assessing the quality of experts in the verification of large language models

D. Teterevenkov[1]
1 Financial University under the Government of the Russian Federation (Moscow, Russia)
1 249453@edu.fa.ru

Abstract

Human evaluation is traditionally used as the "gold standard" for checking the quality of texts generated by large language models (LLMs). However, it is itself subject to subjectivity, variability and systematic errors of perception, which calls into question the reliability and reproducibility of model verification results.

To systematize and analyze modern approaches to assessing the quality of work of expert annotators involved in the verification of the results of large language models, in order to increase the objectivity and reliability of human evaluation.

The main sources of unreliability of expert judgments are analyzed. Methods for improving the reliability of human evaluation are presented and considered in detail: multiple annotation and calculation of agreement coefficients (Cohen's κ, Fleiss' κ, Krippendorff's α, ICC), the use of control ("gold") tasks, blind testing, meta-evaluation and statistical monitoring of evaluators' work. Special attention is paid to probabilistic models of annotator quality, such as the Dawid-Skene model.

The application of systematic calibration and verification of experts is a necessary condition for ensuring the objectivity and reproducibility of human-in-the-loop experiments in natural language processing research and evaluation of large language models. The considered methods make it possible to formalize the process of human evaluation, minimize subjective distortions and increase the reliability of the data obtained, which is critical for the correct comparison and development of LLMs.

Keywords

Large language models, human evaluation, expert verification, annotator agreement, Cohen's kappa, intraclass correlation coefficient, Dawid-Skene model, annotation reliability, meta-evaluation, reproducibility of experiments

For citation

Teterevenkov D. Methods for assessing the quality of experts in the verification of large language models. Neurocomputers. 2025. V. 27. № 6. P. 69–76. DOI: 10.18127/j19997493-202506-07 (in Russian).

References

1. *Chiang C.-H., Lee H.-Y.* Can Large Language Models Be an Alternative to Human Evaluations? Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023). Toronto, Canada, 2023. P. 15607--15631. DOI: 10.18653/v1/2023.acl-long.870.
2. *Gao M.* LLM-based NLG Evaluation: Current Status and Challenges. Computational Linguistics. 2025. V. 51. № 2.
3. *Tam T.Y.C., Chow T.Y.* et al. A framework for human evaluation of large language models in healthcare derived from literature review. npj Digital Medicine. 2024.
4. Anthropic. Challenges in evaluating AI systems. 2023.
5. *Liu S., Wang H., Ma Z., Li X.* How Humans Help LLMs: Assessing and Incentivizing Human Preference Annotators. arXiv:2502.06387. 2025.
6. *Guo Z.* et al. Evaluating Large Language Models: A Comprehensive Survey. arXiv:2310.19736. 2023.
7. *Ouyang L., Wu J., Jiang X.* et al. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems (NeurIPS 2022). 2022.
8. *Yao S.* et al. HUGAGENT: Evaluating LLMs in Simulating Human Reasoning. 2025.
9. TR-Labs. How to Build Reliable Human Annotation Guidelines with LLMs. 2023.
10. OpenAI. GPT-4 – research materials. 2023.
11. *Oltyan N.N.* Metody preobrazovaniya polustrukturirovannyh dannyh v relyacionnye modeli: klassifikaciya, primenenie i ocenka prigodnosti dlya analitiki i mashinnogo obucheniya. Myagkie izmereniya i vychisleniya. 2025. T. 90. № 5. S. 48--67. (in Russian).

Information about the author

Daniil Teterevenkov – Post-graduate Student

The article was submitted 16.10.2025
Approved after reviewing 27.10.2025
Accepted for publication 30.10.2025