Дедушкин Даниил Денисович, Магистрант,
ФГОБУ ВО «Поволжский государственный
университет телекоммуникаций и информатики»

Захарова Оксана Игоревна, к.т.н, доцент,
ФГОБУ ВО «Поволжский государственный
университет телекоммуникаций и информатики»

КРИТИКА СУЩЕСТВУЮЩИХ МЕТРИК ОЦЕНКИ ГЕНЕРАТИВНЫХ МОДЕЛЕЙ И ПРЕДЛОЖЕНИЕ НОВЫХ HUMAN-ALIGNED МЕТОДОВ

Аннотация. В данной статье аргументируется необходимость перехода к метрикам оценки генеративного искусственного интеллекта, ориентированным на человека (human-aligned metrics). Рассматриваются достоинства и недостатки традиционных метрик Inception Score (IS) и Fréchet Inception Distance (FID), а также метрик, ориентированных на человека.

Ключевые слова: Генеративный искусственный интеллект, оценка качества, human-aligned метрики, автоматические метрики, человеческие предпочтения, гибридные системы оценки.

Введение

За последние годы генеративные модели – достигли огромных успехов в создании реалистичных изображений, текстов, аудио и мультимодальных контентов. Эти системы всё чаще применяются в сферах, где взаимодействие с человеком становится центральным. Однако вместе с ростом возможностей обостряется фундаментальная проблема: как объективно и релевантно оценивать качество генераций?

Традиционно для этой цели используются автоматические метрики – такие как Inception Score (IS) и Frechet Inception Distance (FID) в компьютерном зрении или BLEU, ROUGE и BERTScore в обработке естественного языка. Несмотря на широкое распространение, эти метрики страдают от ряда системных недостатков. Во-первых, они часто измеряют лишь статистическое сходство с распределением реальных данных, игнорируя семантическую согласованность, контекстуальную уместность и субъективные аспекты восприятия. Во-вторых, многочисленные исследования показывают слабую корреляцию между значениями этих метрик и оценками, полученными от людей. В-третьих, такие подходы предполагают существование «правильного» ответа, что противоречит самой природе генеративных задач, где допустимо множество решений.

Этот разрыв между автоматизированной оценкой и человеческим восприятием получил название «human misalignment» – несоответствия между тем, как модель оценивается технически, и тем, как её результаты воспринимаются и используются человеком. В условиях, когда генеративные ИИ-системы всё чаще принимают решения, влияющие на пользовательский опыт, безопасность и даже этические нормы, такая несогласованность становится не просто методологической проблемой, а практическим риском.

В данной статье мы проводим критический анализ существующих метрик оценки генеративных моделей с акцентом на их неспособность отражать человеческие предпочтения, намерения и когнитивные особенности. На основе этого анализа мы формулируем принципы построения human-aligned метрик – оценочных систем, согласованных с восприятием, ценностями и целями реальных пользователей.

Чтобы понять масштаб проблемы и сформулировать обоснованные пути ее решения, важно сначала систематизировать существующие подходы, выделить их сильные и слабые стороны, а также проанализировать причины их несогласованности с человеческим восприятием. В следующем разделе мы кратко рассмотрим наиболее распространённые метрики, применяемые в задачах генерации изображений и обозначим ключевые аспекты, в которых они расходятся с целями human-aligned оценки.

1. Обзор существующих метрик оценки генеративных моделей

1.1 Inception Score

Inception Score (IS) – это метрика, предназначенная для оценки качества сгенерированных изображений. Она основывается на пред обученной модели Inception-v3 и анализирует чёткость и разнообразие генерируемых изображений. Оценка производится путём вычисления KL-дивергенции между условным распределением меток для каждого сгенерированного изображения и маргинальным распределением. Высокие значения Inception Score свидетельствуют о том, что сгенерированные изображения обладают высокой степенью резкости и разнообразия, а также сходны с реальными изображениями. IS предполагает, что высокое качество соответствует высокой достоверности классификатора и равномерному распределению по классам, но это игнорирует истинный реализм и часто поощряются модели, исключающие модальные значения, которые генерируют ограниченное количество выборок с высокой степенью достоверности. Это плохо согласуется с человеческими суждениями, поскольку не позволяет обнаружить артефакты или едва заметные недостатки, так как функции Inception-v3 устарели и ориентированы на классы ImageNet [1,2].

Этот показатель нестабилен из-за случайной дисперсии выборки и чувствителен к изменениям в наборе данных, что приводит к ненадежным рейтингам в разных моделях или областях. Он испытывает трудности с условной генерацией, выборками, не являющимися независимыми и одинаково распределенными, и многомодальными результатами, такими как модели диффузии, где разнообразие вводит в заблуждение без эталонных данных [3, 4].

IS оказывается уязвимым для враждебных воздействий или чрезмерной адаптации к исходному состоянию, что позволяет манипулировать выборками для завышения оценок без улучшения восприятия. Эти недостатки подчеркивают, почему IS следует использовать в паре с эталонными показателями, такими как FID, хотя даже они имеют схожие проблемы [5].

1.2 Fréchet Inception Distance

Fréchet Inception Distance (FID) оценивает генеративные модели, измеряя расстояние Вассерштейна-2 между распределениями признаков реальных и сгенерированных изображений, полученных с помощью Inception-v3. FID опирается на устаревшие характеристики Inception-v3, которые плохо отражают богатство современного генеративного контента, что приводит к ненадежным оценкам реализма и предвзятости в отношении шаблонов, обученных на ImageNet. Он объединяет реалистичность и разнообразие в единый показатель, не позволяя отдельно определять коллапс мод или чрезмерное разнообразие [6,7].

Метрика демонстрирует смещение выборки, когда конечные размеры выборки искажают ожидания в зависимости от модели, непредсказуемо завышая или занижая оценки. FID оказывается нестабильным на разных наборах данных, разрешениях или в областях, выходящих за рамки естественных изображений, и уязвим для враждебных манипуляций без улучшения восприятия [8, 9].

Метод FID требует эталонных данных, испытывает трудности с условной генерацией или неизобразительными модальностями и слабо коррелирует с предпочтениями человека в отношении тонких артефактов. Эти проблемы побуждают к поиску альтернатив, таких как варианты с точностью и полнотой или оценщики на основе LLM [2, 10].

IS оптимизирует уверенность классификатора и разнообразие предсказаний, но это не коррелирует с воспринимаемым реализмом или качеством для человека. FID минимизирует расстояние Вассерштейна между распределениями признаков, но этот метод измерения игнорирует перцептивные различия, важные для зрения человека: две изображения могут иметь близкие признаки Inception, но выглядеть совершенно по-разному глазам людей [6, 7].

Люди легко замечают артефакты синтеза (странные структуры, размытость, нарушения физики), но IS и FID часто не обнаруживают эти ошибки, если они не влияют на классификацию Inception или распределение признаков. Это приводит к парадоксу: модель может набрать высокий FID, но генерировать изображения с очевидными для человека дефектами [5].

Метрики не понимают семантического содержания и контекстуального соответствия между входом и выходом, особенно для условной генерации. Например, text-to-image модель может генерировать красивое изображение, но полностью не соответствующее описанию, однако FID может остаться высоким. Люди оценивают такие примеры как неудачу, в то время как метрики молчат [2].

Метрики не обучены на предпочтениях человека или не содержат информации о том, что люди на самом деле выбирают, когда сравнивают изображения. Они основаны на неявных предположениях (высокая классификационная уверенность = качество), которые часто неверны для современных генеративных моделей, особенно диффузионных, где разнообразие может указывать на худшие результаты вместо лучших [2].

2. Принципы построения human-aligned метрик

Human-aligned метрики (например, HPSv3, EvalAlign, G-Eval) обучаются на тысячах человеческих аннотаций и парах предпочтений, явно улавливая то, что люди считают качественным выводом. Они используют современные мультимодальные модели (VLM, MLLM) со способностью к рассуждению и объяснению ошибок. Эти подходы достигают корреляции с человеческими оценками на 20–30% выше, чем традиционные метрики [11, 12].

Продвинутые human-aligned подходы используют Chain-of-Thought (CoT) и Chain-of-Human-Preference (CHP) для итеративного построения обоснований оценок. Вместо просто выдачи числового результата, метрики генерируют объяснения критик (например, GREEN для радиологических отчетов), что делает оценку интерпретируемой и поддающейся проверке. Это отражает то, как люди оценивают качество – они думают о причинах, почему результат хороший или плохой [13].

Вместо сокращения качества до одного скаляра (как FID или IS), human-aligned метрики часто выдают многомерные оценки по разным аспектам. Video-Bench, например, оценивает видеогенерацию по нескольким измерениям (движение, реалистичность, соответствие подсказке), используя MLLM с chain-of-query для каждого аспекта, превосходя бинарные или одномерные суждения [14].

Платформы вроде GenAI Arena используют открытое краудсорсинговое голосование пользователей для сбора человеческих оценок моделей в полузашифрованном формате (Elo-ранжирование), избегая смещения исследователя и обеспечивая масштабируемость. Такие подходы основаны на принципе, что истинное «человеческое выравнивание» требует агрегирования предпочтений многих людей, а не экспертов [15].

Human-aligned метрики, обученные на разнообразных парах, более устойчивы к попыткам манипулирования (например, через adversarial perturbations), которые обманывают IS или FID. Использование MLLM с глубоким пониманием содержания затрудняет создание поддельных высококачественных примеров [5].

Лучшие human-aligned метрики не просто оценивают – они служат сигналом для обучения (reward models, alignment targets). Это позволяет генеративным моделям напрямую оптимизировать на человеческое предпочтение, а не на косвенные прокси [16].

Хотя предложенные human-aligned подходы к оценке генеративных моделей открывают перспективы для более релевантной и человекоцентрический оценки, они не лишены собственных вызовов и ограничений. Переход от чисто автоматических метрик к вовлечению человека в процесс оценки ставит перед исследователями новые методологические, практические и этические вопросы. В следующем разделе мы рассматриваем ключевые недостатки этих подходов.

3. Ограничения human-aligned метрик

Одна из главных преград – финансовая стоимость сбора человеческих аннотаций. HPSv3 обучена на 1+ миллионе пар предпочтений, что требует либо платного краудсорсинга (Amazon Mechanical Turk, Scale AI), либо внутренних аннотаторов компании. Каждая пара может стоить $0,10–$1,00 в зависимости от сложности, что означает $100K–$1M только на разметку данных. Это недоступно для большинства исследовательских групп и создает барьер входа, позволяя только крупным лабораториям (OpenAI, Google) разрабатывать продвинутые human-aligned метрики [11, 12].

По мере появления новых моделей и доменов генерации (video, 3D, audio), потребуется переразметка и переобучение метрик. EvalAlign для text-to-image требует специализированной разметки, а Video-Bench – своей собственной аннотированной базы. Невозможно создать единую метрику, которая масштабируется ко всем будущим применениям. Это приводит к фрагментации: множество специализированных метрик, каждая со своей стоимостью и графиком разработки [11, 14].

Сбор человеческой обратной связи – медленный процесс. К тому времени, как метрика будет обучена и валидирована, технология часто уже дальше: новые архитектуры, новые модальности, новые стили вывода. Это создает постоянную ситуацию «погони за целью», где метрики всегда отстают от границы современного знания. Человеческие предпочтения по своей природе субъективны. Два аннотатора могут оценить одно и то же изображение по-разному в зависимости от личных вкусов, культурного фона, состояния в момент разметки или даже времени суток. В отличие от математических метрик (которые детерминированы), human-aligned подходы наследуют эту вариативность. Попытка скрыть эту субъективность под видом «верной оценки» может быть опасна [13, 17].

Аннотаторы часто рекрутируются из специфичных регионов (США, Индия), говорящих на английском, с определенным культурным контекстом. Их предпочтения в эстетике, репрезентации и контенте отражают этот контекст, но преподносятся как универсальные. Например, краудсорсинговое голосование в GenAI Arena может отражать предпочтения активных интернет-пользователей из высокоразвитых стран, игнорируя остальной мир. Модель, обученная на таких данных, будет системно благоприятствовать визуальным стилям, ценностям и репрезентациям, близким к западным нормам [15, 16].

Существует риск, что аннотаторы будут смещены инструкциями, контекстом или даже платежными структурами. Если платить больше за более быстрые разметки, аннотаторы могут спешить и делать небрежные оценки. Если предоставлять подсказки (например, «это изображение создано состязательной сетью»), это может повлиять на восприятие. Качество human feedback часто не контролируется строго, и даже в крупных проектах возможны ошибки [5].

Человеческие предпочтения меняются со временем и под влиянием тренов. Метрика, обученная на аннотациях 2023 года, может дать неправильные ранжирования в 2025 году, когда вкусы сместились. Human-aligned метрики требуют постоянного переобучения, что снова возвращает к проблеме стоимости.

4. Возможные пути преодоления ограничений

Несмотря на выявленные недостатки human-aligned подходов, их потенциал в создании более надёжных и релевантных систем оценки генеративных моделей остаётся значительным.

Вместо отказа от этих методов, целесообразно сосредоточиться на практических стратегиях смягчения их ограничений. Крупные компании должны публиковать аннотированные датасеты и trained metrics как open source или через API (как начал делать OpenAI с некоторыми своими инструментами). Это снизит барьер входа для исследователей и ускорит развитие стандартов [11].

Вместо усреднения оценок, метрики должны явно моделировать распределение мнений и представлять результаты с неопределенностью. Вывод может быть: «70% аннотаторов предпочли вариант А, 30% – вариант В», вместо единого скалярного рейтинга. Это сохраняет информацию о разногласиях и позволяет пользователям выбрать подходящий им стиль [17].

Human-aligned метрики должны раскрывать:

- Демографию аннотаторов (страны, языки, возраст, опыт).
- Инструкции, которые им давались.
- Процент согласия между аннотаторами (inter-rater agreement).
- Анализ предвзятостей по демографическим характеристикам.

Это позволит исследователям и практикующим специалистам понять, кому может благоприятствовать метрика и кому она может быть несправедлива [16].

Комбинирование human-aligned метрик с автоматическими методами (например, калибровка FID по human feedback как в MetaMetrics) может предложить лучший компромисс между стоимостью и качеством. Использование smaller, специализированных human datasets для fine-tuning больших MLLM может также снизить требуемый объем разметки [18].

Заключение

Быстрое развитие генеративного искусственного интеллекта вывело на первый план фундаментальную проблему: существующие автоматические метрики оценки (такие как IS и FID) неадекватно измеряют качество с точки зрения человека. Они фокусируются на статистическом сходстве, игнорируя семантическую согласованность, контекстуальную уместность и субъективное восприятие, что создает риск «human misalignment» – разрыва между технической оценкой модели и ее реальной полезностью. В качестве ответа на этот вызов формируется новая парадигма human-aligned метрик. Эти подходы, обучаемые на масштабных данных человеческих предпочтений и использующие возможности современных мультимодальных языковых моделей, демонстрируют значительно более высокую корреляцию с человеческими суждениями. Их ключевые преимущества – способность давать интерпретируемые оценки, многомерный анализ качества и устойчивость к манипуляциям, что открывает путь для прямой оптимизации генеративных моделей на основе человеческих ценностей. Однако переход к human-aligned оценке сопряжен с серьезными трудностями: высокой стоимостью и трудоемкостью сбора аннотаций, субъективностью и культурной предвзятостью человеческих суждений, а также проблемой оперативного обновления метрик в условиях быстро меняющейся области. Эти ограничения не отменяют перспективности подхода, но требуют взвешенных стратегий. Таким образом, будущее оценки генеративных моделей лежит не в отказе от human-aligned принципов, а в их ответственной реализации. Наиболее эффективным путем представляется развитие гибридных систем, сочетающих глубину человеческой обратной связи с масштабируемостью автоматических методов, при строгом соблюдении требований к прозрачности, документированию предвзятостей и учету множественности мнений. Только такая сбалансированная и рефлексивная система оценки сможет обеспечить безопасное и ценностно-ориентированное развитие генеративного ИИ, действительно соответствующего потребностям человека.

Список литературы:

1. Barratt, S. A Note on the Inception Score [Электронный ресурс] / S. Barratt, R. Sharma – Электрон. текстовые дан. – Режим доступа: https://arxiv.org/abs/1801.01973 свободный. – Загл. с экрана. – Яз. англ.

2. Stein, G. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models [Текст]: materials conf. / G. Stein, J. C. Cresswell, R. Hosseinzadeh [et al.] // Advances in Neural Information Processing Systems 36 (NeurIPS 2023). — New Orleans, 2023. — P. 1–25

3. Ravuri, S. Classification Accuracy Score for Conditional Generative Models [Текст]: материалы конф. / S. Ravuri, O. Vinyals // Advances in Neural Information Processing Systems 32 (NeurIPS 2019). — Vancouver, 2019. — P. 1–12. — Режим доступа: https://proceedings.neurips.cc/paper_files/paper/2019/file/fcf55a303b71b84d326fb1d06e332a26-Paper.pdf свободный. — Загл. С экрана. — Яз. Англ.

4. Benny, Y. Evaluation Metrics for Conditional Image Generation [Текст] / Y. Benny, T. Galanti, S. Benaim, L. Wolf // International Journal of Computer Vision. — 2021. — Vol. 129. — P. 1712–1731. — Ил., табл. — Текст: непосредственный.

5. Alfarra, M. On the Robustness of Quality Measures for GANs [Текст] : материалы конф. / M. Alfarra, J. C. Pérez, A. Frühstück [et al.] // 17th European Conference on Computer Vision (ECCV 2022). — Tel Aviv, 2022. — P. 1–29. — Ил., табл. — Текст: непосредственный.

6. Jayasumana, S. Rethinking FID: Towards a Better Evaluation Metric for Image Generation [Электронный ресурс] / S. Jayasumana, S. Ramalingam, [et al.] — Электрон. Текстовые дан. — Режим доступа: https://arxiv.org/abs/2401.09603v2свободный. — Загл. с экрана. — Яз. Англ.

7. Naeem, M. F. Reliable Fidelity and Diversity Metrics for Generative Models [Текст]: материалы конф. / M. F. Naeem, S. J. Oh, Y. Uh, Y. Choi, J. Yoo // International Conference on Machine Learning (ICML 2020). — Vienna, 2020. — P. 7176 – 7185

8. Обухов, А. Д. Модифицированный метод оценки качества генеративно-состязательных нейронных сетей [Текст] / А. Д. Обухов // Вестник ВГУ. Серия: Системный анализ и информационные технологии. — 2020. — № 3. — С. 97–107. — Ил., табл. — Текст: непосредственный.

9. Chong M. J., Forsyth D. Effectively Unbiased FID and Inception Score and where to find them [Текст]: материалы конф.// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). – 2020. – P. 6069-6078.

10. Thanh-Tung, H. Toward a Generalization Metric for Deep Generative Models [Электронный ресурс] / H. Thanh-Tung, T. Tran — Электрон. Текстовые дан. — Режим доступа: https://arxiv.org/abs/2011.00754 свободный. — Загл. с экрана. — Яз. Англ.

11. Tan, Z. EVALALIGN: Supervised Fine-Tuning Multimodal LLMs with Human-Aligned Data for Evaluating Text-to-Image Models [Электронный ресурс] / Z. Tan, X. Yang, [et al.] — Электрон. Текстовые дан. — Режим доступа: https://arxiv.org/abs/2406.16562 свободный. — Загл. с экрана. — Яз. Англ.

12. Ma, Y. HPSv3: Towards Wide-Spectrum Human Preference Score [Текст]: материалы конф. / Y. Ma [et al.] // IEEE/CVF International Conference on Computer Vision (ICCV 2025). — Rio de Janeiro, 2025. — P. 1–29. — Ил., табл. — Текст: непосредственный.

13. Sudjianto, A. Human-Calibrated Automated Testing and Validation of Generative Language Models: An Overview [Электронный ресурс] / A. Sudjianto, S. Neppalli. — Электрон. Текстовые дан. — Режим доступа: https://arxiv.org/abs/2411.16391, свободный. — Загл. С экрана. — Яз. Англ.

14. Han, H. Video-Bench: Human-Aligned Video Generation Benchmark [Текст]: материалы конф. / H. Han [et al.] // IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2025). — Seattle, 2025. — P. 1–15. — Ил., табл. — Текст: непосредственный.

15. Jiang, D. GenAI Arena: An Open Evaluation Platform for Generative Models [Текст]: материалы конф. / D. Jiang [et al.] // 38th Conference on Neural Information Processing Systems (NeurIPS 2024). — Vancouver, 2024. — P. 1–10. — Ил., табл. — Текст: непосредственный.

16. Chouldechova, A. A Shared Standard for Valid Measurement of Generative AI Systems' Capabilities, Risks, and Impacts [Электронный ресурс] / A. Chouldechova [et al.]. — Электрон. Текстовые дан. — Режим доступа: https://arxiv.org/abs/2412.01934, свободный. — Загл. С экрана. — Яз. Англ.

17. Park, D. Probabilistic Precision and Recall Towards Reliable Evaluation of Generative Models [Текст]: материалы конф. / D. Park, S. Kim // 11th International Conference on Learning Representations (ICLR 2023). — Paris, 2023. — P. 1–9. — Ил., табл. — Текст: непосредственный.

18. Winata, G. I. MetaMetrics: Calibrating Metrics For Generation Tasks Using Human Preferences [Текст]: материалы конф. / G. I. Winata [et al.] // 13th International Conference on Learning Representations (ICLR 2025). — Singapore, 2025. — P. 1–10. — Ил., табл. — Текст: непосредственный.