

Статья поступила в редакцию: 14.09.2026

Статья принята к публикации: 20.09.2026

Статья опубликована: 28.09.2026

УДК 004.8; 519.816

DOI: 10.32743/UniTech.2026.150.9.23429

Модульный калибровочный ИИ-рецензент для ранжирования научных статей (Flash-статья)

Кравцов Геннадий Григорьевич
директор

НИЦ Прикладная статистика
Российская Федерация, г. Рязань
E-mail: 62abc@mail.ru

A Modular Calibrating AI Reviewer for Ranking Scientific Papers (Flash Paper)

Kravtsov Gennady Grigorievich
Director of the Research Center Applied Statistics
Russian Federation, Ryazan

Аннотация

Представлена модульная архитектура калибровочного ИИ-рецензента, состоящая из шести независимых модулей, последовательно реализующих определение режима рецензирования (STEM/Humanities), оценку научной новизны, методологическую строгость, практическую ценность, качество визуализации и воспроизводимость. ИИ-рецензент предлагается как инструмент первичной экспресс-фильтрации рукописей, поступающих в редакцию. Валидация архитектуры (DeepSeek-V3, 10 прогонов на 7 статьях) с урезанным средним подтвердила снижение вариативности LLM: коэффициент вариации (CV) сильных работ составил 2–3 %. В ходе инвертированного теста Тьюринга ИИ распознал сгенерированный абсурдный текст, присвоив ему минимальный балл (21.62 из 100) при диагностически высоком CV (28.2 %). Дополнительно в рамках работы апробирован формат Flash-статьи: данная статья-якорь обеспечивает видимость в каталогах и цитируемость, полный текст статьи с дополнительными материалами (логи верификации, код) размещён в репозиториях (SSRN, Zenodo с DOI) и на GitHub. Формат сокращает цикл публикации с 6–12 месяцев до 2–4 недель и предназначен для оперативного представления результатов в журналах с ускоренным рецензированием и печатью, что особенно востребовано в быстроразвивающихся научных областях и прикладных исследованиях.

Библиографическое описание: Кравцов Г.Г. Модульный калибровочный ИИ-рецензент для ранжирования научных статей (Flash-статья) // Universum: технические науки : электрон. научн. журн. 2026. 9(150). URL: <https://7universum.com/ru/tech/archive/item/23429>

Abstract

A modular architecture of a calibrating AI reviewer is presented, consisting of six independent modules that sequentially implement the determination of the review mode (STEM/Humanities), assessment of scientific novelty, methodological rigor, practical value, visualization quality, and reproducibility. The AI reviewer is proposed as a tool for primary express filtering of manuscripts submitted to an editorial office. Validation of the architecture (DeepSeek-V3, 10 runs on 7 papers) with a trimmed mean confirmed a reduction in LLM variability: the coefficient of variation (CV) for strong papers was 2–3 %. In an inverted Turing test, the AI recognized the generated absurd text, assigning it a minimum score (21.62 out of 100) with a diagnostically high CV (28.2 %). Additionally, the Flash Paper format was tested within the framework of this work: the anchor paper ensures visibility in catalogs and citability, while the full text of the article with supplementary materials (verification logs, code) is deposited in repositories (SSRN, Zenodo with DOI) and on GitHub. The format reduces the publication cycle from 6–12 months to 2–4 weeks and is intended for the prompt presentation of results in journals with accelerated review and printing, which is especially in demand in fast-moving scientific fields and applied research.

Ключевые слова: калибровочный ИИ-рецензент, модульная архитектура, Flash-статья, оперативная публикация, тест Тьюринга.

Keywords: calibrating AI reviewer, modular architecture, Flash Paper, rapid publication, Turing test.

Введение

Актуальность. Экспоненциальный рост научных публикаций [9], и перегрузка экспертов делают необходимыми инструменты автоматизированной предварительной оценки рукописей. Существующие ИИ-рецензенты страдают контекстной зависимостью [3, 4], стохастической нестабильностью [1], они оценивают работу в рамках имеющейся информации [11], тогда как новаторские идеи могут выходить за эти пределы.

Проблема усугубляется издательским лагом: традиционный цикл рецензирования занимает 6–12 месяцев, за которые исследование теряет актуальность. По этой причине в настоящей работе впервые использован формат, названный Flash Paper: препринт представляет полные материалы [7], Github (<https://github.com/Famimot/calibrating-ai-reviewer-modular>) обеспечивает воспроизводимость результатов, а данная якорная статья с коротким сроком публикации обеспечивает лаконичное и своевременное представление информации.

Простая генерация оценок работы ИИ-рецензента без раскрытия исходных материалов не имеет доказательной силы, а одиночный прогон LLM некорректен в силу стохастической природы модели — необходимы многократные измерения с фильтрацией выбросов.

В прошлой работе [6] предложен монолитный ИИ-рецензент на базе конкурентного ранжирования в замкнутом пакете. Цель работы — спроектировать и валидировать архитектуру из 6 специализированных модулей, отвечающих за отдельный критерий статьи [9, 7]. Валидация проведена на контрастном массиве из 7 статей: 6 реальных публикаций автора с открытым DOI (исключает конфликт интересов) и 1 сгенерированной LLM абсурдной работы (DOI: 10.5281/zenodo.20703374) для инвертированного теста Тьюринга [2].

Это гарантирует полную прозрачность и воспроизводимость, в отличие от исследований на закрытых массивах. Для каждой статьи выполнено 10 прогонов DeepSeek-V3 с расчетом урезанного среднего для фильтрации выбросов.

Бонусный эксперимент проверил способность ИИ распознавать семантический абсурд в формально корректном тексте.

Методология

Принцип калибровки. Рецензент оценивает не отдельную статью по абсолютной шкале, а весь массив одновременно: статьи выступают шкалой друг для друга. Сильная работа получает высокий балл потому, что превосходит остальные, слабая — потому что уступает. Это снимает проблему отсутствия у LLM абсолютной шкалы. Урезанное среднее по 10 прогонам фиксирует устойчивую калибровку, отбрасывая стохастические отклонения.

Архитектура. Выделены фундаментальные сущности оценки научной рукописи, определены их числовые шкалы [5, 8]. Архитектура включает шесть модулей (табл. 1), выполняющих определённую функцию. Это обеспечивает поэтапный контроль корректности, локализацию ошибок и возможность модификации отдельных модулей без переписывания всей системы. Полное описание логики, чек-листов и промптов — в открытом репозитории [7].

Таблица 1.

Модульная структура ИИ-рецензента

Модуль	Название	Функция	Баллы
1	Режим-детектор	Определение STEM/Humanities	0
2	Новизна-анализатор	Оценка новизны, происхождения метода	30
3	Методология-анализатор	Оценка методологической строгости	25
4	Ценность-анализатор	Оценка практической ценности	20
5	Визуализация, воспроизводимость	Оценка визуализации и воспроизводимости	15 + 10
6	Агрегатор-категоризатор	Категоризация, итоговая таблица	100

Процедура рецензирования. Для каждой из $R = 10$ итераций в «чистом» диалоге с LLM (DeepSeek-V3) последовательно выполняются модули 1–6. Каждый последующий модуль использует результаты предыдущих. Фиксируется итоговая таблица с баллами (0–100) для всех статей. После завершения 10 прогонов исключаются два минимальных и два максимальных значения для каждой статьи, вычисляется урезанное среднее [8]:

$$\bar{S}_i = \frac{1}{R-4} \sum_{r \in R_{trim}} S_i^{(r)}$$

R_{trim} — множество прогонов, исключающих два минимальных и два максимальных значения $S_i^{(r)}$.

R_{trim} — множество прогонов, исключающих два минимальных и два максимальных значения .

Финальный балл рассчитывается как сумма оценок по пяти критериям.

Верификация. Надёжность валидации обеспечивается не количеством статей, а кратным повторением прогонов ($R = 10$) с фильтрацией выбросов.

Результаты

Ранжирование. После фильтрации получены итоговые баллы для каждой статьи (табл. 2).

Таблица 2.

Результаты 10 прогонов (отфильтрованные данные)

Статья	Новизна (30)	Методология (25)	Практ. Ценность (20)	Визуализация (15)	Воспроизв. (10)	Итог (100)
2	27.83	24.17	17.33	13.67	8.67	91.67
7	27.33	24.17	15.17	12.67	9.5	88.83
5	28.17	24	13.5	12.33	8.83	86.83
3	25	23.83	11.5	12.5	5.83	78.67
1	25	24	11	12.67	5.67	78.33
4	17.5	16.67	6	10.5	5.5	56.17
6	3.8	5.67	2.4	7.5	2.25	21.62

ИИ-рецензент продемонстрировал устойчивую дифференциацию статей. Разрыв между лучшей работой (91.67) и сгенерированной статьёй (21.62) составил 70 баллов. Распределение итоговых баллов до и после фильтрации представлено на рис. 1 (полные данные всех визуализаций — в препринте [7]).

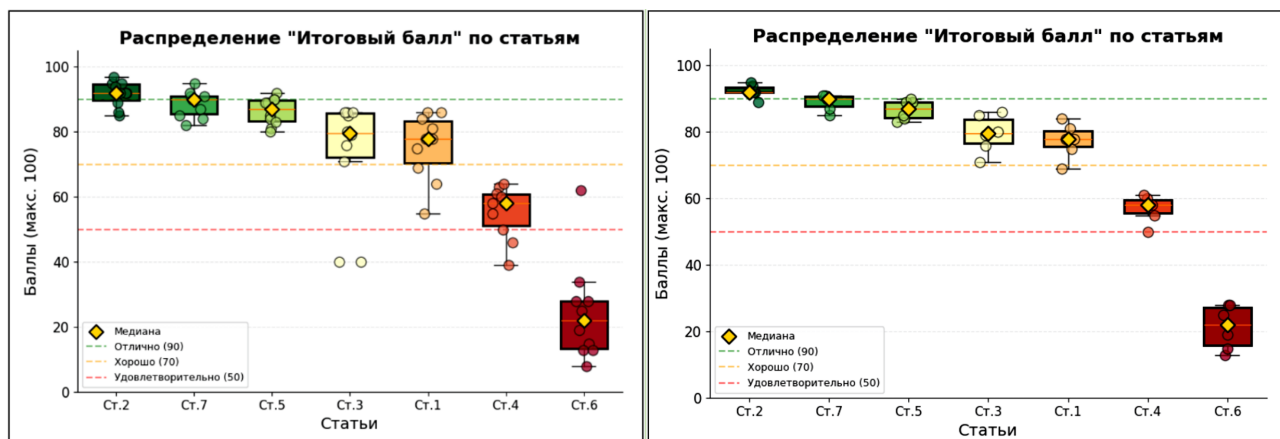


Рисунок 1. Распределение итогового балла: сырые данные (слева) и после фильтрации (справа)

Коэффициент вариации (CV) для сильных статей составляет 2–3 %, для статей средней группы — 6–6.5 %. Для сгенерированной статьи $CV = 28.2$ % (табл. 3), что более чем в 10 раз превышает CV сильных статей.

Таблица 3.

Вариативность результатов

Paper	Mean	Median	Min	Max	Range	Std	CV (%)
2	92.33	92	89	95	6	1.89	2.04
7	89	90	85	91	6	2.24	2.51
5	86.67	87	83	90	7	2.75	3.17
3	79.5	79.5	71	86	15	5.12	6.44
1	77.5	78	69	84	15	4.72	6.09
4	57	58	50	61	11	3.65	6.41
6	21.33	22	13	28	15	6.02	28.21

Воспроизводимость. Все материалы размещены в открытом доступе: дополнительные материалы (скрипт, верификация) [7],

модульные промпты (рус./англ.), — <https://github.com/Famimot/calibrating-ai-reviewer-modular>,

текст сгенерированной ИИ статьи — <https://doi.org/10.5281/zenodo.20703374>,

определение Flash-статьи — <https://doi.org/10.5281/zenodo.22727742>.

Заключение

Разработанная модульная архитектура подтверждает свою эффективность: формализация критериев в виде модульного конвейера обеспечивает поэтапную локализацию ошибок и возможность модификации отдельных модулей.

Ключевые выводы

1. Калибровочный режим. Оценка отдельной рукописи нестабильна из-за стохастической природы LLM. Предпочтителен режим сравнения, где статьи сами выступают калибровочным материалом.

2. Многократность прогонов. Необходимо выполнять несколько независимых прогонов с фильтрацией экстремумов.

3. Дискриминативная способность. Инструмент устойчиво дифференцирует статьи.

4. Практическая ценность. Инструмент предназначен для первичной фильтрации рукописей и снижения нагрузки на экспертов [3], ранжирования журналов [10], а также самооценки.

Представленный подход демонстрирует полную открытость, воспроизводимость и практическую применимость в отличие от закрытых систем, не предоставляющих исходных данных для верификации. В то же время использованный формат Flash-статьи позволяет опубликовать работу через несколько недель, а не 6–18 месяцев, без потери актуальности и практической ценности.

Список литературы

1. Bai J. iRULER: Intelligible Rubric-Based User-Defined LLM Evaluation for Revision / J. Bai [et al.] // Proceedings of the. – 2026. – DOI: 10.1145/3772318.3790539.

2. Bhatnagar Y. Breaking the Turing Test: Testing the relevance of the Turing Test against modern LLMs / Y. Bhatnagar // International Journal for Research in Engineering Application & Management. – 2026. – Т. 11, № 10.
3. Daskaliuk B. Artificial Intelligence in Peer Review: Enhancing Efficiency While Preserving Integrity / B. Daskaliuk [et al.] // Journal of Korean Medical Science. – 2025. – Т. 40, № 7. – DOI: 10.3346/jkms.2025.40.e92.
4. Goyal P. ScholarPeer: A Context-Aware Multi-Agent Framework for Automated Peer Review / P. Goyal [et al.] // arXiv. – 2026. – DOI: 10.48550/arXiv.2601.22638.
5. Institute of Hydrogen Economy. Regulations on the evaluation of scientific works on a 100-point scale based on 10 key criteria [Электронный ресурс]. // International Scientific Journal for Alternative Energy and Ecology (ISJAE). – 2026. (дата обращения: 13.09.2026).
6. Kravtsov G.G. Two-level open-source architecture for independent AI ranking of scientific journals / G.G. Kravtsov // OSF Preprints. – 2026. – DOI: 10.31235/osf.io/vpbk6_v1.
7. Kravtsov G.G. Калибровочный ИИ-рецензент: модульная архитектура для оценки и ранжирования научных статей / Г.Г. Кравцов. – Research Center Applied Statistics. – 2026. – DOI: 10.5281/zenodo.20732844.
8. Menke J. Establishing institutional scores with the Rigor and Transparency Index: Large-scale analysis of scientific reporting quality / J. Menke [et al.] // Journal of Medical Internet Research. – 2022. – Т. 24, № 6. – DOI: 10.2196/37324.
9. Pugliese R. Agentic publications: redesigning scientific publishing in the age of thinking large language models / R. Pugliese [et al.] // Journal of Documentation. – 2026. – Т. 82, № 7. – P. 125–149. – DOI: 10.1108/JD-07-2025-0207.
10. Thelwall M. ChatGPT ranking of business and management journals with article quality scores / M. Thelwall // Aslib Journal of Information Management. – 2025. – DOI: 10.1108/AJIM-11-2024-0927.
11. Zhou Y. Can AI Evaluate Literature at the Level of Experts? A Comparative Analysis of ChatGPT-4o, DeepSeek R1, and Human Ratings Based on Entropy / Y. Zhou, H. Hu // Frontiers in Research Metrics and Analytics. – 2025. – Т. 10. – Art. 1684137. – DOI: 10.3389/frma.2025.1684137.

References

1. Bai J., Cheong W.S., Muller P., Lim B.Y. iRULER: Intelligible Rubric-Based User-Defined LLM Evaluation for Revision // Proceedings of the. – 2026. – DOI: 10.1145/3772318.3790539.
2. Bhatnagar Y. Breaking the Turing Test: Testing the relevance of the Turing Test against modern LLMs // International Journal for Research in Engineering Application & Management. – 2026. – Vol. 11, No. 10.
3. Daskaliuk B., Zimba O., Yessirkepov M. et al. Artificial Intelligence in Peer Review: Enhancing Efficiency While Preserving Integrity // Journal of Korean Medical Science. – 2025. – Vol. 40, No. 7. – DOI: 10.3346/jkms.2025.40.e92.

4. Goyal P., Parmar M., Song Y., Palangi H., Pfister T., Yoon J. ScholarPeer: A Context-Aware Multi-Agent Framework for Automated Peer Review // arXiv. – 2026. – DOI: 10.48550/arXiv.2601.22638.
5. Institute of Hydrogen Economy. Regulations on the evaluation of scientific works on a 100-point scale based on 10 key criteria // International Scientific Journal for Alternative Energy and Ecology (ISJAEE). – 2024. – URL: <https://www.isjaee.com/jour/announcement/view/446> (дата обращения: 01.04.2026).
6. Kravtsov G.G. Two-level open-source architecture for independent AI ranking of scientific journals // OSF Preprints. – 2026. – DOI: 10.31235/osf.io/vpbk6_v1.
7. Kravtsov G. Calibrating AI Reviewer: Modular architecture for evaluating and ranking scientific articles. Research Center Applied Statistics. – 2026. – DOI: 10.5281/zenodo.20732844.
8. Menke J., Eckmann P., Ozyurt I.B., Roelandse M., Anderson N., Grethe J., Gamst A., Bandrowski A. Establishing institutional scores with the Rigor and Transparency Index: Large-scale analysis of scientific reporting quality // Journal of Medical Internet Research. – 2022. – Vol. 24, No. 6. – DOI: 10.2196/37324.
9. Pugliese R., Kourousias G., Venier F., Garlatti Costa G. Agentic publications: redesigning scientific publishing in the age of thinking large language models // Journal of Documentation. – 2026. – Vol. 82, No. 7. – DOI: 10.1108/JD-07-2025-0207.
10. Thelwall M. ChatGPT ranking of business and management journals with article quality scores // Aslib Journal of Information Management. – 2025. – DOI: 10.1108/AJIM-11-2024-0927.
11. Zhou Y., Hu H. Can AI Evaluate Literature at the Level of Experts? A Comparative Analysis of ChatGPT-4o, DeepSeek R1, and Human Ratings Based on Entropy // Frontiers in Research Metrics and Analytics. – 2025. – Vol. 10. – Art. 1684137. – DOI: 10.3389/frma.2025.1684137.