Большие языковые модели как инструмент автоматической оценки качества генеративных систем: анализ подходов и ограничений

Зуев Д.А.

ФГБОУ ВО "МГУ имени М.В. Ломоносова",

г. Москва, Российская Федерация

В современном мире всё стремительнее получают распространение большие языковые модели. В связи с чем растёт потребность в инструментах оценки качества их ответов. Классические метрики, опирающиеся на эталонные тексты, плохо подходят для открытой генерации, допускающей множество корректных формулировок. Человеческая экспертная оценка остаётся с одной стороны наиболее надёжной, а с другой, требующей значительных временных, финансовых затрат и является проблемной для воспроизводимости результатов. Поэтому всё большую популярность получает LLM-as-a-Judge подход, в котором одна языковая модель оценивает результаты другой [1]. Данная стратегия рассматривается как компромисс между строгостью экспертной оценки и масштабируемостью автоматических метрик.

Сущность подхода в том, что сильная модель-судья получает инструкцию пользователя и один (или несколько) ответов, после чего выносит своё решение об их качестве. Чаще всего применяют три режима: попарное сравнение, где судья выбирает лучший из двух представленных ответов; прямую оценку по числовой шкале или рубрике; и оценку относительно эталонного ответа. Попарное сравнение считается лучшим, так как снижает когнитивную нагрузку на модель и разброс оценок. Именно на нём основаны популярные бенчмарки, где модели ранжируются по рейтингу Эло. В ряде задач оценки судей хорошо согласуются с человеческими предпочтениями, что подтверждает практическую применимость подхода.

К достоинствам парадигмы относятся масштабируемость и относительная дешевизна. Один запуск модели-судьи обрабатывает тысячи сравнений за минуты, тогда как сопоставимый объём человеческой экспертизы потребовал бы

несколько недель. Судья не устаёт и не зависит от настроения, что потенциально даёт более стабильные результаты. Но при этом автоматизированная оценка не идеальна и имеет систематические искажения, способные существенно снижать её ценность. Исследования показывают, что стандартные протоколы со шкалой Лайкерта демонстрируют низкую согласованность между разными моделями-судьями [1].

Среди наиболее изученных проблем можно выделить позиционную предвзятость, модель отдаёт предпочтение ответу, стоящему первым в промпте. Систематическое исследование, включающее в себя более 150 тысяч случаев с участием пятнадцати моделей-судей, подтвердило, что позиционная предвзятость не случайна и существенно зависит от судьи и задачи [2]. Длинная предвзятость имеет склонность оценивать более пространные ответы как более качественные, даже если они содержат избыточные фрагменты. Особого внимания заслуживает самопредпочтение, при котором судья лучше оценивает ответы той же модели или моделей того же семейства. Приведенные эффекты устойчивы и подтверждаются на разных наборах данных.

Для русского языка ситуация усугубляется разрывом в качестве между русскоязычными и англоязычными судьями. Исследование REPA (тысяча пользовательских запросов, две тысячи ответов) выявило, что на русском языке модели-судьи согласуются с человеческими оценками заметно хуже, чем на английском [3]. Похожие выводы получены в работе о мультиязычной надёжности судей, из которых видно, что средняя согласованность по коэффициенту Каппы оказалась около 0,3, а для языков с ограниченными ресурсам ещё ниже [4]. Из чего можно сделать вывод, что перенос методологии LLM-as-a-Judge на русскоязычные задачи требует дополнительной проверки и донастройки промптов и критериев.

Для повышения надёжности оценочных суждений нужно учитывать несколько важных сторон. Во-первых, необходим переход от субъективных шкал к декомпозиции критериев на бинарные вопросы, что повысит согласованность судей и интерпретируемость оценок [1]. Во-вторых, требуется

учёт неопределённости задачи оценивания. Для этого предложена методология валидации судей: критерии можно трактовать по-разному (в допустимых пределах), а конфигурации судейских систем - сравнивать и выбирать лучшие. [5]. В-третьих, это использование ансамблей судей, когда итоговая оценка складывается из мнений нескольких моделей, что в ряде случаев улучшает стабильность оценок [4].

На основе изложенного можно сделать вывод, что LLM-as-a-Judge многообещающий, но пока ещё не идеальный инструмент. Данный подход позволяет заметно ускорить и удешевить оценку генеративных систем, однако требует осторожного применения, особенно в случае русскоязычных данных. Оптимальна стратегия, при которой автоматические судьи выполняют предварительную фильтрацию и ранжирование, а в критических случаях финальное решение принимает человек. Дальнейшая работа должна быть направлена на калибровку судей и снижение их предвзятости, а также на создание репрезентативных русскоязычных датасетов для их валидации.

Список использованной литературы

1. Lee Y., Kim J., Kim J., Cho H., Kang J., Kang P., Kim N. CheckEval: *A reliable LLM-as-a-Judge framework for evaluating text generation using checklists* // Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. 2025. P. 15771-15798. URL: https://aclanthology.org/2025.emnlp-main.796/ (дата обращения: 28.09.2026).

2. Shi L., Ma C., Liang W., Diao X., Ma W., Vosoughi S. Judging the Judges: *A Systematic Study of Position Bias in LLM-as-a-Judge* // Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics. 2025. P. 292-314. URL: https://aclanthology.org/2025.ijcnlp-long.18/ (дата обращения: 28.09.2026).

3. Pugachev A., Fenogenova A., Mikhailov V., Artemova E. REPA: *Russian Error Types Annotation for Evaluating Text Generation and Judgment Capabilities* // Proceedings of the 10th Workshop on Slavic Natural Language Processing. 2025. P.

136-150. URL: https://aclanthology.org/2025.bsnlp-1.16/ (дата обращения: 28.09.2026).

4. Fu X., Liu W. *How Reliable is Multilingual LLM-as-a-Judge?* // arXiv preprint. 2025. URL: http://arxiv.org/abs/2505.12201 (дата обращения: 28.09.2026).

5. Guerdan L., Barocas S., Holstein K., Wallach H.M., Wu S.Z., *Chouldechova A. Validating LLM-as-a-Judge Systems under Rating Indeterminacy* // Advances in Neural Information Processing Systems. 2025. URL: https://proceedings.neurips.cc/paper_files/paper/2025/hash/a309239c11a28c597d050bd4a1752d32-Abstract-Conference.html (дата обращения: 28.09.2026).