УДК 004.8

ИССЛЕДОВАНИЕ МЕТОДОВ ОЦЕНКИ LLM В ЗАДАЧАХ КОРПОРАТИВНОЙ ОБРАТНОЙ СВЯЗИ

Бабушкин К.А. (ИТМО), **Тимофеев Н.А.** (ИТМО)
Научный руководитель – Филянин И.В. (ИТМО)

Введение

Интеграция больших языковых моделей (LLM) в корпоративные HR-процессы требует строгой валидации их качества в специфическом контексте обратной связи о сотрудниках [1]. Существующие бенчмарки не отражают особенности HR-задач, где критически важны понимание тонких нюансов межличностной коммуникации [2], мотивационный тон и соответствие организационной культуре. Данное исследование направлено на комплексную оценку современных LLM через методологию слепого тестирования с применением метрик межэкспертного согласия и анализа систематических смещений [3].

Основная часть

Для эксперимента отобраны четыре современные LLM-модели, представляющие различные архитектурные подходы: Gemini 2.5 Pro (флагманская мультимодальная модель для сложной аналитики), Gemini 2.5 Flash (облегчённая версия для быстрого отклика), DeepSeek-R1-0528 (модель с MoE-архитектурой и акцентом на reasoning) и Qwen3-235B-A22B-2507 (крупнейшая открытая мультиязычная модель). Корпус экспериментальных задач включает 40 сценариев корпоративной обратной связи, сгруппированных по категориям: peer-to-peer feedback, manager-to-employee feedback, upward feedback и self-assessment [4].

Применена методология двойного слепого тестирования, где ни человеческие оценщики, ни LLM-судьи не знают авторства оцениваемых ответов. Для каждой задачи четыре модели генерируют ответы на идентичных промптах [5], после чего ответы анонимизируются путём присвоения случайных идентификаторов, варьирующихся между задачами и оценщиками. Оценка проводится по четырём критериям: Fluency (грамматическая корректность), Motivational tone (способность мотивировать), Sentiment match (соответствие тональности) и Final choice (интегральная оценка применимости).

Для количественной оценки согласованности применён комплекс из четырёх метрик Inter-Rater Agreement: Krippendorff's Alpha для общей оценки согласия, Spearman Correlation для измерения монотонной связи, Exact Agreement для процента полного совпадения и Mean Rank Distance для среднего позиционного расстояния [6]. Анализ проводился отдельно для трёх групп: human-human (согласованность между людьми), LLM-LLM (согласованность между моделями-судьями) и human-LLM cross (кросс-согласованность) [7].

Результаты показали, что LLM-модели демонстрируют существенно более высокую внутригрупповую согласованность (Krippendorff's Alpha = 0.71) по сравнению с человеческими оценщиками ($\alpha = 0.52$), что отражает большую консистентность применяемых критериев и отсутствие усталости или эмоциональных факторов. Кросс-согласованность между людьми и LLM составила $\alpha = 0.58$ при анализе только пар human-LLM, что близко к показателю human-human и указывает на сопоставимый уровень расхождений. Gemini 2.5 Pro демонстрировала наивысшую корреляцию с человеческими оценками ($\rho = 0.62$), особенно по критериям Final choice и Sentiment match.

Анализ систематических смещений выявил verbosity bias с умеренной положительной корреляцией между длиной ответа и рангом у людей (0.31) и более сильной у LLM (0.47), особенно выраженной у Qwen3 (0.54). Model bias анализ показал реальное качественное превосходство Gemini 2.5 Pro со средним баллом 2.8 из 4 при ожидаемом 2.5, в то время как Qwen3 получила наименьший балл 2.2. Monotonicity score оставался низким у обеих групп (0.23 у людей, 0.18 у LLM), подтверждая независимость оценки каждой задачи без

механического повторения порядка.

Статистическая значимость различий подтверждена через непараметрические тесты: U-тест Манна-Уитни показал значимые различия между human-human и LLM-LLM согласованностью ($p < 0.001$), а тест Краскела-Уоллиса выявил значимые различия по критериям оценки ($p < 0.01$). Агрегированные результаты ранжируют модели по эффективности: Gemini 2.5 Pro (средний ранг 1.8), Gemini 2.5 Flash (2.3), DeepSeek-R1 (2.6) и Qwen3 (3.3), демонстрируя чёткую градацию воспринимаемого качества в контексте корпоративной обратной связи.

Выводы

Исследование продемонстрировало, что современные LLM достигают умеренного уровня согласия с человеческими оценками качества корпоративной обратной связи, при этом модели показывают большую внутреннюю консистентность, чем люди, но уступают в понимании тонких культурных нюансов. Выявленные систематические смещения (verbosity bias, model-specific preferences) требуют учёта при практическом применении, а разработанная методология слепого тестирования с комплексом IRA-метрик формирует воспроизводимый стандарт для оценки LLM в HR-контексте. Результаты обеспечивают evidence-based основу для обоснованного выбора моделей и определения границ их автономного использования в чувствительных HR-процессах.

Литература

1. Yanamala K.K.R. Integrating Machine Learning and Human Feedback for Employee Performance Evaluation // Journal of Advanced Computing Systems. – vol. 2, no. 1 – 2022. – P. 1-10.
2. Durairaj S., Vetrivel V. The Effect of AI (Artificial Intelligence) in Employee Performance Evaluation on Employee Retention in the Information Technology Sector // International Conference on Digital Transformation in Business: Navigating the New Frontiers Beyond Boundaries (DTBNNF 2024), Atlantis Press. – 2024. – P. 88-108.
3. Islam R., Periaiah N. Overcoming the Pitfalls in Employee Performance Evaluation: An Application of Ratings Mode of the Analytic Hierarchy Process // Journal of Entrepreneurship, Management and Innovation. – vol. 19, no. 2 – 2023. – P. 127-157.
4. Wei J., Wang X., Schuurmans D., Bosma M., Ichter B., Xia F., Chi E., Le Q., Zhou D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models // Advances in Neural Information Processing Systems. – vol. 35 – 2022. – P. 24824-24837.
5. Ouyang L., Wu J., Jiang X., Almeida D., Wainwright C.L., Mishkin P., Zhang C., Agarwal S., Slama K., Ray A., Schulman J., Hilton J., Kelton F., Miller L., Simens M., Askell A., Welinder P., Christiano P., Leike J., Lowe R. Training Language Models to Follow Instructions with Human Feedback // Advances in Neural Information Processing Systems. – vol. 35 – 2022. – P. 27730-27744.
6. Zheng L., Chiang W.-L., Sheng Y., Zhuang S., Wu Z., Zhuang Y., Lin Z., Li Z., Li D., Xing E., Zhang H., Gonzalez J.E., Stoica I. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena // Advances in Neural Information Processing Systems. – vol. 36 – 2023. – P. 46595-46623.
7. Krippendorff K. Computing Krippendorff's Alpha-Reliability // Departmental Papers (ASC). – University of Pennsylvania – 2011. – P. 1-12.