Секция «43.1 Институты и развитие»

Когнитивные искажения российских больших языковых моделей в экономическом контексте: измерение и коррекция через промпты

Заявка № 1683225

В последние годы область применения больших языковых моделей (БЯМ) расширяется. Всё чаще крупные корпорации внедряют свои модели искусственного интеллекта в бизнес-процесс. Люди доверяют большим языковым моделям принятие решений во всех сферах своей деятельности. В экономическом контексте это важно с той точки зрения, что БЯМ выступают в роли инвестиционных консультантов, помощников менеджеров, что оказывает прямое влияние на поведение экономических агентов: как фирм, так и домохозяйств.

Нередко искусственный интеллект воспринимается как рациональный агент, который владеет полной информацией и способен быстро её обрабатывать. Однако эмпирические свидетельства демонстрируют систематическое отклонение БЯМ от рационального поведения. Это начинает играть большую роль в условиях развития ИИ-агентов для финансового консультирования, где наличие предвзятости может приводить к неоптимальным рекомендациям, не учитывающим риск-профиль и индивидуальные цели инверторов. Природа искажений больших языковых моделей отличается от человеческой: человек ограничен когнитивными способностями и вниманием, а БЯМ наследует ограниченно рациональное поведение из обучающих данных и иных особенностей построения модели [4]. Искажения возникают на двух ключевых этапах: предобучении на данных и дообучении во взаимодействии с пользователем [3]. Модели усваивают паттерны ограниченной рациональности из обучающих данных, где могут содержаться предвзятости или дискриминация [9]. Кроме того, искажения возникают из-за специфики формулировки промптов и истории взаимодействия, создающей смещение в сторону предыдущих ответов [1].

Методология исследований опирается на адаптацию классических экспериментов [7; 4]. Современные подходы включают систематический отбор искажений: случайная выборка [2], выбор на основе популярности и важности когнетивного искажения в литературе на оснвое критерия цитируемости работ [5]. Используется многократная генерация сценариев принятия решения для проверки устойчивости результатов работ [5]. Кроме этого, авторы часто применяют методы машинного обучения, например, примирение вектора Шепли для анализа вклада слов промта в итоговое решение [6]. Важным направлением становится изучение мультиагентных систем, где выявлено фундаментальное ограничение: коллективное рассуждение ограничено не индивидуальной рациональностью, а неспособностью агентов осознавать пробелы в знаниях и кооперироваться для их восполнения [8]. Ключевой вывод современных исследований заключается в том, что БЯМ не просто копируют человеческое поведение, что делает их взаимодействие непредсказуемым и требует разработки специфических методов коррекции искажений [1]. Таким образом, большие языковые модели демонстрируют когнитивные искажения, аналогично человеку. Хотя искажения имеют иной характер.

Эмпирическая стратегия данного исследования заключается в систематическом исследовании когнитивных искажений российских БЯМ и их сравнение с зарубежными аналогами. Оценка искажений состоит в проведении экспериментов, схожих с экспериментами для оценки наличия искажений у человека. Эксперименты проводятся с помощью подключения к API больших языковых моделей. Тестируются шесть когнитивных искажений, релевантных для инвестиционных решений: эффект якоря, эффект подражания, избегание потерь, эвристика доступности, эффект диспозиции, предпочтение статуса-кво.

Применяется плацебо-тестирование для оценки устойчивости полученного результата, которое заключается в просьбе выбрать случайно, игнорируя текст задания. Дополнительное направление исследования заключается в оценке эффективности методов коррекции когнитивных искажений. Согласно предварительным результатам, российские большие языковые модели подвержены конитивным искажениям, в частности эффекту якоря.

Источники и литература

1) Echterhoff J. M. et al. Cognitive bias in decision-making with LLMs //Findings of the Association for Computational Linguistics: EMNLP 2024. – 2024. – С. 12640-12653.
2) Geva T. et al. Do LLMs Exhibit Human-Like Cognitive Biases? A Large-Scale Systematic Evaluation //A Large-Scale Systematic Evaluation (September 17, 2025). – 2025.
3) Itzhak I., Belinkov Y., Stanovsky G. Planted in Pretraining, Swayed by Finetuning: A Case Study on the Origins of Cognitive Biases in LLMs //arXiv preprint arXiv:2507.07186. – 2025.
4) Macmillan-Scott O., Musolesi M. (Ir) rationality and cognitive biases in large language models //Royal Society open science. – 2024. – Т. 11. – №. 6.
5) Malberg S. et al. A comprehensive evaluation of cognitive biases in LLMs //Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities. – 2025. – С. 578-613.
6) Shaikh A. et al. CBEval: A framework for evaluating and interpreting cognitive biases in LLMs //arXiv preprint arXiv:2412.03605. – 2024.
7) Yax N., Anlló H., Palminteri S. Studying and improving reasoning in humans and machines //Communications Psychology. – 2024. – Т. 2. – №. 1. – С. 51.
8) Li Y. et al., Systematic Failures in Collective Reasoning under Distributed Information in Multi-Agent LLMs//arXiv preprint arXiv:2505.11556. – 2026.
9) Manyika J., Silberg J., Presten B. What do we do about the biases in AI //Harvard Business Review. – 2019. – Т. 25.