УДК 51-77

Модели социального влияния: от классических агентов до больших языковых моделей

***А. Д. Хлытчиев*[1], *И. В. Козицин*[1,2]**

[1]Московский физико-технический институт (национальный исследовательский университет)
[2]Институт проблем управления им. В.А. Трапезникова РАН (Москва)

В настоящее время при изучении поведения людей в социальных группах все чаще используются большие языковые модели (LLM). Классические многоагентные модели не могут описывать всю комплексность человеческого поведения: к примеру, общение между агентами они описывают как обмен некоторыми численными характеристиками, упрощенно имитирующими взгляды или аргументы реальных людей, что крайне далеко от естественного языка, который используется в коммуникациях [1, 2]. LLM предоставляют достаточно экономичный способ моделировать социального взаимодействия без использования информации о настоящих людях, при этом достаточно правдоподобно воспроизводя реальный обмен мнениями. Более того, такой подход освобождает ученых от проведения дорогостоящих и сложных в организации и реализации лабораторных и натурных экспериментов [4].

Поскольку LLM обучаются на данных, “оставленных” в сети Интернет реальными людьми, мнения LLM могут быть предвзяты, а в научных вопросах даже отличаться от достоверной истины [3]. Ученые полагают, что степень предвзятости мнений LLM можно контролировать и настраивать при помощи специальных инструкций, задаваемых при инициализации моделей (англ. – prompts). Вместе с тем данные аспекты еще мало изучены. В связи с этим, особый интерес представляет поведение LLM при взаимодействии друг с другом и с реальными людьми – поскольку данная технология все чаще используется в качестве ассистента в повседневной и профессиональной деятельности людьми самых разных профессий.

В настоящем докладе будут представлены результаты экспериментов с использованием LLM. Изучение природы валентности различных аргументов и результатов изменений предубеждений исходной модели. Для различных воплощений одной и той же языковой модели при помощи инструкций будет создаваться ее предубежденность. Также в докладе будет представлены результаты по нескольким обсуждаемым моделями темам. Такая постановка позволяет проанализировать природу аргументов, воспринимаемых LLM.

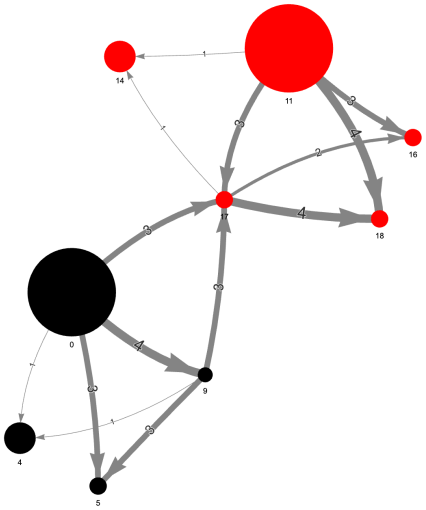

Рис. 1. Граф динамики метрики убедительности аргументов в зависимости от предубежденности модели другим набором аргументов.

Литература

1. *Chuang Y. S. et al.* Simulating opinion dynamics with networks of llm-based agents // arXiv preprint arXiv:2311.09618. – 2023.

2. *Betz G.* Natural-language multi-agent simulations of argumentative opinion dynamics // arXiv preprint arXiv:2104.06737. – 2021.
3. *Bail C. A.* Can Generative AI improve social science? // Proceedings of the National Academy of Sciences. – 2024. – T. 121. – No. 21. – C. e2314021121.
4. *Carpentras D. et al.* Deriving an opinion dynamics model from experimental data // Journal of Artificial Societies and Social Simulation. – 2022. – Vol. 25. – No. 4.
5. Flache A. et al. Models of social influence: Towards the next frontiers //Jasss-The journal of artificial societies and social simulation. – 2017. – T. 20. – No. 4. – C. 2.
6. Mäs M., Flache A. Differentiation without distancing. Explaining bi-polarization of opinions without negative influence //PloS one. – 2013. – T. 8. – №. 11. – C. E74516.
7. Banisch S., Shamon H. Biased processing and opinion polarization: experimental refinement of argument communication theory in the context of the energy debate //Sociological Methods & Research. – 2025. – T. 54. – №. 1. – C. 187-236.