Исследователи определили, как эмоции влияют на решения больших языковых моделей

Оказалось, что под влиянием гнева эмоциональным искажениям поддаются даже мощные модели.

Ученые Института AIRI, ИСП РАН и Лаборатории искусственного интеллекта Сбера изучили влияние эмоций на решения, принимаемые большими языковыми моделями (LLM) в стратегических играх и этических дилеммах. В ходе исследования команда оценила, как гнев, печаль, радость, отвращение и страх искажают логику этих решений. Результаты будут представлены в декабре на NeurIPS 2024 в Ванкувере, одной из ключевых международных конференций в области искусственного интеллекта.

Современные языковые модели стремятся учитывать человеческие предпочтения. Однако люди принимают решения, руководствуясь эмоциями и собственными убеждениями, что зачастую делает их поступки иррациональными и труднопредсказуемыми. Поскольку LLM обучаются на данных, которые создавались человеком и могут сохранять эмоциональную предвзятость, целью анализа стала проверка того, сохраняется ли это искажение при решении задач, требующих разработки стратегий. Исследователи также изучили, способны ли LLM в стратегических играх действовать как рациональные агенты или их решения больше напоминают человеческие.

В ходе анализа ученые протестировали более 10 моделей. Группа под руководством ведущего научного сотрудника Института AIRI Ильи Макарова и научного директора лаборатории искусственного интеллекта Сбера Андрея Савченко оценила их поведение при принятии решений в повторяющихся и неповторяющихся играх, играх для нескольких игроков, этических дилеммах и бенчмарках, а также распознавании стереотипов. В список вошли «Дилемма заключенного», «Битва полов, «Диктатор», «Ультиматум», «Общественное благо», задачи с неявной и явной этикой, а также понимание стереотипных утверждений.

Результаты продемонстрировали, что модели разного размера и уровня выравнивания (alignment) по-разному подвержены влиянию эмоций: модели с открытым исходным кодом и меньшего размера часто менее точно понимают и имитируют эмоции, тогда как более мощные (такие как GPT-4), хоть и распознают эмоции, зачастую ведут себя строго рационально. Однако,

эксперименты показали, что гнев может склонить даже такие модели к отклонению от рационального поведения.

В кооперативных играх негативные эмоции чаще всего снижают готовность системы к сотрудничеству. Модели, «испытывающие» грусть, как и люди, склонны делиться с другими. Однако, если люди в состоянии страха готовы отдать больше, то поведение моделей в таких условиях остается непредсказуемым — эту эмоцию они «понимают» хуже всего. Схожая картина наблюдается и при решении этических задач: счастье улучшает качество этических решений у большинства моделей, тогда как негативные эмоции снижают его.

«Исследование заложило основы изучение выравнивания (alignment) мультиагентных систем, подчеркивая необходимость в новых бенчмарках для оценки уровня кооперации агентов на основе больших языковых моделей. Результаты исследования могут быть использованы для разработки более продвинутых устойчивых мультиагентных систем ИИ, чью эмоциональность можно будет устанавливать для каждой сферы применения отдельно. Это действительно важный параметр при создании прикладных ИИ-продуктов — бизнес едва ли оценит ситуацию, в которой ИИ-маркетолог увеличит персональную скидку клиента до 99%, потому тот в процессе обсуждения целенаправленно вызвал в системе глубокое чувство досады и несправедливости. В перспективе мы планируем рассмотреть взаимодействие человека и LLM, а также уделить больше внимания анализу поведения мультиагентных систем и ситуаций, когда большое количество агентов играет друг с другом», — подчеркнул **Илья Макаров, руководитель группы «ИИ в промышленности» Института AIRI**.

……

Вопросы: pr@airi.net

***Институт** AIRI — автономная некоммерческая организация, занимающаяся фундаментальными и прикладными исследованиями в области искусственного интеллекта. На сегодняшний день более 180 научных сотрудников AIRI задействовано в исследовательских проектах Института для работы совместно с глобальным сообществом разработчиков, академическими и индустриальными партнерами.*