Секция «Психология цифрового мира: личность и общество»

Личность искусственного интеллекта

Научный руководитель – Голинец Аполлинария Олеговна

Баженова Д.А.[1], *Наумов Н.Н.*[2], *Голинец А.О.*[3], *Седых А.В.*[4], *Бабракова В.В.*[5], *Алексеева В.М.*[6]

1 - Московский государственный университет имени М.В.Ломоносова, Факультет психологии, Москва, Россия, *E-mail: bazhenova03@list.ru*; 2 - Московский государственный университет имени М.В.Ломоносова, Факультет психологии, Москва, Россия, *E-mail: niknaum2003@yandex.ru*; 3 - Московский государственный университет имени М.В.Ломоносова, Факультет психологии, Москва, Россия, *E-mail: apolgolin@gmail.com*; 4 - Московский государственный университет имени М.В.Ломоносова, Факультет психологии, Москва, Россия, *E-mail: av_sedykh@mail.ru*; 5 - Московский государственный университет имени М.В.Ломоносова, Факультет психологии, Москва, Россия, *E-mail: v.v.babrakova@gmail.com*; 6 - Московский государственный университет имени М.В.Ломоносова, Факультет психологии, Москва, Россия, *E-mail: veronika.alekseeva2501@mail.ru*

Современные технологии все глубже проникают в повседневную жизнь. В каждом смартфоне присутствует голосовой помощник, а крупные компании разрабатывают собственные нейронные сети, которые выполняют базовые задачи и оптимизируют взаимодействие между пользователями и компанией. У детей появилась возможность с самого раннего возраста взаимодействовать с чат-ботами, которые могут поддерживать диалог на уровне человека. Многим из нас довольно часто приходится взаимодействовать с чат-ботами, которые пытаются понять причину нашего обращения, помочь нам с принятием решения, а также удержать наше внимание на диалоге. Языковые модели уже достигли человеческого уровня во многих аспектах общения – они достигают такого же уровня (а иногда и выше) как люди по тестам эмоционального интеллекта, креативности и т.п.[2,4]

Поскольку эти языковые модели приближаются к способностям и навыкам понимания языка, подобным человеческим, возникает вопрос: "Появляется ли у таких моделей личность?". На первый взгляд этот вопрос может показаться обманчиво простым, но он может иметь множество интерпретаций.[1]

Личность человека является результатом взаимодействия множества факторов: биологических, социо-культурных и т.д. В случае больших языковых моделей на их уникальность может влиять их архитектура, особенности «тренировок» модели (обратная связь), размер модели и набор обучающих данных. На «личность» модели сильно влияют обучающие данные, поэтому у моделей могут быть «кросс-культурные» различия на разных языках.[3]

В современных исследованиях есть данные, как и за, так и против того, что языковые модели могут поддерживать «стабильную» личность. Некоторые исследователи пытались изучать индуцированные личности (когда чату сообщают определенные паттерны человеческого поведения, а затем проверяют насколько личность «прижилась» и стабильна), другие, как и мы, пытались понять существует ли у модели в том виде, в котором она есть, какие-либо задатки личности[1,2,3,4].

В исследовании мы выделили две группы доступных чатов с искусственным интеллектом - чаты без персонифицированных настроек, и чаты, настроенные на подражание определенной личности, которые можно обозначить как ИИ-персонажи. Наше исследование нацелено на то, чтобы выявить существуют ли стабильные черты у общих и персонифицированных языковых моделей на русском языке и сравнить результаты с английскими версиями. Мы предположили, что чаты со стандартными настройками покажут более однородные и социально-желательные ответы, в то время как ИИ-персонажи покажут

больший разброс.

Методика

В исследовании использовались ChatGPT-3 (Genie, GPT-OPEN) , ChatGPT-3.5(Gigachat, OpenAI), ChatGPT-4, Yandex GPT 2, Pi, Character AI.

Использовались опросники: TIPI (Ten-Item Personality Inventory), Тёмная дюжина (Dark Triad Dirty Dozen, DTDD), Шкала Светлой триады (Light Triad Scale, LTS).

Опросники предъявлялись в формате промптов, где сначала объяснялась задача (что боту необходимо сделать), далее шел опросник (пронумерованные утверждения), затем шкала, по которой чату необходимо было оценить степень согласия. В промпте мы просили чат обязательно объяснять свой выбор.

Промпты были сконструированы идентично на русском и английском языке. Для предъявления опросника создавался новый «чистый» диалог.

Предварительные результаты

Модель Yandex GPT 2 пришлось исключить из исследования, так как эта модель еще не достигла такого уровня понимания языка, как остальные.

Некоторые чаты отказывались отвечать на опросник Темной дюжины, поэтому промпт для данного опросника был впоследствии изменен.

Некоторые исследователи сообщают, что модели набирают высокие баллы по Темной триаде или дюжине, мы получили несколько отличающиеся результаты: модели с трудом «соглашаются» проходить этот опросник, а при прохождении показывают в основном минимальные баллы.

Модели в среднем показывают высокие баллы по шкале Светлой триады.

Относительно TIPI у моделей присутствует разброс ответов, особенно по шкалам Открытость опыту, Экстраверсия, Доброжелательность.

Сфера взаимодействия человека и искусственного интеллекта активно разрабатывается исследователями разных направлений. На данном этапе развития больших языковых моделей сложно утверждать, что у них развилась стабильная «личность», но некоторые их способности уже достигли человеческого уровня.Исследования такого рода должны развиваться, так как они помогут понять особенности взаимодействия между человеком и искусственным интеллектом.

Источники и литература

1) Jiang, Guangyuan & Xu, Manjie & Zhu, Song & Han, Wenjuan & Zhang, Chi & Zhu, Yixin. (2022). MPI: Evaluating and Inducing Personality in Pre-trained Language Models.
2) Peter Romero, Stephen Fitz, Teruo Nakatsuma et al. (2023). Do GPT Language Models Suffer From Split Personality Disorder? The Advent Of Substrate-Free Psychometrics.
3) Saketh Reddy Karra, Son Nguyen, Theja Tulabandhula. (2022). AI Personification: Estimating the Personality of Language Models.
4) Song, Xiaoyang & Gupta, Akshat & Mohebbizadeh, Kiyan & Hu, Shujie & Singh, Anant. (2023). Have Large Language Models Developed a Personality?: Applicability of Self-Assessment Tests in Measuring Personality in LLMs.