УДК 004.5

АНТРОПОМОРФИЗМ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ: ИМИТАЦИЯ КОГНИТИВНЫХ ФУНКЦИЙ КАК ЧАСТЬ ПОЛЬЗОВАТЕЛЬСКОГО ИНТЕРФЕЙСА

♦♦♦♦♦

ANTHROPOMORPHISM IN LARGE LANGUAGE MODELS: IMITATION OF COGNITIVE FUNCTIONS AS A USER INTERFACE ELEMENT

Деревянко Иван Владимирович
студент,
Кубанский государственный технологический университет
xdlibraxd@mail.ru

Вонарх Юлия Сергеевна
старший преподаватель
кафедры информационных систем и программирования,
Кубанский государственный технологический университет
jul_nosova@mail.ru

Аннотация. В статье исследуется феномен антропоморфизма в современных больших языковых моделях (LLM). Изучаются передовые технологические основы обучения моделей, включая методы обучения с подкреплением на основе обратной связи от человека (RLHF), которые создают имитацию человеческого поведения. В работе обосновывается концепция, согласно которой антропоморфность больших языковых моделей является не признаком наличия сознания, а высокотехнологичным интерфейсом, оптимизированным для взаимодействия с пользователем. Проанализированы психологические эффекты восприятия искусственного интеллекта, такие как «Эффект Элизы», и их общее влияние на доверие пользователя к системе.

Ключевые слова: большие языковые модели, LLM, антропоморфизм, искусственный интеллект, эффект Элизы, пользовательский интерфейс, RLHF, этика ИИ.

Derevyanko Ivan Vladimirovich
Student,
Kuban State Technological University
xdlibraxd@mail.ru

Vonarkh Yulia Sergeevna
Senior Lecturer of the Department
of Information Systems and Programming,
Kuban State Technological University
jul_nosova@mail.ru

Annotation. The article explores the phenomenon of anthropomorphism in modern large language models (LLMs). Advanced technological foundations of model training are examined, including reinforcement learning from human feedback (RLHF) methods that create an imitation of human behavior. The study primarily substantiates the concept that the anthropomorphic nature of large language models is not a sign of consciousness, but a high-tech interface optimized for user interaction. Psychological effects of artificial intelligence perception, such as the «Eliza effect», and their overall impact on user trust in the system are analyzed.

Keywords: large language models, LLM, anthropomorphism, artificial intelligence, Eliza effect, user interface, RLHF, AI ethics.

В последние годы большие языковые модели (Large Language Models, LLM) совершили качественный переход от узкоспециализированных инструментов к универсальным ассистентам, способным поддерживать диалог с пользователем на естественном для него языке. Широкое распространение данных систем привело к возникновению выраженного феномена антропоморфизма – склонности пользователей приписывать алгоритмам человеческие черты, волю и наличие сознания [1]. Актуальность исследования обусловлена необходимостью разграничения реальных когнитивных способностей искусственного интеллекта и имитационных паттернов поведения, заложенных в их архитектуру для оптимизации пользовательского опыта.

Архитектура Transformer и методы обучения моделей

В основе функционирования больших языковых моделей лежит архитектура Transformer, которая базируется на механизмах самовнимания (self-attenction) [2]. С технической точки зрения процесс генерации ответа представляет собой последовательное статистическое предсказание следующего наиболее вероятного токена в заданной текстовой последовательности. Однако модель, обученная на огромных массивах неструктурированных данных, в основном демонстрирует несвязное поведение, которое пользователь не сможет понять вовсе, вследствие чего контекст диалога с машиной теряется.

Для создания осмысленного диалога применяется технология обучения с подкреплением на основе обратной связи от пользователя (RLHF). На этом этапе живые операторы (асессоры) оценивают ответы модели, ожидая увидеть ответы, которые соответствуют логике.

Формирование имитационных паттернов поведения

В процессе RLHF асессоры не только оценивают логическую связность ответов, но и осуществляют настройку этической стороны модели. Особое внимание уделяется параметрам безопасности и полезности: модель обучается отклонять или игнорировать деструктивные запросы и придерживаться заданного тона общения. На данном этапе происходит формирование имитационных паттернов: если асессоры ставят целью получать ответы, содержащие сочувствие или человеческую индивидуальность, то модель закрепляет эти стратегии как наиболее эффективные для взаимодействия. Таким образом, «личность» искусственного интеллекта становится результатом статистической оптимизации под ожидания человека, но вовсе не продуктом самосознания системы. Модель эффективно «притворяется» субъектом диалога, «персоной», так как именно такая форма интерфейса обеспечивает наивысший уровень удержания внимания и комфорта пользователя.

Психологические аспекты эффекта Элизы

Феномен антропоморфизма в контексте взаимодействия с LLM базируется на естественных психологических установках человеческого сознания [3]. Ключевым фактором здесь выступает известный «эффект Элизы» – склонность человека ошибочно приписывать компьютерной программе одушевлённые качества и понимание смыслов, исключительно основываясь на внешней убедительности текстовых ответов [4].

Наглядно «эффект Элизы» усиливается в условиях современного социального контекста. Рост уровня социальной изоляции и дефицит качественного межличностного общения вызывают потребность в дешёвом, удобном и всегда доступном для диалога субъекте. LLM, благодаря своей архитектуре, заполняют этот социальный вакуум, становясь для пользователя некоторым суррогатным собеседником. Склонность к одушевлению предметов, зачастую неосознаваемая самим разумом, в данном случае переносится на цифровой интерфейс, превращая статистический алгоритм в подобие «личности» в восприятии человека.

Этические риски и проблема доверия к алгоритмам

Центральная этическая проблема – размытие границ восприятия искусственного интеллекта как инструментального средства. Когда пользователь перестаёт воспринимать LLM как алгоритмическую систему, направленную на решение прикладных или творческих задач, возникает феномен избыточного доверия к фактически неполноценно достоверным алгоритмам [5].

Возникает обоснованная, не обязательно намеренная, провокация разработчиками самой модели. Когда пользователь начинает перекладывать ответственность за принятие решений в личной и социальной сферах. Подобная опора на искусственный интеллект в вопросах формирования жизненных стратегий влечёт за собой риск эрозии личной идентичности и деградации навыков самостоятельного выбора. В этом смысле антропоморфизм выступает как манипулятивный фактор: «благодетельный» интерфейс снижает порог критического восприятия, заставляя принимать статистически вероятные ответы за истину, что может привести человека к совсем не тому которое он предполагал ранее, и направить его к выбору деструктивных решений в реальной жизни.

Антропоморфизм больших языковых моделей является не побочным эффектом их развития, а сознательно сконструированным элементом развитого интерфейса. Технология RLHF и психологические механизмы образуют эффективную среду для пользования с огромными массивами данных через уже обыденную форму диалога.

Несмотря на этические риски и проблемы деформации личной идентичности при чрезмерном доверии к алгоритмам, антропоморфизация искусственного интеллекта открывает новые возможности для социальной поддержки или её иллюзии. В условиях растущей изоляции использование нейросетей как суррогатных собеседников не

является деструктивным фактором при условии осознанного подхода пользователя к искусственному интеллекту как к инструменту, а не как к субъекту. Человечество стоит на пороге более глубокой интеграции с искусственным интеллектом, где «человечность» интерфейса станет стандартом многих цифровых продуктов, облегчающих когнитивную нагрузку. Дальнейшее развитие систем должно быть направлено на соблюдение баланса между удобством антропоморфного взаимодействия и сохранением критического мышления пользователя.

Литература

1. Luger G.F. LLMs: Their Past, Promise, and Problems // International Journal of Semantic Computing. – 2024. – Vol. 18. – № 3. – P. 501–544.
2. The rise of transformers – redefining the landscape of artificial intelligence / N.S.D. Ladu [et al.] // Babylonian Journal of Artificial Intelligence. – 2025. – P. 72–76.
3. Гаврилина Е.А. Агентность не-человеков: взаимодействие людей и автономных интеллектуальных систем // Философские проблемы информационных технологий и киберпространства. – 2023. – № 1(21). – С. 44–51.
4. Вейценбаум Дж. Возможности вычислительных машин и человеческий разум. От суждений к вычислениям. – М. : Радио и связь, 1982.
5. Large language model for predicting altered mental status in dementia patients using electronic health records / J. Wu [et al.] // Alzheimer's & Dementia. – 2025. – Vol. 21. – № S4.

References

1. Luger G.F. LLMs: Their Past, Promise, and Problems // International Journal of Semantic Computing. – 2024. – Vol. 18. – № 3. – P. 501–544.
2. The rise of transformers – redefining the landscape of artificial intelligence / N.S.D. Ladu [et al.] // Babylonian Journal of Artificial Intelligence. – 2025. – P. 72–76.
3. Gavrilina E.A. Agency of non-humans: interaction between humans and autonomous intelligent systems. Philosophical Problems of IT and Cyberspace. – 2023. – № 1(21). – P. 44–55.
4. Weizenbaum J. Computer Power and Human Reason: From Judgment to Calculation. – M. : Radio i svyaz, 1982.
5. Large language model for predicting altered mental status in dementia patients using electronic health records / J. Wu [et al.] // Alzheimer's & Dementia. – 2025. – Vol. 21. – № S4.