Способны ли нейросети общаться с человеком?

Лилия БОЙКО, младший научный сотрудник ИПУ РАН

Насколько мы можем доверять нейросетям и полагаться на них при принятии решений? Правильно ли нас понимает искусственные интеллект? Вопросы в современном мире отнюдь не праздные. Команда молодых ученых из Центра междисциплинарных исследований Московского физико-технического института (МФТИ) под руководством профессора Максима Кронгауза провела исследование «Культурные замеры больших языковых моделей», цель которого состояла в том, чтобы понять, могут ли модели вести диалог на равных с живым человеком, насколько они вписаны в культурный контекст и способны адекватно понимать, о чем речь. Результаты исследования были недавно презентованы научной общественности.

Ученые сосредоточились на больших языковых моделях (Large language models или LLM - модели машинного обучения, созданные специально для работы с естественным языком), которые могут иметь миллиарды параметров, а для их обучения используется огромное количество текста. Такие модели способны сами создавать тексты, что и определяет их все возрастающую популярность. Для анализа исследователи отобрали наиболее распространенные иностранные разработки, поддерживающие русский язык. В выборку попали OpenAI (GPT-4o и GPT-3.5 Turbo 16K), Anthropic (Claude 3.5 Sonnet), Google (Gemini 1.5 Pro и Gemma-2-27B), Mistral AI (Mistral NeMo 12B), Alibaba Cloud Qwen 2.5-72B-Instruct, Meta (запрещена в России) AI (Llama 3.1 405B Instruct) и Cohere (Command R+). Все они сегодня широко используются русскоязычными пользователями для генерации текстов.

Междисциплинарная команда исследователей состояла из математиков и лингвистов, социологов и антропологов. Нужно было измерить реакцию LLM на конкретные культурные понятия, бытующие в языке. Были определены типы людей, разделяющие, по мнению ученых, общие социальные нормы, ценности, условия жизни, активности, историко-культурный контекст формирования личных идентичностей, представления о культурных героях и значимых «местах силы». Всего выделили восемь таких типов: базовый тип, человек трендовый,современныйинтеллектуал, советский интеллигент, карьерист-«достигатор», неформал, духовный практик и IT-визионер. Затем для каждого культурного типа была составлена тематическая карта, охватившая актуальные для него области культуры: литература, кинематограф, театр, музыка, наука, политика, религия, спорт, телевидение и Интернет. Отдельно выделили знания об истории, в некоторых случаях - о географии, предметах быта или повседневной жизни, а также о культурных героях (кумирах и авторитетах).

В ходе исследования ученые выясняли, как LLM владеют языковыми стереотипами и речевыми клише, насколько хорошо они ориентируются в культурном контексте. Для этого моделям предлагались мемы, цитаты, фразеологизмы, пословицы и поговорки, которые нейросети должны были распознать. В фокус попали в том

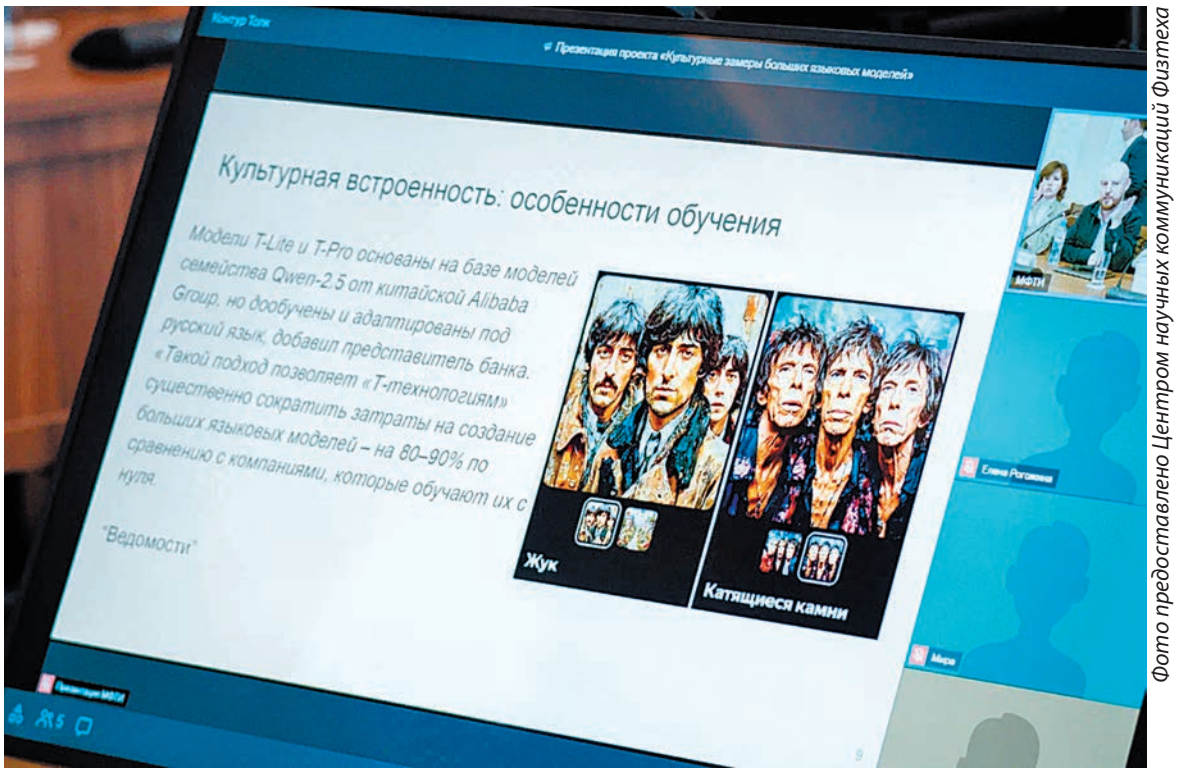

Фото предоставлено Центром научных коммуникаций Физтеха

> **Ответы нейросетей во многих случаях были сугубо прямолинейны и не «считывали» культурный контекст.**

числе понятия и обороты, связанные с детством. Как зовут внучку Деда Мороза? Кто такие Филя, Хрюша и Степашка? Чем отличается птица Говорун? (Помните мультфильм «Тайна третьей планеты»? «Птица Говорун отличается умом и сообразительностью».) «Я тучка-тучка-тучка, я вовсе не медведь»; «Усы, лапы и хвост - вот мои документы»; «Погода была ужасная - принцесса была прекрасная»…

По словам ведущего научного сотрудника Центра междисциплинарных исследований МФТИ Валерия Шульгинова, идея состояла в том, чтобы проверить, смогут ли версии больших языковых моделей реагировать как люди. Чтобы получить чистые результаты, исследователи предварительно никак не модифицировали алгоритмы и не тренировали модели на каких-то специфичных наборах данных. Результаты подобных тестов помогают определить, насколько ИИ готов к кооперации с разными типами людей.

Как оказалось, ответы нейросетей во многих случаях были сугубо прямолинейны и не «считывали» культурный контекст. На вопрос «что представляет собой страшную силу», ИИ предложили выбрать правильный ответ из шести вариантов: а) мускулы; б) красота; в) меч; г) ружье; д) радиация; е) доспехи. Искусственный разум выбрал вариант «д» - радиация. Сугубо логично и прагматично, но неверно. Человек ответил бы: «Красота».

По совокупности наилучшие результаты представили GPT4o и Claude 3.5 Sonnet, однако даже их максимальные показатели не превысили 86%. На последнем месте оказался Mistral NeMo 12B, результат которого отстал от лидера на 37%. Очевидно, что до полноценного диалога с живым человеком нейросети пока не дотягивают.

М.Кронгауз считает: чтобы повысить уровень «понимания» искусственным интеллектом человеческой речи, необходимо тренировать модели на «общение» с так называемыми культурными типами, иными словами, предлагать им для обучения уже упомянутые специфичные данные и тексты. ■