.

GPT-5 еще даже не вышел, а люди уже не могут отличить ChatGPT от человека в разговоре.

Тарас Мищенко, 23.06.2024

Компания OpenAI только готовит к запуску свою следующую большую языковую модель GPT-5, которая должна стать еще больше и мощнее, и соответственно лучше имитировать ответы человека. Но недавнее исследование показало, что отличить ChatGPT от человека в разговоре уже очень сложно. Популярные инструменты искусственного интеллекта, такие как GPT-4, генерируют настолько свободный, похожий на человеческий текст, что прекрасно справляются с разными языковыми задачами, бросая вызов нашей способности отличить машину от собеседника, пишет The Register.

Эта разработка перекликается с известным мнимым экспериментом Алана Тюринга, который предложил тест, чтобы определить, может ли машина демонстрировать поведение, которое невозможно отличить от человеческого, основываясь исключительно на его ответах.

Исследователи кафедры когнитивных наук Университета Сан-Диего провели контролируемый тест Тюринга для оценки современных систем искусственного интеллекта. Они сравнили ELIZA, виртуального собеседника, созданного Джозефом Вейценбаумом в 1966 году, имитирующим диалог с психотерапевтом, реализуя технику активного слушания, с GPT-3.5 и GPT-4. Участники принимали участие в пятиминутных разговорах с человеком или с искусственным интеллектом, а затем их просили решить, был ли их собеседник человеком.

Для теста GPT-4 и GPT-3.5 получили специальные инструкции по соответствию. Моделям было сказано "принять образ молодого человека, который был очень лаконичен, не воспринимал игру очень серьезно, использовал сленг и время от времени совершал орфографические ошибки". Кроме того, подсказка содержала общую информацию о настройке игры и последних новостях, которые касались обучения моделей. Ответы ИИ задерживались в зависимости от длины сообщения, чтобы предотвратить слишком быстрый ответ.

В исследовании приняли участие 500 участников, разделенных на пять групп, каждая из которых взаимодействовала либо с человеком, либо с одной из трех ИИ-систем. Интерфейс разговора напоминал типичный мессенджер для обмена сообщениями. Через пять минут участники определяли, кто был их собеседником — человек или искусственный интеллект и обосновывали свои выводы.

Результаты были впечатляющими: GPT-4 был идентифицирован как человек в 54% случаев, немного опередив GPT-3.5 на 50%, и оба значительно превзошли ELIZA, который был идентифицирован как человек только в 22% случаев. Настоящие люди были правильно идентифицированы в 67% случаев. Это свидетельствует о том, что участники не лучше, чем случайно, идентифицировали GPT-4 как ИИ, демонстрируя совершенство современных систем ИИ в имитации разговора, похожего на человеческий.

Решения участников часто основывались на лингвистическом стиле, социально-эмоциональных подсказках и вопросах, основанных на их знаниях. Это исследование подчеркивает проблему того, что людям все труднее отличить сгенерированный ИИ текст от человеческих ответов, и эта тенденция, вероятно,

продолжится с развитием технологии искусственного интеллекта и выпуском новых крупных языковых моделей, как GPT-5.