Нейросеть *chatgpt* – враг или помощник?

Корчажкина О.М. (olgakomax@gmail.com)

*Институт кибернетики и образовательной информатики им. А.И. Берга*

*ФИЦ «Информатика и управление» РАН, г. Москва*

Аннотация

В статье обсуждаются возможности использования нейронной сети глубокого обучения *ChatGPT* в системе образования. Приводятся примеры результатов работы нейросети по реальным запросам, которые показывают несостоятельность надежды пользователей относительно генерации ею качественных и достоверных текстовых материалов как в языковом, так и в содержательном плане, по крайней мере на современном уровне развития алгоритманекоммерческой версии *ChatGPT* .

**1.** Нейронная сеть глубокого обучения *ChatGPT* была разработана американской компанией OpenAI с целью автоматического создания текстов различного объёма на естественных языках. В описании нейросети *ChatGPT*, которое можно найти в Интернете, говорится, что на основе обширных баз данных она способна в ответ на заданный запрос быстро сгенерировать качественные тексты – эссе, исследования, научные статьи и пр., почти неотличимые от текстов, создаваемых человеком. Нейросеть вышла в открытый доступ в ноябре 2022 года и сразу завоевала огромную популярность – особенно в среде школьной и студенческой молодёжи.

Всем известен пример написания нейросетью *ChatGPT* дипломной работы на тему «Анализ и совершенствование управления игровой организацией» по запросу студента РГГУ Александра Жадана, которая получила оценку «удовлетворительно». Написанию дипломной работы Александр посвятил всего 23 часа личного времени: 15 часов – на формулировку запроса на составление развёрнутого плана работы, согласование полученного от нейросети плана с преподавателем, корректировку плана и формулировку нового запроса по написанию нейросетью полного текста диплома, а также ещё 8 часов «ручного труда» на редактирование полученного текста. Поскольку ни преподаватель, ни дипломная комиссия не отстранили студента от защиты, можно полагать, что «детище *ChatGPT*» было принято за работу, которую не очень прилежный выпускник вуза выполнил сам.

**2.** Исследование способности «машин» маскироваться под человека восходит к 1950-му году, когда знаменитый британский математик Алан Тьюринг в статье «Вычислительные машины и разум» постановкой вопроса «Могут ли компьютеры имитировать человеческий мозг?» обобщил свой десятилетний опыт работы над проблемой, чем заложил гуманитарные – морально-нравственные и социальные – основы искусственного интеллекта (ИИ). Тьюрингу важно было получить ответ на вопрос, способен ли ИИ имитировать реальное человеческое поведение. Для проверки своей гипотезы Тьюринг предложил специальную модель взаимодействия человека и компьютера в виде теста, который был выстроен по аналогии с популярной в то время «игрой в имитацию». Тест имел три модификации, и его должна была пройти именно машина (то есть испытуемой являлась именно она), чем показать – удалось ей или нет обмануть человека-эксперта.

По *первой модификации* у человека-эксперта было два собеседника – человек-ассистент и испытуемый-компьютер. Эксперт в письменной форме задавал вопрос, а ответить на него (так же в письменной форме) мог либо ассистент, либо компьютер. Задача эксперта – угадать, кто же ему ответил.

*Вторая модификация* теста предполагала, что эксперт наблюдает за общением ассистента и компьютера, перебрасывающимися письменными сообщениями. И задача эксперта – понять, кто же из собеседников является человеком, а кто – машиной.

Согласно *третьей модификации* экспертами выступали уже несколько человек – специальное жюри, задающее вопросы компьютеру, который старается убедить их в том, что он человек. Тест считается пройденным компьютером, если ему удаётся обмануть жюри хотя бы в трети случаев.

Известны по крайней мере четыре случая, когда машина прошла одну из модификаций теста Тьюринга: программы Eliza (1966 г.) и Parry (1972 г.), веб-приложение Cleverbot (2011 г.) и чат-бот Eugene Goostman (2014 г.). Американский футуролог, изобретатель систем для распознавания речи Рэймонд Курцвейл ещё в 2002 году высказал мнение, что компьютер сможет пройти все три модификации теста Тьюринга только к 2029 году.

Что касается нейросети *ChatGPT*, то специалисты могут испытать её возможности уже сегодня, предложив человеку-испытуемому пройти обратный тест Тьюринга, когда не компьютер имитирует человека, а человек пытается соответствовать алгоритму *ChatGPT*. Ещё один обратный тест предполагает команде экспертов вызвать нейросеть *ChatGPT* на поединок и сравнить уровень созданных ими текстов согласно аналогичным запросам, причём людям будет разрешено пользоваться определёнными вычислительными и поисковыми сервисами.

**3.** Помимо оперирования фактами по запросу пользователя и владения грамматикой и лексикой естественного языка «машина» должна выстраивать свою письменную речь теми способами, которыми пользуется реальный человек в соответствующих ситуациях, то есть владеть стилистикой текста. Общепринятыми стилями в настоящее время являются научный, официально-деловой, публицистический, художественный и разговорный, причём каждый из них имеет свои различительные элементы: структуру (заголовок, подзаголовки, абзацы, цель изложения, введение, заключение и пр.), специальную или общеупотребительную лексику, синтаксические конструкции, отвечающие стилистике текста. Все перечисленные особенности текстового отклика нейросети, свидетельствующие о её напряжённой «мыслительной деятельности», должны, как представляется, демонстрировать её не рутинные, а творческие возможности.

На самом деле, и это никому не приходит в голову оспаривать, нейросеть путём обработки обширных баз данных всего лишь компилирует сведения, заложенные в текстах, и выполняет только те задачи, на которых она обучалась и продолжает обучение. Запросы, требующие выполнения высокоуровневых когнитивных действий, ставящие цель дать моральную оценку тем или иным фактам и событиям, проявить эмпатию, высказать «свою» точку зрения, демонстрируя прочные убеждения, – всё это находится за пределами дееспособности нейросети, которая не может имитировать человека как психологический феномен: нейросетям не под силу генерировать новое качество – они полагаются только на то, что создано человеком.

**4.** Ограниченные возможности нейросети создавать качественные тексты на русском языке объясняются не только тем, что алгоритм *ChatGPT* базируется преимущественно на англоязычном корпусе текстов и связан с необходимостью выполнять автоматизированный перевод с английского языка на русский, что непременно сопровождается потерей языковой и речевой аутентичности в силу очевидных «трудностей перевода» языков романо-германской группы на языки славянской культуры. Набор датасетов (обработанных и структурированных массивов данных, представленных в виде таблиц), на котором происходит обучение нейросети и который содержит часто непроверенные и спонтанно собранные тексты, может включать недостоверную информацию или явные ошибки, приводящие к последующему недопониманию темы, актуального контекста или, что более существенно – к опоре на заведомо ложные утверждения, приводящие к ложным выводам. Для учащегося, недостаточно хорошо знакомого с предметом, эти детали могут остаться незамеченными, а опытный преподаватель вуза или учитель средней школы при внимательной проверке текста вполне способен определить, кто является его автором – учащийся или нейросеть.

Однако студенты и учащиеся обычно рассматривают современных, даже относительно молодых преподавателей и учителей, особенно гуманитариев, как людей, отставших от жизни и не способных освоить весь спектр новых технологий. Тогда как сами учащиеся, чрезмерно доверяющие и неоправданно часто прибегающие к помощи нейросети, рискуют невольно подпасть под её влияние, грозящее утратой личностного суверенитета и сломом мировоззренческих установок, в особенности связанных с исторической памятью.

Так с целью оценки возможностей нейросети *ChatGPT* работать с источниками по российской и советской истории учащимися одной из московских школ были сделаны запросы о фактах гибели двух бойцов Красной армии периода Великой Отечественной войны: члена диверсионно-разведывательной группы штаба Западного фронта Зои Космодемьянской и генерал-лейтенанта инженерных войск Дмитрия Михайловича Карбышева – каждому из них посмертно было присвоено звание Героя Советского Союза. Каково же было удивление и негодование школьников, когда нейросеть выдала короткие тексты, суть которых сводилась к следующему: Зоя Космодемьянская была расстреляна в ноябре 1941 года на Красной площади в Москве, а генерала Дмитрия Карбышева как исторической личности вообще не существовало. Тогда как школьники знали, что Зоя Космодемьянская – девушка, которой на момент гибели было всего 18 лет, была подвергнута нечеловеческим пыткам, повешена и затем зверски растерзана фашистами на глазах жителей подмосковной деревни Петрищево 29 ноября 1941 года. А генерал Д.М. Карбышев принял мученическую смерть от рук гестаповцев 18 февраля 1945 года на территории концлагеря Маутхаузен.

Таким образом, ненадёжные, непроверенные источники, бесконтрольно привлекаемые для обучения нейросетей, могут содержать не только явную дезинформацию, на основе которой способны генерировать фальшивые обоснования и описания несуществующих событий, но и различного рода манипуляционный контент, негативно влияющий на формирующуюся личность учащихся и восприятие ими окружающего мира.

**5.** Первый опыт обращения учащихся к нейронной сети *ChatGPT* – ещё весьма небезупречной, находящейся на стадии постоянного обучения и совершенствования – показал, что пока она не в состоянии самостоятельно, то есть без участия человека, качественно справиться со сложными заданиями, какие бы обширные на сегодняшний день базы данных ни привлекались ею для удовлетворения запроса пользователя. Поэтому в настоящее время заявка её создателей на предоставление материалов на языке, «неотличимом от написанного человеком», не имеет под собой сколько-нибудь реальных оснований, особенно если это касается текстов, сгенерированных на русском языке путём автоматизированного перевода, или при решении задач, которые ставят в тупик саму нейросеть.

Кроме того, преимущественно англоязычная база данных, привлекаемая нейросетью для создания текстов, может заключать в себе значительный объём недостоверной информации, или заведомо фальсифицированных данных, что делает её ещё менее надёжным инструментом, на который не стоит полагаться русскоязычным пользователям при создании серьёзных исследовательских и образовательных ресурсов.

**6.** Однако специалисты по ИИ не поддерживают столь пессимистические оценки нового машинного алгоритма. Они указывают, что подобные ошибки и демонстрация ограниченных возможностей нейросети характерны только для бесплатных версий *ChatGPT*, тогда как при доступе к коммерческим датасетам и бóльшим вычислительным мощностям, можно получить результаты принципиально иного уровня. То есть при наличии соответствующих ресурсов и создании условий для дальнейшего развития нейросетей эти трудности становятся временными и в обозримой перспективе преодолимыми.

Ведущий аналитик компании «Mobile Research Group» Эльдар Викторович Муртазин 29 мая 2023 года в интервью радио «Комсомольская правда» описал перспективы развития нейросетей, подобных *ChatGPT*, на основе российского контента с использованием отечественных вычислительных мощностей. Специалист сказал, что у России имеется неплохой потенциал для развития этого направления: сильный телеком, большое количество дата-центров, есть суперкомпьютеры для обсчёта больших объёмов данных и относительно дешёвая электроэнергия для их эффективной работы. Наша страна, наравне с США и Китаем, входит в пятёрку мировых лидеров по числу библиотек и датасетов, на основе которых строятся модели и пишутся алгоритмы их работы. Кроме того, у нас есть неоспоримое преимущество перед другими странами в плане развитого академического образования и сильного преподавательского и научно-исследовательского состава, что способствует созданию сильной базы датасетов, без которой нельзя осуществлять развитие систем ИИ.

Это позволит России в ближайшем будущем на основе надёжных датасетов, отвечающих корректным запросам, осуществлять постоянное обучение нейросетей, ориентированных на выполнение задач генерации высококачественных текстов на естественных языках.