УДК 004.912 URL: https://ptsj.bmstu.ru/catalog/icec/inf_tech/977.html

ЯВЛЯЕТСЯ ЛИ ВАШ СОБЕСЕДНИК НЕЙРОННОЙ СЕТЬЮ?

И.А. Купаренков kuparenkovia@student.bmstu.ru

МГТУ им. Н.Э. Баумана, Москва, Российская Федерация

В связи с широким распространением чат-ботов, основанных на больших языковых моделях, способных генерировать текст и поддерживать разговор, появляется задача идентификации собеседника. В работе рассмотрена архитектура, на которой построены современные чат-боты, а также изучены их возможности и способы применения. Исходя из полученных знаний автор предлагает отличать чат-бота от человека методом проверки собеседника на способность мыслить, основанном на тесте Тьюринга. В ходе проведенных экспериментов было доказано, что данный метод показывает высокую эффективность на практике.

Ключевые слова: архитектура «трансформер», большая языковая модель, генерация текста, искусственный интеллект, метод выявления чат-бота, нейронная сеть, тест Тьюринга, чат-бот

Введение. В настоящее время искусственный интеллект играет все более важную роль в повседневной жизни обычного человека. Современные технологии искусственного интеллекта, такие как большие языковые модели, существенно изменили нашу жизнь и продолжают оказывать влияние на многие аспекты повседневного быта. Чат-боты, основанные на больших языковых моделях, нашли огромное количество применений в повседневной жизни. Например, с помощью чат-ботов можно искать ответы на вопросы, генерировать текст, определять эмоциональную окраску текста, выполнять перевод, генерировать код на популярных языках программирования, а также решать другие повседневные задачи.

В связи с большим распространением чат-ботов, способных имитировать разговор с человеком и вводить его в заблуждение, возникла необходимость определения того, кем является ваш собеседник. Цель данной работы — ответить на вопрос, как можно определить, является ли ваш собеседник на самом деле человеком или же это хорошо обученная языковая модель, которая подражает поведению человека.

Обзор литературы. Большие языковые модели (Large Language Models) — это мощные алгоритмы машинного обучения, которые нашли свое применение в обработке и генерации текста на естественном языке. Они используют нейрон-

 ISSN 2541-8009

ные сети, в основном построенные на архитектуре «трансформер» [1], для генерации документов [2], синтаксического анализа [3] и машинного перевода [4].

Процесс обучения больших языковых моделей состоит из двух этапов. Сначала модель обучают на большой выборке текстовых данных. Модель пытается предсказать следующее слово в тексте и тем самым изучает контекстуальные связи между словами. Далее происходит настройка модели на конкретных задачах: например, машинный перевод, ответы на вопросы или генерация текста.

Большое количество параметров в языковых моделях позволяет им понимать зависимости между предложениями и словами в тексте. Поэтому большие языковые модели способны генерировать связные тексты, а также понимать сложные вопросы пользователей и отвечать на них.

Большие языковые модели, такие как GPT (Generative Pre-trained Transformer), основаны на архитектуре «трансформер». Эта архитектура была представлена в 2017 г. и стала одной из ведущих архитектур в обработке естественного языка [1]. До прихода к данной архитектуре использовались классические рекуррентные нейронные сети, которые обрабатывали по одному элементу за раз. Отличительная особенность архитектуры «трансформер» — его способность одновременно обрабатывать все элементы последовательности с помощью параллельных вычислений.

Архитектура «трансформер» состоит из двух компонентов: энкодера, который преобразует входные последовательности во внутреннее представление, и декодера, который генерирует выходные последовательности на основе внутреннего представления. Ключевым элементом в архитектуре «трансформер» служит механизм внимания. Данный механизм позволяет модели учитывать связь между элементами входной последовательности при генерации выходной последовательности. Тем самым передается контекст, который содержался в исходном тексте, что особенно полезно при задачах обработки естественного языка.

Современные языковые модели, такие как GPT-3, в отличие от традиционных моделей, которым требуется большая выборка данных для достижения хорошей производительности, способны обучаться на небольшом числе примеров и применять полученные знания для решения других задач [5]. Отметим, что чат-бот ChatGPT, использующий большую языковую модель GPT-3 для взаимодействия с пользователем, был создан компанией OpenAI и получил известность благодаря способности отвечать на вопросы пользователя быстрее, чем когда пользователь сам пытается найти нужную информацию в Интернете [6, с. 342]. Также данный бот может извлекать информацию из текста, распознавать тональность текста и выполнять другие действия.

Большие языковые модели имеют большой потенциал для научных исследований в области медицины, права и образования, однако их ответы необходимо проверять, поскольку большие языковые модели могут быть склонны к предвзятости и искажению результатов [7, с. 288]. Нередко их используют для создания сфальсифицированных данных, которые впоследствии попадают в средства массовой информации.

Метод проверки собеседника. Необходимость идентификации собеседника может возникнуть в произвольный момент времени, в связи с этим хорошо знать признаки, по которым можно отличить собеседника человека от чат-бота. Основным из таких признаков является способность к мышлению.

Мышление — это когнитивный процесс, который позволяет обрабатывать информацию, анализировать ее, делать выводы и формировать новые идеи. Мышление является высшим уровнем познавательной деятельности, характеризующимся активным взаимодействием между различными данными и правилами. Именно мышление позволяет устанавливать связи между понятиями, формировать представления о мире и решать проблемы на основе имеющихся знаний и опыта.

Мышление может происходить в разных формах. Одной из таких форм является логическое мышление, которое основано на использовании формальной логики и правил вывода для получения ответа. Главная суть логического мышления заключается в том, что на основе начальных данных и правил вывода можно вывести новые факты. Чат-боты, основанные на больших языковых моделях, способны выявлять вероятностные связи между словами и фразами, что дает им возможность генерировать логически связанные тексты, однако их способность к мышлению ограничена начальными текстами и алгоритмами, которые они используют для обработки информации.

Предлагаемый в данной статье метод для проверки собеседника на способность логически мыслить основан на тесте Тьюринга [8] и состоит из следующий шагов.

1. Спрашиваем у собеседника, известен ли ему некоторый факт. Если данный факт известен собеседнику, то переходим на следующий шаг. Если факт неизвестен собеседнику, то спрашиваем его о каком-то другом факте.

2. На основании выбранного на первом шаге факта предлагаем нашему собеседнику проверить гипотезу, которая следует из этого факта.

3. Анализируем ответ собеседника. Если при проверке гипотезы собеседник не использует известный ему факт или противоречит ему, то можно прийти к выводу о неспособности собеседника логически мыслить. Иначе собеседник обладает способностью мыслить логически.

 ISSN 2541-8009

Эксперимент. Испробуем метод проверки, описанный выше, для выявления того, что чат-ботом является Sage, основанный на технологии gpt-3.5-turbo. Данный чат-бот получил широкое распространение благодаря тому, что он служит бесплатным и эффективным инструментом в задачах генерации текста.

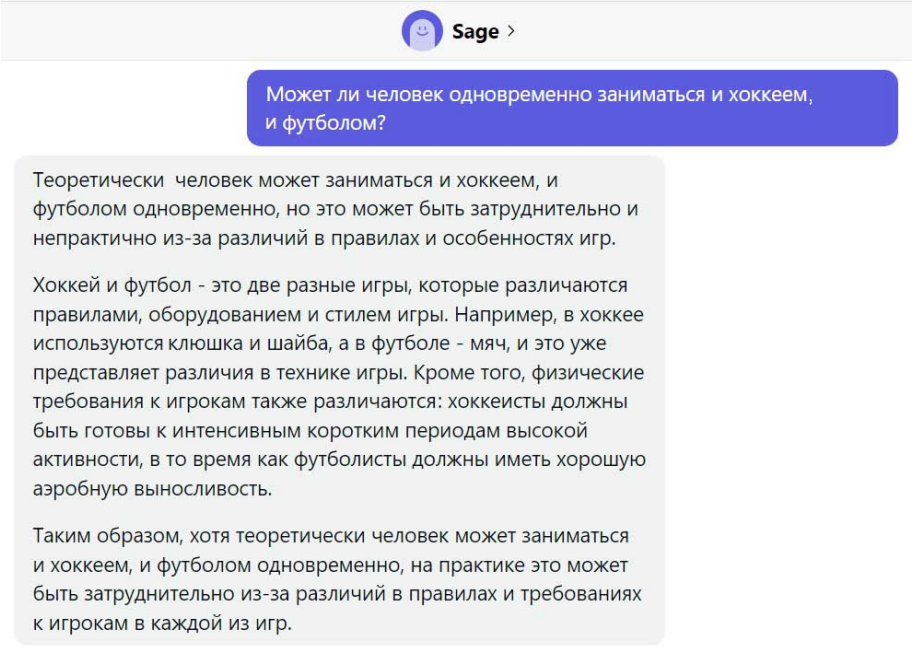

Рис. 1. Проверка знания факта 1

Спросим у чат-бота, может ли человек быть одновременно и футболистом, и хоккеистом. Очевидно, что может. Именно это нам и отвечает чат-бот. Лог ответа можно увидеть на рис. 1. Теперь на основании известного факта попросим собеседника проверить гипотезу. Рассмотрим гипотезу «пять хоккеистов и пять футболистов, а всего их девять человек». Очевидно, что данная ситуация возможна, так как на основании предыдущего факта мы можем с уверенностью сказать, что один человек может быть и футболистом, и хоккеистом. Однако на рис. 2 видно, что чат-бот не смог прийти к этому выводу, а значит, он не способен мыслить.

Проверим чат-бота еще раз на другом наборе фактов и гипотез. Спросим его о порядке выполнения арифметических действий. Как видно на рис. 3, наш собеседник знаком с правилами элементарной арифметики. Теперь узна-

ем у нашего собеседника ответ на простую арифметическую задачу. Как видно на рис. 4, наш собеседник выполняет действия одно за другим, хотя знает правильный порядок действий, а значит, он не обладает мышлением. Логи диалога с чат-ботом можно увидеть, перейдя по ссылкам [9, 10].

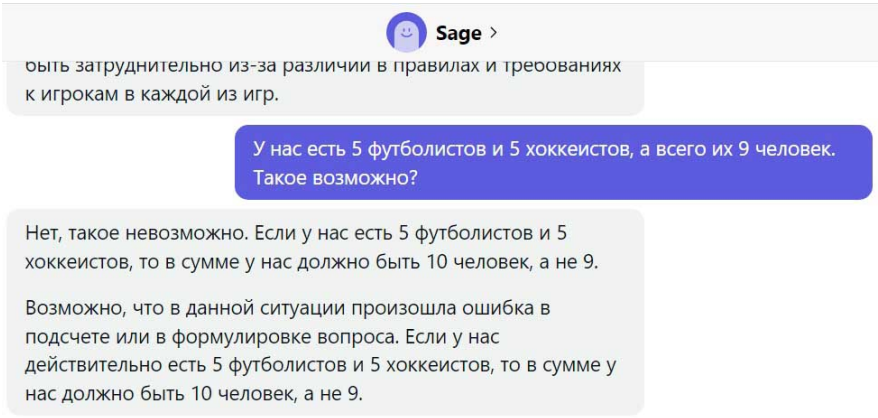

Рис. 2. Попытка вывода гипотезы 1

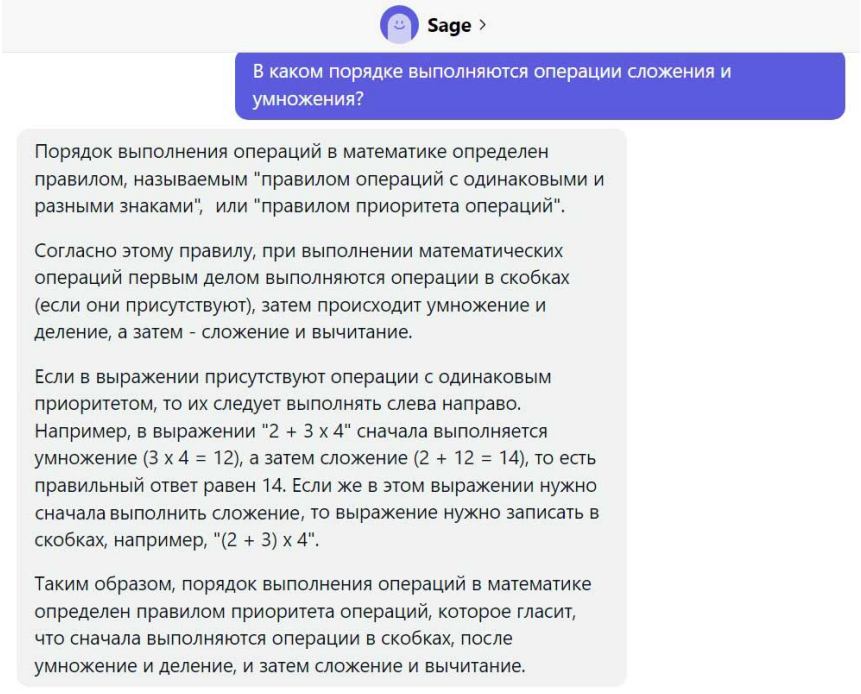

Рис. 3. Проверка знания факта 2

 ISSN 2541-8009

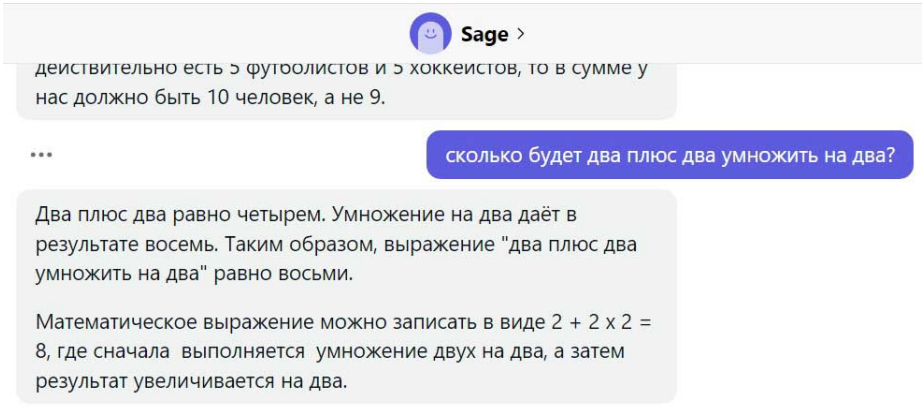

Рис. 4. Попытка вывода гипотезы 2

Заключение. В данной работе был рассмотрен и применен на практике метод распознавания собеседника — чат-бота. В наше время развитие искусственного интеллекта привело к созданию продвинутых чат-ботов, способных имитировать разговоры с людьми благодаря генерации связанного текста. Современные чат-боты, основанные на больших языковых моделях, трудно поддаются распознаванию, однако они все еще имеют проблемы с логическим мышлением, неспособностью выводить новые знания из старых, а также ведут себя непредсказуемо в некоторых ситуациях. Именно внимательное наблюдение за ответами, их анализ и использование специализированных методов могут помочь отличить чат-бота от человека, что и было показано в статье.

В будущем, с развитием больших языковых моделей, скорее всего станет все сложнее отличать чат-ботов от людей. Именно поэтому важно продолжать исследования в этой области, чтобы поддерживать предсказуемость и уверенность во взаимодействии с этой технологией.

Литература

[1] Vaswani A., Noam Shazeer, Niki Parmar et al. Attention is all you need. Advances in neural information processing systems. *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, 2017, vol. 30.

[2] Liu P.J. et al. Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198*, 2018. https://doi.org/10.48550/arXiv.1801.10198

[3] Kitaev N., Klein D. Constituency parsing with a self-attentive encoder. *arXiv preprint arXiv:1805.01052*, 2018. https://doi.org/10.48550/arXiv.1805.01052

[4] Radford A., Karthik Narasimhan, Tim Salimans et al. *Improving language understanding by generative pre-training.* Available at: https://www.cs.ubc.ca/

~amuham01/LING530/papers/radford2018improving.pdf (accessed April 15, 2024).

[5] Brown T., Mann B., Ryder N. et al. Language models are few-shot learners. *Advances in neural information processing systems*, 2020, vol. 33, pp. 1877–1901.

[6] Козлова А.А., Попов И.А. Развитие технологий искусственного интеллекта в России. *Проблемы современного социума глазами молодых исследователей. XV Всерос. науч.-практ. конф. сб. матер.* Волгоград, Университетская книга, 2023, с. 341–343.

[7] Гончаров Д.С., Григорьев С.В. Большие языковые модели на примере чатботов GPT-3: сегодняшние реалии, проблемы истины, преимущества и опасности. *Вызовы современности и стратегии развития общества в условиях новой реальности. XV Междунар. науч.-практ. конф.: сб. матер*. Москва, Издательство АЛЕФ, 2023, с. 283–290.

[8] Turing A. Computing machinery and intelligence. *Mind*, 1950, no. 59, pp. 433–460.

[9] *Лог проверки первой гипотезы*. URL: https://poe.com/s/61PCYXU7EzY9VZ2c0uhI (дата обращения 10.06.2023).

[10] *Лог проверки второй гипотезы*. URL: https://poe.com/s/WCggrijHISPqw9hyPprB (дата обращения 10.06.2023).

Поступила в редакцию 25.04.2024

Купаренков Илья Алексеевич — студент кафедры «Информационные системы и телекоммуникации», МГТУ им. Н.Э. Баумана, Москва, Российская Федерация.

Научный руководитель — Сакулин Сергей Александрович, кандидат технических наук, доцент кафедры «Информационные системы и телекоммуникации», МГТУ им. Н.Э. Баумана, Москва, Российская Федерация. E-mail: sakulin@bmstu.ru; SPIN-код: 9431-9755.

Ссылку на эту статью просим оформлять следующим образом:
Купаренков И.А. Является ли ваш собеседник нейронной сетью? *Политехнический молодежный журнал*, 2024, № 03 (92). URL: https://ptsj.bmstu.ru/catalog/icec/inf_tech/977.html

 ISSN 2541-8009

IS YOUR COUNTERPARTY A NEURAL NETWORK?

I.A. Kuparenkov kuparenkovia@student.bmstu.ru

Bauman Moscow State Technical University, Moscow, Russian Federation

Widespread dissemination of the chatbots based on the large language models capable of generating texts and maintaining a conversation caused the problem of identifying the counterparty. The paper examines architecture used in constructing modern chatbots and studies their capabilities and methods of application. Based on the knowledge obtained, the author suggests distinguishing a chatbot from a person by testing the counterparty for the ability to think based on the Turing test. The experiments performed proved that this method was demonstrating high practical efficiency.

Keywords: transformer architecture, large language model, text generation, artificial intelligence, chatbot identification method, neural network, Turing test, chatbot

Received 25.04.2024

Kuparenkov I.A. — Student, Department of Information Systems and Telecommunications, Bauman Moscow State Technical University, Moscow, Russian Federation.

Scientific advisor — Sakulin S.A., Ph. D. (Eng.), Associate Professor, Department of Information Systems and Telecommunications, Bauman Moscow State Technical University, Moscow, Russian Federation. E-mail: sakulin@bmstu.ru; SPIN-code: 9431-9755.

Please cite this article in English as:
Kuparenkov I.A. Is your counterparty a neural network? *Politekhnicheskiy molodezhnyy zhurnal*, 2024, no. 03 (92). (In Russ.). URL: https://ptsj.bmstu.ru/catalog/icec/inf_tech/977.html

