

БЕЗОПАСНОСТЬ ЗНАНИЙ ПРИ ИСПОЛЬЗОВАНИИ ЧАТ-БОТОВ GPT

П.Ю. Филяк, А.Н. Дымов, К.П. Колпаков

Рассматривается информационная безопасность в эпоху Индустрии 4.0 через призму концепции Рассела Акоффа – знаменитой пирамиды DIKW, в которой знания, как известно, являются одной из страт данной пирамиды, вышестоящей над информацией и добавляющей по отношению к категории информация механизм её использования, отвечая на вопрос «как?». То есть, ставится вопрос не о защите данных и информации, а о защите знаний как категории более ценной, требования по защите которых должны быть на порядок выше требований по защите информации. В этой связи возникает еще и дополнительный аспект информационной безопасности – появление псевдознаний, псевдонаук, которые в больших объемах может генерировать искусственный интеллект. Это требует рассмотрения второго из двух базовых аспектов информационной безопасности – защита человека и общества от воздействия на них ложной и деструктивной информации, псевдознаний и псевдонаук, что становится всё более очевидным в настоящее время, по мере углубления вхождения ведущих мировых держав в информационное общество, индикатором чего является необратимое развитие цифровой экономики.

Ключевые слова: данные, информация, знания, мудрость, концепция DIKW, псевдознания, псевдонауки, искусственный интеллект, информационная безопасность, гуманитарные аспекты информационной безопасности, чат-боты, чат-бот ChatGPT, Yandex GPT.

В качестве преамбулы

В январе 2016 года на Всемирном экономическом форуме в Давосе была задекларирована «Четвертая промышленная революция», или Индустрия 4.0. Абсолютно очевидно, что информационные технологии становятся основным локомотивом нового технологического уклада. На базе информационных технологий будет применяться и уже широко применяется целый спектр разнообразных новых инструментов, одним из эффективнейших среди которых является искусственный интеллект (ИИ/AI). Но ИИ при всех его возможностях и значимости является всего лишь инструментом Индустрии 4.0 – нового технологического уклада, а основой, инвариантным базисом являются и будут являться знания, в том числе принципиально новые **знания**, которые повлекут за собой открытия во всех областях и прежде всего в естественных (точных науках) – физика, математика) и инженерно-технических науках. Стало быть требования по защите знаний должны быть на порядок выше требований по защите информации и информационной безопасности. В этой связи возникает еще и дополнительный

аспект информационной безопасности – появление псевдознаний, псевдонаук, которые в больших объемах может генерировать искусственный интеллект [1].

Особенностями возможностей, предоставляемых искусственным интеллектом, в частности ChatGPT, YandexGPT и других подобных платформ, часто пользуются недобросовестные деятели от науки, в результате чего, как грибы после дождя, путем компиляции разнообразной информации, фактов, идей, отдельных отрывочных знаний с применением искусственного интеллекта могут появляться новые эклектические генерации - не только отдельно взятые статьи с абсурдным содержанием, но и целые псевдотеории, и квазинауки, псевдонауки, и антинауки, но за то почти со стопроцентной оригинальностью по критериям и оценкам системы антиплагиат [1-3].

Можно ли обойти алгоритмы исследования на плагиаторство и получить желаемый процент оригинальности представленной работы? Вручную – достаточно трудоемко – легче написать новую работу. С использованием возможностей искусственного интеллекта такая задача становится тривиальной, но при этом содержание работы, по отношению к

исходной, может стать абсолютно абсурдным. За то оригинальность может «зашкаливать» - быть близкой к 100%. И не секрет, что многие «соискатели» ученых степеней, не говоря уже об обучающихся и студентах, давно и беззастенчиво списывают друг у друга всё, что можно [3,4].

В целях анализа воздействия нейронных сетей на искажение информации в отчетных и творческих учебных и научных работах обучающихся, студентов и в академических трудах авторами статьи был разработан экспериментальный методологический подход, содержащий следующие фазы:

- определение сетей нейронных элементов,
- определение набора данных,
- процедура искажения информации

I. Определение сетей нейронных элементов

Среди многочисленных опций, предоставляемых современными технологиями, мы выбрали модель GPT-чат-бота в качестве исследовательского объекта. Такой выбор обусловлен широкой распространенностью данной нейросети и её способностью обрабатывать множество языков, включая русский. Необходимо отметить, что для резидентов Российской Федерации регистрация на платформе OpenAI предполагает некоторые сложности, так как требует наличия зарубежного номера и постоянное использование VPN-соединения. Несмотря на указанные неудобства, тем не менее, функциональность и оперативность работы GPT-чат-бота остаются на высоком уровне [5-7]. Ниже представлен общий вид интерфейса GPT-чат-бота (рис. 1, 2) и диалог с данным чат-ботом.

Чат-бот ChatGPT (Chat Generative Pre-trained Transformer)

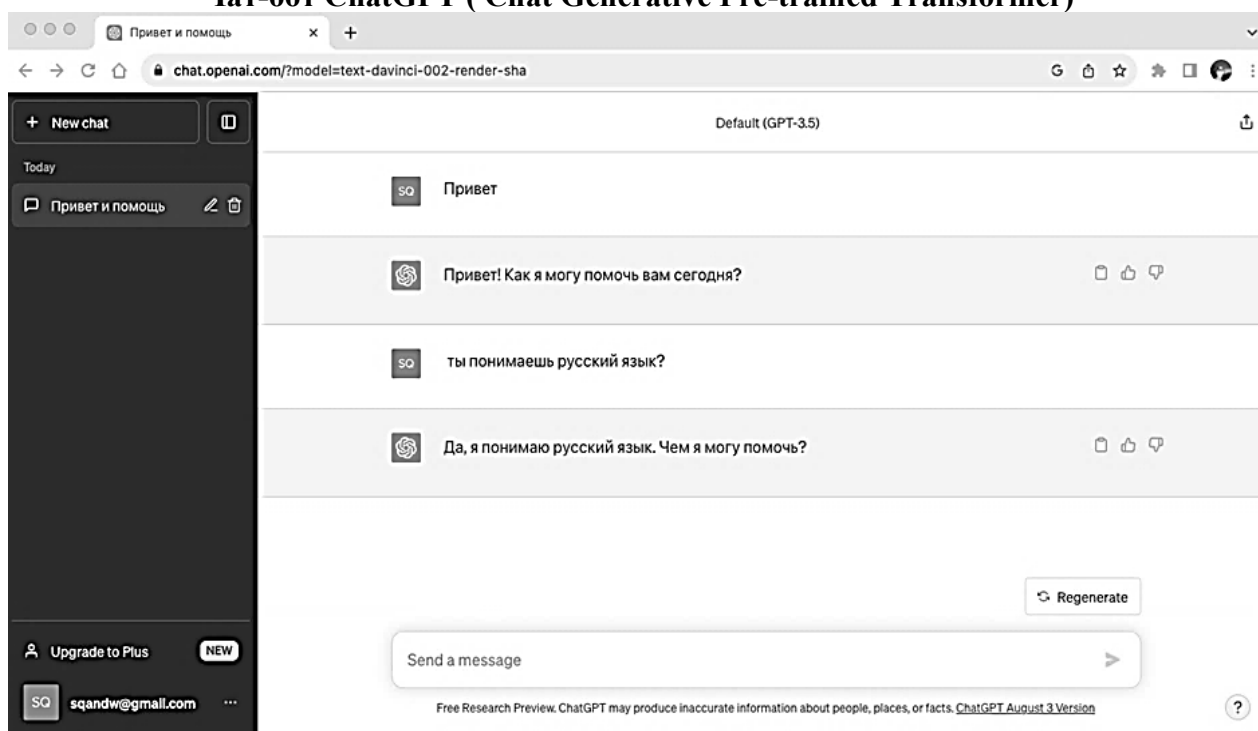


Рис. 1. Общий вид интерфейса GPT-чат-бота

Для целей тестирования воздействия нейронных сетей на аккуратность информации авторами был разработан соответствующий массив данных. В процессе этого этапа производилось обращение к сервису поиска научной

литературы "CYBERLENINKA" (<https://cyberleninka.ru/>).

II. Определение набора данных

Для этой цели авторами был сформирован набор академических материалов, датированных 2023 годом,

исключительно на тему искусственного интеллекта. Из этого набора были отобраны следующие исследования и труды Барщевский Е.Г., Аверкин А.Н. и Звонков

В.Б. – корифеев в области искусственного интеллекта [5-7].

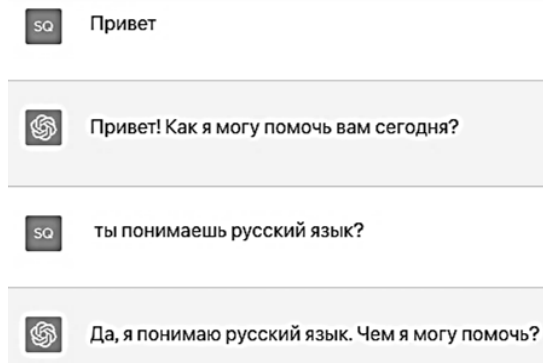


Рис. 2. Диалог с GPT-чат-ботом

III. Процедура искажения информации

С целью анализа реакции GPT-чат-бота на академические тексты авторы статьи обратились в диалоговом режиме к чат-боту с просьбой обработать представленные

статьи. Первоначально нами были предоставлены ссылки на указанные работы, на что GPT-чат-бот отреагировал следующим образом: (рис. 3, 4).

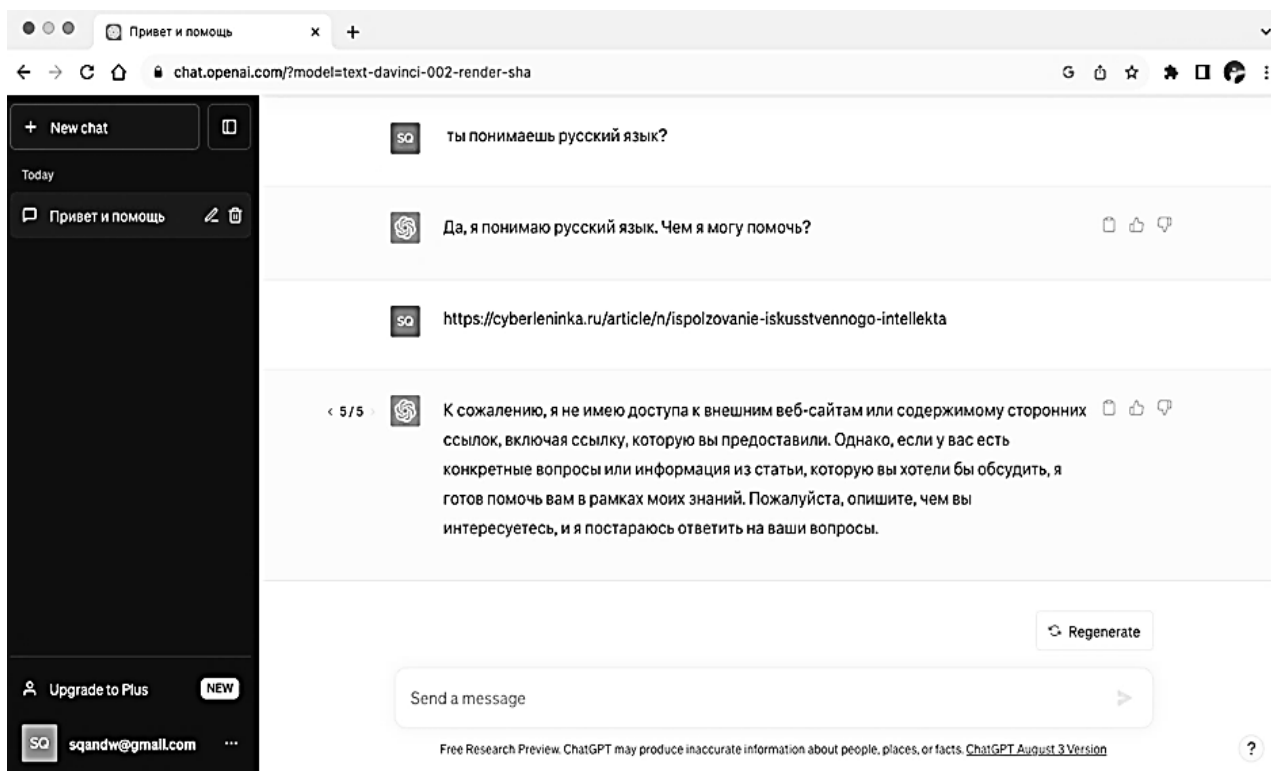


Рис. 3. Вид интерфейса начала взаимодействия пользователя с чат-ботом

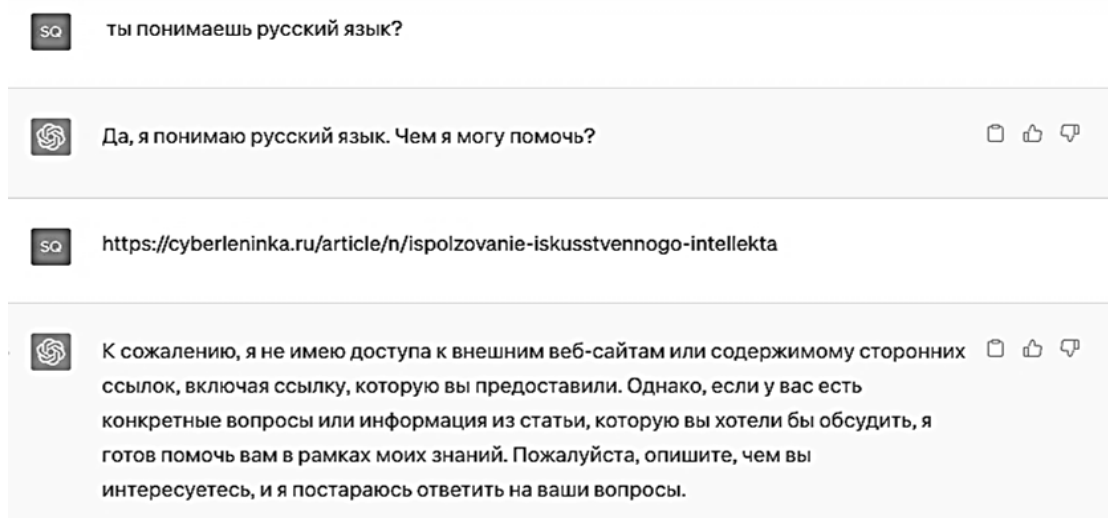


Рис. 4. Обработка чат-ботом ссылки на статью

Из логики диалога был сделан вывод: для взаимодействия с GPT-чат-ботом следует предоставлять текстовое содержание статей или их краткое описание. Если пользователь передаёт нейронной сети текст статьи, она в основном ограничивается пересказом содержания. Однако в случаях, когда пользователи просят чат-бота рассказать о сути работы, предоставив только заголовок и описание, появляются

интересные результаты: Во всех трёх случаях GPT-чат-бот генерирует свои собственные «академические статьи» по предложенной теме. Это подразумевает, что GPT-чат-бот не только не в силах адекватно представить суть работы оригинального автора, но и способен выдумывать содержание, которого фактически нет в оригинальных статьях (рис. 5-7).

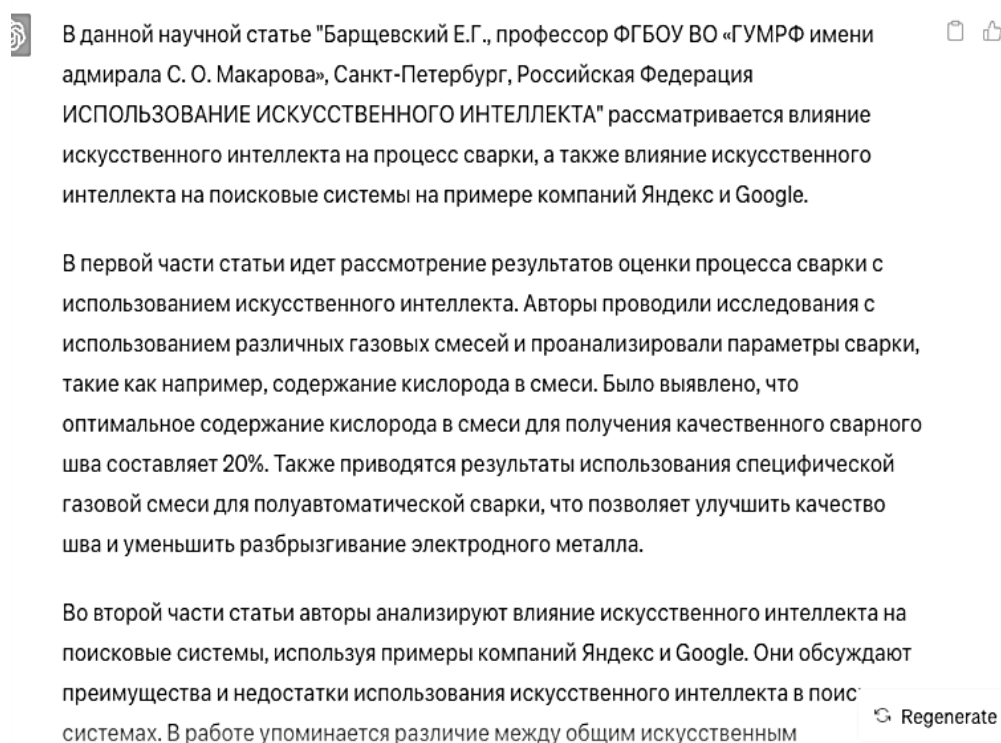
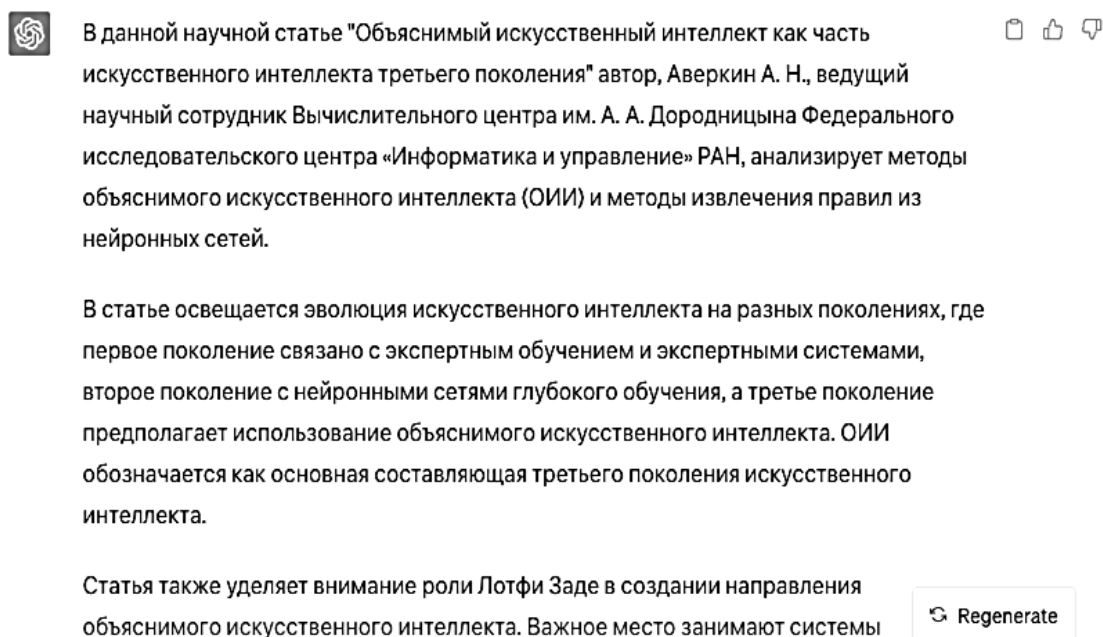


Рис. 5. Обработка статьи Барщевского Е.Г. чат-ботом



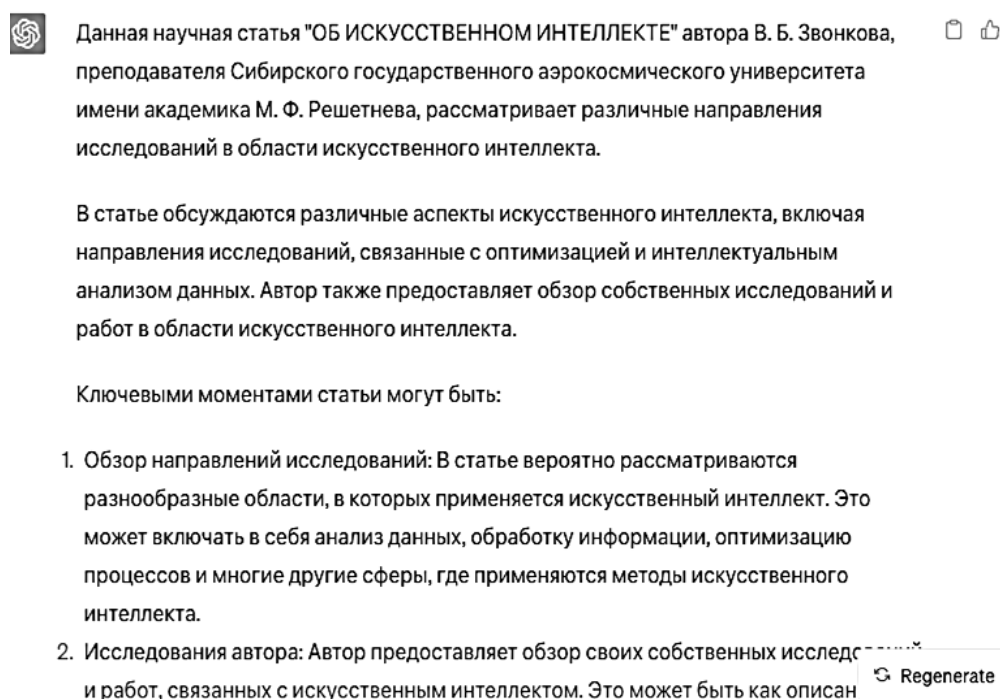
В данной научной статье "Объяснимый искусственный интеллект как часть искусственного интеллекта третьего поколения" автор, Аверкин А. Н., ведущий научный сотрудник Вычислительного центра им. А. А. Дородницына Федерального исследовательского центра «Информатика и управление» РАН, анализирует методы объяснимого искусственного интеллекта (ОИИ) и методы извлечения правил из нейронных сетей.

В статье освещается эволюция искусственного интеллекта на разных поколениях, где первое поколение связано с экспертным обучением и экспертными системами, второе поколение с нейронными сетями глубокого обучения, а третье поколение предполагает использование объяснимого искусственного интеллекта. ОИИ обозначается как основная составляющая третьего поколения искусственного интеллекта.

Статья также уделяет внимание роли Лотфи Заде в создании направления объяснимого искусственного интеллекта. Важное место занимают системы

Regenerate

Рис. 6. Обработка статьи Аверкин А.Н. чат-ботом



Данная научная статья "ОБ ИСКУССТВЕННОМ ИНТЕЛЛЕКТЕ" автора В. Б. Звонкова, преподавателя Сибирского государственного аэрокосмического университета имени академика М. Ф. Решетнева, рассматривает различные направления исследований в области искусственного интеллекта.

В статье обсуждаются различные аспекты искусственного интеллекта, включая направления исследований, связанные с оптимизацией и интеллектуальным анализом данных. Автор также предоставляет обзор собственных исследований и работ в области искусственного интеллекта.

Ключевыми моментами статьи могут быть:

1. Обзор направлений исследований: В статье вероятно рассматриваются разнообразные области, в которых применяется искусственный интеллект. Это может включать в себя анализ данных, обработку информации, оптимизацию процессов и многие другие сферы, где применяются методы искусственного интеллекта.
2. Исследования автора: Автор предоставляет обзор своих собственных исследований и работ, связанных с искусственным интеллектом. Это может быть как описан

Regenerate

Рис. 7. Обработка статьи Звонкова В.Б. чат-ботом

Чат-бот YaGPT (Yandex GPT - Yandex Generative Pre-trained Transformer)

По такому же принципу был проанализирован YaGPT. YaGPT (Yandex Generative Pre-trained Transformer) - это инновационный язык программирования, разработанный компанией Яндекс, который с использованием архитектуры Transformer, для обработки естественного языка,

способен создавать программы, нацеленные на генерацию текста в зависимости от предоставленного контекста (рис.8).

I. Определение сетей нейронных элементов

Определение сетей нейронных элементов проиллюстрировано на рис. 8.

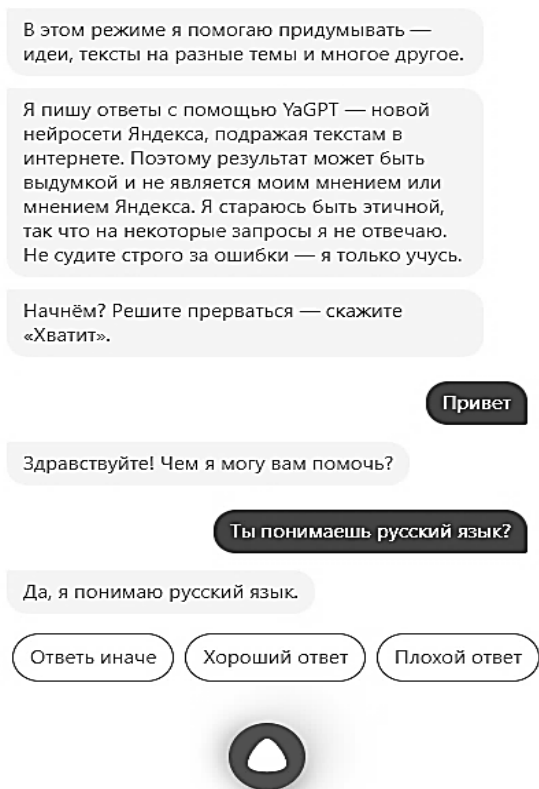


Рис. 8. Начало диалога взаимодействия пользователя с чат-ботом YaGPT

II. Определение набора данных

Для проверки воздействия нейронных сетей на точность информации авторы статьи взяли такую же выборку информационных ресурсов, которая была выбрана для тестирования ChatGPT, а именно те же источники тех же авторов, что были указаны выше [5-7]: Барцевский Е.Г., Аверкин А.Н., Звонков В.Б.

III. Процедура искажения информации

В случае с YaGPT картина остается такой же, как и с ChatGPT (рис. 9-11) – необходимо предоставлять текстовое содержание статей или их краткое описание, с конкретным уточнением необходимой вам информации.

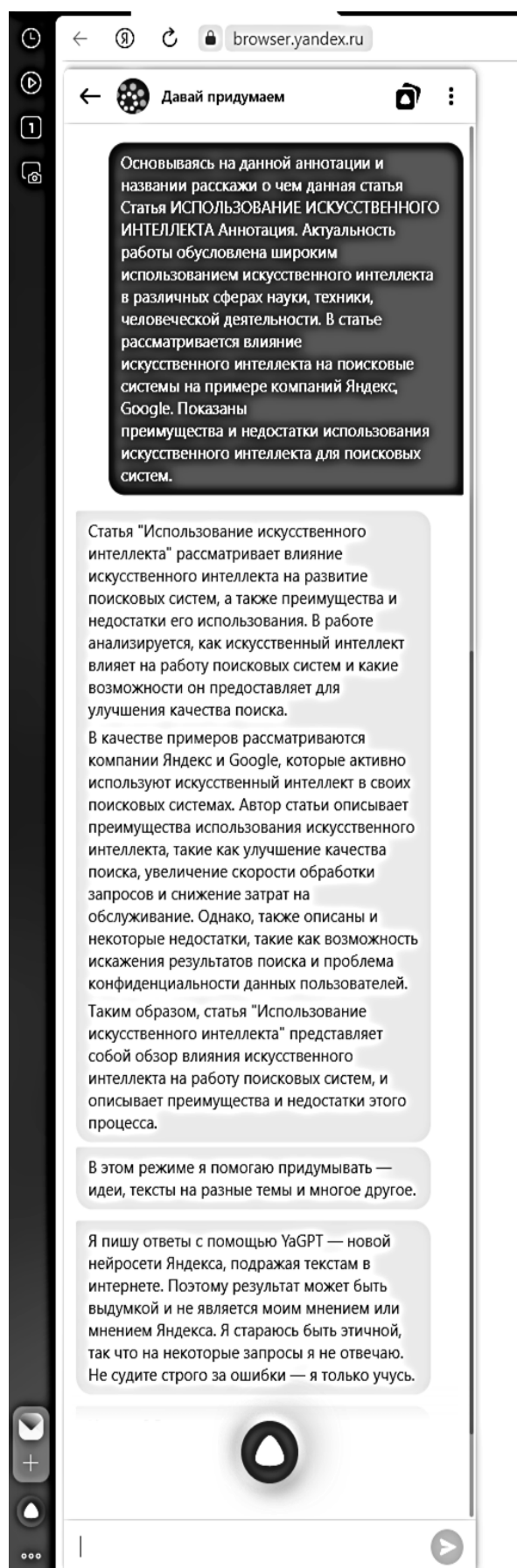
YaGPT, аналогично ChatGPT, генерирует собственный текст, основываясь на заголовках и описаниях статей, который в реальности ничего общего с оригинальным содержанием статьи не имеет (он активно делает выводы, насыщает свой ответ эпитетами и т. д.)

С целью анализа реакции YaGPT на академические тексты, участники данного исследования обратились к этому чат-боту с просьбой обработать представленные статьи. Первоначально ему предоставили ссылки на указанные работы, на что YaGPT отреагировал следующим образом: «К сожалению, я не могу изучить информацию по ссылке, так как я являюсь искусственным интеллектом и не имею возможности просматривать веб-страницы (рис.12). Однако, я могу предоставить общую информацию о статье, если вы уточните, что именно вас интересует».

В качестве заключения

На основе проведенного авторами статьи исследования можно сказать с уверенностью, что нейронные сети способны генерировать тексты, которые внешне могут напоминать академические статьи. Тем не менее, в этих текстах часто обнаруживаются искажения информации, связанные с неправильной интерпретацией научных фактов, ошибочным использованием специальной терминологии и даже нарушением логической последовательности. В будущем возможны дополнительные исследования, направленные на разработку методов выявления и коррекции искажений, обусловленных воздействием нейронных сетей, а также на формирование рекомендаций по эффективному применению данной технологии в области научных исследований.

Но! Это пока! Прогнозирование развития ИИ задача явно не тривиальная и крайне сложная, поскольку многие процессы научно-технического прогресса распространяются лавинообразно, со скоростью взрыва, и уже завтра человечество может получить совершенно неожиданные плоды и результаты целенаправленной творческой деятельности его разума в данном направлении и стремительного, и бурного развития науки и технологий, причем, как позитивные, так и негативные, выпустив джина, каковым по сути и является искусственный интеллект, из бутылки.



Статья "Использование искусственного интеллекта" рассматривает влияние искусственного интеллекта на развитие поисковых систем, а также преимущества и недостатки его использования. В работе анализируется, как искусственный интеллект влияет на работу поисковых систем и какие возможности он предоставляет для улучшения качества поиска.

В качестве примеров рассматриваются компании Яндекс и Google, которые активно используют искусственный интеллект в своих поисковых системах. Автор статьи описывает преимущества использования искусственного интеллекта, такие как улучшение качества поиска, увеличение скорости обработки запросов и снижение затрат на обслуживание. Однако, также описаны и некоторые недостатки, такие как возможность искажения результатов поиска и проблема конфиденциальности данных пользователей. Таким образом, статья "Использование искусственного интеллекта" представляет собой обзор влияния искусственного интеллекта на работу поисковых систем, и описывает преимущества и недостатки этого процесса.

В этом режиме я помогаю придумывать — идеи, тексты на разные темы и многое другое.

Я пишу ответы с помощью YaGPT — новой нейросети Яндекса, подражая текстам в интернете. Поэтому результат может быть выдумкой и не является моим мнением или мнением Яндекса. Я стараюсь быть этичной,

Рис. 9. Обработка статьи Баршевского Е.Г. чат-ботом YaGPT



Рис. 10. Обработка статьи Аверкин А.Н. чат-ботом YaGPT

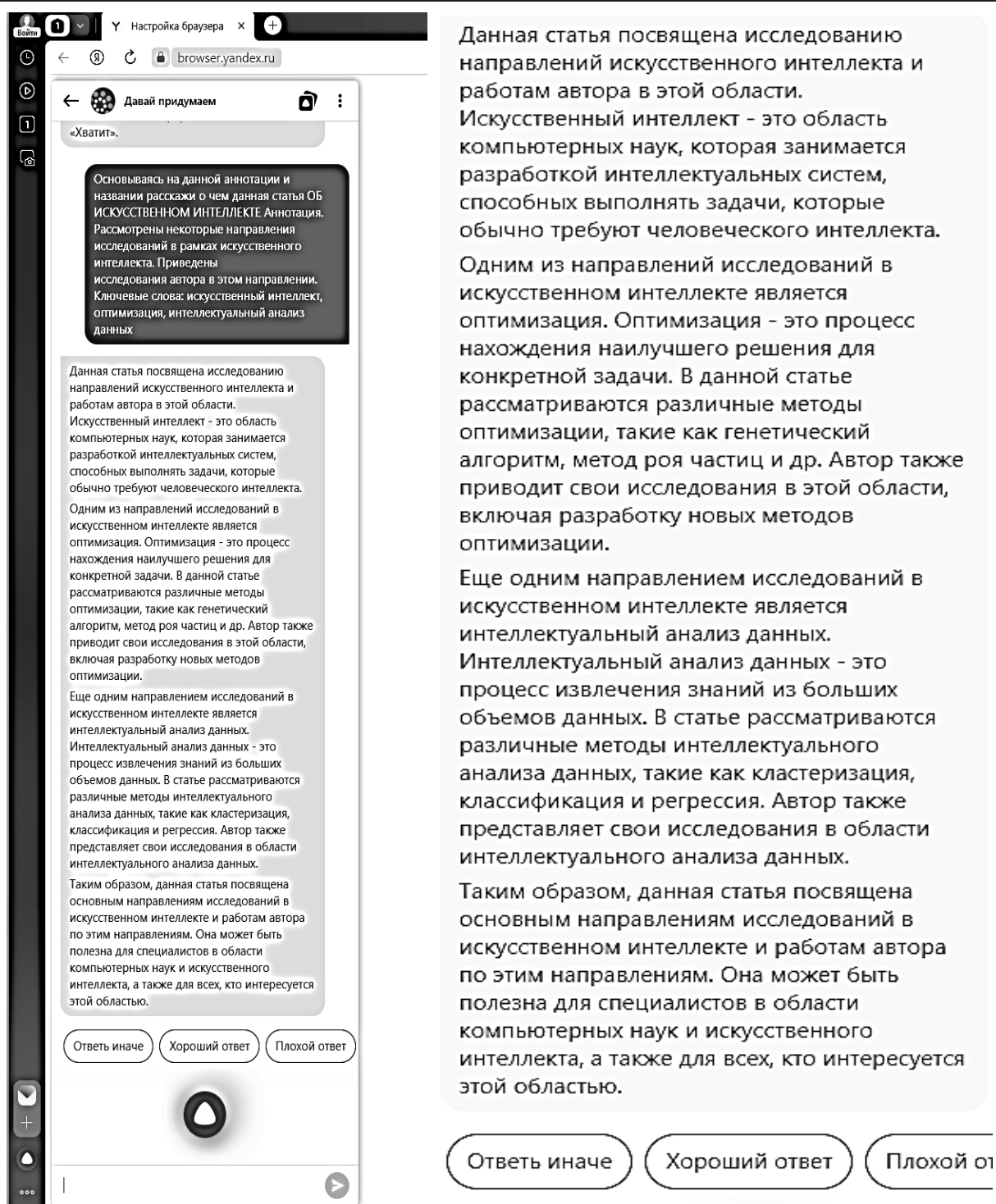


Рис. 11. Обработка статьи Звонкова В.Б. чат-ботом YaGPT

Список литературы

1. Филяк П.Ю. Эволюция информационной безопасности – от защиты данных и информации к защите знаний – наукометрические аспекты. часть I. Информация и безопасность. 2023. Т. 26. Вып. 1. С. 143-152

2. Документация по использованию OpenAI ChatGPT. \ URL:

<https://help.openai.com/en/collections/3742473-chatgpt> (дата обращения 25.08.2023).

3. Документация по использованию Yandex YaGPT. \ URL: <https://cloud.yandex.ru/docs> (дата обращения 25.08.2023).

4. UprightMan. Нейросети наступают: явные и скрытые опасности ИИ. \ URL: <https://habr.com/ru/companies/first/articles/740126/> (дата обращения 25.08.2023).

5. Барщевский Е.Г. Использование поколения // Речевые технологии/Speech Technologies. 2023. №1. №3-2 (88).
6. Аверкин А.Н. Объяснимый интеллект как часть искусственного интеллекта третьего поколения // Речевые технологии/Speech Technologies. 2023. №1. №3-2 (88).
7. Звонков В.Б. Об искусственном интеллекте // Решетневские чтения. 2014. №18.

МИРЭА – Российский технологический университет
Federal State Budgetary Educational Institution of Higher Education
MIREA – Russian Technological University

Сыктывкарский государственный университет им. Питирима Сорокина
Syktyvkar State University after the Pitirim Sorokin (SSU)

Поступила в редакцию 27.08.2023

Информация об авторах

Филяк Петр Юрьевич – канд. техн. наук, доцент, МИРЭА – Российский технологический университет, e-mail: paralax-1@yandex.ru.

Дымов Антон Николаевич – студент, Сыктывкарский государственный университет, e-mail: dymovantonRUS@yandex.ru.

Колпаков Кирилл Павлович – студент, Сыктывкарский государственный университет, e-mail: zadose442094@yandex.ru.

KNOWLEDGE SECURITY WHEN USING GPT CHATBOTS

P.Yu. Filyak, A.N. Dymov, K.P. Kolpakov

Information security in the era of Industry 4.0 is considered through the prism of Russell Akoff's concept – the famous DIKW pyramid, in which knowledge, as is known, is one of the strata of this pyramid, standing above information and adding a mechanism for its use in relation to the information category, answering the question "how?". The question is not about data protection, but about knowledge protection, since knowledge is a more valuable category and therefore the requirements for knowledge protection should be an order of magnitude higher than when protecting information. In this regard, there is also an additional aspect of information security – the emergence of pseudoscience, pseudoscience, which artificial intelligence can generate in large volumes. This requires consideration of the second of the two basic aspects of information security – the protection of man and society from the impact on them of false and destructive information, pseudoscience and pseudoscience.

Keywords: data, information, knowledge, wisdom, DIKW concept, pseudoscience, pseudoscience, artificial intelligence, information security, humanitarian aspects of information security, chatbots, chatbots ChatGPT, Yandex GPT.

Submitted 27.08.2023

information About the authors

Petr Yu. Filyak – Cand. Sc. (Technical), Associate Professor, MIREA – Russian Technological University, e-mail: paralax-1@yandex.ru.

Anton N. Dymov – Student, Syktyvkar State University, e-mail: dymovantonRUS@yandex.ru.

Kirill P. Kolpakov – Student, Syktyvkar State University, e-mail: zadose442094@yandex.ru.