Дистанционный конкурс научно-исследовательских работ обучающихся общеобразовательных организаций

«Физика без границ», «География без границ», «Математика без границ»,

«Астрономия без границ», «Химия без границ», «Биология без границ»,

«Информатика без границ», «Технология без границ»

Научно-исследовательская работа

«CHAT GPT: АНАЛИЗ ЭФФЕКТИВНОСТИ И ЭТИЧЕСКИЕ АСПЕКТЫ ИСПОЛЬЗОВАНИЯ»

Подготовил:

Матвеева Софья Юрьевна,

обучающийся 8Б класса

Руководитель работы:

Склярова Елена Александровна,

 учитель информатики

МБОУ «СОШ № 41 им. В.В. Сизова»

Курск, 2023

**Оглавление**

[План 3](#_Toc152273962)

[Введение 4](#_Toc152273963)

[1 Основные сведения о Chat GPT 6](#_Toc152273964)

[2 Экспериментальная часть исследовательской работы 8](#_Toc152273965)

[2.1. Потенциальные риски. Как этого избежать? 11](#_Toc152273966)

[Заключение 13](#_Toc152273967)

[Список использованной литературы 14](#_Toc152273968)

# **План**

1. Собрать всю имеющуюся информацию для написания исследовательской работы.
2. Запустить Chat GPT для проверки работы бота.
3. Провести исследования по нескольким направлениям.
4. Идентифицировать возможные риски и неблагоприятные последствия, связанные с использованием модели.
5. Предложить стратегии и рекомендации для минимизации этических проблем.
6. Сформулировать заключение по эффективности и этическим аспектам использования Chat GPT.
7. Оформить исследовательскую работу.

# **Введение**

**Тема исследовательской работы** «Chat GPT: анализ эффективности и этические аспекты использования» представляет собой актуальную и важную проблему в области искусственного интеллекта и развития чат-ботов.

С точки зрения эффективности, важно провести анализ и оценку качества работы Chat GPT. Необходимо исследовать, насколько точно и полно модель отвечает на вопросы и предоставляет пользователю нужную информацию. Это позволит определить преимущества и ограничения использования такой модели в различных сценариях.

Однако, помимо эффективности, также актуальны этические аспекты использования Chat GPT. Чат-боты, основанные на этой модели, способны обманывать пользователей, выдавая себя за человека. Это может привести к негативным эмоциональным и психологическим последствиям для пользователей, особенно если они не осознают, что общаются с искусственным интеллектом. Также важным вопросом является безопасность данных и конфиденциальность.

Итак, **актуальность темы** исследовательской работы «Chat GPT: анализ эффективности и этические аспекты использования» заключается в необходимости проведения исследования и обсуждения этих вопросов, чтобы обеспечить эффективное и этичное использование таких чат-ботов в различных сферах деятельности.

**Цель работы:**

Основная цель - оценить эффективность и этические аспекты использования Chat GPT в различных сценариях коммуникации с пользователями.

**Задачи:**

1. Исследовать эффективность модели Chat GPT в сравнении с другими существующими моделями генерации текста для чатов.
2. Оценить качество ответов, сгенерированных моделью Chat GPT, изучив их понятность, релевантность и правдоподобие.
3. Рассмотреть этические вопросы, связанные с применением модели Chat GPT, такие как возможность создания и распространения нежелательного или вредоносного контента.
4. Изучить проблемы конфиденциальности и безопасности, связанные с использованием модели Chat GPT, и предложить меры для защиты персональных данных пользователей.

# **1 Основные сведения о Chat GPT**

Chat GPT – это модель искусственного интеллекта, разработанная OpenAI. Она является одной из наиболее продвинутых моделей для генерации текста на основе заданной информации. Chat GPT обучен на огромном количестве разнообразных данных, чтобы научиться генерировать релевантные и связные ответы на вопросы пользователей.

Chat GPT имеет широкий спектр применений, включая создание диалоговых чат-ботов, помощь в ответах на вопросы пользователей, создание контента и многое другое. Модель может быть использована в различных областях, включая сферу обслуживания клиентов, компьютерные игры, автоматическую генерацию текста и другие.

Chat GPT является продолжением разработки модели GPT-3 (Generative Pre-trained Transformer 3) и представляет собой улучшенную версию предыдущих моделей. Она известна своей способностью генерировать связные и качественные ответы, а также легко адаптируется к различным контекстам и задачам [1].

Однако, важно понимать, что Chat GPT по-прежнему может допускать ошибки и генерировать неправильные или несоответствующие ответы. Это связано с тем, что модель не всегда точно понимает и учитывает контекст и намерения пользователя. Работу модели также могут ограничивать ее недостатки, включая недостаточную обученность на определенных темах или неспособность рассуждать о сложных вопросах.

OpenAI продолжает работу над развитием модели Chat GPT и планирует внесение улучшений и корректировок для ее оптимизации. Использование Chat GPT может быть полезным средством для автоматизации процесса общения с пользователями и генерации текста, но требует внимательности и проверки сгенерированных ответов для достижения желаемых результатов.

Как известно, модели обработки естественного языка, такие как Chat GPT, стали широко применяться в различных областях, включая общение с пользователями через мессенджеры, поддержку клиентов, автоматизацию рутинных задач и многое другое. Однако, вместе с возросшей популярностью этих моделей возникают и вопросы эффективности и этического использования [1].

Chat GPT работает на основе огромного объема информации, собранной из Интернета, что включает в себя личные данные и приватные разговоры людей. Возникает вопрос о защите этой информации и возможных нарушениях приватности.

# **2 Экспериментальная часть исследовательской работы**

Экспериментальная часть исследовательской работы проводится с целью оценить эффективность и выявить потенциальные этические проблемы при использовании модели Chat GPT в различных сценариях общения [3].

*Описание датасетов:* Исследование начинается с составления нескольких датасетов, предназначенных для оценки производительности Chat GPT. Датасеты должны включать в себя широкий спектр тем и контекстов для охвата различных ситуаций общения.

*Оценка качества генерации:* Используя различные метрики, такие как BLEU, ROUGE и Perplexity, оценивается качество ответов, генерируемых Chat GPT. Для этого создается тестовый набор предложений, на которых модель будет проверяться. Сравниваются сгенерированные ответы с эталонными ответами и оценивается степень соответствия.

*Оценка общего уровня понимания:* Для более глубокой оценки эффективности и понимания моделью Chat GPT, проводился эксперимент, где пользователи задавали вопросы или просили выполнить задачу. Оценивали долю правильно выполненных задач или правильно интерпретированных вопросов. (рисунок 1).



Рисунок 1 – Ввод запроса с целью проверки адекватности ответа

*Формирование оценки адекватности модели:* В этом эксперименте оценивается, насколько часто модель генерирует неадекватные, вводящие в заблуждение, оскорбительные или нежелательные ответы. Заданный набор вопросов и провокационных заявлений используются для проверки поведения модели, и результаты анализируются для определения уровня нежелательных ответов.

* *Экспертное мнение:* Можно оценить адекватность модели, предоставив ее использование экспертам в соответствующей области и собрать их мнение о том, насколько хорошо модель отвечает на вопросы и ведет диалог.
* *Оценка пользователей:* Можно попросить пользователей оценить адекватность модели после каждого диалога или после некоторого времени использования и собрать их обратную связь.
* *Сравнение с эталонными ответами:* Если есть набор эталонных ответов, можно сравнить ответы модели с этими эталонами и оценить, насколько близки или совпадают ответы.
* *Обратная связь модераторов:* В случае использования модели в реальных условиях веб-чата, модераторы могут оценить адекватность модели путем активного наблюдения за диалогами и предоставления обратной связи.
* *Процент правильных ответов:* Можно определить процент правильных ответов модели на некотором наборе тестовых вопросов и использовать его в качестве метрики адекватности.
* *Диалоговые метрики:* Можно использовать специальные метрики, такие как BLEU (bilingual evaluation understudy), ROUGE (recall-oriented understudy for gisting evaluation) или F1-score, чтобы измерить сходство ответов модели с эталонными ответами или с ответами, полученными от человека [4].

Важно отметить, что оценка адекватности модели Chat GPT является сложной задачей, так как адекватность может быть субъективной и зависеть от контекста. Поэтому рекомендуется использовать несколько методов и метрик для более всесторонней оценки.

*Эксперименты на пользователях:* Проводились опросы и интервью с пользователями, которые взаимодействовали с моделью Chat GPT в различных сценариях. Они поделились своими впечатлениями и оценками эффективности, а также высказались относительно этических аспектов использования модели.

*Предложение рекомендаций и улучшений:* на основе результатов экспериментального и аналитического анализа, сформулированы рекомендации по улучшению эффективности и чувствительности модели Chat GPT:

1. Увеличить размер обучающего набора данных: для улучшения эффективности и чувствительности модели Chat GPT, необходимо обеспечить ей больше данных для обучения. Увеличение размера обучающего набора поможет модели получить больше информации и лучше понять контекст и намерения пользователя.
2. Ввести ограничения на длину ввода: Ограничение на длину ввода может помочь модели сосредоточиться на существенной информации и выдавать более точные и аккуратные ответы. Определите максимальную длину ввода, которую модель может обрабатывать эффективно, и отбросьте или обрежьте любой ввод, превышающий это значение.
3. Настройка гиперпараметров модели: Изменение гиперпараметров модели, таких как размер слоя, число слоев и коэффициент обучения, может значительно повлиять на ее эффективность и чувствительность. Попробуйте различные комбинации гиперпараметров и выберите те, которые наиболее оптимально справляются с задачей чат-моделирования [2,3].
4. Введение механизма оценки качества ответов: для повышения эффективности модели Chat GPT можно ввести механизм оценки качества ответов. Это может включать в себя оценивание набора тестовых данных специалистами, введение метрик качества или применение ранжирования ответов. Это поможет модели выбирать наиболее подходящие и информативные ответы для пользователя.
5. Фокусировка на конкретных предметных областях: Ограничение области знаний, в которой модель работает, может помочь улучшить ее эффективность и чувствительность. Фокусировка на конкретных предметных областях позволит модели лучше понимать вопросы и контекст, связанные с этой областью, и давать более точные ответы [4].
6. Участие пользователей в обучении модели: Включение обратной связи и участия пользователей в обучении модели может помочь улучшить ее эффективность и чувствительность. Пользователи могут предоставлять информацию о том, насколько точными и полезными были ответы модели, и эту информацию можно использовать для дальнейшего улучшения модели.
7. Сбалансировать генерацию и ранжирование ответов: Модель Chat GPT может быть улучшена путем балансирования генерации и ранжирования ответов. Когда модель имеет несколько вариантов ответов, она должна быть способна выбирать наиболее подходящий и информативный ответ, а не просто генерировать случайный.
8. Разработать эффективный алгоритм ранжирования и обучение модели использовать их может улучшить ее чувствительность и качество предоставляемых ответов [2].

## **2.1. Потенциальные риски. Как этого избежать?**

Еще одно опасение работодателей, связанное с использованием чата GPT — это [конфиденциальность и приватность данных](https://lpgenerator.ru/blog/chto-takoe-utechka-dannyh/). Во время «разговоров» с ИИ сотрудники могут случайно поделиться коммерческой тайной, служебной или конфиденциальной информацией.

Хотя создатели чата заявляют, что чат-бот не сохраняет информацию из разговоров, тем не менее он «учится» на каждом разговоре. И, конечно же, пользователи «общаются» с чатом в Интернете, где отсутствуют какие-либо гарантии безопасности.

Предусмотрительные работодатели будут включать в соглашение и политику конфиденциальности запреты для сотрудников ссылаться или вводить конфиденциальную информацию и коммерческую тайну в чат-боты и языковые модели на базе искусственного интеллекта, такие как чат GPT.

В то же время, поскольку чат обучался на широком спектре данных, сотрудники могут получать и использовать информацию, которая защищена товарным знаком, авторским правом или является интеллектуальной собственностью другого лица или организации, что создает правовой риск для работодателей [1].

 Вывод: всю информацию, которую пользователь сообщает о в чат бот необходимо чательно фильтровать.

# **Заключение**

Данный исследовательский проект направлен на оценку эффективности и этических аспектов использования Chat GPT.

В заключение, исследование позволило оценить эффективность и этические аспекты использования чат-бота GPT. Было выявлено, что данная технология обладает большим потенциалом в различных областях, таких как клиентский сервис, помощь при обучении и развлечения.

Однако, возникают этические вопросы, связанные с безопасностью и приватностью данных пользователей, а также возможностью создания и распространения нежелательного или дезинформационного контента.

Для того чтобы успешно применять чат-боты GPT, необходимо разработать строгие нормы и правила использования, а также активно работать над совершенствованием технологии, чтобы она стала более этичной и надежной.

В целом, GPT представляет собой мощный инструмент, который может быть полезен обществу, но его использование должно осуществляться с осознанием его потенциальных последствий и с соблюдением этических принципов.

Считаю, что поставленная цель достигнута, все задачи решены.

# **Список использованной литературы**

1. Бендер Э. М., Гебру Т., Макмиллан-Мейджор. Об опасностях стохастических попугаев: Могу ли я. – М.: С.П. 2021. – С. 232.
2. Рэдфорд А. Языковые модели - это неконтролируемые многозадачные учащиеся. Блог OpenAI 1(8), 9.
3. Тимур Казанцев. Chat GPT и Революция Искусственного Интеллекта. – М.: Москва, 2022. – С. 135.
4. Чжао, Дж., & Чжу, Дж.// Исследование и контрмеры инициативы по обучению в системах разговорного искусственного интеллекта. – М., Правда, 2021. – С. 1-5.