Министерство образования и науки Российской Федерации

Федеральное государственное бюджетное
образовательное учреждение высшего образования
«Комсомольский-на-Амуре государственный университет»

На правах рукописи

Черезов Никита Сергеевич

Оценка качества и эффективности нейронных сетей в роли чат-ботов

Направление подготовки 01.04.02
«Прикладная математика и информатика»

АВТОРЕФЕРАТ
МАГИСТЕРСКОЙ ДИССЕРТАЦИИ

2024

Работа выполнена в ФГБОУ ВО
«Комсомольский-на-Амуре государственный университет»

Научный руководитель:	кандидат физико-математических наук, доцент кафедры «Прикладная математика» Григорьева Анна Леонидовна
Рецензент:	кандидат физико-математических наук, доцент Анисимов Антон Николаевич

Защита состоится 18 июня 2024 г. в 9:00 часов на заседании государственной экзаменационной комиссии по направлению подготовки 01.04.02 – «Прикладная математика и информатика» в ФГБОУ ВО «КнАГУ» по адресу: 681000, г. Комсомольск-на-Амуре, пр. Ленина, 27, ауд. 204/5.

Автореферат разослан 15 июня 2024 г.

Секретарь ГЭК З.В. Широкова

ОБЩАЯ ХАРАКТЕРИСТИКА ДИССЕРТАЦИОННОЙ РАБОТЫ

Актуальность темы.

Использование нейронных сетей в чат-ботах в современных условиях обусловлено следующими обстоятельствами:

- Рост объема данных, требующих автоматизированной обработки.
- Необходимость улучшения взаимодействия с пользователями.
- Повышение требований к качеству и скорости предоставления информации.

Особенно это актуально для областей, где требуется высокая точность и персонализация, таких как здравоохранение, образование и бизнес.

Актуальность работы заключается в том, что разработка и внедрение высокоэффективных чат-ботов на основе нейронных сетей позволяет значительно улучшить качество обслуживания пользователей, автоматизировать рутинные задачи и повысить общую производительность. *Цель магистерской диссертации* является оценка качества и эффективности различных моделей нейронных сетей, используемых в чат-ботах.

Основные задачи магистерской диссертации

1) изучить принципы работы различных нейронных сетей

2) сравнить эффективность различных моделей нейронных сетей

3) исследовать этические и социальные аспекты использования нейронных сетей.

4) анализ перспектив и будущих направлений развития нейронных сетей

Объектом исследования являются процессы взаимодействия пользователей с чат-ботами, основанными на нейронных сетях.

Предметом исследования являются аналитические методы и модели оценки качества и эффективности нейронных сетей в контексте их использования в чат-ботах.

Научная новизна магистерской диссертации: научная новизна данного исследования заключается в личном подходе к оценке качества и эффективности нейронных сетей в роли чат-ботов в текстовом общение, а также с помощью изображений, аудио и видео. В рамках этого исследования были учтено использование улучшенного бенчмарка MMLU-Pro для более точной оценки способностей моделей в широком спектре задач.

Достоверность и обоснованность результатов исследования. Основана на использовании современных методов анализа данных, экспериментальных тестов и сравнения результатов с существующими моделями.

Практическая ценность магистерской диссертации заключается в разработке рекомендаций по улучшению работы с чат-ботами. Это включает предложения по выбору оптимальных моделей для различных приложений.

Апробация результатов. Результаты работы докладывалась на:

– VII Всероссийской национальной научной конференции студентов, аспирантов и молодых учёных, ФГБОУ ВО «КнАГУ», 2024.

Публикации. По результатам выполненных в диссертации исследований автором опубликовано 2 работы с темами в:

– «Молодёжь и наука: актуальные проблемы фундаментальных и прикладных исследований. Материалы VII Всероссийской национальной научной конференции студентов, аспирантов и молодых учёных. 2024.

Структура и объем. Магистерская диссертация состоит из введения, четырех глав, заключения, списка литературы и приложения. Объем работы – 150 страниц, в том числе 85 рисунков, 42 источника и 1 приложение.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Введение раскрывает актуальность темы, определяются цели и задачи исследования, объект, предмет, указываются научная новизна, практическая значимость, достоверность и обоснованность результатов исследования.

В первой главе описывается история создания нейронных сетей.

Так же описываются применения нейронных сетей, как они используются для обработки и генерации текста, что улучшает качество взаимодействия с пользователями. Основные вызовы включают обеспечение точности и релевантности ответов, а также необходимость в больших вычислительных ресурсах. Так же были описаны принципы работы современных чат-ботов. Современные чат-боты используют методы обработки естественного языка (NLP), механизмы внимания и обучение на больших объемах данных. Эти принципы и этапы развития нейронных сетей и чат-ботов подробно рассмотрены в первой главе, что помогает понять текущие возможности и перспективы этих технологий.

Во второй главе описываются и разбираются некоторые возможности нейронных сетей.

OpenAI ChatGPT-3.5 является одной из версий языковых моделей OpenAI и основывается на многослойных трансформерах. Модель обладает значительным количеством параметров, что позволяет ей генерировать качественные ответы, помогать в создании программного кода и решать повседневные задачи. ChatGPT-3.5 доступна бесплатно, что делает её популярной для широкого круга пользователей.

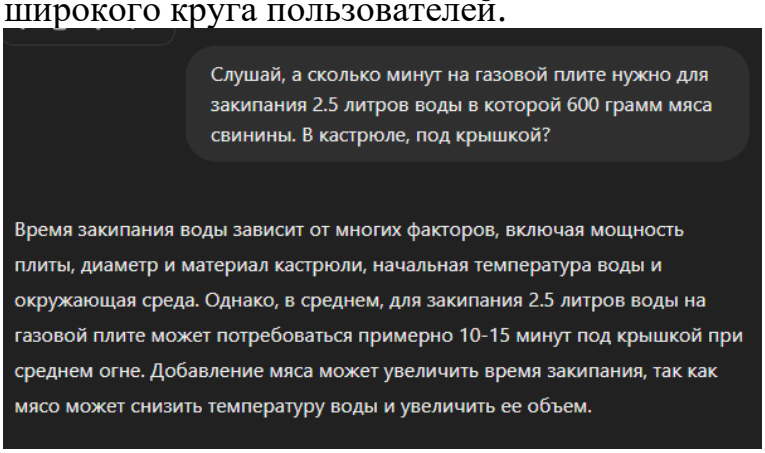

Рисунок 1 - Пример использования

OpenAI GPT-4 улучшена по сравнению с GPT-3.5, имея большее контекстное окно и более высокое количество параметров, что позволяет обрабатывать и запоминать больше информации. Она демонстрирует лучшие результаты в понимании сложных вопросов и генерации ответов. Может получать доступ к интернету в реальном времени для предоставления актуальной информации и анализа веб-ссылок.

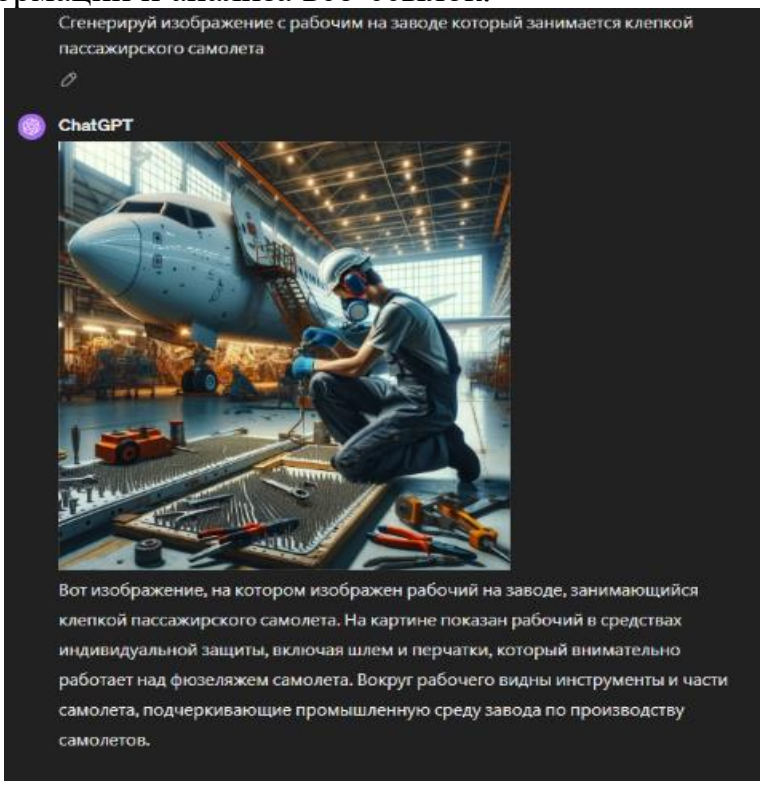

Рисунок 2 - Пример использования

OpenAI GPT-4 Omni поддерживает работу с текстом, аудио, изображениями и видео, что делает её более универсальной по сравнению с GPT-4. Она обладает улучшенной адаптивностью, более высокой скоростью обработки и поддерживает многоязычность, что делает её идеальной для глобальных приложений и реального времени.

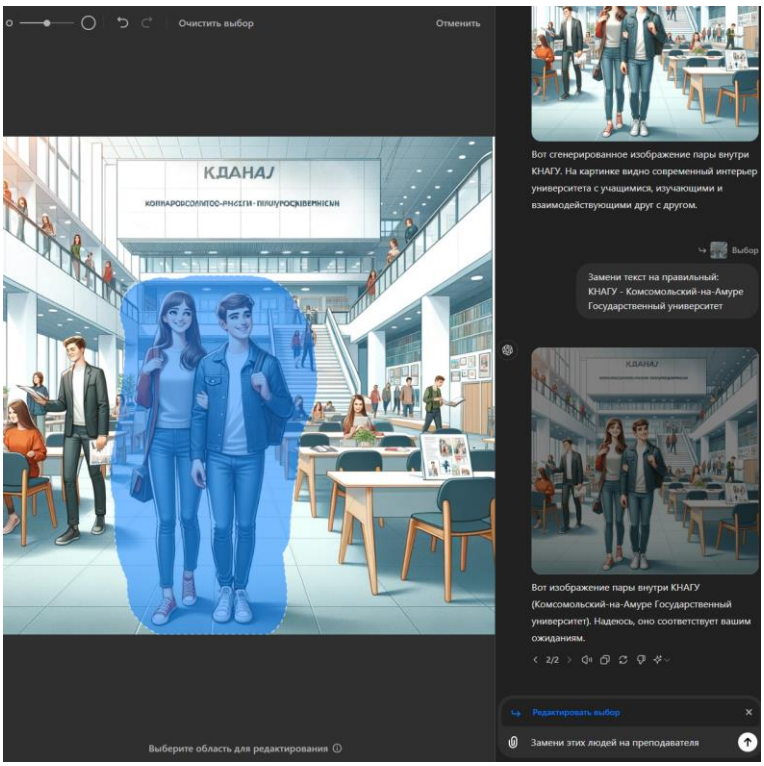

Рисунок 3 - Пример использования

Claude-3-Opus представляет собой продвинутую нейронную сеть для обработки естественного языка, способную обрабатывать и анализировать большие объемы данных. Она поддерживает контекстное окно в 200000 токенов и может генерировать текст, неотличимый от человеческого. Модель может дообучаться на конкретных наборах данных, что делает её полезной для узкоспециализированных задач.

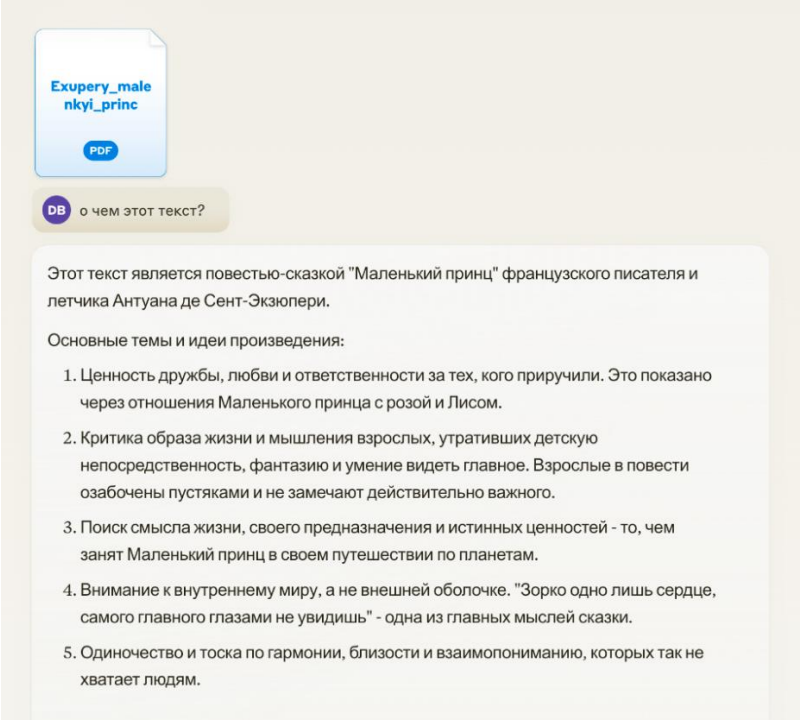

Рисунок 4 - Пример использования

Gemini 1.5 Flash разработана для быстрого выполнения задач с минимальными затратами. Она может обрабатывать различные типы данных, включая аудио, изображения, видео и текст, поддерживает настройку модели и является идеальной для задач, требующих быстрого выполнения и обработки большого объема данных.

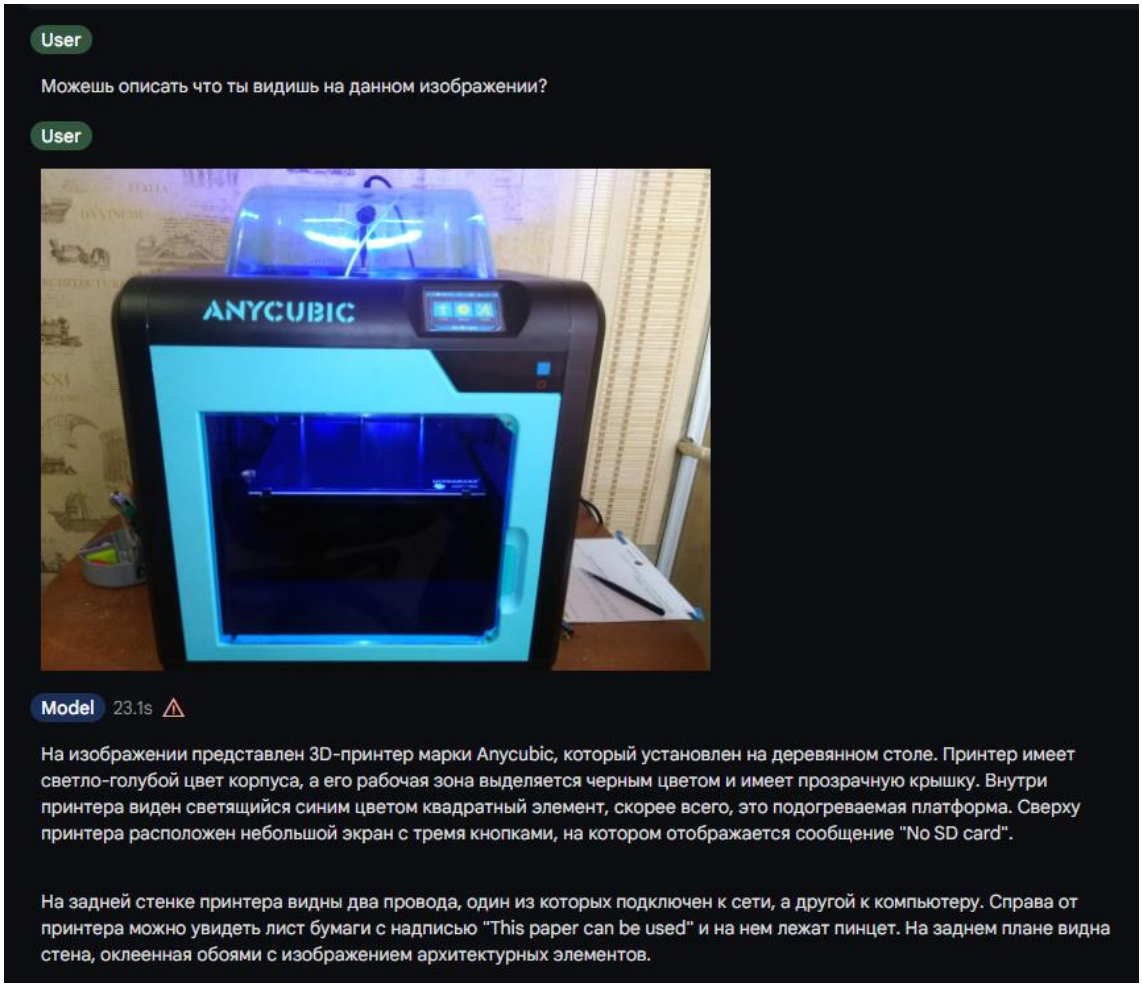

Рисунок 5 - Пример использования

Gemini 1.5 Pro обладает увеличенным контекстным окном и улучшенными возможностями по сравнению с Flash версией. Она показывает улучшения в генерации кода, логических рассуждениях и обработке мультимедийных данных. Pro версия лучше подходит для задач, требующих высоких вычислительных ресурсов и качественного анализа данных.

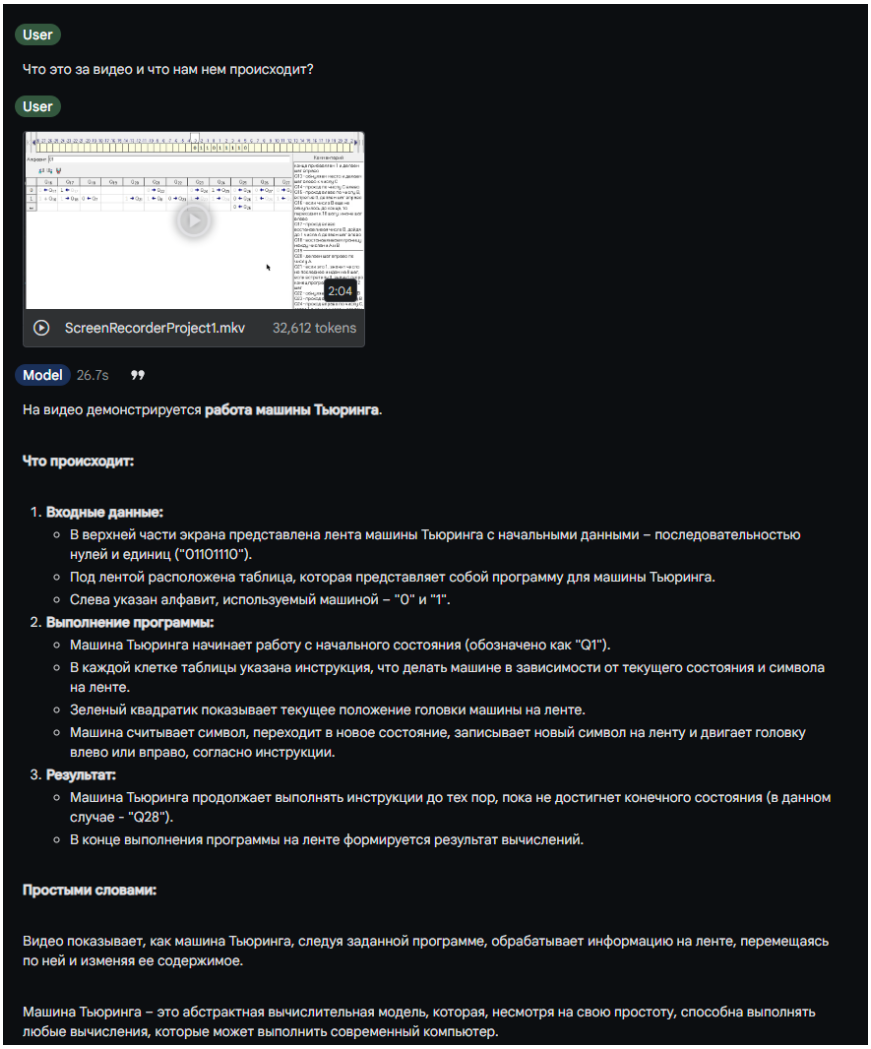

Рисунок 6 - Пример использования

Mistral Large представляет собой мощную нейронную сеть, оптимизированную для низкой задержки и высокой производительности. Она поддерживает контекстное окно до 32,000 токенов, что позволяет обрабатывать большие объемы текстовой информации. Модель обучена на более чем 80 языках программирования, включая Python, Java, C, C++, PHP, Bash и другие, что делает её особенно полезной для задач, связанных с программированием и кодированием. Mistral Large может быть развернута как на облачных платформах, так и локально, что позволяет использовать модель в различных средах с максимальной гибкостью и безопасностью

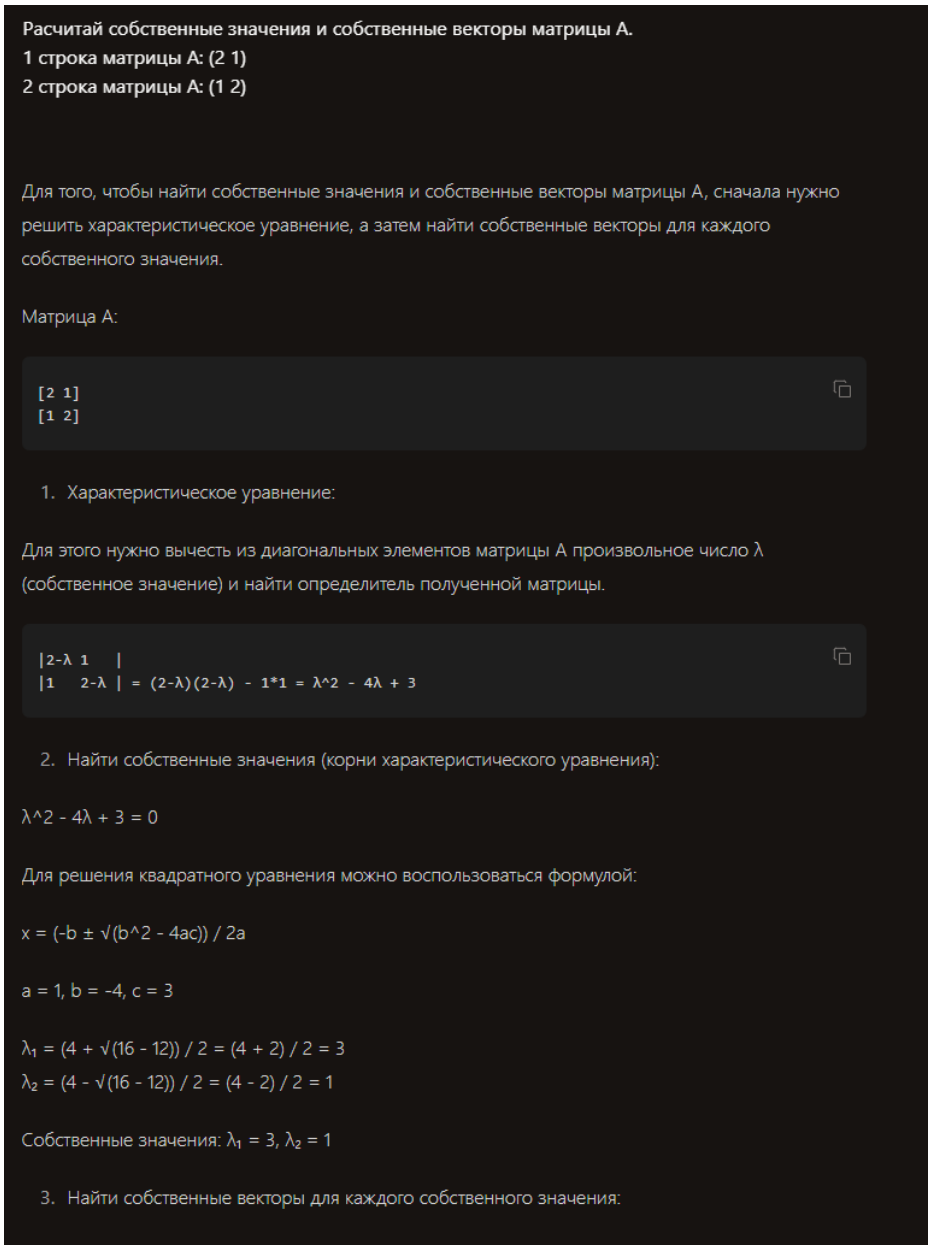

Рисунок 7 - Пример использования

Llama-3-70B-Instruct оптимизирована для задач на основе инструкций и диалогов. Она способна обрабатывать большие объемы текстовой информации и поддерживает многоязычность. Модель использует сложные методы обучения для генерации высококачественных и контекстно подходящих ответов.

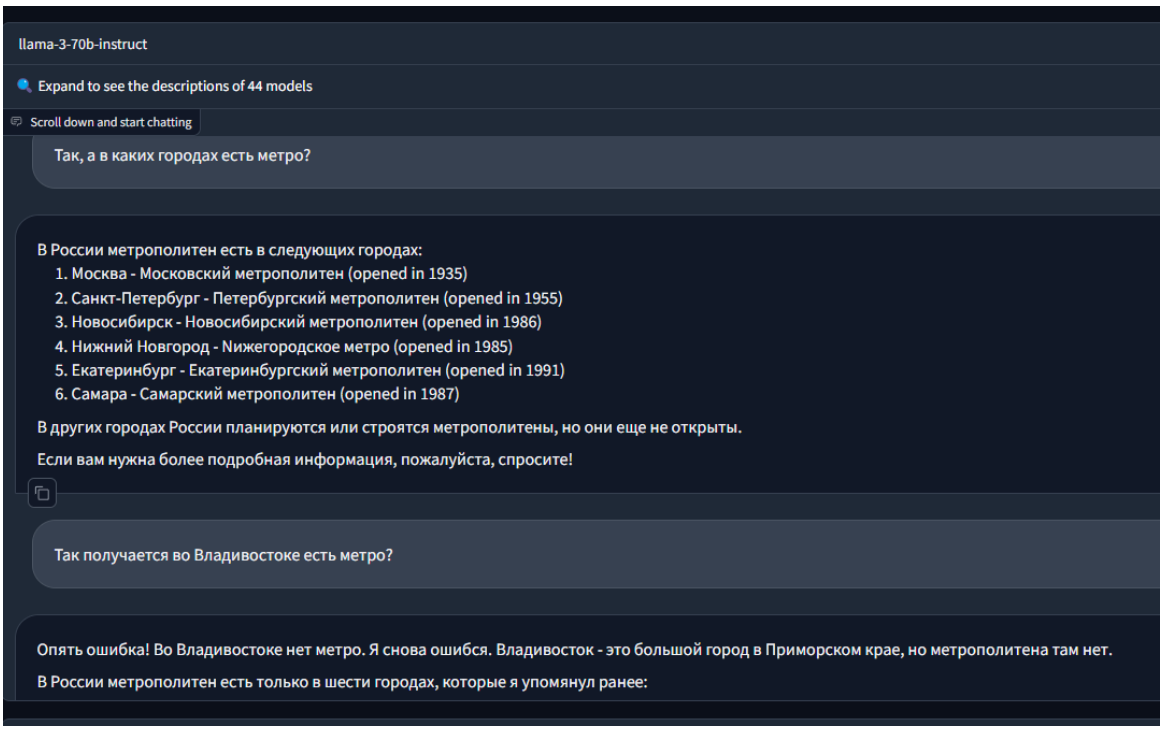

Рисунок 8 - Пример использования

h2oGPT позволяет использовать большие языковые модели локально, обеспечивая высокий уровень приватности и контроля над данными. Модель поддерживает различные типы данных и предлагает возможности для создания чат-ботов, анализа документов, генерации текста и перевода. Она обеспечивает гибкость и адаптивность для специфических нужд пользователей.

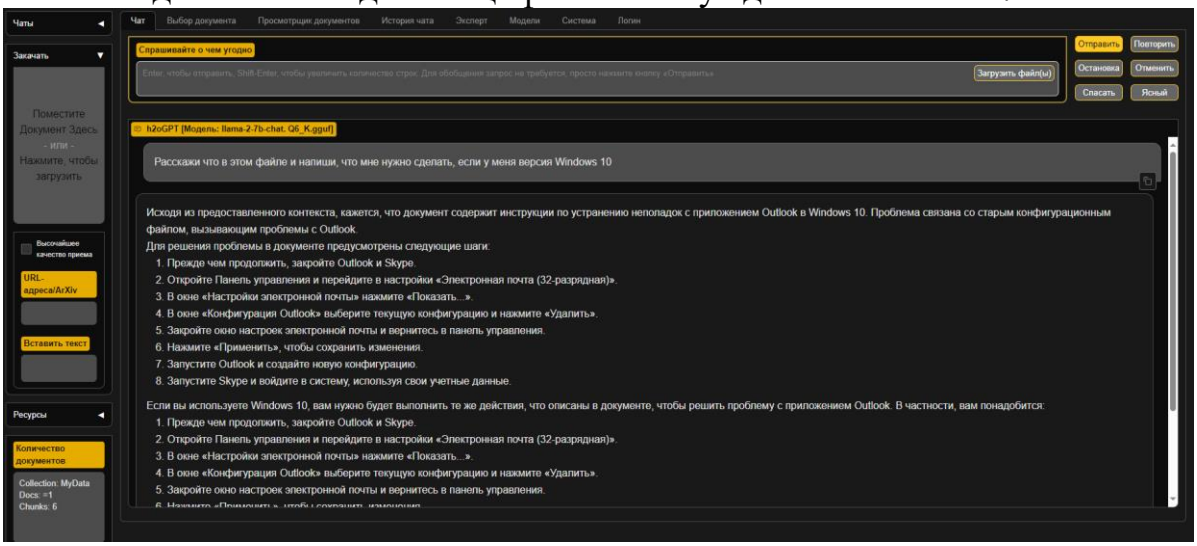

Рисунок 9 - Пример использования

Эта глава демонстрирует сравнительный анализ различных моделей нейронных сетей, их возможности и ограничения, что помогает определить их применение в различных сценариях и задачах, а также вариант использования в локальной системе.

В третьей главе проводится сравнительный анализ качества и эффективности различных моделей нейронных сетей, используемых в чат-ботах.

На основе тестирования и анализа результатов выявлены следующие ключевые аспекты. **Gemini 1.5 Pro** продемонстрировала отличные результаты при работе с разнообразными типами данных, включая видео и аудио, что делает её универсальной и мощной. Однако она уступает ChatGPT-4 Omni и Claude-3 Opus в работе с кодом, особенно при реализации сложных задач, таких как разработка игр (например, тетрис). Несмотря на заявленную поддержку мультимодальных возможностей в ChatGPT-4 Omni, на данный момент не реализует работу с аудио и видео. Однако является одной из лучших нейронных сетей для работы с кодом благодаря своей продвинутости и гибкости. Она также превосходит Gemini 1.5 Pro в детальном анализе изображений. Обе модели хорошо справляются с математическими задачами и могут проверять корректность ответов других нейросетей. Они способны даже консультировать, хотя иногда могут возникать ошибки, связанные с неправильной интерпретацией контекста. В целом, остальные модели тоже подходят для ведения обычных разговоров, хотя каждая имеет свои особенности и ограничения.

Для оценки возможностей нейронных сетей используется усовершенствованный бенчмарк MMLU-Pro. Этот бенчмарк разработан для оценки моделей понимания языка по более широкому и сложному набору задач, включая вопросы, требующие логических рассуждений, и увеличение числа вариантов ответов с четырёх до десяти.

MMLU-Pro включает более 12,000 тщательно отобранных вопросов из академических экзаменов и учебников, охватывающих 14 различных областей, таких как биология, бизнес, химия, компьютерные науки, экономика, инженерия, здравоохранение, история, право, математика, философия, физика, психология и другие. Это делает MMLU-Pro более реалистичным и сложным бенчмарком, значительно уменьшающим вероятность успеха за счёт случайных угадываний.

Эти результаты показывают, что модели имеют разные уровни производительности в зависимости от конкретной области знаний. Например, GPT-4o и Gemini 1.5 Pro показывают лучшие результаты в биологии и математике, тогда как Claude-3-Opus и Higgs-Llama-3-70B сильны в психологии и экономике.

Models	Overall	Biology	Business	Chemistry	Computer Science	Economics	Engineering	Health	History	Law	Math	Phil
GPT-4o	0.7255	0.8675	0.7858	0.7393	0.7829	0.808	0.55	0.7212	0.7007	0.5104	0.7609	0.70
Gemini-1.5-Pro	0.6903	0.8466	0.7288	0.7032	0.7293	0.7844	0.4871	0.7274	0.6562	0.5077	0.7276	0.61
Claude-3-Opus	0.6845	0.8507	0.7338	0.693	0.6902	0.798	0.484	0.6845	0.6141	0.5349	0.6957	0.63
GPT-4-Turbo	0.6371	0.8243	0.673	0.5592	0.6854	0.7476	0.3591	0.7078	0.6772	0.5123	0.6277	0.64
Higgs-Llama-3-70B	0.6316	0.8354	0.6743	0.6034	0.6902	0.7512	0.4737	0.6687	0.6404	0.4432	0.6321	0.55
Gemini-1.5-Flash	0.5912	0.8131	0.667	0.613	0.5951	0.6943	0.4416	0.6039	0.538	0.3732	0.5958	0.49
Yi-large	0.5809	0.6987	0.6413	0.6166	0.6341	0.6813	0.4541	0.6443	0.4961	0.3624	0.6481	0.55
Claude-3-Sonnet	0.568	0.768	0.657	0.5291	0.59	0.709	0.4045	0.6332	0.5721	0.427	0.49	0.51
Llama-3-70B-Instruct	0.562	0.7812	0.6018	0.4681	0.6053	0.6841	0.4362	0.6533	0.5692	0.3991	0.5402	0.54
Phi3-medium-4k	0.557	0.7587	0.616	0.4991	0.5415	0.7038	0.3787	0.6357	0.5722	0.3833	0.5218	0.55
Deepseek-V2-Chat	0.5481	0.6625	0.6375	0.5415	0.5171	0.6363	0.3189	0.5825	0.4528	0.4064	0.5366	0.54
Llama-3-70B	0.5278	0.749	0.4994	0.417	0.5512	0.6528	0.3498	0.6174	0.5774	0.3497	0.4967	0.56

Рисунок 10 - Результат тестирования

Таким образом, третья глава диссертации фокусируется на сравнительном анализе качества и эффективности различных моделей нейронных сетей, а также

на результатах их тестирования с использованием усовершенствованных бенчмарков, таких как MMLU-Pro.

В четвертой главе рассматриваются перспективы развития и применения нейронных сетей в чат-ботах. Основное внимание уделяется текущим трендам в развитии нейронных сетей, таким как использование современных архитектур трансформеров и улучшенных алгоритмов обучения. Обсуждаются возможности дальнейшего улучшения моделей, такие как снижение потребления ресурсов и расширение областей применения, включая здравоохранение, финансы, образование и промышленность.

Также рассматриваются возможности интеграции нейронных сетей с другими передовыми технологиями, такими как IoT, AR и VR, что открывает новые возможности для создания умных систем и улучшения взаимодействия с пользователями.

Отдельное внимание уделяется этическим и социальным аспектам использования нейронных сетей в чат-ботах, включая защиту данных пользователей, предотвращение предвзятости и влияние автоматизации на рынок труда. Обсуждаются меры по обеспечению прозрачности и объяснимости работы нейронных сетей, а также необходимость переподготовки специалистов для адаптации к изменениям на рынке труда.

Таким образом, глава анализирует текущие тренды, прогнозы и перспективы развития нейронных сетей, возможности их интеграции с другими технологиями, а также этические и социальные аспекты использования в чат-ботах и их влияние на рынок труда и образование.

В заключении подводятся основные итоги исследований, проводится анализ полученных результатов.

СПИСОК ПУБЛИКАЦИЙ ПО ТЕМЕ ИССЛЕДОВАНИЯ

1. Черезов Н.С. ОЦЕНКА КАЧЕСТВА И ЭФФЕКТИВНОСТИ НЕЙРОННЫХ СЕТЕЙ В РОЛИ ЧАТ-БОТОВ - РАБОТА С ГОТОВЫМ КОДОМ / Н.С. Черезов, Я.Ю. Григорьев // Молодёжь и наука: актуальные проблемы фундаментальных и прикладных исследований. Материалы VII Всероссийской национальной научной конференции студентов, аспирантов и молодых учёных, Комсомольск-на-Амуре, 08–12 апреля 2024 года. – Комсомольск-на-Амуре: Комсомольский-на-Амуре государственный университет, 2024.
2. Черезов Н.С. ОЦЕНКА КАЧЕСТВА И ЭФФЕКТИВНОСТИ НЕЙРОННЫХ СЕТЕЙ В РОЛИ ЧАТ-БОТОВ / Н.С. Черезов, Я.Ю. Григорьев // Молодёжь и наука: актуальные проблемы фундаментальных и прикладных исследований. Материалы VII Всероссийской национальной научной конференции студентов, аспирантов и молодых учёных, Комсомольск-на-Амуре, 08–12 апреля 2024 года. – Комсомольск-на-Амуре: Комсомольский-на-Амуре государственный университет, 2024