**Министерство образования и науки Российской Федерации**

**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ**

**“САНКТ-ПЕТЕРБУРГСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ**

**УНИВЕРСИТЕТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ,**

**МЕХАНИКИ И ОПТИКИ”**

**ФАКУЛЬТЕТ СРЕДНЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ**

**ОТЧЕТ**

**ПО ЛАБОРАТОРНОЙ РАБОТЕ № 4**

**«Интерфейс, определение ИИ текста»**

|  |  |
| --- | --- |
|  | Выполнили:  студенты группы К33421  Максимов Д. Э.  Ковалев В. Д.  Азаренков Г.Д. |

Санкт-Петербург

2024

ЗАДАЧИ

Для разработанной ранее модели реализовать интерфейсную часть.

ЭТАПЫ ВЫПОЛНЕНИЯ ЛАБОРАТОРНОЙ РАБОТЫ

1. Для модели из лабораторной работы 1-3:
   1. Реализовать интерфейсную часть на выбор студента:
      1. Телеграмм-бот
      2. Бот в Discord
      3. С помощью библиотек Python (Tkinter, PyQT, Yel)
2. Интерфейс должен содержать:
   1. Поле для ввода запроса
   2. Кнопка для отправки запроса
   3. Поле для получения ответа
3. Изучить способы определения текста, который сгенерировал ИИ (см. ссылки). С помощью программных инструментов проверить свои результаты и ответить на вопросы:
   1. Какие конструкции являются явными маркерами, что текст написан ИИ? Привести примеры из своих тестов, явно показать такие места.
   2. Какие есть способы придать тексту больше «человечности»?

ХОД ВЫПОЛНЕНИЯ

Сделали телеграм бота <https://t.me/text_processing_llm_bot>. Будет запущен только в рамках защиты лабораторной работы, думали захостить на сервере, но потребуется использовать сервер с GPU, что сильно удорожает все.

При вводе “/query” команды с запросом отправляет его LLM и выводит ответ.

Изображение выглядит как текст, снимок экрана, программное обеспечение

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана

Автоматически созданное описание

В качестве датасета для примера использовался датасет из лабораторных работ 2-3 – набор описания карточек из игры slay the spire.

Изображение выглядит как текст, снимок экрана, программное обеспечение, Операционная система

Автоматически созданное описание

**Какие конструкции являются явными маркерами, что текст написан ИИ?**

* **Повторяемость и шаблонность:** ИИ может использовать однотипные фразы и структуры предложений, что делает текст предсказуемым и однообразным. Например, частое повторение вводных фраз вроде "в заключение", "однако", "в результате" может указывать на ИИ-генерацию.
* **Несогласованность и несвязность:** Тексты ИИ могут содержать предложения, которые логически не связаны между собой, или переходы между мыслями, которые кажутся неестественными. Это проявляется в резких сменах темы без должных связующих элементов.
* **Ошибки и несуразности:** ИИ может генерировать тексты с фактологическими ошибками или нелогичными утверждениями, которые не соответствуют реальности. Например, утверждение, что "Париж — столица Германии", явно ошибочно.
* **Неестественные обороты речи:** Использование фраз и выражений, которые редко встречаются в человеческой речи, может быть признаком ИИ-текста. Например, фраза "осуществление процесса передвижения" вместо простого "передвижение" выглядит излишне сложной и неестественной.
* **Чрезмерное структурирования:** Разбиение на списки подсписки, абзацы и блоки, излишнее в текущем контексте. Несмотря на то, что такое поведение можно только приветствовать, если нейросеть используется для обучения или ресерча, большинству людей не свойственен такой уровень структурирования своих ответов.

Нужно отметить, что новые поколения больших языковых моделей все сложнее и сложнее отличить от человека – их авторы отлично знают про все проблемы и активно имплементируют улучшения, направленные на очеловечивание ответов. А полагаться лишь на то, что раз ответ высокого качества, грамотно составленный и имеет высокий уровень языковых конструкций и тд – не всегда возможно.

**Привести примеры из своих тестов, явно показать такие места.**

В примерах с используемыми датасетами добиться генерации нечеловечного ответа было достаточно сложно – используемая модель или выдает явные галлюцинации из за своего небольшого размера, или отвечает лаконично и по делу, оставляя лишь самый минимум.

Один из примеров галлюцинаций:



Несмотря на пять поданных примеров, где явно описывается действие карт, нейросеть включает явно не нужный пункт с “Evolve” – пользователь просил перечислить карты, а нейросеть выдает ему общую рекомендацию, с запрошенной механикой ну вообще никак не связанную.

**Какие есть способы придать тексту больше «человечности»?**

* **Внесение разнообразия в структуру предложений**: Использование различных по длине и структуре предложений делает текст более естественным и живым. Чередование коротких и длинных предложений, использование вопросов и восклицаний придают тексту динамику.
* **Добавление личных мнений и эмоций**: Включение субъективных взглядов, эмоций и личного опыта делает текст более человечным и убедительным. Например, выражения вроде "я считаю", "мне кажется", "по моему опыту" добавляют тексту индивидуальности.
* **Использование разговорных выражений и идиом**: Применение фразеологизмов, сленга и разговорных оборотов делает текст более приближенным к живой речи. Например, выражения вроде "офигенно", придают тексту естественность, хотя и ценой некоторого ухудшения общего качества ответа.
* **Включение ошибок и опечаток**: Умеренное количество незначительных ошибок или опечаток может сделать текст более правдоподобным, так как люди склонны к ошибкам при письме. При этом это может вызывать приступы агрессии у другой стороны.

Существуютонлайн-сервисы, которые помогают преобразовать ИИ-сгенерированный текст в более естественный, например, Humanize AI Text, Undetectable AI и другие. Они переписывают текст, добавляя элементы, характерные для человеческой речи.

ВЫВОД

Телеграм боты отлично подходят для имплементации интерфейса для LLM. Примеру этому служит как эта лабораторная работа, так и обилие оберток вокруг таких сервисов, как ChatGPT, и их общая коммерческая успешность.