Распознавание происхождения текстов созданных искусственным интеллектом с помощью больших языковых моделей

Пойманов Дмитрий Романович

Студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: dima.poimanow@yandex.ru

Научный руководитель *— Местецкий Леонид Моисеевич*

С постоянным развитием искусственного интеллекта (ИИ) и его приложений, вопросы, связанные с подделками и манипуляцией информацией, становятся все более актуальными. В последние годы наблюдается значительный прогресс в разработке генеративных моделей, способных автоматически порождать текстовые данные, которые может быть трудно отличить от текстов, созданных людьми. Например, недавно разработанная модель ChatGPT [2] может генерировать человекоподобные тексты для различных задач, таких как написание кодов для компьютерных программ, заполнение документов, ответы на вопросы. В свете этого возникает важная задача - разработка методов и инструментов для распознавания и фильтрации текстов, сгенерированных с использованием искусственного интеллекта.

Среди существующих подходов в данной области можно выделить, во-первых, основанный на предположении, что языковые модели генерируют текст путем частой выборки из высоковероятных слов [5]. Последние исследования [6] показывают, что для языковых моделей с числом параметров больше 1 млрд. подобная гипотеза неверна. Во-вторых, решения [3], [4] на основе предварительной разметки и подмены выхода больших языковых моделей (LLM), что впоследствии облегчает процесс нахождения предложений, созданных определенными нейросетями. Важно, что в этих методах для изучаемого текста имеется информация об LLM, с помощью которой он был сгенерирован, поэтому область применения ограничена.

В работе предлагается подход, идейно основанный на том, что большие языковые модели (LLM) в процессе предобучения на больших объемах текстовых данных обретают способность обнаружения текстов [1], сгенерированных с помощью подобных LLM. Предлагается дополнительное обучение open-source модели на специально собранных инструкциях для приведения выхода нейросети к формату: "человек"или "ИИ". Алгоритм выглядит следующим образом:

1. Подготовка датасета D для задачи бинарной классификации в виде ответов на вопросы из разных сфер как от реальных людей, так и сгенерированных с помощью LLM.

2. Дополнительное обучение параметров (finetuning) open-source языковой модели под задачу 1 максимизации вероятности следующего слова:

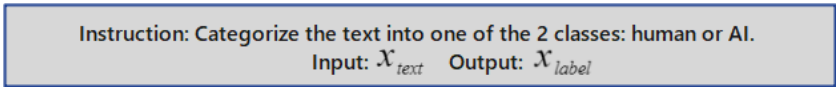
$$\theta^{'} = \underset{\theta}{argmax} \sum_{(x_{text}, x_{label}) \in D} \log P(x_{label} | x_{instruction}; x_{text}; \theta) \quad (1)$$

   где $\theta, \theta^{'}$ - изначальные и обновленные параметры модели, (x_{text}, x_{label}) - текст и его метка из подготовленных данных D, $x_{instruction}$ - инструкция для модели на иллюстрации ниже.

3. Оценка результатов классификации на датасете $HC3$ как для английского, так и для русского языков.

Преимуществом предлагаемого метода является обобщаемость под любую open-source модель. На данный момент работоспособность показана для схожего способа детектирования [1] на китайском языке, где точность распознавания свыше 90%.

Иллюстрации



Пример инструкции из обучающего датасета на английском языке

Литература

1. Wang R., Chen H. LLM-Detector: Improving AI-Generated Chinese Text Detection with Open-Source LLM Instruction Tuning // 2024
2. Сайт большой языковой генеративной модели ChatGPT `https://chat.openai.com`
3. Kirchenbauer J., Geiping J. A Watermark for Large Language Models // 2023

4. Kuditipudi R., Thickstun J. Robust Distortion-free Watermarks for Language Models // 2023
5. Gehrmann S. GLTR: Statistical Detection and Visualization of Generated Text // 2019
6. Ahmed M. Elkhatat, Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text // 2023