Использование методов подсчета неопределенности для борьбы с атаками на детекторы машинно-сгенерированного текста

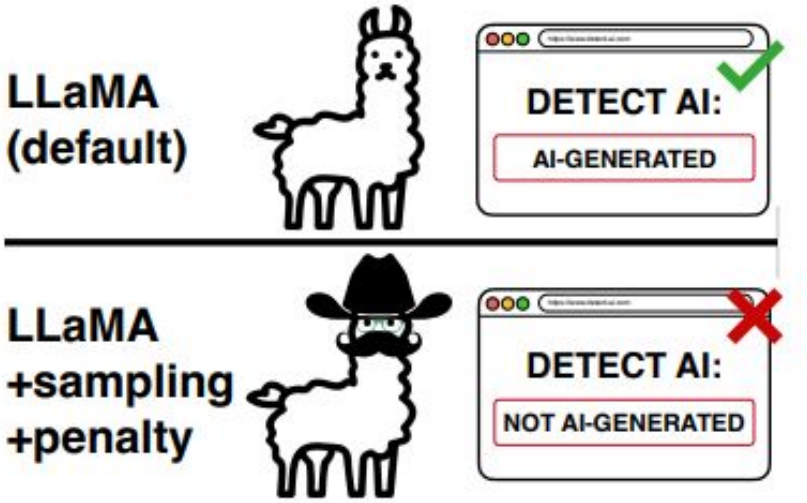
студент: Леванов В.Д. (МФТИ)
научный руководитель: Вознюк А. Е. (МФТИ)

14.04.2025

Постановка проблемы

Сейчас LLM-модели показывают невероятные результаты в генерации текста, поэтому необходимо иметь способы обнаружения машинно-сгенерированного текста, например, чтобы выявлять дезинформацию или списанные домашние работы студентов. Для этого и нужны AI-детекторы. Однако многие из них легко обмануть простыми манипуляциями с генеративной моделью или результатом генерации.Необходимо предложить метод обнаружения машино-сгенерированного текста устойчивого к различным атакам

Возможное решение - попробовать классифицировать тексты используя различные методы оценки неопределенности



Что было сделано?

На текстах части датасета M4GT предназначенной для бинарной классификации с помощью логитов контекста модели Llama-3.1-8B-Instruct вычислены 4 метода подсчета неопределенности

$$PPL = \exp\left(-\frac{1}{L}\sum_{l=1}^{L}\log P(w_l|w_{<l})\right)$$

Perplexity

$$H = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j} P(w_j|w_{<i})\log P(w_j|w_{<i})$$

Mean token entropy

$$H_S(x;\theta) = -\frac{1}{K}\sum_{k=1}^{K}\log P(y^{(k)}|x,\theta)$$

Monte Carlo Sequence Entropy

$$MD(x) = (h(x)-\mu)^T\Sigma^{-1}(h(x)-\mu)$$

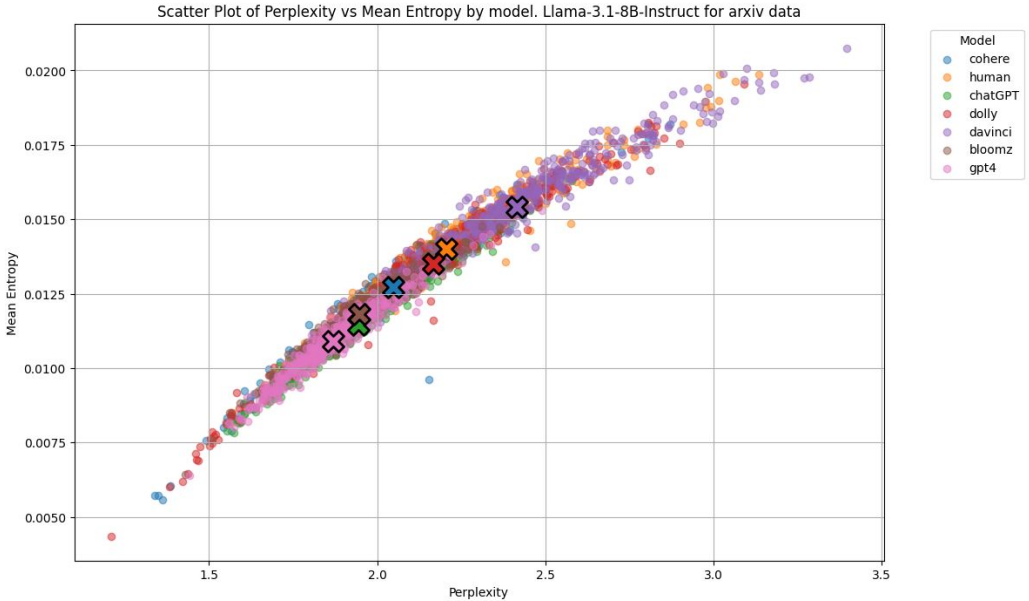
Mahalanobis Distance

Что было сделано?

На основе подсчитанных метрик свести задачу к задаче бинарной классификации. Таким образом, появляется способ, изначально имея только тексты, определять, являются ли тексты рукописными или машино-сгенерированными

Результаты

Кластеризация текстов сгенерированных одной моделью по значениям метрик Perplexity и Mean Token Entropy



Результаты

Обучены модели, которые могут за небольшое время обучения могут показывать хорошие значения точности

Data from Reddit

Модель	Accuracy	Precision	Recall	Время обучения (с)
BERT Classifier	0.9921	0.9925	0.9983	600.0193
Neural Classifier with Uncertainty	0.9114	0.9325	0.9667	103.1544
Logistic Regression with Uncertainty	0.9093	0.9289	0.9683	0.0119

Data from Arxiv

Модель	Accuracy	Precision	Recall	Время обучения
BERT Classifier	0.9942	1.0000	0.9912	1485.2992 с (24.7 мин)
Neural Classifier with Uncertainty	0.8206	0.8630	0.8688	219.0326 с (3.7 мин)
Logistic Regression with Uncertainty	0.7744	0.8126	0.8600	0.0136 с (13.6 мс)

Дальнейшие планы

1. Обучиться на данных из датасета RAID с атаками
2. Увеличить число источников рассматриваемых текстов
3. Перебрать большее число классифицирующих моделей для увеличения точности
4. (опционально) Добавить новые рассматриваемые методы подсчета неопределенности