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

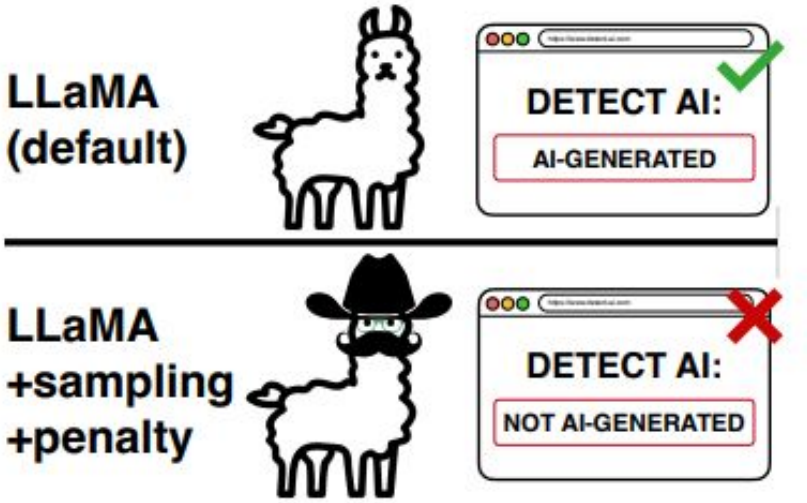
студент: Леванов В.Д. (МФТИ)
научный руководитель: Вознюк А. Е. (МФТИ)

16.05.2025

# 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



# Uncertainty estimation

Функция над логитами контекста некоторой LLM модели. В нашем случае исследовалось 4 вида метрик. Первые три оценки основаны на информации. Расстояние Махаланобиса основано на плотности.

**Perplexity**

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

**Mean token entropy**

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

**Monte Carlo Sequence Entropy**

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

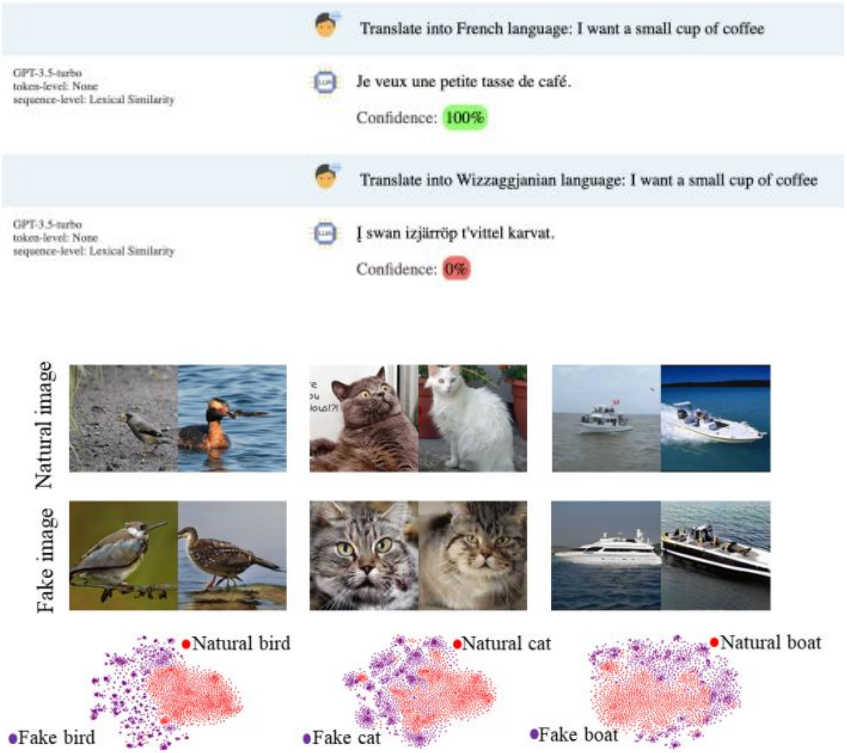
**Mahalanobis Distance**

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

# Обзор литературы

1. Исследование различных методов подсчета неопределенности для задач NLP

2. Применение неопределенности для обнаружения отличий настоящих картинок от сгенерированных



# План действий

1. Выбрать тексты и для них подсчитать метрики оценки неопределенности, предварительно получив логиты контекста для них из LLM с открытой архитектурой
2. По полученным метрикам решить задачу бинарной классификации. Таким образом, появляется способ, изначально имея только тексты, определять, являются ли тексты рукописными или машино-сгенерированными. Хочется получить модель показывающую хорошую точность и требующую небольшое время на обучение.

# Параметры исследования

1. Модель для получения логитов: Llama-3-8B-Instruct
2. Датасеты: 1) RAID - огромный датасет с атаками. 2) M4GT - датасет для задач детектинга, классификации модели генерации, определения процента сгенерированности с текстами на разных языках.
3. Данные: 18,000 из M4GT ⅔ сгенерированных. 30000 из RAID ½ сгенерированных
4. Модели для классификации:

   1) ROBERTa-Base - baseline.

   2) Логистическая регрессия

   3) Случайный лес (300 деревьев, максимальная глубина дерева = 10)

   4) Нейросетевой Классификатор (4 слоя, Adam, Binary Cross-Entropy loss, 300 эпох)

# Результаты

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

M4GT, arXiv

| Model | Accuracy | ROC-AUC | Train Time (s) |
|---|---|---|---|
| BERT Classifier | 0.9942 | 0.9954 | 1489.0528 |
| Neural Classifier with uncertainty | 0.8183 | 0.7942 | 208.9576 |
| Random forest with uncertainty | 0.8103 | 0.7831 | 6.7727 |
| Logistic Regression with uncertainty | 0.7744 | 0.7317 | 0.0134 |

RAID, Reddit

| Model | Accuracy | ROC-AUC | Train Time (s) |
|---|---|---|---|
| BERT Classifier | 0.9538 | 0.9532 | 2362.1725 |
| Neural Classifier with uncertainty | 0.8987 | 0.8977 | 378.1838 |
| Random Forest with uncertainty | 0.8992 | 0.8987 | 10.7419 |
| Logistic Regression with uncertainty | 0.7271 | 0.7258 | 0.0350 |

# Вывод

Использование методов неопределенности показало хороший результат в задаче детектинга машино-сгенерированного текста. Дальнейшие исследования в теме могут позволить создать точный детектор, устойчивый к многим видам атак.