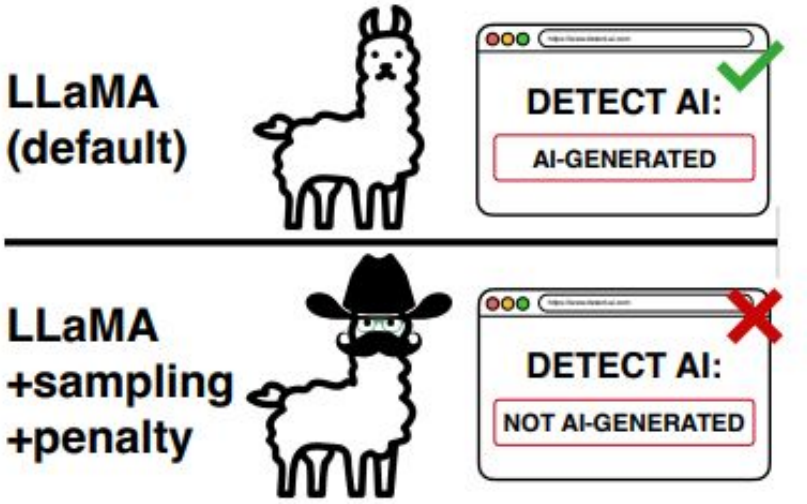
Использование методов подсчета неопределенности для борьбы с атаками на детекторы машинно-сгенерированного текста

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

18.03.2025

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.

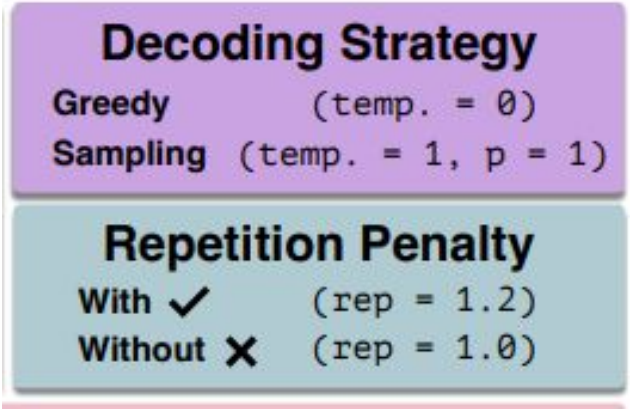


Постановка проблемы

Необходимо предложить метод обнаружения машино-сгенерированного текста устойчивого к различным атакам

1. **Alternative Spelling**: Use British spelling
2. **Article Deletion**: Delete ('the', 'a', 'an')
3. **Add Paragraph**: Put \n\n between sentences
4. **Upper-Lower**: Swap the case of words
5. **Zero-Width Space**: Insert the zero-width space U+200B every other character
6. **Whitespace**: Add spaces between characters
7. **Homoglyph**: Swap characters for alternatives that look similar, e.g. e → e (U+0435)
8. **Number**: Randomly shuffle digits of numbers
9. **Misspelling**: Insert common misspellings
10. **Paraphrase**: Paraphrase with the fine-tuned T5-11B model from Krishna et al. (2023)
11. **Synonym**: Swap tokens with highly similar BERT (Devlin et al., 2019) candidate tokens

атаки после генерации



атаки перед генерацией

Метод и Задача

Метод: Оценка неопределённости (Uncertainty Estimation) - распространенный подход к работе с моделями и их предсказаниями в NLP. Различные методы подсчёта UE помогают понять уверенность модели в своих предсказаниях. UE хорошо зарекомендовала себя в других задачах NLP (QA, TS, MT).

Задача: Исследовать различные методы подсчета неопределенности. Проверить гипотезу, что с их помощью можно уверенно различать рукописные и машинно-сгенерированные тексты, даже при наличии атак.

Особенности работы

Новизна: применение неопределенности для задачи детектинга обширно не исследовалась

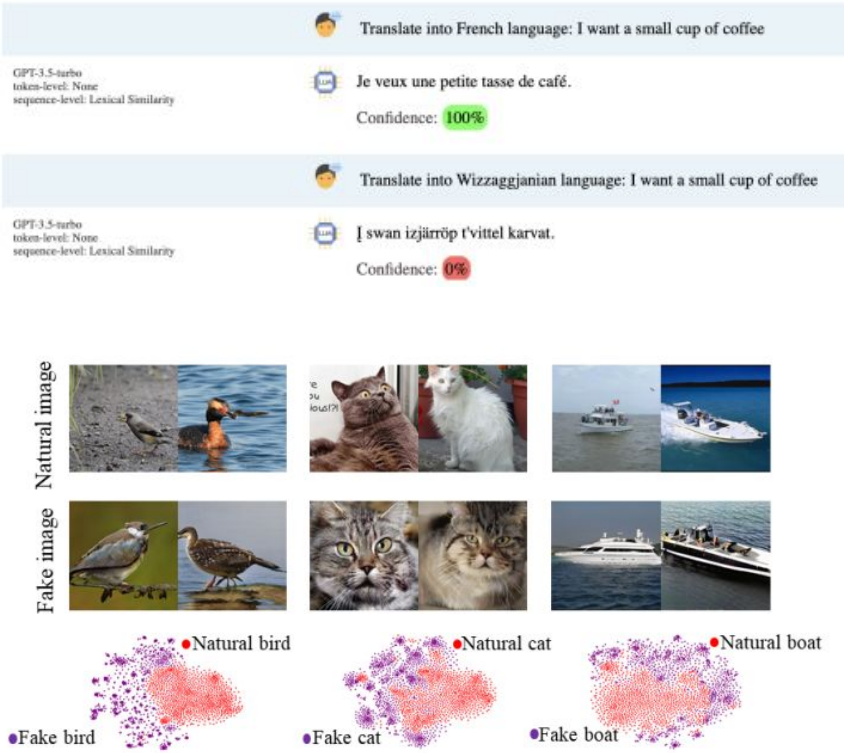
Несколько разных подходов: исследуются два типа методов: **White-box** и **Black-box**. Способы подсчетов и использование моделей различаются

Особое внимание на робастность: требуется, чтобы методы показывали хорошие результаты на датасетах с атаками

Обзор литературы

1. Исследование различных методов подсчета неопределенности для задач NLP

2. Применение неопределенности для обнаружения отличий настоящих картинок от сгенерированных



Датасеты

RAID: огромный датасет с атаками

MAGE: датасет для бинарной классификации с текстами из множества областей

M4GT: датасет для задач детектинга, классификации модели генерации, определения процента сгенерированности с текстами на разных языках

raid-bench.xyz

Применение

Исследование может помочь создать хороший детектор машино-сгенерированного текста устойчивого к атакам.