AI Generated Text Detection Using Deep Learning

Presented By: Sabbir Hossain Ujjal

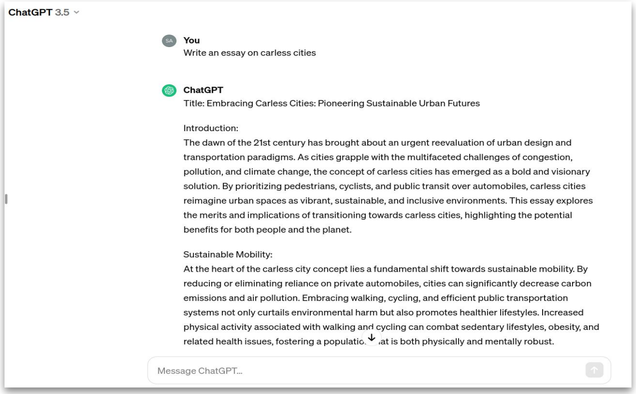
Problem Statement

In this project I developed a Deep Learning based model that can accurately detect whether an essay was written by a student or an LLM which may help the evaluator to take proper action.

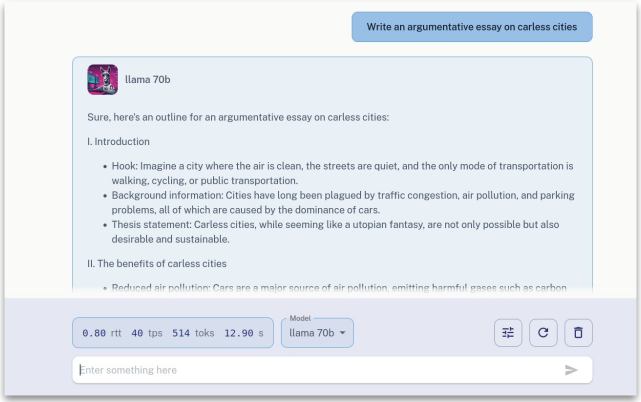
Motivation

- In recent years, large language models (LLMs) have become increasingly sophisticated, capable of generating text that is difficult to distinguish from human-written text.
- Modern LLM are so powerful that students could use LLMs to generate essays that are not their own, missing crucial learning keystones, which also bring significant changes in education system.
- Available AI generated text classifier prone to give false positive which eventually leads to poor performance of students who genuinely written their task.

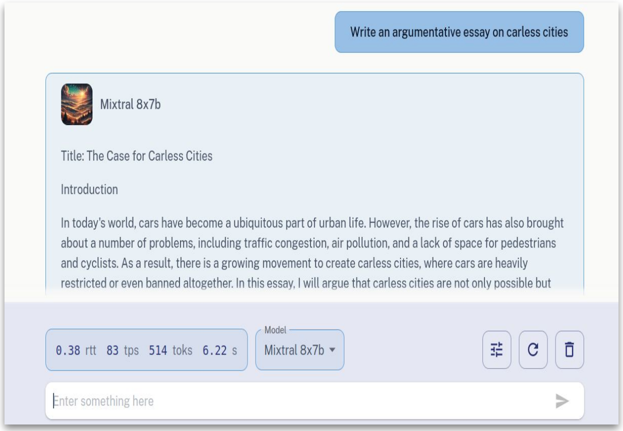
Different Writing Platform that uses different LLMs



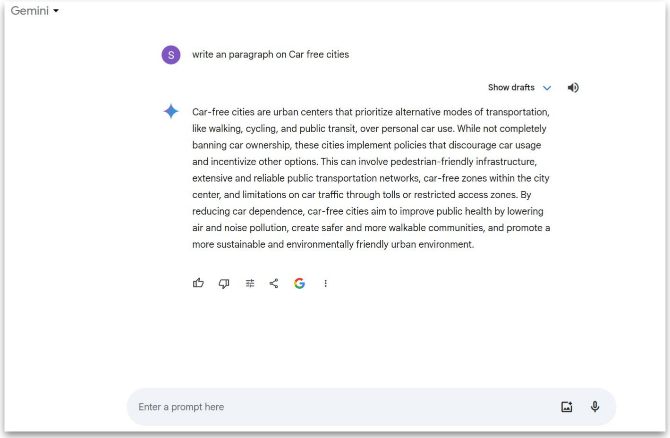
[ChatGPT](#)



[LLAMA Chat](#)



[Mistral Chat](#)



[Gemini Chat](#)

Related Works

- Herbold, S., Hautli-Janisz, A., Heuer, U. *et al.* **A large-scale comparison of human-written versus ChatGPT-generated essays**.
  - **Journal:** *Scientific Reports by Nature*-**2023**
  - In this article authors tried to make comparison between human written essay and **ChatGPT** generated essays by extracting linguistic features and show that how much different is LLM generated text than human written text.

- Heather Desaire, Aleesa E. Chua, Min-Gyu Kim, David Hua, "**Accurately detecting AI text when ChatGPT is told to write like a chemist**,"
  - **Journal:** Cell Reports Physical Science by Science Direct- **2023**
  - In this research authors tried to generate chemical related scientific papers using LLM and compare these with existing paper with the same topic and tried to detect AI generated text using linguistic features.
  - They uses different linguistic features like number of words in the paragraph, sentence complexity etc. for classifying ai generated text.

Dataset

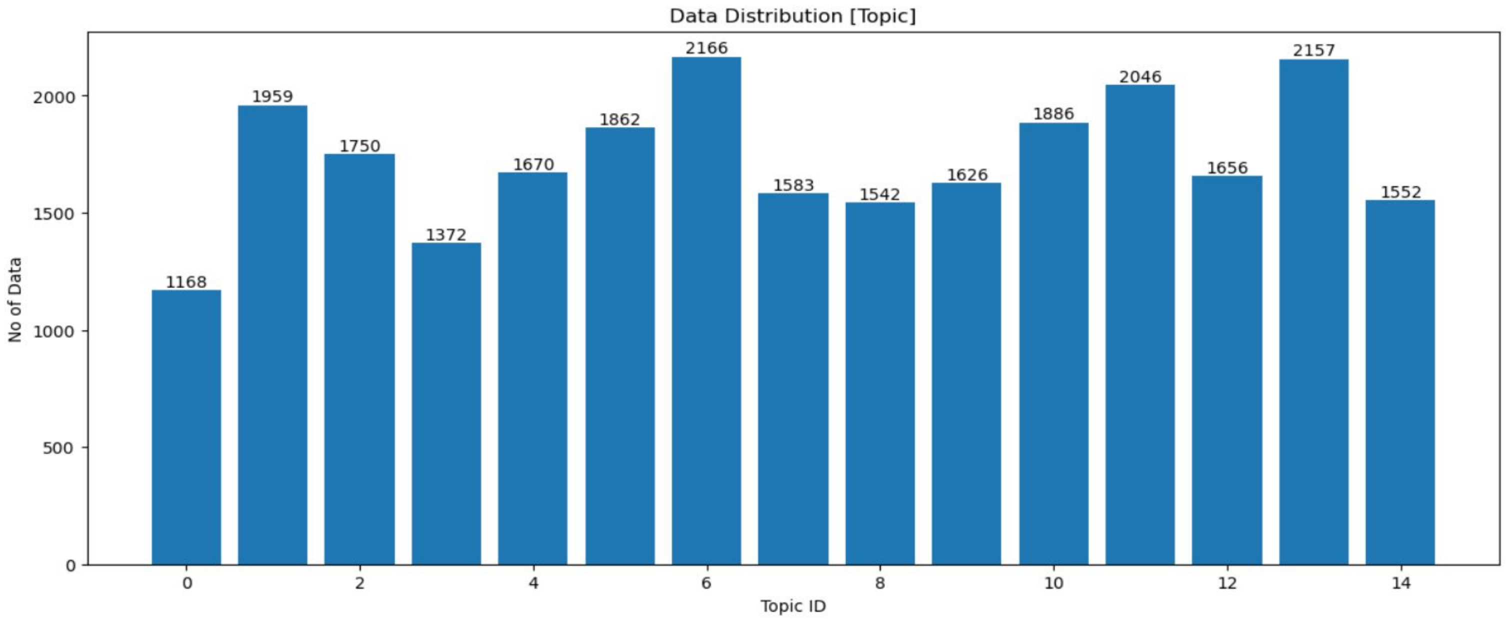
- ❖ As the task is classification task with 2 label we will need data that are human written and data that are AI generated.

For our task we will use two different dataset:

- ❖ Human written text data:
  - ➢ **persuade corpus 2.0 [**https://github.com/scrosseye/persuade_corpus_2.0**]**
  - ➢ This dataset comprises over **25,000 argumentative essays** produced by 6th-12th grade students in the United States for **15 topics**.

Topics of the argumentative essays:

- Phones and driving
- Driverless cars
- Does the electoral college work?
- Cell phones at school
- Seeking multiple opinions
- Car-free cities
- Distance learning
- A Cowboy Who Rode the Waves
- Mandatory extracurricular activities
- Exploring Venus
- Facial action coding system
- The Face on Mars
- Community service
- Grades for extracurricular activities
- Summer projects



Human written Essay counts for each topic

Dataset

- AI generated Data:
  - **For AI generated Data we used different available LLM models for the same topics as the human written text.**
  - **As there is huge rise in LLM and Chat models we tried to use different famous chat models for creating dataset**

Model Used for Data generation:

- **Chat-GPT-3.5**
- **Chat-GPT-4**
- **LLAMA-2**
- **Mistral**
- **Gemini**

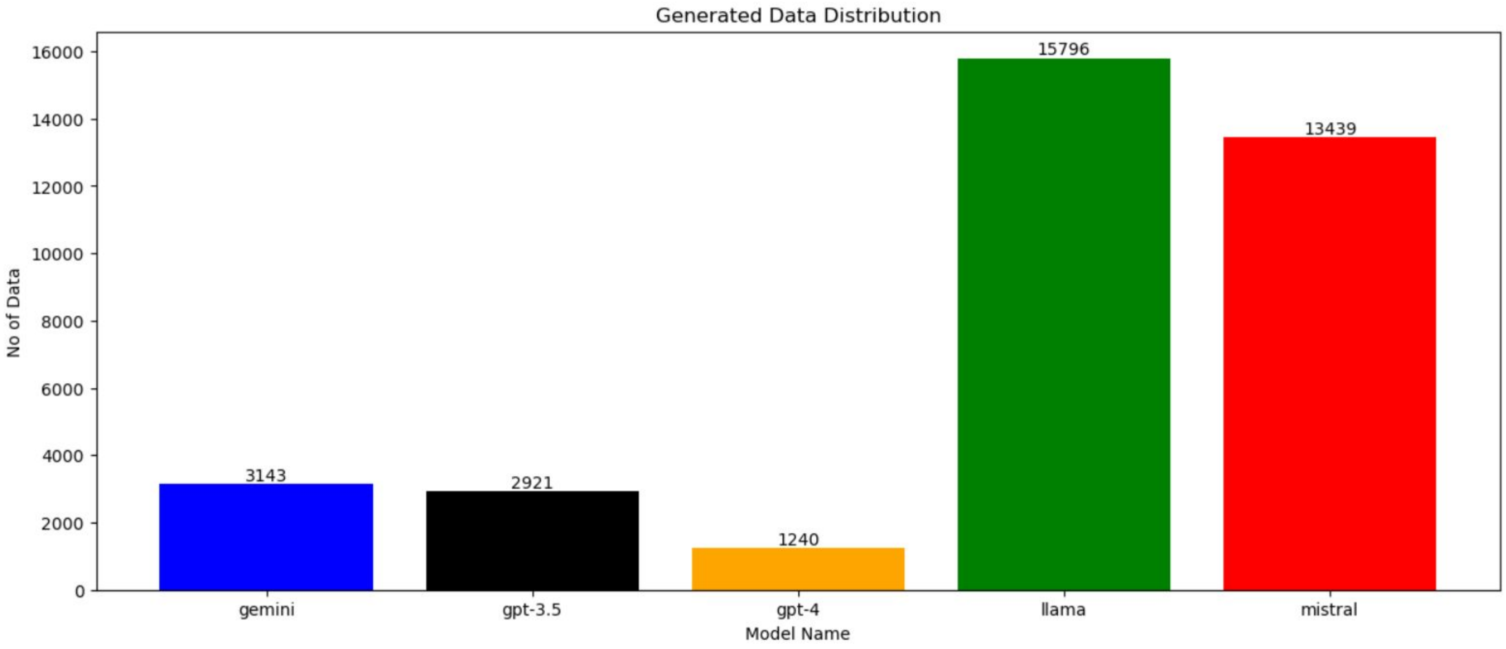
Dataset

Considerations during Data generation:

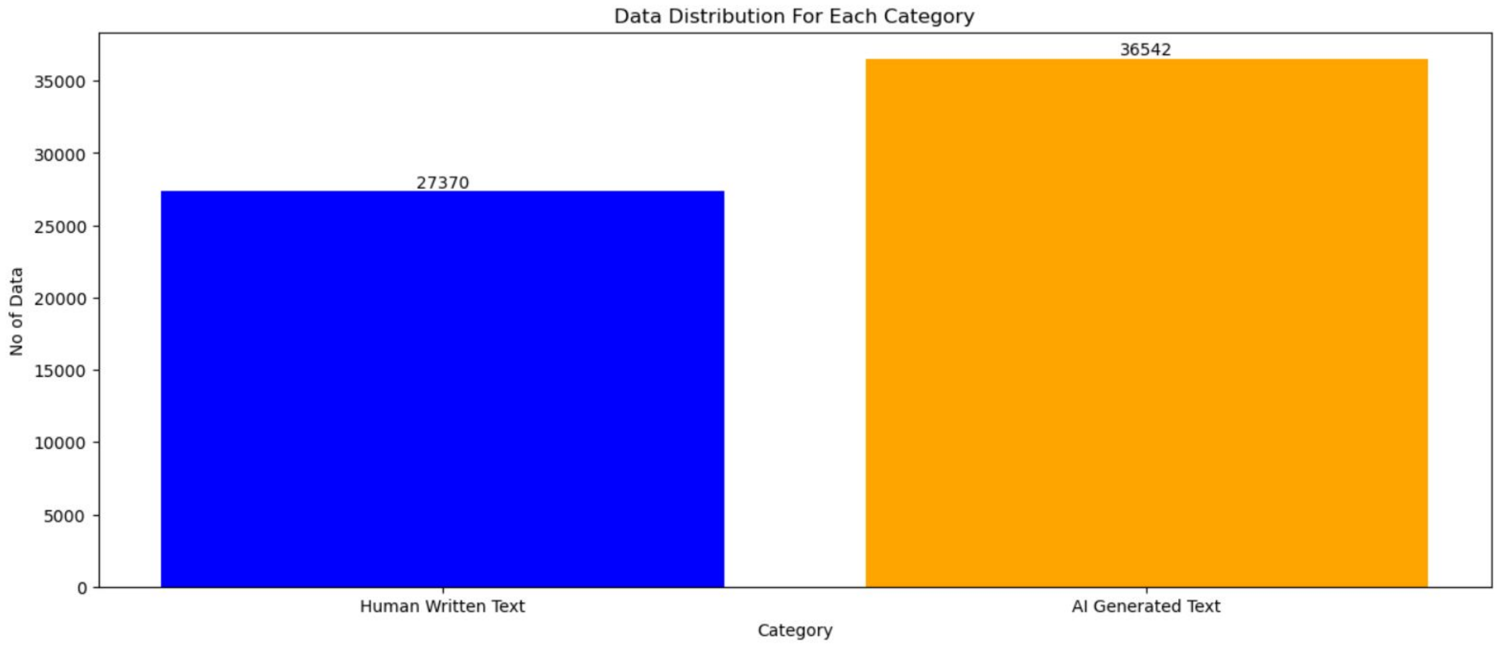
During Data generation we followed the following configuration. During the process we tried to ensure that the generated data are not too much different and driven away from the human generated data.

- Zero shot prompting.
- Few shot prompting for in context generation.
- Instruction Prompting to write as student [With grade level, without grade level]
- Defined Minimum words [150 words]
- Defined and undefined maximum words [500 words]
- Temperature during generation. Randomly choose between 0.5 to 1.5
- Top K = 10 -20
- Top P = 1

AI Generated Text From Different Models



Overall Data Distribution



Modeling Approach

- The task is a **Binary classification task**.
- We used two type of modeling approach for the task.
  - Feature Based ML Model
  - Deep Learning Based Model

ML Modeling:

- For conventional ML model we extracted different features from the dataset. We extracted feature on different level for the model.
  - Paragraph level features
  - Sentence level features
  - Word level features

Word Level Features:

- Number of Words
- Number of Unique Words
- Number of Stop Words
- Number of Uppercase letters
- Number of Nouns
- Number of Verbs
- Number of Adverbs
- Number of Adjectives
- Number of Conjunctions
- Number of Interjections
- Number of other parts of speech

Sentence Level Features:

- Number of Sentence
- Mean Sentence Length
- Mean words per Sentence
- Standard Deviation of words per sentence

Paragraph Level Features:

- Number of Punctuations
- Number of Digits
- Number of Misspelled words
- Lexical Diversity
- Sentence Complexity
- Number of Nominalisation
- Number of Modal Verbs

Modeling Approach

ML Modeling:

We used different ML model for the task.

- Random Forest
- XGBoost
- Support Vector Machine Classifier

Modeling Approach

Deep learning based Modeling:

Deeplearning based natural language processing have advance a lot because of recent invention of transformer based models. Transformer based model are very efficient for capturing context of given text and because of attention mechanism it can give focus on most relevant features of given text.

For our task we leverage different transformer base models

- Bert-base-cased
- Bert-small
- Deverta-V3-small

Training Strategy

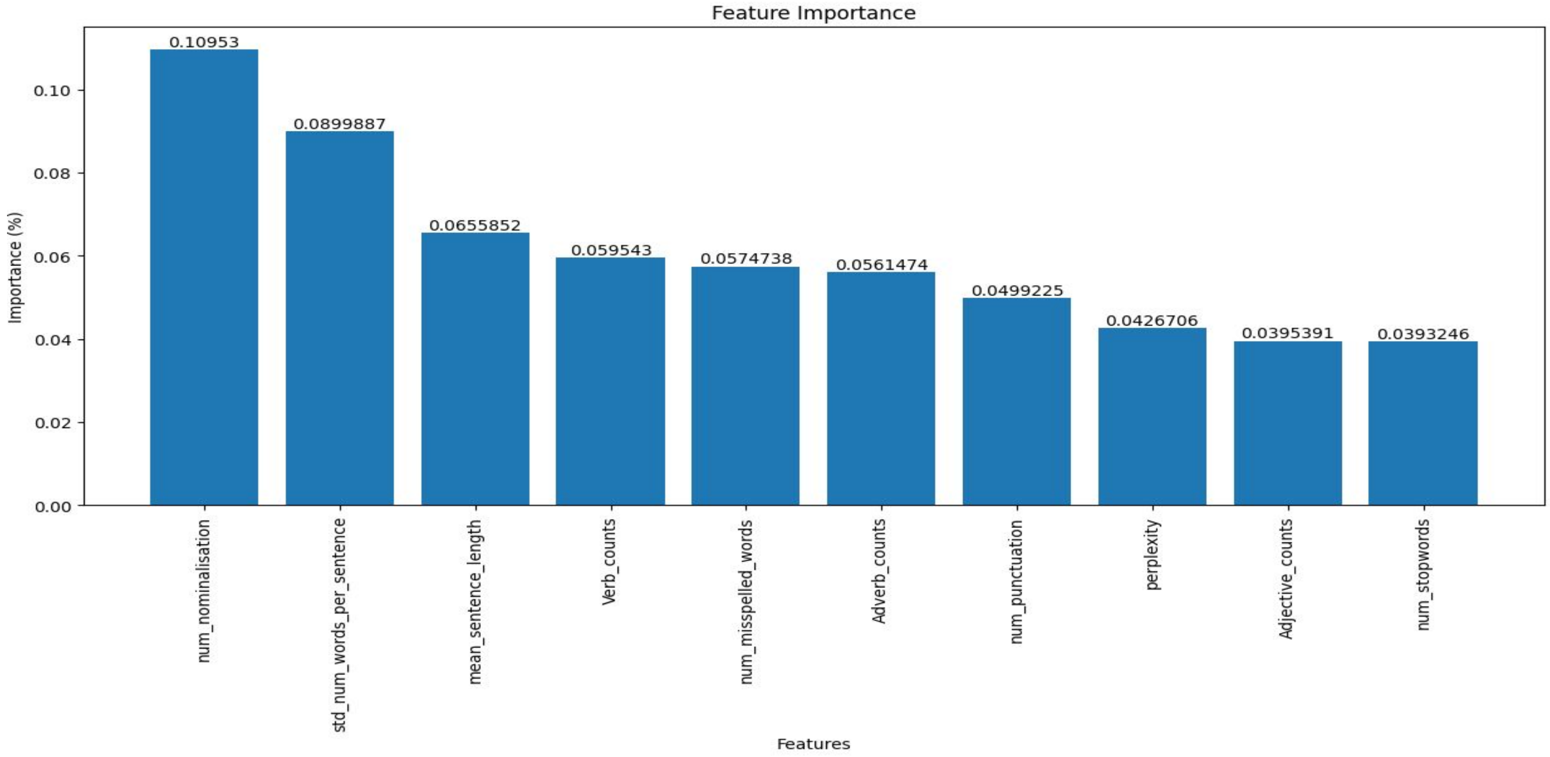
- As we want a robust model that can identify ai-generated text correctly without giving false positive and works on diverse topics, we trained our models on data from 10 topics out of 15 topics.
- We take 3 topics that are not present in training data for validation and rest of the 2 topics for testing purpose.
- By leveraging this training process we can ensure that our model is robust and not depend only the given topics that are present in the dataset.

- For evaluation we used **Accuracy** and **F1 score** as evaluation metrics for model robustness

Result Analysis

Modeling Strategy	Model	Valid Acc	Valid F1	Test Acc	Test F1
ML Based Models	**Random Forest**	0.9187	0.9319	0.8981	0.8981
	XGBoost	0.9170	0.9314	0.9076	0.901
	SVC	0.8965	0.9164	0.8586	0.8460
Deep Learning Based Models	**bert-base-cased**	0.9228	0.9421	0.9139	0.9285
	bert-small	0.8741	0.9035	0.9202	0.9313
	deberta-v3-small	0.9593	0.9764	**0.9645**	**0.9693**

Feature Importance Inspection




THANK YOU!