Meta-Learning to Evade AI-Text Detection with Minimal Edits

Winnie Chow and Ryan Zhao
Department of Computer Science, Stanford University

Summary

Our goal is to train a model that can **evade AI-text detectors** using **minimal edits**.

- We generate a dataset consisting of pairs of **human-written** and **AI-generated** text.
- We adopt **Meta-Agnostic Meta-Learning (MAML)** to train a sentence paraphraser on tasks with different number of edits by finetuning the paraphraser on the dataset.
- Our method is shown to be effective at evading DetectGPT, a zero-shot AI text detector. We are able to decrease the **AUROC w.r.t. human-text vs. AI-text from 0.6795 to 0.3287.**
- Most paraphrases generated are indeed with **minimal edits per sentence.**

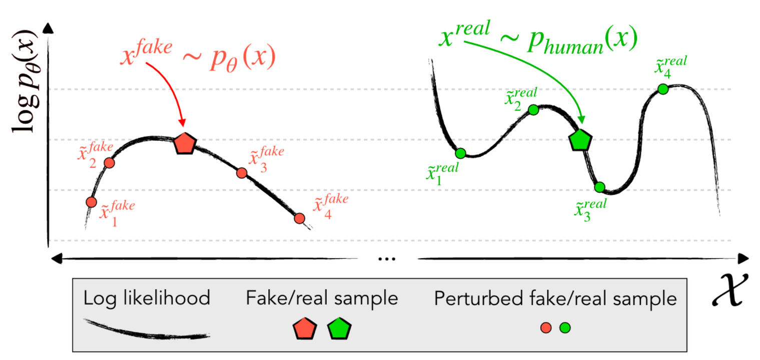
Relevant Background

Problem Setup

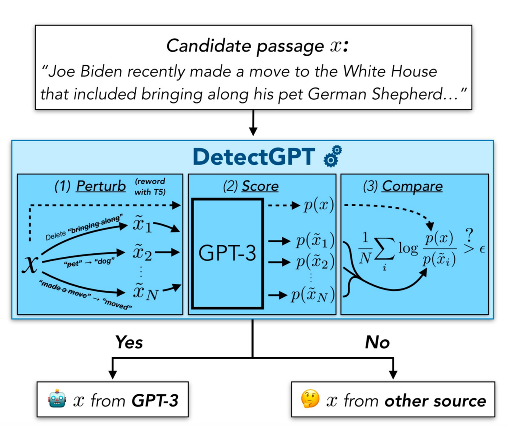
Given AI-generated text X_{ai}, we want to train a model f_θ to generate an edited version X_{edited}, using a minimal number of edits, such that it can evade AI-text detection.

DetectGPT[3]

DetectGPT is a prominent AI-text detection algorithm. It is based on the assumption that, since machine learning models optimize for maximum likelihood, **AI-generated texts are more likely to lie in the negative curvature region of the log probability** of text.



To measure this curvature in log probability space, DetectGPT makes a series of **perturbations** to each sample text, calculates the log likelihood of each perturbation, and assigns each text a **z-score** based on how much the perturbation log likelihoods differ from the original sample. **AI text is expected to have a high score**, and human text is expected to have a score near 0.



Methodology

Methodology Overview

Our approach is to take **AI-generated text as input** and pick a **paraphrasing** of the input that most **resembles human text as output**

Our model is trained on **human-text/AI-text pairs** and learns to recreate the human text given the AI text. We train a **sentence-level paraphrasing model** using meta-learning to accomplish this goal.

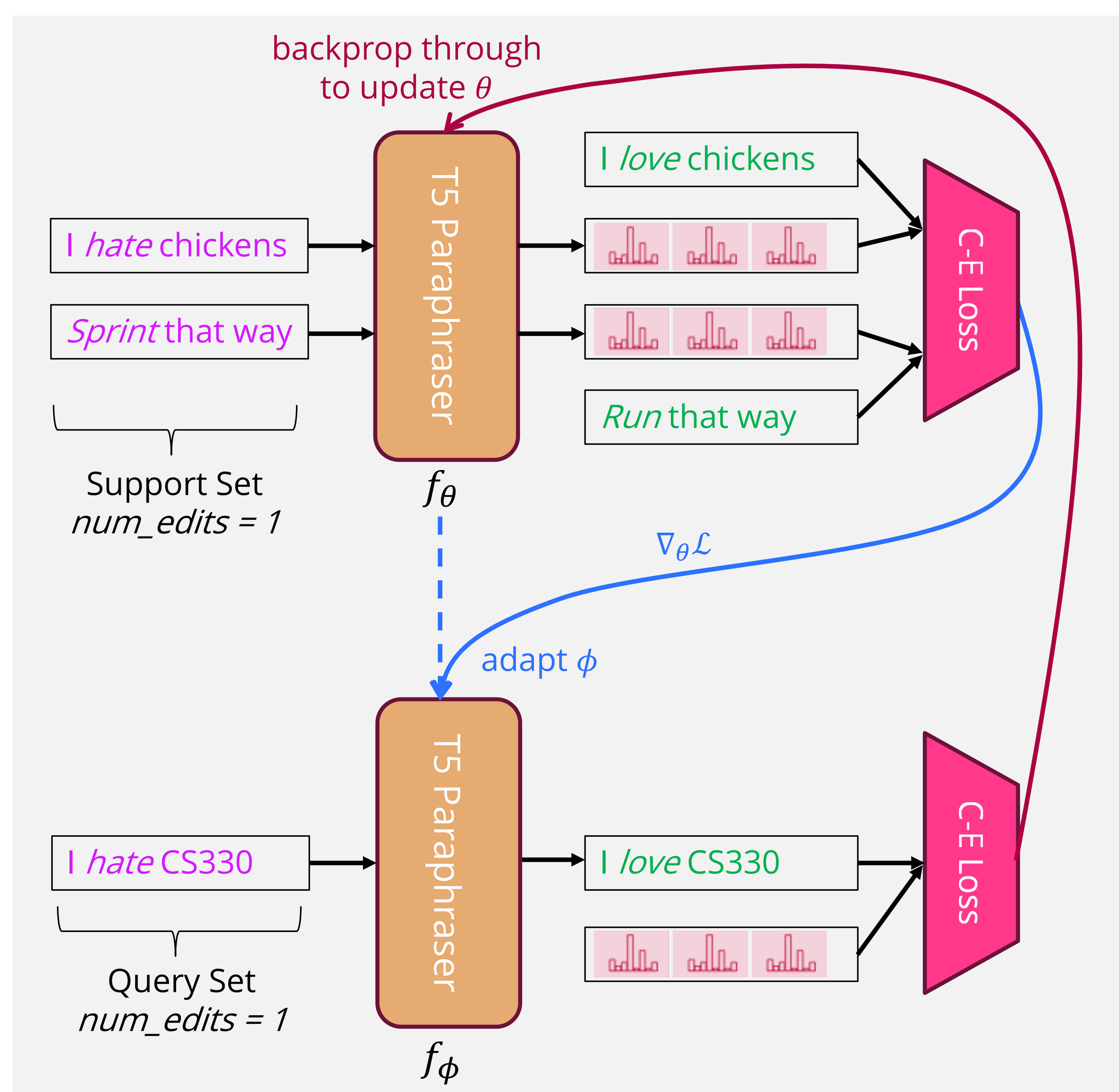
Dataset

The data we use for model training and evaluation consists of **Wikipedia article. introductions**[1] To generate our training data, we perform the following for each introduction:

1. Separate each text $X_{human} \in \boldsymbol{X}_{human}$ into sentences $\{x_{human}\} = X_{human}$
2. Pass each sentence x_{human} into an AI paraphrasing model to get x_{ai}
3. Create **dataset of sentence pairs** (x_{human}, x_{ai})

For paraphrasing, we use a T5 model pretrained on the paraphrasing task using the Google PAWS dataset.

To generate our test data, we use a held-out portion of the Wikipedia dataset $\boldsymbol{Z}_{human}$. For each text $Z_{human} \in \boldsymbol{Z}_{human}$, we prompt **GPT-3-davinci to generate a "Wikipedia-style intro"** and include the first 7 tokens of the human text, yielding AI-text $\boldsymbol{Z}_{ai}$.

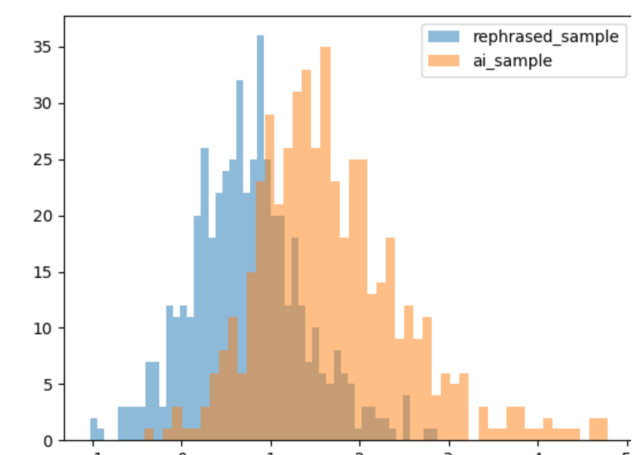


Results

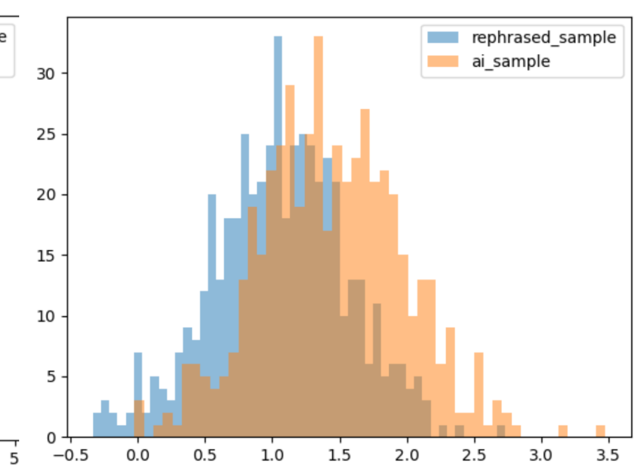
We ran our trained model on the **200 held-out examples** $\boldsymbol{Z}_{ai}$ from the Wikipedia dataset and evaluated the resulting paraphrased text on its ability to evade DetectGPT.

1. Separate each text $Z_{ai} \in \boldsymbol{Z}_{ai}$ into sentences $\{z_{ai}\} = Z_{ai}$
2. Pass each sentence z_{ai} through our model to generate $\tilde{x}_{edited} = f_\theta(z_{ai})$
3. Recombine $\{z_{edited}\}$ into Z_{edited} and run through DetectGPT to calculate score

DetectGPT scores were generated using **GPT-Neo 1.3B** as the scoring model



DetectGPT scores: Trained Model

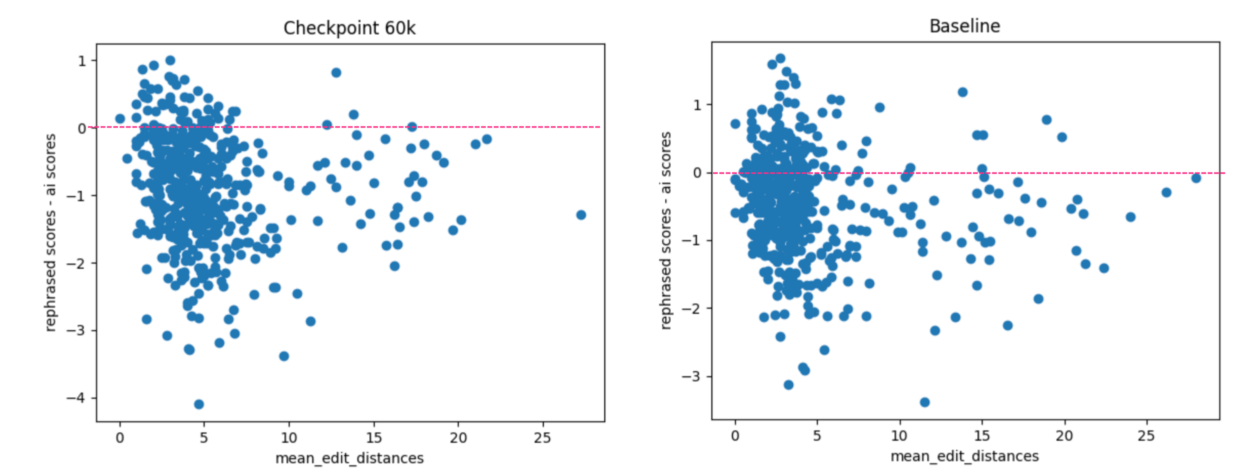


DetectGPT scores: Pretrained T5 Paraphraser

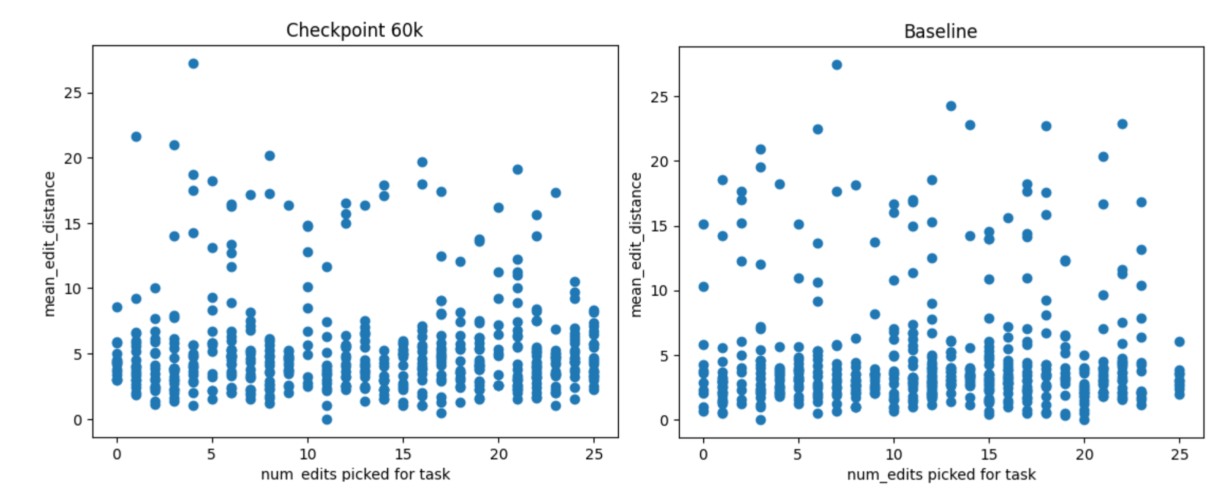
Qualitatively, we can see that the output of our model has a lower distribution of DetectGPT scores than passing AI text through a generic paraphraser. We can also measure this effect quantitatively with AUROC

Human vs.	Our Model	Pretrained T5 Paraphraser	Unedited GPT-3-davinci
AUROC	**0.3287**	0.5148	0.6795

We also examined how our model performed relative to the number of edits it made. We found that a vast **majority of the edits made by our model were low in edit distance**. Additionally, despite only changing a small number of words, they almost always resulted in lower DetectGPT scores.



Finally, we investigated if the **support set effected the number of edits** performed by our model. Unfortunately, there does not appear to be a significant correlation between the number of edits chosen for the support set and the number of edits performed on the query. This is likely due to **"number of edits" being too weak of a signal for the model to learn** from only 4 support examples.



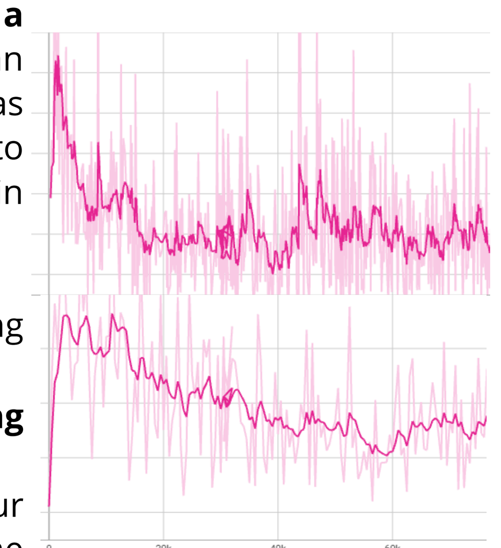
Methodology (continued)

Model Training

Our model uses the **architecture of a 223M-parameter T5 model**. We began with a pretrained T5 model that was finetuned on the Google PAWS dataset to generate sentence-level paraphrases in English.

We perform supervised finetuning using **Meta-Agnostic Meta-Learning (MAML)[2]**.

- Different **tasks are defined as editing with different numbers of edits**
- Each outer loop iteration consists of four support examples (x_{human}, x_{ai}) and one query example
- We use **cross-entropy loss between** $\boldsymbol{x_{human}}$ **and** $\boldsymbol{x_{edited}} = f_\theta(x_{ai})$



Top: Training Loss
Bottom: Validation Loss

To stay within compute constraints, we use **LoRA** to update only low rank matrices within each layer.

References

1. Aaditya Bhat. Gpt-wiki-intro (revision 0e458f5), 2023. URL https://huggingface.co/datasets/aadityaubhat/GPT-wiki-intro.

2. Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks, 2017.

3. Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. Detectgpt: Zero-shot machine-generated text detection using probability curvature, 2023.

4. Sai Vamsi Alisetti. Paraphrase-Generator. https://github.com/Vamsi995/Paraphrase-Generator, 2020.

5. Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In International Conference on Learning Representations, 2022.