EMNLP 2024

Detecting LLM-Assisted Cheating on Open-Ended Writing Tasks on Language Proficiency Tests

Chenhao Niu, Kevin P. Yancey, Ruidong Liu,
Mirza Basim Baig, André Kenji Horie, James Sharpnack

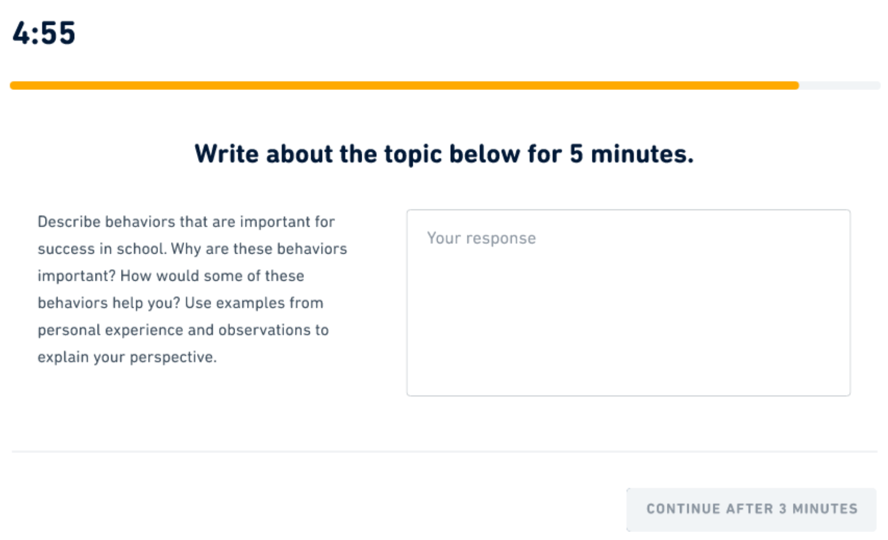
Duolingo, Inc.

Motivation

MOTIVATION

Context: Open-Ended Writing

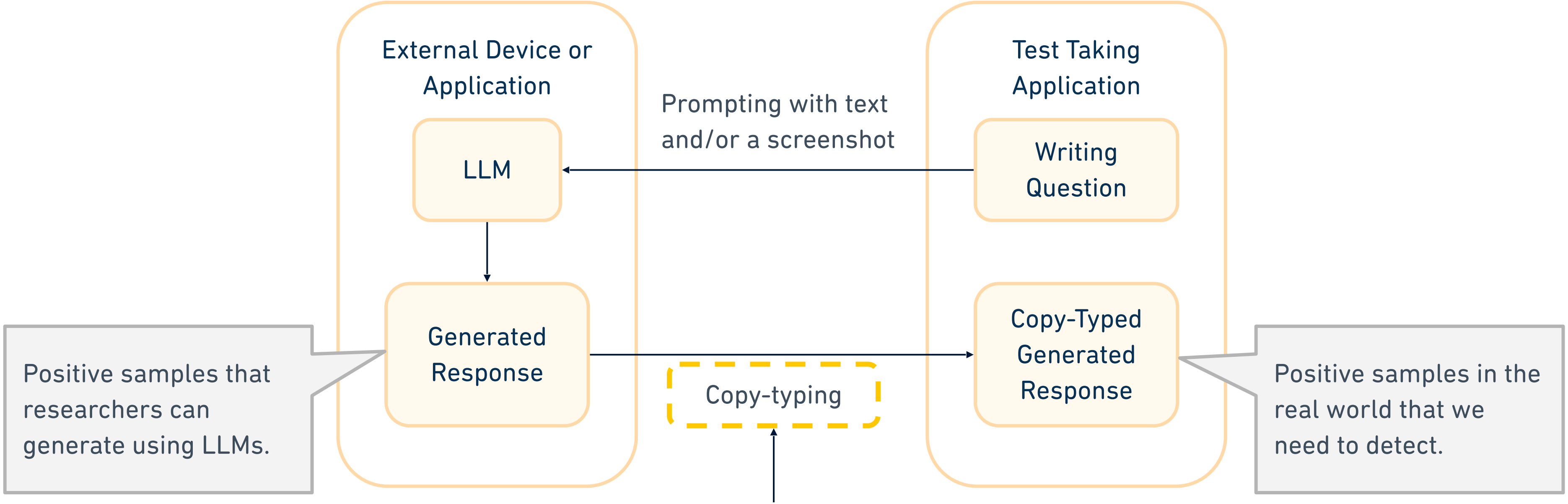
- **Application**: Duolingo English Test (Cardwell et al., 2024), a high-stakes online English proficiency test.
- **Security features include**:
    - Screen and video recording.
    - Remote asynchronous proctoring.
    - Copy-pasting is disabled.



A screenshot of an open-ended writing question in the Duolingo English Test

MOTIVATION

Problem: Detecting LLM-Assisted Cheating



Various modifications can be introduced: typos, omissions, word replacements, and/or being cut off due to time limit.

MOTIVATION

Example: Copy-Typing Modifications

Question: Describe behaviors that are important for success in school. Why are these behaviors important? How would some of these behaviors help you?

Generated by GPT-4 (OpenAI, 2023):

Key behaviors for success in school include time management and active participation. Managing time well helps complete tasks without stress, while engaging in class discussions improves understanding.

In my experience, planning study time boosted my performance, and students who participated more grasped the material better. These habits are crucial for academic success.

Additional text

Typos, omissions, replacement

Being cut off due to time limit

Copy-typed within 5 minutes:

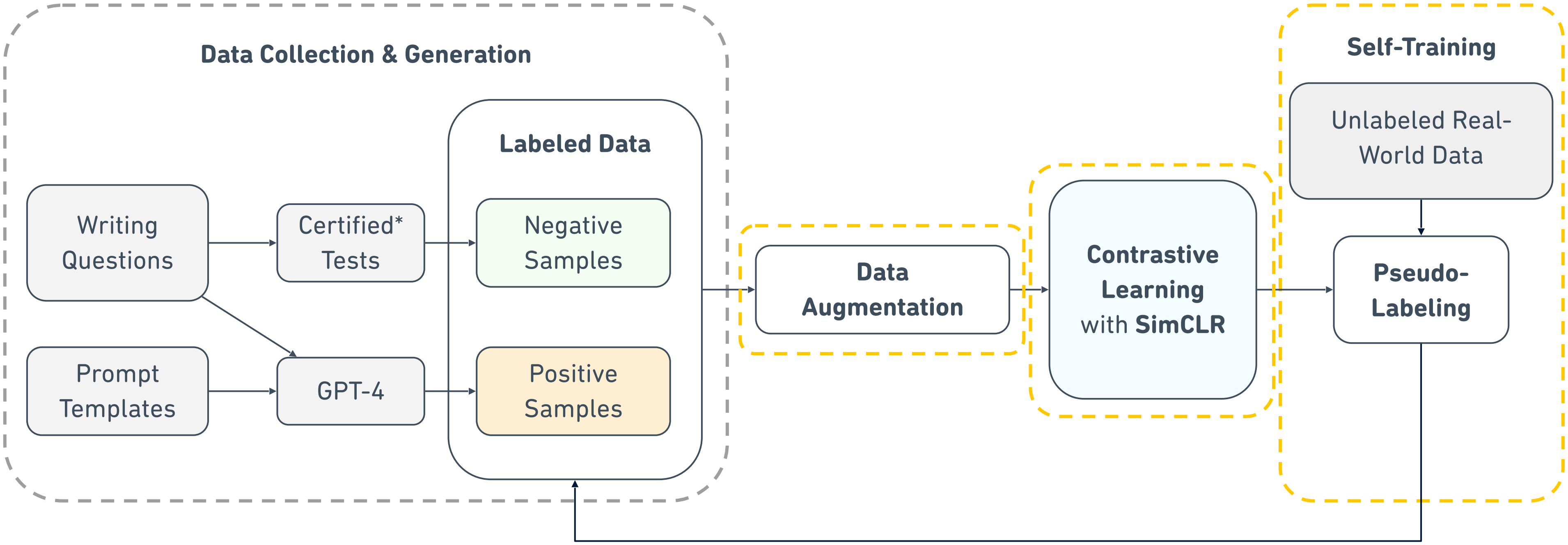
That is an interesting questi**no**. In my opinion, key behavior~~s~~ for success in school inc**ul**de time management and active partic~~i~~pation. Managing time ~~**well**~~ helps complete task~~s~~ without stress, **when** engaging in class discus~~s~~ion improves understanding.

In my experience, planning study time boost~~**ed**~~ my perform**e**nce, and students who participate~~d~~ more graspe**dd** ~~the material better. These habits are crucial for academic success.~~

Method

METHOD

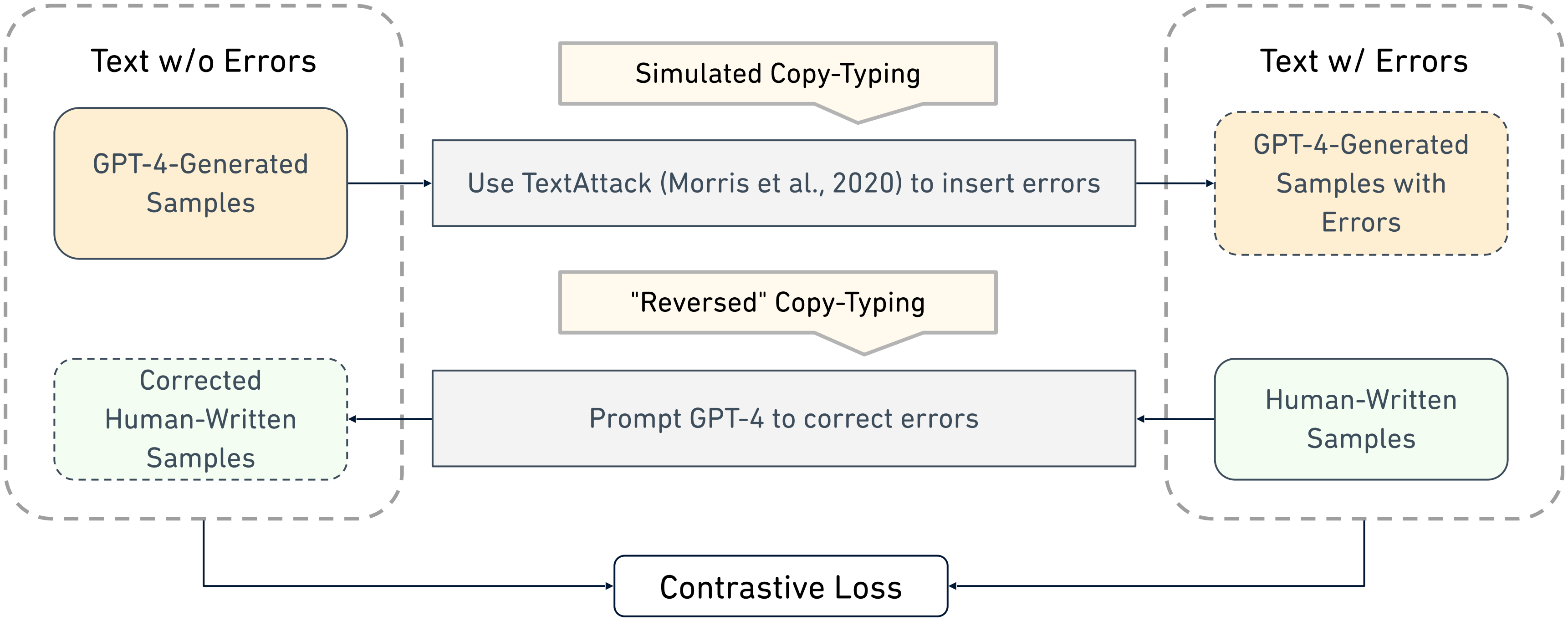
Method Overview



*: A test is certified if no violation is found during the proctoring process.

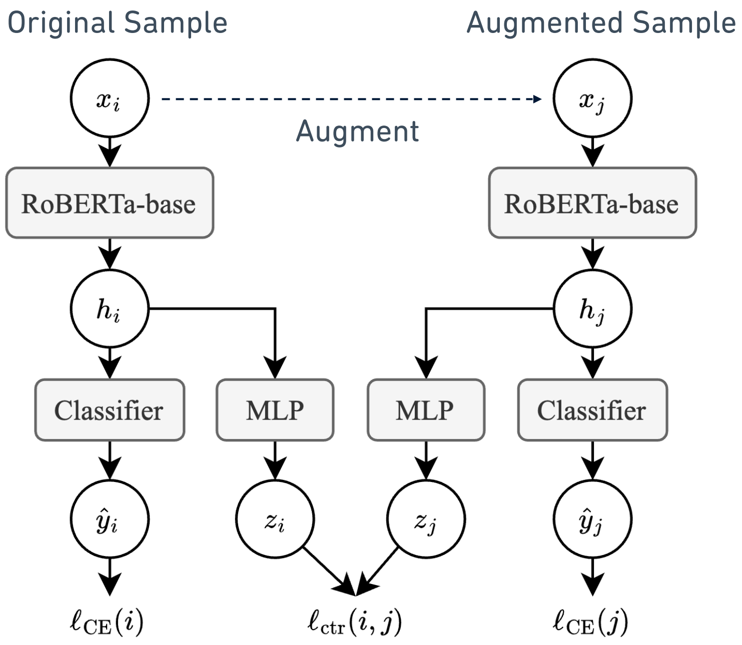
METHOD

Data Augmentation



Contrastive Learning

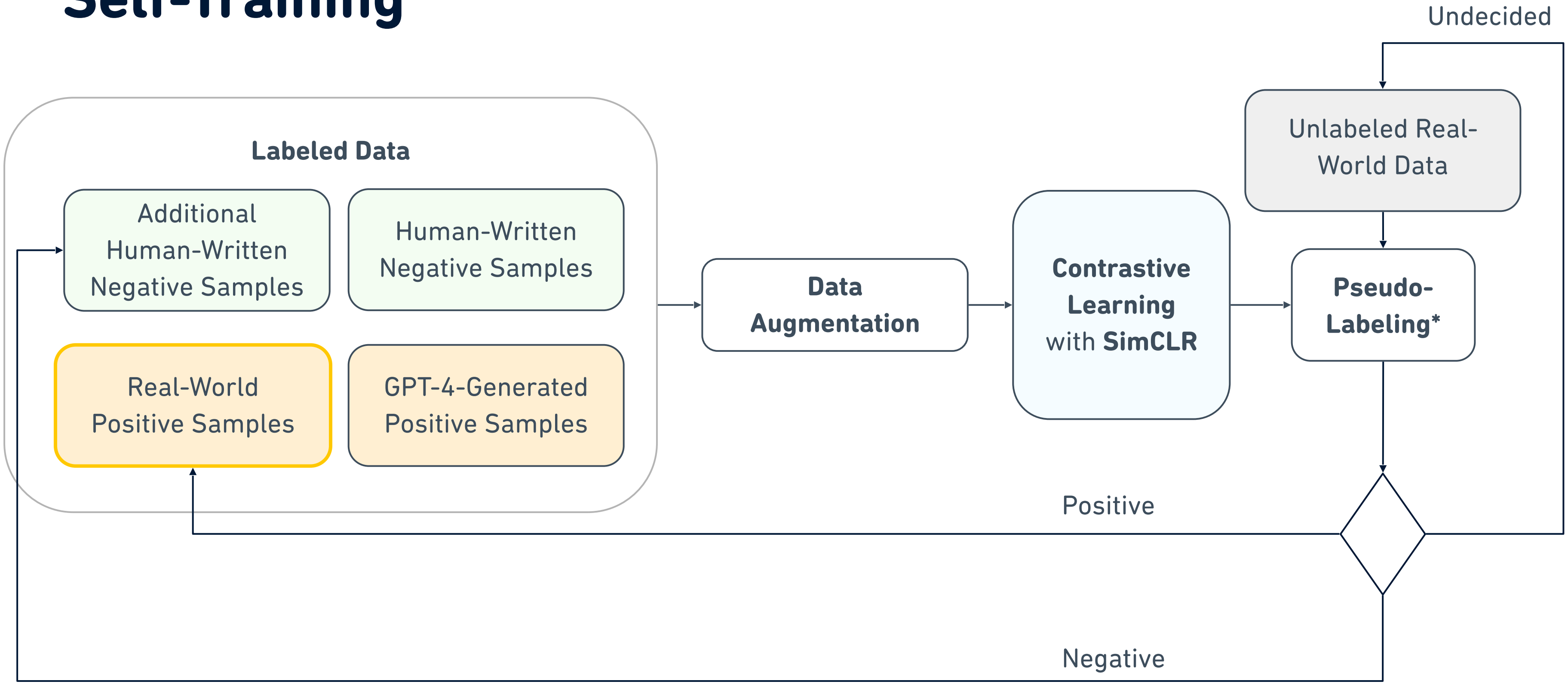
- **Base Model**: pretrained RoBERTa-base (Liu et al., 2019).
- **Framework**: SimCLR (Chen et al., 2020).
- **Intuition**:
  Text embeddings should be similar regardless of copy-typing errors.
- **Goal**:
  The classification result should be robust towards copy-typing errors.



Model Architecture

METHOD

Self-Training



*: pseudo-labeling using class-balanced self-training (Zou et al., 2018).

Results

RESULTS

Baselines and Models

- **Baselines**: OpenAI Detector (Solaiman et al., 2019), GPTZero (Tian and Cui, 2023)
- **Evaluated Models**:

	Model	Data Augmentation	Loss	Self Training
Ablation Study →	RoBERTa_{naive}	None	Binary Cross Entropy	No
	RoBERTa_{err}	Error insertion	Binary Cross Entropy	No
	RoBERTa_{ctr}	Error insertion & correction	SimCLR Loss	No
Proposed →	RoBERTa_{ctr+st}	Error insertion & correction	SimCLR Loss	Yes

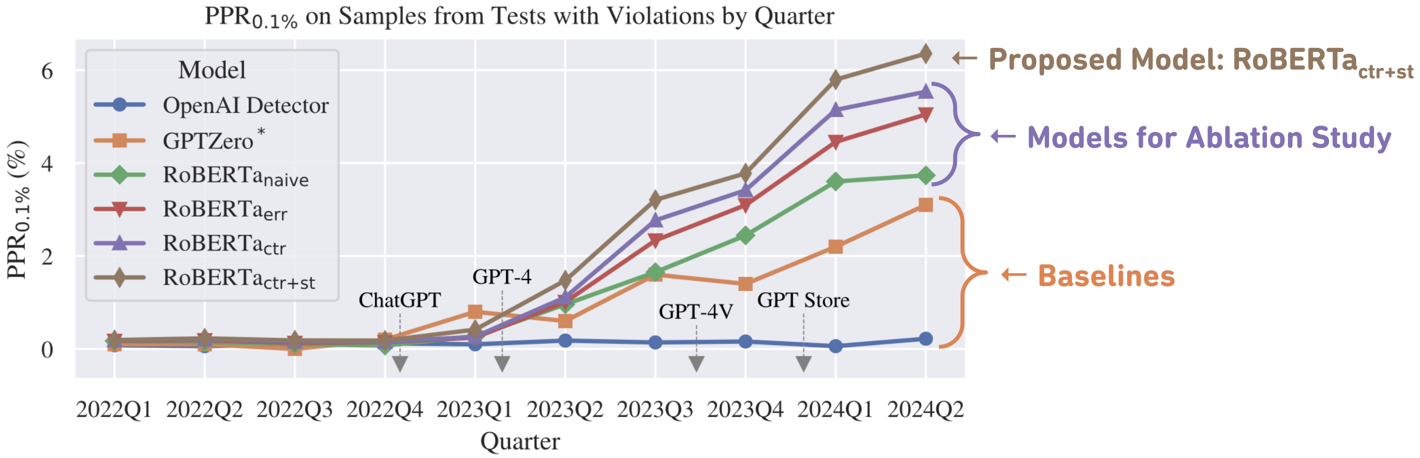
RESULTS

Datasets

	Purpose	# Positive	# Negative	# Unlabeled
Train Set	Training	5,338	5,338	None
Val Set	HP tuning and early stopping	1,776	1,776	None
Test Set	Evaluation on unmodified LLM-generated samples	1,786	100,000	None
Dev Set	Unlabeled data for self-training	None	None	150,000
Samples from Tests with Violations	Evaluation on real-world copy-typed LLM-generated samples	None	None	5,000 per quarter since 2022

RESULTS

Positive Predictions Over Time



Dataset: 5,000 samples from tests with violations in each quarter.

$PPR_{0.1\%}$: the proportion of positive predictions at 0.1% false positive rate.

Observations:

- Upward trends of $PPR_{0.1\%}$ Over time.
- In-domain fine-tuning is useful.
- Improvement from contrastive learning and self-training.

Summary

- LLM-assisted cheating often involves manual modifications due to copy-typing, making it harder to detect than unaltered LLM-generated text.

- We proposed a framework for training classifiers to detect the LLM-generated samples even after they have been copy-typed.

- Evaluation on a real-world dataset from the Duolingo English Test shows our improved model outperforms the original transformer-based classifier and other baselines.

Reference

1. Ramsey Cardwell, Ben Naismith, Geoffrey T. LaFlair, and Steven Nydick. 2024. Duolingo english test: technical manual. Duolingo Research Report.
2. Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In International conference on machine learning, pages 1597–1607. PMLR.
3. John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 119–126.
4. OpenAI. 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774.
5. Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. arXiv preprint arXiv:1908.09203.
6. Edward Tian and Alexander Cui. 2023. Gptzero: Towards detection of ai-generated text using zero-shot and supervised methods.
7. Yang Zou, Zhiding Yu, B.V.K. Vijaya Kumar, and Jinsong Wang. 2018. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In Proceedings of the European Conference on Computer Vision (ECCV).

EMNLP 2024

Thanks for Listening!