ReMoDetect: Reward Models Recognize Aligned LLM's Generations

Hyunseok Lee*, Jihoon Tack*, Jinwoo Shin *Equal contribution

TL; DR. Observe reward models recognize aligned LLM-generated texts (LGTs) and continually train reward model for effective and robust aligned LLM-generated text detection.

Introduction

Societal risk of LLM-generated text (LGT)

- e.g., fake news generation, academic corruption

Generalizability of Detecting LGTs

- Need to detect vast numbers of unseen LLMs.

Common Characteristics of LLMs

- Modern LLMs are aligned to human preference.

Research Question: Identify common characteristics of LGTs and find effective ways to detect them.

ReMoDetect: SOTA LGT Detector

(a) Fast-DetectGPT benchmark [12]: PubMed, XSum, and WritingPrompts-small (WP-s)

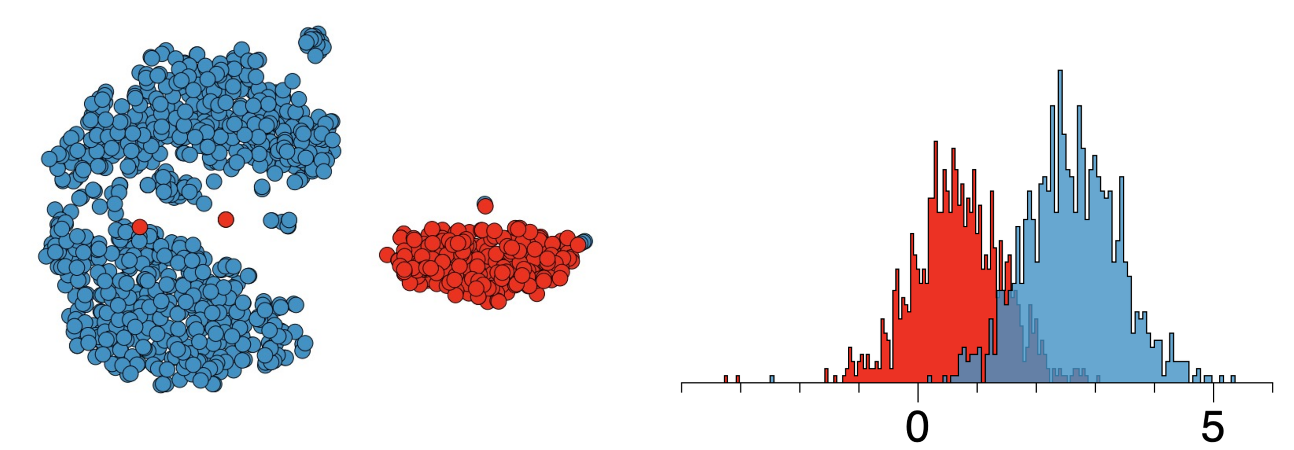
Model	Domain	Loglik.	Rank	D-GPT	LRR	NPR	FD-GPT	Open-D	Chat-D	Ours
GPT3.5 Turbo	PubMed	87.8	59.8	74.4	74.3	67.8	90.2	61.9	21.9	**96.4**
	XSum	95.8	74.9	89.2	91.6	86.6	99.1	91.5	9.7	**99.9**
	WP-s	97.4	80.7	94.7	89.6	94.2	99.2	70.9	27.5	**99.8**
GPT4	PubMed	81.0	59.7	68.1	68.1	63.3	85.0	53.1	28.1	**96.1**
	XSum	79.8	66.4	67.1	74.5	64.8	90.7	67.8	50.3	**98.7**
	WP-s	85.5	71.5	80.9	70.3	78.0	96.1	50.7	45.3	**98.8**
GPT4 Turbo	PubMed	86.5	60.8	63.6	73.5	63.7	88.8	55.8	31.0	**97.0**
	XSum	90.9	73.4	83.2	87.9	81.8	97.4	88.2	4.4	**100.0**
	WP-s	97.6	80.8	92.8	92.9	92.5	99.4	72.3	22.5	**99.8**
Llama3 70B	PubMed	85.4	60.9	66.0	71.3	65.0	90.8	52.9	35.1	**96.3**
	XSum	97.9	74.9	93.2	95.5	93.8	99.7	96.2	7.1	**99.8**
	WP-s	97.1	77.9	95.5	90.1	95.8	**99.9**	77.5	28.1	99.5
Gemini pro	PubMed	83.0	58.3	63.2	75.0	66.8	82.1	57.3	39.3	**86.4**
	XSum	78.6	44.5	72.8	73.0	**79.6**	79.5	72.2	54.7	74.5
	WP-s	75.8	63.0	77.8	72.7	81.1	78.0	70.2	48.0	**86.4**
Calude3 Opus	PubMed	85.5	60.3	66.3	74.3	64.4	88.2	48.9	33.1	**96.4**
	XSum	95.9	71.1	85.3	89.7	84.7	96.2	86.2	5.3	**99.9**
	WP-s	93.8	75.0	91.9	86.5	91.8	93.5	65.7	24.1	**99.5**
Average	-	88.6	67.4	79.2	80.6	78.7	91.9	68.9	28.6	**95.8**

ReMoDetect outperforms other detectors in detecting LGTs in unseen LLM, unseen domains.

Motivation & Observations

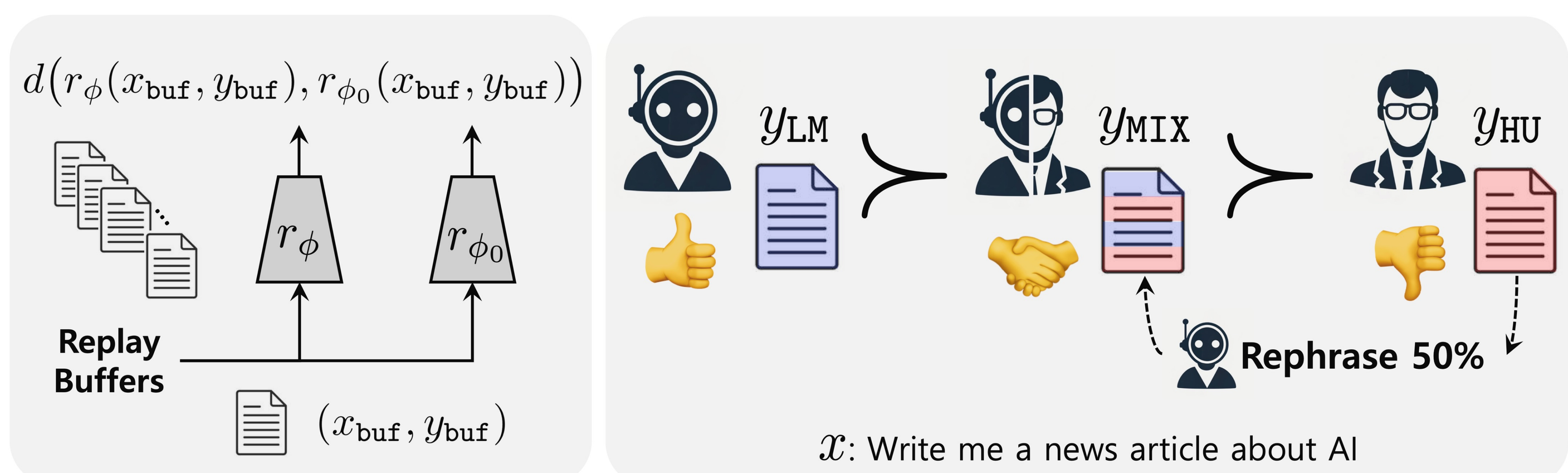
Aligned LLMs optimized to maximize preferences.
↔ LGTs have higher rewards than human-written texts.

Visualization of Vanilla Reward Model Features



Method: Continual Preference Tuning

Improve the detection ability of the reward model



Continual Preference Tuning | Reward Modeling with Mixed Responses

1. Continual Preference Tuning.

Finetune the reward model with LLM/Human text pairs.

$$L_{RM}(x, y_w, y_l) := -\log\sigma\left(r_\phi(x, y_w) - r_\phi(x, y_l)\right)$$

$$L_{\text{cont}} := L_{RM}(x, y_{LM}, y_{HU}) + \lambda d\left(r_\phi(x_{\text{buf}}, y_{\text{buf}}), r_{\phi_0}(x_{\text{buf}}, y_{\text{buf}})\right)$$

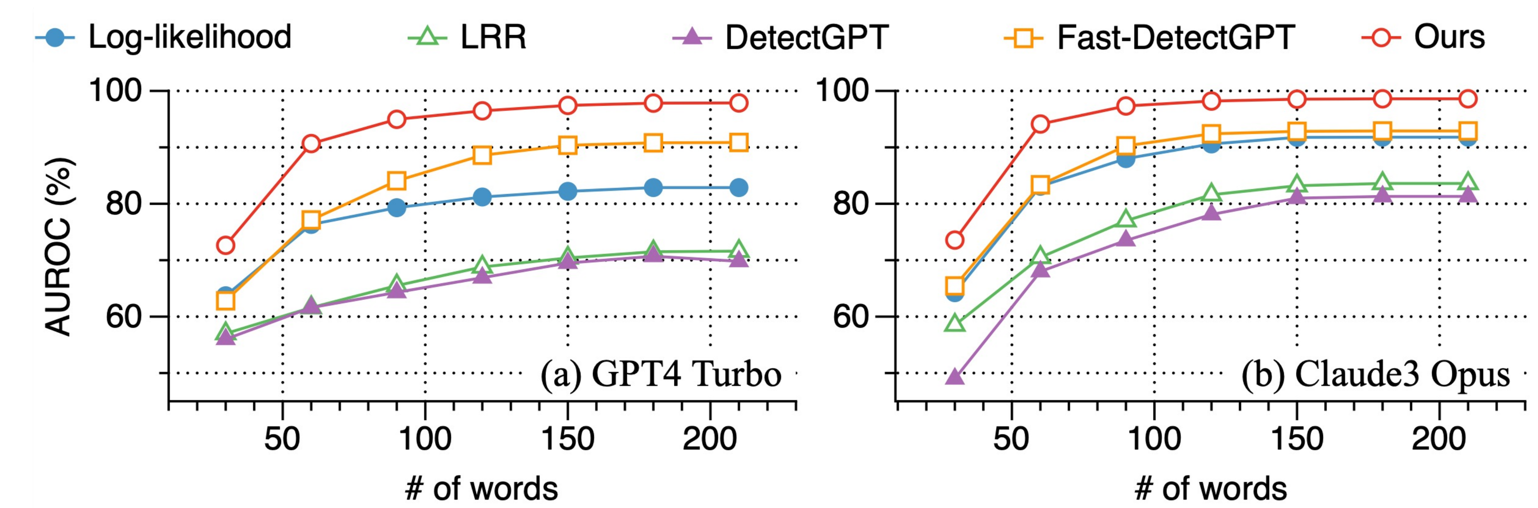
2. Reward Modeling with Mixed Responses

Synthesize the Human/LLM mixed text to enable the detector to learn a better decision boundary.

$$L_{\text{ours}} := L_{\text{cont}} + \beta_1 L_{RM}(x, y_{MIX}, y_{HU}) + \beta_2 L_{RM}(x, y_{LM}, y_{MIX})$$

Experimental Results

Robustness against text Length



Robustness against Paraphrasing Attack

Model	Accuracy	Loglik.	D-GPT	NPR	FD-GPT	Ours
GPT3.5 Turbo	Original	93.6	86.1	82.9	96.1	**98.7**
	Attacked	80.5 (-14.0%)	60.3 (-30.0%)	73.5 (-11.3%)	87.2 (-9.3%)	**91.4 (-7.4%)**
GPT4 Turbo	Original	91.7	79.9	79.4	95.2	**98.9**
	Attacked	80.0 (-12.7%)	50.3 (-37.0%)	61.3 (-22.8%)	87.3 (-8.3%)	**94.6 (-4.4%)**
Claude3 Opus	Original	91.7	81.1	80.3	92.6	**98.6**
	Attacked	80.5 (-15.8%)	55.2 (-32.0%)	60.1 (-25.2%)	81.6 (-11.9%)	**91.1 (-7.1%)**

ReMoDetect is efficient

Method	Detection Time (secs)	Model Parameters	AUROC
Log-likelihood	11.7	2.7B	88.6
DetectGPT	7738.8	3B & 2.7B	79.2
NPR	7837.3	3B & 2.7B	78.7
Fast-DetectGPT	62.7	6B & 2.7B	91.9
Ours	**8.7**	**0.5B**	**95.8**

Outperforms Commercial Detector

Model	GPT 3.5 Turbo	GPT4	GPT4 Turbo	Llama3 70B	Gemini pro	Claude3-Opus
GPTZero	93.5	88.5	95.7	96.6	**82.9**	95.7
Ours	**98.7**	**97.9**	**98.9**	**98.5**	82.4	**98.6**

Summary of Contribution

1. Hypothesized and proven LGT commonly have higher rewards than humans.
2. ReMoDetect outperforms other methods.
3. ReMoDetect is robust and efficient.