ReMoDetect: Reward Models Recognize Aligned LLM's Generations

NeurIPS 2024 Poster

Hyunseok Lee*, Jihoon Tack*, Jinwoo Shin

Presenter: Hyunseok Lee

Hyunseok Lee

Education

- M.S. in KAIST AI, Mar. 2024 - current
  Advisor: Prof. Jinwoo Shin
- B.S. in KAIST EE and CS(double major), Mar. 2018 - Feb. 2024

Research Interest:

- LLM Safety, LLM Agent, Korean LLM

Publication:

- "ReMoDetect: Reward Models Recognize Aligned LLM's Generations", NeurIPS 2024
  Hyunseok Lee*, Jihoon Tack*, Jinwoo Shin

Experience:

- Korean LLM leaderboard 1st place (Oct 2023)
  hyunseoki/ko-en-llama2-13b

Contact: hs.lee@kaist.ac.kr

Objective. Detect LLM-generated texts (LGTs)

ReMoDetect: Reward Models Recognize Aligned LLM's Generations

NeurIPS 2024 Poster

Hyunseok Lee*, Jihoon Tack*, Jinwoo Shin

Presenter: Hyunseok Lee

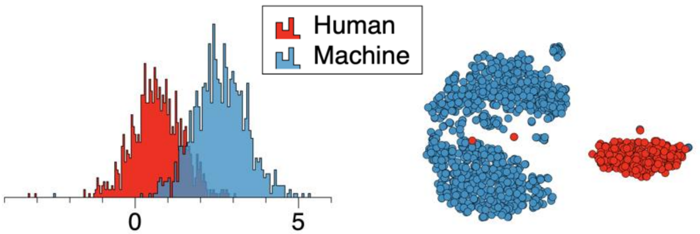
Overview

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

Motivation & Observations

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

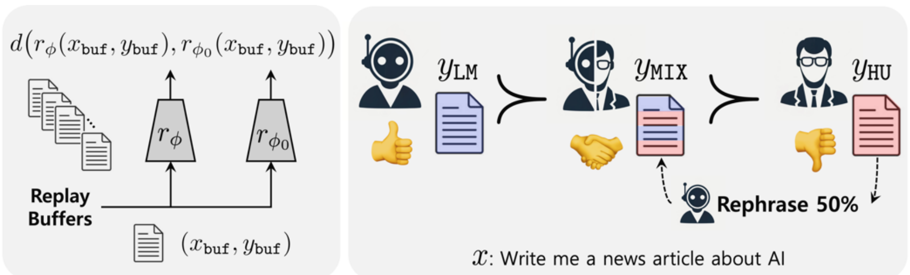
✓ Reward Model recognize LLM



Reward Score Distribution

t-SNE

Method: Continual Preference Tuning



Continual Preference Tuning

Reward Modeling with Mixed Responses

✓ Continual training with Human/LLM texts.

✓ Train with Mixed Human/LLM texts.

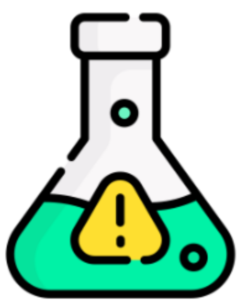
ReMoDetect is SOTA in Unseen LLMs and Unseen domains.

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**

Introduction

Nowadays, LLMs generate **fluent and convincing text**, which gives people many benefits.

- ✓ The quality of generated text is comparable to human specialists,
- ✓ They are **difficult to distinguish** from human-written content.
- ✓ This phenomenon will grow further and further.
- ✓ However, this also increases the potential for misuse.



Toxicity

Harmful or discriminatory language or content

Hallucination

Factually incorrect content

Legal Aspects

Data Protection, Intellectual Property, and the EU AI Act

LLM generated text Detection

Detecting LGT is a challenging problem in many aspects.

- **Accuracy**
  - LGT needs to be detected while minimizing false positives.
- **Generalizability**
  - Domain generalizability
  - LLM generalizability
- **Robustness**
  - Length Robustness
  - Paraphrasing Robustness

Prior Works

Model based Method [1],[2]

- Training supervised classification model for the detection of LLM-generated texts (LGT).

Metric Based Method (Zero-shot) [3],[4],[5]

- Scoring the text with entropy, perplexity, and log probability.

Watermark [6]

- Generating perturbations to a model's output and catching them in the outputs.

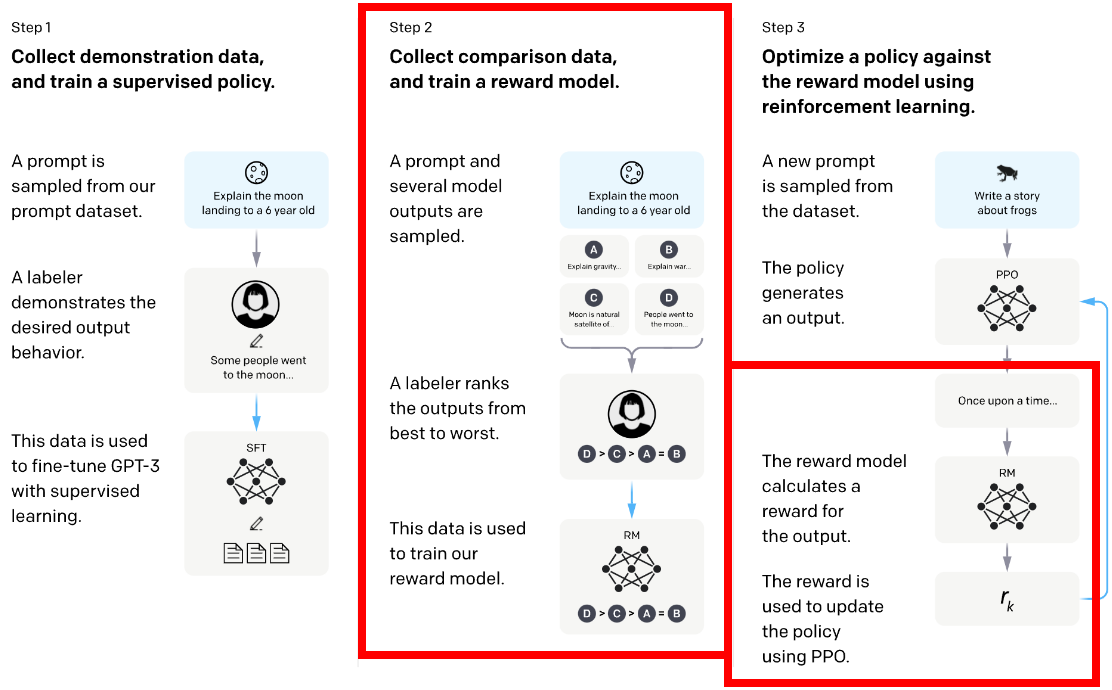
[1] Open AI, New AI classifier for indicating AI-written text, 2023 (end of service on July 2023 due to low accuracy.)
[2] Daphne Ippolito et al. Automatic Detection of Generated Text is Easiest when Humans are Fooled, ACL, 2020
[3] DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature, ICML 2023
[4] DetectLLM: Leveraging Log Rank Information for Zero-Shot Detection of Machine-Generated Text, arxiv2023
[5] Intrinsic Dimension Estimation for Robust Detection of AI-Generated Texts, Neurips 2023
[6] Kirchenbauer, J., et al. A watermark for large language models, arXiv 2023.

Motivation – Human Preference Align

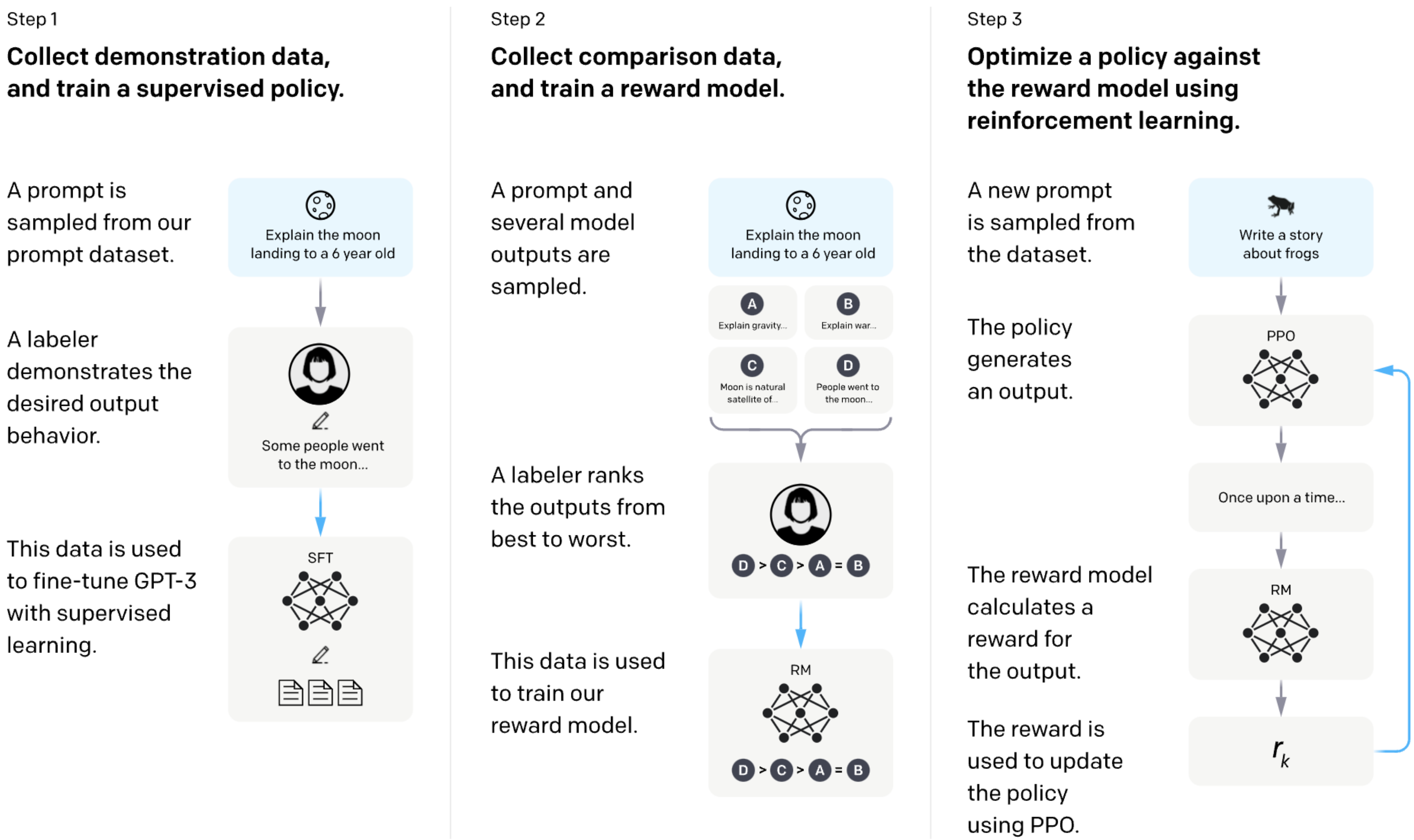
To solve the challenging problem: Accuracy, Generalizability, Robustness

Let's find **common characteristics of LLMs**!!

Common Characteristics of LLMs : Finetuned to fit human preferences using RLHF.
+ using the Reward Model as a proxy of human preference.



Common Characteristics of LLMs : Finetuned to fit human preferences using RLHF.
+ using the Reward Model as a proxy of human preference.



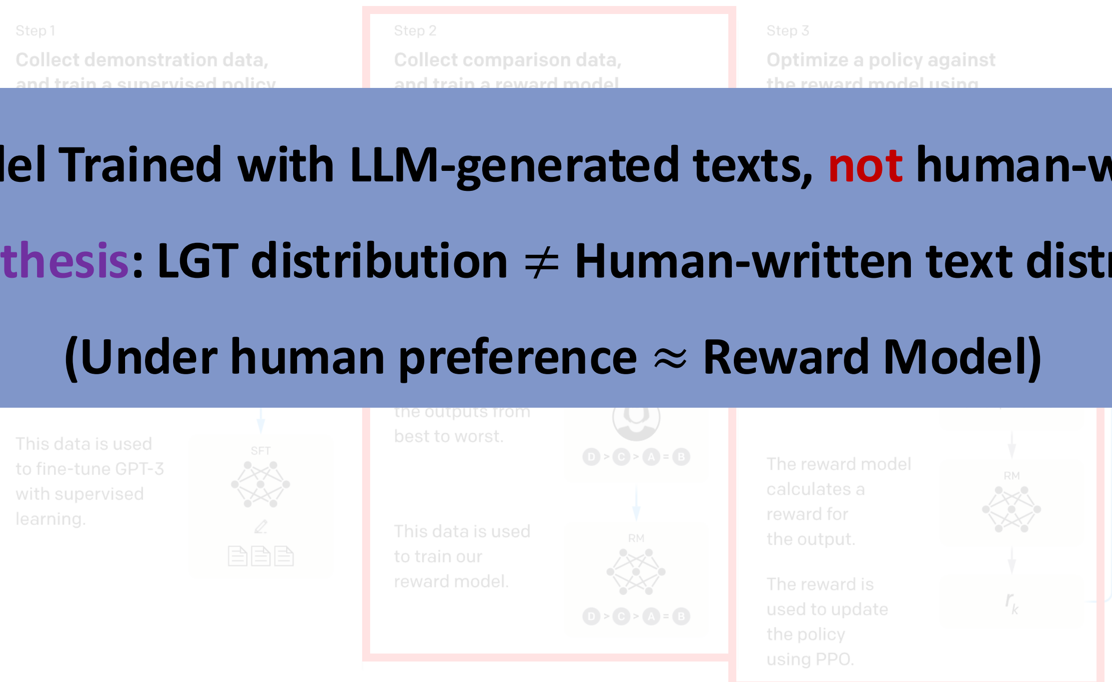
[1] OpenAI, InstructGPT: training language models to follow instructions with human feedback, 2022

Motivation – Human Preference Align

To solve the challenging problem: Accuracy, Generalizability, Robustness

Let's find common characteristics of LLMs!!

Common Characteristics of LLMs : Finetuned to fit human preferences using RLHF.
+ using the Reward Model as a proxy of human preference.



Reward Model Trained with LLM-generated texts, not human-written texts.

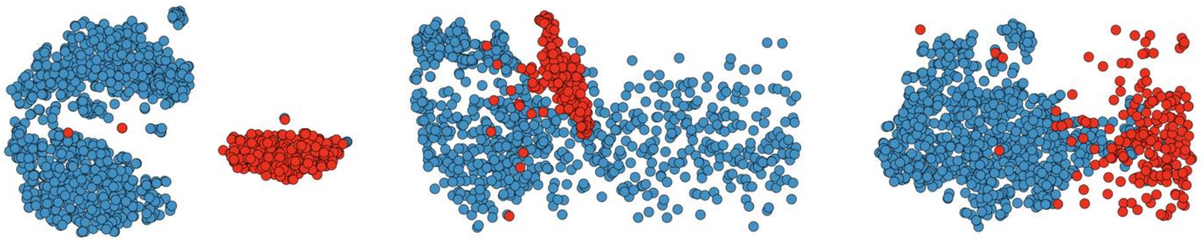
↔ Hypothesis: LGT distribution ≠ Human-written text distribution

(Under human preference ≈ Reward Model)

Observation: LGT distribution ≠ Human-written text distribution

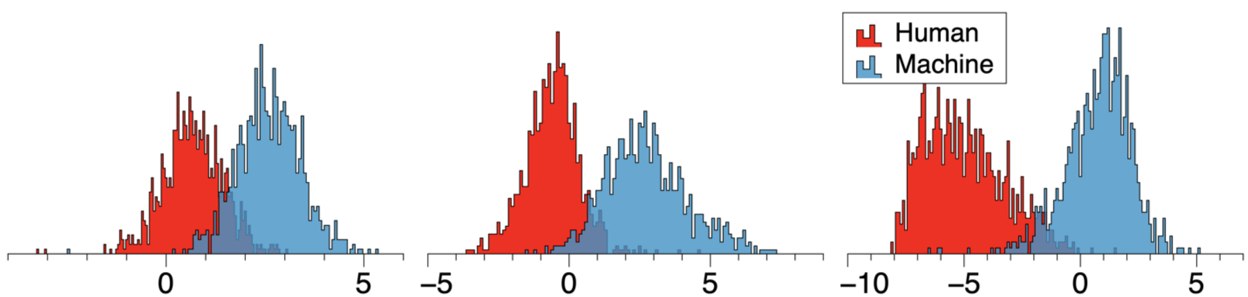
And LGTs have higher rewards than human-written texts.

t-SNE of Reward Model



(a) Reuters (b) Essay (c) WritingPrompts

Reward Score Distribution of Reward Model



(a) Reuters (b) Essay (c) WritingPrompts

Observation: LGTs have higher rewards than human-written texts.

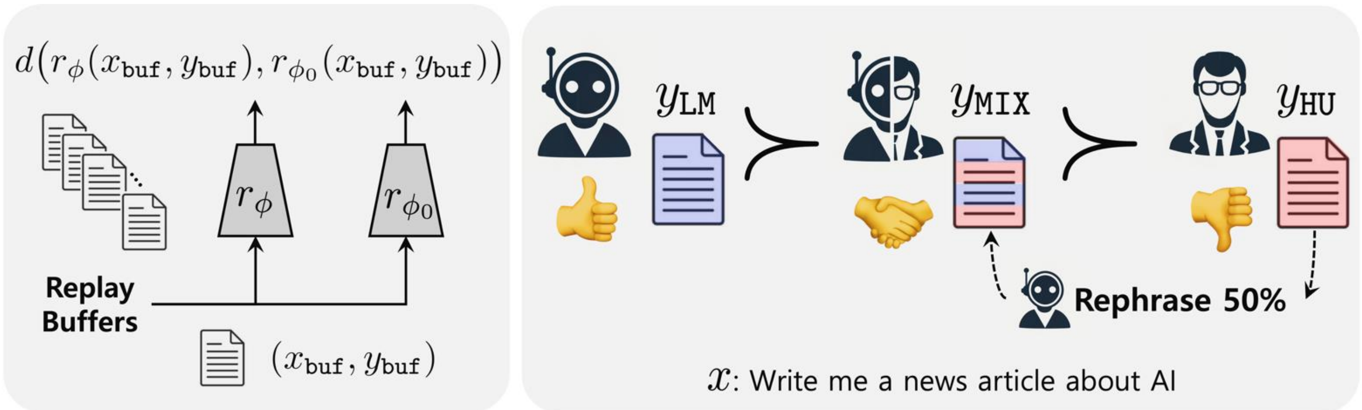
And LGT distribution $\neq$ Human-written text distribution

t-SNE of Reward Model

Reward Model Recognize LLM!!

Reward Score Distribution of Reward Model

Method



Continual Preference Tuning

Reward Modeling with Mixed Responses

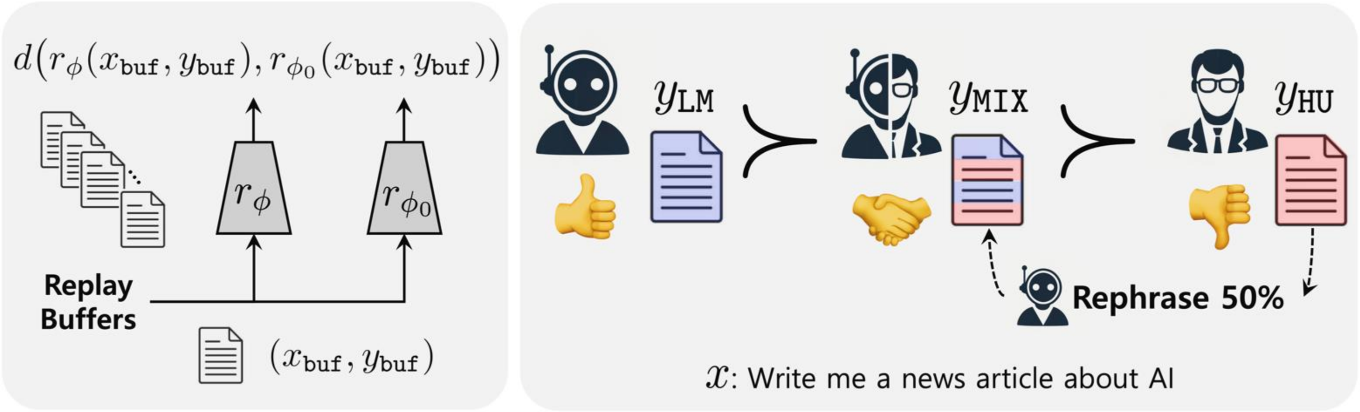
Two training components to improve the detection ability of the reward model.

1. **Continual Preference Tuning**

- Finetune the **reward model** with **LLM/Human text** pairs.
- Mitigate forgetting using **replay buffer**.

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

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



Continual Preference Tuning

Reward Modeling with Mixed Responses

Two training components to improve the detection ability of the reward model.

2. Reward Modeling with Mixed Responses

- **Partially rephrase** the human-written text using LLM.
- Mixed texts are used as a median preference. **P(LLM ≻ Mixed ≻ Human)**
- Enabling the detector to learn a **better decision boundary**.

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

Experiments

Baselines:

- Statistic Metrics (Loglikelihood, Rank)
- Detect GPT-style (Detect GPT, Fast-DetectGPT, LLR, NPR)
- Binary Classifier (OpenAI-Detector ChatGPT-Detector)

Trained Model:

- OpenAssistant/reward-model-deberta-v3-large-v2 (700M parameters)

Trained Dataset:

- HC3 (**Human / ChatGPT3.5 pairs**): 4400 samples

Evaluation Dataset (AUROC)

- MGTBench, Fast-DetectGPT Bench
- **Unseen Domains: Pubmed, Xsum, WritingPrompt, Essay, Reuters**
- **Unseen LLMs: Llama, Gemini, GPT4, Claude, Phi …**

Robustness Evaluation:

- Shorter passage lengths, paraphrasing attack

Results - Fast-DetectGPT Benchmark

(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 significantly outperforms prior detection methods.

Detection performance is consistent among various LLMs and domains.

Results - MGTBench

(b) MGTBench [14]: Essay, Reuters, and WritingPrompts (WP)

Model	Domain	Loglik.	Rank	D-GPT	LRR	NPR	FD-GPT	Open-D	Chat-D	**Ours**
GPT3.5 Turbo	Essay	97.3	95.7	57.8	97.8	48.1	99.6	57.5	81.5	**100.0**
	Reuters	98.2	94.8	50.5	98.7	51.1	**99.9**	98.5	97.2	**99.9**
	WP	89.8	90.2	52.9	77.2	48.3	91.7	50.8	66.3	**100.0**
GPT4 Turbo	Essay	96.5	93.9	58.9	93.9	62.4	98.9	55.8	77.1	**99.9**
	Reuters	95.8	93.1	52.6	94.9	53.3	99.4	87.5	92.4	**99.9**
	WP	94.2	91.0	53.5	85.2	55.3	93.0	68.2	67.9	**99.9**
Llama3 70B	Essay	98.3	95.3	56.2	98.9	57.8	99.5	83.9	91.7	**100.0**
	Reuters	99.9	89.7	58.9	98.7	59.2	**100.0**	96.7	90.8	**100.0**
	WP	97.3	90.8	57.2	91.1	60.4	99.1	86.6	77.3	**99.8**
Gemini pro	Essay	98.3	93.6	64.4	97.7	65.5	98.3	48.9	65.9	**100.0**
	Reuters	99.9	83.1	73.0	99.3	74.9	**100.0**	95.3	91.5	**100.0**
	WP	91.7	82.0	63.9	76.7	67.3	99.2	68.8	73.4	**99.8**
Claude	Essay	91.6	85.9	44.2	82.7	48.7	83.6	32.4	19.6	**99.7**
	Reuters	91.3	79.5	68.1	79.2	68.7	87.8	65.5	25.6	**99.8**
	WP	88.4	80.0	60.0	71.2	60.7	74.1	46.2	26.7	**99.1**
Average	-	95.2	89.2	58.1	89.5	58.8	94.9	69.5	69.7	**99.9**

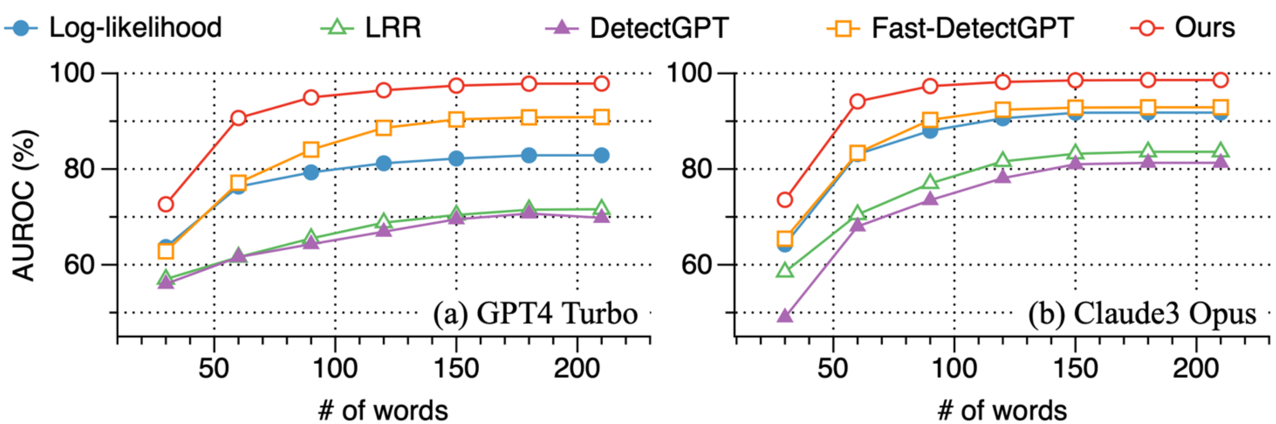
ReMoDetect significantly outperforms prior detection methods.

Detection performance is consistent among various LLMs and domains

Robustness on 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%)**

Robustness on Passage Length



ReMoDetect is relatively more **robust to rephrasing attacks** and **various length** of passage lengths than other detection methods.

Results

Comparison with Commercial Models

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**

Evaluation Results in DPO-trained LLMs.

Model	Domain	Loglik.	Rank	D-GPT	LRR	NPR	FD-GPT	Open-D	Chat-D	**Ours**
Phi-3 mini	PubMed	65.0	56.2	46.8	48.8	45.4	63.7	37.7	80.7	**94.5**
	XSum	70.3	64.1	69.0	61.7	70.5	91.0	82.7	23.4	**97.6**
	WP-s	82.4	73.3	89.6	68.8	87.8	96.7	60.0	31.1	**99.3**
Phi-3 small	PubMed	57.2	50.4	48.1	51.2	48.1	59.9	31.9	82.7	**91.7**
	XSum	81.1	69.7	70.0	68.8	72.7	95.6	79.3	19.5	**98.7**
	WP-s	84.0	72.3	86.7	67.1	83.2	97.2	58.6	32.2	**97.4**
Phi-3 medium	PubMed	65.4	55.4	51.2	50.3	37.6	61.7	34.2	15.8	**95.2**
	XSum	64.5	61.2	80.7	79.0	81.3	85.4	75.0	18.1	**98.0**
	WP-s	83.1	73.6	90.3	70.6	90.2	95.7	53.9	38.5	**98.8**

ReMoDetect outperform commercial model GPTZero

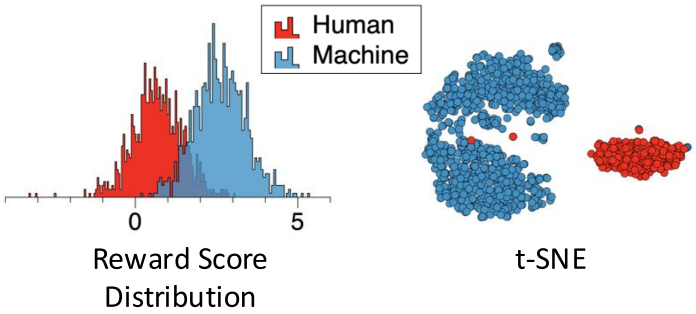
Detection performance is consistently outperforming even DPO-trained models and smaller models

Summary

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

Motivation & Observations

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



Method: Continual Preference Tuning

- ✓ Continual training with Human/LLM texts.
- ✓ Train with Mixed Human/LLM texts.

Take Away Messages

- ✓ **LGT distribution ≠ Human-written text distribution**
- ✓ ReMoDetect is **SOTA** for detecting most unseen domains and LLMs.

Discussion Point : Why LGT distribution ≠ Human-written text distribution?

- ✓ Hypothesis 1: LLMs are trained with LLM-generated text and model-annotated data.
- ✓ Hypothesis 2: Human writing styles vary individually, while LLMs are optimized to average.

Additional Discussions

Homepage | **Paper** | **Demo**