IPAD: Inverse Prompt for AI Detection - A Robust and Interpretable LLM-Generated Text Detector

Zheng Chen[1]* **Yushi Feng**[2]* **Jisheng Dang**[3] **Yue Deng**[1] **Changyang He**[4]
Hongxi Pu[5] **Haoxuan Li**[6]† **Bo Li**[1]†
[1]Computer Science and Engineering, Hong Kong University of Science and Technology
[2]School of Computing and Data Science, The University of Hong Kong
[3]School of Information Science & Engineering, Lanzhou University
[4]Max Planck Institute for Security and Privacy
[5]Computer Science, The University of Michigan
[6]Center for Data Science, Peking University
zchenin@connect.ust.hk, fengys@connect.hku.hk, dangjisheng@lzu.edu.cn,
ydengbi@connect.ust.hk, changyang.he@mpi-sp.org,
hongxi@umich.edu, hxli@stu.pku.edu.cn, bli@cse.ust.hk

Abstract

Large Language Models (LLMs) have attained human-level fluency in text generation, which complicates the distinguishing between human-written and LLM-generated texts. This increases the risk of misuse and highlights the need for reliable detectors. Yet, existing detectors exhibit poor robustness on out-of-distribution (OOD) data and attacked data, which is critical for real-world scenarios. Also, they struggle to provide interpretable evidence to support their decisions, thus undermining the reliability. In light of these challenges, we propose **IPAD (Inverse Prompt for AI Detection)**, a novel framework consisting of a **Prompt Inverter** that identifies predicted prompts that could have generated the input text, and two **Distinguishers** that examine the probability that the input texts align with the predicted prompts. Empirical evaluations demonstrate that IPAD outperforms the strongest baselines by 9.05% (Average Recall) on in-distribution data, 12.93% (AUROC) on out-of-distribution data, and 5.48% (AUROC) on attacked data. IPAD also performs robustly on structured datasets. Furthermore, an interpretability assessment is conducted to illustrate that IPAD enhances the AI detection trustworthiness by allowing users to directly examine the decision-making evidence, which provides interpretable support for its state-of-the-art detection results.

1 Introduction

Large Language Models (LLMs), characterized by their massive scale and extensive training data [Chen et al., 2024, Feng et al., 2025a, Cheng et al., 2025], have achieved significant advances in natural language processing (NLP) [Ouyang et al., 2022, Veselovsky et al., 2023, Wu et al., 2025]. However, with the advanced capabilities of LLMs, they are subject to frequent misused in various domains, including academic fraud, the creation of deceptive material, and the generation of fabricated information [Ji et al., 2023, Pagnoni et al., 2022, Mirsky et al., 2023, Chen et al., 2025], which underscores the critical need to distinguish between human-written text (HWT) and LLM-generated text (LGT) [Pagnoni et al., 2022, Yu et al., 2025, Kirchenbauer et al., 2023].

*Equal contribution.
†Corresponding authors.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

 https://doi.org/10.52202/085713-5580

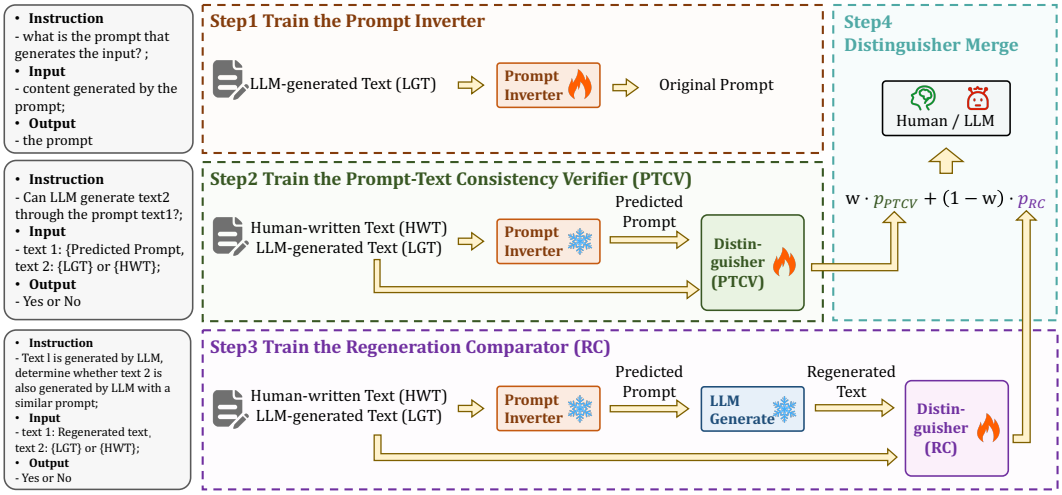

Figure 1: The overall workflow of our proposed IPAD framework

However, due to their sophisticated functionality, LLMs pose significant challenges in the robustness of current AI detection systems [Wu et al., 2025]. The existing detection systems, including commercial ones, frequently misclassify texts as HWT [Price and Sakellarios, 2023, Walters, 2023] and generate inconsistent results when analyzing the same text using different detectors [Chaka, 2023, Weber-Wulff et al., 2023]. Studies show false positive rates reaching up to 50% and false negative rates as high as 100% in different tools [Weber-Wulff et al., 2023] when dealing with out-of-distribution (OOD) datasets.

Another critical issue with the existing AI detection systems is their lack of verifiable evidence [Halaweh and Refae, 2024, Feng et al., 2025b], as these tools typically provide only simple outputs like *"likely written by AI"* or percentage-based predictions [Weber-Wulff et al., 2023]. The lack of evidence prevents users from defending themselves against false accusations [Chaka, 2023] and hinders organizations from making judgments based solely on the detection results without convincing evidences [Weber-Wulff et al., 2023]. This problem is particularly troublesome not only because the low accuracy of such systems as mentioned before, but also due to the consequent inadequate response to LLM misuse, which can lead to significant societal harm [Stokel-Walker and Van Noorden, 2023, Porsdam Mann et al., 2023, Shevlane et al., 2023, Wu et al., 2025]. These limitations highlight the pressing need for more reliable, explainable and robust detectors.

In this paper, we propose **IPAD** (Inverse Prompt for AI Detection), a novel and interpretable framework for detecting AI-generated text. As illustrated in Figure 1, IPAD consists of two main components: a **Prompt Inverter**, which reconstructs the underlying prompts from input texts, and two **Distinguishers**—the **Prompt-Text Consistency Verifier (PTCV)**, which measures the alignment between the predicted prompt and input text, and the **Regeneration Comparator (RC)**, which compares the input with the corresponding regenerated text for consistency. By explicitly modeling the reasoning path from prompt inversion to final classification, IPAD introduces a paradigm shift in AI-generated content detection, significantly enhancing both detection robustness and user interpretability.

Empirical results show that IPAD outperforms state-of-the-art baselines by 9.05% in Average Recall on in-distribution datasets, 12.93% in AUROC on out-of-distribution (OOD) datasets, and 5.48% in AUROC under adversarial attacks. IPAD also generalizes well to structured data. A user study further reveals that IPAD improves trust and usability in detection tasks by presenting concrete decision evidence, including predicted prompts and regenerated texts. Code is available at `https://github.com/Bellafc/IPAD-Inver-Prompt-for-AI-Detection`.

Our contributions can be summarized as follows:

- We introduce a novel fine-tuned inverse-prompt-based detection framework that integrates prompt reconstruction and dual consistency evaluation.

- We achieve superior detection performance on in-distribution, OOD, adversarially attacked, and prompt-structured datasets.
- We demonstrate through an interpretability assessment that IPAD improves human trust and interpretability in AI text detection.

2 Methodology

2.1 Preliminaries

Modules. Our method comprises a **Prompt Inverter** $f_{\textbf{inv}}$, and two Distinguishers, namely the **Prompt-Text Consistency Verifier (PTCV)** $f_{\textbf{PTCV}}$ and the **Regeneration Comparator (RC)** $f_{\textbf{RC}}$. Given an input text T, the task is to determine whether it is human-written (HWT) or generated by an LLM (LGT). We denote by $\mathcal{D}_{\text{PI}}$ the training set for f_{inv}, consisting of pairs (T, P) where T is an LLM-generated text and P is its original prompt. The two distinguishers are trained using disjoint datasets: $\mathcal{D}_{\text{LGT}}$ contains LLM-generated samples and $\mathcal{D}_{\text{HWT}}$ contains human-written ones. All components are fine-tuned using Microsoft's Phi3-medium-128k-instruct modelAbdin et al. [2024].

Softmax-Based probability for Binary Classification in LLM. To estimate the fine-tuned model's binary classification probability (i.e., the probability of predicting "yes" or "no"), we follow the logit-based estimation approach [Yoshikawa and Okazaki, 2023]. Given the model input x, and the output logits z, the model's probability assigned to $\hat{y}$ is computed through the softmax function σ :

$$\text{Confidence}_{\text{yes}} = P(\hat{y} = \text{“yes”} \mid x) = \sigma(z)_{\text{yes}}, \quad \text{Confidence}_{\text{no}} = P(\hat{y} = \text{“no”} \mid x) = \sigma(z)_{\text{no}}$$

Since the fine-tuned model will only output"yes" or "no", we further calculate the probability for this binary classification as:

$$\text{Probability}_{\text{yes}} = \frac{\text{Confidence}_{\text{yes}}}{\text{Confidence}_{\text{yes}} + \text{Confidence}_{\text{no}}}, \quad \text{Probability}_{\text{no}} = 1 - \text{Probability}_{\text{yes}}$$

2.2 Workflow

Our framework follows a multi-stage fine-tuning pipeline with the following four steps, as illustrated in Figure 1. The details of the datasets for fine-tuning is illustrated in Appendix A.

Step 1: Training Prompt Inverter. We first fine-tune a model f_{inv} on dataset $\mathcal{D}_{\text{PI}}$, with the data structure shown in Figure 1. For any input text T, f_{inv} predicts the most likely prompt P that could have generated it, i.e. $P = f_{\text{inv}}(T)$. The resulting Prompt Inverter is then frozen and reused in the following downstream steps.

Step 2: Training the Prompt-Text Consistency Verifier (PTCV). Given the predicted prompt P in step 1, and the input text $T \in \{\text{HWT}, \text{LGT}\}$, the verifier f_{PTCV} is trained to predict whether the text T could plausibly be generated by an LLM using the prompt P. The fine-tuning datasets $\mathcal{D}_{\text{LGT}}$ and $\mathcal{D}_{\text{HWT}}$ share the same structure, with output labels "yes" for $\mathcal{D}_{\text{LGT}}$ and "no" for $\mathcal{D}_{\text{HWT}}$, as shown in the Figure 1.

After fine-tuning this module, we applied it to the validation set and computed the probability score $p_{\text{PTCV}} = f_{\text{PTCV}}(T, P)$, where the confidence value was estimated using the softmax-based method described in Section 2.1.

Step 3: Training the Regeneration Comparator (RC). With the same predicted prompt P in step 1, we use an LLM to generate a regenerated text $T' \leftarrow LLM(P)$. By default, the LLM we use is `gpt-3.5-turbo`. Then, the comparator f_{RC} is trained to assess whether T and T' can be generated by LLM with a similar prompt. This step uses the same dataset as in Step 2, but applies a different structural formatting, as shown in Figure 1.

After fine-tuning this module, we also applied it to the validation set and computed the probability score $p_{\text{RC}} = f_{\text{RC}}(T, P)$.

Step 4: Distinguisher Merge. To determine the final classification, we combine the two probability scores, p_{PTCV} and p_{RC}, obtained from Step 2 and Step 3 on the validation set. Specifically, we

 https://doi.org/10.52202/085713-5580

compute a weighted ensemble as $\hat{p} = w \cdot p_{\text{PTCV}} + (1 - w) \cdot p_{\text{RC}}$, and assign the prediction $\hat{Y} = \text{LGT}$ if $\hat{p} > \tau$, or $\hat{Y} = \text{HWT}$ otherwise. The weight $w \in [0, 1]$ and the threshold $\tau \in [0, 1]$ are treated as hyperparameters and selected via grid search on the validation set. The selected values were $w = 0.45$ and $\tau = 0.54$.

Inference. We perform inference on unseen input texts T by sequentially applying the trained modules. Given an input text T, we first use the prompt inverter f_{inv} to recover the most plausible prompt P. The prompt is then used to regenerate a candidate text T' via the an LLM. Next, we compute two probability scores: p_{PTCV}, indicating whether T is consistent with P, and p_{RC}, assessing the likelihood that T and T' originate from the same prompt. These scores are fused into a final decision score $\hat{p}$ using the gird-searched weight w, and the predicted label is determined by comparing $\hat{p}$ against the threshold τ. The complete inference pipeline is summarized in Algorithm 1.

Algorithm 1 IPAD Detection Procedure

Require: Input text T; trained modules $\mathbf{f}_{\text{inv}}, \mathbf{f}_{\text{PTCV}}, \mathbf{f}_{\text{RC}}$; LLM $\mathbf{f}_{\text{LLM}}$; fusion weight $w \in [0, 1]$; threshold $\tau \in [0, 1]$

$P \leftarrow \mathbf{f}_{\text{inv}}(T)$ ▷ Inverse-prompt prediction
$T' \leftarrow \mathbf{f}_{\text{LLM}}(P)$ ▷ Regenerate text using P
$\boldsymbol{z}^{\text{PTCV}} \leftarrow \mathbf{f}_{\text{PTCV}}(P, T)$
$p_{\text{PTCV}} \leftarrow \frac{\sigma(\boldsymbol{z}^{\text{PTCV}}_{\text{yes}})}{\sigma(\boldsymbol{z}^{\text{PTCV}}_{\text{yes}}) + \sigma(\boldsymbol{z}^{\text{PTCV}}_{\text{no}})}$
$\boldsymbol{z}^{\text{RC}} \leftarrow \mathbf{f}_{\text{RC}}(T', T)$
$p_{\text{RC}} \leftarrow \frac{\sigma(\boldsymbol{z}^{\text{RC}}_{\text{yes}})}{\sigma(\boldsymbol{z}^{\text{RC}}_{\text{yes}}) + \sigma(\boldsymbol{z}^{\text{RC}}_{\text{no}})}$
$\hat{p} \leftarrow w \cdot p_{\text{PTCV}} + (1 - w) \cdot p_{\text{RC}}$
if $\hat{p} > \tau$ **then**
    $\hat{Y} \leftarrow \text{LGT}$
else
    $\hat{Y} \leftarrow \text{HWT}$
end if
$\mathcal{E} \leftarrow (P,\ p_{\text{PTCV}},\ p_{\text{RC}},\ \hat{p})$
return $(\hat{Y},\ \mathcal{E})$

2.3 Computational Complexity and Deployment Considerations

The inference procedure of the IPAD framework consists of three calls through a light-weight open-sourced LLM `phi-3-medium-128k-instruct`. `Phi-3` is a decoder-only Transformer, whithin which, the self-attention complexity per layer is $\mathcal{O}(n^2 \cdot d)$, where n is the sequence length and d is the hidden dimension [Vaswani et al., 2017]. The additional api call to `gpt-3.5-turbo` for regenerating texts introduces fixed latency but no local computation cost. Therefore, the overall computational cost is bounded by $\mathcal{O}(3 \cdot L \cdot n^2 \cdot d + \texttt{OpenAI}_{\text{api}})$, where $L = 32$ is the number of layers in `phi-3` [Abdin et al., 2024], which is relatively small. All three `phi-3` calls can be deployed in an Nvidia V100 GPU as the minimum requirement. This demonstrates that IPAD is not computationally expensive and can be deployed with relatively modest hardware requirements.

2.4 Training

The supervised fine-tuning [Wei et al., 2022] process is performed on a Microsoft's open model, `phi3-medium-128k-instruct`, and we use low-rank adaptation (LoRA) method [Hu et al., 2022] on the `LLaMA-Factory` framework [Zheng et al., 2024a]. We train it using six A800 GPUs for 20 hours for **Prompt Inverter**, 7 hours for **PTCV**, and 9 hours for **RC**.

3 Experiments

We investigate the following questions through our experiments:

- Assess the robustness of IPAD, which includes using various LLMs as generators, comparing IPAD with other detectors, and evaluating on out-of-distribution (OOD), attacked datasets, and prompt-structured datasets.
- Independently analyze the necessity and effectiveness of the **Prompt Inverter**, the **PTCV**, and the **RC**.
- Explore the user-friendliness of IPAD through an interpretability assessment.

Table 1: Detection Accuracy (HumanRec, MachineRec, AvgRec, and AUROC %) of IPAD across Various LLMs on In-Distribution Data

Original Generator	Re-Generator	HumanRec	MachineRec	AvgRec	AUROC
gpt-3.5-turbo	gpt-3.5-turbo	98.50	100	99.25	100
gpt-4	gpt-4	98.70	100	99.35	100
	gpt-3.5-turbo	96.10	100	98.05	99.96
Qwen-turbo	Qwen-turbo	98.60	99.80	99.20	99.96
	gpt-3.5-turbo	98.40	99.50	98.95	99.86
LLaMA-3-70B	LLaMA-3-70B	98.70	100	99.35	100
	gpt-3.5-turbo	98.60	100	99.30	100

3.1 Robustness of IPAD

3.1.1 Evaluation Baselines and Metrics

The in-distribution experiments refer to the testing results presented in [Koike et al., 2024], where the data aligns with the training data used for the IPAD, thereby serving as our baseline. This baseline assesses how the RoBERTa classifiers (base and large) [Park et al., 2021], the HC3 detector [Guo et al., 2023], and the OUTFOX detector [Koike et al., 2024] perform on standard data as well as under DIPPER [Alkanhel et al., 2023] and OUTFOX attacks.

The OOD experiments refer to the DetectRL baseline [Wu et al., 2024], which is a comprehensive benchmark, which includes four datasets: (1) academic abstracts from the arXiv Archive (covering the years 2002 to 2017) , (2) news articles from the XSum dataset [Narayan et al., 2018], (3) creative stories from Writing Prompts [Fan et al., 2018], and (4) social reviews from Yelp Reviews [Zhang et al., 2015]. It also employs three attack methods to simulate complex real-world detection scenarios, which include (1) the prompt attacks, (2) paraphrase attacks, and (3) perturbation attacks [Wu et al., 2024]. DetectRL evaluates three classifiers on the OOD dataset: DetectLLM (LRR) [Su et al., 2023], Fast-DetectGPT [Bao et al., 2024], RoBERTa Classifier (Base). We included two more strong classifiers in our evaluation DetectLLM (NPR) [Su et al., 2023] and Binoculars [Hans et al., 2024]. All the testing sets have 1,000 samples in our experiments.

We further evaluate its performance on OOD datasets with **structured prompts**. The LongWriter dataset [Bai et al., 2025], featuring an average prompt length of 1,501 tokens, reflects IPAD's capability to handle long-form prompts. The Code-Feedback[Zheng et al., 2024b] and Math datasets [Hendrycks et al., 2021] contain highly structured prompts, in contrast to typical essay-like writing. We compare IPAD with baseline detectors from DetectRL to assess its relative performance under these challenging conditions.

The **Area Under Receiver Operating Characteristic curve (AUROC)** is widely used for assessing detection method [Mitchell et al., 2023a]. Since our models predict binary labels, we follow the *Wilcoxon-Mann-Whitney* statistic [Calders and Jaroszewicz, 2007], and the formula is shown in Appendix B. The **AvgRec** is the average of **HumanRec** and **MachineRec**, which refers to the recall of the Human-written texts and the LLM-generated texts [Li et al., 2024].

3.1.2 Robustness across different LLMs

As shown in Table 1, IPAD achieves consistently strong performance across all combinations of original generators and re-generators, which shows its robustness to diverse LLM as generators. The best results are generally observed when the original generator and the re-generator are the same, while the `gpt-3.5-turbo` serves as an effective universal re-generator: it performs well even when the original generator differs. In real-world applications where the identity of the original generator is unknown, using `gpt-3.5-turbo` as a fixed re-generator provides a practical and reliable solution.

3.1.3 Comparison of IPAD with other detectors in and out of distribution

In Distribution. For the in-distribution data, as shown in Figure 2, the baseline detectors like RoBERTa, HC3, and OUTFOX perform well on standard data but degrade significantly under DIPPER and OUTFOX attacks. In contrast, IPAD maintains high accuracy across all scenarios, which surpasses the strongest baseline **9.05%** in AvgRec.

 https://doi.org/10.52202/085713-5580

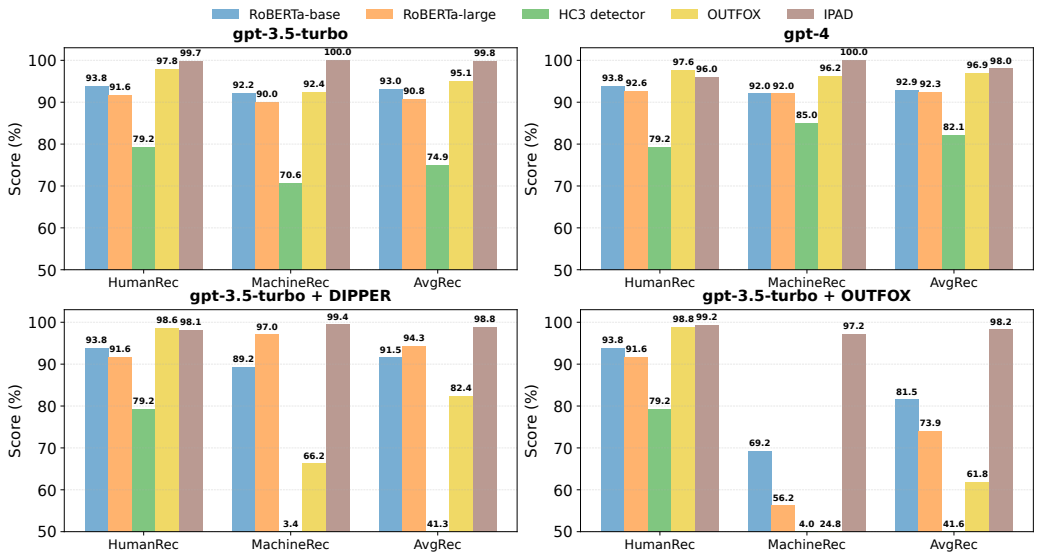

Figure 2: The In-distribution data performance of IPAD and the baseline detectors. Since Koike et al. [2024] only presents the AvgRec data for the baselines, we also calculate AvgRec data for IPAD to compare.

Out of Distribution. Table 2 reports detection accuracy across four benchmark datasets, which shows that IPAD significantly outperforms prior baselines. Table 3 further evaluates robustness under three attack types, where IPAD again demonstrates superior resilience. Compared to the strongest baseline, IPAD achieves a **12.93%** relative improvement on standard datasets in AUROC and a **5.48%** improvement on attack datasets.

Table 2: Detection Accuracy (AUROC %) on four diverse OOD datasets

Method	Arxiv	XSum	Writing	Review	Average
DetectLLM (LRR)	48.17	48.41	58.70	58.21	53.37
DetectLLM (NPR)	53.85	34.59	54.96	50.09	48.37
Binoculars	84.03	77.39	94.38	90.00	86.45
Fast-DetectGPT	42.00	45.72	51.13	54.55	48.35
Rob-Base	81.06	76.81	86.29	87.84	83.00
IPAD Merge	**100**	**99.85**	**99.40**	**98.25**	**99.38**

Table 3: Detection Accuracy (AUROC %) on three attacked OOD datasets

Method	Prompt Attack	Paraphrase Attack	Perturbation Attack	Average
DetectLLM (LRR)	54.97	49.23	53.62	52.61
DetectLLM (NPR)	77.15	56.94	6.78	46.96
Binoculars	93.45	88.34	76.89	86.23
Fast-DetectGPT	43.89	41.15	44.38	43.14
Rob-Base	92.81	90.02	92.12	91.65
IPAD	**97.30**	**96.00**	**98.10**	**97.13**

Structured Prompts. The results are shown in Table 4, while these datasets lack HWT references and are thus only evaluated using MachineRec, the strong scores suggest that IPAD maintains robustness even on structured diverse inputs, with an improvement of 9.87% against the strongest baseline in MachineRec.

Table 4: Detection Accuracy (MachineRec %) on three structured OOD datasets

Method	LongWriter	Code-Feedback	Math	Average
DetectLLM (LRR)	32.1	29.0	30.2	30.43
DetectLLM (NPR)	41.2	45.9	56.0	47.7
Binoculars	82.1	84.6	89.4	85.4
Fast-DetectGPT	12.0	11.1	15.1	12.7
Rob-Base	81.5	89.2	82.1	84.3
IPAD	**97.5**	**92.7**	**95.6**	**95.27**

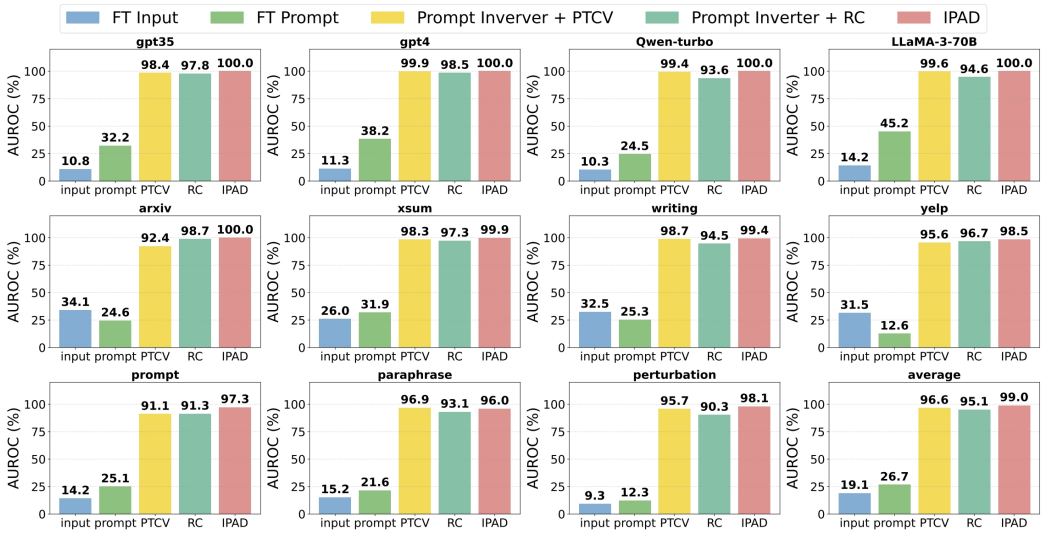

Figure 3: Ablation study. Evaluating **Fine-tune only on Input**, **Fine-tune only on Prompt**, **Prompt Inverter + PTCV**, **Prompt Inverter + RC**, and **IPAD** on In-distribution datasets, standard OOD datasets, and attacked OOD datasets.

3.2 Necessity and Effectiveness of the Prompt Inverter, PTCV, and RC

3.2.1 Necessity

To prove that it is necessary to fine-tune on IPAD with IPAD with **PTCV** and **RC**, we conducted ablation study to use the same finetune method on only **input texts** and only **predicted prompts**, with the finetune data format shown in Appendix C. We only experimented on **Prompt Inverter + PTCV** and **Prompt Inverter + RC** to compare with the three-moduled IPAD.

Based on the ablation study results as shown in Figure 3, fine-tuning only on input texts or only on predicted prompts performs poorly across all datasets in AUROC scores. While using **Prompt Inverter + PTCV** or **Prompt Inverter + RC** individually significantly improves performance, neither approach consistently excels across both HWT-style and LGT-style generations. In contrast, the full IPAD framework achieves consistently high performance across all settings, which demonstrates the necessity of the **Prompt Inverter**, **PTCV**, and **RC** modules.

3.2.2 Effectiveness

Prompt Inverter. We use DPIC [Yu et al., 2024] and PE [Zhang et al., 2024a] as baseline methods for prompt extraction. DPIC employs a zero-shot approach using the prompt states in Appendix D, while PE uses adversarial attacks to recover system prompts. In our evaluation, we tested 1000 LGT and 1000 HWT samples. We use only in-distribution data for testing since only these datasets include original prompts. The metrics are all tested on comparing the similarity of the original prompts and the predicted prompts. The results shown in Table 5 illustrate that IPAD consistently outperforms both DPIC and PE across all four metrics (BartScore [Yuan et al., 2021], Sentence-Bert Cosine

https://doi.org/10.52202/085713-5580

Table 5: Comparison of prompt inverters on the similarities of the original prompts and the predicted prompts on LGT and HWT.

Metric	LGT			HWT		
	DPIC	PE	IPAD	DPIC	PE	IPAD
Bart-large-cnn	-2.12	-2.23	**-1.84**	-2.47	-2.39	**-2.22**
Sentence-Bert	0.46	0.58	**0.69**	0.42	0.53	**0.57**
BLEU	5.61E-05	3.21E-04	**0.24**	8.75E-06	2.56E-08	**0.13**
ROUGE-1	0.04	0.25	**0.51**	0.06	0.13	**0.39**

Table 6: Comparison of distinguishers on HumanRec, MachineRec, and AvgRec (%).

Distinguish Method	HumanRec	MachineRec	AvgRec
Sentence-Bert (Thr. 0.67)	61.20	95.20	78.20
Bart-large-cnn (Thr. -2.52)	42.60	97.20	69.90
Prompt to ChatGPT	33.20	64.50	48.85
IPAD	**98.50**	**100.00**	**99.25**

Similarity [Reimers and Gurevych, 2019], BLEU [Papineni et al., 2002], and ROUGE-1 [Lin, 2004]), which highlight the effectiveness of the IPAD **Prompt Inverter**.

PTCV and RC. We conducted a comparison study using the frozen Prompt Inverter but different distinguishing methods. The first and second methods employed `Sentence-Bert` [Reimers and Gurevych, 2019] and `Bart-large-cnn` [Yuan et al., 2021] to compute the similarity score between the input texts and the regenerated texts. We selected thresholds that maximized AvgRec, which were 0.67 for Sentence-Bert and -2.52 for Bart-large-cnn. The classification rule is that the texts with scores greater than the threshold will be classified as LGT, while the texts with scores less than or equal to the threshold will be classified as HWT. The third method is to directly prompt ChatGPT in Appendix C, which mimic the fine-tuning process of **PTCV** and **RC**. The final results shown in Table 6 demonstrate that the other distinguishing methods performed worse than IPAD, highlighting the superior effectiveness of the IPAD **Distinguishers**.

Compare with DPIC. DPIC first uses a zero-shot prompt inverter to generate prompts, then applies a Siamese encoder and classifier to measure similarity between the embeddings of the original and regenerated texts. However, the classifier's reliance on embedding similarity is ambiguous, as similar texts may stem from different prompts. IPAD addresses this by fine-tuning directly on raw texts and reformulating the task as a logical reasoning problem as shown in the instructions of **PTCV** and **RC**. Our trained **Prompt Inverter** outperforms DPIC's generic zero-shot method as shown in Table 5, and IPAD also achieves better performance than DPIC overall, as results shown in Appendix E.

3.3 Interpretability Assessment of IPAD

To assess the explainability improvement of IPAD, we designed an interpretability assessment with ten participants evaluating one HWT and one LGT article. We used IPAD version 2 due to its superior OOD performance and attack resistance. Participants compared three online detection platforms (i.e., Scribbr, QuillBot, GPTZero) with IPAD's process (which displayed input texts, predicted prompts, regenerated texts, and final judgments). After evaluation, participants rated IPAD on four key explainability dimensions. Transparency received strong ratings (40%:5, 60%:4), with participants appreciating the visibility of intermediate processes. Trust scores were more varied (10%:3, 70%:4, 20%:5), but IPAD was generally considered more convincing than single-score detectors. Satisfaction was mixed (30%:3, 30%:4, 40%:5), with participants acknowledging better detection but raising concerns about energy efficiency since IPAD runs three LLMs. Debugging received unanimous 5s, as participants could easily analyze the predicted prompt and regenerated text to verify the decision-making process. If needed, users could refine the generated content by adjusting instructions, such as specifying a word count, making IPAD a more effective and user-friendly tool compared to black-box detectors.

4 Related Work

4.1 AI detectors Methods and challenges

Recent studies have explored diverse strategies for detecting AI-generated text. **Watermarking** embeds identifiable patterns during training Gu et al. [2022], Shevlane et al. [2023] or inference Lucas and Havens [2023], but requires model access and is vulnerable to erasure attacks Hou et al. [2024]. **Statistics-based methods** treat output distributions as detection signals. DetectGPT Mitchell et al. [2023b] and Fast-DetectGPT Bao et al. [2024] locate LGT in regions of negative curvature of log-probability; Lastde Xu et al. [2025] and Glimpse Bao et al. [2025] exploit token-probability dynamics and partial-distribution prediction. Other statistical approaches rely on n-gram divergence or revision similarity Hamed and Wu [2023], Kalinichenko et al. [2003], Zhu et al. [2023], Mao et al. [2024], Yang et al. [2024a], though robustness remains limited Wu et al. [2025]. **Regeneration-based methods** compare model rewrites with originals: RAIDAR Mao et al. [2024] and MAGRET Huang et al. [2025] observe stronger edits on human text; DNA-GPT Yang et al. [2024a] and TOCSIN Veselovsky et al. [2023] measure continuation or deletion-based differences. **Neural approaches** fine-tune large encoders (e.g., RoBERTa Liu et al. [2019], BERT Devlin et al. [2019], XLNet Yang et al. [2019]) with adversarial or contrastive objectives Pagnoni et al. [2022], Yang et al. [2024b], while **human-in-the-loop methods** provides complementary semantic judgment and explainability Chaka [2023], Dugan et al. [2023], Uchendu et al. [2023].

4.2 Prompt Inverter techniques and applications

Prompt extraction techniques aim to reverse-engineer the prompts that generate specific outputs from LLMs. Approaches include black-box methods like output2prompt Zhang et al. [2024b], which extracts prompts based on model outputs without access to internal data, and logit-based methods like logit2prompt Mitka [2024], which rely on next-token probabilities but are constrained by access to logits. Adversarial methods can bypass some defenses but are model-specific and fragile Zhang et al. [2024c]. Despite the success of some zero-shot LLM-inversion based methods Li and Klabjan [2024], Yu et al. [2024], they are mostly naive usage of prompting LLMs, which makes them poor in prompt extraction accuracy and robustness.

5 Conclusion

This paper introduces **IPAD (Inverse Prompt for AI Detection)**, a framework consisting of a **Prompt Inverter** that identifies predicted prompts that could have generated the input text, and two **Distinguishers** that examines how well the input texts align with the predicted prompts. One is the *Prompt-Text Consistency Verifier (PTCV)* which evaluates direct alignment between predicted prompts and input text, and the other is *Regeneration Comparator (RC)* that examines content similarity by comparing input texts with the corresponding regenerated texts. Empirical evaluations demonstrate that IPAD outperforms the strongest baselines by 9.05% (Average Recall) on in-distribution data, 12.93% (AUROC) on out-of-distribution (OOD) data, and 5.48% (AUROC) on attacked data. IPAD also performs robustly on structured datasets. While the local alignment in RC approach provides explicit interpretability, it is more sensitive to adversarial attacks. In contrast, the global distribution in **PTCV** matching approach implicitly learns generative LLM's distributional properties, which offers more robustness while maintaining explainability. The combination of the two modules suggests that combining self-consistency checks of generative models with multi-step reasoning for evidential explainability holds promise for future AI detection systems in real-world scenarios. An interpretability assessment reveals that IPAD enhances trust and transparency by allowing users to examine decision-making evidence. Overall, IPAD establishes a new paradigm for more robust, reliable, and interpretable AI detection systems to combat the misuse of LLMs.

While IPAD demonstrates SOTA performance, two limitations warrant discussion: (1) The **Prompt Inverter** may not fully reconstruct prompts containing explicit in-context learning examples, as it prioritizes semantic alignment over precise syntactic replication. (2) While IPAD achieves strong performance across diverse datasets, it relies on LLMs, making it more computationally expensive compared to lightweight detectors such as RoBERTa or HC3. However, compared other detectors compared with LLMs, such as DPIC, IPAD is more lightweight since it calls the open-sources light-weight `Phi-3` model.

 https://doi.org/10.52202/085713-5580

Acknowledgments

This work was supported in part by NSFC grants (62432008 and 623B2002), in part by RGC RIF grant R6021-20, in part by RGC TRS grant T43-513/23N-2, in part by RGC CRF grants (C7004-22G, C1029-22G and C6015-23G), in part by NSFC/RGC grant CRS_HKUST601/24, and in part by RGC GRF grants (16207922, 16207423 and 16203824).

References

Zheng Chen, Di Zou, Haoran Xie, Huajie Lou, and Zhiyuan Pang. Facilitating university admission using a chatbot based on large language models with retrieval-augmented generation. *Educational Technology & Society*, 27(4):pp. 454–470, 2024. ISSN 11763647, 14364522.

Yushi Feng, Tsai Hor Chan, Guosheng Yin, and Lequan Yu. Democratizing large language model-based graph data augmentation via latent knowledge graphs. *Neural Networks*, 191:107777, 2025a.

Fengxiang Cheng, Haoxuan Li, Fenrong Liu, Robert Van Rooij, Kun Zhang, and Zhouchen Lin. Empowering llms with logical reasoning: A comprehensive survey. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, 2025.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.

Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks. *arXiv preprint arXiv:2306.07899*, 2023.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. A survey on llm-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, pages 1–65, 01 2025. ISSN 0891-2017.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Artidoro Pagnoni, Martin Graciarena, and Yulia Tsvetkov. Threat scenarios and best practices to detect neural fake news. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1233–1249, 2022.

Yisroel Mirsky, Ambra Demontis, Jaidip Kotak, Ram Shankar, Deng Gelei, Liu Yang, Xiangyu Zhang, Maura Pintor, Wenke Lee, Yuval Elovici, et al. The threat of offensive ai to organizations. *Computers & Security*, 124:103006, 2023.

Zheng Chen, Zhaoxin Feng, Jianfei Ma, Jiexi Xu, and Bo Li. Can LLMs recognize their own analogical hallucinations? evaluating uncertainty estimation for analogical reasoning. In Yuji Zhang, Canyu Chen, Sha Li, Mor Geva, Chi Han, Xiaozhi Wang, Shangbin Feng, Silin Gao, Isabelle Augenstein, Mohit Bansal, Manling Li, and Heng Ji, editors, *Proceedings of the 3rd Workshop on Towards Knowledgeable Foundation Models (KnowFM)*, pages 84–93, Vienna, Austria, August 2025. Association for Computational Linguistics. ISBN 979-8-89176-283-1.

Peipeng Yu, Jiahan Chen, Xuan Feng, and Zhihua Xia. Cheat: A large-scale dataset for detecting chatgpt-written abstracts. *IEEE Transactions on Big Data*, 2025.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *International Conference on Machine Learning*, pages 17061–17084. PMLR, 2023.

Gregory D. Price and Marc Sakellarios. The effectiveness of free software for detecting ai-generated writing. *International Journal of Teaching, Learning and Education*, 2(6), 2023.

William H. Walters. The effectiveness of software designed to detect ai-generated writing: A comparison of 16 ai text detectors. *Open Information Science*, 7(1):20220158, 2023.

Chaka Chaka. Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic: The case of five AI content detection tools. *Journal of Applied Learning and Teaching*, 6(2): 94–104, 2023.

Debora Weber-Wulff, Alla Anohina-Naumeca, Sonja Bjelobaba, Tomáš Foltỳnek, Jean Guerrero-Dib, Olumide Popoola, Petr Šigut, and Lorna Waddington. Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, 19(1):26, 2023.

Mohanad Halaweh and Ghaleb El Refae. Examining the accuracy of ai detection software tools in education. In *2024 Fifth International Conference on Intelligent Data Science Technologies and Applications (IDSTA)*, pages 186–190, 2024.

Yushi Feng, Junye Du, Yingying Hong, Qifan Wang, and Lequan Yu. PASS: probabilistic agentic supernet sampling for interpretable and adaptive chest x-ray reasoning. abs/2508.10501, 2025b.

Chris Stokel-Walker and Richard Van Noorden. What chatgpt and generative ai mean for science. *Nature*, 614(7947):214–216, 2023.

Sebastian Porsdam Mann, Brian D Earp, Sven Nyholm, John Danaher, Nikolaj Møller, Hilary Bowman-Smart, Joshua Hatherley, Julian Koplin, Monika Plozza, Daniel Rodger, et al. Generative ai entails a credit–blame asymmetry. *Nature Machine Intelligence*, 5(5):472–475, 2023.

Toby Shevlane, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, Lewis Ho, Divya Siddarth, Shahar Avin, Will Hawkins, Been Kim, Iason Gabriel, Vijay Bolina, Jack Clark, Yoshua Bengio, Paul F. Christiano, and Allan Dafoe. Model evaluation for extreme risks. abs/2305.15324, 2023.

Marah I Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat S. Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Dan Iter, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Chen Liang, Weishung Liu, Eric Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Xia Song, Masahiro Tanaka, Xin Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Michael Wyatt, Can Xu, Jiahang Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone. abs/2404.14219, 2024.

Hiyori Yoshikawa and Naoaki Okazaki. Selective-LAMA: Selective prediction for confidence-aware evaluation of language models. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2017–2028, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022.

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 *The Tenth International Conference on Learning Representations*. OpenReview.net, 2022.

 https://doi.org/10.52202/085713-5580

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410, Bangkok, Thailand, August 2024a. Association for Computational Linguistics.

Ryuto Koike, Masahiro Kaneko, and Naoaki Okazaki. OUTFOX: llm-generated essay detection through in-context learning with adversarially generated examples. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 21258–21266. AAAI Press, 2024.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won-Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Youngsook Song, Tae Hwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Eunjeong Lucy Park, Alice Oh, Jung-Woo Ha, and Kyunghyun Cho. KLUE: Korean language understanding evaluation. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 2021*, December 2021.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. abs/2301.07597, 2023.

Reem Alkanhel, El-Sayed M El-kenawy, Abdelaziz A Abdelhamid, Abdelhameed Ibrahim, Mostafa Abotaleb, and Doaa Sami Khafaga. Dipper throated optimization for detecting black-hole attacks in manets. *Computers, Materials & Continua*, 74(1), 2023.

Junchao Wu, Runzhe Zhan, Derek F. Wong, Shu Yang, Xinyi Yang, Yulin Yuan, and Lidia S. Chao. DetectRL: Benchmarking LLM-generated text detection in real-world scenarios. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, Vancouver, BC, Canada, December 2024.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia, July 2018. Association for Computational Linguistics.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 649–657, 2015.

Jinyan Su, Terry Yue Zhuo, Di Wang, and Preslav Nakov. DetectLLM: Leveraging log rank information for zero-shot detection of machine-generated text. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12395–12412, Singapore, December 2023. Association for Computational Linguistics.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. Fast-DetectGPT: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In *The Twelfth International Conference on Learning Representations*, Vienna, Austria, May 2024.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Spotting LLMs with binoculars: Zero-shot detection of machine-generated text. In *Forty-first International Conference on Machine Learning*, Vienna, Austria, July 2024.

Yushi Bai, Jiajie Zhang, Xin Lv, Linzhi Zheng, Siqi Zhu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. Longwriter: Unleashing 10,000+ word generation from long context LLMs. In *The Thirteenth International Conference on Learning Representations*, 2025.

Tianyu Zheng, Ge Zhang, Tianhao Shen, Xueling Liu, Bill Yuchen Lin, Jie Fu, Wenhu Chen, and Xiang Yue. Opencodeinterpreter: Integrating code generation with execution and refinement. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 12834–12859. Association for Computational Linguistics, 2024b.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning*, pages 24950–24962. PMLR, 2023a.

Toon Calders and Szymon Jaroszewicz. Efficient auc optimization for classification. In *European conference on principles of data mining and knowledge discovery*, pages 42–53. Springer, 2007.

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. MAGE: Machine-generated text detection in the wild. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–53, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

Xiao Yu, Yuang Qi, Kejiang Chen, Guoqiang Chen, Xi Yang, Pengyuan Zhu, Xiuwei Shang, Weiming Zhang, and Nenghai Yu. Dpic: Decoupling prompt and intrinsic characteristics for llm generated text detection. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Yiming Zhang, Nicholas Carlini, and Daphne Ippolito. Effective prompt extraction from language models. In *First Conference on Language Modeling (COLM 2024)*, 2024a.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277, 2021.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

Chenxi Gu, Chengsong Huang, Xiaoqing Zheng, Kai-Wei Chang, and Cho-Jui Hsieh. Watermarking pre-trained language models with backdooring. *arXiv preprint arXiv:2210.07543*, 2022.

Evan Lucas and Timothy Havens. Gpts don't keep secrets: Searching for backdoor watermark triggers in autoregressive language models. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing*, pages 242–248, 2023.

Abe Bohan Hou, Jingyu Zhang, Tianxing He, Yichen Wang, Yung-Sung Chuang, Hongwei Wang, Lingfeng Shen, Benjamin Van Durme, Daniel Khashabi, and Yulia Tsvetkov. SemStamp: A semantic watermark with paraphrastic robustness for text generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4067–4082, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning*, pages 24950–24962. PMLR, 2023b.

 https://doi.org/10.52202/085713-5580

Yihuai Xu, Yongwei Wang, Yifei Bi, Huangsen Cao, Zhouhan Lin, Yu Zhao, and Fei Wu. Training-free LLM-generated text detection by mining token probability sequences. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=vo4AHjowKi`.

Guangsheng Bao, Yanbin Zhao, Juncai He, and Yue Zhang. Glimpse: Enabling white-box methods to use proprietary models for zero-shot LLM-generated text detection. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=an3fugFA23`.

Ahmed Abdeen Hamed and Xindong Wu. Improving detection of chatgpt-generated fake science using real publication text: Introducing xfakebibs a supervised-learning network algorithm. abs/2308.11767, 2023.

Leonid A. Kalinichenko, Vladimir V. Korenkov, Vladislav P. Shirikov, Alexey N. Sissakian, and Oleg V. Sunturenko. Digital libraries: Advanced methods and technologies, digital collections: Report on RCDL'2002 – the 4th all-Russian scientific conference, dubna, 15–17 october 2002. *D-Lib Magazine*, 9(1), January 2003.

Biru Zhu, Lifan Yuan, Ganqu Cui, Yangyi Chen, Chong Fu, Bingxiang He, Yangdong Deng, Zhiyuan Liu, Maosong Sun, and Ming Gu. Beat llms at their own game: Zero-shot llm-generated text detection via querying chatgpt. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7470–7483, 2023.

Chengzhi Mao, Carl Vondrick, Hao Wang, and Junfeng Yang. Raidar: generative AI detection via rewriting. In *The Twelfth International Conference on Learning Representations*, 2024.

Xianjun Yang, Wei Cheng, Yue Wu, Linda Ruth Petzold, William Yang Wang, and Haifeng Chen. DNA-GPT: divergent n-gram analysis for training-free detection of gpt-generated text. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024a.

Yifei Huang, Jiuxin Cao, Hanyu Luo, Xin Guan, and Bo Liu. MAGRET: Machine-generated text detection with rewritten texts. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8336–8346, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5754–5764, 2019.

Lingyi Yang, Feng Jiang, Haizhou Li, et al. Is chatgpt involved in texts? measure the polish ratio to detect chatgpt-generated text. *APSIPA Transactions on Signal and Information Processing*, 13(2), 2024b.

Liam Dugan, Daphne Ippolito, Arun Kirubarajan, Sherry Shi, and Chris Callison-Burch. Real or fake text?: Investigating human ability to detect boundaries between human-written and machine-generated text. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 12763–12771. AAAI Press, 2023.

Adaku Uchendu, Jooyoung Lee, Hua Shen, Thai Le, Dongwon Lee, et al. Does human collaboration enhance the accuracy of identifying llm-generated deepfake texts? In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 11, pages 163–174, 2023.

Collin Zhang, John Xavier Morris, and Vitaly Shmatikov. Extracting prompts by inverting LLM outputs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14753–14777, Miami, Florida, USA, November 2024b. Association for Computational Linguistics.

Krystof Mitka. Stealing part of a production language model. B.S. thesis, University of Twente, 2024.

Yiming Zhang, Nicholas Carlini, and Daphne Ippolito. Effective prompt extraction from language models. In *First Conference on Language Modeling*, 2024c.

Hanqing Li and Diego Klabjan. Reverse prompt engineering. *arXiv preprint arXiv:2411.06729*, 2024.

John Xavier Morris, Wenting Zhao, Justin T Chiu, Vitaly Shmatikov, and Alexander M Rush. Language model inversion. In *The Twelfth International Conference on Learning Representations*, 2024.

Yiming Zhang, Nicholas Carlini, and Daphne Ippolito. Effective prompt extraction from language models. In *First Conference on Language Modeling*, 2024d.

A Fine-tune Dataset

Prompt Inverter Dataset. We use the following four datasets, with the first three datasets enhance the model's generalization to recover the prompts, while the last dataset improves performance on essay-related tasks.

- **Instructions-2M** Morris et al. [2024], a collection of 2 million user prompts and system prompts, from which we used 30,000 prompts.
- **ShareGPT** Zhang et al. [2024d], an open platform where users share ChatGPT prompts and responses, from which we used 500 samples.
- **Unnatural Instructions** Zhang et al. [2024d], a dataset of creative instructions generated by OpenAI's models, from which we used 500 samples.
- **OUTFOX dataset** Koike et al. [2024], which contains 15,400 essay problem statements, student-written essays, and LLM-generated essays.

The first three datasets aims to enhance the general querying capability of the **Prompt Inverter**, and are all released under the MIT license. All the samples we used are the same to the samples randomly selected in Zhang et al. [2024b]. The last dataset aims to enhance the familiarity of the **Prompt Inverter** with the data of the essay to detect the LLM-generated essays, and are created and examined by Koike et al. [2024], We specifically used the LLM-generated essays and problem statements for this supervised fine-tuning (SFT). There are 45,400 training pairs in total.

Given that essay data are diverse, we utilize only the OUTFOX dataset Koike et al. [2024]. To adapt this dataset for training our **Distinguisher**, we enhance it to align with the model's requirements. The original dataset consists of 14,400 training triplets of essay problem statements, student-written essays, and LLM-generated essays. To further process the data, we apply the **Prompt Inverter** to both student-written and LLM-generated essays, generating corresponding *Predicted Prompts*. These *Predicted Prompts* are then used to regenerate texts via **ChatGPT**, i.e. **gpt-3.5-turbo**. Following this procedure, we construct a total of 28,800 training samples, with an equal distribution of positive and negative examples (14,400 each).

The final dataset is structured as follows:

 https://doi.org/10.52202/085713-5580

Table 7: Instruction, input/output structure, and inference outputs of each fine-tuned module. T is the input text, P the predicted prompt, and T' the regenerated text.

Field	Prompt Inverter	PTCV	RC
Instruction	"What is the prompt P that generates the Input Text T?"	"Can LLM generate the input text T through the prompt P?"	"T' is generated by LLM, determine whether T is also generated by LLM with a similar prompt."
Input	T	(P, T)	(T', T)
Output	P	"yes"/"no"	"yes"/"no"
Output in Inference	P	p_{PTCV}	p_{RC}

B AUROC formula

Since our model predicts binary labels, we follow the *Wilcoxon-Mann-Whitney* statistic Calders and Jaroszewicz [2007] to calculate the Area Under Receiver Operating Characteristtic curve (AUROC):

$$\text{AUC}(f) = \frac{\sum_{t_0 \in \mathcal{D}^0} \sum_{t_1 \in \mathcal{D}^1} \mathbf{1}[f(t_0) < f(t_1)]}{|\mathcal{D}^0| \cdot |\mathcal{D}^1|}$$

where $\mathbf{1}[f(t_0) < f(t_1)]$ denotes an indicator function which returns 1 if $f(t_0) < f(t_1)$ and 0 otherwise. $\mathcal{D}^0$ is the set of negative examples, and $\mathcal{D}^1$ is the set of positive examples.

C Ablation study data structures

Input-only fine-tuning data instructions. "Is this text generated by LLM?"

Prompt Only fine-tuning data instructions. "Prompt Inverter predicts prompt that could have generated the input texts. Is this prompt predicted by an input texts written by LLM?"

Ablation Prompt. "Text A is generated by an LLM. Determine whether Text B is also generated by an LLM using a similar prompt. Meanwhile, determine whether Text B could have been generated from Prompt C using an LLM. Answer with YES or NO."

D DPIC (decouple prompt and intrinsic characteristics) Prompt Extraction Zero-shot Prompts Yu et al. [2024]

"I want you to play the role of the questioner. I will type an answer in English, and you will ask me a question based on the answer in the same language. Don't write any explanations or other text, just give me the question. <TEXT>.".

E Comparison with DPIC

Since DPIC has not released its code, data, or models, we are unable to independently evaluate the performance of its classifier. Consequently, we rely on the reported results in the DPIC paper and construct a comparable dataset following their described settings to enable a fair comparison with IPAD. However, due to these limitations, we are unable to apply DPIC to additional datasets for broader evaluation.

To assess the generalization of IPAD, we reconstruct the following datasets, each containing 200 randomly sampled examples: **XSum**, **WritingPrompts**, and **PubMedQA**. For each dataset, we generate texts using three large language models: `ChatGPT (gpt-3.5-turbo)`, `GPT-4 (gpt-4)`, and `Claude 3 (claude-3-opus-20240229)`. Furthermore, the XSum datasets generated by these three models are augmented using two attack methods—**DIPPER** and **Back-Translation**—resulting in a total of 15 evaluation datasets.

Table 8: AUROC comparison across tasks (XSum, Writing, PubMed) for ChatGPT, GPT-4, and Claude 3 using various prompt extraction methods.

Method	**ChatGPT**				**GPT-4**				**Claude 3**			
	XSum	Writing	PubMed	Avg.	XSum	Writing	PubMed	Avg.	XSum	Writing	PubMed	Avg.
DPIC (ChatGPT)	1.0000	0.9821	0.9092	0.9634	0.9996	0.9768	0.9438	0.9734	1.0000	0.9950	0.9686	0.9878
DPIC (Vicuna-7B)	0.9976	0.9708	0.8990	0.9558	0.9986	0.9644	0.9394	0.9674	0.9992	0.9943	0.9690	0.9875
IPAD (Version 1)	0.9850	0.9800	0.9250	0.9633	1.0000	0.9700	0.9700	0.9800	1.0000	0.9800	0.9750	0.9850
IPAD (Version 2)	1.0000	0.9850	0.9800	0.9883	1.0000	0.9800	0.9500	0.9767	1.0000	0.9950	1.0000	1.0000

Table 9: AUROC comparison under generation perturbation settings (DIPPER, Back-translation) for each model.

Method	**ChatGPT**			**GPT-4**			**Claude 3**		
	Ori.	DIPPER	Back-trans.	Ori.	DIPPER	Back-trans.	Ori.	DIPPER	Back-trans.
DPIC (ChatGPT)	1.0000	1.0000	0.9972	0.9996	0.9991	0.9931	1.0000	0.9996	0.9878
DPIC (Vicuna-7B)	0.9976	0.9980	0.9889	0.9986	0.9969	0.9903	0.9992	0.9996	0.9979
IPAD (Version 1)	0.9850	0.8900	0.9850	1.0000	0.8950	0.9900	1.0000	0.9250	0.9950
IPAD (Version 2)	1.0000	0.9750	0.9950	0.9800	0.9750	0.9950	1.0000	1.0000	1.0000

Based on the experimental results, IPAD performs well and exhibits notable resistance to adversarial attacks.

IPAD open-sourced all the fine-tuned models, including the Prompt Inverter, and the two versions of distinguishers. Therefore, all the experiment results can be validated and reproduced.

F IPAD and DPIC prompt inverter examples

https://doi.org/10.52202/085713-5580

Table 10: IPAD and DPIC prompt inverter examples

Input	IPAD	DPIC
The IPC opened proceedings against the National Paralympic Committee of Russia after a report claimed the country had operated a widespread doping programme.A decision on any ban will come in the week commencing 1 August.The International Olympic Committee (IOC) has opted against a blanket ban."I can assure you that our board will take the right decision in the interest of sport and the interest of the Paralympic movement moving forward," said Craven.On Sunday, the IOC said it would leave it up to the governing bodies of individual sports to decide if Russian competitors are clean and should be allowed to take part.But Craven, himself a member of the IOC, was critical of that decision and said the IPC would not necessarily follow suit."I am disappointed in their decision, but that is a personal view," he added."We have to acknowledge their right to take such a decision. This is ultra-serious. I don't think there has been a situation in the past where you have had institutional doping on such a scale."We believe the Russian NPC is either unwilling or unable to uphold the IPC anti-doping code, which is in line with the World Anti-Doping Agency code, so that is what they have to respond to."Canadian law professor Richard McLaren's report, published last week, claimed Russia operated a state-sponsored doping programme from 2011 to 2015.The IPC said it acted after McLaren provided the names of the athletes associated with the 35 "disappearing positive samples" from the Moscow laboratory highlighted in the report.Nineteen samples potentially doctored as part of the sample-swapping regime during the 2014 Sochi Paralympic Winter Games have been sent for further analysis.Russia will have up to 21 days to appeal against any IPC decision, with the Rio Paralympics due to begin on 7 September.	Explain the actions taken by the International Paralympic Committee against the Russian Paralympic Committee and the reasons behind these actions.	What were the key findings of Richard McLaren's report regarding Russia's doping programme?
The world came crashing down in minutes. Many of us were asleep when it happened, and did n' t find out about it until later. When we awoke, we saw the carnage spread through the land, and we wept. There were n' t many of us left, but what few there were managed to find each other over the Internet. We gathered together in what remained of a major city on the East Coast of what was once the United States. It took us time, but we eventually began to rebuild. The brightest among those who survived thought to ask " Why, " while most of us were content with just surviving. Years passed, and no link was found between us. Eventually, those who had the question resigned themselves to the fact that they would never know. They went to their new homes, and tried to integrate themselves as best they could into the new society. It was n' t until 14 years after the event happened that the connection was discovered, quite by accident. One of the former questioners had taken a job as a mover, and was helping a fellow survivor move into a newly cleaned house. Sticking out of one of the boxes was the missing link. " Oh, you used to shop at Bad Dragon too? " Moments later, the realization struck him.	Write an essay about a time when you and your friends were the only survivors of a catastrophic event that wiped out most of the world's population. Describe how you and your friends coped with the aftermath and the challenges you faced in rebuilding society.	What was the event that caused the world to come crashing down?

Input	IPAD	DPIC
In an alternate timeline, a second sentient race evolved in parallel with humans. These beings, known as the Avralians, possessed extraordinary abilities and resided in the hidden corners of the Earth. For centuries, unaware of each other's existence, humans and Avralians progressed independently. However, fate intervened one fateful day when an Avralian girl named Elara stumbled upon a group of human explorers deep in the forest. With wide-eyed wonder, she approached them cautiously. The humans, initially startled, soon recognized Elara's unique nature. They befriended her, and through her, the two races began to unravel the secrets of their intertwined existence. As time passed, harmony emerged between the two races, and mutual respect solidified their bond. The Avralians shared their ancient knowledge and wisdom, while humans brought innovation and technology. Together, they worked towards a world where unity triumphed over differences. However, not all were convinced of this newfound alliance. A faction within the human population, fueled by fear and mistrust, sought to exploit Avralian abilities for personal gain. Conflict brewed, threatening the fragile equilibrium. Elara, fueled by her pure heart and belief in unity, emerged as a beacon of hope. With a small band of supporters, she embarked on a journey to bridge the gap and foster understanding. Battles were fought, sacrifices made, but ultimately, Elara's message prevailed. Humans and Avralians learned to cherish their diversity and forge a future marked by collaboration and empathy. The world transformed into a tapestry of coexistence, where magnificent cities stood as testaments to unity and cultural exchange. Humans and Avralians moved freely through bustling markets, sharing knowledge, stories, and laughter. Together, they faced global challenges, from climate crises to epidemics, with unwavering determination.	Write an essay describing an alternate timeline in which a second sentient race evolved in parallel with humans, exploring the potential interactions and conflicts between the two species.	How did Elara manage to convince both races to embrace unity despite the conflict?
Both times I had the banana pepper appetizer, which is great and goes really well with the FRESH and delicious bread and cheese they give you at the start of your meal. nnFor entrees, me and my girlfriend have had mixed experience. I've had the fish sandwich (very good) and the eggplant parm sandwich (okay). My girlfriend got the salad with bread and basil on it, but the basil was over powering and the bread was soggy with the dressing. nnThe service is also a mixed bag. The first time our server went out of her way to take care of us and even MADE me cocktail sauce for my fish sandwich. The second time, the server was lackluster, didn't know anything about the menu and wasn't able to take proper care of us. nnI would return to Papa J's, but I my terrible experience last time isn't enough to say it would be my first pick of places to eat around Carnegie/Robinson.	This was a great place to stop for a quick lunch. The lines were not too long for the sandwiches they had and they had a wide selection of bagels if you wanted a bagel sandwich. With a great front patio for enjoying your food, it was a relaxing place to stop. Write a review for it.	What made the banana pepper appetizer stand out to you compared to other starters?

https://doi.org/10.52202/085713-5580

Input	IPAD	DPIC
Abstract: This article explores the longstanding debate between Einstein's theory of general relativity and Maxwell's theory of electromagnetism regarding the nature of gravitation. The central question addressed is whether gravitation is best understood as a curvature of space, a field in flat space, or perhaps a combination of both concepts. Drawing upon a comprehensive analysis of the theoretical framework and empirical evidence, the article presents a nuanced examination of the arguments put forth by Einstein and Maxwell.The article begins by discussing Einstein's general theory of relativity, which proposes that gravitation arises from the curvature of spacetime caused by mass and energy. It outlines the mathematical formalism used to describe this curvature and highlights the key predictions and experimental confirmations of the theory. Conversely, the article delves into Maxwell's electromagnetic theory, which suggests that gravitation may be explained as a fundamental force mediated by a field propagating through flat space, similar to electromagnetic fields.Further, the article explores the distinctive features and limitations of each theory. It scrutinizes the conceptual foundations, mathematical rigor, and empirical support for both approaches, highlighting their respective strengths and weaknesses. Moreover, the article examines attempts to reconcile the two theories into a unified framework, such as the development of theories of quantum gravity.By critically evaluating the arguments and evidence from both camps, this article aims to offer a comprehensive assessment of the question regarding the nature of gravitation. Based on the analysis presented, it becomes evident that both Einstein's theory of general relativity and Maxwell's theory of electromagnetism provide valuable insights into the phenomenon of gravitation.	Write a paper abstract to explain the debate between Einstein's theory of general relativity and Maxwell's theory of electromagnetism regarding the nature of gravitation, and argue for which theory is more likely to be correct based on the evidence presented in the essay statement.	What are the main challenges in reconciling Einstein's theory of general relativity with Maxwell's theory of electromagnetism in explaining gravitation?

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We explained the contributions listed in the abstract and introduction in Section 2-4.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We present the limitation in the section Conclusion.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

https://doi.org/10.52202/085713-5580

Justification: There is no theorem or lemma in this paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The codes are presented in the anonymous github, and the results can be reproduced by following the readme.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The codes are presented in the anonymous github, and the results can be reproduced by following the readme. The dataset for this paper are publicly available.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental details are presented.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We set the temperature parameter of the large language models to 0, so the results are deterministic.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

https://doi.org/10.52202/085713-5580

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We use one A800-80GB gpu and one Nvidia v100 to conduct our experiments.

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This paper has no ethics problem.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the potential societal impacts mainly in section 3.2.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate

to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: There are no such risks in our paper.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: : We have cited the original papers and included proper license.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

https://doi.org/10.52202/085713-5580

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: We do not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: We have no crowdsourcing experiments and human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: We have no crowdsourcing experiments and human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: Pretrained LLMs are used as backbones in our method, which is clearly stated in this paper.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

https://doi.org/10.52202/085713-5580