MoRE: A Mixture-of-Representation-Experts Model for LLM Text Detection Under PGD Attacks

Shifali Agrahari*, Subhashi Jayant*, Sanasam Ranbir Singh
Department of Computer Science & Engineering
Indian Institute of Technology Guwahati
Guwahati, Assam, India - 781039

***Abstract*—The ability of modern LLMs to craft text that closely aligns with human writing styles has intensified discussions about academic integrity and the credibility of online material. Existing detectors perform well on directly LLM-generated text but degrade sharply under simple adversarial edits such as paraphrasing or grammar-based perturbations. To address these limitations, we propose MoRE, a Mixture of Representation Experts framework that integrates stylometric, semantic, and contextual experts combined with PGD-based adversarial augmentation and dual loss. We further curate a comprehensive essay dataset containing human-written essays and essays generated by six major LLM families (Gemini, LLaMA, Mistral, Phi, Gemma, and DeepSeek), along with four adversarial attack variants that emulate real student evasion strategies. Extensive experiments across eight datasets show that MoRE achieves state-of-the-art performance in both binary and multi-class settings, maintaining high F1 scores across attacks, LLMs, and cross-model evaluations. Overall, MoRE establishes a strong, practical foundation for adversarially LLM-generated text detection in education and real-world scenarios.** [1]

***Index Terms*—MoE, LLM-Generated text detection, Essay**

I. Introduction

Large Language Models (LLMs) [1]–[3] have developed rapidly in recent years, resulting in their text generation capabilities being on par with human writing. Modern LLMs are now routinely used to generate product reviews [4], [5], write news articles [6], and are widely adopted in academic contexts, including essay writing, homework and assignment assistance, and automated content creation [7]–[9]. The widespread availability of LLMs has raised substantial concerns regarding academic integrity and the authenticity of content. Within the educational domain, the potential misuse of LLMs poses a significant concern [10]. For example, students may rely on LLMs to generate essays and complete writing tasks, potentially achieving high grades without developing the essential skills associated with critical thinking and academic writing. Such risks further underscore the necessity of addressing the ethical and practical implications of LLM usage and of establishing effective safeguards to mitigate the potential negative consequences of AI-assisted academic work.

Although several studies have explored LLM-generated essay detection [7], [8], most existing methods perform well when the essay is directly generated by an LLM. However, in real-world academic settings, students often apply various forms of adversarial modifications [11] to evade detection. For example, they may intentionally introduce grammatical errors, spelling mistakes, sentence shuffling, or minor paraphrasing to make the generated essay appear more human-like. Such adversarial manipulations can significantly degrade the performance of current detectors, highlighting their limited robustness in practical educational environments [10]. Beyond essay detection, broader research on identifying LLM-generated text falls into four categories: (i) **zero-shot detectors** [12]–[15], which rely on perplexity or rank-based signals but degrade under simple adversarial edits; (ii) **fine-tuned classifiers** [16]–[18], which struggle to generalize to unseen LLMs or out-of-domain settings; (iii) **watermarking approaches** [19], which require control over the generating model; and (iv) **adversarial learning methods** [7], [20]–[22], which improve robustness but still face challenges on complex manipulations.

Despite these advances, existing detectors demonstrate strong performance primarily on unattacked text but struggle to generalize across diverse attack types, languages, or LLM families [19]. Essay-style writing makes these challenges worse because its length and discourse structure amplify the impact of small, meaning-preserving edits. Even minor changes can break the cues single-stream detectors rely on, highlighting the need for models that fuse multiple linguistic signals for robust essay-level detection. In the literature, only limited work examines feature-based detectors [8], [23], which alone are insufficient for reliable detection. A few studies explore sentence-level detection [18], while others focus on contextual-level training [24], but each faces its own limitations. Researchers have also explored token-level adversarial training [7], [21], but it often faces limitations such as generating unrealistic perturbations or failing to preserve semantic consistency. Also some studies [10], [24] report high accuracy despite low class-wise F1-scores, detectors misclassify well-written, grammatically perfect human text as LLM-generated.

To overcome these limitations, we propose **MoRE**, a Mixture of Representations Expert framework designed specifically for the challenges posed by essay-style LLM-generated text detection. MoRE integrates three experts: stylometric, semantic, and contextual to capture the full spectrum of linguistic cues present in text also prevent overfitting. The

*Equal contribution.
Corresponding author: shifali@iitg.ac.in
[1]The dataset and code repository are publicly available at Github repository

stylometric expert incorporates interpretable features [25] such as readability indices, which help identify subtle deviations typically found in LLM-generated writing. The semantic expert captures sentence-level meaning and coherence, making the model less sensitive to paraphrasing or word-level edits. The contextual expert leverages transformer-based token-level representations to differentiate deeper distributional patterns associated with LLM outputs. By combining these diverse signals within a unified architecture and strengthening robustness through Projected Gradient Descent (PGD) [26] based adversarial augmentation, MoRE addresses three key limitations: (i) PGD operates within the semantic and contextual experts, preventing unrealistic or meaning-breaking adversarial samples, (ii) stylometric features constrain the perturbation search space, ensuring linguistically coherent and human-like perturbations, and (iii) it moves beyond superficial lexical cues and instead learns stable, content-level signatures that persist under adversarial perturbations. Most existing essay datasets use only one or a few LLM-generated texts and do not include adversarial attacks. To support robust LLM-generated essay detection, we curated a comprehensive dataset comprising both human-written and six LLMs Gemini 2 [27], LLaMA 2 [1], Mistral 7b [2], Phi 2 [28], Gemma 7b [29], and DeepSeek 7b [3] generated essays. To evaluate robustness against evasion, we further constructed adversarially perturbed versions of the AI-generated essays using four attack strategies [11] such as Paraphrase Attack, Word Substitution Attack, Sentence Shuffling Attack and Grammar & Spelling Error Attack. Our dataset mirrors real student evasion strategies and serves as a challenging benchmark for essay-level adversarial robustness.

Furthermore, we conduct an in-depth analysis of how LLM-generated essays differ from human-written essays across diverse stylometric features, reinforcing the necessity of incorporating specialized expert representations. Extensive experiments benchmark MoRE on the binary classification task (Human vs. LLM) against six state-of-the-art detectors across eight existing datasets, adversarial attack settings, and cross-domain evaluation scenarios. We further evaluate MoRE in a multi-class setting, where each LLM is treated as an individual class, demonstrating its ability to capture distinct generative signatures across model families. This study also explores three core research questions: **RQ1:** Can MoRE stay accurate when the text fights back through adversarial attacks? **RQ2:** How well does MoRE read across different LLM "personalities" and still spot the machine? **RQ3:** Can MoRE detect AI text even when trained on one model and tested on another? Finally, a detailed ablation study quantifies the contribution of each expert module and the impact of adversarial augmentation. Experiments show that MoRE performs consistently across attacks, LLMs, and domains, demonstrating that robust, fusion-based architectures meaningfully improve practical LLM-generated text detection in practical educational scenarios.

II. Related Work

Tackling the malicious uses of LLMs, recent studies [10] have proposed detectors to identify LLM-generated texts. LLM-generated text detectors are mainly categorized into four different categories. (i) Zero-shot or statistical outlier detection methods [12]–[15] includes n-gram frequencies [15], entropy [15], perplexity [15], token log probabilities [30], token ranks [31], and negative curvature regions of the model's log probability [14]. Supervised classifiers are models specifically trained to discriminate between human-written and LLM-generated texts with labels. (ii) supervised classifiers [17], [32], [33] The classifiers range from classical methods [31], [34]to neural-based methods [24], [30]. Finetuned detectors frame detection as a binary classification task. GPTZero performed best among off-the-shelf detectors, and fine-tuned RoBERTa reached 99% accuracy at the document level for argumentative essay detection [24]. (iii) Watermarking algorithms [19] approaches are in-domain but tend to overfit and lose cross-domain robustness. Water-marking algorithms use token-level secret markers in texts that humans cannot recognize for detection. In order to embed the markers into texts, the probabilities of selected tokens by a hash function are modified to be higher in text generation at each time step [19] and (iv) adversarial learning methods [21], [35].

Adversarial learning has been applied to increase detector robustness against paraphrased or style-altered text. RADAR [21] employs a detector-paraphraser competition, improving AUROC by 31.64% against unseen paraphrasers. OUTFOX [7] uses mutual evaluation between an attacker generating undetectable essays and a detector adapting via adversarial examples, outperforming prior statistical and RoBERTa-based methods. Other studies examine the impact of adversarial attacks on detectors. Most detectors perform well on unattacked content but lose generality across attack types and language models, highlighting the need for detection systems that are both accurate and attack-resilient. Despite recent progress, existing AI-text detectors perform well only on clean inputs but fail to generalize across paraphrasing, word substitutions, style transfer attacks, languages, and LLM families. Prior work offers limited exploration of feature-based [8], [23] or sentence-level detectors [18]. However, to the best of our knowledge, no prior work integrates a mixture of heterogeneous representations nor incorporates PGD-based adversarial training for LLM text detection. To address this gap, MoRE integrates complementary stylometric, semantic, and contextual experts with PGD-based adversarial augmentation to achieve robust, attack-resilient LLM-generated text detection.

III. Proposed Model

We propose **MoRE** Mixture of Representations Expert, an adversarially robust framework model for LLM-generated text detection. MoRE integrates with three expert representation modules. Each expert captures a distinct dimension of textual information such that (i) the Stylometric Expert (E_1) interpretable linguistic statistics, (ii) Semantic Expert (E_2) global sentence-level semantics, and (iii) the Contextual Expert (E_3) deep token-level transformer embeddings. Together with Projected Gradient Descent (PGD)-based adversarial augmentation, the model is optimized to remain robust under

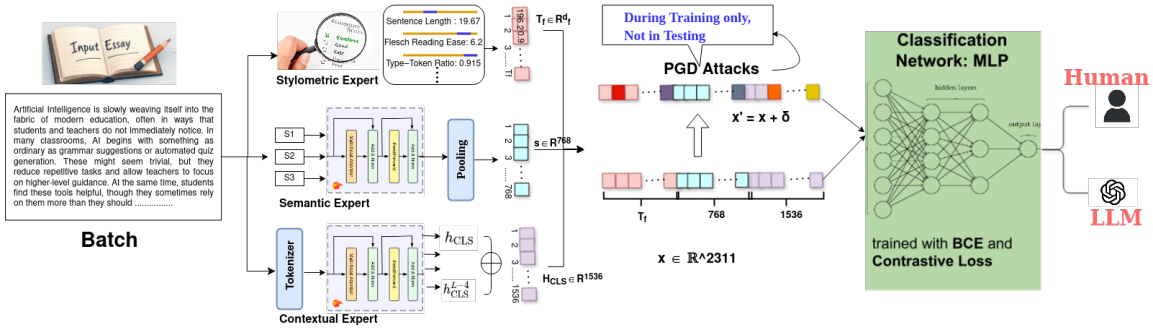

Fig. 1. Overview of our Proposed Mixture-of-Representation-Experts (MoRE) model. Stylometric, semantic, and contextual experts extract complementary features, which are fused and trained using BCE and contrastive objectives. PGD attacks are applied only during training to enhance robustness.

paraphrased, manipulated, or adversarially modified inputs. The MoRE pipeline consists of three components, such as (i) expert-level representation extraction, (ii) PGD adversarial augmentation, and (iii) a dual-objective classification network. A detailed description of each component follows.

A. Expert Representations

Given an input essay t, MoRE extracts three distinct representation vectors using the stylometric expert E_1, the semantic expert E_2, and the contextual expert E_3. These experts are designed to model complementary properties of the text.

(i) Stylometric Expert (E_1) The Stylometric Expert captures interpretable linguistic patterns across multiple dimensions including statistical, stylometric, and LLM based [23], [36]. A large set of handcrafted features is initially extracted and aggregated into a raw feature vector. $T_F \in \mathbb{R}^{d_1}$ Through discriminative feature selection and visualization methods (Fig. 2), the most informative subset of features is retained, forming the final stylometric representation as $T_f \in \mathbb{R}^{d_f}$., here d_f is 7. The expert helps capture linguistic deviations typically exhibited by LLM-generated text, such as reduced stylistic variance or unnatural distribution of lexical features.

(ii) Semantic Expert (E_2) To extract high-level semantic meaning that remains relatively stable under paraphrasing or word substitutions, we encode each essay using a sentence-level transformer encoder. The semantic expert outputs a fixed-size embedding that $E_2(t) = s \in \mathbb{R}^{768}$, represents the holistic semantic structure of the text. Its representation captures global coherence signals, topic consistency, and semantic similarity patterns characteristic of human vs. LLM written essays.

(iii) Contextual Expert (E_3) To obtain deep contextual information and token-level distributional signatures associated with LLM generation patterns, each essay is passed through a pre-trained transformer model. We extract two CLS token embeddings from different layers as defined by Equation 1

$$h_{\text{CLS}}^{(L)} \in \mathbb{R}^{768}, \qquad h_{\text{CLS}}^{(L-4)} \in \mathbb{R}^{768}. \tag{1}$$

These capture surface-level and deeper contextual dependencies. The two embeddings are concatenated $H_{\text{CLS}} = \left[h_{\text{CLS}}^{(L)} \,\|\, h_{\text{CLS}}^{(L-4)}\right] \in \mathbb{R}^{1536}$ as mention [37] Finally, The unified MoRE representation x is obtained by concatenating all three expert representation outputs as represented in Equation 2.

$$x = \left[h_{\text{CLS}}^{(L)} \,\|\, h_{\text{CLS}}^{(L-4)} \,\|\, s \,\|\, T_f\right] \in \mathbb{R}^{2311}. \tag{2}$$

The representation leverages the advantages of all experts, such as (i) T_f providing interpretable human stylistic features, (ii) s encoding sentence-level semantic coherence, and (iii) H_{CLS} capturing deep contextual signals produced by LLMs.

B. Adversarial Augmentation (PGD)

To improve robustness under adversarial attacks such as paraphrasing, synonym substitution, and embedding-level attacks, we incorporate **Projected Gradient Descent (PGD)** [22] during training. Given the fused representation x and classifier parameters θ, PGD computes an adversarial perturbation δ by maximizing the classification loss as mention in Equation 3

$$\delta^{(k+1)} = \Pi_{\|\delta\|_p \le \epsilon}\left[\delta^{(k)} + \alpha \cdot \text{sign}\left(\nabla_x L(f(x+\delta^{(k)};\theta), y)\right)\right], \tag{3}$$

Where α is the step size, and Π is a projection operator enforcing the perturbation constraint.

The adversarially perturbed representation is $x' = x + \delta$, $\text{PGD}(x) = x'$. To ensure realistic adversarial modifications, the perturbation is constrained within an ϵ-ball as defined. $\delta \in \{\delta : \|\delta\|_p \le \epsilon\}$. MoRE is trained using both clean representations x and adversarial representations x', allowing the classifier to learn decision boundaries that remain stable under adversarial transformations. PGD simulates worst-case perturbations in embedding space. Training with these examples helps the model maintain detection reliability even when an adversary deliberately modifies text to evade detection.

C. A Dual-Objective Classification Network

The final stage of MoRE is a dual-objective classification network designed to jointly optimize (i) binary classification between human- and LLM-generated text and (ii) discriminative representation learning through supervised contrastive learning. This design ensures that the classifier not only separates the two classes effectively but also learns a representation space that clusters similar samples while pushing apart dissimilar ones, thereby improving robustness under adversarial perturbations.

1) Main Classification Path: Given the unified MoRE representation $x \in \mathbb{R}^{2311}$, we first apply layer normalisation $\tilde{x} = \text{LayerNorm}(x)$, to stabilize training and reduce feature-scale variations across experts. The normalized vector is then passed through a two-layer feed-forward network with hidden sizes 512 and 256, using ReLU activations and dropout (0.3,

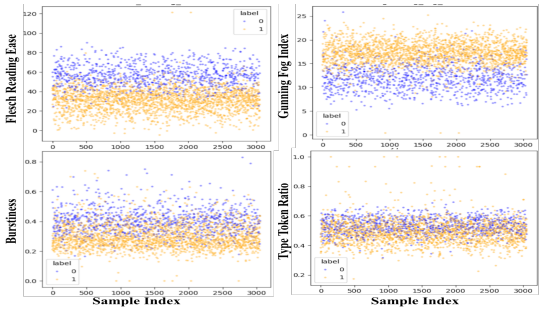

Fig. 2. Visualization of key stylometric feature distributions for human-written (label 0) and LLM-generated (label 1) essays. Each scatter plot illustrates how specific linguistic characteristics vary across samples, highlighting measurable differences between human and machine-generated text.

0.2) to reduce overfitting. A final sigmoid layer outputs the probability that the input text is LLM-generated.

2) Contrastive Projection Head: MoRE includes a supervised contrastive projection head to improve class separability during training. The MoRE representation x is passed through a two-layer projection module ($2311 \rightarrow 128 \rightarrow 128$) to obtain a compact embedding z used only for the supervised contrastive loss. The projection head is discarded during inference.

3) Binary Cross-Entropy Loss: Classification is trained using class-balanced Binary Cross-Entropy (BCE) loss in Equation 4

$$L_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(p_i) + (1-y_i)\log(1-p_i)\right]. \tag{4}$$

Loss encourages correct classification of humans vs LLMs.

4) Supervised Contrastive Loss: Let z_i the contrastive embedding for sample i, with label $y_i \in \{0,1\}$ Equation 5.

$$P(i) = \{p : y_p = y_i,\ p \neq i\}, \quad A(i) = \{a : a \neq i\}. \tag{5}$$

The supervised contrastive loss for the sample i is shown in Equation 6.

$$L_{\text{Con}}(i) = -\frac{1}{|P(i)|}\sum_{p\in P(i)} \log \frac{\exp\left(\frac{z_i \cdot z_p}{\tau}\right)}{\sum_{a\in A(i)} \exp\left(\frac{z_i \cdot z_a}{\tau}\right)}, \tag{6}$$

Where τ is the temperature hyperparameter. If a minibatch contains only one class (e.g., due to augmentation), it L_{Con} is skipped for that batch.

5) Joint Optimization: MoRE is trained using a weighted combination of BCE and contrastive loss as defined in Equation 7

$$L_{\text{Total}} = L_{\text{BCE}} + \lambda\, L_{\text{Con}} \tag{7}$$

where λ controls the strength of the contrastive learning objective. The dual-objective setup encourages the classifier to leverage both supervised decision boundaries and contrastive structure in the representation space, improving generalization and robustness to adversarially manipulated inputs. To extend MoRE to a multi-class setting, we replace the BCE loss with the standard multi-class softmax cross-entropy loss.

IV. Experimental Setup

In this section, we describe our dataset preparation, baseline detectors, evaluation metrics, and training configuration.

A. Dataset Preparation

We collect a total of 6672 human-written essays sourced from publicly available TOEFL and IELTS preparation materials. To generate the corresponding LLM-generated essays, we use six LLM such that Gemini (gemini-2.0-flash) [27], LLaMA (LLaMA-2-7b-chat-hf) [1], Hermes-Mistral (Nous-Hermes-2-Mistral-7B-DPO) [2], Phi (phi-2) [28], Gemma (gemma-7b-it) [29], and DeepSeek (deepseek-llm-7b-base) [3]. Each model is prompted to generate an essay for every topic using the following instruction:

> "Write a detailed essay on the {topic}:. Only provide the full essay content. Do not include outlines, instructions, sample headings, vocabulary, or author information."

Furthermore, to assess the detector's robustness under evasion attempts, we generate adversarially perturbed versions of the essays using four diverse attack strategies [11] including (i) **Paraphrase Attack:** Each essay is paraphrased by prompting the same set of LLMs to rewrite the text while preserving its meaning and coherence. The prompt used is: *"Paraphrase the following essay while maintaining its meaning and context. Use different phrasing and sentence construction to make it appear written by a different person."*, (ii) **Word Substitution Attack:** Individual words are replaced with synonyms or semantically related alternatives using WordNet from the NLTK library, ensuring grammatical correctness is maintained, (iii) **Sentence Shuffling Attack:** Sentences within each essay are randomly reordered to disrupt global coherence without altering the sentence-level content, and (iv) **Grammatical Errors & Spelling Mistakes Attack:** Artificial grammatical errors and typographical mistakes are introduced to emulate human-like writing imperfections, using NLTK resources such as Punkt, the Averaged Perceptron Tagger, and OMW-1.4.

This diverse adversarial dataset allows for a comprehensive evaluation of detection models across four attack types and six different LLM sources, better simulating real-world evasion strategies. To compare our proposed MoRE framework with existing literature, we evaluate its performance against three widely used essay datasets: the Outfox dataset [7], the ChatGPT Generated Text Detection Corpus [8], and the Kaggle IELTS Writing Dataset[2]. Beyond essay-focused dataset, we also assess cross-domain generalization using the HC3 dataset [24], which spans five diverse domains: HC3 Finance, HC3 Medicine, HC3 OpenQA, HC3 Reddit ELI5, and HC3 Wiki. In total, our analysis incorporates eight publicly available benchmark datasets, enabling a comprehensive and rigorous comparison of MoRE against state-of-the-art detectors across multiple domains and writing styles.

B. Baselines and Metric

To comprehensively evaluate the effectiveness of MoRE, we compare it against a diverse set of existing detection approaches. Specifically, we include four zero-shot statistical detectors Log Probability [15], Rank [15], rank [15], and

[2]https://www.kaggle.com/datasets/mazlumi/ielts-writing-scored-essays

TABLE I

COMPARISON OF THE PROPOSED MORE MODEL ON OUR CURATED DATASET AND BASELINE DATASETS, AND BASELINE DETECTORS EVALUATED ACROSS BOTH OUR DATASET AND EXISTING BENCHMARK DATASETS.

Detector	Create Dataset		Outfox (Essay)		ChatGPT Essay		Kaggle Essay		HC3 Finance		HC3 Medicine		HC3 OpenQA		HC3 Reddit ELI5		HC3 Wiki	
	F1(H)	F1(M)	F1(H)	F1(M)	F1(H)	F1(M)	F1(H)	F1(M)	F1(H)	F1(M)	F1(H)	F1(M)	F1(H)	F1(M)	F1(H)	F1(M)	F1(H)	F1(M)
Log p	0.167	0.513	0.44	0.44	0.65	0.65	0.21	0.22	0.35	0.39	0.13	0.16	0.01	0.41	0.44	0.06	0.48	0.48
Rank	0.167	0.518	0.44	0.44	0.65	0.65	0.21	0.22	0.35	0.39	0.13	0.16	0.01	0.41	0.44	0.06	0.48	0.48
LogRank	0.148	0.432	0.48	0.38	0.65	0.07	0.26	0.11	0.40	0.18	0.14	0.15	0.04	0.15	0.44	0.06	0.52	0.20
Entropy	0.12	0.187	0.50	0.26	0.08	0.03	0.24	0.18	0.04	0.08	0.03	0.04	0.07	0.19	0.41	0.00	0.05	0.06
OpenAI	0.78	0.735	0.68	0.16	0.96	0.96	0.80	0.71	0.91	0.92	0.95	0.95	0.72	0.90	0.97	0.92	0.92	0.93
RADAR	0.823	0.952	0.83	0.84	1.00	1.00	0.98	0.98	0.64	0.76	0.36	0.73	0.78	0.83	0.49	0.48	0.84	0.85
MoRE	0.915	0.977	0.94	0.94	0.95	0.95	0.81	0.87	0.81	0.79	0.60	0.78	0.72	0.55	0.82	0.49	0.73	0.88

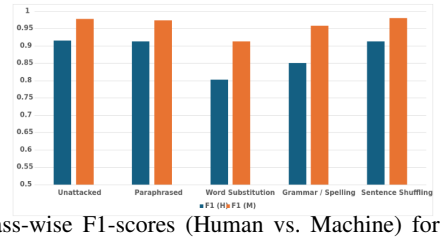

Fig. 3. Class-wise F1-scores (Human vs. Machine) for MoRE proposed detection model under five adversarial attack settings.

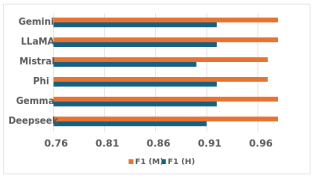

Fig. 4. Class-wise F1-scores of MoRE for binary detection (Human vs. each individual LLM) across six LLM sources.

Entropy [15] where all decision thresholds are determined using Youden's J-index for optimal class separation. In addition, we benchmark against a supervised based detector, OpenAI RoBERTa [24], and an adversarially robust method, RADAR [21]. Performance is assessed using **class-wise F1-scores**, reporting separate F1 values for human-written text and machine-generated (LLM) text to capture asymmetric detection difficulty across classes. In cases where a unified metric is required, we additionally employ the **Macro F1-score**, which provides an unbiased measure of balanced classification performance by equally weighting both classes.

TABLE II

CLASS-WISE F1 SCORES OF MORE TRAINED IN A 7-CLASS (MULTI-CLASS) SETTING, WHERE EACH LLM IS TREATED AS A SEPARATE CLASS INSTEAD OF PERFORMING BINARY HUMAN VS. MACHINE DETECTION.

	Phi	LLaMA	Human	Mistral	Gemma	DeepSeek	Gemini
F1	0.87	0.89	0.87	0.58	0.675	0.97	0.83

TABLE III

MACRO F1 SCORES OF OUR PROPOSED MORE AND BASELINE DETECTION METHODS UNDER FIVE ATTACK SETTINGS.

Method	Unattacked	Paraphrased	Substitution	Spelling	Shuffling
Log p	0.340	0.386	0.274	0.339	0.363
Rank	0.343	0.388	0.278	0.342	0.365
Log-Rank	0.290	0.302	0.168	0.202	0.235
Entropy	0.154	0.194	0.280	0.213	0.174
OpenAI	0.758	0.706	0.264	0.321	0.748
RADAR	0.888	0.876	0.478	0.892	0.892
MoRE	0.946	0.943	0.857	0.904	0.946

C. Stylometric Expert Features that Differentiate LLM-Generated and Human-Written Text

The Stylometric Expert is built upon a set of well-established, interpretable linguistic and statistical features known to reveal differences between human written and LLM-generated text. We incorporate three statistical indicators average number of sentences per essay capture surface-level structural variation. In addition, a rich suite of readability metrics is used to assess how easy or difficult a text is to comprehend. These include the Flesch–Kincaid (FK) Grade and the Automated Readability Index (ARI), both of which estimate the school grade level required to understand the text. Gunning Fog scores quantify the proportion of complex words. Lexical diversity is further captured through the Type–Token Ratio (TTR), and Language Model Fluency based Metrics perplexity and burstiness features, which collectively reflect vocabulary richness, word distribution irregularity, and fluency. Together, seven stylometric indicators provide a strong foundation for distinguishing human-written essays from those produced by large language models. Fig. 1 provides a visual comparison of four feature distributions such as Flesch–Kincaid Grade, Gunning Fog, Burstiness and, Type–Token Ratio revealing consistent stylistic differences between the human vs LLM.

D. Training Procedure

To address class imbalance, the human-written essays are oversampled during training to ensure equal representation of both classes. Each training batch contains both the original fused feature vector x and its adversarially perturbed counterpart x', enabling the model to jointly learn from clean and manipulated samples. 70:10:20 is the distribution of train, validation and test dataset. The model parameters are optimized using the Adam optimizer with a learning rate of 1×10^{-3}. Training is performed for 15 epochs with a batch size of 16. In the PGD adversarial augmentation loop, the perturbation is bounded by $\epsilon = 0.01$, with a step size of $\alpha = 0.002$, and is iteratively updated for three PGD steps. This setup ensures a controlled yet effective perturbation training.

V. RESULTS AND ANALYSIS

Binary classification: Table I presents a comprehensive comparison of MoRE with existing baseline detectors across our dataset and widely used eight existing datasets. The results clearly indicate that MoRE achieves the highest overall performance, particularly on our curated dataset, where it attains F1(H)=0.915 and F1(M)=0.977, outperforming all Zero-shot methods demonstrate inconsistent performance across domains, especially struggling with stylistically diverse or noisy datasets such as Kaggle Essay and HC3 Reddit ELI5 . While advanced detectors like OpenAI and RADAR achieve competitive results in more structured domains (e.g., HC3 Finance and HC3 Medicine). In contrast, MoRE maintains

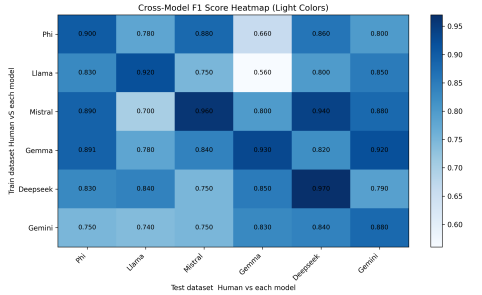

Fig. 5. Cross-model macro F1 heatmap. Rows represent the LLM used to train a Human vs. LLM binary classifier, and columns represent the LLM used as the test set. Darker colors indicate higher cross model generalization performance.

strong performance across heterogeneous datasets, confirming the effectiveness of expert and PGD adversarial training.

Multi Class Classification: Table II reports the class-wise F1-scores when MoRE is trained in a seven class setting, where each LLM source is treated as an independent class rather than performing binary Human vs. LLM detection. The results highlight MoRE's ability to discriminate between stylistic and semantic patterns unique to different LLM families. Notably, DeepSeek and LLaMA achieve the highest F1-scores, suggesting that MoRE can clearly capture their distinctive generative signatures. Mistral displays the lowest F1-score (0.58), indicating higher overlap with other model distributions. Overall, these results demonstrate MoRE's capability to extend beyond binary detection and effectively identify source-specific LLM writing styles in a multi-class.

How robust is MoRE under adversarial attacks? To evaluate the robustness of MoRE against adversarial manipulation, Table III and Fig. 3 present its performance across unattacked and five different attack settings. In Table III, MoRE consistently outperforms all baseline detectors in terms of macro F1 score, although performance decreases under more challenging perturbations such as word substitution and grammar or spelling corruption. although OpenAI and RADAR exhibit noticeable drops in performance. In contrast, MoRE retains macro F1 scores above 0.90 across all attacks except word substitution (0.857), demonstrating both resilience and strong generalization capability. The class-wise results in Fig. 3 further support this finding; MoRE preserves high separability between human-written and machine-generated text. These results highlight the strength of MoRE's, establishing it as a robust and attack-resilient detector that outperforms.

How consistently does MoRE detect text from different LLM families? To evaluate the performance of MoRE on the Human vs. LLM detection task, Table IV and Fig. 4 together provide a detailed assessment of MoRE's ability to distinguish human-written text from outputs generated by six different LLM families. The macro F1 scores in the table show that MoRE maintains consistently high performance across all sources. This narrow range reflects MoRE's ability to capture stable, generalizable writing signatures that are not tied to any specific model architecture. Fig. 4 reinforces these observations by presenting class-wise F1-scores for binary detection (Human vs. each LLM). MoRE achieves near-perfect

TABLE IV

MACRO F1 SCORES OF MORE AND BASELINE DETECTION METHODS ACROSS SIX LLM SOURCES. BINARY CLASSIFICATION TASK HUMAN VS EACH LLMS

Method	DeepSeek	Gemma	Phi	Mistral	LLaMA	Gemini
Log p	0.47	0.31	0.35	0.36	0.31	0.25
Rank	0.47	0.31	0.35	0.36	0.32	0.25
Log-Rank	0.47	0.14	0.35	0.33	0.27	0.18
Entropy	0.22	0.21	0.10	0.10	0.10	0.20
OpenAI	0.995	0.43	0.96	0.965	0.995	0.20
RADAR	0.895	0.895	0.89	0.885	0.89	0.87
MoRE	0.945	0.95	0.945	0.935	0.95	0.95

TABLE V

ABLATION STUDY ON THE MORE EVALUATING (I) ONLY STYLOMETRIC EXPERT, (II) ONLY SEMANTIC EXPERT, (III) ONLY CONTEXTUAL EXPERT, (IV) EXPERTS WITHOUT PGD, (V) EXPERTS WITH PGD, AND (VI) EXPERTS WITH PGD AND LOSS. BCE LOSS IS APPLIED IN ALL. A "✓" INDICATES COMPONENT IS ENABLED, WHILE "–" DISABLED.

Stylometric	Semantic	Contextual	PGD	Contrastive	BCE	MacroF1
✓	–	–	–	–	✓	0.786
–	✓	–	-	–	✓	0.801
–	–	✓	–	–	✓	0.815
✓	✓	✓	–	–	✓	0.868
✓	✓	✓	✓	–	✓	0.894
✓	✓	✓	–	✓	✓	0.912
✓	✓	✓	✓	✓	✓	0.960

machine-class F1 values, confirming its ability to reliably identify LLM-generated text regardless of the underlying LLM. Human-class F1-scores also remain consistently high (0.88–0.92), indicating that MoRE avoids overfitting toward machine detection and preserves balanced classification. The slight differences across LLMs reflect varying degrees of stylistic similarity to human writing; for example, Mistral and Gemini produce more human-like, narrowing separation margin.

How well does MoRE generalize to unseen LLM sources? Figure 5 presents a cross-model macro F1 analysis that examines how effectively a classifier trained on Human vs. a specific LLM transfers to other unseen LLM sources. Models trained on Mistral, Gemma, and DeepSeek demonstrate the strongest cross-model generalisation, achieving high macro F1 scores across most test LLMs. This suggests that these LLMs produce stylistic and semantic patterns diverse enough for MoRE to learn highly transferable representations. In contrast, classifiers trained on LLaMA and Gemini show reduced generalization, particularly when evaluated on structurally different LLMs such as Gemma or DeepSeek, indicating larger stylistic gaps among these systems. Overall, the result shows that while MoRE exhibits strong cross-model robustness generalization.

A. ABLATION STUDY

Table V presents the ablation study assessing the contribution of each component in the MoRE architecture. Using only a single expert module Stylometric, Semantic, or Contextual relatively modest performance (F1 scores between 0.786 and 0.815), indicating that no individual representation channel is sufficient for reliable detection. When all three experts are combined, the performance increases substantially to 0.868, demonstrating the complementary nature of stylistic, semantic, and contextual features in capturing diverse LLM-generation patterns. Adding PGD-based adversarial augmentation further improves the F1 score to 0.894, showing that robustness against perturbation-based attacks significantly strengthens classifier generalization. Introducing the supervised contrastive objective also yields a clear gain (F1 = 0.912), confirming that contrastive separation helps the model learn more discriminative representation clusters. The full MoRE configuration

with all experts, PGD augmentation, and contrastive learning achieves the best overall performance (F1 = 0.960), highlighting the importance of each component and demonstrating that the combination of multi-expert representations with adversarial and contrastive training provides the strongest and most robust detection capability.

VI. Conclusion

This study tackles the challenge of detecting both direct and adversarially manipulated LLM-generated text. We propose MoRE, a framework that fuses stylometric, semantic, and contextual experts with PGD-based adversarial augmentation and supervised contrastive learning to improve robustness. Using a curated dataset spanning six LLMs and four attack types, MoRE achieves state-of-the-art performance across eight benchmark datasets, demonstrating strong generalization to unseen models and resilience against diverse perturbations. Ablation studies highlight the complementary strengths of each expert and the key role of adversarial and contrastive objectives. However, MoRE may fail under advanced adaptive attacks (human-in-the-loop editing) and on newly emerging LLMs. Future work will improve robustness and expand the framework to other domains.

Acknowledgment

We acknowledge the use of AI-based grammar checking and rephrasing tools to improve the language and readability of this manuscript.

References

[1] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, et al. Llama: Open and efficient foundation language models. 2023.

[2] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Chaplot, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[3] Xiao Bi, Deli Chen, Guanting Chen, and Chenothers. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024.

[4] Joni Salminen, Chandrashekhar Kandpal, Ahmed Mohamed Kamel, Soon-gyo Jung, and Bernard J Jansen. Creating and detecting fake reviews of online products. *Journal of Retailing and Consumer Services*.

[5] Shifali Agrahari, Sujit Kumar, and Ranbir Singh Sanasam. Can you really trust that review? protofewroberta and detectairev: A prototypical few-shot method and multi-domain benchmark for detecting ai-generated reviews. In *Proceedings of the 14th IJCNN and the 4th AACL*.

[6] Zecong Wang, Jiaxi Cheng, Chen Cui, and Chenhao Yu. Implementing bert and fine-tuned roberta to detect ai generated news by chatgpt.

[7] Ryuto Koike, Masahiro Kaneko, and Naoaki Okazaki. Outfox: Llm-generated essay detection through in-context learning with adversarially generated examples. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21258–21266, 2024.

[8] Shifali Agrahari, Subhashi Jayant, Saurabh Kumar, and Sanasam Ranbir Singh. Essaydetect at genai detection task 2: Guardians of academic integrity: Multilingual detection of ai-generated essays.

[9] Shifali Agrahari and Sanasam Ranbir Singh. Osint at genai detection task 1: Multilingual mgt detection: Leveraging cross-lingual adaptation for robust llms text identification. In *Proceedings of the 1stWorkshop on GenAIDetect*.

[10] 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*.

[11] Liam Dugan, Alyssa Hwang, Filip Trhlik, Josh Magnus Ludan, et al. Raid: A shared benchmark for robust evaluation of machine-generated text detectors. *arXiv preprint arXiv:2405.07940*, 2024.

[12] A Hans, A Schwarzschild, V Cherepanova, H Kazemi, A Saha, et al. Spotting llms with binoculars: Zero-shot detection of machine-generated text, 2024. *URL: https://arxiv. org/abs/2401.12070*.

[13] Shixuan Ma and Quan Wang. Zero-shot detection of llm-generated text using token cohesiveness. *arXiv preprint arXiv:2409.16914*, 2024.

[14] 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, 2023.

[15] Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. Gltr: Statistical detection and visualization of generated text. *arXiv preprint arXiv:1906.04043*, 2019.

[16] 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.

[17] Amrita Bhattacharjee, Tharindu Kumarage, Raha Moraffah, and Huan Liu. Conda: Contrastive domain adaptation for ai-generated text detection. *arXiv preprint arXiv:2309.03992*.

[18] Pengyu Wang, Linyang Li, Ke Ren, Botian Jiang, Dong Zhang, and Xipeng Qiu. Seqxgpt: Sentence-level ai-generated text detection.

[19] Zhengyuan Jiang, Moyang Guo, Yuepeng Hu, and Neil Zhenqiang Gong. Watermark-based attribution of ai-generated content. *arXiv preprint arXiv:2404.04254*, 2024.

[20] Ahmed K Kadhim, Lei Jiao, Rishad Shafik, and Ole-Christoffer Granmo. Adversarial attacks on ai-generated text detection models: A token probability-based approach using embeddings.

[21] Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. Radar: Robust ai-text detection via adversarial learning. *Advances in neural information processing systems*, 36:15077–15095, 2023.

[22] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[23] Lorenz Mindner, Tim Schlippe, and Kristina Schaaff. Classification of human-and ai-generated texts: Investigating features for chatgpt. In *International Conference on Artificial Intelligence in Education Technology*, pages 152–170. Springer, 2023.

[24] 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. 2023.

[25] Shifali Agrahari, Samridhi Bisht, and Ranbir Singh Sanasam. Text authorship attribution: Stylometric insights into human and llm-generated text. In *Proceedings of the 12th ACM IKDD CODS and 30th COMAD*.

[26] Hetvi Waghela, Jaydip Sen, and Sneha Rakshit. Enhancing adversarial text attacks on bert models with projected gradient descent. *arXiv preprint arXiv:2407.21073*, 2024.

[27] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[28] Mojan Javaheripi, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Sebastien Bubeck, et al. Phi-2: The surprising power of small language models. *Microsoft Research Blog*, 1(3):3, 2023.

[29] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Pathak, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

[30] Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, and Wu others. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*, 2019.

[31] Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. Automatic detection of generated text is easiest when humans are fooled. *arXiv preprint arXiv:1911.00650*, 2019.

[32] Shifali Agrahari and Sanasam Ranbir Singh. Tracing thought: Using chain-of-thought reasoning to identify the llm behind ai-generated text.

[33] Shifali Agrahari, Hemanth Prakash Simhadri, and OTther. Osint at arageneval shared task: Fine-tuned modeling for tracking style signatures and ai generation in arabic texts. In *Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks*, pages 82–87.

[34] Jiwei Luo, Guofang Nan, Dahui Li, and Yong Tan. Ai-generated review detection. *Available at SSRN 4610727*, 2023.

[35] Shifali Agrahari, Prabhat Mishra, and Sujit Kumar. Random at genai detection task 3: A hybrid approach to cross-domain detection of machine-generated text with adversarial attack mitigation. In *Proceedings of the 1stWorkshop on GenAIDetect*, pages 365–370.

[36] Tharindu Kumarage, Joshua Garland, Amrita Bhattacharjee, Kirill Trapeznikov, Scott Ruston, and Huan Liu. Stylometric detection of ai-generated text in twitter timelines. *arXiv preprint arXiv:2303.03697*.

[37] Ting-Yun Chang and Yun-Nung Chen. What does this word mean? explaining contextualized embeddings with natural language definition.