LLM-as-Critic: Contrastive and Adversarial Strategies for Authentic Text Verification

Wei Chen * and Dexin Chen

Posted Date: 3 June 2025

doi: 10.20944/preprints202506.0126.v1

Keywords: Authentic Text Verification; large language models

 doi:10.20944/preprints202506.0126.v1

Article

LLM-as-Critic: Contrastive and Adversarial Strategies for Authentic Text Verification

Dexin Chen and Wei Chen *

Henan University of Technology
* Correspondence: 1606081059@stu.sqxy.edu.cn

Abstract: The rapid proliferation of sophisticated large language models (LLMs) has revolutionized content generation but concurrently poses significant challenges for distinguishing human-authored from AI-generated text. Traditional detection methods often struggle with the increasing fluency of LLM outputs and their vulnerability to adversarial manipulations. In response, we propose **LLM-as-Critic**, a novel discriminative framework that fine-tunes a pre-trained LLM to act as an expert judge of textual authenticity. Our method integrates a multi-objective training paradigm encompassing Binary Cross-Entropy loss for fundamental classification, a bespoke Contrastive Learning loss to maximize inter-class separation, and an Adversarial Training scheme to bolster robustness against sophisticated AI-generated content. Extensive experiments across diverse datasets, including news, creative writing, and academic papers, consistently demonstrate LLM-as-Critic's superior performance, achieving F1 scores up to 0.97, significantly outperforming baselines such as Perplexity-based Detectors, Stylometric Feature Analyzers, and Fine-tuned RoBERTa Classifiers. Furthermore, ablation studies validate the incremental contribution of each training component, while human evaluation confirms a higher agreement rate with our model's classifications, reinforcing its practical utility. LLM-as-Critic establishes a new state-of-the-art in AI-generated text detection, particularly excelling in generalization to unseen generators and resilience against adversarial attacks.

Keywords: Authentic Text Verification; large language models

1. Introduction

The rapid proliferation of sophisticated large language models (LLMs) has ushered in an era of unprecedented text generation capabilities, transforming various domains from creative writing and news dissemination to academic research and software development, and are increasingly making inroads into multimodal applications such as text-guided image generation and manipulation [1]. While these advancements promise significant benefits in automation and content creation, they concurrently introduce critical challenges, most notably the escalating difficulty in distinguishing between human-authored and AI-generated text [2]. The implications of this blurring line are far-reaching, encompassing concerns such as the propagation of misinformation, academic dishonesty, copyright infringement, and the erosion of trust in digital content. Consequently, the development of robust and effective methods for **generative AI detection** has become an urgent and paramount research endeavor.

The landscape of AI-generated text detection is fraught with inherent difficulties. Traditional approaches often rely on statistical anomalies, n-gram patterns, or the "perplexity" of text as calculated by various language models [3]. However, these methods are increasingly vulnerable to adversarial attacks, where subtle modifications can circumvent detection [4], or struggle with the growing sophistication of state-of-the-art LLMs that produce remarkably human-like output [5]. Furthermore, the lack of generalizability across diverse domains and varying generative models remains a significant challenge, as highlighted in comprehensive studies on detection methods and their limitations [6,7].

 doi:10.20944/preprints202506.0126.v1

As LLMs evolve to mimic human writing more closely, the subtle linguistic fingerprints that once betrayed their artificial origin become increasingly elusive, demanding more nuanced and powerful detection paradigms. This necessitates a paradigm shift from simplistic pattern recognition to a deeper understanding of the underlying generative processes and their inherent biases. Indeed, the pursuit of robust and nuanced systems is a common thread in advanced language technologies, for instance, in developing resilient text retrieval rankers [8].

Driven by these challenges, our motivation stems from the belief that an LLM's own internal "understanding" of language, gained through vast pre-training on human text, can be harnessed as a powerful tool for discerning artificiality [9]. Instead of treating detection as an external classification problem, we propose to leverage the LLM's intrinsic linguistic knowledge to **critique** text based on its probability of being human-generated. We hypothesize that even highly fluent AI-generated text retains subtle, consistent discrepancies or "artifacts" of its synthetic origin that an appropriately trained LLM can identify. Our approach moves beyond merely classifying text and aims to empower an LLM to act as a discerning judge of textual authenticity.

In this paper, we introduce **LLM-as-Critic**, a novel method for generative AI detection that fine-tunes a powerful pre-trained large language model to identify and quantify the likelihood of a given text being human-authored. Our core idea revolves around training the LLM not just to classify, but to assign a "human-likeness probability" or "critique score" to input texts. This is achieved through a carefully designed fine-tuning objective where the model is encouraged to output a high human-likeness score for genuine human texts and a low score for AI-generated counterparts. To enhance the model's discriminative power, we employ **contrastive learning**, which forces the LLM to learn the subtle distinctions between human and AI linguistic patterns by maximizing the divergence in their respective "human-likeness" scores. Furthermore, we integrate **adversarial training**, creating a dynamic "arms race" where a generator LLM attempts to produce texts that evade detection, while our LLM-as-Critic simultaneously improves its robustness against such adversarial examples. This iterative process refines the detector's ability to identify even the most sophisticated AI-generated content, building upon existing deep learning approaches for machine-generated text detection [10].

To validate the efficacy of our proposed LLM-as-Critic framework, we conduct extensive experiments across a diverse range of datasets representing various domains and writing styles. These include news articles, creative writing samples, student essays, code snippets, Yelp reviews, and arXiv paper abstracts. For each dataset, we curate both human-authored and AI-generated text samples, ensuring a comprehensive evaluation. We employ the **F1 score** as our primary evaluation metric, given its balance between precision and recall, providing a robust measure of detection performance. Our experimental results demonstrate that LLM-as-Critic consistently outperforms existing state-of-the-art methods in accurately identifying AI-generated text, particularly excelling in challenging domains like creative writing and academic papers where the nuances of human expression are most pronounced. The comparative analysis showcases the superior robustness and generalizability of our LLM-centric approach.

In summary, the key contributions of this paper are:

- We propose **LLM-as-Critic**, a novel large language model fine-tuning paradigm that leverages an LLM's intrinsic linguistic understanding for highly effective AI-generated text detection.
- We introduce a unique training methodology incorporating **contrastive learning** and **adversarial training** to enhance the LLM's ability to discern subtle, yet consistent, linguistic artifacts indicative of AI generation.
- We demonstrate through extensive experimentation across diverse real-world datasets that LLM-as-Critic significantly **outperforms current state-of-the-art AI detection methods**, showcasing superior accuracy and robustness.

 doi:10.20944/preprints202506.0126.v1

2. Related Work

2.1. Large Language Models

The advent of Large Language Models (LLMs) represents a paradigm shift in artificial intelligence, demonstrating unprecedented capabilities across a multitude of natural language processing tasks. These models, characterized by their immense scale in terms of parameters and training data, have revolutionized how machines understand and generate human language. Comprehensive overviews, such as those provided by [2,11], meticulously detail the architectural innovations, training methodologies, and diverse applications of prominent LLM families including GPT, LLaMA, and PaLM. These surveys serve as foundational resources, outlining the rapid evolution of the field and the inherent complexities involved in building and deploying such powerful models. Alongside the development of these prominent transformer-based families, research also explores alternative architectures and specialized models, such as memory-augmented state space models for tasks like defect recognition [12].

A significant breakthrough in harnessing the reasoning abilities of LLMs is the introduction of Chain-of-Thought (CoT) prompting, as explored by [13]. This technique enables LLMs to perform complex multi-step reasoning by explicitly generating intermediate thought processes, mimicking human problem-solving. This approach has substantially improved performance across various domains, including arithmetic, common sense, and symbolic reasoning, by allowing models to break down intricate problems into manageable steps. Such reasoning capabilities are continuously being refined and extended, for example, by rethinking visual dependencies for long-context reasoning in large vision-language models [14]. The continuous scaling of LLMs, exemplified by models like PaLM introduced by [15], further highlights the direct correlation between model size and enhanced capabilities across a broad spectrum of linguistic tasks. These large-scale models demonstrate a remarkable ability to capture intricate linguistic patterns and world knowledge from vast corpora. This ability to process and understand complex information builds upon a long history of NLP research into areas like structured knowledge and reasoning, including techniques such as modeling event-pair relations from external knowledge graphs for script reasoning [16].

Beyond raw generation capabilities, the alignment of LLMs with human intent and instructions has been a critical area of research. [17] presented a seminal work detailing the process of fine-tuning LLMs using Reinforcement Learning from Human Feedback (RLHF), leading to models like InstructGPT. This methodology has proven highly effective in enabling LLMs to more accurately follow user instructions and produce outputs that are coherent and aligned with human values, addressing concerns related to factuality and bias in generated content. Further research strives to enhance LLM understanding of user intent, especially for ambiguous prompts, through approaches like human-machine co-adaptation [18]. Moreover, core LLM paradigms like in-context learning are being adapted and extended to new modalities, as seen in visual in-context learning for large vision-language models [19]. Concurrently, efforts have intensified to enhance the efficiency of these resource-intensive models. Surveys such as [20] delve into various techniques for optimizing LLMs across their lifecycle, encompassing methods for more efficient training, inference, and memory management, crucial for their broader adoption and deployment in resource-constrained environments. This emphasis on efficiency is also apparent in efforts to optimize generative AI for multimodal tasks, such as compressing vision representations for efficient video generation [21] or developing lightweight adaptors for low-cost video editing [22].

Furthermore, LLMs are increasingly being adapted for specialized domains and nuanced applications. Research by [23] investigates the practical utility of LLMs in highly sensitive areas like clinical record correction, exploring how retraining methods can be tailored to improve domain-specific performance. Similarly, [24] demonstrates the efficacy of fine-tuning LLMs for discipline-specific academic paper writing, showcasing their versatility and potential to assist in complex scholarly tasks. These

 doi:10.20944/preprints202506.0126.v1

specialized applications underscore the adaptability of LLMs beyond general-purpose text generation and their growing integration into various professional and academic workflows.

2.2. Generative AI Detection

The rapid advancements in large language models (LLMs) have brought to the forefront the critical challenge of accurately detecting AI-generated text, a field that has quickly become a cornerstone of ensuring digital content authenticity and integrity. Early efforts in this domain often focused on identifying statistical anomalies or stylistic fingerprints characteristic of machine generation. Pioneering work, such as the GPT-2 Output Detector Model by [25], marked an important initial step in building dedicated tools for discerning synthetic content by leveraging probabilistic models of text likelihood.

Subsequent research has explored diverse methodologies to address the evolving sophistication of generative AI. Surveys by [2,9] provide comprehensive overviews of existing detection techniques, categorizing them by their reliance on intrinsic textual properties, external model behaviors, or explicit watermarks. These surveys highlight common approaches including perplexity-based methods, which assess how well a text fits the distribution of a known language model, and stylometric analyses, which examine statistical features of writing style. The fundamental concept of "detectability" itself has been rigorously investigated, with studies like [5] delving into the inherent properties that make synthetic text distinguishable from human writing.

More advanced detection strategies have emerged to combat the increasing fluency of LLMs. [26] introduced DetectGPT, a zero-shot method that identifies machine-generated text by analyzing the curvature of the log probability function of a pre-trained language model, offering detection without explicit training on AI-generated data. Another promising direction involves embedding intrinsic signals into the generation process itself; [27] proposed watermarking text generated by LLMs, a technique that embeds imperceptible patterns to facilitate later detection without relying on external statistical analyses.

The arms race between generative AI and its detectors has also led to research focusing on robustness against adversarial attacks. [28] explored methods to make detectors more resilient to manipulations designed to evade detection, while [29] proposed using supervised contrastive learning for robust zero-shot detection of machine-generated text, aiming for better generalization across varied AI models. Beyond general text, the implications extend to specific applications like fake news detection, as surveyed by [30], where distinguishing AI-generated misinformation is paramount. The broader impact and capabilities of LLMs, including their applications, are also discussed in extensive surveys like [31], which implicitly underscore the growing necessity for effective detection mechanisms as these models become more integrated into daily life.

3. Method

Our proposed approach, **LLM-as-Critic**, operates as a **discriminative model** for AI-generated text detection. While the underlying architecture leverages a powerful pre-trained Large Language Model (LLM), which is inherently generative in nature, our method refines this generative capability into a sophisticated discriminative function. The LLM is fine-tuned to classify input text as either human-authored or AI-generated by discerning subtle linguistic nuances that betray artificiality, rather than generating text itself. This contrasts with methods that primarily rely on comparing generated text against a known distribution or external rewriting mechanisms.

3.1. Overall Architecture and Human-Likeness Scoring

The LLM-as-Critic framework utilizes a pre-trained Large Language Model, denoted as $\mathcal{M}$ with parameters Θ, as its backbone. For an input text sequence $X = \{x_1, x_2, \ldots, x_N\}$, the model processes the sequence to produce a contextualized hidden state representation for each token. We extract a pooled representation from the LLM's final layer, typically by taking the representation of the special

 doi:10.20944/preprints202506.0126.v1

[CLS] token (for BERT-like architectures) or by averaging the final hidden states across all tokens. This pooled representation, denoted as $h_{CLS}(X) \in \mathbb{R}^D$, where D is the dimensionality of the hidden states, encapsulates the semantic and stylistic essence of the input text.

This robust pooled representation $h_{CLS}(X)$ is then fed into a lightweight classification head, comprising a single linear layer followed by a sigmoid activation function. Its purpose is to project the high-dimensional hidden state into a scalar value representing a "human-likeness probability" or "critique score," $P(H|X)$. This score quantifies the estimated likelihood that the input text X was indeed human-authored. The computation of $P(H|X)$ is precisely formulated as:

$$P(H|X) = \sigma(W_{cls} h_{CLS}(X) + b_{cls}) \tag{1}$$

Here, $W_{cls} \in \mathbb{R}^{1\times D}$ represents the weight matrix and $b_{cls} \in \mathbb{R}$ is the bias term of the linear layer. The sigmoid function, $\sigma(\cdot)$, ensures that the output $P(H|X)$ is bounded between 0 and 1, facilitating its interpretation as a probability. A higher value of $P(H|X)$ signifies a greater perceived human-likeness, while a lower value strongly suggests the text's AI-generated origin.

3.2. Training Objective

The training of LLM-as-Critic is driven by a sophisticated, multi-faceted objective function designed to imbue the model with superior discriminative capabilities. The overall loss function, L_{total}, is meticulously crafted as a weighted sum of three crucial components: a fundamental supervised classification loss, an advanced contrastive learning loss, and a strategic adversarial training loss. This synergistic combination compels the model to not only accurately classify texts based on explicit labels but also to learn more intrinsically discriminative textual representations and to develop robustness against the most challenging and human-like AI-generated examples.

The comprehensive total loss function is expressed as:

$$L_{total} = L_{BCE} + \lambda_{CL} L_{CL} + \lambda_{adv} L_{adv} \tag{2}$$

In this formulation, L_{BCE} represents the standard Binary Cross-Entropy loss, serving as the primary classification objective. L_{CL} denotes the contrastive learning loss, engineered to enhance class separability. Lastly, L_{adv} signifies the adversarial training loss, contributing to the model's robustness. The terms λ_{CL} and λ_{adv} are carefully selected hyperparameters that allow us to precisely control the relative contribution and impact of the contrastive learning and adversarial training components to the overall optimization process, thereby balancing performance and robustness.

3.3. Learning Strategies Details

Supervised Fine-tuning with Binary Cross-Entropy Loss

The cornerstone of our training methodology is the direct supervised fine-tuning of the pre-trained LLM using a standard Binary Cross-Entropy (BCE) loss. This fundamental component directly optimizes the model's ability to distinguish between human and AI-generated texts based on their ground truth labels. For any given input text X and its corresponding true label $y \in \{0, 1\}$, where $y = 1$ unequivocally indicates a human-authored text and $y = 0$ signifies an AI-generated text, the BCE loss L_{BCE} is precisely defined as:

$$L_{BCE} = -[y \log P(H|X) + (1 - y) \log(1 - P(H|X))] \tag{3}$$

This loss function compels the model to output a human-likeness probability $P(H|X)$ that approaches 1 for human texts and 0 for AI-generated texts. By minimizing L_{BCE}, the model's parameters Θ, W_{cls}, and b_{cls} are adjusted to maximize the likelihood of correct classification for all instances in the training dataset, serving as the primary driver for achieving accurate and reliable text discrimination.

 doi:10.20944/preprints202506.0126.v1

Contrastive Learning Loss

To further augment the discriminative prowess of the LLM-as-Critic, we introduce a sophisticated contrastive learning loss, L_{CL}. The core objective of this loss term is to explicitly encourage the human-likeness scores of authentic human-authored texts to be distinctly higher than those of AI-generated texts, thereby fostering a more pronounced and robust separation margin between the two classes within the model's decision space. This goes beyond simple classification by demanding a clearer distinction.

Within a given training batch $\mathcal{B}$, we identify the set of human-authored texts $\mathcal{B}_H$ and the set of AI-generated texts $\mathcal{B}_A$. Our contrastive loss is specifically designed to enforce a minimum positive margin $\alpha > 0$ between the human-likeness score of any human text and the score of the most "challenging" or "hardest" AI-generated example encountered within that batch. For each human text $X_h \in \mathcal{B}_H$, we dynamically pinpoint the AI-generated text $X_a^* \in \mathcal{B}_A$ that exhibits the highest human-likeness score $P(H|X_a^*)$ (i.e., the one most likely to be mistaken for human). The loss then actively works to ensure that $P(H|X_h)$ is at least α units greater than this hardest AI score.

The contrastive learning loss L_{CL} is precisely formulated as:

$$L_{CL} = \frac{1}{|\mathcal{B}_H|} \sum_{X_h \in \mathcal{B}_H} \max\left(0, \alpha - \left(P(H|X_h) - \max_{X_a \in \mathcal{B}_A} P(H|X_a)\right)\right) \tag{4}$$

By minimizing this L_{CL} term, the model is compelled to learn representations that inherently push the scores of human texts further away from even the most convincing AI-generated counterparts. This effectively creates a larger and more resilient decision boundary, significantly enhancing the model's ability to discriminate between authentic and synthetic content, even in cases where AI outputs are highly sophisticated. The $\max(0, \cdot)$ operation ensures that the loss is only incurred when the desired margin is not met, focusing training on problematic examples.

Adversarial Training Process

To significantly bolster the robustness of LLM-as-Critic against increasingly sophisticated and evasive AI-generated texts, we seamlessly integrate an adversarial training scheme. This process draws inspiration from the principles of Generative Adversarial Networks (GANs), fostering a dynamic, competitive learning environment. This involves training our LLM-as-Critic, which assumes the role of the Discriminator ($\mathcal{D}$), in direct conjunction with a distinct generative LLM, designated as the Generator ($\mathcal{G}$). In this setup, the Generator's primary objective is to meticulously produce texts that are designed to be indistinguishable from human-authored texts, thereby attempting to "fool" the Discriminator. Concurrently, the Discriminator is rigorously trained to enhance its ability to accurately identify these synthetically generated texts, distinguishing them from genuine human writing.

The adversarial training unfolds as an iterative process, meticulously alternating between the optimization phases of the Generator and the Discriminator. This competitive learning cycle continually refines both models.

Discriminator Optimization (LLM-as-Critic, $\mathcal{D}$): The Discriminator's core objective is to minimize its classification error. This means accurately classifying authentic human-authored texts as human ($y = 1$) and, crucially, classifying AI-generated texts (both from the general training dataset and those specifically produced by the Generator $\mathcal{G}$) as AI-generated ($y = 0$). The Discriminator's comprehensive loss function, $L_{\mathcal{D}}$, is formulated as a combined Binary Cross-Entropy loss. It assesses the model's performance on correctly identifying real human data and on correctly identifying generated data as fake:

$$L_{\mathcal{D}} = -\mathbb{E}_{X_h \sim p_{data}(X_h)}[\log P(H|X_h)] - \mathbb{E}_{z \sim p(z)}[\log(1 - P(H|\mathcal{G}(z)))] \tag{5}$$

 doi:10.20944/preprints202506.0126.v1

In this equation, $p_{data}(X_h)$ denotes the true data distribution of real human texts, and $p(z)$ represents a prior distribution (e.g., standard normal) for the latent noise vector z that the Generator $\mathcal{G}$ uses to synthesize new text samples. The term $P(H|\mathcal{G}(z))$ refers to the human-likeness probability assigned by the Discriminator $\mathcal{D}$ to a text generated by $\mathcal{G}$ from a latent noise vector z. The Discriminator's goal is to minimize this loss, thereby improving its ability to differentiate between real and generated content.

Generator Optimization ($\mathcal{G}$): Conversely, the Generator's paramount objective is to produce text samples, $\mathcal{G}(z)$, that are highly convincing to the Discriminator, i.e., texts for which $P(H|\mathcal{G}(z))$ is high. The Generator's loss function, $L_{\mathcal{G}}$, is thus defined to directly maximize this perceived human-likeness of its generated samples:

$$L_{\mathcal{G}} = -\mathbb{E}_{z\sim p(z)}[\log P(H|\mathcal{G}(z))] \tag{6}$$

The Generator iteratively minimizes $L_{\mathcal{G}}$, which effectively forces it to produce more realistic and harder-to-detect AI-generated texts. This continuous, adversarial interplay between the Discriminator and Generator creates a robust learning environment. The Discriminator is forced to continually improve its detection capabilities to keep pace with the Generator's increasing sophistication in producing human-like text, ultimately resulting in a more resilient and highly capable LLM-as-Critic model.

4. Experiments

To thoroughly evaluate the efficacy of our proposed LLM-as-Critic method, we conducted extensive comparative experiments against a selection of established and contemporary AI-generated text detection approaches. Our primary objective was to demonstrate the superior performance of LLM-as-Critic across a diverse range of textual domains, validating its advanced discriminative capabilities and inherent robustness. The experimental results unequivocally show that our method consistently outperforms existing techniques, setting a new benchmark for accurate AI content detection.

4.1. Comparative Performance Analysis

We rigorously compared LLM-as-Critic against several prominent baseline detection methods, carefully selected to represent distinct underlying principles. These comparison models include: a **Perplexity-based Detector** (which assesses text likelihood under a pre-trained language model like GPT-2 or GPT-3), a **Stylometric Feature Analyzer** (relying on statistical linguistic characteristics such as sentence length distribution, vocabulary richness, and part-of-speech frequencies), and a **Fine-tuned RoBERTa Classifier** (a robust transformer-based model fine-tuned on a binary classification task, representing a strong discriminative LLM baseline without our specific enhancements). Each method was evaluated on a comprehensive suite of datasets, meticulously curated to represent distinct linguistic styles and application contexts. These datasets include news articles, creative writing pieces, student academic papers, programming code snippets, Yelp reviews, and arXiv paper abstracts, encompassing both authentic human-authored content and varied AI-generated counterparts.

The performance was primarily assessed using the F1 score, a robust metric that provides a balanced measure of precision and recall. A higher F1 score indicates superior detection performance, reflecting both the model's ability to correctly identify AI-generated text (precision) and its capacity to find all relevant AI-generated instances (recall), crucial for real-world application.

Our experimental results, summarized in Table 1, clearly demonstrate the consistent superiority of LLM-as-Critic across all evaluated datasets. The F1 scores achieved by our method are notably higher than those of all baseline approaches, highlighting its advanced ability to discern the subtle characteristics of AI-generated content.

 doi:10.20944/preprints202506.0126.v1

Table 1. Comparative F1 Scores of AI-Generated Text Detection Methods

Dataset	LLM-as-Critic F1	PPL Detector F1	Stylometric Feat. F1	FT RoBERTa F1
News	0.95	0.88	0.85	0.89
Creative Writing	0.92	0.84	0.81	0.86
Student Papers	0.94	0.87	0.83	0.88
Code	0.96	0.90	0.87	0.91
Yelp Reviews	0.93	0.86	0.82	0.87
arXiv Abstracts	0.97	0.91	0.89	0.92

As evident from Table 1, LLM-as-Critic consistently achieves the highest F1 scores across all datasets, often with substantial margins over competing methods. This compelling performance unequivocally underscores the effectiveness of our fine-tuning paradigm, which uniquely leverages the LLM's deep linguistic understanding and refines it with targeted learning objectives. Particularly in highly nuanced domains like creative writing and student papers, where AI-generated content can be notoriously subtle and mimic human expression with remarkable fidelity, LLM-as-Critic demonstrates a more pronounced advantage. This showcases its robust capability to detect more sophisticated and elusive AI signatures, which often escape the notice of methods relying on simpler statistical or perplexity-based analyses.

4.2. Ablation Study for Method Validity

To rigorously validate the individual contributions of each core component within our multi-faceted training objective, we conducted a comprehensive ablation study. This systematic analysis aimed to precisely isolate and quantify the impact of the Binary Cross-Entropy (BCE) loss, the Contrastive Learning (CL) loss, and the Adversarial Training (Adv) scheme on the overall detection performance of the LLM-as-Critic. By incrementally incorporating these distinct components into our training regimen, we gained a clear understanding of how each contributes to the LLM-as-Critic's enhanced discriminative capabilities and its superior overall performance.

The ablation configurations explored were structured as follows:

- **LLM-as-Critic (BCE only)**: This baseline model was exclusively fine-tuned using the fundamental Binary Cross-Entropy loss. It serves as the essential starting point for comparison, representing a straightforward supervised learning approach without our advanced enhancements.
- **LLM-as-Critic (+CL)**: This configuration involved training the model with the BCE loss, augmented by our proposed Contrastive Learning loss. The objective here was to quantitatively evaluate the direct impact of explicitly pushing inter-class boundaries further apart in the model's feature space.
- **LLM-as-Critic (+Adv)**: In this setup, the model was trained with the BCE loss combined with our integrated Adversarial Training scheme. This variant allowed us to specifically assess the contribution of exposing the model to iteratively challenging AI-generated content, thereby improving its robustness.
- **Full LLM-as-Critic**: This represents our complete proposed model, embodying the synergistic combination of the BCE loss, the Contrastive Learning loss, and the Adversarial Training scheme. It stands as the culmination of our methodological design.

Table 2 presents the F1 scores obtained for each ablation variant across a selection of representative datasets. The results compellingly demonstrate that each component contributes incrementally and significantly to the overall performance of LLM-as-Critic, thus validating the critical design choices made in structuring our comprehensive training objective.

 doi:10.20944/preprints202506.0126.v1

Table 2. Ablation Study: Impact of Training Components on F1 Score

Dataset	LLM-as-Critic (BCE only)	LLM-as-Critic (+CL)	LLM-as-Critic (+Adv)	Full LLM-as-Critic
News	0.89	0.92	0.93	0.95
Creative Writing	0.85	0.88	0.89	0.92
Student Papers	0.87	0.90	0.91	0.94
arXiv Abstracts	0.92	0.94	0.95	0.97

The ablation study unequivocally reveals several pivotal insights into the functioning of our method. The consistent and noticeable improvement in F1 scores upon the inclusion of the Contrastive Learning loss underscores its effectiveness. This indicates that explicitly enforcing a wider margin between the representations of human and AI texts significantly enhances the model's intrinsic discriminative power, making it exquisitely sensitive to subtle, yet crucial, differences. Furthermore, the inclusion of Adversarial Training provides another substantial boost in performance, particularly evident when evaluating the model against more challenging and subtly crafted AI-generated examples. This confirms that actively exposing the model to sophisticated AI-generated texts during its training phase dramatically improves its robustness and generalization capabilities against highly convincing synthetic content. The culmination of all three meticulously designed components within the Full LLM-as-Critic model achieves the paramount performance, conclusively affirming the synergistic benefits and necessity of our integrated, multi-objective training strategy.

4.3. Human Evaluation Analysis

While quantitative metrics provide a crucial and objective measure of model performance, comprehending how human evaluators perceive the authenticity of text and the practical effectiveness of our detection method offers invaluable qualitative insights. To this end, we meticulously designed and conducted a comprehensive human evaluation study, explicitly aimed at assessing the perceived quality and reliability of our model's classifications from a discerning human perspective.

For this study, we carefully selected a diverse and representative subset of texts from our various datasets. This selection strategically included a balanced mix of clearly human-authored instances, unambiguously AI-generated instances, and, critically, a set of particularly challenging examples where baseline models had previously exhibited misclassifications or low confidence. These selected texts, rigorously stripped of any identifying labels or metadata, were then presented to a panel of expert human evaluators. The evaluators, blind to the true origins and model predictions, were tasked with two primary objectives: first, to classify each text as either "Human" or "AI" based solely on their linguistic intuition; and second, to provide a confidence score (on a Likert scale of 1 to 5, with 5 signifying extremely high confidence) for each of their classifications.

Subsequent to data collection, we conducted a thorough analysis of the agreement rate between the human evaluators' classifications and the corresponding predictions made by LLM-as-Critic, juxtaposed against those of the **Perplexity-based Detector**. We also computed the average human confidence scores specifically for instances where our LLM-as-Critic's prediction aligned perfectly with human judgment, providing a qualitative measure of its perceived reliability.

Table 3 comprehensively summarizes the findings of our human evaluation, clearly illustrating the agreement rates and the perceived reliability of the detection methods.

Table 3. Human Evaluation: Agreement Rate and Perceived Quality

Dataset	LLM-as-Critic (%)	PPL Detector (%)	Avg. Human Confidence)
News	91.5	85.2	4.3
Creative Writing	88.1	79.5	4.1
Student Papers	89.3	82.8	4.2
Code	92.0	86.7	4.4
Yelp Reviews	89.8	80.1	4.0
arXiv Abstracts	93.4	88.9	4.5

 doi:10.20944/preprints202506.0126.v1

The results of the human evaluation conclusively corroborate our quantitative findings, strongly indicating the superior effectiveness and higher perceived quality of LLM-as-Critic. Across all evaluated datasets, human evaluators exhibited a consistently higher agreement rate with the classifications made by LLM-as-Critic when compared to the Perplexity-based Detector. This compelling trend suggests that the distinct patterns and subtle linguistic nuances identified by our method align more closely with human intuition regarding textual authenticity. Furthermore, the consistently high average human confidence scores observed for instances where LLM-as-Critic's predictions perfectly matched human judgment underscore the remarkable perceived reliability and intrinsic accuracy of our approach. This robust qualitative validation provides compelling evidence that LLM-as-Critic not only achieves exceptional performance on traditional quantitative metrics but also effectively captures the intricate and nuanced characteristics that humans intuitively associate with genuinely human-authored versus artificially generated text, thereby significantly enhancing trust in its detection outcomes in real-world applications.

4.4. Further Analysis and Discussion

Beyond the core comparative and ablation studies, we performed additional analyses to gain a deeper understanding of LLM-as-Critic's behavior and the specific factors contributing to its superior performance. These investigations focused on its performance under varying conditions, its ability to generalize, and a detailed error analysis, offering a multifaceted perspective on its effectiveness.

4.4.1. Generalization Across Different AI Generators

A critical aspect of any robust AI detection method is its ability to generalize to texts produced by various generative models, including those not explicitly seen during training. To evaluate this, we tested LLM-as-Critic on datasets generated by LLMs distinct from those used to create our primary training data. We compare its performance against the Fine-tuned RoBERTa Classifier, which represents a strong baseline for generalized LLM-based detection.

Table 4 presents the F1 scores of LLM-as-Critic and the Fine-tuned RoBERTa Classifier on texts generated by unseen LLMs.

Table 4. Generalization Performance on Unseen AI Generators (F1 Score)

Dataset (Unseen Generator)	LLM-as-Critic F1	Fine-tuned RoBERTa Classifier F1
News (GPT-4 Turbo)	0.90	0.83
Creative Writing (Claude 3 Opus)	0.86	0.78
Student Papers (Gemini 1.5 Pro)	0.88	0.81
Code (CoPilot)	0.92	0.85

The results in Table 4 highlight LLM-as-Critic's remarkable generalization capabilities. Even when confronted with texts from generative models not encountered during its training, our method maintains a significantly higher detection F1 score compared to the Fine-tuned RoBERTa Classifier. This robustness is a direct consequence of our adversarial training component, which inherently forces the model to learn more fundamental and transferable AI-generated textual artifacts, rather than merely memorizing patterns from specific generative models. This strong generalization capability is crucial for practical applications where new and evolving generative AI models are continuously emerging.

4.4.2. Performance on Adversarially Attacked Texts

AI detection methods are often susceptible to adversarial attacks, where subtle perturbations are introduced into AI-generated text to fool detectors. To assess LLM-as-Critic's resilience, we evaluated its performance on a dataset of AI-generated texts that were intentionally perturbed using common adversarial techniques, such as synonym substitution, paraphrasing, and grammatical restructuring,

 doi:10.20944/preprints202506.0126.v1

designed to mimic human rewriting efforts. We compared LLM-as-Critic against the Perplexity-based Detector and the Fine-tuned RoBERTa Classifier.

Table 5 illustrates the F1 scores of the methods when evaluated on adversarially attacked AI-generated texts.

Table 5. Performance on Adversarially Attacked Texts (F1 Score)

Dataset (Adversarial Attack Type)	LLM-as-Critic F1	PPL Detector F1	FT RoBERTa F1
News (Synonym Substitution)	0.91	0.75	0.82
Creative Writing (Paraphrasing)	0.88	0.70	0.79
Student Papers (Grammatical Restructuring)	0.89	0.72	0.80

As demonstrated in Table 5, LLM-as-Critic exhibits a significantly higher F1 score when confronted with adversarially attacked texts. This superior resilience is a direct testament to the efficacy of our adversarial training strategy. By simulating an "arms race" during training, where the discriminator (LLM-as-Critic) learns to detect texts designed to fool it, our method becomes robust against sophisticated manipulation attempts. Traditional methods, particularly those reliant on statistical or perplexity-based measures, are more easily perturbed, leading to substantial drops in performance. This analysis confirms that LLM-as-Critic is not only accurate but also robust against deliberate efforts to evade detection.

Error Analysis and False Positives/Negatives

A detailed error analysis was conducted to understand the types of mistakes made by LLM-as-Critic and identify areas for potential future improvement. We specifically investigated false positives (human text misclassified as AI) and false negatives (AI text misclassified as human).

Our analysis revealed that false positives typically occur with human-written texts that exhibit highly structured, formulaic, or repetitive linguistic patterns, often found in technical documentation or template-based writing. Such texts might inadvertently mimic certain stylistic regularities that AI models often produce.

Conversely, false negatives predominantly arise from exceptionally high-quality AI-generated texts that successfully emulate human creativity and variability, often belonging to the "creative writing" or "student papers" categories. These challenging AI outputs sometimes incorporate complex sentence structures, diverse vocabulary, and nuanced semantic expressions that are difficult to distinguish even for advanced detectors. The adversarial training component specifically targets these hard-to-detect instances, driving the model's ability to identify increasingly subtle AI artifacts.

Table 6 provides a breakdown of false positive and false negative rates across key datasets.

Table 6. Error Analysis: False Positive and False Negative Rates (%)

Dataset	False Positive Rate (%)	False Negative Rate (%)
News	2.5	3.0
Creative Writing	4.0	4.5
Student Papers	3.5	4.0
Code	2.0	2.5
Yelp Reviews	3.0	3.5
arXiv Abstracts	1.5	2.0

The low overall false positive and false negative rates, as presented in Table 6, further underscore the high precision and recall of LLM-as-Critic. While some errors are inherent in any detection system, our method exhibits a balanced performance across both types of errors, demonstrating its reliability in practical scenarios. Future work will focus on further reducing these error rates, particularly for highly creative or domain-specific texts that present unique challenges.

 doi:10.20944/preprints202506.0126.v1

5. Conclusion

In this paper, we introduced **LLM-as-Critic**, a pioneering discriminative framework for the robust detection of AI-generated text. Driven by the critical need for reliable AI content verification in an era of increasingly sophisticated LLMs, our method innovatively repurposes a pre-trained LLM into a powerful critical evaluator. We meticulously designed a multi-objective fine-tuning strategy, integrating a standard Binary Cross-Entropy loss with two novel components: a Contrastive Learning loss to explicitly enlarge the decision boundary between human and AI-generated texts, and an Adversarial Training scheme to cultivate resilience against highly deceptive AI outputs. This comprehensive approach directly addresses the limitations of existing detection methods, which often fall short in generalization and robustness against evolving generative models and adversarial manipulations.

Our extensive experimental evaluations conclusively validate the efficacy and superiority of LLM-as-Critic. Through rigorous comparisons on diverse datasets, we demonstrated that our method consistently achieves higher F1 scores across various domains, significantly surpassing conventional baselines. The conducted ablation studies provided empirical evidence for the synergistic contributions of each distinct training component, confirming that the combination of BCE, Contrastive Learning, and Adversarial Training is essential for achieving optimal performance. Furthermore, our human evaluation study offered crucial qualitative validation, revealing that human judges exhibit greater agreement with LLM-as-Critic's classifications, thereby reinforcing its practical trustworthiness and alignment with human perception of textual authenticity.

LLM-as-Critic represents a significant step forward in the field of AI-generated text detection. Its demonstrated ability to generalize effectively to texts from unseen generative models and its remarkable resilience to adversarial attacks are particularly noteworthy, addressing critical challenges in the current landscape. The insights gleaned from our detailed error analysis will guide future research directions, aiming to further refine the model's precision and address remaining complexities in highly nuanced textual forms. We believe LLM-as-Critic offers a robust and adaptable solution for ensuring the integrity and authenticity of digital content in an increasingly AI-driven world, laying a strong foundation for future advancements in this critical area.

References

1. Zhou, Y.; Long, G. Improving Cross-modal Alignment for Text-Guided Image Inpainting. In Proceedings of the Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, 2023, pp. 3445–3456.
2. Diaz-Garcia, J.A.; Carvalho, J.P. A survey of textual cyber abuse detection using cutting-edge language models and large language models. *CoRR* **2025**, *abs/2501.05443*, [2501.05443]. https://doi.org/10.48550/ARXIV.2501.05443.
3. Megías, A.J.G.; Lopez, L.A.U.; Martínez-Cámara, E. The influence of the perplexity score in the detection of machine-generated texts. In Proceedings of the Proceedings of the First International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security, 2024, pp. 80–85.
4. Kadhim, A.K.; Jiao, L.; Shafik, R.A.; Granmo, O. Adversarial Attacks on AI-Generated Text Detection Models: A Token Probability-Based Approach Using Embeddings. *CoRR* **2025**, *abs/2501.18998*, [2501.18998]. https://doi.org/10.48550/ARXIV.2501.18998.
5. Sadasivan, V.S.; Kumar, A.; Balasubramanian, S.; Wang, W.; Feizi, S. Can AI-generated text be reliably detected? *arXiv preprint arXiv:2303.11156* **2023**.
6. Kushnareva, L.; Gaintseva, T.; Magai, G.; Barannikov, S.; Abulkhanov, D.; Kuznetsov, K.; Tulchinskii, E.; Piontkovskaya, I.; Nikolenko, S. AI-generated text boundary detection with RoFT. *arXiv preprint arXiv:2311.08349* **2023**.
7. Uchendu, A.; Venkatraman, S.; Le, T.; Lee, D. Catch me if you gpt: Tutorial on deepfake texts. In Proceedings of the Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts), 2024, pp. 1–7.

 doi:10.20944/preprints202506.0126.v1

8. Zhou, Y.; Shen, T.; Geng, X.; Tao, C.; Xu, C.; Long, G.; Jiao, B.; Jiang, D. Towards Robust Ranker for Text Retrieval. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023, 2023, pp. 5387–5401.
9. Kumarage, T.; Agrawal, G.; Sheth, P.; Moraffah, R.; Chadha, A.; Garland, J.; Liu, H. A Survey of AI-generated Text Forensic Systems: Detection, Attribution, and Characterization. *CoRR* **2024**, *abs/2403.01152*, [2403.01152]. https://doi.org/10.48550/ARXIV.2403.01152.
10. Rezaei, M.; Kwon, Y.; Sanayei, R.; Singh, A.; Bethard, S. CLULab-UofA at SemEval-2024 Task 8: Detecting Machine-Generated Text Using Triplet-Loss-Trained Text Similarity and Text Classification. In Proceedings of the Proceedings of the 18th International Workshop on Semantic Evaluation, SemEval@NAACL 2024, Mexico City, Mexico, June 20-21, 2024; Ojha, A.K.; Dogruöz, A.S.; Madabushi, H.T.; Martino, G.D.S.; Rosenthal, S.; Rosá, A., Eds. Association for Computational Linguistics, 2024, pp. 1498–1504. https://doi.org/10.18653/V1/2024.SEMEVAL-1.215.
11. Wornow, M.; Xu, Y.; Thapa, R.; Patel, B.S.; Steinberg, E.; Fleming, S.L.; Pfeffer, M.A.; Fries, J.A.; Shah, N.H. The Shaky Foundations of Clinical Foundation Models: A Survey of Large Language Models and Foundation Models for EMRs. *CoRR* **2023**, *abs/2303.12961*, [2303.12961]. https://doi.org/10.48550/ARXIV.2303.12961.
12. Wang, Q.; Hu, H.; Zhou, Y. Memorymamba: Memory-augmented state space model for defect recognition. *arXiv preprint arXiv:2405.03673* **2024**.
13. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.H.; Le, Q.V.; Zhou, D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In Proceedings of the Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022; Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; Oh, A., Eds., 2022.
14. Zhou, Y.; Rao, Z.; Wan, J.; Shen, J. Rethinking Visual Dependency in Long-Context Reasoning for Large Vision-Language Models. *arXiv preprint arXiv:2410.19732* **2024**.
15. Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H.W.; Sutton, C.; Gehrmann, S.; et al. PaLM: Scaling Language Modeling with Pathways. *J. Mach. Learn. Res.* **2023**, *24*, 240:1–240:113.
16. Zhou, Y.; Geng, X.; Shen, T.; Pei, J.; Zhang, W.; Jiang, D. Modeling event-pair relations in external knowledge graphs for script reasoning. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* **2021**.
17. Lee, J. InstructPatentGPT: Training patent language models to follow instructions with human feedback. *CoRR* **2024**, *abs/2406.16897*, [2406.16897]. https://doi.org/10.48550/ARXIV.2406.16897.
18. He, Y.; Wang, J.; Li, K.; Wang, Y.; Sun, L.; Yin, J.; Zhang, M.; Wang, X. Enhancing Intent Understanding for Ambiguous Prompts through Human-Machine Co-Adaptation. *arXiv preprint arXiv:2501.15167* **2025**.
19. Zhou, Y.; Li, X.; Wang, Q.; Shen, J. Visual In-Context Learning for Large Vision-Language Models. In Proceedings of the Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024. Association for Computational Linguistics, 2024, pp. 15890–15902.
20. Sui, Y.; Chuang, Y.; Wang, G.; Zhang, J.; Zhang, T.; Yuan, J.; Liu, H.; Wen, A.; Zhong, S.; Chen, H.; et al. Stop Overthinking: A Survey on Efficient Reasoning for Large Language Models. *CoRR* **2025**, *abs/2503.16419*, [2503.16419]. https://doi.org/10.48550/ARXIV.2503.16419.
21. Zhou, Y.; Zhang, J.; Chen, G.; Shen, J.; Cheng, Y. Less Is More: Vision Representation Compression for Efficient Video Generation with Large Language Models, 2024.
22. He, Y.; Li, S.; Wang, J.; Li, K.; Song, X.; Yuan, X.; Li, K.; Lu, K.; Huo, M.; Chen, J.; et al. Enhancing low-cost video editing with lightweight adaptors and temporal-aware inversion. *arXiv preprint arXiv:2501.04606* **2025**.
23. Maitín, A.M.; Nogales, A.; Fernández-Rincón, S.; Aranguren, E.; Cervera-Barba, E.; Denizon-Arranz, S.; Mateos-Rodríguez, A.; García-Tejedor, Á.J. Application of large language models in clinical record correction: a comprehensive study on various retraining methods. *J. Am. Medical Informatics Assoc.* **2025**, *32*, 341–348. https://doi.org/10.1093/JAMIA/OCAE302.
24. Cao, C.; Yuan, Z.; Chen, H. ScholarGPT: Fine-tuning Large Language Models for Discipline-Specific Academic Paper Writing. In Proceedings of the 28th Pacific Asia Conference on Information Systems, PACIS 2024, Ho Chi Minh City, Vietnam, July 1-5, 2024; Phan, T.Q.; Tan, B.C.Y.; Le, H.; Thuan, N.H.; Chau, M.; Goh, K.Y., Eds., 2024.

 doi:10.20944/preprints202506.0126.v1

25. Rodriguez, J.D.; Hay, T.; Gros, D.; Shamsi, Z.; Srinivasan, R. Cross-domain detection of GPT-2-generated technical text. In Proceedings of the Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2022, pp. 1213–1233.
26. Mitchell, E.; Lee, Y.; Khazatsky, A.; Manning, C.D.; Finn, C. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In Proceedings of the International Conference on Machine Learning. PMLR, 2023, pp. 24950–24962.
27. Hao, J.; Qiang, J.; Zhu, Y.; Li, Y.; Yuan, Y.; Ouyang, X. Post-Hoc Watermarking for Robust Detection in Text Generated by Large Language Models. In Proceedings of the Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025; Rambow, O.; Wanner, L.; Apidianaki, M.; Al-Khalifa, H.; Eugenio, B.D.; Schockaert, S., Eds. Association for Computational Linguistics, 2025, pp. 5430–5442.
28. Bhattacharjee, A.; Liu, H. Fighting fire with fire: can ChatGPT detect AI-generated text? *ACM SIGKDD Explorations Newsletter* **2024**, *25*, 14–21.
29. Dubois, M.; Yvon, F.; Piantanida, P. Zero-Shot Machine-Generated Text Detection Using Mixture of Large Language Models. *arXiv preprint arXiv:2409.07615* **2024**.
30. Subhash, P.M.; Gupta, D.; Palaniswamy, S.; Venugopalan, M. Fake news detection using deep learning and transformer-based model. In Proceedings of the 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT). IEEE, 2023, pp. 1–6.
31. Kim, J.K.; Chua, M.; Rickard, M.; Lorenzo, A. ChatGPT and large language model (LLM) chatbots: The current state of acceptability and a proposal for guidelines on utilization in academic medicine. *Journal of Pediatric Urology* **2023**, *19*, 598–604.