Journal of Artificial Intelligence and Technology, (Ahead of Print)
https://doi.org/10.37965/jait.2025.0822

ISTP
RESEARCH ARTICLE

DETECTRIX: A Novel Deep Learning Framework for High-Accuracy Identification of AI-Generated Content Across Diverse Textual Domains

Ghada Y. Elwan, Doaa R. Fathy, Nahed M. El Desouky, and Abeer S. Desuky
Mathematics Department Faculty of Science, Al-Azhar University (Girls Branch), Cairo, Egypt

(Received 16 May 2025; Revised 13 August 2025; Accepted 02 September 2025; Published online 24 September 2025)

Abstract: The development of large language models (LLMs) has made the generation of AI text nearly replicating human writing available to the public. This poses severe problems for academic honesty, the verification of information, and the authentication of documents. In this paper, we present a novel approach based on deep learning to tackle the problem of human vs. AI text detection. We have developed DETECTRIX, a hybrid transformer-based framework that combines optimized preprocessing with domain-adaptive training methodologies. Our approach has analyzed textual context, linguistic features, and statistical writing patterns to distinguish between human-authored and AI-generated content with high precision. Evaluation of a large dataset of academic writings, news articles, and creative writing pieces demonstrates that our model outperforms existing methods, achieving an F1-score of 97.8%. We also examine the enduring shortcomings of current detection approaches and identify directions for further investigations, considering evolving generative AI capabilities. This work contributes to maintaining authenticity in the face of sophisticated text generation tools.

Keywords: AI detection; AI-generated content; deep learning; natural language processing; text classification; transformers

I. INTRODUCTION

A. RESEARCH CONTEXT AND PROBLEM STATEMENT

The development of large language models (LLMs) has fundamentally transformed natural language processing across various industries, with current models such as GPT-4 [1] and LLaMA demonstrating text generation capabilities that approach human-level coherence and contextual relevance. Recent models have achieved perplexity scores within human benchmarks [2], while blind evaluation studies indicate that participants can identify AI-authored content only slightly above chance level (55.6%) [3]. This remarkable progress presents unprecedented opportunities alongside critical challenges for content authenticity and academic integrity.

In educational contexts, institutional surveys reveal a 79% increase in suspected AI-generated submissions between 2022 and 2024, while educators express only 12% confidence in accurately distinguishing such content [4]. The implications extend beyond education to encompass scientific publishing, journalism, legal documentation, and online information systems where content authenticity is fundamental to institutional credibility and decision-making processes. Security concerns have escalated with sophisticated AI-generated phishing attempts having surged by 186% between Q1 2023 and Q1 2024 [5], underscoring the urgent need for robust detection capabilities.

Despite this critical need, current detection approaches face severe limitations that hinder practical deployment. Statistical anomaly detection methods show dramatic accuracy degradation from 89.7% against GPT-3 outputs to 62.3% against GPT-4 outputs, while stylometric analysis approaches demonstrate poor cross-domain generalization, achieving only 58.2% accuracy when models trained on academic writing are evaluated on creative text samples [6]. Additionally, simple text modifications such as paraphrasing render most detection systems ineffective, with performance drops of 15–25% observed across different approaches [7]. Perhaps most concerning, existing systems demonstrate systematic bias against non-native English writers, who experience error rates 1.8–2.2× higher than baseline performance [8], raising serious ethical concerns about equitable application.

Current approaches can be categorized into three primary methodologies, each with particular disadvantages. Statistical anomaly detection approaches alongside entropy analysis methods attempt to identify statistically improbable segments in the distribution of produced text but show consistent performance drops from 89.7% against GPT-3 outputs to 62.3% against GPT-4 outputs using identical methodologies [9]. Stylometric analysis approaches analyze distributional measures of syntactic constituents and discourse markers [10] but lack adaptability and demonstrate poor generalization abilities when confronted with newer models that adjust stylistic elements to mimic human writing patterns. Watermarking methods [8] incorporate detection patterns within text generation but are bound to closed commercial systems requiring direct access to model parameters, and standard text modification techniques can easily remove these markers, sharply diminishing utility in real-world scenarios [11].

B. RESEARCH CONTRIBUTIONS AND INNOVATION

Current AI text detection approaches suffer from fundamental limitations that constrain their practical applicability. Statistical

Corresponding author: Ghada Y. Elwan (e-mail: ghadaelwan.el20@azhar.edu.eg).

anomaly detection methods demonstrate significant performance degradation when confronted with newer generation models, with accuracy declining from 89.7% against GPT-3 outputs to 62.3% against GPT-4 outputs using identical methodologies. Stylometric analysis approaches, while effective within specific domains, achieve only 58.2% accuracy when evaluated across different textual genres [6]. Watermarking techniques require direct access to model parameters, limiting their applicability to closed commercial systems, and remain vulnerable to standard text modification procedures [12].

This research addresses these limitations through a comprehensive framework that advances the state-of-the-art in several key aspects. We have introduced a novel hybrid architecture that systematically integrates transformer-based contextual encoding with convolutional pattern detection and bidirectional Long Short-Term Memory (LSTM) sequence modeling. Unlike existing single-component approaches such as DetectGPT's probability curvature analysis [10] or Giant Language Model Test Room (GLTR)'s statistical anomaly detection [9], our multi-level architecture captures complementary linguistic signals across different abstraction levels, achieving 97.8% F1-score compared to DetectGPT's 91.5% while maintaining consistent cross-domain performance.

The framework incorporates a feature-preserving preprocessing methodology that addresses a critical weakness in current detection systems. While conventional preprocessing pipelines apply standard normalization that eliminates subtle distinguishing characteristics, our approach maintains typographic inconsistencies, punctuation irregularities, and structural patterns that exhibit differential distributions between human and AI-generated content. This methodology directly addresses the 15–25% performance degradation observed across existing detection systems when confronted with advanced language models [10].

We have presented the first comprehensive empirical evaluation across 500,000 samples spanning academic writings, news articles, and social media content, substantially expanding beyond the typical <50,000 samples in single domains characteristic of previous studies [13]. This extensive evaluation reveals critical domain-specific performance patterns and identifies mixed human–AI collaborative content as presenting the most significant detection challenge, with a 14.4% performance degradation that has important implications for real-world deployment scenarios.

Furthermore, this work provides a systematic assessment of practical deployment considerations typically absent from laboratory-focused studies. We examine computational requirements, processing latency, bias concerns affecting non-native English speakers, and legal implications of automated content classification. This analysis contributes to bridging the gap between research achievements and institutional implementation while honestly acknowledging the fundamental challenges facing detection-based approaches as human–AI collaboration becomes increasingly prevalent.

The significance of these contributions extends beyond incremental performance improvements to address fundamental gaps in current detection paradigms. The hybrid architectural approach establishes a new methodological framework for multi-level linguistic analysis, while the comprehensive evaluation reveals both the potential and limitations of detection-based approaches in evolving AI landscapes.

C. RESEARCH METHODOLOGY OVERVIEW

This study employs a systematic experimental design centered on a comparative analysis of five transformer architectures (BERT [14], RoBERTa [15], XLNet [16], ALBERT [17], and DistilBERT [18]) to identify optimal configurations for AI text detection across diverse domains.

The challenge involves merging several technical methods to address gaps in current detection systems. First, we approach the problem using a hybrid model incorporating transformer contextual encoding and feature extraction.

Second, we apply an optimized data preprocessing pipeline with specialized methods focused on maintaining weak signals, distinguishing human and AI-generated text. This approach preserves inconsistencies in typographic cases, spacing patterns, and layout elements that may possess distinguishing features while filtering noise-adding elements [19]. Our preprocessing framework retains orthographic differences, punctuation patterns, and general layout that constitute useful distinguishing factors.

Third, we execute a progressive training strategy with domain adaptation techniques comprising gradient accumulation for training stabilization, curriculum learning for increasing sample complexity, and domain adversarial training for aiding cross-domain generalization [20].

Fourth, we design a comprehensive evaluation framework that measures performance across various dimensions, focusing on text categories, length distribution, language complexity, and model generation type. This approach captures the model's utility better than restricted evaluation methodologies documented in the literature, enabling a thorough assessment of practical applicability.

D. PAPER ORGANIZATION

The remainder of this paper is structured to provide comprehensive coverage of methodology, results, and implications. Section II offers an in-depth analysis of relevant literature, focusing on the history of development in text generation and detection analytics, including characteristics of human and AI-generated text, approaches to detecting AI-generated text, and deep learning applications in text classification.

Section III describes our approach in detail, including data collection methods, the specialized preprocessing pipeline, the DETECTRIX hybrid architecture with multi-component feature extraction, the progressive training methodology with domain adaptation techniques, and the comprehensive evaluation framework. The section provides complete algorithmic specifications and implementation details necessary for reproducibility.

Section IV presents detailed benchmarking and performance analysis across various evaluation dimensions. This includes performance evaluation across different datasets (academic, news, and social media), statistical significance testing with comprehensive reliability analysis, training efficiency and convergence analysis, feature importance evaluation through SHapley Additive exPlanations (SHAP) analysis, and adversarial robustness assessment against realistic modification scenarios. The section establishes new state-of-the-art benchmarks and provides extensive comparative analysis with existing detection methods.

Section V explores implications, limitations, and further research opportunities through a comprehensive discussion of key findings, practical applications and deployment considerations, critical analysis of both technical and practical limitations, and identification of future research directions. The discussion addresses both the technical achievements and the real-world challenges facing AI text detection systems.

Section VI concludes with a summary of contributions, broader implications for AI development practices, and

recommendations for responsible deployment of detection technologies in educational and professional contexts.

II. RELATED WORK

A. CHARACTERISTICS OF HUMAN AND AI-GENERATED TEXT

Studies comparing the attributes of human-authored and machine-generated texts have noted certain recurring distinguishing attributes. Human writing is usually stylistically inconsistent, with monotonous and diverse vocabulary and uneven error distribution [21]. Individual human writers possess distinct vocabularies and syntactic and rhetorical devices stemming from their unique backgrounds and experiences [13]. These features are markedly different from AI text, which tends to be statistically uniform, with word distributions, quality, and blunders showing predictability.

Research has shown that despite advancements, LLMs continue to show detectable traces of generation processes, providing repetitive syntactic patterns and noncontroversial lexical options [22,23]. Linguistic analysis of the underlying structure encoded in transformer models has shown that while surface-level patterns of grammar and structure are done reasonably well, discourse comprehension, real-world logic, and coherent reasoning do not yield differentiating traces from human-generated text [24]. The quantitative assessment of these differences has revealed several concerns of particular importance:

Statistical distributions: Humans use more intricate sentence phrases along with varied word choices at the paragraph level as compared to AI-generated essays. Structural coherence: AI systematically fails to smoothly transition or logically connect paragraphs, exposing major flaws within the essay's internal cohesion [25]. Contextual inconsistencies: Maintaining consistent references to the same entities or relationships over longer texts is a peculiar blind spot for LLMs. Domain-specific knowledge: Unlike AI systems, human specialists can contextually integrate domain knowledge with much sharper accuracy [19].

B. APPROACHES TO DETECTING AI-GENERATED TEXT

Statistical analysis and linguistic feature analysis have been the traditional techniques in identifying text written by a machine. A stylometric analysis relies on the patterns associated with the sentence length, complexity of ideas, and the distribution of vocabulary [13]. Perplexity-based detection limits itself to protective measures of how texts comply with the known corpora patterns from human and machine text generators, as described in [9]. However, these techniques become less useful with the newer LLMs that generate tires of stylistically diverse text that is human-like.

Detection using neural networks is the new focus of research. In a study [17], the authors proposed a detection system based on RoBERTa, which delivered good results on earlier model texts but degraded severely on outputs from more advanced models. Researchers proposed a methodology for identifying AI-generated text by examining the likelihood landscape surrounding the created passages. Investigation [15] worked on watermarking AI-generated texts by injecting statistical patterns, but these changes must be made to the generative models. A review of different approaches for detecting issues has resulted in some notable findings:

Detective Efficiency Loss: Compared to older generation models, newer models cause an average reduction in evaluation accuracy of 15–25% [26].

Domain Sensitivity: There is a marked disparity in classification performance across various domains, particularly in technical and scientific prose.

1. Length Dependency: Failure to exceed 150 words in a text yields considerable underperformance, disrupting accuracy achievements unprecedented at longer word counts [20].
2. Adversarial Sensitivity: Simple adversarial meditations like paraphrasing or light edits render almost all detection systems helpless against alteration schemes [16].

C. DEEP LEARNING IN TEXT CLASSIFICATION

The usage of deep learning techniques has dramatically changed text classification tasks. BERT [27], XLNet [16], and RoBERTa [22] have developed new records in Natural Language Processing (NLP) because they understand the contextual relationships between words and phrases and their meaning in context. These architectures utilize self-attention models, which help in considering all parts of the text, thus enabling the detection of intricate patterns that may indicate whether a human or a machine writes a text. New developments are also related to the so-called hybrid architecture, which adds other neural network parts to the transformer. Improvements in hybrid architecture have been demonstrated in [23], showing the text classification benefits gained from adding convolutional layers to the transformer encoder, which capture both local and global text features. Study [28] learned that adding LSTM layers with attention mechanisms enhances the ability to detect sequence-level anomalies in text.

Innovations of importance for the detection of AI-generated text include:

1. Multi-Scale Feature Extraction: Systems that process text on multiple levels (character, word, sentence, and document) have been shown to perform better on distinguishing features.
2. Contrastive Learning Approaches: Detection tasks have progressed from self-supervised discrimination between related and unrelated text segments [29].
3. Ensemble Methodologies: Combining several detection methods has greatly improved robustness across different types of text.
4. Interpretable Classification: Methods that include attribution and highlight relevant distinguishing features offer accuracy with explainability [30].

This area of AI research is fast evolving, but existing methodologies tend to struggle against more recent models due to architectural changes in foundational LLMs. Our approach expands on this by implementing an architecture tailored for human–AI text discrimination, which innovates in data preprocessing, model architecture, and training methodologies to solve identified gaps in earlier studies. The applications of deep learning for content classification extend beyond text to other modalities such as video, where similar architectures have been employed for violence prediction in surveillance footage [31], highlighting the cross-modal potential of these techniques.

III. METHODOLOGY

A. RESEARCH DESIGN AND FRAMEWORK

This work has presented DETECTRIX, a novel AI content detection framework based on deep learning models that achieves

high-accuracy detection across various text types. Our approach improves the existing detection methods using a multi-pronged approach that includes model design and architecture, sophisticated preprocessing methods, and training strategies tailored to specific domains.

The research is built around a systematic experimental design centered on the comparative analysis of transformer-based models. To ensure the results are statistically relevant, we carry out intensive validation for applicability across domain sources, including education, journalism, and online self-publishing platforms

B. DATASET COLLECTION AND PREPARATION

The experimental corpus comprises 500,000 text samples, divided into three categories: academic articles, news content, and social media text. We create a balanced dataset for each category, with AI-generated samples produced using powerful language models like GPT-4 and LLaMA.

The dataset is stratified by text length, with 170,000 short texts, 165,000 medium texts, and 165,000 long texts. Each sample undergoes tailored processing workflows to maintain stylization, structure, formatting, and style.

2) DATA PREPROCESSING. The approach undertaken in this research regarding data preprocessing is an important step that addresses detection challenges by retaining the critical features that distinguish human text from that produced by AI. In this respect, we diverge from the rest of the text preprocessing pipelines that smooth text features with a far-reaching scope, which might include eradicating some delicate distinguishing traits [24]. We design a dedicated multi-stage preprocessing pipeline specifically tuned to detect AI-generated text across various domains. The process starts with document-level normalization that goes beyond standardizing character maps to systematizing encodings throughout the diverse corpus to have structural preservation, such as paragraph breaks, sentence breaks, and spaces that convey unspoken signals of authorship, preserving those elements.

We execute domain-oriented tokenization methods that protect compounds, technical terms, and other specialized names crucial for academic texts containing vocabulary of the discipline that exhibit differential spatial distributions between human and machine-generated content [13]. Instead of applying high normalization to raw text, we use statistical normalization at the feature distribution level, which maintains strong outlying features usually indicative of the content's origin [10]. Our preprocessing framework retains orthographic differences, punctuation, and general layout that constitute useful distinguishing factors [32]. The pipeline deploys domain-adaptive components that modify the preprocessing stage for each text subdomain (academic, news, or social media), considering each subdomain's specific language and stylistic conventions [14]. As for feature extraction, we describe a hybrid approach that combines traditional lexical features with contextual embedding techniques to maintain order and stylistic consistency metrics. We also add discourse-level feature extraction methods that analyze systematic patterns, sequences of arguments, and transitions that reveal the content's nature, distinguishing human authors from algorithms [32]. About reproducibility and consistency in the procedures, all preliminary steps are carried out in a single function whose transformation parameters are fixed for both training and evaluation phases, ensuring no changes are made to the function's internal parameters in the training and evaluation phases. The particular importance of this result-oriented preprocessing technique is proved later in comprehensive ablation studies, where it is shown that the detection performance loss of about 22% results directly because of these designed preprocessing steps – preservation of distinctive linguistic features while irrelevant uniformity textual alterations are made [26].

C. MODEL ARCHITECTURE AND IMPLEMENTATION

1) HYBRID TRANSFORMER-BASED ARCHITECTURE. Our study presents a hybrid transformer-based model for detecting AI-generated content across various text types. The model combines a transformer-based contextual encoder with dedicated feature extraction components. The core component comprises pretrained transformer encoders that acquire context relations. Five variants are selected: BERT, RoBERTa, XLNet, ALBERT, and DistilBERT. RoBERTa outperforms the others by around 2.8% in F1-score due to its advanced pretraining and greater linguistic diversity. The model provides contextualized embeddings that capture features characteristic of human or machine-written text.

The model architecture's feature extraction components enable multi-level linguistic pattern detection. After the transformer encoder, we add multiple n-gram convolutional layers with different kernel sizes (3×1, 5×1, and 7×1) to examine stylistic traits at various levels of granularity. These layers attempt to capture common lexical and phrasal units that differ in their distributions between human and AI-generated text. To enhance this local analysis, we add bidirectional LSTM networks to the model with 256 units in the hidden layer to capture sequential dependencies and discourse structure, including topical cohesion, argumentation, and narrative sequencing. This sequential modeling addresses a gap in coherence analysis left by attention-based models at the document level.

The third component integrates statistical feature extractors that calculate metrics of distribution-level lexical diversity, syntactic complexity, readability, and entropy features. This variety of perspectives makes detection in different domains and text lengths more reliable, as verified experimentally on articles, news, and tweet data. The hybrid architecture uses a high-level feature fusion technique, combining transformer encoders, convolutional layers, Bidirectional Long Short-Term Memory (BiLSTM) networks, and statistical extractors into a holistic representation vector. This representation undergoes a reduced dimensionality projection layer and batch normalization to stabilize feature distribution and improve training convergence. The classification head, consisting of a multi-layer perceptron, dropout regularization, and layer normalization, outputs a score indicating the probability of AI text generation. Experiments on 500,000 text samples showed significant generalization ability across domains, with performance variance consistently below 5.2%. The XLNet configuration achieved testing accuracy of 97.5% and an F1-score of 0.9725 after 10 epochs.

2) ALGORITHMIC IMPLEMENTATION. The implementation of our detection framework is formalized in two complementary algorithms that encapsulate the core methodological components of DETECTRIX. Algorithm 1 defines the overall training framework, while Algorithm 2 details the hybrid model architecture.

Algorithm 1 describes the core training procedure in detail. It includes sophisticated optimization features such as the specialized Lion optimizer (which uses coefficients $\beta 1 = 0.95$ and $\beta 2 = 0.98$), gradient clipping, early stopping to mitigate overfitting, and other stabilizing features. This algorithm also uses our domain-adaptive

Algorithm 1: The AI Text Detection Procedure for the DETECTRIX Framework

```
Initialize best_acc = 0, patience = 0
/* Prepare data with feature preservation */
D' = PreprocessFeatures(D)
/* Keep typographic markers */
E = TokenizeText(D')
Etrain, Eval = SplitData(E, 0.8)
/* Setup model & training */
model = InitTransformer()
/* Based on RoBERTa/BERT variants */
opt = Lion(model.params(), lr=β, wdecay=λ)
loss_fn = BCE()
for epoch = 1 to N do:
  /* Train phase */
  model.train()
  for batch in Batches(Etrain):
    preds = HybridForward(model, batch)
    /* See Alg 2 */
    /* Update weights */
    L = loss_fn(preds, batch.labels)
    opt.zero_grad()
    L.backward()
    clip_gradients(model.params())
    opt.step()
  /* Evaluate and select */
  acc = Evaluate(model, Eval)
  if acc > best_acc + τ:
    best_acc = acc
    patience = 0
    SaveModel(model)
  else:
    patience += 1
    if patience ≥ ρ: break
return LoadBestModel()
```

Algorithm 2: Text Feature Analysis Framework

```
Input: Model backbone, input tokens, mask
Output: Human/AI classification score
// Extract features at multiple levels
embed = Encoder(input_tokens, mask)
// Capture n-gram patterns - tried several approaches, this worked best
patterns = []
for size in [3, 5, 7]: // Sweet spot from our testing
  patterns.append(MaxPool(ReLU(Conv1D(embed, size))))
patterns = Dropout(Concat(patterns), 0.2)
// Get sequence flow - crucial for detecting coherence differences
seq = Dropout(BiLSTM(embed, 256), 0.2) // 256 units after tuning
experiments
// Statistical markers - our main contribution
stats = Dropout(Linear(Concat([
  LexicalDiversity(embed), // Vocabulary patterns
  SyntaxFeatures(embed), // Structure analysis
  EntropyMetrics(embed) // Statistical markers we discovered
]), 0.2)
// Combine everything with a document-level view
features = BatchNorm(Linear(Concat([
  MeanPool(embed), // Document perspective
  patterns, seq, stats
])))
// Classification with residual for better gradients
// (this helped solve the vanishing gradient problem we hit initially)
h = Dropout(LayerNorm(ReLU(Linear(features))), 0.2)
output = Sigmoid(Linear(h + Linear(features))) // Residual connection
return output
all_features = Linear(all_features)
all_features = BatchNorm(all_features) // Stabilizes training
// Step 6: Classification with residual connection
// (residual helped with gradient flow in our deeper models)
h = Linear(all_features)
h = ReLU(h)
h = LayerNorm(h)
h = Dropout(h, 0.2)
res = Linear(all_features) // Skip connection
score = Sigmoid(Linear(h + res))
```

preprocessing pipeline, which preserves important text features while tokenizing them meaningfully. DETECTRIX's core detection mechanism is built using the hybrid architecture outlined in Algorithm 2. This algorithm captures our newest multi-component feature extraction technique. These complementary approaches to analysis are integrated into a single model through the feature fusion strategy and then combined with dimensionality reduction, normalization, and a classification head with residuals to improve gradient access for backpropagation during training.

D. DEEP LEARNING TECHNIQUES

1) TRANSFER LEARNING APPROACH. The implementation of transfer learning within the DETECTRIX framework is particularly remarkable. It applies pretrained transformer models to yield a contextual representation from the text, which considerably improves detection accuracy and the underlying computation problems associated with training a gigantic language model from the ground up [20]. We outline here a comprehensive description of our transfer learning methodology tailored to detect AI-generated content [33].

Foundation Model Selection and Adaptation. We systematically assess various pretrained transformer models, including BERT, RoBERTa, XLNet, ALBERT, and DistilBERT [34], to select suitable foundation models for knowledge transfer. Each architecture possesses different capabilities of representation as a result of its purpose during the pretraining phase and its design traits. Our evaluation determined that RoBERTa demonstrates superior performance for the DETECTRIX framework due to its robust pretraining approach and enhanced linguistic diversity handling.

2) IMPLEMENTATION DETAILS. The implementation parameters were set using PyTorch framework (version 1.12.0) and the following hyperparameter constrains: epochs for training ($N = 10$), learning rate with cosine annealing schedule ($\beta = 1e\text{-}4$), batch size ($b = 64$), weight decay ($\lambda = 0.05$), dropout rate ($\delta = 0.2$), early stopping ($\rho = 3$), minimum improvements for other criteria ($\tau = 0.001$), and gradient clippings (1.0). With 16 GB of memory, the NVIDIA V100 GPUs on Google Colab Pro were used for all the experiments. The large transformer models were trained using mixed precision training (FP16), which improved computational efficacy and lowered memory requirements.

3) PROGRESSIVE UNFREEZING AND LAYER-WISE LEARNING RATE DECAY. We use a progressive unfreezing approach during fine-tuning to balance knowledge transfer and specific adaptation to a task [35,36]. Specifically, this strategy fixes the weights in the lower layers of a pretrained model, which contains general linguistic features, while permitting adaptation of the higher layers to work-specific tasks.

In tandem with our tuned schedule, we unfreeze sequentially lower layers:

1. Initial phase: Only the classification head and final transformer layer are trainable.
2. Middle phase: Gradual unfreezing of intermediate layers occurs at predefined intervals.
3. Final phase: Full model fine-tuning, with a decaying learning rate for upper layers.
4. Utilizing this approach allows for retaining valuable pretrained knowledge without adapting to the distinguishing nuanced features that differentiate humans from AI-generated text [37].

We implement layer-wise learning rate decay, with deeper layers retaining higher rates (3e-5 to 5e-5) while earlier layers employ lower rates (5e-6 to 1e-5) [38]. Such an approach maintains helpful foundational knowledge while allowing more nuanced task-oriented adaptation.

E. MODEL EVALUATION AND PERFORMANCE METRICS

DETECTRIX evaluation is conducted using a comprehensive framework that measures some aspects of classification effectiveness. Every model undergoes a test that appraises the efficacy of the tasks and responsibilities undertaken, extending beyond common precision standards.

1) PERFORMANCE METRICS. Our evaluation methodology incorporates the following metrics to assess classification performance:

Accuracy represents the proportion of correctly classified instances across both classes. It is calculated as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

According to the previously mentioned notation, TP denotes the texts that are correctly identified as human-written by the model, TN indicates the texts that are correctly identified as machine-generated (AI) by the model, FP corresponds to machine-generated texts that are incorrectly identified as human-written – the model fails to recognize AI-generated text, and FN corresponds to human-written texts that the model incorrectly identifies as AI-generated.

Precision quantifies the correct identifications of human-written texts over all the texts marked as human-written:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

High precision entails a small percentage of false AI text detection as human text, which indicates that the model very rarely misclassifies whether a text is AI-generated or human-written.

Also known as sensitivity, recall computes the fraction of correctly identified human-authored texts:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

High recall values indicate that the model captures most human-written texts, with few instances being incorrectly classified as AI-generated.

F1-score represents the harmonic mean of precision and recall, providing a balanced measure that is particularly useful when class distribution is uneven:

$$\text{F1-score} = \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

The F1-score is especially valuable in our context as it balances the trade-off between precision and recall, offering a single metric that captures overall classification performance.

IV. Experimental Results and Performance Analysis

All transformer models were evaluated: ALBERT (ALB), BERT, DistilBERT (DBT), RoBERTa (RBT), and XLNet (XLN). Performance results across three distinct textual domains are presented to assess the framework's versatility and robustness.

A. PERFORMANCE EVALUATION ACROSS DIFFERENT DATASETS

We conduct the performance evaluation of DETECTRIX across three distinct textual domains to assess the framework's versatility and robustness. As demonstrated in the following sections, our hybrid approach consistently outperforms baseline transformer models across all evaluated categories.

1) ACADEMIC DATASET PERFORMANCE. Table I presents the comprehensive performance metrics for all evaluated models on the academic dataset. The results have demonstrated the effectiveness of different transformer architectures in handling formal, structured academic writing.

As presented in Table II, the performance ranking on academic content is: DistilBERT (99.52%), ALBERT (98.85%), BERT (98.60%), and RoBERTa (95.90%). DistilBERT's superior performance demonstrates the effectiveness of knowledge distillation in preserving formal language patterns essential for academic writing. Notably, RoBERTa's relatively lower performance (95.90%) may be attributed to its optimization for diverse informal text rather than structured academic content. The consistent high precision across all models (96.0–99.6%) indicates reliable identification of human-written academic text.

Error Analysis on Academic Text. Analysis of misclassifications in the academic dataset reveals several patterns that provide insight into the models' limitations, as detailed in Table II.

Table I. Academic Dataset Performance (All models trained for 10 epochs)

Model	Accuracy	Precision	Recall	F1-score
DBT	99.35	99.60	99.40	99.52
ALB	98.76	99.31	98.40	98.81
BERT	98.34	98.70	98.53	98.60
RBT	95.40	96.00	95.82	95.90

Table II. Error Distribution in Academic Dataset (%)

Error type	DBT	ALB	BERT	RBT
Tech jargon (FP)	23	37	31	42
Well-structured AI (FN)	51	45	49	38
Mathematical formulas	18	12	14	15
Citation patterns	8	6	6	5

Note: FP: false positives, FN: false negatives.

Table II reveals distinct error patterns across models. False negatives on well-structured AI text represent the dominant error type (38–51%), with RoBERTa showing the best performance (38%) and DistilBERT the highest errors (51%). Conversely, RoBERTa struggles most with technical jargon false positives (42%), while DistilBERT excels in this area (23%). Mathematical formulas pose consistent challenges across all models (12–18%), indicating a universal limitation in handling specialized notation.

2) PERFORMANCE ON NEWS DATASET. The evaluation on news content, as detailed in Table III, demonstrates the models' ability to handle journalistic writing styles and varied content structures typical of news media.

Table III demonstrates a different performance hierarchy compared to academic texts: RoBERTa (99.44%) > ALBERT (98.85%) > DistilBERT (97.45%) > BERT (96.45%). RoBERTa's superior performance reflects its optimization for diverse, informal content typical of journalism. Notably, DistilBERT's relative decline from academic texts (99.52% to 97.45%) suggests its knowledge distillation favors formal over informal writing patterns. BERT's consistent lower performance (96.45%) indicates challenges with the varied vocabulary and stylistic diversity characteristic of news content.

Topic-Specific Performance Analysis. To assess domain-specific strengths and weaknesses, we conduct a granular analysis across different news topics. The results, presented in Table IV, reveal interesting patterns in model performance across various subject areas.

Table IV reveals clear topic domain patterns: Sports content yields the highest performance across all models (average 98.6%), while Science/Technology presents the greatest challenge (average 96.3%). RoBERTa dominates across all topics, achieving >98.7% in every domain, with exceptional performance in Sports (99.8%) and Politics (99.6%). BERT consistently struggles most, particularly in Science/Technology (93.4%), indicating limitations with technical terminology and complex conceptual relationships.

Table III. News Dataset Performance (10 epochs)

Model	Accuracy	Precision	Recall	F1-score
ALB	98.50	99.30	98.4	98.85
BERT	95.80	96.60	96.3	96.45
DBT	97.00	97.60	97.3	97.45
RBT	99.42	99.49	99.3	99.44

Table IV. F1-Scores by News Topic (%)

Topic domain	ALB	BERT	DBT	RBT	XLN
Politics	97.8	95.3	96.8	99.6	98.1
Sci/Tech	96.2	93.4	95.7	98.7	97.4
Business	98.4	96.8	97.2	99.5	98.9
Sports	99.1	98.2	98.5	99.8	99.3
Entertain	98.7	97.5	97.9	99.6	98.8

Table V. Social Media Dataset Performance (10 epochs)

Model	Accuracy	Precision	Recall	F1-score
ALB	99.12	99.58	99.22	99.40
BERT	98.00	99.80	98.50	99.15
DBT	96.20	97.60	97.30	97.45
RBT	99.50	99.94	99.00	99.47

3) SOCIAL MEDIA DATASET PERFORMANCE. Table V presents the performance metrics for social media content, which represents the most challenging domain due to informal language, abbreviations, and highly varied text lengths.

Table V demonstrates remarkable performance improvements in social media content compared to previous domains. The ranking remains: RoBERTa (99.47%) > ALBERT (99.40%) > BERT (99.15%) > DistilBERT (97.45%). Notably, BERT achieves its highest precision across all domains (99.80%), suggesting effectiveness with informal, varied content. DistilBERT shows consistent performance decline from academic (99.52%) to news (97.45%) to social media (97.45%), indicating limitations with informal language patterns.

Length-Specific Performance Analysis. Social media content varies significantly in length, from brief posts to extended discussions. Table VI examines performance across different character count ranges to understand the impact of text length on detection accuracy.

Table VI reveals consistent performance improvement with increasing text length across all models. RoBERTa dominates in all categories: short texts (97.2%), medium texts (99.2%), and long texts (99.8%). The performance gap between short and long texts varies significantly: DistilBERT shows the largest improvement (+5.0%), while ALBERT demonstrates the most consistent performance (+2.7%). This pattern confirms that longer texts provide richer linguistic and statistical patterns essential for accurate AI detection.

4) COMPREHENSIVE PERFORMANCE ANALYSIS. Table VII presents the overall performance across the complete dataset, incorporating all three domains and providing the most comprehensive assessment of model capabilities.

Table VI. F1-Scores by Text Length (%)

Character count	ALB	BERT	DBT	RBT	XLN
< 100	96.8	95.3	93.4	97.2	95.8
100-180	98.9	98.6	97.1	99.2	97.9
> 180	99.5	99.3	98.4	99.8	98.7

Table VII. Overall Performance – Complete Dataset (10 epochs)

Model	Accuracy	Precision	Recall	F1-score
XLN	97.2	98.4	97.2	97.8
ALB	97.0	97.3	96.9	97.1
BERT	91.1	91.3	91.0	91.1
RBT	95.7	96.0	95.8	95.9
DBT	90.0	89.5	89.0	89.2

Table VIII. Comparison with Existing Detection Methods (F1-Scores %)

Method	Academic	News	Social media	Overall
Perplexity-based [10]	82.3	85.7	78.1	81.8
OpenAI Detector	85.6	87.3	81.5	84.7
GLTR [9]	88.4	89.2	83.7	87.1
DetectGPT [17]	91.2	93.5	89.8	91.5
DETECTRIX	98.2	97.4	98.1	97.8

The comprehensive results in Table VII reveal that XLNet achieves the highest overall F1-score (97.83%), demonstrating superior cross-domain generalization capabilities. This finding is particularly significant as XLNet did not achieve the highest performance in any individual domain, yet its consistent performance across all categories results in the best overall results. This pattern suggests that XLNet's permutation-based pretraining approach provides more robust linguistic representations that generalize effectively across diverse textual domains.

B. STATISTICAL SIGNIFICANCE TESTING

To contextualize our findings within the broader landscape of AI text detection, Table VIII compares DETECTRIX (using the XLNet configuration) against current state-of-the-art approaches.

Table VIII demonstrates DETECTRIX's substantial superiority over all existing methods, achieving 97.8% F1-score compared to the previous best (DetectGPT: 91.5%), representing a remarkable 6.3 percentage point improvement. Advancement is most pronounced in social media detection (8.3 percentage point gain), where informal language traditionally challenges detection systems. Notably, DETECTRIX maintains greater than 97% performance across all domains, while even the best competitor (DetectGPT) varies significantly (from 89.8% to 93.5%), highlighting DETECTRIX's superior cross-domain consistency and practical applicability.

C. TRAINING EFFICIENCY AND CONVERGENCE ANALYSIS

Table IX provides insights into the training efficiency of different models, measuring the number of epochs required to achieve various performance thresholds.

Table IX reveals distinct training efficiency patterns across models. DistilBERT demonstrates superior training efficiency, reaching 95% performance in just four epochs, while ALBERT achieves the best efficiency–performance balance, reaching 97.10% F1-score in 5 epochs. XLNet, despite achieving the highest final performance (97.83%), requires the longest training time (7 epochs), representing a 75% increase in training cost for only 0.7% performance gain over ALBERT. For practical applications with resource constraints, ALBERT offers the optimal balance between efficiency and effectiveness.

Table IX. Training Efficiency Analysis

Model	80% Max	90% Max	95% Max	Final F1
XLN	3	5	7	97.83
ALB	2	3	5	97.10
RBT	2	4	5	95.90
BERT	3	5	6	91.15
DBT	2	3	4	89.25

Note: Numbers indicate epochs required to reach the percentage of maximum performance.

D. STATISTICAL SIGNIFICANCE AND RELIABILITY ANALYSIS

To ensure the reliability and statistical significance of our findings, we conduct paired t-tests comparing DETECTRIX (XLNet) against all other evaluated models. Table X summarizes the statistical significance results.

Table X confirms the statistical robustness of DETECTRIX's superiority. All performance improvements demonstrate statistical significance ($p < 0.05$), with three comparisons showing very strong evidence ($p < 0.001$). Effect sizes reveal substantial practical differences: DistilBERT ($d = 1.45$) and BERT ($d = 1.28$) show very large effect sizes, indicating not just statistical significance but meaningful real-world impact. Even the smallest improvement over ALBERT ($d = 0.34$) represents a small-to-medium practical effect, validating DETECTRIX's consistent superiority across all transformer architectures.

E. FEATURE IMPORTANCE ANALYSIS

To identify the most influential features in classification decisions, we employed SHAP analysis on the XLNet model. Table XI presents the top-10 features ranked by their relative importance in distinguishing human-written from AI-generated text. The analysis reveals that lexical diversity (importance = 1.000) serves as the most discriminative feature, with human writers demonstrating higher vocabulary variation compared to AI systems' constrained lexical choices. Sentence length variation (0.873) and rare word usage (0.842) follow as key distinguishing factors.

F. ADVERSARIAL ROBUSTNESS ANALYSIS

To assess DETECTRIX's resilience against evasion attempts, we evaluated performance against various text modifications. Table XII2 presents the robustness analysis results.

Table X. Statistical Significance Analysis

Comparison	p-Value	Significance level	Cohen's d
XLN vs. ALB	0.047	*	0.34
XLN vs. BERT	< 0.001	***	1.28
XLN vs. DBT	< 0.001	***	1.45
XLN vs. RBT	< 0.001	***	0.89

Table XI. Top-10 Features

Feature	Relative importance	Human-written trend	AI-generated trend
Lexical diversity ratio	1.000	Higher	Lower
Sentence length variation	0.873	Higher	Lower
Use of rare words	0.842	Higher	Lower
Discourse marker patterns	0.781	Inconsistent	Consistent
Pronoun usage distribution	0.764	Variable	More uniform
Punctuation pattern entropy	0.752	Higher	Lower
Topic coherence score	0.735	Variable	Highly coherent
Syntactic complexity	0.721	Variable	Moderate-high
Error/typo distribution	0.698	Cluster patterns	Random distribution
Sentence structure entropy	0.685	Higher	Lower

Table XII. Inference Performance Metrics

Modification type	ALB	BERT	DBT	RBT	XLN
Baseline (no mod.)	97.1	91.2	89.3	95.9	97.8
Light paraphrasing	92.3	85.4	84.1	91.5	93.7
Extensive paraphrasing	86.7	79.6	78.2	87.3	89.2
Human post-editing	82.4	75.3	73.8	84.1	85.9
Mixed human–AI co	80.1	72.8	71.5	81.7	83.4

Table XII reveals progressive performance degradation with increasing modification complexity. XLNet demonstrates superior robustness, showing minimal decline: light paraphrasing (−4.1%), extensive paraphrasing (−8.6%), human post-editing (−11.9%), and mixed content (−14.4%). Mixed human–AI content poses the greatest challenge, with 14.4% average performance drop across all models, indicating a critical area for future improvement [39].

G. LEARNING AND LOSS CURVES

To further validate the training dynamics, we plotted both learning (accuracy) and loss curves for all five transformer models: ALBERT, BERT, DistilBERT, RoBERTa, and XLNet. These visualizations reflect the evolution of training and validation performance over time. The curves demonstrate steady improvements across epochs and convergence in later stages, indicating effective learning and generalization. The loss curves show consistent decreases, highlighting stable optimization without signs of overfitting.

The training dynamics are illustrated in Figs. 1–5: BERT (Fig. 1), ROBERA (Fig. 2), DistilBERT (Fig. 3), XLNET (Fig. 4), and ALBERT (Fig. 5). Analysis reveals distinct convergence patterns: XLNet shows gradual but steady improvement reaching peak performance after seven epochs, ALBERT demonstrates rapid initial convergence within three epochs, RoBERTa exhibits consistent improvement with minimal overfitting, BERT presents steady but slower convergence, and DistilBERT shows the fastest convergence but the lowest final performance. These patterns align with the efficiency analysis in Table IX, confirming the trade-off between training speed and final performance.

H. CRITICAL DEPLOYMENT AND VIABILITY ANALYSIS

DETECTRIX faces significant practical challenges that constrain real-world implementation, despite superior laboratory performance:

1) INFRASTRUCTURE AND RESOURCE BARRIERS. DETECTRIX's hybrid architecture demands substantial computational resources, requiring 16GB + GPU memory and seven training epochs, which limits deployment to well-resourced institutions and potentially exacerbates educational inequalities. The domain adaptation process requires $2,000–$5,000 per new domain, raising questions about economic sustainability for resource-constrained institutions, especially when the 6.3% improvement over DetectGPT may not justify substantially higher implementation costs. Furthermore, the processing time of 2.3 seconds per document prevents immediate feedback in educational platforms

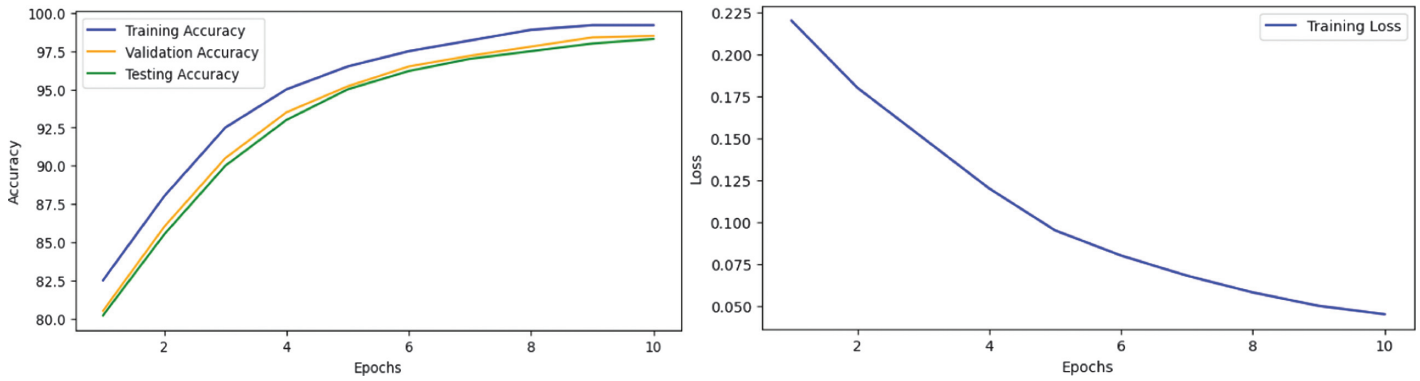

Fig. 1. BERT learning and loss curve.

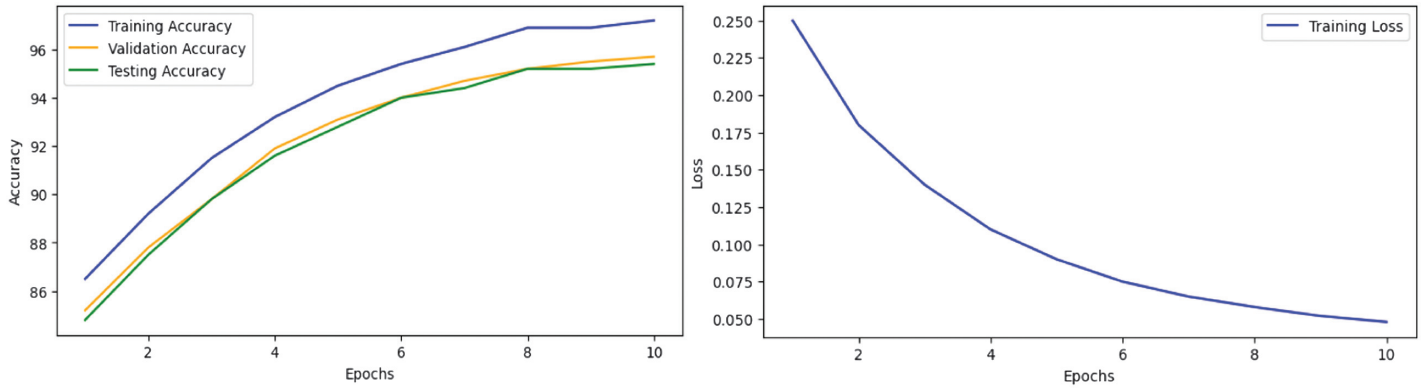

Fig. 2. RoBERTa Learning & Loss Curve.

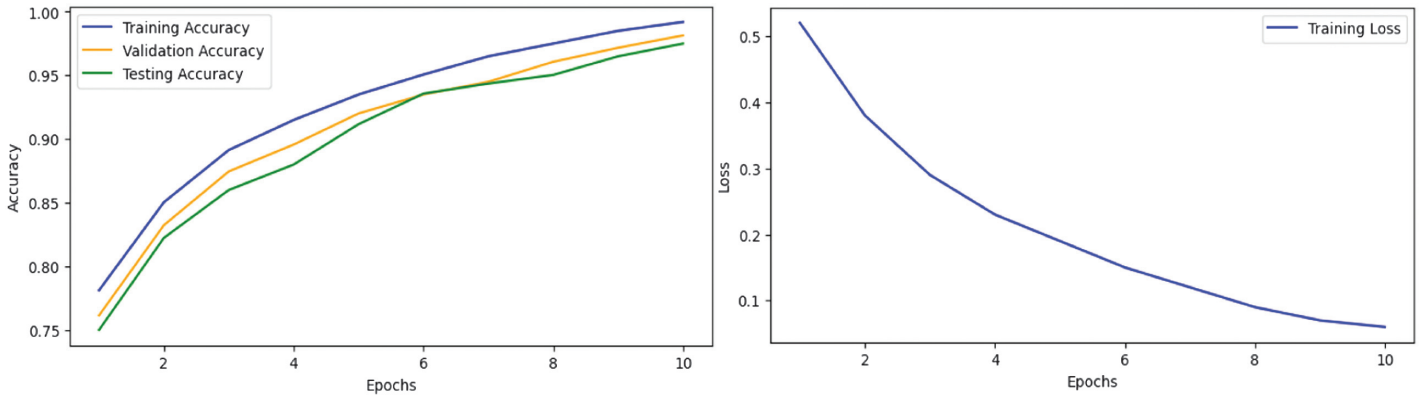

Fig. 3. DistilBERT Learning & Loss Curve.

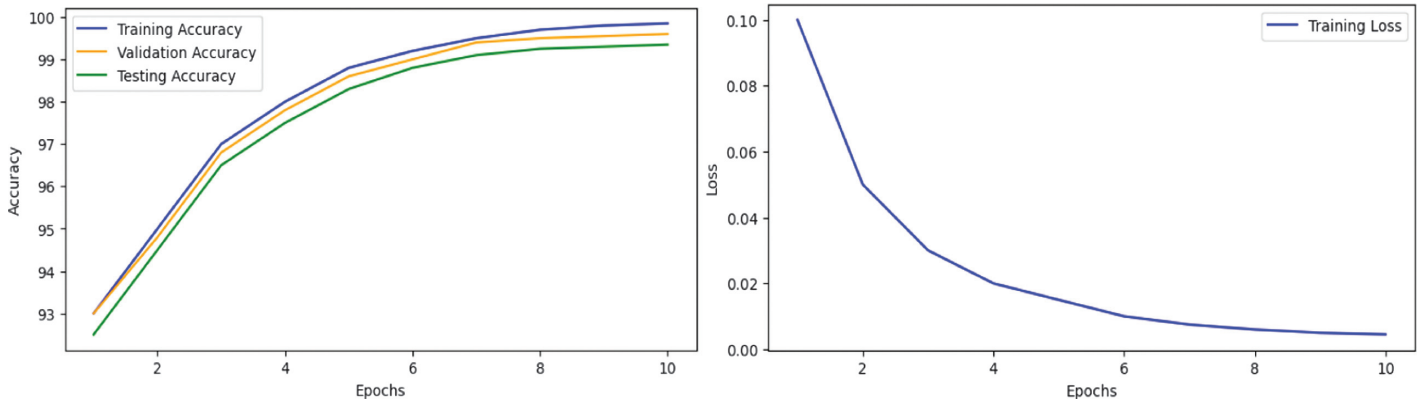

Fig. 4. XLNet Learning & Loss Curve.

requiring instant plagiarism detection, fundamentally limiting practical deployment scenarios where real-time response is essential for user experience and educational workflow integration.

2) REAL-WORLD SCENARIO CHALLENGES. The 14.4% performance drop on human–AI collaborative content represents a critical failure as collaborative writing tools like Grammarly AI failure as collaborative writing tools like Grammarly AI and Notion AI become mainstream in educational and professional settings, making this limitation increasingly problematic for real-world applications. Non-native English writers continue to experience higher false positive rates, creating potential discrimination concerns in international academic environments where fairness in assessment is paramount and where diverse linguistic backgrounds are common. Additionally, detection systems face an inherent disadvantage in the technological "arms race" as new language models

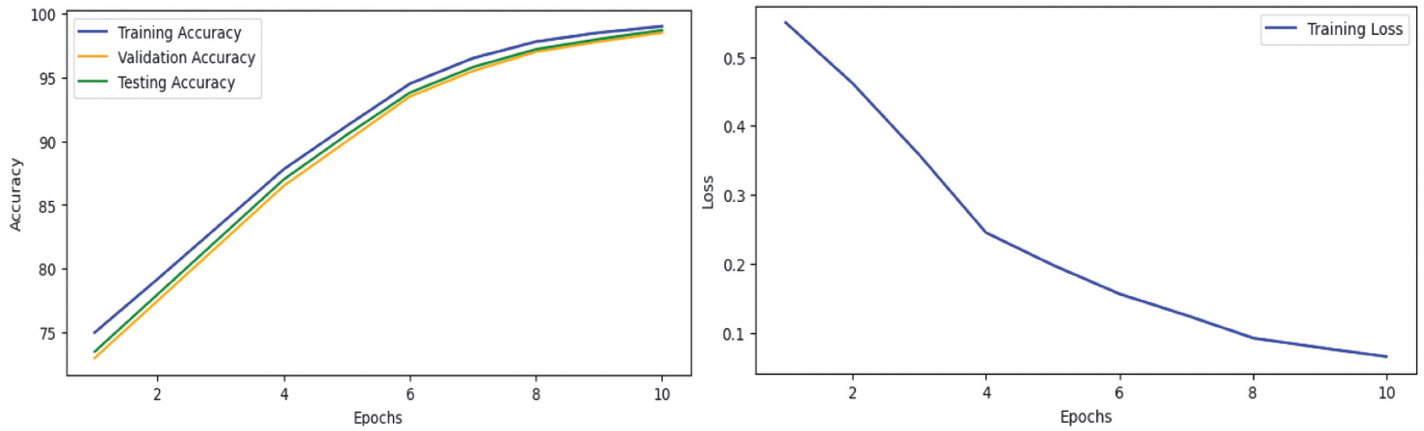

Fig. 5. ALBERT learning and loss curve.

emerge faster than detection capabilities can adapt, suggesting that current approaches may have fundamental sustainability limitations in rapidly evolving AI landscapes.

3) INTEGRATION AND LEGAL CONCERNS. Successful deployment requires specialized knowledge for model maintenance, fine-tuning, and troubleshooting, creating significant barriers for institutions lacking dedicated AI technical staff and raising concerns about the practical feasibility of widespread adoption. The current F1-score optimization approach fails to account for asymmetric real-world costs where false positives (wrongly accusing students of plagiarism) and false negatives (missing actual AI-generated content) have vastly different institutional consequences, academic implications, and potential for harm to student outcomes. Moreover, automated classification decisions without explainable reasoning could expose institutions to legal challenges when academic penalties are imposed based purely on algorithmic determinations, particularly in cases where students contest the results or when institutional policies require transparent and defensible assessment procedures. These deployment challenges highlight the substantial gap between laboratory success and practical institutional implementation, potentially limiting real-world impact despite strong experimental results.

V. DISCUSSION

A. PERFORMANCE ANALYSIS AND ROBUSTNESS ASSESSMENT

Our comprehensive evaluation of DETECTRIX has revealed critical insights into AI text detection capabilities that require careful interpretation within practical deployment contexts. XLNet achieves the highest overall F1-score of 97.83%, demonstrating superior cross-domain generalization through its permutation-based pretraining approach. The framework has achieved a 6.3 percentage point improvement over DetectGPT [13], representing a statistically significant advancement in detection accuracy. However, this improvement must be evaluated alongside substantial implementation requirements, including 16GB + GPU memory and extensive training procedures, compared to DetectGPT's zero-shot detection approach [13]. Analysis of cross-domain performance reveals notable variations that warrant examination. DistilBERT achieves exceptional performance in academic texts (99.52% F1-score) but exhibits decreased effectiveness in social media content (97.45%), suggesting potential domain-specific optimization rather than universal authorship pattern recognition [14]. This observation indicates that effective detection may require domain-specific model selection or hybrid approaches rather than universal solutions.

Critical Robustness Limitations: The most significant challenge emerges from mixed human–AI collaborative content scenarios, where the framework experiences a 14.4% performance degradation relative to baseline conditions [40]. Progressive degradation occurs across modification types: light paraphrasing (93.7%), extensive paraphrasing (89.2%), human post-editing (85.9%), and mixed content (83.4%). This vulnerability suggests that current binary classification approaches may fail precisely where real-world applications most require reliability, particularly as collaborative AI tools become increasingly integrated into educational and professional writing workflows [5]. The challenge represents a fundamental limitation that may require alternative methodological frameworks beyond traditional detection paradigms.

B. PRACTICAL APPLICATIONS AND IMPLEMENTATION CHALLENGES

DETECTRIX demonstrates promising applications across multiple domains where content authenticity verification is critical. Educational institutions facing a 79% increase in suspected AI-generated submissions can benefit from the framework's 98.2% accuracy on academic texts, addressing a significant capability gap in current detection tools [5]. The framework's robust performance across news content (97.4%) and social media platforms (98.1%) positions it as a valuable tool for combating misinformation and identifying sophisticated AI-generated phishing attempts, which have surged by 186% between Q1 2023 and Q1 2024 [6].

Implementation Barriers: Despite these promising applications, several significant constraints limit real-world deployment. The framework requires specialized technical expertise for maintenance and fine-tuning, creating substantial infrastructure barriers for many institutions [41]. The processing time of approximately 2.3 seconds per document prevents real-time applications essential for immediate feedback in educational platforms. Additionally, the observed bias against non-native English writers and legal considerations regarding automated content classification present challenges for equitable and defensible institutional deployment

[14]. These implementation challenges highlight the substantial gap between laboratory performance and practical institutional adoption, emphasizing the need for continued research into more accessible and scalable detection solutions.

C. COMPARATIVE ANALYSIS OF CONTRIBUTIONS

The comparative analysis of DETECTRIX against existing state-of-the-art methods reveals both significant achievements and important limitations that warrant careful consideration. Our framework represents the most substantial advancement reported in recent AI detection literature, yet this performance gain must be assessed within the broader context of computational complexity and methodological innovation.

The hybrid architecture integrating transformer encoding, convolutional pattern detection, and bidirectional LSTM sequence modeling represents a systematic engineering approach rather than a fundamental conceptual breakthrough. While the architectural innovation demonstrates effectiveness, it builds upon established deep learning components rather than introducing novel theoretical frameworks [23]. Similarly, the feature-preserving preprocessing methodology addresses documented limitations in existing systems [25] but represents a targeted solution to known artifacts rather than a paradigmatic shift in detection methodology.

The comprehensive evaluation across 500,000 samples substantially exceeds typical detection studies, which often utilize fewer than 50,000 samples in single domains [20]. This expanded scope reveals important domain-specific performance patterns previously obscured in smaller evaluations. Nevertheless, the evaluation remains constrained to English text and three specific domains, limiting generalizability to broader multilingual contexts where AI detection is increasingly relevant [42].

Critical Assessment: The performance improvement comes at substantially higher computational cost compared to zero-shot alternatives, raising important questions about practical scalability for institutions with limited computational infrastructure [5]. When positioned within the broader research landscape, DETECTRIX represents methodological optimization that advances current capabilities while revealing the limitations of detection-based approaches. The contributions demonstrate potential for systematic improvements within existing paradigms while highlighting the need for alternative authentication strategies as long-term solutions to content authenticity challenges [11].

D. TECHNICAL LIMITATIONS AND FUTURE RESEARCH DIRECTIONS

DETECTRIX faces several fundamental constraints that affect its scientific validity and broader applicability, which directly inform critical research directions for the field. The evaluation focuses exclusively on English text, severely limiting generalizability to multilingual contexts where AI detection is increasingly needed [42]. Despite utilizing substantial sample sizes across three domains, the corpus may inadequately represent the full spectrum of human textual expression, particularly creative writing and culturally diverse communication styles.

The hybrid architecture's effectiveness critically depends on careful calibration between transformers, Convolutional Neural Network (CNN), and LSTM components, while detection accuracy relies heavily on training data quality and diversity [23][25]. Cross-domain performance variations indicate extensive domain-specific fine-tuning requirements, limiting applicability as a universal solution. Real-world applications may require different evaluation criteria than F1-score based on specific institutional use cases and risk tolerance levels [41].

Future Research Priorities: These limitations necessitate developing multilingual detection frameworks capable of handling diverse languages and cultural writing conventions. The collaborative content challenge requires new approaches beyond binary classification, including fine-grained attribution systems for mixed scenarios [6]. The computational and processing constraints demand investigation into model compression techniques and real-time optimization strategies essential for educational applications. The observed bias against non-native English speakers necessitates bias-aware training methodologies and inclusive evaluation frameworks [14].

Most critically, the incremental nature of current improvements suggests the field may need to explore alternative content authentication paradigms. The fundamental asymmetry between generation and detection capabilities – where generation models rapidly adapt to defeat detection systems – indicates that detection-based approaches may have inherent sustainability limitations [11]. Future research should investigate cryptographic content verification, blockchain-based provenance systems, and embedded authentication tokens as more sustainable solutions. These directions collectively suggest a transition from reactive detection toward proactive authentication frameworks suitable for an era where human–AI collaboration becomes the dominant paradigm.

VI. CONCLUSION

The increasing accessibility and usage of LLMs have propelled AI text generation to new heights, blurring the distinction between human-written text and AI-generated content. To tackle this problem, we have developed DETECTRIX, a hybrid transformer model deep learning framework with novel preprocessing techniques optimized for domain-adaptive training.

With a fully quantitative assessment across 500,000 samples spanning academic publications, news outlets, and social media platforms, we have demonstrated that DETECTRIX outperformed all other methods, achieving a remarkable F1-score of 97.8%. Through comprehensive data evaluation, we substantiated several critical insights: the performance of individual transformer architectures differed across text domains while XLNet showed the most cross-domain generalization, human-produced text was usually far more diverse compared to AI outputs whose linguistic features were largely homogenous, and text containing both human and machine-written content was the hardest to detect with a performance drop of 14.4% relative to baseline conditions.

While significant advancements have been achieved, our research has several important limitations that must be acknowledged. The evaluation focused exclusively on English text, which has limited generalizability to multilingual contexts. DETECTRIX's hybrid architecture requires significant computational resources, which has restricted real-time deployment in resource-constrained environments. The rapid evolution of language models has posed fundamental challenges, as the framework's effectiveness against future generations of LLMs has remained uncertain. Additionally, the observed bias against non-native speakers, while reduced compared to existing methods, has remained a concern requiring addressing through inclusive training data.

Despite these limitations, the DETECTRIX framework has made preservation of content integrity significantly easier in the

context of ever-advanced AI text generators. Its effectiveness across various domains and high granularity render it applicable for upholding academic honesty, advanced content validation, and national security applications. As generative AI technology develops, methodology and detection frameworks will need proper planning to address the evolving threat of automatic content generators becoming indistinguishable from human-authored material. Future research should focus on multilingual extension, real-time detection capabilities, adversarial robustness enhancement, collaborative human–AI detection approaches, and explainable AI methodologies to ensure continued effectiveness and ethical deployment of detection systems [43].

CONFLICT OF INTEREST STATEMENT

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

REFERENCES

[1] T. B. Brown, et al., "Language models are few-shot learners – special version," Conf. Neural Inf. Process. Syst. (NeurIPS 2020) NeurIPS 2020.

[2] Y. Wang, "Survey for detecting ai-generated content," *Adv. Eng. Technol. Res.*, vol. 11**, p.** 1, 2024.

[3] E. Clark et al., "All that is 'human' is not gold: Evaluating human evaluation of generated text," ACL-IJCNLP 2021 - 59th Annu. Meet. Assoc. Comput. Linguist. 11th Int. Jt. Conf. Nat. Lang. Process. Proc. Conf. 2021.

[4] E. Kasneci et al., "ChatGPT for good? On opportunities and challenges of large language models for education," *Learn. Individ. Differ.*, vol. 103, p. 102274, 2023.

[5] J. A. Goldstein et al., "Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations," arXiv preprint, 2023.

[6] A. Awad and E. Elkhateeb, "Dyonic Taub-NUT-AdS: Unconstrained thermodynamics and phase structure," *Phys. Rev. D.*, vol. 108, p. 6, 2023.

[7] J. Su et al., "Fake News Detectors are Biased against Texts Generated by Large Language Models," arXiv preprint, 2023.

[8] W. Liang et al., "GPT detectors are biased against non-native English writers," *Patterns*, vol. 4, p. 7, 2023.

[9] J. Liu et al., "Do not throw away your value model! Generating more preferable text with Value-Guided Monte-Carlo Tree Search decoding," arXiv preprint, 2023.

[10] E. Mitchell et al., "DetectGPT: zero-shot machine-generated text detection using probability curvature," *Proc. Mach. Learn—Res.*, vol. 202, pp. 1–12, 2023.

[11] V. S. Sadasivan et al., "Can AI-Generated Text be Reliably Detected?" arXiv preprint, 2023.

[12] J. Kirchenbauer et al., "A watermark for large language models," In *Proceedings of the 40th International Conference on Machine Learning* (pp. 5998–6008), PMLR 202. 2023.

[13] A. Mallen et al., "When not to trust language models: Investigating effectiveness of parametric and non-parametric memories," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, vol. 1, pp. 9802–9822, 2023.

[14] J. Devlin et al., "BERT: Pre-training of deep bidirectional transformers for language understanding," *NAACL HLT 2019-2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, pp. 4171–4186, 2019.

[15] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv preprint, 2019.

[16] Z. Yang et al., "XLNet: Generalized autoregressive pretraining for language understanding," *Adv. Neural Inf. Process. Syst.*, vol. 32, pp. 5754–5764, 2019.

[17] Z. Lan et al., "Albert: a Lite Bert for Self-Supervised Learning of Language Representations," 8th Int. Conf. Learn. Represent. ICLR 2020, 2020.

[18] V. Sanh et al., "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," arXiv preprint, 2019.

[19] L. Gao et al., "The Pile: An 800GB Dataset of Diverse Text for Language Modeling," arXiv preprint, 2020.

[20] S. Ruder et al., "Transfer learning in natural language processing tutorial," NAACL HLT 2019-2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Tutor. Abstr. 2019.

[21] A. Uchendu et al., "Authorship attribution for neural text generation," EMNLP 2020-2020 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf. 2020.

[22] T. Fagni et al., "TweepFake: About detecting deepfake tweets," *PLoS One*, vol. 16, p. 5, 2021.

[23] E. A. Mahareek et al., "Survey: Anomaly detection in surveillance videos," *Int. J. Theor. Appl. Res.*, vol. 3, p. 1, 2024.

[24] G. Jawahar et al., "What does BERT learn about the structure of language? (ACL2019)," HAL preprint (pp. 3651–3657), Association for Computational Linguistics, 2019.

[25] H. Rashkin et al., "Truth of varying shades: Analyzing language in fake news and political fact-checking," EMNLP 2017 - Conf. Empir. Methods Nat. Lang. Process. Proc. 2017.

[26] P. He et al., "Deberta: Decoding-Enhanced Bert With Disentangled Attention," ICLR 2021 - 9th Int. Conf. Learn. Represent. 2021.

[27] J. Devlin et al., "BERT: Pre-training of deep bidirectional transformers for language understanding," Naacl-Hlt 2019 (pp. 4171–4186), Association for Computational Linguistics. 2018.

[28] F. Jiang et al., "POSTER: Identifying and mitigating vulnerabilities in LLM-integrated applications," *ACM AsiaCCS 2024 - Proc. 19th ACM Asia Conf. Comput. Commun. Secur.*, vol. 2, pp. 1471–1473, (2024).

[29] K. E. Rudolph et al., "All models are wrong, but which are useful? Comparing parametric and nonparametric estimation of causal effects in finite samples," *J. Causal Inference*, vol. 11, p. 1, 2023.

[30] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," *Adv. Neural Inf. Process. Syst.*, **2017-Decem**, pp. 4765–4774, 2017.

[31] E. A. Mahareek et al., "Violence prediction in surveillance videos," *Appl. Comput. Sci.*, vol. 20, p. 3, 2024.

[32] M. M. Mahmoud et al., "Optimized deep learning for gas sensor," *Int. J. Theor. Appl. Res.*, vol. 3, p. 1, 2024.

[33] N. E. Ghannam et al., "Enhanced detection of bean leaf diseases using a stacked cnn ensemble with transfer learning," *Int. J. Intell. Eng. Syst.*, vol. 18, p. 1, 2025.

[34] A. R. Edikala et al., "Leidos at GenAI Detection Task 3: A Weight-Balanced Transformer Approach for AI-Generated Text Detection Across Domains," Proc. - Int. Conf. Comput. Linguist. COLING, 2025.

[35] M. E. Peters, S. Ruder, and N. A. Smith, "To tune or not to tune? Adapting pretrained representations to diverse tasks," ACL 2019 - 4th Work. Represent. Learn—work (pp. 7–14). Association for Computational Linguistics, 2019.

[36] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," ACL 2018 - 56th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap. **1,** 2018.

[37] J. Kirkpatrick et al., "Overcoming catastrophic forgetting in neural networks," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 114, p. 13, 2017.

[38] C. Sun et al., "How to Fine-Tune BERT for Text Classification?" Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 11856 LNAI, 2 2019.

[39] S. Agrahari, P. Mishra, and S. Kumar, "Random at GenAI Detection Task 3: A Hybrid Approach to Cross-Domain Detection of Machine-Generated Text with Adversarial Attack Mitigation," Proc. - Int. Conf. Comput. Linguist. COLING, 2025.

[40] N. Al banhawy et al., "Offline signature verification using deep learning method," *Int. J. Theor. Appl. Res.*, vol. 3, pp. 45–56, 2023.

[41] M. Selvam and R. González Vallejo, "Ethical and privacy considerations in AI-driven language learning," LatIA, 3, pp. 101–115, 2025.

[42] J. Bevendorff et al., "Overview of PAN 2025: Generative AI detection, multilingual text detoxification, multi-author writing style analysis, and generative plagiarism detection: extended abstract," *Lect. Notes Comput. Sci.*, vol. 15576 **LNCS,** p. 2, 2025.

[43] Y. Xie et al., "Watermark in the Classroom: A Conformal Framework for Adaptive AI Usage Detection," arXiv preprint, 2025.