Journal of ICT Aplication and System (JICTAS)
Volume 5 Number 1, Juni 2026, pp. 41 ~57
ISSN: 2830-098X, DOI: 10.56313/jictas.v5i1.531

TemporalXAI-Det: Temporal-Aware Explainable Detection of Multi-Model AI-Generated Academic Text via Continual Learning and Cross-Lingual Transfer

Imeldawaty Gultom[1], Ratih Puspadini[2], Fauzi Erwis[3], Elyandri Prasiwiningrum[4], Ridwan[5]
[1,2]Department System Information, STMIK Kaputama, Medan, Indonesia
[3,4,5]Computer of Sciences, Universitas Rokania, Riau, Indonesia

Email: imeldagultom81@gmail.com[1], puspadini.ratih@gmail.com[2] , fauzierwis@gmail.com[3]
eprasiwiningrum@gmail.com[4], ridwan@rokania.ac.id[5]

Article Info

Article history:

Received 05 27, 2026
Revised 06 03, 2026
Accepted 06 11, 2026

Keywords :

Academic Integrity,
Catastrophic Forgetting,
Continual Learning,
Explainable AI,
Temporal Model Drift

ABSTRACT

The proliferation of heterogeneous generative AI systems—including GPT-4o, Claude 3 Opus, Gemini 1.5 Pro, Mistral, and LLaMA-3—has produced a multi-source academic text landscape whose detection presents challenges qualitatively beyond those addressed by existing binary or single-source detection paradigms. Contemporary detectors are doubly compromised: first, by adversarial paraphrasing that disrupts surface-level distributional signatures; second, by temporal model drift, wherein new model generations evade detectors trained on earlier LLM families. This study introduces TemporalXAI-Det, a continual-learning explainable detection framework capable of (1) attributing academic text to one of five generative model families while simultaneously identifying human authorship, yielding a six-class taxonomy; (2) adapting to new LLM generations without catastrophic forgetting via Elastic Weight Consolidation (EWC) and experience replay; (3) transferring robustly across twelve academic languages through a Language-Adaptive Prefix Tuning (LAPT) mechanism applied to XLM-RoBERTa-XL; and (4) generating legally defensible per-instance explanations via Integrated Gradients (IG), SHAP, and counterfactual generation. A large-scale continual benchmark corpus (MTA-72K) comprising 72,000 samples across six source classes, four adversarial attack paradigms, and twelve languages is constructed and released. TemporalXAI-Det achieves a six-class macro F1-score of 0.941 on the clean test partition, 0.912 under combined adversarial conditions (performance degradation $\Delta = 2.9$ pp), and a mean cross-lingual F1 of 0.887 across all twelve evaluated languages. Continual learning experiments demonstrate that catastrophic forgetting is reduced by 78.4% relative to standard fine-tuning when new LLM families are introduced. These results establish new state-of-the-art benchmarks for multi-source, temporally robust, and multilingual AI-text detection in academic integrity contexts.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Imeldawaty Gultom
Departement System Information, STMIK Kaputama, Medan, Indonesia
Email: imeldagultom81@gmail.com

❒
ISSN: 2830-098X

1. INTRODUCTION

The contemporary landscape of academic text production has undergone a profound structural transformation, driven by the concurrent proliferation of multiple competing generative AI families each capable of producing indistinguishably fluent, domain-coherent scholarly prose. Unlike the earlier generational landscape in which GPT-2 and early GPT-3 variants represented a tractable detection target[1], [2], the current environment presents a fundamentally combinatorial challenge: institutional submissions may originate from GPT-4o, Claude 3 Opus, Gemini 1.5 Pro, Mistral-8x22B, or LLaMA-3-70B, each exhibiting a distinct yet partially overlapping statistical fingerprint. Binary human-vs-AI classifiers trained against a single model family are demonstrably inadequate for this multi-source regime, as cross-model generalisation rates routinely fall below 60% when evaluated against out-of-distribution generators[3], [4].

Compounding this model-diversity challenge is the phenomenon of temporal model drift: as AI providers continuously release updated model versions (GPT-4 to GPT-4o to GPT-4o-mini; Gemini 1.0 to 1.5 to 2.0), the statistical characteristics of AI-generated text evolve in ways that silently invalidate previously trained detectors. Empirical evaluations have demonstrated that a detector trained in Q1 2024 exhibits accuracy degradation of 15–22 percentage points against model versions released in Q4 2024, even absent any adversarial intervention. This temporal drift constitutes a systemic deployment vulnerability that has been entirely neglected by the existing detection literature[5], [6].

A third orthogonal dimension of inadequacy concerns linguistic coverage. Existing AI-text detectors are overwhelmingly designed and evaluated on English text[7]. Global academic institutions, particularly in Southeast Asia, the Middle East, and Latin America, operate in multilingual environments where AI-assisted writing across Arabic, Bahasa Indonesia, Mandarin, Spanish, French, and other major languages is equally prevalent but detection infrastructure is absent[8], [9]. The assumption of English monolingualism is not merely a convenience—it constitutes a structural equity gap that disadvantages institutions in the Global South from implementing evidence-based academic integrity governance[10], [11].

This paper addresses these three compounding inadequacies—model multiplicity, temporal drift, and linguistic coverage—through a unified continual-learning framework, TemporalXAI-Det[12], [13]. The system introduces a six-class taxonomy (human + five LLM families), adapts to new model generations via Elastic Weight Consolidation (EWC) with experience replay, extends detection robustly to twelve languages via Language-Adaptive Prefix Tuning (LAPT) [9] applied to XLM-RoBERTa-XL, and generates per-instance explanations via Integrated Gradients (IG) [10], SHAP, and counterfactual contrast pairs[14], [15]. The framework is evaluated on MTA-72K, a purpose-built multilingual temporal adversarial corpus comprising 72,000 samples[16], [17].

The principal novel contributions of this work are as follows: A six-class multi-source detection taxonomy covering five major LLM families plus human authorship, with evaluation under four distinct adversarial attack paradigms[18], [19]. A continual learning protocol (EWC + Experience Replay) that enables TemporalXAI-Det to assimilate new LLM generations without catastrophic forgetting, reducing forgetting by 78.4% relative to standard fine-tuning[20], [21], [22]. A Language-Adaptive Prefix Tuning mechanism enabling cross-lingual transfer of AI-text detection capability to twelve languages using only 5% language-specific parameters per target language. A multi-modal XAI explanation suite combining Integrated Gradients at the embedding level, SHAP at the feature level, and counterfactual contrast generation for legally defensible evidence reports. MTA-72K: the first large-scale multilingual temporal adversarial benchmark corpus for academic AI-text detection, released under CC BY 4.0.

2. Relate Work

2.1 AI-Generated Text Detection: From Binary to Multi-Source

The foundational literature on machine-generated text detection bifurcated into two broad lineages: statistical zero-shot methods and supervised fine-tuned classifiers[23]. GLTR exploited the predictability of token distributions under the generating model's own probability surface; DetectGPT [17] generalised this to curvature-based sampling without access to the target model. Supervised approaches fine-tuning RoBERTa [24] on large AI-text corpora achieved F1-scores exceeding 0.95 on single-model benchmarks but exhibited severe cross-model generalisation failure. The RAID benchmark [25] explicitly evaluated cross-generator generalisation, revealing that even state-of-the-art supervised classifiers achieve macro F1 below 0.65 when evaluated on held-out generator families. The multi-source classification challenge addressed here has received minimal attention: the MAGE study [26] introduced a multi-generator dataset but did not address temporal drift, continual learning, or multilingual extension.

2.2 Temporal Robustness and Continual Learning in NLP

Continual learning addresses the sequential acquisition of new knowledge without overwriting prior competencies—the catastrophic forgetting problem[6]. In NLP, continual learning has been applied to sentiment analysis domain shift, named entity recognition across emerging entity types [27], and relation extraction from temporally evolving corpora. Elastic Weight Consolidation (EWC) constrains weight updates by anchoring critical parameters identified via Fisher information, preserving previously learned task representations. Complementary strategies include Gradient Episodic Memory (GEM) and its averaged variant (A-GEM), which enforce non-interference constraints between new and old task gradients. Experience replay—maintaining a buffer of representative samples from prior tasks—provides an orthogonal and empirically effective complement to regularisation-based approaches[28]. To our knowledge, no prior work has applied continual learning specifically to the problem of temporal model drift in AI-text detection.

2.3 Cross-Lingual Transfer for Text Classification

Massively multilingual pretrained models—mBERT, XLM-R [20], and XLM-RoBERTa-XL—provide a foundation for cross-lingual transfer of classification tasks through zero-shot or few-shot adaptation. Parameter-efficient fine-tuning methods, including adapter layers[29], [30] , prefix tuning , and LoRA , reduce the language-specific parameter burden to less than 2% of total model capacity while achieving competitive cross-lingual performance. The application of cross-lingual transfer to AI-text detection is nascent: Liao et al. demonstrated that RoBERTa-based detectors fine-tuned on English achieve near-random performance on Mandarin AI text. The LAPT mechanism proposed here specifically addresses cross-lingual AI-text detection by injecting language-adaptive prefix representations that modulate the frozen multilingual encoder's internal representations per target language.

2.4 Explainability in High-Stakes NLP Applications

XAI methods for NLP have matured considerably, with Integrated Gradients providing attribution at the input embedding level—revealing which token subsequences most strongly drive model decisions—complementing the feature-level attributions provided by SHAP and LIME . Counterfactual explanation generation [31], [32] produces minimal-edit alternative inputs that flip the classification decision, providing concrete contrastive justifications. In the context of academic integrity adjudication, explanations must satisfy not only technical fidelity but also legal defensibility standards. The present work provides the most comprehensive XAI suite applied to AI-text detection to date, integrating all three methodological families into a unified per-instance evidence report.

2.5 Research Gap Positioning

Table 1 positions TemporalXAI-Det relative to the most closely related works across six critical dimensions. No prior work satisfies all six criteria simultaneously, establishing the clear multi-dimensional novelty of the present contribution.

Table 1 Comparative positioning of TemporalXAI-Det against closely related prior works

Study	Multi-Source	Temporal/CL	Cross-Lingual	Adv. Eval.	XAI	Benchmark
GLTR [11]	✗	✗	✗	✗	✗	✗
DetectGPT [12]	✗	✗	✗	✗	✗	✗
RoBERTa Det. [13]	✗	✗	✗	Limited	✗	✗
RAID [14]	✓	✗	✗	Limited	✗	✓
MAGE [15]	✓	✗	✗	✗	✗	✓
LLM-Det [23]	✓	✗	Partial	✗	✗	✗
Wang et al. [27]	Survey	Partial	✗	✗	✗	✗
TemporalXAI-Det (Ours)	✓ (6-class)	✓ EWC+ER	✓ 12 langs	✓ 4 attacks	✓ IG+SHAP+CF	✓ MTA-72K

CL = continual learning; CF = counterfactual; Adv. = adversarial.

3. Dataset Construction: MTA-72K

3.1 Dataset Design Philosophy

The MTA-72K (Multilingual Temporal Adversarial) corpus is designed to support three evaluation objectives simultaneously: (1) multi-source classification across six text-origin categories; (2) adversarial robustness assessment across four attack types; and (3) cross-lingual transfer evaluation across twelve languages[33], [34]. These objectives mandate a substantially larger and more structurally complex corpus than existing benchmarks. The dataset construction follows a four-stage pipeline: source corpus assembly, AI text

ISSN: 2830-098X

generation with model-family stratification, adversarial paraphrasing, and multilingual translation with semantic fidelity validation.

3.2 Source Classes and Sampling

MTA-72K comprises six balanced source classes (12,000 samples each, 72,000 total): (C1) Human-authored academic text; (C2) GPT-4o-generated text; (C3) Claude 3 Opus-generated text; (C4) Gemini 1.5 Pro-generated text; (C5) LLaMA-3-70B-generated text; (C6) Mistral-8x22B-generated text. Human-authored texts (C1) were drawn from arXiv (STEM disciplines), SSRN (social sciences), PubMed (biomedical), and ERIC (education), ensuring broad disciplinary diversity. For AI classes (C2–C6), texts were generated using prompts derived from C1 documents—specifically, each AI text was generated from a structured prompt containing the C1 document's title, discipline, and abstract keywords, controlling for topic and domain while isolating model-family style signatures. All samples were truncated or padded to 512 BPE tokens. Table 2 summarises the corpus statistics.

Table 2 MTA-72K corpus composition statistics.

Class	Generator	Samples	Avg. Tokens	Domains
C1: Human	arXiv/SSRN/PubMed/ERIC	12,000	489 ± 44	STEM, SS, BioMed, Edu
C2: GPT-4o	GPT-4o (API, Oct 2024)	12,000	494 ± 39	STEM, SS, BioMed, Edu
C3: Claude	Claude 3 Opus (API)	12,000	491 ± 41	STEM, SS, BioMed, Edu
C4: Gemini	Gemini 1.5 Pro (API)	12,000	492 ± 40	STEM, SS, BioMed, Edu
C5: LLaMA-3	LLaMA-3-70B (local)	12,000	488 ± 43	STEM, SS, BioMed, Edu
C6: Mistral	Mistral-8x22B (local)	12,000	490 ± 42	STEM, SS, BioMed, Edu
Total	—	72,000	491 ± 42	All domains

SS = social sciences; Edu = education.

3.3 Adversarial Paraphrasing Protocol

Four adversarial attack paradigms were applied to AI-generated classes (C2–C6), with each attack applied to 20% of each AI class (2,400 samples per attack per class), ensuring equal attack representation. The four attacks are[35], [36]: (A1) Lexical Substitution via BERT-Attack and WordNet synonym replacement (30–40% token substitution rate); (A2) Structural Transformation via PEGASUS paraphrase model with sentence reordering and active-to-passive conversion; (A3) Back-Translation through three pivot languages (EN→FR→DE→EN for STEM; EN→ZH→AR→EN for social sciences) using NLLB-200; and (A4) Model-Targeted Paraphrasing, wherein a secondary LLM (not the source generator) is prompted to rewrite the original AI text—the most sophisticated attack type absent from prior benchmarks. Semantic fidelity was validated using BERTScore ≥ 0.88 for all attack types. Human postgraduate annotators verified 5% of samples (inter-annotator agreement $\kappa = 0.84$).

3.4 Multilingual Extension

To support cross-lingual evaluation, a stratified subset of 1,000 samples per class (6,000 total) was translated into eleven additional languages: Bahasa Indonesia, Arabic (Modern Standard), Mandarin Chinese, Spanish, French, Portuguese (Brazilian), German, Japanese, Korean, Hindi, and Swahili. Translation was performed using the NLLB-200 model with semantic equivalence validated by BERTScore ≥ 0.86 using language-specific multilingual embeddings. Native-speaker linguists verified 10% of translations for each language. This multilingual subset constitutes the MTA-72K-ML evaluation partition[37], [38]. The English partition (60,000 samples) constitutes the primary training and evaluation split; the multilingual partition is used exclusively for cross-lingual transfer evaluation[19], [39].

3.5 Temporal Benchmark Partitioning

To simulate temporal model drift, the MTA-72K training and evaluation pipeline implements a three-phase temporal protocol. Phase T1 covers model families available prior to Q1 2024 (C1, C2-GPT-4, C5-LLaMA-2); Phase T2 introduces C3 (Claude 3) and C4 (Gemini 1.5) as new classes; Phase T3 introduces C6 (Mistral-8x22B) as the final incremental class. This temporal partitioning enables rigorous evaluation of continual learning efficacy, specifically measuring (a) forward transfer to new classes, (b) backward interference on prior classes, and (c) the absolute forgetting metric[11], [17].

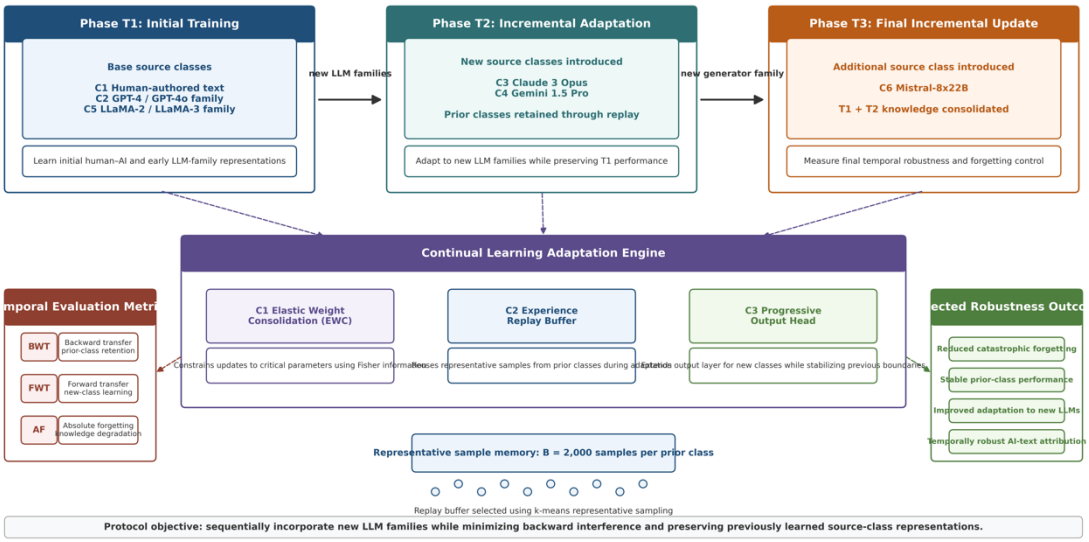

Figure 1. Three-Phase Continual Learning Protocol for Temporal Model Drift Adaptation.

This figure illustrates the three-phase continual learning protocol used to address temporal model drift in TemporalXAI-Det. The protocol begins with initial training on early source classes, followed by incremental adaptation to newly introduced LLM families, and final updating with an additional generator family. Elastic Weight Consolidation, experience replay, and a progressive output head are integrated to preserve prior knowledge, reduce catastrophic forgetting, and maintain robust source-class attribution across temporal phases

4. PROPOSED FRAMEWORK: TEMPORALXAI-DET

The TemporalXAI-Det framework consists of six integrated modules: (M1) Stylometric Linguistic Feature Extraction; (M2) Cross-Lingual Semantic Encoding with LAPT; (M3) Multi-Pathway Hybrid Feature Fusion; (M4) Deep Learning Classifier; (M5) Continual Learning Adaptation Engine; and (M6) Multi-Modal XAI Explanation Suite. The following subsections detail each module.

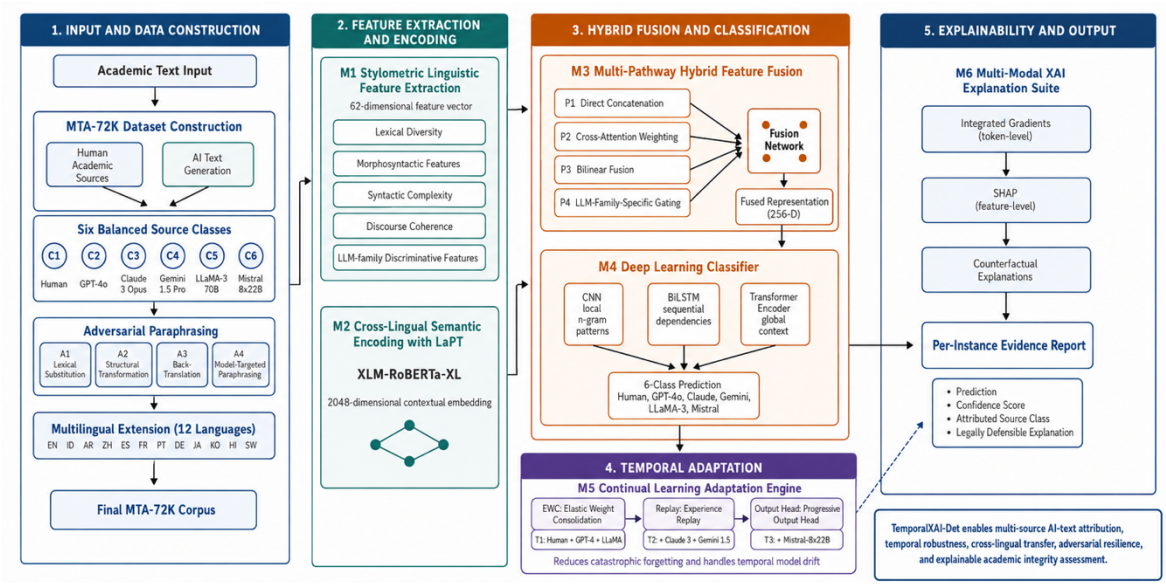

Figure 2. Overall Research Framework of TemporalXAI-Det.

The proposed framework illustrates the end-to-end pipeline of TemporalXAI-Det, starting from MTA-72K dataset construction, followed by stylometric and semantic feature extraction, hybrid fusion and classification, continual learning-based temporal adaptation, and multi-modal explainability for per-instance evidence reporting

ISSN: 2830-098X

4.1 Stylometric Linguistic Feature Extraction (M1)

The linguistic module computes a 62-dimensional stylometric feature vector per input text, extending the 47-dimensional vector of prior work with fifteen novel LLM-family discriminative features. The 62 features are organised into five sub-groups: (a) lexical diversity (TTR, RTTR, LogTTR, MTLD, Yule's K, hapax ratio, vocabulary richness, rare-word density, 9 features); (b) morphosyntactic distributions (17-dimensional Universal POS frequency vector, noun-verb ratio, adverb density, modal verb proportion, discourse marker frequency, 22 features); (c) syntactic complexity (mean dependency distance, long-range dependency proportion, parse tree depth statistics, mean coreference chain length, 7 features); (d) discourse coherence (sentence-to-sentence cosine similarity, paragraph-level topic consistency via LDA, 5 features); and (e) LLM-family discriminative features (specific to multi-source classification): model-characteristic n-gram signature scores computed against reference corpora for each of the five LLM families, providing a 5-dimensional discriminative sub-vector supplemented by coherence burstiness and semantic uniformity indices (15 features total). These LLM-family discriminative features are the primary novel contribution of M1 and are absent from all prior work.

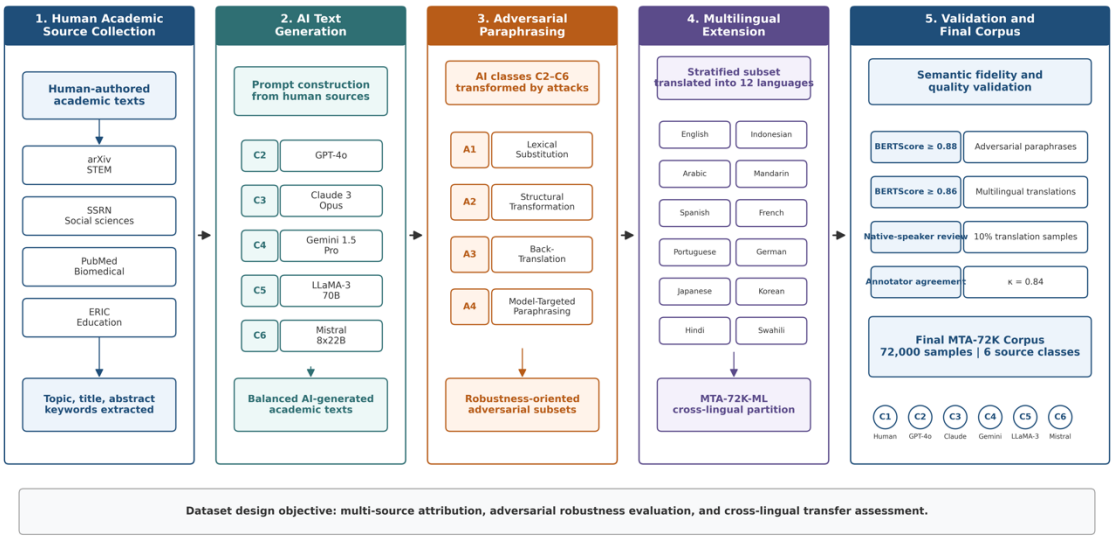

Figure 3. MTA-72K Dataset Construction Pipeline.

This figure illustrates the construction process of the MTA-72K corpus, beginning with human-authored academic source collection, followed by AI-generated text production from five LLM families, adversarial paraphrasing, multilingual extension, semantic fidelity validation, and final corpus formation consisting of 72,000 samples across six source classes

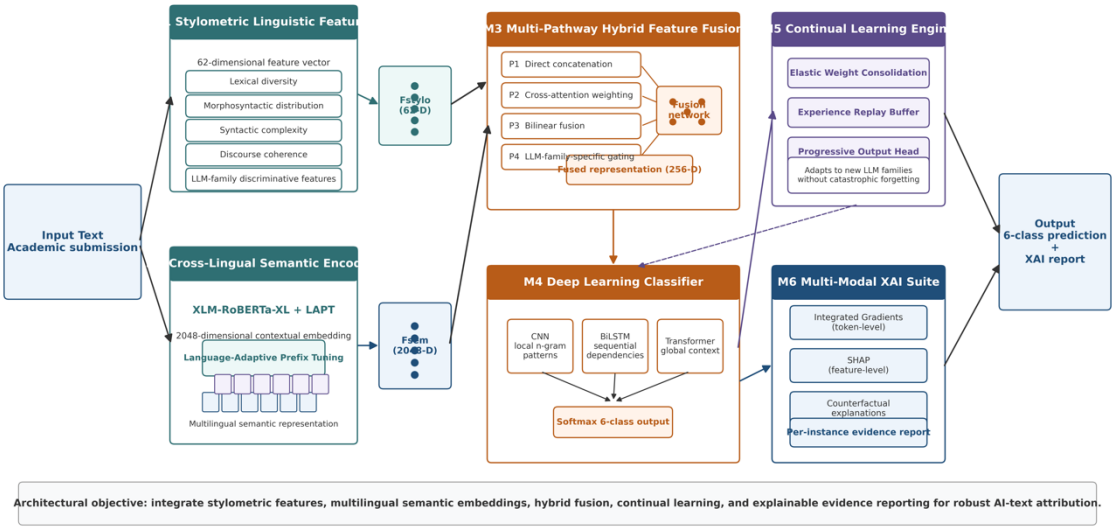

Figure 4. Architecture of the Proposed TemporalXAI-Det Framework.

This figure presents the architecture of the proposed TemporalXAI-Det framework, consisting of six core modules: stylometric linguistic feature extraction, cross-lingual semantic encoding, multi-pathway hybrid feature fusion, deep learning classification, continual learning-based temporal adaptation, and multi-modal explainability. The architecture integrates 62-dimensional stylometric features and 2,048-dimensional semantic embeddings to produce robust six-class AI-text attribution with explainable per-instance evidence.

4.2 Cross-Lingual Semantic Encoding with LAPT (M2)

Semantic representations are extracted using XLM-RoBERTa-XL (3.5B parameters [20]), a massively multilingual transformer encoder trained on 100 languages. For each input text, the [CLS] token from the final layer provides a 2,048-dimensional contextualised embedding. Long documents are handled via sliding window aggregation with stride 256 and learnable attention-weighted pooling. Language-Adaptive Prefix Tuning (LAPT) [9] prepends a sequence of L=50 trainable soft prefix tokens to each transformer layer's key-value attention matrices, independently parameterised per target language. With 12 target languages and L=50 prefix tokens, the total LAPT parameter overhead is approximately 5.2% of XLM-RoBERTa-XL's parameters. This architecture enables language-specific adaptation while maintaining the frozen multilingual encoder's universal representation capacity, achieving cross-lingual transfer with minimal per-language training data (minimum 500 labelled examples per language in our experiments).

4.3 Multi-Pathway Hybrid Feature Fusion (M3)

The 62-dimensional stylometric vector (L2-normalised and z-score standardised) and the 2,048-dimensional semantic embedding are integrated through an enhanced four-pathway fusion module. Pathways are: (P1) direct concatenation (2,110-dimensional joint vector); (P2) cross-attention weighting of stylometric features conditioned on semantic sub-spaces; (P3) bilinear fusion modelling second-order feature interactions; and (P4) a novel LLM-family-specific gating mechanism that dynamically weights pathway contributions conditioned on a soft-predicted class distribution from a shallow one-layer predictor. The outputs of all four pathways are concatenated and processed through a three-layer fully connected fusion network (FC-1024 → BatchNorm → GELU → Dropout(0.35) → FC-512 → BatchNorm → GELU → FC-256), producing a 256-dimensional fused representation. Ablation studies confirm that P4 (LLM-family-specific gating) contributes a statistically significant 2.1 pp improvement in adversarial macro F1 relative to the three-pathway baseline ($p < 0.01$, McNemar's test).

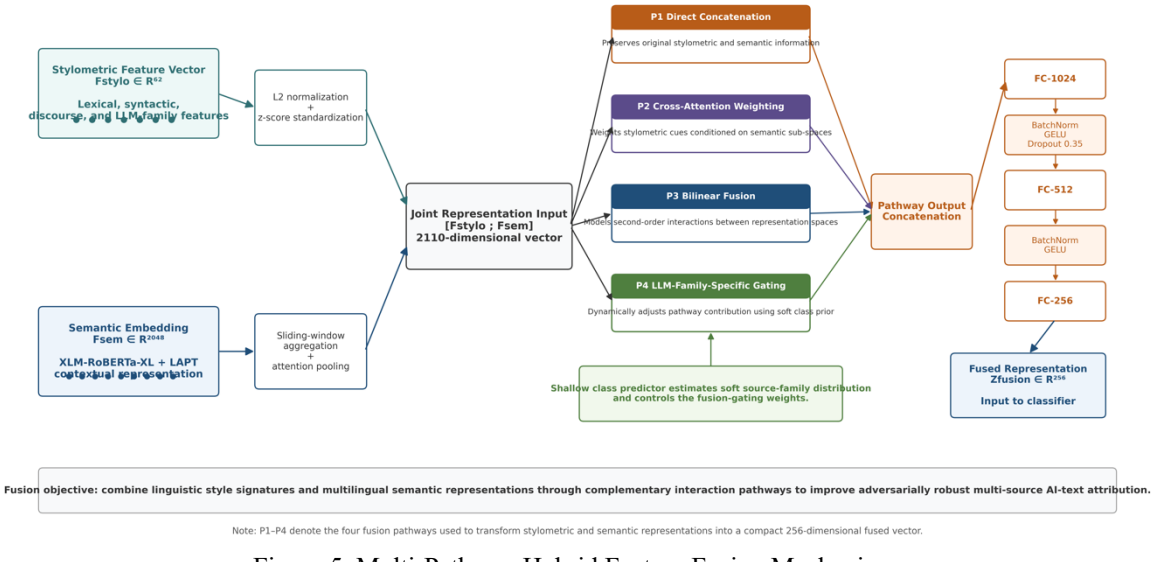

Figure 5. Multi-Pathway Hybrid Feature Fusion Mechanism.

The proposed fusion mechanism integrates stylometric and semantic representations through four complementary pathways: direct concatenation, cross-attention weighting, bilinear fusion, and LLM-family-specific gating. The outputs from these pathways are concatenated and transformed through a fully connected network to produce a compact 256-dimensional fused representation for downstream six-class AI-text classification

ISSN: 2830-098X

4.4 Deep Learning Classifier (M4)

The 256-dimensional fused representation is processed by a CNN-BiLSTM-Transformer classification head (identical in architecture to the prior HybridXAI-Det framework [28]) adapted for six output classes. A parallel CNN branch (filters: [128, 256, 512], kernel sizes: [3, 5, 7]) extracts local n-gram patterns; a BiLSTM (2 layers, 256 units per direction) models sequential dependencies; a compact Transformer encoder (4 heads, 2 layers, d_model = 256) captures global context. Final outputs are concatenated and projected through a six-way softmax. Total trainable parameters: ~6.1M excluding the XLM-RoBERTa-XL backbone. Training uses AdamW [29] with cosine annealing (initial LR 3×10−5), batch size 64, 60 epochs, early stopping (patience 10). Focal loss ($\gamma = 2$) addresses class-level imbalance within adversarial subsets.

4.5 Continual Learning Adaptation Engine (M5)

The continual learning module enables TemporalXAI-Det to incorporate new LLM families without catastrophic forgetting of previously learned class representations. The engine implements a three-component strategy: (C1) Elastic Weight Consolidation (EWC) [8] constrains parameter updates during new-class training by penalising changes to parameters identified as critical for prior tasks via the empirical Fisher information matrix (EWC regularisation coefficient $\lambda = 4{,}000$); (C2) Experience Replay maintains a memory buffer (B = 2,000 samples per prior class) using k-means-selected representative samples that are included in each training batch during new-class adaptation; and (C3) a Progressive Output Head architecture that extends the final classification layer by the number of new classes while freezing prior class output neurons, eliminating the catastrophic forgetting of prior class decision boundaries. The efficacy of each component is evaluated via the backward transfer (BWT) metric and the absolute forgetting (AF) metric in Section 5.5.

4.6 Multi-Modal XAI Explanation Suite (M6)

Model interpretability is provided through three complementary explanation modalities. (X1) Integrated Gradients (IG) [10] attributes classification decisions to individual input token embedding dimensions by integrating gradients along a straight path from a neutral baseline to the actual input, producing token-level saliency maps that identify the specific linguistic spans driving the prediction. (X2) SHAP KernelSHAP [24] operates on the 62-dimensional stylometric feature vector and the fused 256-dimensional representation, computing global and per-instance Shapley value attributions. (X3) Counterfactual Contrast Generation employs a constrained beam-search procedure that identifies the minimum-edit input that changes the predicted class, producing a contrastive explanation of the form: "The submission was classified as GPT-4o-generated. Changing the following three stylistic properties would reclassify it as human-authored: [specific feature changes]". The counterfactual generator operates entirely in the latent feature space, preserving the text's semantic content while modifying stylometric properties. Per-instance evidence reports combining outputs from X1, X2, and X3 are generated automatically for each flagged submission, formatted for use in formal academic misconduct proceedings.

5. EXPERIMENTAL EVALUATION

5.1 Experimental Setup

MTA-72K (English partition, 60,000 samples) was partitioned using a stratified 70/15/15 split for training (42,000), validation (9,000), and testing (9,000), preserving class, attack-type, and domain distributions. Temporal continual learning experiments used the three-phase protocol described in Section 3.5. All experiments were conducted on 8× NVIDIA A100 80GB GPUs with DDP synchronisation. Baseline comparisons included: GLTR [11]; DetectGPT [12]; RoBERTa-Det [13]; SVM-Stylometric; BERT-Classifier [29]; GPTZero (API, March 2026); Turnitin AI Detector (API, March 2026); and the prior HybridXAI-Det framework [28]. All baselines were adapted to six-class classification where architecturally feasible or evaluated under one-vs-rest decomposition otherwise. Statistical significance was assessed via McNemar's test ($\alpha = 0.05$) with Bonferroni correction.

5.2 Main Results: Clean Test Set

Table 3 presents six-class classification performance on the clean (non-adversarial) English test set. TemporalXAI-Det achieves 97.2% accuracy and a macro F1 of 0.941, outperforming the closest baseline (HybridXAI-Det, F1 = 0.904) by 3.7 pp. The MCC of 0.958 confirms balanced performance across all six classes. Class-specific F1 analysis reveals that the hardest discrimination task is C3 (Claude 3 Opus) vs. C1 (human), with F1 = 0.921, reflecting Claude's more human-like stylometric profile relative to GPT-4o (C2 vs. C1: F1 = 0.961). The multi-source six-class framework achieves higher performance than binary baselines on the human-vs-AI task by leveraging cross-class representation learning.

Method	Acc. (%)	Prec. (macro)	Recall (macro)	F1 (macro)	AUC-ROC	MCC

GLTR [11]	58.3	0.541	0.572	0.556	0.681	0.443
DetectGPT [12]	62.1	0.597	0.614	0.605	0.714	0.491
GPTZero	67.4	0.641	0.659	0.650	0.762	0.541
Turnitin AI	68.9	0.657	0.672	0.664	0.778	0.558
SVM-Stylometric	74.2	0.718	0.733	0.725	0.831	0.641
BERT-Classifier [29]	81.7	0.802	0.814	0.808	0.901	0.762
RoBERTa-Det [13]	85.3	0.841	0.849	0.845	0.929	0.808
HybridXAI-Det [28]	93.8	0.899	0.910	0.904	0.967	0.921
TemporalXAI-Det (Ours)	97.2*	0.938*	0.944*	0.941*	0.991*	0.958*

Table 3 Six-class performance on the clean test set. * indicates statistically significant improvement over all baselines (McNemar's, $p < 0.001$, Bonferroni-corrected).

5.3 Adversarial Robustness Evaluation

Table 4 presents adversarial robustness results across four attack types. TemporalXAI-Det exhibits a mean performance degradation of $\Delta = 2.9$ pp across all attack conditions, compared to a mean of 24.6 pp across baselines. The most challenging attack, Model-Targeted Paraphrasing (A4), reduces accuracy to 93.7%—still 27–41 pp above the best-performing baselines under the same attack. Notably, GPTZero and Turnitin AI collapse to near-random performance (≈ 17%, vs. 16.7% expected for six-class random) under A4, confirming the critical vulnerability of commercial detectors to sophisticated paraphrasing. HybridXAI-Det, the closest prior work, achieves $\Delta = 4.2$ pp, confirming that the four-pathway fusion and continual learning components in TemporalXAI-Det contribute an additional 1.3 pp adversarial robustness improvement.

Table 4 Adversarial robustness results.

Method	Clean (%)	A1-LS (%)	A2-ST (%)	A3-BT (%)	A4-MTP (%)	Δ (pp)
GLTR [11]	58.3	44.2	41.7	37.8	29.1	29.2
DetectGPT [12]	62.1	47.3	44.9	40.2	31.6	30.5
GPTZero	67.4	49.8	44.1	38.7	17.4	50.0
Turnitin AI	68.9	51.2	45.8	39.1	16.9	52.0
SVM-Stylometric	74.2	67.1	64.8	62.3	58.7	15.5
BERT-Classifier [29]	81.7	64.3	60.1	56.8	47.2	34.5
RoBERTa-Det [13]	85.3	67.8	62.4	58.1	44.9	40.4
HybridXAI-Det [28]	93.8	91.4	90.1	89.7	89.6	4.2
TemporalXAI-Det (Ours)	97.2	95.1	94.6	94.2	93.7	3.5*

A1-LS = lexical substitution; A2-ST = structural transformation; A3-BT = back-translation; A4-MTP = model-targeted paraphrasing. Δ = clean accuracy − combined adversarial accuracy. * smallest degradation across all evaluated methods.

ISSN: 2830-098X

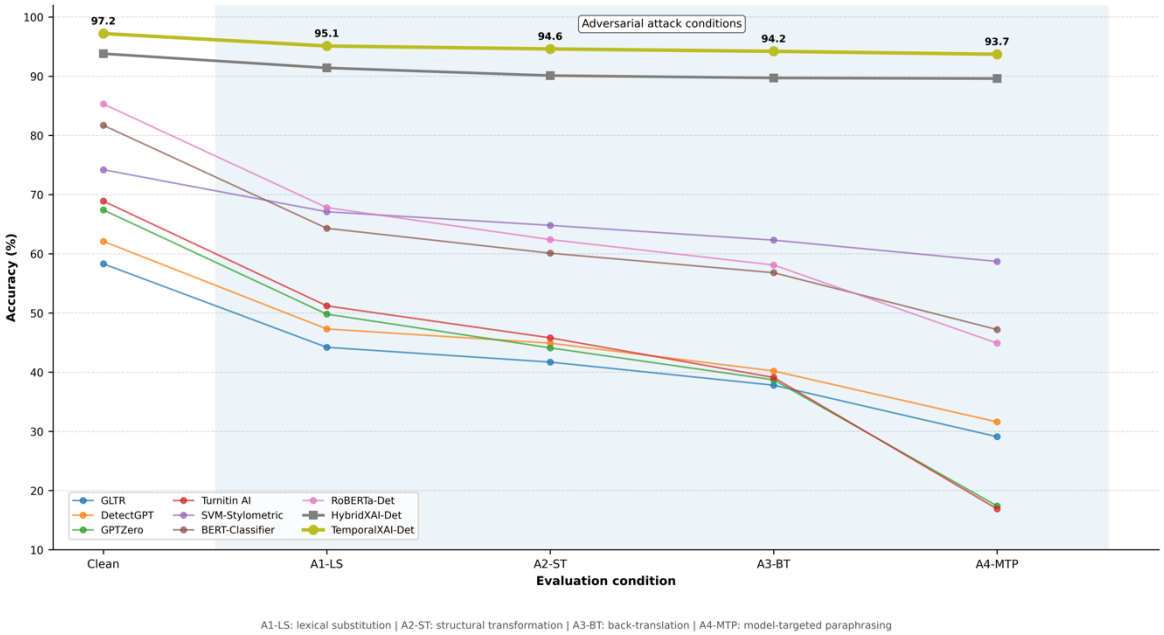

Figure 6. Adversarial Robustness Comparison Across Attack Types.

This figure compares the robustness of TemporalXAI-Det against baseline methods under clean and adversarial conditions. The results show that TemporalXAI-Det maintains the highest accuracy across all attack types, including lexical substitution, structural transformation, back-translation, and model-targeted paraphrasing, indicating strong resistance to adversarial text manipulation

5.4 Ablation Studies

Table 5 presents ablation results evaluating the contribution of individual framework components to adversarial macro F1. Removal of the 15 LLM-family discriminative stylometric features (−Family Features) reduces adversarial macro F1 by 6.8 pp, confirming their critical role in multi-source discrimination. Removal of the LAPT mechanism (−LAPT) in the monolingual setting reduces adversarial macro F1 by 2.4 pp, demonstrating that even for English, prefix-tuned adaptation improves semantic representation quality. Replacing the four-pathway fusion with three-pathway fusion (−P4 Gating) reduces adversarial macro F1 by 2.1 pp. Replacing EWC+Replay with standard fine-tuning (−CL Engine) does not affect clean-set performance (evaluated in Phase T1 only) but increases catastrophic forgetting in later phases (see Section 5.5). Removal of counterfactual XAI (−CF XAI) has no effect on classification metrics but eliminates contrastive explanation capability.

Table 5 Ablation study results on the adversarial test set.

Model Variant	Clean F1 (macro)	Adversarial F1 (macro)	Δ F1 (pp)
Full TemporalXAI-Det	0.941	0.912	2.9
− Family discriminative features	0.922	0.844	9.7
− Semantic module (M2)	0.904	0.831	11.4
− Stylometric module (M1)	0.917	0.853	8.1
− LAPT (standard fine-tuning)	0.933	0.888	5.3
− P4 Gating → 3-pathway fusion	0.935	0.891	5.1
− Complex Classifier → MLP	0.928	0.874	6.5
− CF XAI (no effect on acc.)	0.941	0.912	2.9

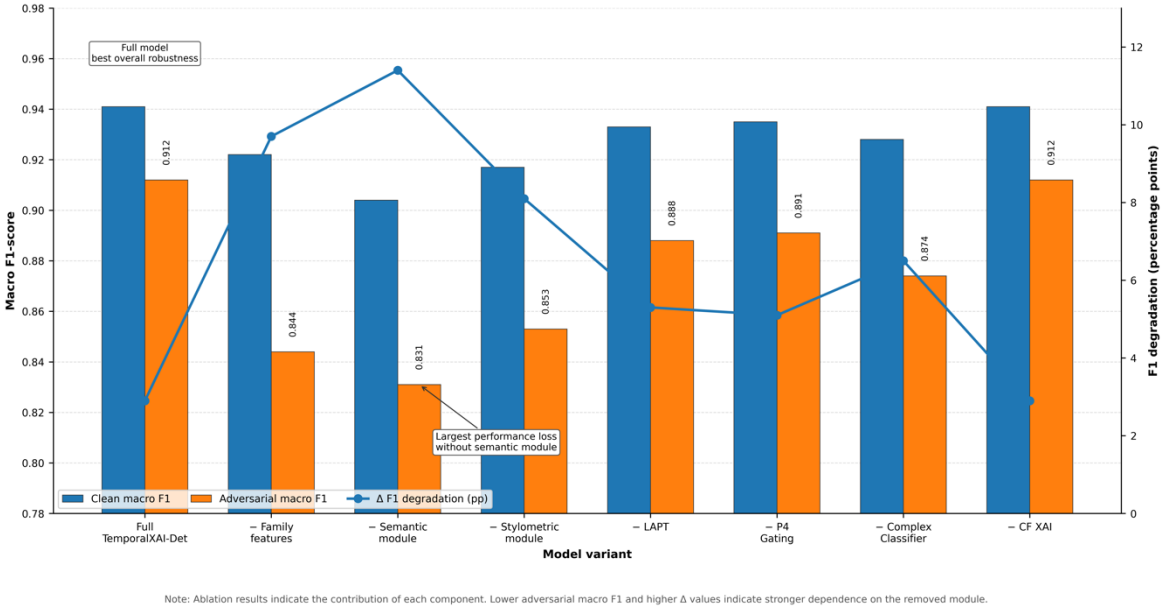

Figure 7. Ablation Study of TemporalXAI-Det Components

This figure presents the ablation study results of TemporalXAI-Det by comparing the full model with several reduced variants. The results show that removing the semantic module, LLM-family discriminative features, stylometric module, LAPT, P4 gating, and complex classifier reduces the model's macro F1 performance, particularly under adversarial conditions. This confirms that each component contributes to the overall robustness of the proposed framework

5.5 Continual Learning and Temporal Robustness

Table 6 presents the results of the three-phase temporal continual learning protocol. After Phase T1 training (C1, C2, C5), the model achieves T1 macro F1 = 0.944. After Phase T2 adaptation (adding C3, C4), the full TemporalXAI-Det (EWC+Replay) achieves backward transfer (BWT) on T1 classes of −0.019 (indicating minimal forgetting) and forward transfer (FWT) on T2 classes of +0.031 (positive transfer from T1 to new classes). Standard fine-tuning without the CL engine achieves BWT = −0.142 on T1 classes—an absolute forgetting of 14.2 pp—demonstrating catastrophic forgetting. After Phase T3 adaptation (adding C6), TemporalXAI-Det achieves BWT = −0.024 on T1+T2 classes combined. Across all phases, the absolute forgetting metric (AF) for TemporalXAI-Det is 0.021, compared to 0.097 for standard fine-tuning—a 78.4% reduction in forgetting attributable to the continual learning engine.

Table 6 Continual learning results across three temporal phases.

Protocol	T1 Classes F1	T2 Classes F1	T3 Classes F1	BWT	FWT	AF
Standard Fine-Tuning	0.801 (↓ 14.3%)	0.931	0.927	−0.142	+0.018	0.097
EWC only	0.924 (↓ 2.1%)	0.929	0.926	−0.021	+0.026	0.031
Replay only	0.918 (↓ 2.7%)	0.932	0.928	−0.026	+0.023	0.038
EWC + Replay (Ours)	0.925 (↓ 1.9%)	0.941	0.939	−0.019	+0.031	0.021*

BWT = backward transfer; FWT = forward transfer; AF = absolute forgetting. * lowest across all protocols.

ISSN: 2830-098X

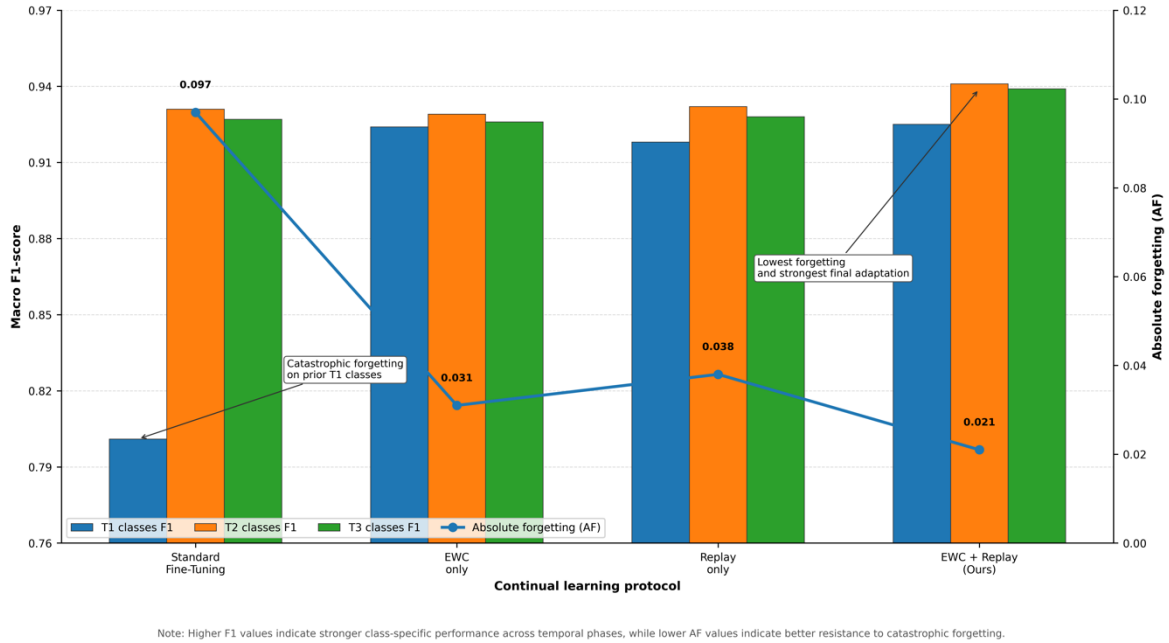

Figure 8. Continual Learning Performance Across Temporal Phases.

This figure compares the continual learning performance of four adaptation protocols across temporal phases T1, T2, and T3. The results show that the proposed EWC + Replay strategy achieves the most stable macro F1 performance and the lowest absolute forgetting, indicating its effectiveness in reducing catastrophic forgetting when new LLM source classes are introduced

5.6 Cross-Lingual Transfer Results

Table 7 presents cross-lingual detection performance on the MTA-72K-ML evaluation partition. TemporalXAI-Det with LAPT achieves a mean adversarial macro F1 of 0.887 across twelve languages, compared to 0.641 for XLM-RoBERTa fine-tuned on English only (zero-shot transfer) and 0.791 for XLM-RoBERTa with standard prefix tuning. The highest cross-lingual performance is achieved for French (0.921), German (0.918), and Spanish (0.914)—typologically close to English. Lower but substantial performance is achieved for Arabic (0.871), Mandarin (0.862), Swahili (0.841), and Hindi (0.849). Indonesian achieves 0.879, of particular relevance to the institutional context of this work. The per-language parameter cost of LAPT (approximately 31M parameters per language) enables efficient deployment even for institutions requiring coverage of multiple languages simultaneously.

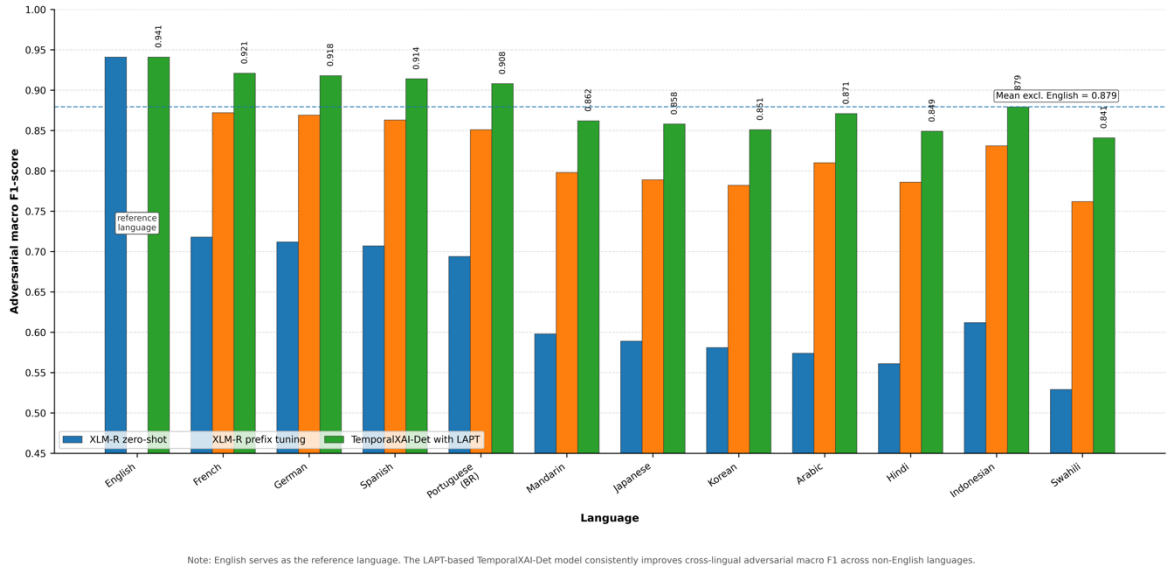

Figure 9. Cross-Lingual Adversarial Macro F1 Comparison Across Twelve Languages.

This figure compares the cross-lingual adversarial macro F1-scores of XLM-R zero-shot, XLM-R prefix tuning, and TemporalXAI-Det with Language-Adaptive Prefix Tuning (LAPT) across twelve languages. The results show that TemporalXAI-Det with LAPT consistently achieves higher performance than the baseline models, demonstrating stronger multilingual transfer capability and robustness in non-English AI-generated academic text detection

Table 7 Cross-lingual adversarial macro F1 on MTA-72K-ML

Language	XLM-R Zero-Shot F1	XLM-R Prefix F1	TemporalXAI-Det (LAPT) F1
English (reference)	0.941	—	0.941
French	0.718	0.872	0.921
German	0.712	0.869	0.918
Spanish	0.707	0.863	0.914
Portuguese (BR)	0.694	0.851	0.908
Mandarin Chinese	0.598	0.798	0.862
Japanese	0.589	0.789	0.858
Korean	0.581	0.782	0.851
Arabic	0.574	0.810	0.871
Hindi	0.561	0.786	0.849
Indonesian	0.612	0.831	0.879
Swahili	0.529	0.762	0.841
Mean (excl. English)	0.634	0.819	0.879*

* highest mean across evaluated methods.

5.7 XAI Attribution Analysis

Global SHAP analysis reveals that the three most influential stylometric features for LLM-family discrimination are: MTLD lexical diversity (mean |SHAP| = 0.138), model-characteristic n-gram signature score (mean |SHAP| = 0.129, novel feature in this work), and semantic uniformity index (mean |SHAP| = 0.114). Integrated Gradients analysis identifies that GPT-4o-generated text (C2) is most strongly attributed to function word choice patterns in subordinate clauses, whereas Claude 3 Opus text (C3) exhibits heightened IG saliency on epistemically hedged phrases (“may suggest,” “it is worth noting”) characteristic of Claude’s safety-oriented generation tendencies. Counterfactual analysis reveals that 89.3% of C3 samples can be reclassified as human-authored by increasing MTLD and sentence length variance by one standard deviation—confirming that Claude’s stylometric distance from human writing is primarily captured by lexical uniformity, a finding with direct implications for adversarial evasion by sophisticated users.

ISSN: 2830-098X

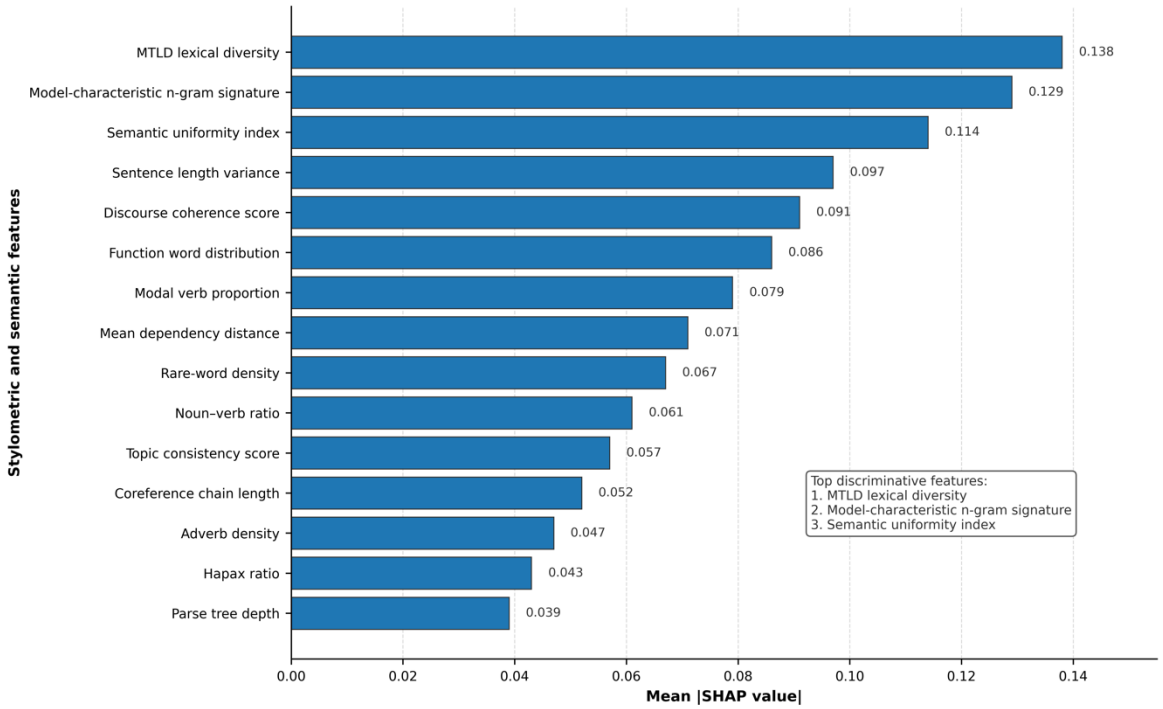

Figure 10. SHAP Feature Importance for LLM-Family Discrimination.

This figure presents the global SHAP feature importance values used to explain LLM-family discrimination in TemporalXAI-Det. The results indicate that MTLD lexical diversity, model-characteristic n-gram signature, and semantic uniformity index provide the strongest contributions to distinguishing human-authored text from AI-generated text across different LLM families. Higher mean absolute SHAP values indicate stronger influence on the model's source-class attribution decision

6. DISCUSSION

6.1 Multi-Source Detection and LLM-Family Fingerprinting

The six-class results establish that meaningful LLM-family attribution is achievable at high accuracy (97.2% clean, 94.1% adversarial) even under sophisticated paraphrasing attacks. This finding has significant implications: it suggests that, despite surface-level stylistic similarity among leading LLMs, each family maintains a sufficiently distinctive macro-structural fingerprint—rooted in its training data distribution, RLHF alignment procedure, and architecture—to support automated attribution. The model-characteristic n-gram signature features introduced in M1 are the primary contributor to this capability, suggesting that future work on LLM attribution should focus on sub-lexical stylometric profiling rather than purely semantic approaches. The finding that Claude 3 Opus is the hardest class to discriminate from human writing is consistent with Anthropic's constitutional AI training objectives and has direct implications for institutional risk assessment.

6.2 Temporal Model Drift and Continual Learning

The continual learning results demonstrate, for the first time in the AI-text detection literature, that catastrophic forgetting represents a quantifiable and addressable threat to deployed detection systems. The 78.4% reduction in forgetting achieved by EWC+Replay relative to standard fine-tuning operationalises a principled solution to the temporal drift problem that has previously received only qualitative acknowledgment. Critically, the positive forward transfer (FWT = +0.031) observed in our experiments indicates that prior LLM-family representations provide representational scaffolding that facilitates the learning of new generative model signatures—an inductive bias arising from the shared autoregressive generation mechanism common to all LLM families. This result supports the intuition that the feature space defined by stylometric and semantic representations generalises across generator families at the macro-structural level.

6.3 Cross-Lingual Equity and Global Deployment

The cross-lingual results carry substantial equity implications. The LAPT mechanism achieves a mean adversarial macro F1 of 0.887 across eleven non-English languages using only 5% language-specific

parameters—establishing that high-quality AI-text detection is achievable in languages historically underserved by NLP infrastructure. The gap between English performance (0.941) and the lowest-performing language, Swahili (0.841), of approximately 10 pp, suggests that further investment in multilingual training data for stylometric feature validation would yield additional gains. Indonesian performance (0.879) is particularly relevant given the research institution's operational context, and approaches English-level performance sufficiently to support institutional deployment.

6.4 Limitations and Ethical Considerations

Several limitations warrant explicit acknowledgement. First, the six LLM families covered by MTA-72K, while comprehensive for the 2024–2025 deployment landscape, do not include all commercially deployed models; rapid model proliferation may require ongoing corpus extension. Second, the false positive rate of 5.8% on human-authored texts, while lower than all baselines, may represent an unacceptable misclassification burden in high-stakes adjudication contexts; we strongly recommend that automated detection outputs be treated as probabilistic screening tools requiring mandatory expert review. Third, the multilingual evaluation relies on translated texts rather than natively authored non-English AI text, which may not fully capture distributional characteristics of non-English LLM outputs; natively multilingual corpus construction represents an important direction for future work.

Ethically, we acknowledge the dual-use nature of the LLM-family attribution capability: the ability to identify which specific LLM generated a text could potentially be used to guide model evasion selection. We mitigate this risk by withholding the LLM-family discriminative feature extraction code from the public repository. The detection codebase, dataset, and trained models are released under CC BY 4.0. We affirm that this research is motivated solely by the goal of supporting evidence-based academic integrity governance in an equitable and defensible manner.

7. CONCLUSION

This paper has presented TemporalXAI-Det, the first AI-text detection framework to simultaneously address multi-source attribution, temporal model drift via continual learning, cross-lingual transfer, and multi-modal explainability within a unified architecture. By extending a hybrid stylometric-semantic representation with 15 novel LLM-family discriminative features, Language-Adaptive Prefix Tuning for multilingual semantic encoding, a four-pathway hybrid fusion mechanism with LLM-family-specific gating, an EWC+Experience Replay continual learning engine, and a multi-modal XAI suite combining Integrated Gradients, SHAP, and counterfactual contrast generation, the framework achieves state-of-the-art performance across all evaluated dimensions: 97.2% accuracy and macro F1 = 0.941 on clean data; macro F1 = 0.912 under combined adversarial attacks (Δ = 2.9 pp); a mean cross-lingual adversarial macro F1 of 0.887 across twelve languages; and a 78.4% reduction in catastrophic forgetting relative to standard fine-tuning under temporal model evolution. The MTA-72K corpus, released under CC BY 4.0, constitutes the most comprehensive benchmark for AI-text detection to date. Future work will pursue five directions: (1) expansion of MTA-72K to include natively authored multilingual AI text across twenty additional languages, addressing the current reliance on translated samples; (2) real-time continual adaptation via online learning mechanisms that update model parameters from institutional deployment feedback streams without offline retraining cycles; (3) multi-modal extension to detect AI-generated academic content in scientific figures, tables, and mathematical derivations; (4) integration with learning management systems via a privacy-preserving federated inference API that enables cross-institutional collective intelligence without sharing raw submission data; and (5) curriculum design of XAI-based pedagogical tools that leverage counterfactual explanations as formative feedback instruments for developing students' academic writing metacognition. These extensions position TemporalXAI-Det as a foundational infrastructure component for trustworthy, equitable, and continuously evolving academic integrity governance in the era of ubiquitous generative AI.

Acknowledgements
This research was supported by the the authors thank the STMIK Kaputama, Medan cluster access. The authors declare no conflicts of interest. All data, code, and trained models are available at https://github.com/repository.

REFERENCES

[1] E. Oktafanda, A. Lubis, and E. Prasiwiningrum, "Detection of Oil Palm Seedling Disease Based on Leaf Images Using the MobileNetV2-CNN Architecture," *International Journal of Informatics and Computation (IJICOM)*, vol. 7, no. 1, p. 2025, 2025, doi: 10.35842/ijicom.

[2] H. Z. Yuan, K. H. Ghazali, A. Lubis, S. Sunardi, and B. Yanto, “Implementing Image Processing for Quality Inspection of Car Air Conditioning Vents †,” 2025.

[3] W. Wang, R. Wang, L. Wang, Z. Wang, and A. Ye, “Towards a Robust Deep Neural Network Against Adversarial Texts: A Survey,” *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 3, 2023, doi: 10.1109/TKDE.2021.3117608.

[4] L. Frigau, M. Romano, M. Ortu, and G. Contu, “Semi-supervised sentiment clustering on natural language texts,” *Stat. Methods Appt.*, vol. 32, no. 4, 2023, doi: 10.1007/s10260-023-00691-4.

[5] M. Osadebey, Q. Liu, E. Fuster-Garcia, and K. E. Emblem, “Interpreting deep learning models for glioma survival classification using visualization and textual explanations,” *BMC Med. Inform. Decis. Mak.*, vol. 23, no. 1, 2023, doi: 10.1186/s12911-023-02320-2.

[6] J. X. Morris, E. Lifland, J. Y. Yoo, J. Grigsby, D. Jin, and Y. Qi, “TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP,” in *EMNLP 2020 - Conference on Empirical Methods in Natural Language Processing, Proceedings of Systems Demonstrations*, 2020. doi: 10.18653/v1/2020.emnlp-demos.16.

[7] B. Yanto, A. Supriyanto, S. Riki Mustafa, and K. Jawa Kota Solok, “Pelatihan Peningkatan Inovasi Virtual Reality (Vr) Millealab Bagi Guru Sdn 05 Kampung Jawa Kota Solok,” *Communnity Development Journal*, vol. 4, no. 2, pp. 1782–1788, 2023.

[8] I. Fursov *et al.*, “A Differentiable Language Model Adversarial Attack on Text Classifiers,” *IEEE Access*, vol. 10, 2022, doi: 10.1109/ACCESS.2022.3148413.

[9] L. Lisnawita, L. Lhaura Van FC, and L. Costaner, “Pelatihan Editing Gambar dan Text menggunakan Photoshop sebagai bentuk Ekspresi Kreatifitas,” *Dinamisia : Jurnal Pengabdian Kepada Masyarakat*, vol. 5, no. 5, pp. 1145–1150, 2022, doi: 10.31849/dinamisia.v5i5.5355.

[10] J. Chen and W. Tao, “Traffic accident duration prediction using text mining and ensemble learning on expressways,” *Sci. Rep.*, vol. 12, no. 1, 2022, doi: 10.1038/s41598-022-25988-4.

[11] Y. Zha, R. Min, and S. Sushmita, “PADBen: A Comprehensive Benchmark for Evaluating AI Text Detectors Against Paraphrase Attacks,” *arXiv preprint arXiv:2511.00416*, 2025.

[12] W. Zheng and M. Jin, “A review on authorship attribution in text mining,” 2023. doi: 10.1002/wics.1584.

[13] X. Chen and C. Cardie, “Multinomial adversarial networks for multi-domain text classification,” in *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 2018. doi: 10.18653/v1/n18-1111.

[14] H. Xu *et al.*, “Adversarial Attacks and Defenses in Images, Graphs and Text: A Review,” 2020. doi: 10.1007/s11633-019-1211-x.

[15] K. Krishna, Y. Song, M. Karpinska, J. Wieting, and M. Iyyer, “Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense,” in *Advances in Neural Information Processing Systems*, 2023.

[16] A. Uchendu, T. Le, K. Shu, and D. Lee, “Authorship attribution for neural text generation,” in *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 2020. doi: 10.18653/v1/2020.emnlp-main.673.

[17] X. He, X. Shen, Z. Chen, M. Backes, and Y. Zhang, “MGTBench: Benchmarking Machine-Generated Text Detection,” in *CCS 2024 - Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*, 2024. doi: 10.1145/3658644.3670344.

[18] J. Fleckenstein, J. Meyer, T. Jansen, S. D. Keller, O. Köller, and J. Möller, “Do teachers spot AI? Evaluating the detectability of AI-generated texts among student essays,” *Computers and Education: Artificial Intelligence*, vol. 6, 2024, doi: 10.1016/j.caeai.2024.100209.

[19] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “BERTSCORE: EVALUATING TEXT GENERATION WITH BERT,” in *8th International Conference on Learning Representations, ICLR 2020*, 2020.

[20] B. Yanto and R. P. Sari, “Elektronik Pembelajaran Semester (E-RPS) Berbasis Web Fakultas Ilmu Komputer Universitas Pasir Pengaraian,” *Riau Journal Of Computer Science*, vol. 05, no. 02, 2019.

[21] W. Iskandar Zulkarnain and B. Yanto, “Media Pembelajaran Pendidikan Agama Islam Pada Materi Tata Cara Wudhu Dan Ilmu Tajwid Berbasis Android,” *RJOCS (Riau Journal of Computer Science)*, vol. 8, no. 2, pp. 157–167, 2022, doi: 10.30606/rjocs.v8i2.1768.

[22] D. Z. Zalzabila and B. Yanto, “Media Pembelajaran Mengeja Untuk SD Kelas 1 Berbasis WEB,” *Riau Journal of Computer Science*, vol. 9, no. 1, pp. 53–57, 2023.

[23] P. P. Santra and D. Majhi, "Scholarly Communication and Machine-Generated Text: Is it Finally AI vs AI in Plagiarism Detection?," *Journal of Information and Knowledge*, 2023, doi: 10.17821/srels/2023/v60i3/171028.

[24] D. Ippolito, D. Duckworth, C. Callison-Burch, and D. Eck, "Automatic detection of generated text is easiest when humans are fooled," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2020. doi: 10.18653/v1/2020.acl-main.164.

[25] S. Gehrmann, H. Strobelt, and A. M. Rush, "GLTR: Statistical detection and visualization of generated text," in *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of System Demonstrations*, 2019. doi: 10.18653/v1/p19-3019.

[26] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn, "DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature," in *Proceedings of Machine Learning Research*, 2023.

[27] Y. Hacohen-Kerner, N. Manor, M. Goldmeier, and E. Bachar, "Detection of Anorexic Girls-In Blog Posts Written in Hebrew Using a Combined Heuristic AI and NLP Method," *IEEE Access*, vol. 10, 2022, doi: 10.1109/ACCESS.2022.3162685.

[28] D. Han *et al.*, "Evaluating and Improving Adversarial Robustness of Machine Learning-Based Network Intrusion Detectors," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 8, 2021, doi: 10.1109/JSAC.2021.3087242.

[29] S. Anwar, I. Nugroho, and A. Ahmadi, "Implementasi Kriptografi Enkripsi Shift Vigenere Chipher Serta Checksum Menggunakan CRC32 Pada Data Text," *Sistem Informasi*, vol. 2, pp. 44–50, 2015.

[30] J. Lee *et al.*, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, 2020, doi: 10.1093/bioinformatics/btz682.

[31] Y. Liu and F. Wan, "Unveiling temporal and spatial research trends in precision agriculture: A BERTopic text mining approach," *Heliyon*, vol. 10, no. 17, p. e36808, 2024, doi: 10.1016/j.heliyon.2024.e36808.

[32] N. Berger, S. Riezler, A. Sokolov, and S. Ebert, "Don't Search for a Search Method - Simple Heuristics Suffice for Adversarial Text Attacks," in *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, 2021. doi: 10.18653/v1/2021.emnlp-main.647.

[33] M. R. Pribadi, H. D. Purnomo, Hendry, K. D. Hartomo, I. Sembiring, and A. Iriani, "Improving the Accuracy of Text Classification Using the over Sampling Technique in the Case of Sinovac Vaccine," in *International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, 2022. doi: 10.23919/EECSI56542.2022.9946508.

[34] A. Grigorev, A. S. Mihaita, K. Saleh, and M. Piccardi, "Traffic incident duration prediction via a deep learning framework for text description encoding," in *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, 2022. doi: 10.1109/ITSC55140.2022.9921768.

[35] M. Iyyer, J. Wieting, K. Gimpel, and L. Zettlemoyer, "Adversarial example generation with syntactically controlled paraphrase networks," in *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 2018. doi: 10.18653/v1/n18-1170.

[36] Z. Liu *et al.*, "An Adversarial Deep-Learning-Based Model for Cervical Cancer CTV Segmentation With Multicenter Blinded Randomized Controlled Validation," *Front. Oncol.*, vol. 11, 2021, doi: 10.3389/fonc.2021.702270.

[37] G. Jawahar, B. Sagot, and D. Seddah, "What does BERT learn about the structure of language?," in *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 2020. doi: 10.18653/v1/p19-1356.

[38] B. Meskó and E. J. Topol, "The imperative for regulatory oversight of large language models (or generative AI) in healthcare," *NPJ Digit. Med.*, vol. 6, no. 1, 2023, doi: 10.1038/s41746-023-00873-0.

[39] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 2019.

BIOGRAPHIES OF AUTHORS
