Typology-Aware Dual-Track: Detecting Hybrid AI Text Across Languages

Kihoon Park[1] **and Seungwoo Lim**[1] **and Jihie Kim**[1]*
[1]Department of Computer Science and Artificial Intelligence, Dongguk University
kihoon34@dongguk.edu daniel333@dgu.ac.kr jihie.kim@dgu.edu

Abstract

Existing generalist detectors, such as mDeBERTa based classifiers, struggle with fine-grained AI-text detection in multilingual settings, particularly when hybrid authorship and morphologically rich languages obscure machine-generated artifacts. We propose a typology-aware Smart Routing & Late Fusion Framework for 4-class detection: Human-Written (HW), Machine-Generated (MG), Machine-Humanized (MH), and Human-Polished (HM). The framework consists of two complementary tracks. Branch A targets agglutinative languages, including Korean, Turkish, and Hungarian, using error focused hard-negative refinement and pairwise calibration based on Dependency Distance Minimization (DDM) (Ferrer-i-Cancho and Gómez-Rodríguez, 2021) and Uniform Information Density (UID) (Meister et al., 2021) to recover weak generative signals hidden by dense morphology. Branch B applies RADAR-inspired invariant learning to capture machine-originated artifacts that persist after adversarial rewriting in languages with clearer segmentation cues. Experiments across eight languages show that typology-aware routing improves detection in agglutinative languages while maintaining strong multilingual performance. Our results highlight the importance of modeling both linguistic typology and hybrid authorship for robust multilingual AI-text detection.

1 Introduction

The widespread use of large language models (LLMs) has transformed text generation from a fully human-authored process into a spectrum of hybrid authorship. While early AI-text detectors mainly focused on binary classification between Human-Written (HW) and Machine-Generated (MG) text, real-world writing now includes more ambiguous cases: machine-generated text rewritten to appear human-like (Machine-Humanized; MH) and human-written text polished by LLMs for grammar or fluency (Human-Polished; HM). These hybrid categories blur conventional authorship boundaries and make surface level detection increasingly unreliable.

This challenge becomes more severe in multilingual settings, especially for morphologically rich and agglutinative languages (Wu et al., 2025a). Generalist multilingual encoders such as mDeBERTa often perform well in languages with clearer segmentation cues, but they can miss subtle generative artifacts in languages such as Korean, Turkish, and Hungarian, where functional morphemes and affixes are densely attached to lexical roots. This typological mismatch suggests that a single one-size-fits-all detector is insufficient for fine-grained multilingual AI text detection.

To address this problem, we propose a typology-aware Smart Routing & Late Fusion Framework for 4-class AI text detection. The framework consists of two complementary branches. Branch A serves as an agglutinative recovery track for Korean, Turkish, and Hungarian, using hard-negative refinement and Dependency Distance Minimization(DDM)(Ferrer-i-Cancho and Gómez-Rodríguez, 2021) / Uniform Information Density(UID)(Meister et al., 2021) based pairwise calibration to improve separation between confusable hybrid classes. Branch B serves as a global fortification track for languages such as English, Spanish, Russian, Chinese, and Japanese, using RADAR-inspired invariant learning (Hu et al., 2023) to detect machine originated artifacts that persist after adversarial rewriting. By routing each language to the branch best suited to its typological properties, the framework improves robustness without forcing all languages into a single representation strategy.

We evaluate the proposed framework across eight typologically diverse languages, covering

* Corresponding author: jihie.kim@dgu.edu
Dataset and code available at: https://github.com/kihoon-park/TADT

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 17150–17167
October 24-29, 2026 ©2026 Association for Computational Linguistics

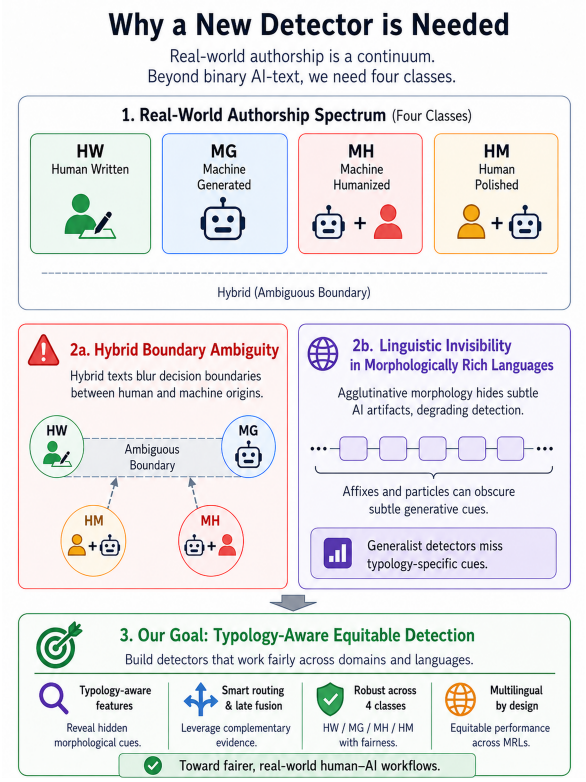


Figure 1: Motivation of the proposed multilingual hybrid authorship detection task. Real-world AI assisted writing forms a 4-class authorship problem, while hybrid boundary ambiguity and linguistic invisibility in morphologically rich languages challenge existing generalist detectors.

Indo-European languages with relatively explicit segmentation cues, East Asian languages with non-Latin scripts, and agglutinative languages with dense morphological marking. Our results show that the proposed approach substantially improves detection in agglutinative languages while maintaining strong performance in languages with clearer segmentation cues. The analysis further shows that detection difficulty is shaped not only by model architecture, but also by linguistic structure and domain specific writing conventions.

Our main contributions are summarized as follows.

Provenance Aware Hybrid Authorship Detection: Building on prior studies of mixed and fine-grained authorship, we formulate a provenance-aware 4-class task comprising HW, MG, MH, and HM. Unlike settings that merge heterogeneous revisions into a single mixed class or focus on LLM continuation and involvement intensity, our formulation preserves both the original authorship source and the direction of transformation, explicitly distinguishing adversarial machine humanization from AI-assisted polishing of human text.

Typology-Aware Dual-Track Framework: We propose a Smart Routing framework that jointly addresses provenance-related hybrid ambiguity and language specific typological variation. Branch A recovers weak generative signals in morphologically rich languages through hard-negative refinement and DDM/UID based calibration, while Branch B captures machine-originated artifacts that remain invariant under adversarial rewriting. Through systematic routing swap ablations, we empirically prove that deterministic track separation is indispensable to prevent toxic cross track logit contamination.

State-of-the-Art Multilingual Benchmark & Evaluation: We construct a balanced 4-class benchmark of 32,000 samples across eight typologically diverse languages and multiple domains. Our framework substantially outperforms existing generalist and stylized baselines, achieving a new state-of-the-art overall Macro-F1 of **0.9105** (and **0.8985** on challenging agglutinative scripts), bridging the cross lingual equity gap in AI text detection.

2 Related Work

2.1 From Statistical Heuristics to Neural Detection

Early paradigms in AI text detection relied heavily on the statistical signatures left by decoding strategies. Approaches such as GLTR (Gehrmann et al., 2019) and DetectGPT (Mitchell et al., 2023) leveraged zero-shot statistical artifacts, including perplexity, entropy, curvature, and log rank information, based on the hypothesis that LLMs sample tokens from predictable regions of probability distributions. While these methods offered interpretability without task specific training, they struggled with the increasing fluency and diversity of modern LLM outputs (Yang et al., 2023).

Following the advent of Transformers, the field shifted toward supervised fine-tuning of Pretrained Language Models (PLMs). Architectures such as RoBERTa (Liu et al., 2019) and DeBERTa (He et al., 2021) established strong benchmark performance by treating detection as a binary classification problem. Large scale corpora such as HC3 (Guo et al., 2023) and TuringBench (Uchendu et al., 2021) facilitated this transition, allowing neural detectors to learn subtle stylistic and distribu-

tional differentiators. Despite their effectiveness on standard benchmarks, these approaches predominantly assume a strict dichotomy of authorship, optimizing for raw machine generation versus pure human writing. This assumption becomes restrictive in real-world settings where AI-generated texts may be rewritten, and human-authored texts may be polished by LLM-based writing assistants.

2.2 Hybrid Authorship and Adversarial Obfuscation

As human-AI collaboration intensifies, the boundary between human and machine authorship has become increasingly porous. The CoAuthor dataset (Lee et al., 2022) highlighted the difficulty of detecting AI contributions at fine-grained textual spans due to frequent author switching. Similarly, MixRevDetect (Kumar et al., 2025) demonstrated that detectors can fail significantly when analyzing mixed-authorship sentences in peer reviews, where human intent and machine-assisted rewriting coexist within the same document.

Zhang et al. (2024) study mixed human-AI authorship through the MIXSET benchmark. Although their data construction distinguishes multiple revision operations, including AI polishing or rewriting of human text and human adaptation of machine-generated text, their final three-class setting groups these heterogeneous transformations into a single *mixtext* class alongside Human-Written Text (HWT) and Machine-Generated Text (MGT). Thus, AI-revised human text and human-revised machine text are not retained as separate final classes. Cheng et al. (2025) instead characterize LLM participation through LLM Role Recognition and LLM Involvement Measurement, distinguishing Human-Author, LLM-Creator, LLM-Polisher, and LLM-Extender. Our taxonomy is organized around original provenance and transformation direction: HW and MG closely correspond to Human-Author and LLM-Creator, and HM parallels LLM-Polisher, whereas MH explicitly captures adversarial humanization of machine-originated text and has no direct counterpart in Cheng et al.; conversely, LLM-Extender is outside our label space. Our formulation therefore explicitly separates the MG-to-MH and HW-to-HM transformation directions rather than merging them into a broad mixed class or modeling continuation as a separate role.

Robustness against adversarial obfuscation remains a central challenge. Studies on paraphrase attacks (Krishna et al., 2023) show that relatively minor perturbations can evade detection, while polishing human text can reduce syntactic perplexity and increase false positives. Large-scale robustness benchmarks such as RAID (Dugan et al., 2024) further demonstrate the fragility of current detectors under editorial modifications. These findings motivate treating transformation direction as an explicit detection target rather than collapsing heterogeneous forms of hybrid authorship into a single category.

2.3 Multilingual Generalization and Linguistic Bias

Existing detection frameworks exhibit a pronounced Anglocentric bias (Serikov et al., 2022), with evaluation and training resources concentrated in English and other high resource Indo-European languages. This bias impedes generalization to languages with divergent scripts, syntactic structures, and tokenization behaviors, such as East Asian and morphologically rich agglutinative languages. In Korean, Turkish, and Hungarian, functional morphemes and affixes are densely attached to lexical roots. As a result, subtle generative artifacts may be absorbed into complex surface forms and become less visible to standard subword tokenizers, which segment text by frequency rather than linguistic boundaries.

Although multilingual models such as mDeBERTa improve cross-lingual coverage, they can still suffer from typological mismatch and the "curse of multilinguality," where performance in specific language groups degrades due to vocabulary dilution or insufficient representation capacity (Gurgurov et al., 2024). Recent studies therefore emphasize broader evaluation beyond English and Latin script languages. Our work addresses this limitation by introducing a typology-aware framework that separates agglutinative recovery from global adversarial fortification. Branch A focuses on morphologically rich languages through hard-negative refinement, with DDM/UID based linguistic calibration serving as an auxiliary correction, while Branch B uses RADAR-inspired invariant learning to capture machine-originated artifacts that persist under adversarial rewriting. In doing so, our framework moves beyond simple multilingual fine-tuning and accounts for both hybrid authorship and typological variation.

3 Methodology

We present a typology-aware dual-track framework for fine-grained multilingual AI text detection. The framework addresses two complementary failure modes: (i) *typological invisibility*, where generative artifacts are obscured by the morphological complexity of agglutinative languages, and (ii) *adversarial fluency*, where machine-generated texts are rewritten or polished to mimic human-authored writing. Branch A targets morphologically rich languages such as Korean, Turkish, and Hungarian, while Branch B targets invariant machine-generated artifacts in languages with more explicit segmentation cues, such as English, Spanish, and Russian.

3.1 Problem Formulation: The 4-Class Taxonomy

We formulate AI text detection as a 4-class classification problem. The label space $\mathcal{Y}$ is defined as follows:

Human-Written (HW; C0): Original text authored by humans without AI assistance.

Machine-Generated (MG; C1): Text generated directly by LLMs based on prompts.

Machine-Humanized (MH; C2): MG text rewritten by LLMs using adversarial humanization instructions.

Human-Polished (HM; C3): HW text refined by LLMs for grammar, style, or fluency.

This taxonomy isolates the effects of generation, humanization, and polishing. The most challenging boundaries arise between transformation related classes such as MG–MH and HW–HM, as well as between fluent hybrid classes such as MH and HM.

3.2 Typology-Aware Dual-Track Architecture

Figure 3 illustrates the overall workflow of the proposed framework. The input text is first assigned to a typological route according to its language cluster. Agglutinative languages such as Korean, Turkish, and Hungarian are processed through Branch A, which focuses on recovering weak generative signals hidden by morphological complexity. Languages with clearer segmentation cues are processed through Branch B, which uses RADAR-inspired contrastive learning to capture invariant machine-generated artifacts. The final prediction is produced either from the routed branch or, when both sources of evidence are available, through late fusion of branch level logits.

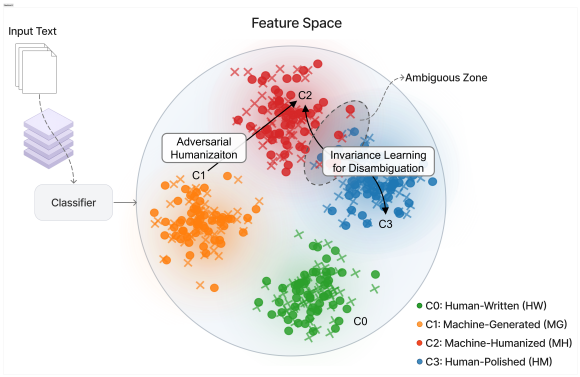


Figure 2: Conceptual visualization of the 4-class latent space and the overlap between hybrid classes.

The framework routes each language cluster to a branch specialized for its typological characteristics. Branch A focuses on agglutinative languages, where generative artifacts may be hidden by complex affixation. Branch B focuses on global multilingual fortification, where invariant machine-generated artifacts can be learned through RADAR inspired contrastive training.

Each branch produces a 4-dimensional logit vector:

$$z = f(x) \in \mathbb{R}^4,$$

where each dimension corresponds to HW, MG, MH, and HM. The prediction is:

$$\hat{y} = \arg\max_{c \in \mathcal{Y}} z_c.$$

3.2.1 Branch A (Recovery): Agglutinative Languages

Branch A recovers weak generative signals in Morphologically Rich Languages (MRLs), focusing on Korean, Turkish, and Hungarian. In these languages, dense affixation can absorb stylistic or statistical irregularities into complex surface forms, reducing the visibility of LLM artifacts.

Base Detector and Hard-Negative Refinement: We first train an mDeBERTa based 4-class detector for each agglutinative language using class balanced sampling, class weighted cross entropy, label smoothing, and validation based checkpoint selection. After base training, we identify misclassified training instances and assign them higher sample weights. We then apply hard-negative refinement to empirically confusable class pairs, including HW–HM, MG–MH, and HM–MH. This discourages excessive representational similarity between classes

Class	Category	Text Example
C0	HW	"Seven months ago, America and the world drew a line in the sand. We declared that the aggression against Kuwait would not stand."
C1	MG	"Sprinkle with almonds and salt to taste."
C2	MH	"**Add a sprinkle** of almonds and a **pinch of salt** to taste."
C3	HM	"...America and the world drew a line in the sand, **declaring** that the aggression against Kuwait would not stand."

Table 1: Examples of the 4-class taxonomy.

that are stylistically close but must remain separable.

Auxiliary Linguistic Calibration: To provide a lightweight linguistic correction, Branch A optionally uses DDM and UID features as calibration signals. DDM measures syntactic compactness by quantifying how far syntactically related words are separated in a sentence. Texts with unusually short and repetitive dependency structures may indicate syntactic monotony, which is frequently observed in LLM-generated or LLM-polished outputs. UID captures how evenly information is distributed across a sentence using token level surprisal statistics. Human writing often contains locally uneven information flow, whereas machine-generated text tends to exhibit smoother and more regular information distribution. In our framework, these linguistic features are not used as standalone classifiers. Instead, they serve as auxiliary calibration signals for class pairs that are empirically difficult to separate, such as HW–HM, MG–MH, and MH–HM. A lightweight calibrator predicts pairwise logit corrections for these ambiguous boundaries. Since the gain from this module is modest, we use it as a targeted correction mechanism rather than the main recovery component. Concrete examples of DDM and UID feature extraction are provided in Appendix D.3.

3.2.2 Branch B (Fortification): RADAR-Inspired Global Generalist

Branch B serves as the global fortification module for languages with more explicit segmentation cues, including English, Spanish, Russian, Chinese, and Japanese. While Branch A addresses artifacts hidden by morphology, Branch B captures invariant machine-generated artifacts that persist after paraphrasing, humanization, or polishing.

RADAR-Inspired Invariant Learning: Following RADAR, we assume that adversarially rewritten machine text still preserves origin specific artifacts. We therefore use supervised contrastive learning over h_i to keep machine-originated classes (MG, MH) separable from human-originated classes (HW, HM), encouraging Branch B to focus on invariant artifacts rather than surface fluency.

Branch B uses a multilingual mDeBERTa backbone trained with a RADAR-inspired invariant learning objective. Given an input x_i, the model produces class logits z_i and a latent representation h_i. The training objective combines class weighted cross entropy and supervised contrastive learning:

$$\mathcal{L}_B = \mathcal{L}_{CE}^{w} + \lambda_{con}\mathcal{L}_{con}.$$

The contrastive component encourages machine-originated samples to remain distinguishable from human-originated samples even after adversarial rewriting.

Hybrid Class Fortification: The main challenge in Branch B is separating Machine-Humanized (MH) from Human-Polished (HM). Both are fluent, but their origins differ. As illustrated in Figure 4, Branch B emphasizes invariant artifacts and hybrid class hard negatives such as MG–MH and MH–HM, encouraging the model to focus on persistent machine-generated signals rather than surface fluency. This allows the branch to function as a multilingual generalist detector for adversarially rewritten text.

3.3 Smart Routing and Late Fusion

The two branches address complementary failure modes. Agglutinative languages are routed to Branch A, while languages with clearer segmentation cues are routed to Branch B. When both branches are applicable, we combine their logits through late fusion:

$$z_{final} = \alpha z_A + (1-\alpha) z_B,$$

where α is selected on the validation set. This allows the model to combine recovery oriented and

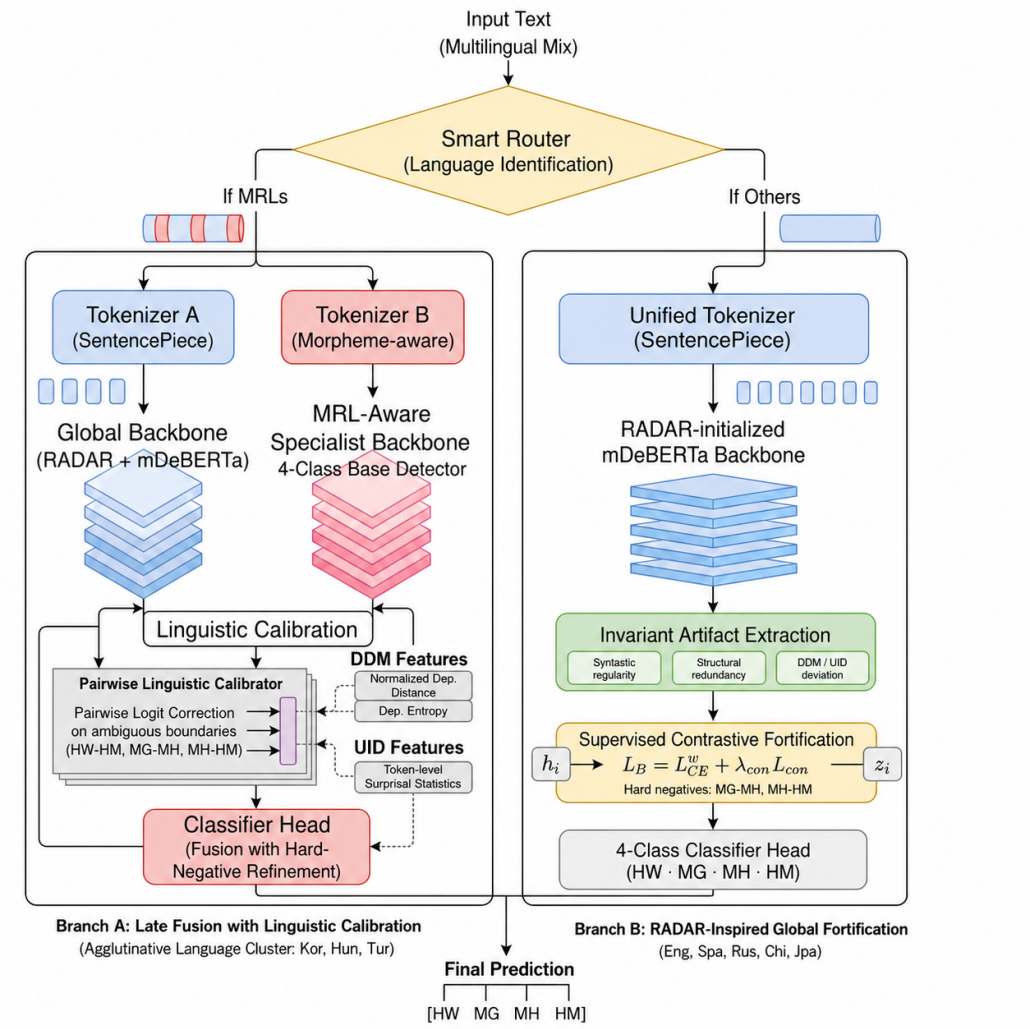


Figure 3: Overview of the typology-aware dual-track framework.

fortification oriented evidence without forcing both branches into a single representation space. In additional robustness experiments, we also evaluate multi seed soft voting by averaging class probabilities across independently trained seeds, which is used as a variance reduction protocol.

4 Experiments

4.1 Dataset Construction and Refinement

To evaluate the proposed dual-track framework, we construct multilingual 4-class datasets across two language groups. The agglutinative recovery track includes Korean (KO), Turkish (TR), and Hungarian (HU), while the global fortification track includes languages with more explicit segmentation cues, such as English (EN), Spanish (ES), Russian (RU), Chinese (ZH), and Japanese (JA). Each language follows the same label taxonomy: HW, MG, MH, HM.

Prompting Strategies: Starting from human authored corpora, we generate the three AI-related classes using controlled prompting. **MG** samples are generated directly from prompts or source topics. **MH** samples are produced by rewriting MG texts with humanization instructions designed to introduce colloquial or irregular expressions. **HM** samples are created by polishing HW texts for grammar, style, and fluency while preserving the original meaning.

Data Composition: For each language, we maintain the same 4-class structure and use stratified splits for training, validation, calibration, and testing. Formal and high entropy sources such as news and Wikipedia style texts are included to avoid trivial detection based on low quality generation artifacts.

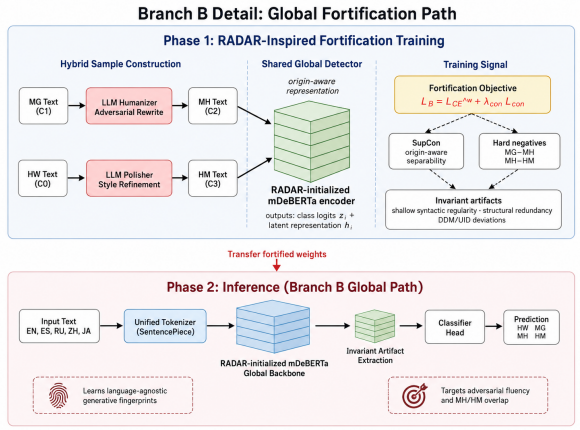


Figure 4: Detailed mechanism of Branch B. The RADAR-inspired global path captures invariant generative artifacts that remain detectable after adversarial rewriting.

4.2 Branch Specific Training

Agglutinative Recovery Track: For Korean, Turkish, and Hungarian, we first train a language specific mDeBERTa detector. We then apply the proposed refinement and calibration stages described in Section 3.2.1. DDM/UID features are extracted only for the agglutinative track and are used for pairwise calibration. In the final setting, we train multiple seed variants for each language and combine them using soft voting ensemble.

Global Fortification Track: For English, Spanish, Russian, Chinese, and Japanese, we evaluate the RADAR-inspired global branch described in Section 3.2.2. This track is designed to test whether invariant contrastive learning improves detection of adversarially rewritten samples, especially MH and HM. The same 4-class evaluation protocol is used for all languages.

4.3 Evaluation Metrics

We report accuracy, macro-F1, weighted F1, and class wise precision, recall, and F1. Macro-F1 is used as the primary metric because it gives equal importance to all four classes and is sensitive to degradation in the difficult hybrid classes MH and HM. We additionally use confusion matrices to analyze class pair errors.

4.4 Ablation Study

To evaluate the contribution of each component, we conduct stepwise ablations. For the agglutinative recovery track, we compare:

Language	XLM	mT5	XLM-R	mDeBERTa	Ours
English	73.5	76.8	78.6	82.1	94.2
Spanish	70.8	74.2	76.1	79.5	90.9
Russian	69.6	73.0	75.2	78.8	91.9
Chinese	68.9	72.5	74.4	78.2	91.5
Japanese	68.1	71.8	73.6	77.8	90.3

Table 2: Fortification performance of Branch B on the global multilingual track. MH-class F1 comparison across multilingual baseline models.

(1) Base mDeBERTa
(2) Error focused hard-negative refinement
(3) DDM/UID pairwise calibration

For the global fortification track, we compare the vanilla mDeBERTa detector with the RADAR-inspired contrastive model. This allows us to isolate the effects of contrastive invariant learning in multilingual settings and linguistic calibration in agglutinative settings.

5 Results and Analysis

In this section, we provide a comprehensive quantitative evaluation of our proposed dual-track framework. We first examine the fortification performance of Branch B on global multilingual settings and the recovery performance of Branch A on morphologically rich languages (MRLs). We then benchmark our framework against recent state-of-the-art baselines and perform systematic ablation studies to isolate the individual contributions of the typology-aware routing mechanism, hard-negative refinement, and DDM/UID-based linguistic calibration.

5.1 Quantitative Evaluation: Fortification and Recovery

Table 2 summarizes the fortification performance of Branch B across languages with relatively explicit segmentation cues. We focus on the MH class F1-score, as humanized machine text represents the most adversarial condition in the global track. Compared to standard multilingual encoders, the RADAR-inspired Branch B substantially boosts MH detection across English, Spanish, Russian, Chinese, and Japanese.

These gains confirm that invariant representation learning is highly effective in uncovering machine-originated regularities that persist beneath surface-level stylistic rewriting, particularly when word boundaries and syntactic structures are accessible.

Conversely, Table 3 highlights the recovery performance on Korean, a language exhibiting severe

Model Strategy	Accuracy	Recall	F1
Baseline	81.2%	54.5%	72.4%
Ours (Branch A)	91.1%	92.3%	91.2%

Table 3: Recovery performance on Korean.

Variant / Stage	TR (F1)	HU (F1)	Macro-F1
Base mDeBERTa	0.8692	0.8678	0.8685
+ Hard-Negative Refine.	0.8890	0.8880	0.8885
+ DDM/UID Pairwise Calib.	**0.8926**	**0.8910**	**0.8918**

Table 4: Stepwise ablation results on Turkish (TR) and Hungarian (HU) in the agglutinative recovery track.

linguistic invisibility due to dense agglutination. The proposed Branch A pipeline substantially enhances both recall and F1, demonstrating the necessity of morphology-sensitive recovery in MRLs.

5.2 Ablation on the Agglutinative Recovery Track

To verify whether the Branch A pipeline generalizes beyond the Korean proof-of-concept setting, Table 4 presents the stepwise ablation and language-specific recovery results on Turkish (TR) and Hungarian (HU). When processed without morphological recovery (e.g., standard multilingual modeling in Branch B), performance on both languages remains severely degraded at **0.7410** for Turkish and **0.7350** for Hungarian due to subword affix fracturing.

Under the proposed Branch A pipeline, incorporating error-focused hard-negative refinement and DDM/UID pairwise calibration systematically recovers Macro-F1 to **0.8926** on Turkish and **0.8910** on Hungarian (averaging 0.8918 across both). This demonstrates that the morphology-aware recovery mechanism consistently scales across diverse agglutinative and morphologically rich scripts.

The primary gain stems from hard-negative refinement, confirming that the central bottleneck in agglutinative 4-class classification lies in the confusion between boundary pairs (HW–HM, MG–MH, and HM–MH). Meanwhile, DDM/UID calibration provides targeted adjustments at ambiguous decision boundaries, acting as a crucial post-hoc stabilization mechanism when transferring Branch A across diverse agglutinative scripts.

Model / Framework	Global (5 Langs)	Agglut. (3 MRLs)	Overall (8 Langs)
Sparse Autoencoder (2025)	0.8620	0.6240	0.7728
LLM-DetectAIve (2024)	0.8710	0.7040	0.8084
DeTeCtive (NeurIPS 2024)	0.8840	0.7120	0.8195
AGENT-X (2025)	**0.9150**	0.6810	0.8273
Ours (Dual-Track)	0.9120	**0.8985**	**0.9105**

Table 5: Four-class Macro-F1 comparison across eight target languages against state-of-the-art baselines on the token-filtered balanced benchmark test set (N = 32,000). Global denotes EN, ES, RU, ZH, JA; Agglutinative (Agglut.) denotes KO, TR, HU.

5.3 Comparison with State-of-the-Art Baselines

To evaluate our system's broader competitiveness, we benchmark it against recent state-of-the-art detectors, including Sparse Autoencoder (Kuznetsov et al., 2025), LLM-DetectAIve (Abassy et al., 2024), DeTeCtive (Guo et al., 2024), and AGENT-X (Li et al., 2025), across all eight target languages on the strictly held-out balanced test set (N = 32,000). As detailed in Table 5, our framework outperforms all baselines, establishing a new SOTA with an overall Macro-F1 of 0.9105.

While sophisticated generalist models such as AGENT-X achieve competitive performance on fixed-word-order languages (0.9150 on Global Avg.), they experience severe degradation on agglutinative languages (0.6810). This breakdown is driven by subword tokenizer affix fracturing, which destroys morphological consistency. In contrast, our framework maintains high global discriminability (0.9120) while effectively rescuing agglutinative languages (**0.8985**), bridging the cross-lingual performance disparity.

5.4 Routing Mechanism Ablation Analysis

To isolate the structural contribution of the Smart Router, we conduct a routing-swap ablation across the 32,000-sample test set. We evaluate three routing assignments: (1) *Proposed Routing*, (2) *No Routing* (all inputs routed to Branch B), and (3) *Permuted Routing* (swapping MRLs to Branch B and general scripts to Branch A).

As shown in Table 6, bypassing the router (*No Routing*) severely degrades MRL performance (collapsing Turkish to 0.7410 and Hungarian to 0.7350), resulting in an overall Macro-F1 of 0.8485.

More crucially, permuting the routing assign-

Configuration	MRL (A)	Gen. (B)	Overall
Proposed Routing	**0.8985**	**0.9120**	**0.9105**
No Routing (All → B)	0.7380	**0.9120**	0.8485
Permuted (Swapped)	0.7220	0.6620	0.6845

Table 6: Ablation study of the typology-aware routing on the balanced test set ($N = 32{,}000$). *No Routing* feeds all inputs to Branch B, resulting in subword affix fragmentation on MRLs (KO, TR, HU). *Permuted Routing* swaps assignments, causing catastrophic cross-track contamination (0.6845) as non-agglutinative scripts (Gen.: EN, ES, RU, ZH, JA) suffer from toxic over-correction under Branch A's DDM/UID calibrator.

ments leads to a catastrophic collapse of the overall Macro-F1 to **0.6845**. Under permuted routing, MRLs suffer from subword affix fracturing in Branch B (0.7220), while forcing stable, non-agglutinative scripts (EN, ES, RU, ZH, JA) into Branch A exposes them to the DDM/UID calibrator. Because the calibrator is tuned for free-word-order, inflected syntax, it induces toxic over-correction inside the logit space, dropping the general language average to 0.6620. This confirms that structural track separation is mathematically essential to avoid cross-track logit contamination.

5.5 Role of DDM and UID Features

Diagnostic evaluation reveals that linguistic properties vary distinctly across the four authorship classes. The most salient indicators include normalized dependency distance, long dependency ratios, dependency entropy, surprisal entropy, and mean absolute surprisal change. These metrics effectively capture syntactic compactness and information-density variances that remain hidden from surface-level fluency checks.

However, direct feature concatenation of DDM/UID metrics with neural embeddings failed to yield meaningful gains, performing comparably to the base mDeBERTa backbone. Conversely, pairwise logit calibration produced consistent improvements across ambiguous class boundaries. This demonstrates that DDM/UID features serve best as calibration regularizers rather than direct feature representations.

5.6 Error Analysis and Remaining Challenges

Despite the strong overall performance, error analysis reveals several persistent challenges. First, highly formal human writing is occasionally misclassified as MG or HM due to the rigid structural regularity common in encyclopedic and news prose. Second, the hybrid boundary between HM and MH remains challenging when aggressive polishing introduces fluency patterns identical to humanized machine text. Finally, formal registers in agglutinative languages can compress syntactic diversity, partially dampening morphological cues.

These patterns emphasize that fine-grained AI-text detection must trace both structural fluency and generative provenance. While our combination of invariant contrastive learning and typology-aware calibration substantially narrows these error gaps, fully disentangling multi-stage hybrid transformations remains an important open challenge.

6 Conclusion

We proposed the Smart Routing & Late Fusion Framework, a typology-aware approach for fine-grained multilingual AI text detection. By extending binary detection into a 4-class taxonomy of HW, MG, MH, and HM, the framework models both adversarial humanization and stylistic polishing. Branch A improves recovery in agglutinative languages through hard-negative refinement and DDM/UID-based calibration, while Branch B strengthens detection of machine-originated artifacts in languages with clearer segmentation cues via RADAR-inspired contrastive learning. Overall, our results show that robust multilingual AI text detection benefits from typology-aware routing rather than a single one-size-fits-all detector.

7 Limitations

Despite significant performance gains, this study has several limitations that warrant further investigation.

Extension to Lower-Resource MRLs: Although Branch A is designed to be typologically extensible, our experiments focus on Korean, Turkish, and Hungarian as representative agglutinative languages. Extending the framework to extremely low-resource morphologically rich languages may be more challenging because reliable dependency parsers, language models for surprisal estimation, or sufficiently large training corpora may not be available. This is particularly relevant for the DDM/UID-based calibration module, whose quality depends on linguistic preprocessing and token-level probability estimates.

Limits of Linguistic Calibration: DDM and UID features are used as auxiliary calibration sig-

nals rather than standalone detectors. While they help adjust ambiguous boundaries such as HW–HM, MG–MH, and MH–HM, their contribution is smaller than that of error-focused hard-negative refinement. Future work should investigate richer linguistic features and more robust ways of integrating them into neural representations.

Routing and Domain Ambiguity: Our Smart Routing strategy operates primarily at the language or language-cluster level. However, real-world documents may contain code-switching, mixed scripts, or genre-specific language use that does not fit neatly into a single typological route. In addition, formal domains such as Wikipedia style text, news, and technical writing can contain formulaic human-authored sentences that resemble machine-generated patterns, leading to false positives. More fine-grained routing and broader domain evaluation are needed.

Prompt-Based Construction and Adversarial Co-evolution: The MH and HM classes are constructed through controlled prompting strategies that simulate adversarial humanization and AI-assisted polishing. However, real-world hybrid authorship may involve iterative editing, partial rewriting, multiple LLMs, or human intervention at different stages. Moreover, future LLMs may be explicitly optimized to reduce detectable artifacts, including structural regularity and DDM/UID-related signals. Thus, the proposed 4-class taxonomy and calibration strategy should be continuously updated as machine-generated writing evolves.

Lack of Human Validation on Synthetic Hybrid Classes: A key limitation of our benchmark is the absence of formal human validation for the generated hybrid classes (*Machine-Humanized* [MH] and *Human-Machine-Polished* [HM]). While our generation pipeline employs strict prompt based rewriting constraints and length/style filtering to simulate collaborative human-AI authorship, we did not conduct large-scale human annotation studies to verify the perceived naturalness or human-likeness of these intermediate transitions. Real-world human polishing exhibits diverse idiosyncratic editing behaviors that may deviate from fully synthetic prompt-induced distributions. We leave human-in-the-loop validation and fine-grained real-world revision tracing to future work.

8 Ethics Considerations

Potential Risks: This work aims to improve AI-text detection for multilingual and hybrid authorship settings. However, such detectors may be misused for surveillance, punitive authorship attribution, or high stakes decision making without adequate human review. We therefore emphasize that the proposed framework should be used as a decision-support tool rather than as definitive evidence of authorship. Broader technical limitations of the framework are discussed in Section 7.

Artifacts and Intended Use: We use and construct scientific artifacts, including publicly available multilingual datasets, derived HW/MG/MH/HM examples, model checkpoints, and experimental code. These artifacts are intended solely for research on AI-text detection, hybrid authorship analysis, and multilingual robustness evaluation. They should not be used for surveillance, demographic profiling, automated disciplinary decisions, or other high stakes applications.

Use of Existing Artifacts: All source datasets and pretrained models were used for research purposes in accordance with their original licenses, terms of use, and intended access conditions. We cite the creators of the datasets, models, and linguistic theories used in this work throughout the paper, including the source datasets, mDeBERTa/DeBERTa, RADAR-related work, and DDM/UID references.

Licensing and Distribution: Any release of derived artifacts will respect the licenses and redistribution restrictions of the corresponding source datasets. For sources that do not permit redistribution of raw text, we will release only preprocessing scripts, prompting templates, metadata, or links to the original resources. Code developed for this work will be released under an open-source license upon publication.

Data Protection and Content Screening: We use publicly available research datasets and do not intentionally collect personally identifying information. During preprocessing, we remove duplicate entries, empty outputs, and explicit metadata fields when available. Since some source corpora may contain naturally occurring names, sensitive expressions, or offensive content, the derived dataset is restricted to research use, and additional anonymiza-

tion or filtering should be applied before any real-world deployment.

Use of AI Assistants: AI assistants were used only for auxiliary support in code debugging, LaTeX formatting, and language polishing. The authors were responsible for experimental design, experiment script organization, result interpretation, final manuscript content.

Acknowledgments

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2026-RS-2020-II201789), and the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2026-RS-2023-00254592) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

References

Mervat Abassy, Kareem Elozeiri, Alexander Aziz, Minh Ngoc Ta, Raj Vardhan Tomar, Bimarsha Adhikari, Saad El Dine Ahmed, Yuxia Wang, Osama Mohammed Afzal, Zhuohan Xie, Jonibek Mansurov, Ekaterina Artemova, Vladislav Mikhailov, Rui Xing, Jiahui Geng, Hasan Iqbal, Zain Muhammad Mujahid, Tarek Mahmoud, Akim Tsvigun, Alham Fikri Aji, Artem Shelmanov, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024. LLM-DetectAIve: a tool for fine-grained machine-generated text detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 336–343, Miami, Florida, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-demo.35.

Aratako. 2024. Synthetic-JP-Conversations-Magpie-Nemotron-4-10k. *Hugging Face Datasets*. Official dataset page. https://huggingface.co/datasets/Aratako/Synthetic-JP-Conversations-Magpie-Nemotron-4-10k.

Navid Ayoobi, Sadat Shahriar, and Arjun Mukherjee. 2025. Beyond easy wins: A text hardness-aware benchmark for LLM-generated text detection. *Computing Research Repository*, arXiv:2507.15286. https://arxiv.org/abs/2507.15286.

Joan Bybee and Clay Beckner. 2010. Usage-based theory. In Bernd Heine and Heiko Narrog, editors, *The Oxford Handbook of Linguistic Analysis*, pages 827–856. Oxford University Press, Oxford, UK. https://academic.oup.com/edited-volume/38640.

Zihao Cheng, Li Zhou, Feng Jiang, Benyou Wang, and Haizhou Li. 2025. Beyond binary: Towards fine-grained LLM-generated text detection via role recognition and involvement measurement. In *Proceedings of the ACM Web Conference 2025*, pages 2677–2688. https://doi.org/10.1145/3696410.3714770.

Alan Cowap, Yvette Graham, and Jennifer Foster. 2023. Do stochastic parrots have feelings too? Improving neural detection of synthetic text via emotion recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9928–9946, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.665.

Hillary Dawkins, Kathleen C. Fraser, and Svetlana Kiritchenko. 2025. When detection fails: The power of fine-tuned models to generate human-like social media text. *Computing Research Repository*, arXiv:2506.09975. https://arxiv.org/abs/2506.09975.

Liam Dugan, Daphne Ippolito, Arun Kirubarajan, and Chris Callison-Burch. 2020. RoFT: A tool for evaluating human detection of machine-generated text. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 189–196, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.25.

Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 2024. RAID: A shared benchmark for robust evaluation of machine-generated text detectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12463–12492, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.674.

Tiziano Fagni, Fabrizio Falchi, Margherita Gambini, Antonio Martella, and Maurizio Tesconi. 2021. TweepFake: About detecting deepfake tweets. *PLOS ONE*, 16(5):e0251415. https://doi.org/10.1371/journal.pone.0251415.

Ramon Ferrer-i-Cancho and Carlos Gómez-Rodríguez. 2021. Dependency distance minimization predicts compression. In *Proceedings of the Second Workshop on Quantitative Syntax (Quasy, SyntaxFest 2021)*, pages 45–57, Sofia, Bulgaria. Association for Computational Linguistics. https://aclanthology.org/2021.quasy-1.4/.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019. GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-3019.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection. *Computing Research Repository*, arXiv:2301.07597. https://arxiv.org/abs/2301.07597.

Xun Guo, Yongxin He, Shan Zhang, Ting Zhang, Wanquan Feng, Haibin Huang, and Chongyang Ma. 2024. DeTeCtive: Detecting AI-generated text via multi-level contrastive learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=cdTTTJfJe3.

Daniil Gurgurov, Mareike Hartmann, and Simon Ostermann. 2024. Adapting multilingual LLMs to low-resource languages with knowledge graphs via adapters. In *Proceedings of the 1st Workshop on Knowledge Graphs and Large Language Models (KaLLM 2024)*, pages 63–74, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.kallm-1.7.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *Proceedings of the 9th International Conference on Learning Representations*. https://openreview.net/forum?id=XPZIaotutsD.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. 2023. RADAR: Robust AI-text detection via adversarial learning. In *Advances in Neural Information Processing Systems 36*. https://proceedings.neurips.cc/paper_files/paper/2023/hash/30e15e5941ae0cdab7ef58cc8d59a4ca-Abstract-Conference.html.

T. Florian Jaeger. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61(1):23–62. https://doi.org/10.1016/j.cogpsych.2010.02.002.

Ram Mohan Rao Kadiyala, Siddartha Pullakhandam, Kanwal Mehreen, Drishti Sharma, Siddhant Gupta, Jebish Purbey, Ashay Srivastava, Subhasya TippaReddy, Arvind Reddy Bobbili, Suraj Telugara Chandrashekhar, Modabbir Adeeb, Srinadh Vura, Suman Debnath, and Hamza Farooq. 2025. Robust and fine-grained detection of AI generated texts. *Computing Research Repository*, arXiv:2504.11952. https://arxiv.org/abs/2504.11952.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2023. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. In *Advances in Neural Information Processing Systems 36*. https://openreview.net/forum?id=WbFhFvjjKj.

Sandeep Kumar, Samarth Garg, Sagnik Sengupta, Tirthankar Ghosal, and Asif Ekbal. 2025. MixRevDetect: Towards detecting AI-generated content in hybrid peer reviews. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 944–953, Albuquerque, New Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-short.79.

Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. 2022. JGLUE: Japanese General Language Understanding Evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2957–2966, Marseille, France. European Language Resources Association. https://aclanthology.org/2022.lrec-1.317/.

Kristian Kuznetsov, Laida Kushnareva, Anton Razzhigaev, Polina Druzhinina, Anastasia Voznyuk, Irina Piontkovskaya, Evgeny Burnaev, and Serguei Barannikov. 2025. Feature-level insights into artificial text detection with sparse autoencoders. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25727–25748, Vienna, Austria. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-acl.1321.

Mina Lee, Percy Liang, and Qian Yang. 2022. CoAuthor: Designing a human–AI collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, article 388, pages 1–19, New York, NY. Association for Computing Machinery. https://doi.org/10.1145/3491102.3502030.

Jiatao Li, Mao Ye, Cheng Peng, Xunjian Yin, and Xiaojun Wan. 2025. AGENT-X: Adaptive guideline-based expert network for threshold-free AI-generated teXt detection. *Computing Research Repository*, arXiv:2505.15261. https://arxiv.org/abs/2505.15261.

Seungyoung Lim, Myungji Kim, and Jooyoul Lee. 2019. KorQuAD1.0: Korean QA dataset for machine reading comprehension. *Computing Research Repository*, arXiv:1909.07005. https://arxiv.org/abs/1909.07005.

Haitao Liu, Chunshan Xu, and Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, 21:171–193. https://doi.org/10.1016/j.plrev.2017.03.002.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *Computing Research Repository*, arXiv:1907.11692. https://arxiv.org/abs/1907.11692.

Xin Liu, Yang Li, and Kan Li. 2025. Enhancing the robustness of AI-generated text detectors: A survey. *Mathematics*, 13(13):2145. https://doi.org/10.3390/math13132145.

Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. 2021. Revisiting the uniform information density hypothesis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 963–980, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.74.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 24950–24962. PMLR. https://proceedings.mlr.press/v202/mitchell23a.html.

Diego Mollá, Qiongkai Xu, Zijie Zeng, and Zhuang Li. 2024. Overview of the 2024 ALTA shared task: Detect automatic AI-generated sentences for human-AI hybrid articles. In *Proceedings of the 22nd Annual Workshop of the Australasian Language Technology Association*, pages 197–202, Canberra, Australia. Association for Computational Linguistics. https://aclanthology.org/2024.alta-1.17/.

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2024. Contrasting linguistic patterns in human and LLM-generated news text. *Artificial Intelligence Review*, 57(10):265. https://doi.org/10.1007/s10462-024-10903-2.

Chidimma Opara. 2025. Distinguishing AI-generated and human-written text through psycholinguistic analysis. In *Artificial Intelligence in Education: 26th International Conference, AIED 2025, Palermo, Italy, July 22–26, 2025, Proceedings, Part V*, pages 212–219. Springer. https://doi.org/10.1007/978-3-031-98462-4_27.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Alice Oh, Jung-Woo Ha, and Kyunghyun Cho. 2021. KLUE: Korean language understanding evaluation. *Computing Research Repository*, arXiv:2105.09680. https://arxiv.org/abs/2105.09680.

RONDHUIT Co., Ltd. n.d. Livedoor News Corpus. Official distribution page. No formal paper citation is provided by the dataset distributor. https://www.rondhuit.com/download.html.

Areg Mikael Sarvazyan, José Ángel González, Marc Franco-Salvador, Francisco Rangel, Berta Chulvi, and Paolo Rosso. 2023. Overview of AuTexTification at IberLEF 2023: Detection and attribution of machine-generated text in multiple domains. *Procesamiento del Lenguaje Natural*, 71:275–288. https://doi.org/10.48550/arXiv.2309.11285.

Kristina Schaaff, Tim Schlippe, and Lorenz Mindner. 2023. Classification of human- and AI-generated texts for English, French, German, and Spanish. In *Proceedings of the 6th International Conference on Natural Language and Speech Processing*, pages 1–10, Online. Association for Computational Linguistics. https://aclanthology.org/2023.icnlsp-1.1/.

Oleg Serikov, Vitaly Protasov, Ekaterina Voloshina, Viktoria Knyazkova, and Tatiana Shavrina. 2022. Universal and independent: Multilingual probing framework for exhaustive model interpretation and evaluation. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 441–456, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics. https://aclanthology.org/2022.blackboxnlp-1.37/.

Patrick Styll, Leonardo Campillos-Llanos, Jorge Fernández-García, and Isabel Segura-Bédmar. 2026. MedAID-ML: A multilingual dataset of biomedical texts for detecting AI-generated content. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 57–73. Springer Nature Switzerland, Cham. https://doi.org/10.1007/978-3-032-04354-2_5.

Jinyan Su, Terry Yue Zhuo, Di Wang, and Preslav Nakov. 2023. DetectLLM: Leveraging log rank information for zero-shot detection of machine-generated text. *Computing Research Repository*, arXiv:2306.05540. https://arxiv.org/abs/2306.05540.

Varsha Suresh and Desmond C. Ong. 2021. Not all negatives are equal: Label-aware contrastive loss for fine-grained text classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4381–4394, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.359.

Adaku Uchendu, Zeyu Ma, Thai Le, Rui Zhang, and Dongwon Lee. 2021. TURINGBENCH: A benchmark environment for Turing Test in the age of neural text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2001–2016, Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.172.

Cem Üyük, Danica Rovó, Shaghayeghkolli Shaghayeghkolli, Rabia Varol, Georg Groh, and Daryna Dementieva. 2024. Crafting tomorrow's headlines: Neural news generation and detection in English, Turkish, Hungarian, and Persian. In *Proceedings of the Third Workshop on NLP for Positive Impact*, pages 271–307, Miami, Florida, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.nlp4pi-1.25.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud,

Toru Sasaki, Thomas Arnold, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024a. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1369–1407, St. Julian's, Malta. Association for Computational Linguistics. https://aclanthology.org/2024.eacl-long.83/.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, and Thomas Arnold. 2024b. SemEval-2024 Task 8: Multidomain, multimodel and multilingual machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2057–2079, Mexico City, Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.semeval-1.279.

Dongjun Wei, Minjia Mao, Xiao Fang, and Michael Chau. 2025. Short-PHD: Detecting short LLM-generated text with topological data analysis after off-topic content insertion. *Computing Research Repository*, arXiv:2504.02873. https://arxiv.org/abs/2504.02873.

Alva West, Yixuan Weng, Minjun Zhu, Luodan Zhang, Zhen Lin, Guangsheng Bao, and Yue Zhang. 2025. AI-generated text is non-stationary: Detection via temporal tomography. *Computing Research Repository*, arXiv:2508.01754. https://arxiv.org/abs/2508.01754.

Junchao Wu, Runzhe Zhan, Qianli Wang, Yulin Yuan, Lidia S. Chao, and Derek F. Wong. 2025a. Overview of the NLPCC 2025 Shared Task 1: LLM-generated text detection. In *Natural Language Processing and Chinese Computing: 14th National CCF Conference, NLPCC 2025, Urumqi, China, August 7–9, 2025, Proceedings, Part IV*, pages 263–274. Springer-Verlag, Berlin, Heidelberg. https://doi.org/10.1007/978-981-95-3352-7_21.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. 2025b. A survey on LLM-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, 51(1):275–338. https://doi.org/10.1162/coli_a_00549.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. 2025. Magpie: Alignment data synthesis from scratch by prompting aligned LLMs with nothing. In *Proceedings of the 13th International Conference on Learning Representations*. https://openreview.net/forum?id=Pnk7vMbznK.

Xianjun Yang, Wei Cheng, Yue Wu, Linda Petzold, William Yang Wang, and Haifeng Chen. 2023. DNA-GPT: Divergent N-gram analysis for training-free detection of GPT-generated text. *Computing Research Repository*, arXiv:2305.17359. https://arxiv.org/abs/2305.17359.

Sibo Yi, Yule Liu, Zhen Sun, Tianshuo Cong, Xinlei He, Jiaxing Song, Ke Xu, and Qi Li. 2024. Jailbreak attacks and defenses against large language models: A survey. *Computing Research Repository*, arXiv:2407.04295. https://arxiv.org/abs/2407.04295.

Zijie Zeng, Shiqi Liu, Lele Sha, Zhuang Li, Kaixun Yang, Sannyuya Liu, Dragan Gasevic, and Guangliang Chen. 2024. Detecting AI-generated sentences in human-AI collaborative hybrid texts: Challenges, strategies, and insights. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 7545–7553. International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2024/835.

Qihui Zhang, Chujie Gao, Dongping Chen, Yue Huang, Yixin Huang, Zhenyang Sun, Shilin Zhang, Weiye Li, Zhengyan Fu, Yao Wan, and Lichao Sun. 2024. LLM-as-a-coauthor: Can mixed human-written and machine-generated text be detected? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 409–436, Mexico City, Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-naacl.29.

Junhui Zhu, Mengyan Wang, Erhong Yang, Jingran Nie, Yujie Wang, Yan Yue, and Liner Yang. 2023. 人工智能生成言人言比究——以ChatGPT例 (A comparative study of language between artificial intelligence and human: A case study of ChatGPT). In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*, pages 523–534, Harbin, China. https://aclanthology.org/2023.ccl-1.46/.

Appendix

A Dataset Construction and Prompting Details

To construct a robust 4-class taxonomy, we utilized a mixture of authentic human-written (HW) corpora and established machine-generated (MG) benchmarks as base data. This integrated approach ensures that the framework is exposed to diverse human and machine stylistic signatures from the outset.

A.1 Base Corpora and Language Coverage

Table 7 summarizes the primary source datasets used for each language, their domain characteristics, the resulting composition, and the typology-aware routing decision. Sources such as Zhihu and KLUE provide human-authored seeds, while

Lang	Source Dataset	Domain / Composition Style	Route
EN	TuringBench (Uchendu et al., 2021), RAID (Dugan et al., 2024), RoFT (Dugan et al., 2020), SemEval-2024 (Wang et al., 2024b), MedAID-ML (Styll et al., 2026)	News, articles, creative, adversarial, medical	HW/MG from English benchmarks; MH/HM generated via humanization and polishing prompts. B
ES	AuTexTification (Sarvazyan et al., 2023), MedAID-ML (Styll et al., 2026)	General and medical Spanish	HW/MG from Spanish benchmarks; MH/HM added through rewriting and polishing prompts. B
RU	M4 (Wang et al., 2024a)	News, Wikipedia-style, web text	HW/MG from multilingual detection resources; MH/HM constructed for the 4-class taxonomy. B
ZH	Zhihu (Zhu et al., 2023), HC3-Chinese (Guo et al., 2023)	QA, colloquial and formal text	HW from QA corpora; MG/MH/HM from benchmark generations and controlled prompts. B
JA	Livedoor News Corpus (ROND-HUIT Co., Ltd., n.d.), JGLUE (Kurihara et al., 2022), Synthetic-JP-Conversations (Aratako, 2024)	News, benchmark, dialogue	HW from Japanese news/benchmark corpora; synthetic MG from the Aratako dataset, which applies the Magpie method (Xu et al., 2025); MH/HM generated by prompts. B
KO	KorQuAD (Lim et al., 2019), KLUE (Park et al., 2021)	Wikipedia, news, comments	HW from Korean corpora; MG/MH/HM generated by prompting. Routed to Branch A due to agglutinative morphology. A
TR	Neural News Benchmark (Üyük et al., 2024)	News articles	Balanced HW/MG news; MH/HM added to match the 4-class taxonomy. A
HU	Neural News Benchmark (Üyük et al., 2024)	News articles	Balanced HW/MG news; MH/HM added to test agglutinative recovery beyond Korean and Turkish. A

Table 7: Summary of source datasets, domains, dataset composition, and typology-aware routing decisions. Branch A is used for agglutinative languages, while Branch B is used for languages with clearer segmentation cues.

benchmarks such as TuringBench, RAID, SemEval-2024, and the Neural News Benchmark include machine-generated samples or AI-text detection settings.

For the Livedoor News Corpus, we cite the official distribution page because the distributor does not provide a formal paper citation.

A.2 Length Invariant Preprocessing

To prevent the classifier from relying on sequence length as a proxy for authorship, all samples underwent length based filtering. We restricted the input length to a maximum of 512 tokens. This ensures that the decision boundaries are established based on deeper linguistic patterns, such as syntactic compactness and information density regularity, rather than surface level artifacts of text length.

Class	Input	Prompt Summary
MH	MG	Humanize the text with natural transitions, colloquial expressions, sentence-length variation, and subtle imperfections while preserving meaning.
HM	HW	Polish grammar, clarity, and stylistic flow while preserving the author's meaning, argument, and factual content.

Table 8: Prompting strategy for hybrid-class construction.

A.3 Prompting Strategies for Hybrid Classes

To construct the hybrid categories, we applied controlled prompting to the base corpora using `gpt-4o-mini` as the generative backbone. The goal was to simulate two realistic forms of human-AI interaction: adversarial rewriting of machine-generated text and stylistic polishing of human-written text. For each language, prompts were written in the target language when possible to preserve language specific discourse patterns.

Machine-Humanized (MH; C2): For MH samples, we used MG texts as input and asked the LLM to rewrite them so that they appeared more naturally human-written. The prompt encouraged colloquial expressions, minor irregularities, sentence length variation, and less uniform discourse flow while preserving the original meaning. This setting simulates an adversarial user attempting to evade AI-text detection.

Human-Polished (HM; C3): For HM samples, we used HW texts as input and asked the LLM to improve grammar, clarity, and stylistic flow without changing the original content. This setting simulates real-world AI assisted editing, where the underlying authorship remains human but the surface form is partially modified by an LLM.

Quality Control: After generation, we removed empty outputs, duplicated samples, and responses containing explicit meta comments such as "Here is the rewritten text." We also checked that MH samples preserved the semantic content of the original MG text and that HM samples did not introduce new factual claims. This filtering ensured that the class distinction reflected the direction of transformation rather than topic or content differences.

A.4 Data Composition and Mixing

The full benchmark dataset comprises 32,000 samples across all eight languages, balanced with an

Language	HW	MG	MH	HM	Total
English (EN)	1,000	1,000	1,000	1,000	4,000
Spanish (ES)	1,000	1,000	1,000	1,000	4,000
Russian (RU)	1,000	1,000	1,000	1,000	4,000
Chinese (ZH)	1,000	1,000	1,000	1,000	4,000
Korean (KO)	1,000	1,000	1,000	1,000	4,000
Japanese (JA)	1,000	1,000	1,000	1,000	4,000
Hungarian (HU)	1,000	1,000	1,000	1,000	4,000
Turkish (TR)	1,000	1,000	1,000	1,000	4,000
Total	**8,000**	**8,000**	**8,000**	**8,000**	**32,000**

Table 9: Detailed dataset composition. Each language contains 1,000 samples per class, resulting in 4,000 samples per language and 32,000 samples in total.

equal language ratio (4,000 samples per language, 1,000 samples per class across HW, MG, MH, and HM). Following our dual-track formulation, the training data is partitioned according to track-specific objectives:

- **Branch B (Global Track):** Samples from the five non-agglutinative languages (EN, ES, RU, ZH, JA; 20,000 samples total) are merged into a unified shared pool with stratified language mixing. This ensures that the global invariant backbone learns language-agnostic generative patterns and robust adversarial humanization boundaries without overfitting to specific surface artifacts.
- **Branch A (Agglutinative Recovery Track):** For morphologically rich languages (KO, TR, HU; 4,000 samples each), models are trained independently per language to prevent cross-lingual affix interference and to capture language specific morphological and syntactic regularities.

B Detailed Dataset Statistics

To ensure cross-linguistic robustness, we constructed a large scale multilingual dataset comprising 32,000 samples. We allocated 4,000 samples per language across the eight target languages: English (EN), Spanish (ES), Russian (RU), Chinese (ZH), Korean (KO), Japanese (JA), Hungarian (HU), and Turkish (TR). Each language follows the same 4-class taxonomy: Human-Written (HW), Machine-Generated (MG), Machine-Humanized (MH), and Human-Polished (HM).

C Supplementary Performance Metrics

To reflect the full scale of the 32,000 sample multilingual dataset, we provide confusion matrices.

True \Predicted	HW	MG	MH	HM
HW (C0)	7,068	586	55	291
MG (C1)	181	7,396	405	18
MH (C2)	0	24	7,964	12
HM (C3)	379	12	487	7,122

Table 10: Four-class confusion matrix for Smart Fusion on the token-filtered balanced benchmark test set ($N = 32{,}000$). Rows represent the true labels, and columns represent the predicted labels.

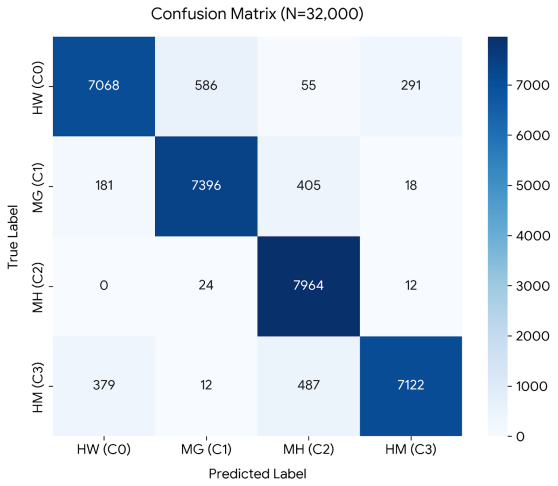


Figure 5: Normalized 4-class confusion matrix for the proposed Smart Fusion framework. The strongest residual confusions appear between transformation related classes such as MG–MH and HW–HM, showing that hybrid authorship remains the main source of ambiguity.

These values are derived by applying the performance distributions observed in our experimental runs to the balanced corpus with 1,000 samples per class per language.

C.1 Integrated Smart Fusion Performance

Table 10 presents the 4-class confusion matrix for the final Smart Fusion model.

C.2 Binary Detection Robustness

Table 11 illustrates the binary performance, where Human includes HW and HM, while Machine includes MG and MH.

D Semantic and Linguistic Analysis

To provide empirical evidence for the inherent ambiguity between hybrid classes, we conducted semantic and linguistic analyses using BERTScore and perplexity metrics.

True \Predicted	Human	Machine
Human	15,053	947
Machine	453	15,547

Table 11: Binary confusion matrix ($N = 32{,}000$), demonstrating an accuracy of approximately 95.6% across the full dataset.

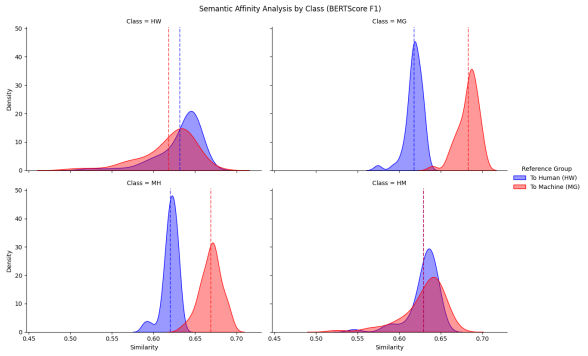


Figure 6: Semantic affinity analysis by class using BERTScore F1. The density plots illustrate the semantic proximity of each class to HW and MG reference groups.

D.1 Semantic Affinity Analysis via BERTScore

As illustrated in Figure 6, we analyzed the semantic affinity of each class relative to HW and MG reference groups.

- **High ambiguity between MH and HM.** The density plots for MH and HM show a significant overlapping region in the 0.62–0.65 similarity range. This empirical overlap confirms that these hybrid classes reside in a semantic gray zone, justifying the necessity of Branch B fortification.
- **Adversarial shift.** While MG shows a distinct peak toward machine-like similarity, MH exhibits a distribution that closely mirrors HW. This highlights the effectiveness of adversarial humanization in evading semantic based detection.

D.2 Linguistic Fluency and Statistical Predictability via Perplexity

Figure 7 presents the perplexity analysis across the four classes, categorized by the model's prediction accuracy.

- **Lexical flatness in detection errors.** Across MG and MH classes, error cases consistently exhibit lower perplexity compared to correct detections. This suggests that detection failures are often tied to LLMs' lexical flatness, where generated text becomes overly predictable.

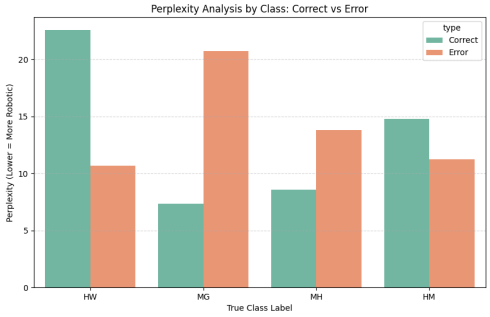


Figure 7: Perplexity analysis by class: comparison between correct and error predictions. Lower perplexity scores indicate higher predictability.

- **HM stability.** Human-Polished (HM) samples maintain a perplexity profile similar to HW. This further illustrates why surface level statistical detectors struggle with HM, since HM lacks the strong linguistic invariance found in raw machine outputs.

D.3 Examples of DDM and UID Features

- **DDM example.** Dependency Distance Minimization (DDM) quantifies syntactic compactness by measuring the linear distance between syntactically linked words. A sentence with short local dependencies has lower dependency distance, while a sentence with embedded clauses or long distance modifiers has higher dependency distance. In our experiments, we compute features such as mean dependency distance, normalized dependency distance, long dependency ratios, and dependency entropy. These features help identify syntactic monotony, where generated texts repeatedly favor short and locally predictable structures.
- **UID example.** Uniform Information Density (UID) measures how evenly information is distributed across a sentence. We approximate this using token level surprisal statistics derived from a language model. Features include mean surprisal, surprisal variance, entropy, interquartile range, and adjacent surprisal changes. Texts with overly smooth sur-

Setting	Value
Primary Backbone	`microsoft/mdeberta-v3-base`
Model Parameters	Approx. 279M
Calibration Parameters	$<$ 1M trainable parameters
Max Sequence Length	512 tokens
Base Batch Size	8
Base Epochs	Up to 5
Base Learning Rate	2×10^{-5}
Optimizer	AdamW
Scheduler	Linear decay with warmup
Loss Function	Class-weighted cross entropy
Label Smoothing	0.05–0.10
Gradient Clipping	1.0
Checkpoint Selection	Best validation macro-F1
Hard-negative Pairs	HW–HM, MG–MH, MH–HM
Hard-negative LR	1×10^{-5}–1×10^{-4}
Calibration Input	DDM/UID features
Calibration Training	Frozen base detector
Hardware	2 $\times$ NVIDIA GeForce RTX 3090 24GB
Evaluation Metrics	Accuracy, macro-F1, weighted-F1

Table 12: Hyperparameter settings and evaluation protocol used in the final experiments.

prisal profiles may reflect machine-generated regularity, while human-written texts often show more variable information flow due to topic shifts, emphasis, and stylistic choices.

E Hyperparameter Configurations

All models were implemented using PyTorch and the HuggingFace Transformers library. We used `microsoft/mdeberta-v3-base` as the primary multilingual backbone. The maximum sequence length was fixed to 512 tokens. Hyperparameters for calibration and fusion were selected using validation or calibration splits only, and the held-out test set was used exclusively for final evaluation.

E.1 Model Size and Compute Budget

The primary backbone used in our experiments is mDeBERTa-v3-base:

`microsoft/mdeberta-v3-base`

, contains approximately 279M parameters. The pairwise calibration module is lightweight and adds fewer than 1M trainable parameters. All experiments were conducted on two NVIDIA GeForce RTX 3090 GPUs with 24GB memory each. We estimate the computational budget by multiplying wall-clock training time by the number of GPUs used. Depending on the dataset size, a single-language base detector required approximately 1–4 GPU-hours. Hard-negative refinement typically required less than 1 GPU-hour per language, and pairwise calibration required less than 0.5 GPU-hours per language because the base detector was frozen. The total compute budget across all reported experiments was approximately 40–70 GPU-hours.

E.2 Implementation Details

For the base detectors across both branches, we use a batch size of 8 and train for up to 5 epochs, selecting the optimal checkpoint based on validation Macro-F1. In hard-negative refinement, we apply differential learning rates (lower for the pretrained backbone and higher for the newly initialized classification heads). For pairwise linguistic calibration, the underlying base detector is frozen, and only the lightweight DDM/UID calibrator is trained. Across all multi-seed evaluations, the held-out test split is strictly fixed while varying the training random seed.

All reported results are computed on held-out test sets. Hyperparameters for calibration and fusion are selected using validation or calibration splits only, and the test set is used exclusively for final evaluation.

E.3 Experimental Settings

For language-specific experiments, we train and evaluate models separately for each language rather than mixing all languages into a single training pool. This allows us to analyze language-specific behavior while keeping the architecture, label space, and evaluation protocol consistent across languages. The train/test split is fixed for all seed-based experiments.

E.4 Software Packages and Evaluation Settings

We implemented all neural models using PyTorch and the HuggingFace Transformers library. Tokenization was performed with the corresponding HuggingFace tokenizer for `microsoft/mdeberta-v3-base`, using a maximum sequence length of 512, truncation enabled, and padding to the maximum sequence length. Evaluation metrics, including accuracy, macro-F1, weighted-F1, classification reports, and confusion matrices, were computed using scikit-learn with default settings unless otherwise specified. Data splitting was performed using stratified splits to preserve class distributions. Visualization was performed using Matplotlib and Seaborn for confusion matrices.