BHAID: A Dataset of LLM-Generated Text in Indian Languages with a Dual Contrastive Learning Framework for Detection

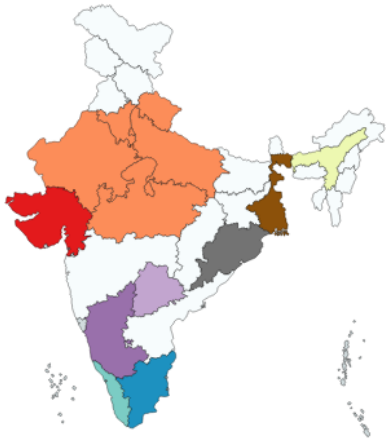
Shifali Agrahari[1], Sujit Kumar[1] Sanasam Ranbir Singh[1],
[1]Department of Computer Science and Engineering,
Indian Institute of Technology Guwahati, India
{a.shifali, ranbir}@iitg.ac.in, kumar.sujit474@gmail.com

Abstract

The rapid growth of multilingual Large Language Models (LLMs) has enabled the generation of fluent and culturally nuanced text across Indian languages, posing new challenges in distinguishing LLM-generated from human written content. Existing datasets and methodologies have primarily focused on AI-generated text detection in the English language. Existing datasets and methodologies, which are predominantly trained on English data, exhibit limited generalisation to low-resource and morphologically rich Indian languages, underscoring the need for linguistically diverse and culturally grounded benchmark datasets and methodologies. To address this gap, we introduce BHAID *BHaratiya AI-generated text Detection*, the first large scale multilingual benchmark for AI-text detection spanning nine Indian languages: Hindi, Bengali, Tamil, Telugu, Kannada, Malayalam, Gujarati, Odia, and Assamese. Furthermore, we propose a Dual Hierarchical Contrastive Learning framework *DualCLF* that integrates self-supervised and supervised contrastive objectives with focal loss to jointly enhance semantic robustness and inter-class discriminability. Observations from our experiments on the BHAID and benchmark multilingual datasets from the literature reveal that our method consistently outperforms recent state-of-the-art baseline approaches from the literature in detecting AI-generated text [1].

1 Introduction

Recent advances in large language models (LLMs) enable the automatic generation of coherent, human-like text (Singh et al., 2024b; Kumar et al., 2026), raising several concerns, including trust erosion (Bo et al., 2025; Peng et al., 2025), misinformation amplification (Vykopal et al.; Chen and Shu, 2023), Fake new (Kumar et al., 2025a; Kumar and Singh, 2025) academic integrity loss (Dou et al., 2024; Kim et al., 2023), factual hallucinations (Huang et al., 2025), ethical and legal violations (Samuelson, 2023), cultural homogenization (Anderson et al., 2024), automation bias (Laurito et al., 2025), privacy leakage (Meeus et al., 2025), economic disruption (Lin and Li, 2025), accountability gap (Peterson, 2025), and societal degradation (Bahov). While rapid advancements in multilingual (Qin et al., 2025) and Indic-centric LLMs have extended the human-like text generation capabilities of LLMs beyond English, enabling fluent, context-aware, and culturally nuanced text generation across diverse Indian languages (Gala et al., 2024; Singh et al., 2024a; Sankalp et al.; Singh et al., 2025).



Train	Test (en)	Test (Indian)	Difference
en	0.982	0.533	↓45.7%

Table 1: Performance drop from English to Indian languages test sets of *BHAID* Dataset.

Despite remarkable advancements in multilingual and Indic-centric LLMs, the literature lacks comprehensive datasets and methodologies for detecting LLM-generated text in Indian languages. To bridge this gap, we propose BHaratiya AI-generated text Detection (BHAID), the first large-scale benchmark dataset for detecting LLM-

[1]Data and code are publicly available. Dataset: Hugging Face BHAID Dataset Code: GitHub:DualCLF

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 1850–1871
November 6-10, 2026 ©2026 Association for Computational Linguistics

Dataset	Size	Domain	Languages	LLM Sources	Adv. Attacks	Multilingual	Multi-LLM
MGTBench (2021)	3K	Essays/News	EN	GPT-family	✓	✗	✗
MULTITuDE (2023)	44K	News	EN, ES, RU	GPT	✓	✓	✗
HC3 (2023)	84K	QA	EN, ZH	GPT-3	✗	Limited	✗
M4 (2023)	147K	Wiki/Reddit	EN, ZH, RU	Multiple	✗	✓	Limited
TuringBench	200K	News	EN	GPT, FAIR	✗	✗	Limited
DetectAIRev	256K	Reviews	EN	7+ LLMs	✗	✗	✓
BHAID	150K	News/Wiki	**Hi, As, Bn, Gu, Ml, Kn, Or, Ta, Te**	**4 LLMs**	✓	✓	✓

Table 2: Comparison of AI-generated text detection benchmarks. BHAID introduces a multilingual adversarial benchmark across 9 Indic languages using multiple LLM families.

generated text across nine major Indian languages: Hindi, Bengali, Tamil, Telugu, Kannada, Malayalam, Gujarati, Odia, and Assamese.

AI generated text detection methods (Mao et al., 2024; Su et al., 2023) in the literature have the following key limitations (i) often overlook fine-grained detection objectives, such as language-specific characteristics of generated text, and perform poorly in unsupervised settings (Liu et al., 2022), (ii) most existing approaches primarily focus on binary classification (i.e., distinguishing between human-written and machine-generated text) (Mao et al., 2024), and (iii) detectors often fail in language adaptation, as detectors (Mitchell et al., 2023; Macko et al., 2024) trained or fine-tuned on a single language struggle to generalize effectively across linguistically diverse datasets, as demonstrated in Table 1. Detection performance varies across languages, particularly between English and other languages, revealing that existing AI-generated text detectors do not generalize well to multilingual or low-resource settings. To address this, we propose DualCLF, a Dual Contrastive Learning framework that combines hierarchical contrastive learning with focal loss to improve multilingual robustness and discriminative power. Leveraging XLM-RoBERTa representations, DualCLF jointly maximises semantic consistency and inter-class separability, achieving robust, language-agnostic detection of LLM-generated text across diverse Indian languages.

2 Related Work

Recent studies have surveyed datasets and methods for detecting LLM-generated text (AIGT) (Ghosal et al., 2023; Yang et al., 2025; Fariello et al., 2025; Wu et al., 2025). These surveys highlight rapid advances in dataset construction, multilingual adaptation, and evaluation frameworks for distinguishing human-written from LLM-generated text.

Datasets: Numerous benchmark datasets have been developed to support AI-generated text detection across domains, languages, and LLMs (Guo et al., 2023; Salim and Hossain, 2024; Wang et al., 2023; Uchendu et al., 2021). Several datasets have been developed beyond English, such as Urdu (Ali et al., 2025), Korean (Park et al., 2025), Spanish (Sarvazyan et al., 2023), Russian (Shamardina et al., 2022), and Hindi (Kavathekar et al., 2024). Appendix Section A.3 and Table 2 summarize the key differences between BHAID and existing datasets. Several shared tasks further advance this area, including the multilingual track at SemEval-2024 Task 8 (Wang et al., 2024a) and the multilingual GenAI content detection task (Wang et al., 2025). Recent multilingual benchmarks for LLM-generated text detection include M4GT-Bench (Wang et al., 2024b) (covering Russian, Chinese, Urdu, Indonesian, and Arabic), MULTITuDE (Macko et al., 2023) (spanning 11 languages such as Arabic, Catalan, Chinese, Czech, Dutch, English, German, Portuguese, Russian, Spanish, and Ukrainian); and MultiSocial (Macko et al., 2024), which further expands multilingual coverage for this task. However, to the best of our knowledge, there is no publicly available dataset for AI-generated text detection in diverse Indian languages. To fill this gap, we introduce BHAID. The key differences between the proposed dataset and the existing dataset are highlighted in Appendix Subsection A.3 and Table 2.

Detection Methods: LLM-generated text detection methods can be broadly classified into supervised and zero-shot approaches. Supervised models (Fagni et al., 2021; Su et al., 2023; Liyanage et al., 2024; Liu et al., 2023; Agrahari et al., 2025b) other Semi-supervised Model (Agrahari et al., 2026, 2024; Agrahari and Singh, 2025; Agrahari et al., 2025a,c) perform well in-domain but degrade significantly under out-of-distribution (OOD) settings (Wang et al., 2023), especially for unseen models, languages, or do-

mains. In contrast, zero-shot methods such as GLTR (Gehrmann et al., 2019) (statistical features), DetectGPT (Mitchell et al., 2023) (token perturbation), SimLLM (Nguyen-Son et al., 2024) (semantic similarity), DART (Park et al., 2024) (semantic gap), and RAIDAR (Mao et al., 2024) (rewrite-based detection) rely on intrinsic model behaviours to identify AI-generated text without supervised training. Multilingual LLM-generated text detection remains challenging due to limited language-agnostic methods, performance gaps across languages (Papi et al., 2025), and the lack of standardized multilingual benchmarks (Macko et al., 2023). To address this, we propose DualCLF, a dual contrastive learning framework that combines hierarchical contrastive learning with focal loss for better multilingual discrimination.

3 Proposed Dataset

We begin our dataset curation process by collecting **Wikipedia articles**, and **news articles** across nine **Indian languages**, including **Hindi** (Hi), **Assamese** (As), **Bengali** (Bn), **Gujarati** (Gu), **Malayalam** (ML), **Kannada** (Kn), **Odia** (Or), **Tamil** (Ta), and **Telugu** (Te), to represent **human-authored text**. To ensure diversity and authenticity of human-authored text across nine languages, we collect news and Wikipedia articles from multiple publicly available and credible sources. (i) *News:* We consider the AI4Bharat-IndicNLP news corpus (Kunchukuttan et al., 2020)[2], providing large-scale monolingual news text across Indic languages. (ii) *Wikipedia:* We scrape Indian-language articles from Wikipedia archives (Hindi Wikipedia, 2025), restricting content to pre-2017 publications to minimise contamination from recent LLM-generated text. Subsequently, we employ prompting with **GPT** (Yenduri et al., 2023) [3], **Grok** [4], **Gemini** (Comanici et al., 2025) [5], and **Aya** (Üstün et al., 2024) [6] to obtain LLM-generated text corresponding to human-authored text, by providing each LLM with a prompt instruction(Kumar et al., 2025b). Alongside text generated by LLM using prompt instructions, we include three types of adversarial attacks detailed in the study (Dugan et al., 2024): *Rewrite* (Krishna et al., 2023), *Style Transfer* (Gagiano et al., 2021), and *Synonym Replacement* (Pu et al., 2023). Incorporating LLM-generated text from *Rewrite* (Krishna et al., 2023), *Style Transfer* (Gagiano et al., 2021), and *Synonym Replacement* (Pu et al., 2023) attacks, along with prompt-based generations, makes BHAID more robust for training detectors to distinguish AI- from human-generated text in real-world settings. Details are provided in Appendix Section A.1 and Appendix Table 2. (Example of dataset present in Appendix Figure 4).

Domain	Human	LLM	Total
News	20,000	80,000	100,000
Wikipedia	10,000	40,000	50,000
Total	**30,000**	**120,000**	**150,000**

Table 3: Domain-wise statistics of the proposed dataset.

Prompt and Adversarial Attack Diversity: We guide LLMs with prompts to generate text exhibiting language-specific stylistic, linguistic, tonal, and discourse-level features across Indian languages. For Wikipedia-style text, we provide the page title and target word count, while for news articles we supply the headline, publication details (location and date), and target word count. For adversarial attacks, including rewrite, style transfer, and synonym replacement, we provide human-written text along with attack-specific prompts to ensure semantic and contextual alignment between human and LLM-generated content, thereby enhancing the authenticity and diversity of the generated text. Table 3 presents the characteristics of BHAID dataset.

3.1 Human Annotation & Quality Evaluation

To assess the reliability of **BHAID** and the quality of multilingual LLM-generated text, we randomly selected 200 LLM-generated samples per language and assigned them to four native-speaker graduate annotators. Each language had two annotators with engineering backgrounds and two with linguistics backgrounds to ensure annotation quality.

Human annotators assigned scores for Fluency, Origin, Coherence, and Factuality of the LLM-generated text, where Fluency measures grammatical correctness, readability, and naturalness in word choice and sentence structure. *Origin:* Assesses whether the text exhibits linguistic and stylistic cues typical of human writing or machine-generated output. *Coherence:* Measures the logical flow and semantic consistency of ideas and sentences within the generated text. *Factuality:*

[2] AI4Bharat/indicnlp_corpus
[3] openai.gpt
[4] x.ai/grok
[5] gemini ai apiaya
[6] huggingface.co/CohereLabs/aya-101

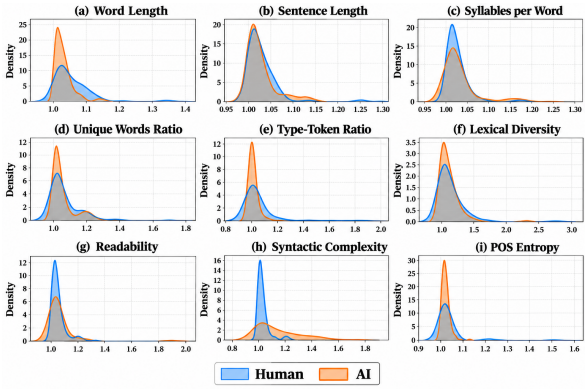


Figure 1: Density plots showing word-length distribution vs. perplexity for human (blue) and LLM-generated (orange) text across Indian languages.

Assesses how accurately the generated text reflects real-world facts or aligns with the source content. Appendix Table 10 presents our annotation instructions and scoring criteria for evaluating the Fluency, Origin, Coherence, and Factuality of LLM-generated text. For multilingual annotation, we compute Krippendorff alpha (Krippendorff, 2011) and Fleiss kappa (Fleiss, 1971) separately for each language to capture inter-annotator reliability within language-specific contexts, and report a macro-averaged κ as an overall reliability measure (Table 4). Language-wise human-annotation scores are reported in Appendix Table 11. The results show moderately high inter-annotator agreement, suggesting consistent evaluation and generally fluent LLM-generated text. Annotators show high agreement in assessing coherence and factuality, indicating consistent judgments and factually coherent LLM-generated text in BHAID. The Origin metric identifies whether a text is human-authored or LLM-generated. As shown in Table 4, low inter-annotator agreement for Origin indicates that distinguishing human from AI-written text was difficult. This suggests that LLM-generated samples in BHAID closely mimic human writing, posing challenges for both human judgment and automated detection models. Detail analysis is presented in Appendix Section A.2, B.

3.2 Linguistic Features

We consider the following linguistic features (Agrahari et al., 2024) to study the differences between LLM-generated and human-written text in the nine languages. Figure 1 shows density plots of word-count distributions, revealing that LLM-generated texts exhibit more uniform lengths with sharper peaks round 1.0, while human-written texts display greater variability and longer tails. This indicates that LLMs follow more regular word-length patterns than natural human writing. We further examine linguistic features to characterize these differences (Appendix Section C).

Metric	Fluency	Fact	Coherence	Origin
Krippendorff α	0.742	0.683	0.711	0.462
Fleiss κ	0.812	0.745	0.768	0.598

Table 4: Krippendorff's Alpha (α) and Fleiss' Kappa (κ) scores for reviews by metric, as mentioned in Table 10.

Lexical and Structural Features. Lexical features capture surface-level vocabulary characteristics, such as word diversity, length, and repetition, helping analyze linguistic richness and stylistic differences between LLM-generated and human-written text across Indian languages. We analyze the following features to study lexical diversity: (i) *Average Number of Sentences:* the average number of sentences per paragraph, (ii) *Average Words per sentence:* Measures sentence length to assess linguistic complexity and formality, enabling comparison of whether LLM-generated text uses simpler or more complex structures than human-written text, (iii) *Number of Unique Words:* count of unique words in the text, (iv) *Type-Token Ratio (TTR):* Measures the proportion of unique words (types) to total words (tokens), serving as an indicator of lexical diversity and enabling comparison of vocabulary richness and repetition between LLM-generated and human-written text, (v) *Measure of Textual Lexical Diversity (MTLD):* a robust metric of lexical diversity that accounts for text length by calculating the average length of word sequences required for TTR to drop below a specified threshold (typically 0.72). Higher MTLD indicate greater lexical diversity and less repetition, providing a more stable basis for comparison across texts of varying lengths, and (vi) *Hyper geometric Distribution Diversity (HD-D):* a probabilistic measure of lexical diversity that estimates the likelihood of encountering unique words within random samples of fixed size (commonly 42 words). HD-D is less sensitive to text length variations, provides a reliable indicator of lexical variety between human-written and LLM-generated content.(DDetails are provided in Appendix Section C).

Table 5 presents a comparison between LLM-generated and human-written text in terms of lexical diversity features. The following conclusions can be drawn from Table 5 (i) Human-authored

Metric	Type	Hi	Bn	Gu	Ml	Ta	Te	As	Or	Kn
#Sent.	Human	22.73	31.46	39.01	46.26	76.33	36.41	53.17	52.49	46.56
	LLM	26.83	33.75	40.51	51.00	29.90	63.75	46.43	62.30	55.05
#Word	Human	17.16	16.10	10.13	9.18	16.87	12.90	8.42	13.05	12.22
	LLM	11.18	18.62	13.98	10.00	16.93	22.76	10.16	17.10	18.56
Uni.	Human	337.00	402.83	510.59	402.77	595.33	312.42	408.57	297.68	344.53
	LLM	266.32	375.22	460.70	341.00	333.53	276.05	232.11	160.25	234.67
TTR	Human	0.66	0.59	0.57	0.48	0.60	0.61	0.59	0.63	0.56
	LLM	0.43	0.42	0.72	0.40	0.32	0.35	0.69	0.64	0.65
MTLD	Human	148.47	124.26	100.12	72.55	129.65	108.95	99.90	155.51	123.45
	LLM	35.24	69.01	62.51	56.01	60.90	75.71	72.28	75.44	94.45
HD-D	Human	0.86	0.88	0.85	0.84	0.86	0.85	0.84	0.90	0.83
	LLM	0.65	0.71	0.83	0.60	0.47	0.64	0.72	0.86	0.68

Table 5: Lexical and structural feature comparison between human- and LLM-generated text across nine Indian languages.

Metric	Direction	Hi	Bn	Gu	Ml	Ta	Te	As	Or	Kn
KL	Human → LLM	13.58	13.97	16.46	21.93	14.71	12.91	15.72	10.50	16.50
	LLM → Human	14.53	15.33	15.92	17.31	14.50	13.92	14.80	11.93	13.23
JS	Human → LLM	0.58	0.61	0.64	0.69	0.60	0.56	0.61	0.48	0.54
	LLM → Human	0.58	0.61	0.64	0.69	0.60	0.56	0.61	0.48	0.58

Table 6: Comparison of KL and JS divergence between human-written and LLM-generated texts across Indian languages.

texts exhibit higher lexical diversity than LLM-generated texts across most languages, as reflected by TTR, MTLD, and HD-D scores, (ii) LLM-generated texts exhibit lexical repetition and lower variation, reflected by substantially lower MTLD and HD-D values. (iii) TTR values are generally lower for LLMs, suggesting a more limited and repetitive vocabulary use. (vi) Human texts demonstrate greater richness and balance in vocabulary, especially evident in languages like Hi, Ta, and Bn. The gap is most evident in Hi and Be, where human texts achieve much higher MTLD and HD-D values. In Gujarati, the slightly higher TTR in LLM outputs likely reflects shorter or simpler sentences rather than genuine diversity. Malayalam and Tamil show reduced lexical variety in LLM text, with Tamil displaying the sharpest contrast. Human-authored text has more than twice the MTLD of LLM outputs. Telugu LLM-authored texts are longer but less diverse, suggesting verbosity without richness. Odia, and Kannada human texts record the highest MTLD and HD-D scores, highlighting natural lexical depth and stylistic flexibility. While LLMs produce grammatically coherent text, it still fall short of capturing the nuanced lexical diversity characteristic of human writing.

Information Theoretic Divergence Metrics: We use Kullback–Leibler (KL) and Jensen–Shannon (JS) divergences to quantify distributional differences between human and LLM-generated text via token frequency distributions. Lower divergence indicates closer lexical and stylistic alignment, while higher values reflect greater deviation, JS capturing overall dissimilarity and KL emphasizing directional divergence from human writing patterns.(Detail mention in Appendix Section C)

Metric	Type	Hi	Bn	Gu	Ml	Ta	Te	As	Or	Kn
Flesch	Human	78.16	72.53	80.48	75.83	82.53	80.92	81.12	73.12	79.34
	LLM	65.60	68.31	70.47	69.22	73.13	77.49	74.12	70.72	72.16
Fog	Human	9.74	11.12	8.53	10.08	7.85	8.41	8.07	10.82	8.94
	LLM	12.71	13.60	10.71	11.40	9.54	9.66	9.41	11.46	10.23
SMOG	Human	7.35	8.69	6.25	7.71	6.13	6.91	6.44	8.30	6.87
	LLM	9.05	9.41	7.67	8.84	7.28	7.44	7.14	8.98	7.92
Dale-Chall	Human	6.77	6.36	5.67	6.12	5.55	5.92	5.62	6.52	5.84
	LLM	7.91	7.38	6.93	6.84	6.72	6.23	6.25	6.89	6.48

Table 7: Readability metrics comparison between human and LLM-generated texts across Indian languages.

Table 6 reveals that LLM-generated texts exhibit moderate divergence from human-authored ones across all languages. Odia and Kannada show the lowest KL and JS values, suggesting that LLMs closely approximate human lexical distributions in these languages. Telugu, Tamil, and Hindi demonstrate moderate divergence, implying partial similarity in vocabulary and stylistic patterns. In contrast, Malayalam, Gujarati, Bengali, and Assamese display higher divergence scores, revealing that LLMs struggle to replicate the lexical richness and balanced word distributions found in human writing, particularly in morphologically complex languages. Accordingly, BHAID dataset, is diverse and well-balanced, making it suitable for developing robust multilingual LLM-generated text detection models in Indic languages.

Readability Metrics: Assess text comprehensibility based on sentence length, word complexity, and syntax; higher readability indicates simpler, more accessible text, while greater sentence complexity reflects richer grammatical structure and denser information. We consider the following readability and sentence complexity features: (i) *Flesch Reading Ease* (Fleiss, 1971): A readability metric with higher scores indicating easier-to-read text. (ii) *Fog Index*: Estimates the years of formal education required to understand the text. A higher Fog Index implies more complex sentence structures and vocabulary, typical in academic or machine-generated text. (iii) *SMOG Index* (Mc Laughlin, 1969): A readability metric based on polysyllabic words that estimates the education level required to understand a text, enabling comparison of text complexity between LLM-generated and human-written content. (iv) *Dale-Chall:* Measures readability using the propor-

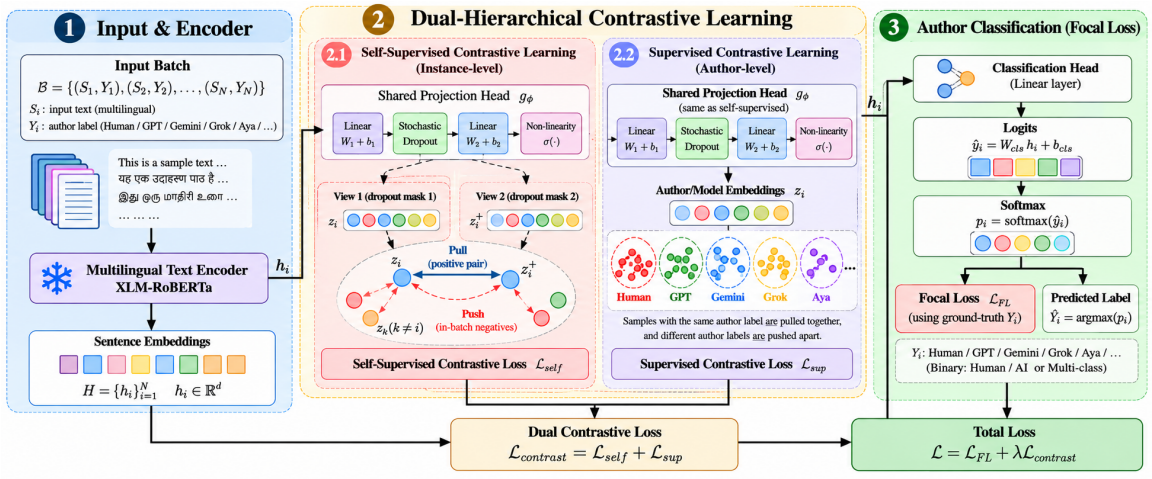


Figure 2: Proposed method Dual-CLF a Dual Contrastive Learning framework with contrastive learning loss.

tion of unfamiliar words, with higher scores indicating more complex vocabulary and higher reading grade levels. Table 7 shows that human-written texts are more readable and more straightforward than LLM-generated texts.

Across languages, human-written texts are consistently more readable than LLM-generated ones, with higher Flesch Reading Ease and lower Fog, SMOG, and Dale–Chall scores. LLM outputs tend to use longer sentences and denser vocabulary, reducing readability. The gap is largest in Hindi, Bengali, and Gujarati, while Telugu, Assamese, and Tamil show closer alignment, indicating comparatively better LLM performance in these languages. Overall, while LLMs maintain grammatical correctness, they tend to generate more lexically rich and stylistically dense text that is slightly less reader-friendly than human-authored content.

4 Proposed Methodology

Given a batch $\mathcal{B} = \{(\mathcal{S}_i, \mathcal{Y}_i)\}_{i=1}^{N}$, where $\mathcal{S}_i$ denotes the input text and $\mathcal{Y}_i$ is the author label, the objective is to predict $\mathcal{Y}_i$ for each ($\mathcal{S}_i \in \mathcal{B}$. while N denotes the batch size. Considering a batch $\mathcal{B}$, we first obtain the encoded representations $\mathbf{h}i_{i=1}^{N}$ of the input texts $\mathcal{S}i_{i=1}^{N}$ using the multilingual text encoder XLM-RoBERTa[7]. Subsequently, these representations are fed into the dual-hierarchical contrastive learning framework and the focal loss function for author classification.

4.1 Dual-Hierarchical Contrastive Learning

The Dual-Hierarchical Contrastive Learning module maximizes mutual information between representations and class labels by combining self-supervised learning, which enforces invariance across augmented views, with supervised learning that clusters same-class samples and separates different classes. These objectives are jointly optimized with focal loss to enhance LLM-generated text detection, as shown in Fig. 2.

4.1.1 Self-Supervised Contrastive

We adopt self-supervised contrastive learning to enable effective data augmentation and richer semantic representations, capturing span-level and contextual features beyond token semantics to better distinguish human-written from LLM-generated text. This study considers the positive and negative sampling strategy SimCSE (Gao et al., 2021) and the stochastic dropout method (Srivastava et al., 2014) to generate semantically consistent augmented views of human text for self-supervised contrastive learning. Given the encoded representation $\mathbf{h}_i \in \mathbb{R}^d$ of each input sentence S_i, we generate two augmented views, (z_i, z_i^+), from $\mathbf{h}_i$ following Equation 1:

$$\begin{aligned} \mathbf{z}_i &= \sigma\left(\mathbf{W}_2\, \text{Dropout}_1(\mathbf{W}_1 \mathbf{h}_i + \mathbf{b}_1) + \mathbf{b}_2\right), \\ \mathbf{z}_i^+ &= \sigma\left(\mathbf{W}_2\, \text{Dropout}_2(\mathbf{W}_1 \mathbf{h}_i + \mathbf{b}_1) + \mathbf{b}_2\right), \end{aligned} \tag{1}$$

where $\{\mathbf{W}_1, \mathbf{b}_1, \mathbf{W}_2, \mathbf{b}_2\}$ are learnable parameters. The independent stochastic dropout operations, Dropout_1 and Dropout_2, are applied over the encoded representation $\mathbf{h}_i$ using independent binary masks drawn from $\mathbf{m}_1, \mathbf{m}_2 \sim \text{Bernoulli}(1-p)$:

$$\begin{aligned} \text{Dropout}_1(\mathbf{h}_i) &= \mathbf{h}_i \odot \mathbf{m}_1, \quad \mathbf{m}_1 \sim \text{Bernoulli}(1-p), \\ \text{Dropout}_2(\mathbf{h}_i) &= \mathbf{h}_i \odot \mathbf{m}_2, \quad \mathbf{m}_2 \sim \text{Bernoulli}(1-p), \end{aligned} \tag{2}$$

where p is the dropout probability and $\odot$ denotes element-wise multiplication. Both $\mathbf{z}_i$ and $\mathbf{z}_i^+$ are

[7] huggingface.co/FacebookAI/xlm-roberta-base

derived from the same representation $\mathbf{h}_i$, differing only due to independently sampled dropout masks, while a shared projection head enforces view-invariant semantic representations. Subsequently, we define self-supervised contrastive loss:

$$\mathcal{L}_{\text{self}} = -\frac{1}{2N}\sum_{i=1}^{N}\log\frac{\exp(\text{sim}(z_i, z_i^+)/\tau)}{\sum_{k=1}^{2N}\mathbf{1}_{[k\neq i]}\exp(\text{sim}(z_i, z_k)/\tau)}, \tag{3}$$

where N denotes the batch size before augmentation, τ represents the temperature, and $\text{sim}(\cdot)$ computes the cosine similarity between two embeddings. Intuitively, the self-supervised objective drives the model to maximize agreement between dropout-induced views of the same sentence, promoting the retention of semantic information and improving representation robustness and generalization in LLM-generated text detection.

4.1.2 Supervised Contrastive Learning

We integrate supervised signals to refine the embedding by leveraging author labels $\mathcal{Y}_i \in \{\text{human}, \text{GPT}, \text{Grok}, \text{Gemini}, \text{Aya}\}$ to enhance inter-class separability and intra-class compactness, thereby improving the discriminative capability of the learned representations across text samples. For a batch of N samples, the supervised contrastive loss is defined in Equation 4:

$$\mathcal{L}_{\text{sup}} = -\sum_{i=1}^{N}\frac{1}{|P(i)|}\sum_{j=1}^{N}\mathbf{1}_{[i\neq j]}\cdot\mathbf{1}_{[y_i=y_j]} \cdot\log\frac{\exp\left(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau\right)}{\sum_{k=1}^{N}\mathbf{1}_{[k\neq i]}\exp\left(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau\right)}. \tag{4}$$

Here for each anchor sample i within a batch of size N, the set $P(i)$ contains all other samples j sharing the same author label $y_i = y_j$. The indicator function $\mathbf{1}_{[y_i=y_j]}$ equals 1 if samples i and j share the same label i.e., form a positive pair, and zero otherwise. Similarly, $\mathbf{1}_{[i\neq j]}$ ensures that self-comparisons are excluded from the summation. Together, these indicator functions guarantee that only distinct positive pairs, i.e., embeddings belonging to the same author class, contribute to the numerator of the loss. The numerator $\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)$ measures the similarity between the anchor embedding $\mathbf{z}_i$ and a positive embedding $\mathbf{z}_j$, while the denominator aggregates similarities between the anchor $\mathbf{z}_i$ and all other samples $\mathbf{z}_k$ in the batch. The temperature hyperparameter τ controls the sharpness of the similarity distribution. Intuitively, the supervised contrastive objective $\mathcal{L}_{\text{sup}}$ maximizes the agreement between embeddings of the same author (promoting intra-class compactness) and minimizes the similarity between embeddings of different authors (inter-class separability), to enhance the discriminative capability of the learned representation space for text author detection. The self-supervised contrastive objective encourages two stochastic projections of the same encoder representation h_i to remain close while separating representations originating from different instances, thereby promoting dropout-invariant and instance-discriminative encoder representations.

4.2 Final Training Objective

We jointly optimize classification and representation learning using a dual-contrastive framework with focal loss (Lin et al., 2017). The objective combines focal loss for author classification, self-supervised contrastive loss for semantic robustness, and supervised contrastive loss for author-level discriminability, as defined in Equation 5.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{FL}} + \lambda\left(\mathcal{L}_{\text{self}} + \mathcal{L}_{\text{sup}}\right). \tag{5}$$

where λ is a weighting coefficient that balances the contributions of the self-supervised and supervised contrastive objectives relative to the focal loss. We consider two settings: binary *Human vs. AI* classification and multi-class *author detection*, with labels $\mathcal{Y}_i \in \text{human}, \text{GPT}, \text{Grok}, \text{Gemini}, \text{Aya}$.

For each input text $\mathcal{S}_i$, the multilingual encoder produces a sentence-level representation $\mathbf{h}_i \in \mathbb{R}^d$, which is shared by the contrastive and classification branches. In the contrastive branch, $\mathbf{h}_i$ is passed through the projection head to obtain the representations used by the self-supervised and supervised contrastive objectives, whereas the classification branch operates directly on $\mathbf{h}_i$. During joint training, gradients from both branches fine-tune the multilingual encoder, such that $\mathbf{h}_i$ is learned under both contrastive and classification supervision. In particular, the supervised contrastive objective encourages representations from the same author to become more similar while separating those from different authors, whereas the focal-loss objective promotes author-discriminative classification. The classification head maps this shared encoder representation into class logits through a linear transformation, as defined in Equation 6:

$$\mathbf{o}_i = \mathbf{W}_{\text{cls}}\mathbf{h}_i + \mathbf{b}_{\text{cls}}, \tag{6}$$

where $\mathbf{W}_{\text{cls}} \in \mathbb{R}^{n\times d}$ and $\mathbf{b}_{\text{cls}} \in \mathbb{R}^n$ are learnable parameters, and n denotes the number of author

Based	Model	BHAID			M4 (Wang et al., 2023)			COLING (Wang et al., 2025)		
		Acc	F1-H	F1-AI	Acc	F1-H	F1-AI	Acc	F1-H	F1-AI
Baselines	Log-Likelihood (Gehrmann et al., 2019)	0.35	0.28	0.33	0.38	0.29	0.38	0.38	0.33	0.44
	Rank (Gehrmann et al., 2019)	0.44	0.41	0.45	0.46	0.44	0.46	0.45	0.43	0.45
	Log-Rank (Gehrmann et al., 2019)	0.47	0.44	0.47	0.49	0.46	0.49	0.48	0.46	0.48
	Entropy (Gehrmann et al., 2019)	0.50	0.48	0.50	0.52	0.49	0.52	0.51	0.49	0.51
	NPR (Su et al., 2023)	0.61	0.59	0.61	0.74	0.62	0.74	0.65	0.60	0.72
	DetectGPT (Mitchell et al., 2023)	0.67	0.65	0.67	0.70	0.67	0.69	0.69	0.66	0.68
	Fast-DetectGPT (Bao et al., 2023)	0.72	0.70	0.73	0.75	0.73	0.75	0.74	0.71	0.73
	J-Guard (Kumarage et al., 2023)	0.78	0.76	0.78	0.80	0.79	0.80	0.81	0.78	0.80
	RAIDAR (Mao et al., 2024)	0.83	0.82	0.83	0.85	0.84	0.86	0.84	0.83	0.85
	BISCOPE (Guo et al., 2024)	0.86	0.84	0.87	0.88	0.86	0.89	0.87	0.85	0.88
Ablation	Self-Supervised $\mathcal{L}_{\text{self}}$	0.81	0.74	0.82	0.82	0.85	0.83	0.82	0.74	0.82
	Supervised $\mathcal{L}_{\text{sup}}$	0.88	0.87	0.89	0.89	0.88	0.90	0.87	0.86	0.89
	Dual Contrastive (w/o Focal)	0.93	0.91	0.94	0.90	0.91	0.90	0.92	0.89	0.91
Proposed	DualCLF (Dual + Focal)	**0.96**	**0.95**	**0.97**	**0.92**	**0.93**	**0.94**	**0.94**	**0.92**	**0.93**

Table 8: Performance comparison of the proposed DualCLF model and baselines across multilingual datasets. *Acc* denotes accuracy; *F1-H* and *F1-AI* represent F1-scores for human and LLM-generated texts.

classes. The dimensionality of $\mathbf{o}_i$ depends on the specific classification setting. For binary classification (Human vs. AI), the output layer consists of $n = 2$ units, yielding the class probabilities

$$\mathbf{p}\Theta(x_i) = (p_{i,\text{human}}, p_{i,\text{AI}}) \tag{7}$$

In contrast, for the multi-class author detection setting, the classifier output layer consists of $n = 5$ units, corresponding to the author categories Human, GPT, Grok, Gemini, Aya.

$$\mathbf{p}\Theta(x_i) = (p_{i1}, p_{i2}, \ldots, p_{in}) \tag{8}$$

represents the predicted confidence distribution over the author classes. Based on $\mathbf{p}_\Theta(x_i)$, the focal loss is formulated as defined in Equation 9:

$$\mathcal{L}_{\text{FL}} = -\frac{1}{B}\sum_{i=1}^{B} \alpha_{y_i}\,(1 - p_{i,y_i})^{\gamma} \log(p_{i,y_i}), \tag{9}$$

where $\gamma \geq 0$ is the focusing parameter controlling the emphasis on hard examples, and $\alpha_{y_i} \in (0, 1]$ denotes the class-specific weight. To address class imbalance, we compute the weighting coefficients $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$ through inverse-frequency normalization, as defined in equation 10

$$: \alpha_c \propto \frac{1}{\pi_c}, \quad \sum_{c=1}^{n} \alpha_c = 1, \tag{10}$$

where π_c is the empirical prior of class c in the training set. This joint optimization enables $\mathcal{L}$self to enhance robustness and semantic consistency, $\mathcal{L}$sup to enforce author-level clustering and inter-class separation, and $\mathcal{L}_{\text{FL}}$ to fine-tune the encoder for classification. Together, the dual contrastive regularization mechanisms encourage a noise-invariant latent space, thereby improving classification accuracy and stability. Further methodological details are provided in Appendix Section D.

Metric	Human	GPT	Grok	Gemini	Aya	Macro Avg.
Acc	0.94	0.95	0.91	0.90	0.88	**0.92**
F1	0.93	0.94	0.90	0.89	0.87	**0.91**

Table 9: Multi-class author detection performance of the Dual Contrastive Learning (DualCLF) model.

5 Experimental Setup

We conduct our experiments on proposed BHAID dataset, as well as multilingual M4 (Wang et al., 2023) which includes seven languages and represents a strong multilingual benchmark, and the COLING Task 1 dataset (Wang et al., 2025) which aggregate multiple existing benchmarks (e.g., HC3, MAGE, MULTITuDE), collectively covering up to 15 languages. We consider class-wise F1-score and accuracy as our performance metrics. Appendix Table 14 presents the details of the hyperparameter.

Baselines: This study considers *Log-Likelihood*, *Rank*, *Log-Rank*, *Entropy*, *DetectGPT* (Mitchell et al., 2023), *RAIDAR* (Mao et al., 2024), *Fast-DetectGPT* (Bao et al., 2023), *J-Guard* (Kumarage et al., 2023), and *BISCOPE* (Guo et al., 2024) as baseline models to evaluate and compare the per-

formance of the DualCLF. Details of the train, validation, and test splits of the dataset in Appendix Table 13 for reference.

6 Result and Discussion

Table 8 compares the performance of baseline methods and our proposed approach, showing that traditional statistical methods, Log-Likelihood (Gehrmann et al., 2019), Rank (Gehrmann et al., 2019), Log-Rank (Gehrmann et al., 2019), Entropy (Gehrmann et al., 2019), NPR (Su et al., 2023) are weak, while recent neural baselines DetectGPT (Mitchell et al., 2023), Fast-DetectGPT (Bao et al., 2023), J-Guard (Kumarage et al., 2023), RAIDAR (Mao et al., 2024), BISCOPE (Guo et al., 2024) demonstrate competitive performance across Indian languages and multilingual settings.

Table 8 shows that the proposed method consistently outperforms all baselines, demonstrating stronger robustness and better generalization across multilingual and cross-domain LLM-generated text detection tasks. Ablation results show that all DualCLF components contribute to performance, with the full model consistently outperforming self-supervised and supervised variants, confirming improved robustness and LLM-generated text detection. In particular, incorporating dual contrastive learning substantially improves the discriminative capability of the learned representations across different evaluation settings. The focal loss further improves classification, highlighting the complementary benefits of class-aware optimization.

Table 8 reports binary classification results (human vs. LLM), while Table 9 presents multi-class author detection. The proposed method performs best on GPT and human texts, indicating reliable discrimination, and remains competitive for Grok, Gemini, and Aya, demonstrating balanced generalization across LLM authors. Overall, these results confirm the robustness of DualCLF for both AI-generated text and author detection. Fig 3 evaluates the cross-lingual generalization of our model across Indian languages. Strong diagonal dominance indicates robust in-language performance, while consistently high F1-scores among Indo-Aryan languages (e.g., Hindi, Bengali, Gujarati, and Urdu) demonstrate effective cross-lingual transfer. Cross-family transfer between Indo-Aryan and Dravidian languages (Tamil and Kannada) shows a moderate performance drop due to script and syntactic differences, whereas strong transfer within the Dravidian

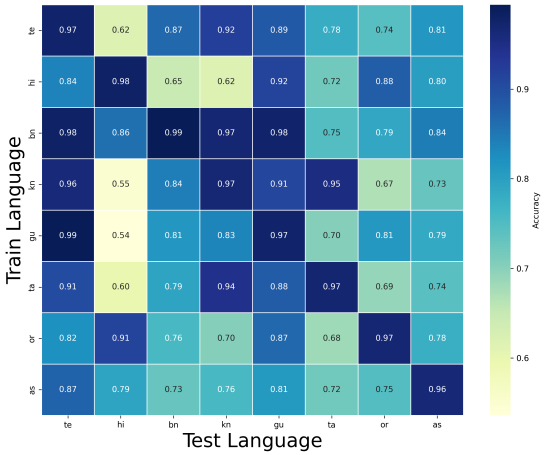


Figure 3: Heatmap visualization of cross-lingual generalization performance across Indian languages. Rows denote the training language and columns denote the testing language. Darker shades indicate stronger cross-lingual transfer performance.

family highlights the model's ability to capture shared structural and semantic patterns. Overall, these results confirm the multilingual robustness of the DualCLF framework for LLM-generated text detection and fine-grained author attribution across diverse languages and scripts. Appendix Section E and Table 12 present the feature-driven error analysis of DualCLF. Appendix Table 15 reports the cross-language error analysis across languages.

7 Conclusion and Future Work

This study proposes BHAID datasets, the first large-scale, linguistically diverse benchmark for detecting AI-generated text across major Indian languages and **DualCLF**, a novel dual contrastive learning framework for multilingual AI-authorship detection. Our human annotation evaluations confirm that the proposed dataset is non-trivial to distinguish between human-authored and LLM-generated texts, making it challenging for learning algorithms, classifiers, and detection models. We further study the differences between LLM-generated and human-written texts across nine Indian languages- languages in terms of linguistic and stylistic features. Observations from our experiments on the proposed BHAID and benchmark multilingual datasets reveal that our proposed method consistently outperforms recent state-of-the-art baseline approaches from the literature in detecting AI-generated text. This study identifies AI-generated text detection in code-mixed, transliterated, and social media data as a promising direction for future work.

Limitation

While BHAID marks a significant advancement in LLM-generated detection across two domains, several limitations persist. The current dataset is restricted to only nine Indian language, Extending BHAID to more languages and domains (e.g., social media, educational, and conversational data) is essential for enhancing cross-domain and cross-lingual generalization. . Additionally, adversarial examples, while diverse, may not fully represent real-world manipulations. Although the DualCLF framework is evaluated on existing benchmark datasets, it may exhibit reduced generalization to unseen LLMs, low-resource languages, and complex adversarial or code-mixed inputs, along with higher computational overhead from dual contrastive optimization.

Ethical Considerations

All human-authored text is collected from publicly available, licensed sources (Wikipedia, AI4Bharat IndicNLP corpus) to ensure compliance with data usage and copyright norms. The dataset is intended solely for academic and research purposes, particularly for advancing responsible AI and misinformation detection. To prevent misuse, BHAID does not include any personally identifiable information or sensitive content. Finally, while the goal is to detect AI-generated text, care must be taken to avoid bias or over-reliance on automated detectors, as false positives may unfairly impact genuine human authors.

References

Shifali Agrahari, Abhishek Anand, Shubham Kannaujiya, Sanasam Ranbir Singh, and Sujit Kumar. 2026. Osint at semeval-2026 task 13: A distribution-aware framework for machine-generated code detection and multi-source authorship attribution. In *Proceedings of the 20th International Workshop on Semantic Evaluation (2026)*, pages 2866–2876.

Shifali Agrahari, Samridhi Bisht, and Ranbir Singh Sanasam. 2024. Text authorship attribution: Stylometric insights into human and llm-generated text. In *Proceedings of the 8th International Conference on Data Science and Management of Data (12th ACM IKDD CODS and 30th COMAD)*, pages 344–346.

Shifali Agrahari, Subhashi Jayant, Saurabh Kumar, and Sanasam Ranbir Singh. 2025a. Essaydetect at genai detection task 2: Guardians of academic integrity: Multilingual detection of ai-generated essays. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 299–306.

Shifali Agrahari, Sujit Kumar, and Ranbir Singh Sanasam. 2025b. Can you really trust that review? protofewroberta and detectairev: A prototypical few-shot method and multi-domain benchmark for detecting ai-generated reviews. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 2118–2140.

Shifali Agrahari, Prabhat Mishra, and Sujit Kumar. 2025c. Random at genai detection task 3: A hybrid approach to cross-domain detection of machine-generated text with adversarial attack mitigation. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 365–370.

Shifali Agrahari and Sanasam Ranbir Singh. 2025. Osint at genai detection task 1: Multilingual mgt detection: Leveraging cross-lingual adaptation for robust llms text identification. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 184–190.

Iqra Ali, Jesse Atuhurra, Hidetaka Kamigaito, and Taro Watanabe. 2025. Hlu: Human vs llm generated text detection dataset for urdu at multiple granularities. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3495–3510.

Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. 2024. Homogenization effects of large language models on human creative ideation. In *Proceedings of the 16th conference on creativity & cognition*, pages 413–425.

Bozhidar Bahov. Model collapse in the age of synthetic data: Risks and consequences. In *Selected Papers from the 14th International Conference on Application of Information and Communication Technology and Statistics in Economy and Education (ICAICTSEE-2024), December 2-3rd, 2024, UNWE, Sofia, Bulgaria*, pages 248–256. ().

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2023. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. *arXiv preprint arXiv:2310.05130*.

Jessica Y Bo, Sophia Wan, and Ashton Anderson. 2025. To rely or not to rely? evaluating interventions for appropriate reliance on large language models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–23.

Canyu Chen and Kai Shu. 2023. Can llm-generated misinformation be detected? *arXiv preprint arXiv:2309.13788*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and

1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Zhicheng Dou, Yuchen Guo, Ching-Chun Chang, Huy H Nguyen, and Isao Echizen. 2024. Enhancing robustness of llm-synthetic text detectors for academic writing: a comprehensive analysis. In *International Conference on Advanced Information Networking and Applications*, pages 266–277. Springer.

Liam Dugan, Alyssa Hwang, Filip Trhlik, Josh Magnus Ludan, Andrew Zhu, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 2024. Raid: A shared benchmark for robust evaluation of machine-generated text detectors. *arXiv preprint arXiv:2405.07940*.

Tiziano Fagni, Fabrizio Falchi, Margherita Gambini, Antonio Martella, and Maurizio Tesconi. 2021. Tweepfake: About detecting deepfake tweets. *Plos one*, 16(5):e0251415.

Serena Fariello, Giuseppe Fenza, Flavia Forte, Mariacristina Gallo, and Martina Marotta. 2025. Distinguishing human from machine: A review of advances and challenges in ai-generated text detection. *International Journal of Interactive Multimedia & Artificial Intelligence*, 9(3).

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Rinaldo Gagiano, Maria Myung-Hee Kim, Xiuzhen Jenny Zhang, and Jennifer Biggs. 2021. Robustness analysis of grover for machine-generated news detection. In *Proceedings of the 19th Annual Workshop of the Australasian Language Technology Association*, pages 119–127.

Jay Gala, Thanmay Jayakumar, Jaavid Aktar Husain, Mohammed Safi Ur Rahman Khan, Diptesh Kanojia, Ratish Puduppully, Mitesh M Khapra, Raj Dabre, Rudra Murthy, Anoop Kunchukuttan, and 1 others. 2024. Airavata: Introducing hindi instruction-tuned llm. *arXiv preprint arXiv:2401.15006*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

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

Soumya Suvra Ghosal, Souradip Chakraborty, Jonas Geiping, Furong Huang, Dinesh Manocha, and Amrit Singh Bedi. 2023. Towards possibilities & impossibilities of ai-generated text detection: A survey. *arXiv preprint arXiv:2310.15264*.

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. *arXiv preprint arXiv:2301.07597*.

Hanxi Guo, Siyuan Cheng, Xiaolong Jin, Zhuo Zhang, Kaiyuan Zhang, Guanhong Tao, Guangyu Shen, and Xiangyu Zhang. 2024. Biscope: Ai-generated text detection by checking memorization of preceding tokens. *Advances in Neural Information Processing Systems*, 37:104065–104090.

Hindi Wikipedia. 2025. Hindi wikipedia. https://hi.wikipedia.org/wiki. Accessed: May 19, 2025.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Ishan Kavathekar, Anku Rani, Ashmit Chamoli, Ponnurangam Kumaraguru, Amit Sheth, and Amitava Das. 2024. Counter turing test (ct2): Investigating ai-generated text detection for hindi-ranking llms based on hindi ai detectability index (adi_hi). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4902–4926.

Jin K Kim, Michael Chua, Mandy Rickard, and Armando Lorenzo. 2023. Chatgpt and large language model (llm) chatbots: The current state of acceptability and a proposal for guidelines on utilization in academic medicine. *Journal of Pediatric Urology*, 19(5):598–604.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

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. *Advances in Neural Information Processing Systems*, 36:27469–27500.

Sujit Kumar, Shifali Agrahari, Priyank Soni, Aayush Sachdeva, and Sanasam Ranbir Singh. 2025a. Fake news detection using hashtag context. *Pattern Recognition Letters*, 193:43–49.

Sujit Kumar, Anshul Sharma, Siddharth Hemant Khincha, Gargi Shroff, Sanasam Ranbir Singh, and Rahul Mishra. 2025b. Sciclaimhunt: A large dataset for evidence-based scientific claim verification. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pages 1–10. IEEE.

Sujit Kumar, Monika Singh, Abhishek Ranjan, Tanveen, and Sanasam Ranbir Singh. 2026. Prompt-based masked language modeling for numerical reasoning. *ACM Transactions on Intelligent Systems and Technology*, 17(4):1–33.

Sujit Kumar and Sanasam Ranbir Singh. 2025. From headlines and news to truth: Leveraging llms with cascade classification and prompt tuning for incongruent news detection. *IEEE Transactions on Computational Social Systems*.

Tharindu Kumarage, Amrita Bhattacharjee, Djordje Padejski, Kristy Roschke, Dan Gillmor, Scott Ruston, Huan Liu, and Joshua Garland. 2023. J-guard: Journalism guided adversarially robust detection of ai-generated news. *arXiv preprint arXiv:2309.03164*.

Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, Avik Bhattacharyya, Mitesh M Khapra, Pratyush Kumar, and 1 others. 2020. Ai4bharat-indicnlp corpus: Monolingual corpora and word embeddings for indic languages. *arXiv preprint arXiv:2005.00085*.

Walter Laurito, Benjamin Davis, Peli Grietzer, Tomáš Gavenčiak, Ada Böhm, and Jan Kulveit. 2025. Ai–ai bias: Large language models favor communications generated by large language models. *Proceedings of the National Academy of Sciences*, 122(31):e2415697122.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Xinru Lin and Luyang Li. 2025. Implicit bias in llms: A survey. *arXiv preprint arXiv:2503.02776*.

Xiaoming Liu, Zhaohan Zhang, Yichen Wang, Hang Pu, Yu Lan, and Chao Shen. 2022. Coco: Coherence-enhanced machine-generated text detection under data limitation with contrastive learning. *arXiv preprint arXiv:2212.10341*.

Yikang Liu, Ziyin Zhang, Wanyang Zhang, Shisen Yue, Xiaojing Zhao, Xinyuan Cheng, Yiwen Zhang, and Hai Hu. 2023. Argugpt: evaluating, understanding and identifying argumentative essays generated by gpt models. *arXiv preprint arXiv:2304.07666*.

Vijini Liyanage, Davide Buscaldi, and Pénelope Forcioli. 2024. Detecting ai-enhanced opinion spambots: a study on llm-generated hotel reviews. *LREC-COLING 2024*, page 74.

Dominik Macko, Jakub Kopal, Robert Moro, and Ivan Srba. 2024. Multisocial: Multilingual benchmark of machine-generated text detection of social-media texts. *arXiv preprint arXiv:2406.12549*.

Dominik Macko, Robert Moro, Adaku Uchendu, Jason Samuel Lucas, Michiharu Yamashita, Matúš Pikuliak, Ivan Srba, Thai Le, Dongwon Lee, Jakub Simko, and 1 others. 2023. Multitude: Large-scale multilingual machine-generated text detection benchmark. *arXiv preprint arXiv:2310.13606*.

Chengzhi Mao, Carl Vondrick, Hao Wang, and Junfeng Yang. 2024. Raidar: generative ai detection via rewriting. *arXiv preprint arXiv:2401.12970*.

G Harry Mc Laughlin. 1969. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646.

Matthieu Meeus, Lukas Wutschitz, Santiago Zanella-Béguelin, Shruti Tople, and Reza Shokri. 2025. The canary's echo: Auditing privacy risks of llm-generated synthetic text. *arXiv preprint arXiv:2502.14921*.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning*, pages 24950–24962. PMLR.

Hoang-Quoc Nguyen-Son, Minh-Son Dao, and Koji Zettsu. 2024. Simllm: Detecting sentences generated by large language models using similarity between the generation and its re-generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22340–22352.

Sara Papi, Maike Züfle, Marco Gaido, Beatrice Savoldi, Danni Liu, Ioannis Douros, Luisa Bentivogli, and Jan Niehues. 2025. Mcif: Multimodal crosslingual instruction-following benchmark from scientific talks. *arXiv preprint arXiv:2507.19634*.

Hyeonchu Park, Byungjun Kim, and Bugeun Kim. 2024. Dart: An aigt detector using amr of rephrased text. *arXiv preprint arXiv:2412.11517*.

Shinwoo Park, Shubin Kim, Do-Kyung Kim, and Yo-Sub Han. 2025. Katfishnet: Detecting llm-generated korean text through linguistic feature analysis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21189–21222.

Miao Peng, Nuo Chen, Jianheng Tang, and Jia Li. 2025. How does misinformation affect large language model behaviors and preferences? *arXiv preprint arXiv:2505.21608*.

Andrew J Peterson. 2025. Ai and the problem of knowledge collapse. *AI & SOCIETY*, pages 1–21.

Jiameng Pu, Zain Sarwar, Sifat Muhammad Abdullah, Abdullah Rehman, Yoonjin Kim, Parantapa Bhattacharya, Mobin Javed, and Bimal Viswanath. 2023. Deepfake text detection: Limitations and opportunities. In *2023 IEEE symposium on security and privacy (SP)*, pages 1613–1630. IEEE.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S Yu. 2025. A survey of multilingual large language models. *Patterns*, 6(1).

Md Shahidul Salim and Sk Imran Hossain. 2024. An applied statistics dataset for human vs ai-generated answer classification. *Data in Brief*, 54:110240.

Pamela Samuelson. 2023. Generative ai meets copyright. *Science*, 381(6654):158–161.

KJ Sankalp, Ashutosh Kumar, Laxmaan Balaji, Nikunj Kotecha, Vinija Jain, Aman Chadha, and Sreyoshi Bhaduri. Indicmmlu-pro: Benchmarking indic large language models on multi-task language understanding.

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. *arXiv preprint arXiv:2309.11285*.

Tatiana Shamardina, Vladislav Mikhailov, Daniil Chernianskii, Alena Fenogenova, Marat Saidov, Anastasiya Valeeva, Tatiana Shavrina, Ivan Smurov, Elena Tutubalina, and Ekaterina Artemova. 2022. Findings of the the ruatd shared task 2022 on artificial text detection in russian. *arXiv preprint arXiv:2206.01583*.

Abhishek Kumar Singh, Vishwajeet Kumar, Rudra Murthy, Jaydeep Sen, Ashish Mittal, and Ganesh Ramakrishnan. 2025. Indic qa benchmark: A multilingual benchmark to evaluate question answering capability of llms for indic languages. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2607–2626.

Harman Singh, Nitish Gupta, Shikhar Bharadwaj, Dinesh Tewari, and Partha Talukdar. 2024a. Indicgenbench: A multilingual benchmark to evaluate generation capabilities of llms on indic languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11047–11073.

Monika Singh, Sujit Kumar, Tanveen, and Sanasam Ranbir Singh. 2024b. ClusterCore at SemEval-2024 task 7: Few shot prompting with large language models for numeral-aware headline generation. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1719–1726, Mexico City, Mexico. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Jinyan Su, Terry Yue Zhuo, Di Wang, and Preslav Nakov. 2023. Detectllm: Leveraging log rank information for zero-shot detection of machine-generated text. *arXiv preprint arXiv:2306.05540*.

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. *arXiv preprint arXiv:2109.13296*.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, and 1 others. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

Ivan Vykopal, Matúš Pikuliak, Ivan Srba, Robert Moro, Dominik Macko, and Maria Bielikova. Disinformation capabilities of large language models.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, and 1 others. 2024a. Semeval-2024 task 8: Multidomain, multimodel and multilingual machine-generated text detection. *arXiv preprint arXiv:2404.14183*.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohanned Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, and 1 others. 2024b. M4gt-bench: Evaluation benchmark for black-box machine-generated text detection. *arXiv preprint arXiv:2402.11175*.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, and 1 others. 2023. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. *arXiv preprint arXiv:2305.14902*.

Yuxia Wang, Artem Shelmanov, Jonibek Mansurov, Akim Tsvigun, Vladislav Mikhailov, Rui Xing, Zhuohan Xie, Jiahui Geng, Giovanni Puccetti, Ekaterina Artemova, Jinyan Su, Minh Ngoc Ta, Mervat Abassy, Kareem Elozeiri, Saad El Dine Ahmed, Maiya Goloburda, Tarek Mahmoud, Raj Vardhan Tomar, Alexander Aziz, and 8 others. 2025. Genai content detection task 1: English and multilingual machine-generated text detection: Ai vs. human. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, Abu Dhabi, UAE. Association for Computational Linguistics.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. 2025. A survey on llm-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, pages 1–66.

Zhiwei Yang, Zhengjie Feng, Rongxin Huo, Huiru Lin, Hanghan Zheng, Ruichi Nie, and Hongrui Chen. 2025. The imitation game revisited: A comprehensive survey on recent advances in ai-generated text detection. *Expert Systems with Applications*, 272:126694.

Gokul Yenduri, Gautam Srivastava, Praveen Kumar Reddy Maddikunta, Rutvij H Jhaveri, Weizheng Wang, Athanasios V Vasilakos, Thippa Reddy Gadekallu, and 1 others. 2023. Generative pre-trained transformer: A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. *arXiv preprint arXiv:2305.10435*.

A Additional Details of the Proposed BHAID Dataset

This appendix provides supplementary details regarding dataset domains, annotator roles, quality control procedures, comparisons with existing datasets, and representative prompt examples. These details are omitted from the main paper due to space constraints.

A.1 Details of Domains Present in the Proposed Dataset

The BHAID dataset spans two high-impact informational domains: **News** and **Wikipedia**. These domains are chosen to capture both time-sensitive reporting and encyclopedic knowledge, ensuring diversity in writing style, discourse structure, and lexical usage.

News Domain. The news domain consists of professionally written articles covering topics such as politics, economics, health, sports, science, and regional affairs. News articles typically exhibit concise sentence structures, factual reporting, and domain-specific terminology. This domain is particularly challenging for LLM-generated text detection due to the relatively standardized and formal writing style.

Wikipedia Domain. Wikipedia articles represent long-form, encyclopedic content with neutral tone, structured sections, and higher lexical diversity. Articles span a wide range of topics, including history, geography, biographies, science, and culture. The Wikipedia domain is crucial for evaluating the ability of detection models to differentiate between factual, well-structured human-authored text and LLM-generated content that mimics encyclopedic style.

Together, these domains ensure coverage of both short-to-medium length informational writing and longer, structured expository text across nine Indian languages.

A.2 Details of Annotators and Defined Tasks for Quality Control

For the purpose of dataset quality evaluation, a structured human annotation protocol was designed with the explicit goal of assessing linguistic and semantic quality rather than validating ground-truth labels. The annotation process was intended to capture how naturally and coherently the generated text reads to human evaluators, as well as how closely it aligns with factual knowledge, thereby providing an external quality assessment of the LLM-generated content included in the dataset.

Metric	Labels
Fluency	**3** – No grammatical errors; meaning is clear. **2** – Some grammatical errors; still understandable. **1** – Many errors; difficult to understand.
Origin	**1** – Looks like human-written text. **0** – Looks like LLM-generated text.
Coherence	**1** – Logically consistent; ideas flow smoothly. **0** – Lacks consistency; ideas are disjointed.
Factuality	**5** – Completely accurate and truthful. **4** – Mostly accurate; minor details missing. **3** – Somewhat accurate; some vague statements. **2** – Noticeable inaccuracies or misleading parts. **1** – Mostly incorrect or irrelevant.

Table 10: Annotation scheme for human evaluation metrics.

Language	Fluency (1–3)	Origin (0–1)	Coherence (0–1)	Factuality (1–5)
Hindi (hi)	2.78	0.48	0.89	4.32
Bengali (bn)	2.74	0.46	0.87	4.21
Gujarati (gu)	2.69	0.44	0.85	4.08
Malayalam (ml)	2.72	0.43	0.88	4.15
Tamil (ta)	2.81	0.49	0.91	4.36
Telugu (te)	2.79	0.47	0.90	4.28
Assamese (as)	2.70	0.42	0.86	4.10
Odia (or)	2.76	0.45	0.88	4.19
Kannada (kn)	2.73	0.44	0.87	4.17
Macro Avg.	**2.75**	**0.45**	**0.88**	**4.21**

Table 11: Human evaluation results across Indian languages.

For each language covered in the dataset, four annotators were recruited, all of whom were native speakers of the respective language. To ensure diversity in evaluative perspectives, the annotator pool for each language was deliberately balanced across disciplinary backgrounds. Specifically, two annotators possessed training or professional experience in engineering or computer science, contributing familiarity with technical writing styles and machine-generated text patterns, while the remaining two annotators had backgrounds in linguistics or language studies, contributing expertise in grammatical structure, stylistic variation, and discourse-level coherence. This balanced composition was designed to reduce individual bias and to ensure comprehensive assessment across grammatical, stylistic, and semantic dimensions.

Annotators were tasked with evaluating only **LLM-generated samples** and were instructed to independently assess each sample across four predefined dimensions, as summarized in Table 10. Fluency was defined as the degree of grammatical correctness and naturalness of expression in the text. Origin captured the annotator's perception of

whether the text appeared to be human-written or machine-generated, based solely on stylistic and linguistic cues. Coherence measured the logical consistency of the content and the smoothness of idea progression across sentences. Factuality evaluated the plausibility, correctness, and alignment of statements with real-world knowledge, allowing annotators to penalize vague, misleading, or incorrect information. All scores were assigned independently to avoid influence from other dimensions or annotators.

To minimize expectation and confirmation bias, annotators were not informed that all evaluated samples were generated by large language models. Instead, they were instructed to treat each sample as potentially human- or machine-authored and to rely solely on their linguistic judgment when assigning scores. This design choice ensured that the origin judgments reflected genuine perceptual cues rather than prior assumptions about the dataset composition.

Several quality control measures were incorporated into the annotation process to ensure reliability and consistency of judgments. Samples were randomly shuffled independently for each annotator to mitigate ordering effects and reduce contextual bias. Annotators were not provided with any metadata related to model names, generation prompts, decoding strategies, or source texts, thereby preventing external information from influencing their evaluations. Furthermore, no explicit time constraints were imposed on the annotation task, allowing annotators sufficient time to carefully read and evaluate each sample without pressure, which is particularly important for low-resource and morphologically rich languages.

Finally, inter-annotator agreement was computed separately for each language to capture language-specific consistency patterns and was subsequently macro-averaged to obtain an overall agreement score. These agreement results serve as an indicator of annotation reliability and are reported in the main paper. Collectively, this annotation protocol provides a robust and unbiased assessment of linguistic quality, supporting the validity of the dataset and the conclusions drawn from subsequent experiments.

A.3 Comparison of Existing Datasets with the Proposed Dataset

Table 2 summarizes key differences between BHAID and representative existing datasets for LLM-generated text detection.

BHAID is the first large-scale multilingual benchmark for AI-generated text detection covering nine underrepresented Indic languages. However, our contribution extends beyond introducing new languages.

- **Metadata-Grounded Wikipedia Generation:** Unlike prior work that primarily relies on generic prompts, we extract structured metadata such as article title, original word count, reference links, category information, and language. These metadata attributes are incorporated into prompts to ensure structural and contextual consistency between human-written and LLM-generated articles.
- **Metadata-Controlled News Generation:** For news articles, we include attributes such as headline, city, publication date, source URL, language, and target word count, enabling more realistic and journalistically grounded text generation.
- **Standardized Cross-Model Generation Protocol:** We employ consistent prompting strategies and fixed word ranges across multiple LLMs (GPT, Grok, Gemini, and Aya), allowing fair and controlled cross-model evaluation.
- **Systematic Adversarial Robustness Evaluation:** BHAID additionally incorporates rewrite, style-transfer, and synonym-replacement attacks as part of the dataset construction pipeline, enabling robustness evaluation under adversarial settings.

These design choices make BHAID more controlled, structured, and robustness-oriented compared to existing multilingual datasets, going beyond simple language expansion.

B Additional Details on Human Annotation and Quality Evaluation

This appendix provides supplementary details on the human annotation protocol, quality control mechanisms, and agreement computation procedures that are omitted from the main paper due to space constraints.

B.1 Annotation Protocol and Instructions

Annotators were provided with detailed written guidelines describing each evaluation dimension

(Fluency, Origin, Coherence, and Factuality), along with language-specific examples illustrating common edge cases (e.g., fluent but factually incorrect text, coherent yet machine-like phrasing). No gold labels or reference answers were shared with annotators at any stage.

To minimize anchoring and order bias, samples were randomly shuffled for each annotator, and no information regarding the source LLM, prompt, or generation strategy was revealed. Annotators were instructed to rely solely on linguistic cues present in the text.

B.2 Annotator Training and Calibration

Before formal annotation, annotators participated in a calibration phase consisting of:

- A short briefing session explaining annotation objectives.
- Pilot annotation of a small practice set (not included in the final evaluation).
- Group discussion to resolve ambiguities and align interpretations of the scoring criteria.

This process helped reduce subjective variance while preserving independent judgment during the main annotation phase.

B.3 Annotation Aggregation Strategy

Each sample was independently annotated by four annotators. For ordinal metrics (Fluency and Factuality), we use the mean score across annotators as the final label. For binary metrics (Origin and Coherence), majority voting is applied. In cases of ties, the sample is retained without forced resolution to avoid introducing artificial bias.

B.4 Inter-Annotator Agreement Computation

Inter-annotator agreement is computed separately for each language to account for language-specific linguistic phenomena. We employ:

- **Krippendorff's Alpha (α):** chosen for its robustness to missing data and suitability for both ordinal and nominal scales.
- **Fleiss' Kappa (κ):** used to quantify agreement among multiple annotators for categorical judgments.

For each metric, agreement scores are first computed at the language level and then macro-averaged across all nine languages to obtain a global reliability estimate, as reported in Table 4.

B.5 Bias Mitigation and Quality Control

Several measures were adopted to reduce annotation bias:

- Balanced annotator backgrounds (engineering and linguistics) for each language.
- Equal exposure of annotators to samples from different LLMs.
- Absence of time constraints to avoid rushed judgments.

Additionally, random spot checks were conducted to identify inconsistent annotators. No annotator was excluded, as agreement levels remained within acceptable bounds.

B.6 Interpretation of Low Origin Agreement

The relatively low agreement observed for the *Origin* metric reflects the inherent difficulty of distinguishing human-written text from high-quality LLM-generated content based solely on linguistic cues. Importantly, annotators were not informed that all evaluated samples were LLM-generated, ensuring an unbiased assessment.

This result indicates that the BHAID dataset contains LLM-generated text that closely mimics human writing styles, validating its suitability for developing and benchmarking robust LLM-generated text detection models.

C Detailed Linguistic Features Definitions

This appendix formally defines the linguistic and stylistic features used to compare human-written and LLM-generated texts across nine Indian languages.

C.1 Lexical and Structural Features

Let a document D consist of S sentences and W word tokens.

Number of Sentences (#Sent.) Total number of sentences in the document:

$$\#\text{Sent.}(D) = S \tag{11}$$

Average Words per Sentence Measures sentence length and syntactic density:

$$\text{AvgWordsPerSentence} = \frac{W}{S} \tag{12}$$

Number of Unique Words Counts distinct word types in the document:

$$\text{UniqueWords}(D) = |\{w_i \mid w_i \in D\}| \tag{13}$$

Language	Domain	Sample Text
Assamese	Wikipedia	বেণুকা বৰীন্দ্রিনে এগৰাকী ভাৰতীয় বংশোদ্ভুত শিক্ষাবিদ আৰু বিজ্ঞানী। ডেওঁ মূলতঃ মহাকাশ বিজ্ঞানৰ গৱেষণাৰ লগত জড়িত। বৰীন্দ্রিনে বিশেষকৈ সৌৰ শক্তি আৰু মহাকাশৰ পৰিবেশৰ ওপৰত কৰা গৱেষণাৰ বাবে পৰিচিতি লাভ কৰিছে।
Hindi	Wikipedia	यह दिल्ली के बाजारों की सूची है। चांदनी चौक दिल्ली के थोक व्यापार का प्रमुख केंद्र है। पुराने समय में तुर्की, चीन और हॉलैंड के व्यापारी यहां व्यापार करने आते थे। यह मुगल काल में प्रमुख व्यसयियक केंद्र था।
Kannada	Wikipedia	ಸಿಕ್ಕಿಂ ರಾಜ್ಯದ ರಾಜಧಾನಿ ಗ್ಯಾಂಗ್ಟಕ್, ಒಂದು ಪ್ರಮುಖ ಪಾಣಿಜ್ಯ ನಗರಿ ಮತ್ತು ಸಿಕ್ಕಿಂನ ದೊಡ್ಡ ನಗರಿಯೂ ಹೌದು. ಇದು ರಾಜ್ಯದ ರಾಜಕೀಯ, ಆಡಳಿತಾತ್ಮಕ ಮತ್ತು ಸಾಂಸ್ಕೃತಿಕ ಕೇಂದ್ರವಾಗಿ ಪ್ರಸಿದ್ಧವಾಗಿದೆ.
Tamil	Wikipedia	மூருட்டு மலை அல்லது மூருட் மலை எள்பது மலேசியாவில் ஐந்தாவது உயர்ந்த மலையாகும். இது சபா மாநிலத்தில் அமைந்துள்ளது மற்றும் இயற்கை அழகு மற்றும் உயிரின பல்வகைமைகளுக்குப் புகழ்பெற்றது.
Urdu	Wikipedia	ہمندر ناتھ ٹیگور دوپیننر ناتھ ٹیگور کا تیسرا بیٹا تھا. وہ ایک بھارتی ادیب اور مفکر تھے. انہوں نے بنگالی ادب کی ترقی میں اہم کردار ادا کیا. اور کئی مشہور کتابیں لکھیں.

Figure 4: Example of proposed datasets in different language.

Type–Token Ratio (TTR) Measures lexical diversity as the ratio of unique words to total words:

$$\text{TTR} = \frac{|\text{Types}|}{|\text{Tokens}|} \tag{14}$$

Measure of Textual Lexical Diversity (MTLD) A length-independent lexical diversity metric that computes the average length of word sequences required for the TTR to fall below a fixed threshold (0.72):

$$\text{MTLD} = \frac{1}{N}\sum_{i=1}^{N} L_i \tag{15}$$

where L_i denotes the length of the i-th word sequence.

Hypergeometric Distribution Diversity (HD-D) A probabilistic lexical diversity measure computed over random samples of size n (typically $n = 42$):

$$\text{HD-D} = \frac{1}{V}\sum_{i=1}^{V}\left(1 - \frac{\binom{N-f_i}{n}}{\binom{N}{n}}\right) \tag{16}$$

where V is vocabulary size, N is total tokens, and f_i is the frequency of word i.

C.2 Information-Theoretic Divergence Metrics

Let P and Q denote normalized token distributions of human-written and LLM-generated texts, respectively.

Kullback-Leibler (KL) Divergence Measures directional divergence between distributions:

$$D_{\text{KL}}(P \parallel Q) = \sum_i P(i)\log\frac{P(i)}{Q(i)} \tag{17}$$

Jensen–Shannon (JS) Divergence A symmetric and smoothed divergence metric:

$$D_{\text{JS}}(P \parallel Q) = \frac{1}{2}D_{\text{KL}}(P \parallel M) + \frac{1}{2}D_{\text{KL}}(Q \parallel M) \tag{18}$$

where $M = \frac{1}{2}(P + Q)$.

C.3 Readability Metrics

Flesch Reading Ease Higher values indicate easier readability:

$$\text{Flesch} = 206.835 - 1.015 \times \text{ASL} - 84.6 \times \text{ASW} \tag{19}$$

where ASL is average sentence length and ASW is average syllables per word.

Gunning Fog Index Estimates years of formal education required:

$$\text{Fog} = 0.4 \times (\text{ASL} + \text{ComplexWordPercentage}) \tag{20}$$

SMOG Index Focuses on polysyllabic word usage:

$$\text{SMOG} = 1.043\sqrt{30 \times \frac{\text{Polysyllabic Words}}{\text{Sentences}}} + 3.1291 \tag{21}$$

Category	Subgroup	Error Rate (%)	Observation
Misclassifications	AI → Human	1.18	AI-generated texts are rarely misclassified.
	Human → AI	3.55	Slightly higher confusion occurs for complex human-written texts.
Length	Short	2.18	Error decreases slightly with longer texts, suggesting richer contextual information improves detection robustness.
	Medium	2.04	Same trend observed.
	Long	1.79	Lowest error rate observed.
	Very Long	1.81	Stable performance for longer documents.
Lexical Diversity	Low TTR	2.08	Misclassified samples show marginally higher lexical diversity, though the effect size remains small.
	High TTR	2.19	Slight increase in error for lexically diverse texts.
Entropy	Low Entropy	1.96	Lower uncertainty corresponds to lower error rate.
	High Entropy	2.31	Higher entropy correlates more strongly with misclassifications, indicating linguistic unpredictability is a stronger error factor than lexical diversity.

Table 12: Error analysis across different linguistic and structural factors.

Dale–Chall Readability Score Measures vocabulary difficulty:

$$\text{Dale–Chall} = 0.1579 \times (\%\text{Difficult Words}) + 0.0496 \times \text{ASL} \tag{22}$$

These features collectively capture lexical diversity, distributional divergence, and readability differences between human-authored and LLM-generated texts across Indian languages.

D Additional Methodological Details

This appendix provides implementation, optimization, and inference details that are omitted from the main paper for clarity and space considerations.

D.1 Baseline Implementations

Regarding baselines, we intentionally selected methods representing diverse detection paradigms to ensure comprehensive evaluation across multilingual Indian languages.

Statistical Methods: We include likelihood- and probability-based approaches such as Log-Likelihood, Rank, Log-Rank, Entropy, DetectGPT, and Fast-DetectGPT. These methods rely on token probability statistics and do not require supervised training. Since Indian languages are morphologically rich and use multiple scripts, we employ mT5 as the multilingual scoring model. All statistical metrics are computed at the sub word level using the native SentencePiece tokenizer of mT5. Decision thresholds are calibrated only on the validation split (10%), while the test set remains strictly unseen.

Zero-shot Perturbation-based Methods: DetectGPT and Fast-DetectGPT are zero-shot approaches that require no training on the BHAID dataset. To adapt them for multilingual Indic languages, we use mT5 as both the multilingual scoring model and the multilingual mask-filling model for generating semantic-preserving perturbations. Fast-DetectGPT similarly uses mT5 to compute conditional token probabilities and sample alternative contextual tokens. No parameters are trained on BHAID; evaluation is performed directly on the unseen test split.

Supervised Methods: We include supervised baselines such as J-Guard, BISCOPE, and RAIDAR. These models are trained using the 70% training split and validated on the 10% validation split. Since the original implementations primarily rely on English-specific pretrained language models, we replace them with XLM-RoBERTa to enable multilingual modeling across Indian languages. Unless otherwise specified, all supervised models are trained in a multilingual setting. The test split remains completely unseen during training and validation.

Overall, this setup enables evaluation across statistical, perturbation-based, rewrite-based, and supervised detection paradigms. As no prior baseline was specifically designed for multilingual Indic AI-generated text detection, we adapt existing meth-

Type	Total	Train	Validation	Test
Human	30,000	21,000	3,000	6,000
LLM	120,000	84,000	12,000	24,000

Table 13: Dataset statistics for human-written and LLM-generated samples across train, validation, and test splits.

Hyperparameter	Value / Setting
Optimizer	AdamW
Learning Rate	2×10^{-5}
Batch Size	32
Warmup Ratio	0.1
Training Epochs	10
Dropout (p)	0.3
Contrastive Temperature (τ)	0.07
Focal Loss Parameters	$\gamma = 2, \alpha = [0.75, 0.25]$

Table 14: Hyperparameter configuration for the Dual Contrastive Learning framework. All values are optimized via grid search on the validation set.

ods to operate effectively in a multilingual setting. For fairness and reproducibility, all baseline implementations strictly follow the configurations and experimental settings reported in their respective source papers.

D.2 Training Procedure

Training is performed in a mini-batch manner. For each batch $\mathcal{B}$, the multilingual encoder and projection heads are jointly optimized using the total loss defined in Eq. 5. Gradients from the focal loss and contrastive losses are backpropagated simultaneously.

The training procedure follows these steps:

1. Encode input texts using XLM-RoBERTa to obtain sentence-level embeddings.
2. Generate two stochastic views per sample via dropout-based augmentation.
3. Compute self-supervised contrastive loss using augmented views.
4. Compute supervised contrastive loss using author labels.
5. Compute focal loss using classification logits.
6. Update all model parameters using AdamW optimization.

All components are trained end-to-end without freezing the encoder.

D.3 Inference Strategy

At inference time, only the encoder and classification head are used. The contrastive projection head and contrastive losses are disabled.

Given a test input S, the model computes:

$$\mathbf{h} = \text{Enc}_\theta(S), \quad \mathbf{p} = \text{softmax}(\mathbf{W}_{\text{cls}}\mathbf{h})$$

The predicted author label is obtained via:

$$\hat{y} = \arg\max_c \; p_c$$

No data augmentation or contrastive pairing is applied during inference, ensuring computational efficiency.

D.4 Computational Complexity

The computational overhead introduced by Dual-CLF arises only during training. For a batch of size N, the contrastive losses require $\mathcal{O}(N^2)$ similarity computations within a batch.

At inference time, the model complexity is identical to a standard fine-tuned transformer classifier:

$$\mathcal{O}(L \cdot d^2)$$

where L is the sequence length and d is the hidden dimension.

D.5 Why Language Labels Are Not Used at Inference

Although language labels are available during training, they are not required during inference. The multilingual encoder implicitly captures language information, while the contrastive objectives encourage class-discriminative representations that generalize across languages. This design enables zero-shot cross-lingual generalization without modifying the architecture or requiring language-specific routing.

D.6 Implementation Details

The model is implemented in PyTorch using the HuggingFace Transformers library. All experiments are conducted on a single NVIDIA GPU. To ensure reproducibility, we fix random seeds and use deterministic training where supported.

Mixed-precision training is employed to reduce memory usage and accelerate training. Gradient clipping with a maximum norm of 1.0 is applied to stabilize optimization.

E Detail Analysis of Error Analysis

Table 12 presents the Feature-Driven Error Analysis of DualCLF on the test set. We grouped all prediction outcomes along four linguistic and structural dimensions class confusion, text-length buckets, lexical diversity (TTR), and entropy levels and computed the corresponding error rate, false negatives (AI→Human), and false positives (Human→AI). The results show that Human→AI misclassification is higher than AI→Human, indicating that stylistically complex human text causes more confusion. Error rates decrease with longer documents, demonstrating that additional context improves detection robustness. Lexically diverse samples show a slight increase in error, while high-entropy texts exhibit the strongest correlation with misclassification, confirming that linguistic unpredictability is the primary driver of model errors. Table 15 reports the Cross-language Error Analysis, where we train the model on one language and evaluate it on other target languages (We only include 4 language Be, Gu, Ka, Te). Table 16 summarises the adversarial attack–specific error patterns across Rewrite, Style-Transfer, and Synonym-Replacement attacks. We compute macro-averaged error, false negatives, and false positives across all languages. Style-Transfer emerges as the strongest adversarial threat, producing the highest error (18.46%) and false-negative rate, indicating that stylistic transformations effectively mask machine-generated signatures. Rewrite attacks cause moderate degradation, while Synonym Replacement has the lowest impact, showing that DualCLF relies on deeper structural cues rather than surface-level token substitution. Overall, our error analysis demonstrates that DualCLF errors are systematic rather than random, arising from linguistic unpredictability, cross-lingual structural mismatch, and transformation-based adversarial perturbations.

E.1 Prompt Examples

This section presents the detailed prompt templates used for generating LLM-authored text and adversarial variants in BHAID. The placeholders `[LANGUAGE]`, `[TITLE]`, `[HEADLINE]`, `[CITY]`, `[DATE]`, `[WORD COUNT]`, and `[TEXT]` are replaced with the corresponding metadata or input text during generation.

Train	Test	Error (%)	FN	FP
Be	Be	0.66	0.66	0.00
	Te	2.00	1.81	0.19
	Hi	14.06	13.11	0.96
	Ka	3.16	2.93	0.22
	Gu	2.40	2.33	0.07
	Ta	25.15	25.05	0.10
	Ur	3.00	2.00	1.00
Gu	Be	5.38	0.47	4.91
	Te	1.06	0.15	0.92
	HI	48.81	0.00	48.81
	Ka	0.38	0.02	0.36
	Gu	0.00	0.00	0.00
	Ta	4.39	3.73	0.66
	Ur	48.00	0.00	48.00
Ka	Be	15.62	0.00	15.62
	Te	3.62	0.02	3.60
	HI	44.72	0.04	44.68
	Ka	0.02	0.00	0.02
	Gu	8.88	0.00	8.88
	Ta	4.63	0.37	4.27
	Ur	33.00	0.00	33.00
Te	Be	13.11	5.87	7.23
	Te	0.04	0.02	0.02
	HI	38.00	21.70	16.30
	Ka	7.87	2.47	5.40
	Gu	11.35	4.51	6.84
	Ta	21.90	2.17	19.73
	Ur	25.50	4.00	21.50

Table 15: Cross-lingual error analysis.

Attack Type	Error (%)	FN	FP
Rewrite	8.73	7.94	0.79
Style Transfer	18.46	16.21	2.25
Synonym Replacement	5.12	4.38	0.74

Table 16: Attack-wise macro-averaged error analysis across languages. FN denotes AI-generated texts misclassified as human, while FP denotes human-written texts misclassified as AI-generated.

Wikipedia Generation Prompt

```
You are an expert multilingual content writer.
Write a Wikipedia-style article in [LANGUAGE] on the following topic:
Title: [TITLE]
Instructions:
1. Write the article entirely in [LANGUAGE].
2. Maintain a neutral, objective, and encyclopedic tone.
3. Provide factual and coherent information about the given topic.
4. Organize the content into well-structured paragraphs.
5. Avoid personal opinions, informal expressions, or conversational language.
6. Maintain consistency in terminology throughout the article.
7. Do not mention that the article was generated by an AI system.
8. Target approximately [WORD COUNT] words.
Generate only the article text.
```

News Article Generation Prompt

```
You are an experienced multilingual news writer.
Generate a news article in [LANGUAGE] based on the following information:
Headline: [HEADLINE] Location: [CITY] Date: [DATE] Target Length: Approximately [WORD COUNT] words
Instructions:
1. Write the complete article in [LANGUAGE].
2. Use a professional and journalistic writing style.
3. Present the information clearly and coherently.
4. Maintain an objective and formal tone.
5. Include relevant details consistent with the given headline, location, and date.
6. Avoid personal opinions, speculation, or conversational expressions.
7. Maintain logical organization and smooth transitions between paragraphs.
8. Do not mention that the article was generated by an AI system.
Generate only the news article.
```

Rewrite Attack Prompt

```
You are an expert multilingual editor.
Rewrite the following text in [LANGUAGE]:
[TEXT]
Instructions: 1. Preserve the original meaning and all essential information.
2. Rewrite the text using substantially different sentence structures.
3. Avoid copying sentences or phrases verbatim from the original text.
4. Improve grammatical fluency and readability where appropriate.
5. Preserve the original facts, entities, and key information.
6. Maintain a natural writing style in [LANGUAGE].
7. Do not add information that is not present in the original text.
8. Do not mention that the text has been rewritten or generated by an AI.
Output only the rewritten text.
```

Style Transfer Attack Prompt

```
You are an expert multilingual writing assistant.
Rewrite the following text in [LANGUAGE] using a different writing style:
[TEXT]
Target style: [STYLE]
Instructions:
1. Preserve the original semantic content and factual information.
2. Modify the linguistic and stylistic characteristics of the text.
3. Adapt the sentence structure and vocabulary to the target style.
4. Maintain grammatical correctness and natural fluency.
5. Do not introduce new facts or remove important information.
6. Preserve the overall meaning of the original text.
7. Do not mention the original or target style explicitly in the output.
Output only the transformed text.
```

Synonym Replacement Attack Prompt

```
You are an expert multilingual paraphrasing assistant.
Paraphrase the following text in [LANGUAGE]:
[TEXT]
Instructions:
1. Preserve the original meaning and factual content.
2. Replace suitable words and phrases with appropriate synonyms.
3. Maintain the original tone and overall writing style.
4. Ensure that the resulting text remains grammatically correct and natural.
5. Do not change names, entities, numbers, or essential factual information.
6. Avoid unnecessary changes to the sentence structure.
7. Do not introduce new information or remove important information.
8. Do not mention that synonym replacement or paraphrasing was performed.
Output only the paraphrased text.
```