

# Meridian: A Bivariate Perplexity Architecture for Trustworthy and Equitable AI Text Detection

Pranil Raichura

Mantis AI Research, MIT CSAIL

Granite Bay High School

raichura@mit.edu

## Abstract

AI text detectors are widely deployed in educational institutions to enforce academic integrity, yet they systematically misidentify authentic writing by non-native English speakers as AI-generated, with false positive rates reaching 61–72% on English as a Second Language (ESL) student essays. This is a population-scale harm that falls disproportionately on the most linguistically vulnerable learners. Retraining detectors on learner data cannot close this gap: large language models can be prompted to mimic ESL writing patterns, and single-stream detectors are theoretically bounded to degrade toward random chance under such style-mimicry attacks. We present **Meridian**, a dual-stream architecture that maps each essay into a bivariate perplexity space with a Native-Expert (DistilGPT-2) axis measuring sub-word predictability and an ESL-Expert (character-level LSTM) axis modeling learner-conditioned character-sequence patterns. Meridian then classifies with an interpretable logistic boundary, deployable in under 2 seconds on CPU. Validated across three LLM families and three learner corpora ( $n=1,590$ ), Meridian reduces wrongful ESL flags by  $15.7\times$  on held-out TOEFL11 with zero retraining, achieving 3.9% overall false positive rate and 0.0% for low-proficiency writers, while maintaining 98.3% detection of AI-mimicked ESL text. These results demonstrate that equitable AI text detection for diverse, non-WEIRD populations is achievable through principled multi-stream feature design that encodes complementary linguistic signals rather than a single native-English reference norm.

## 1 Introduction

### 1.1 A Population-Scale Harm in Education

The deployment of AI text detectors across educational institutions has created an unintended population-scale harm: systematic misidentification of authentic writing by non-native English

speakers as AI-generated (Liang et al., 2023; Pratama, 2025). At the 61.2% false positive rate documented by Liang et al. (2023), an institution processing 1,000 authentic ESL submissions would wrongly flag approximately 612 students for academic dishonesty they did not commit. These students (still developing English proficiency, already navigating unfamiliar academic systems, often from under-resourced educational backgrounds) represent precisely the population that AI bias research has consistently identified as most harmed by WEIRD (Western, Educated, Industrialized, Rich, and Democratic) system defaults (Bender et al., 2021; Blodgett et al., 2020).

The emergence of highly capable generative large language models (LLMs) has intensified this challenge (Guo et al., 2023). AI text detection tools have been rapidly deployed by schools and publishers, functioning as consequential arbiters of academic integrity. A prominent family of detectors relies on likelihood-derived signals: they measure how predictable a piece of text is to a language model trained on standard English, a metric called *perplexity*. AI-generated text is expected to be highly predictable (low perplexity) because the model optimizes for fluent token sequences (Radford et al., 2019). Zero-shot methods such as DetectGPT (Mitchell et al., 2023) and Fast-DetectGPT (Bao et al., 2024) extend this by measuring log-probability curvature, but all remain anchored to a single native-English reference model, which encodes the bias described below.

### A governance failure, not just a technical one.

These tools operate in an institutional context that amplifies their harms: an AI flag can trigger formal academic misconduct proceedings, grade penalties, or expulsion, outcomes with lasting consequences. Proceedings governed by due-process norms demand reliable evidence (Bender et al., 2021); a detector with a 61% false positive rate on a specific

demographic group does not meet this standard. Bassett et al. (2026) argue that AI detection should not be used in education at all, because detectors produce unverifiable probabilistic outputs violating procedural fairness norms. The ESL bias problem is therefore not merely a technical performance deficit but a *governance failure*: the deployment of AI systems in high-stakes institutional settings without adequate auditing for disparate impact. It is also a concrete opportunity for NLP to deliver positive social impact in education: the harm is measurable, the affected population is identifiable, and architectural interventions can be evaluated directly against it.

## 1.2 The ESL Bias Problem

A learner still developing English proficiency naturally uses simpler vocabulary, more predictable sentence structures, and syntax constrained by their native language, a phenomenon documented in Second Language Acquisition (SLA) theory as *interlanguage* (Selinker, 1972; Corder, 1967; Ellis, 1994). This produces text that appears highly predictable to a standard model, *not because it was generated by AI, but because the writer is still learning*. Liang et al. (2023) showed that commercial detectors misclassified 61.2% of authentic TOEFL essays as AI-generated; Pratama (2025) confirmed that modernized tools still flag up to 25% of non-native writing.

To verify this bias on our own corpora, we conducted a systematic red-team audit of five publicly available detectors plus a lexical baseline on our ELLIPSE evaluation set (full results in Table 1, §4.1). The results reveal a *simultaneous double failure*: roberta-base-openai-detector (Solaiman et al., 2019) achieves 72.7% ESL FPR alongside only 37.5% TPR on AI-ESL mimics, worse than random chance, and most audited systems trade one failure mode for the other (§4.1 discusses the two that do not). **The gap it exposes is structural: single-axis perplexity detectors trained on native English cannot simultaneously protect ESL writers from false accusation and catch AI text that mimics ESL patterns, because both failure modes share the same root cause, a single reference distribution that conflates learner errors with AI-generated smoothness.**

## 1.3 The Adversarial Gap

One intuitive mitigation is to retrain detectors on ESL data. However, modern LLMs can be prompted to mimic ESL writing patterns (Gu et al., 2026; Zhou et al., 2024). Even a character-level LSTM trained exclusively on authentic ESL data fails to distinguish real student writing from adversarial AI mimicry. Sadasivan et al. (2023) proved theoretically that any single-stream detector degrades toward random chance under capable paraphrasing attacks, and Cheng et al. (2025) demonstrated universal attacks reducing true positive rates by 87.88%. Our empirical evaluation in Section 4.1 confirms this at scale: the most sophisticated zero-shot detector we tested flags authentic ESL students more than four times as often as it flags AI-generated mimicry, a structural inversion rather than an implementation flaw.

The gap no existing system addresses is this: *simultaneous low false positive rate on ESL writers and high detection rate on AI text that mimics ESL patterns*. Retraining on learner data improves fairness but sacrifices robustness; anchoring to native English improves robustness but destroys fairness. No single perplexity axis can satisfy both constraints at once. A fundamentally different representation is required.

## 1.4 Our Contribution

We make four contributions, prioritized by their societal significance:

1. **A systematic red-team audit of five deployed AI text detectors plus a lexical baseline** across a controlled ESL fairness benchmark, documenting a fairness-robustness tension that is systematic among likelihood-based detectors anchored to a single native-English reference: most either flag ESL writers at rates exceeding 60% or miss a large share of adversarial AI mimics, and even the strongest performer still flags ESL writers at four times its native-writer rate.
2. **A fairness-first problem formulation** that frames AI-text detection on ESL writing as an NLP-for-social-impact problem: current detectors cause measurable population-scale harm to an identifiable learner population, and architectural choices can mitigate or exacerbate it.

3. **A simple, interpretable dual-stream geometry** (Meridian) mapping essays into a bivariate perplexity space where human ESL and AI-mimicked ESL occupy separable regions (Fraser et al., 2025; Ellis, 1994), grounded in SLA interlanguage theory.
4. **Population-scale evidence of harm reduction**: a  $15.7\times$  reduction in false flags on held-out TOEFL11 (3.9% overall FPR; 0.0% for low-proficiency writers), with 98.3% detection of AI-mimicked ESL from three LLM families, with zero retraining on the target population.

**Scope.** Meridian is not presented as a universal AI-text detector. Its contribution is an interpretable, theory-grounded representation for ESL-sensitive detection. Broader adversarial settings, additional baselines, and domain transfer remain future work (see Limitations).

**Three-phase evaluation:** (1) *Does the geometry work?* (ELLIPSE,  $n=300$ ) — controlled same-domain test of geometric separability. (2) *Does it generalize?* (COREFL + 3 LLM families,  $n=900$ ) — cross-corpus and cross-model validation on data entirely disjoint from LSTM training. (3) *Is it fair?* (Zero-shot TOEFL11,  $n=330$ , no retraining) — the primary fairness evaluation on a fully held-out population stratified by L1 background and proficiency level.

## 2 Related Work

### 2.1 Perplexity as a Detection Heuristic and the Field’s Current State

Given a language model  $M$  and token sequence  $X = (x_1, \dots, x_t)$ , perplexity is:

$$\text{PPL}(X) = \exp\left(-\frac{1}{t} \sum_{i=1}^t \log p_M(x_i | x_{<i})\right). \quad (1)$$

DetectGPT (Mitchell et al., 2023) and Fast-DetectGPT (Bao et al., 2024) extend this by measuring log-probability curvature, but both remain anchored to a single native-English reference model and thus inherit the ESL fairness problem.

Wu et al. (2025) provide the field’s most comprehensive survey of LLM-generated text detection, identifying demographic fairness for non-native writers as a critical open problem. Major 2024–2025 benchmarks — M4GT-Bench (Wang et al., 2024), RAID (Dugan et al., 2024), DetectRL (Wu

et al., 2024), and GEDE (Gehring and Paaßen, 2025) — substantially expanded adversarial and educational evaluation, but none evaluates detection fairness for ESL student writing or addresses the simultaneous fairness-robustness constraint. Meridian directly targets this gap.

### 2.2 Demographic Bias and Adversarial Robustness

Liang et al. (2023) documented 61.2% misclassification of TOEFL essays, the finding establishing the ESL fairness problem, and it persists: Pratama (2025), Lege (2025), and Shue et al. (2025) confirm the gap across all tested commercial tools through 2024–2025, with Shue et al. (2025) showing non-native authors were flagged more than US authors even before the post-2022 rise of AI writing. Al Ali et al. (2026) found no systematic bias in Czech detectors (Czech morphological errors *increase* entropy), confirming the English-specificity of the failure, consistent with the broader WEIRD bias documented by Bender et al. (2021) and Blodgett et al. (2020). Basu et al. (2025) introduced BAID (AAAI 2026), the most systematic bias audit to date (200k+ samples, seven sociolinguistic categories including ELL status). BAID documents the problem; Meridian is, to our knowledge, the first architecture to target the *simultaneous* ESL-FPR and adversarial-TPR constraint.

On the adversarial side, Sadasivan et al. (2023) proved that sufficiently powerful paraphrasing degrades any single-stream detector toward random chance; Cheng et al. (2025) demonstrated universal attacks reducing TPR at 1% FPR by 87.88%; and Su et al. (2023) showed style-preserving rewrites degrade detection from 98.65% to 59.91%, precisely the attack modeled by our AI-ESL Mimic class. Wu et al. (2025) identify fairness *and* adversarial robustness as the field’s hardest joint constraint, which is exactly the constraint Meridian addresses. DivEye (Basani and Chen, 2025) achieves strong detection via surprisal diversity but does not address ESL fairness or adversarial mimicry.

### 2.3 Character-Level Modeling and Interlanguage Theory

Sub-word (BPE) tokenization operates above the character level, which can obscure fine-grained morphological structure, a general limitation of sub-word representations noted in recent detectability surveys (Fraser et al., 2025). Whether this specifically disadvantages detection of interlan-

guage error patterns is, to our knowledge, untested in prior work; it motivates our design choice rather than being an established finding we cite as prior evidence. In SLA theory, learner errors are rule-governed outputs of an internal cognitive grammar (Corder, 1967): L1 transfer and overgeneralization produce errors that are *bursty*, *idiosyncratic*, and *non-Markovian* at the character level (Ellis, 1994). Character-level LSTMs (Hochreiter and Schmidhuber, 1997; Graves, 2013) are one natural architecture for learning this kind of structure directly from character sequences. We do not compare against a token-level model trained on the same ESL data to confirm that character-level granularity, rather than simply ESL-population-matched training data, is the operative factor (see Limitations). This interdisciplinary grounding, connecting a documented SLA phenomenon to a measurable NLP feature space, is central to Meridian’s design and to its interpretability.

### 3 Methodology

**Core idea.** Each essay is mapped to a point in a 2D plane: one axis measures how natural the text looks to a character LSTM trained on ESL learner writing ( $P_{\text{esl}}$ ); the other measures how surprising it looks to a native-English language model ( $P_{\text{native}}$ ). Human ESL writers land upper-left (familiar to the ESL model, surprising to the standard one, because their L1 transfer errors produce token sequences no fluent-English model anticipates); AI mimics land lower-left (predictable to both, because surface errors cannot mask the underlying smooth LLM probability landscape). A straight line separates them.

#### 3.1 Datasets

**Phase 1** ( $n=300$ ): 100 authentic ESL essays from ELLIPSE (Crossley et al., 2023) (grades 8–12, 5 persuasive writing prompts; writers are students enrolled in US schools representing diverse L1 backgrounds including Spanish, Arabic, Vietnamese, and other home languages, though L1 labels are not provided in the public release); 100 AI-Native essays (GPT-4.1, fluent AP-level English); 100 AI-ESL Mimic essays (GPT-4.1 prompted to simulate B1-level ESL (Gu et al., 2026; Su et al., 2023)). Full prompts are in Appendix C.

**Phase 2** ( $n=900$ ): 300 COREFL essays (Díaz-Negrillo et al., 2019) (9 L1 backgrounds, all 6 CEFR levels) for human ESL, entirely disjoint from the ELLIPSE corpus used to train the ESL-Expert

LSTM, eliminating any data leakage concern for this phase. AI-ESL mimics from three independent LLM families: GPT-4.1 (OpenAI, 2025), Gemini 2.5 Flash (Google DeepMind, 2025), and Claude 3.5 Sonnet (Anthropic, 2024) (100 each).<sup>1</sup>

**Phase 3** ( $n=330$ ): Stratified sample from TOEFL11 (Liang et al., 2023): 10 essays  $\times$  11 L1 groups  $\times$  3 proficiency bands (low/medium/high). **No TOEFL data were used in training. Both streams and the classifier are frozen from Phase 2.**

#### 3.2 Stream 1: Native-Expert (DistilGPT-2)

We use the Hugging Face `distilgpt2` checkpoint (Wolf et al., 2020), a distilled version of GPT-2 (Radford et al., 2019) produced using the knowledge-distillation methodology of Sanh et al. (2019): a 6-layer autoregressive Transformer with BPE tokenization pre-trained on WebText. We chose a small model deliberately, rather than one of the larger models typical of DetectGPT-family methods, because the Native-Expert’s role here is narrower than in a standalone zero-shot detector: it need only provide a stable native-English perplexity signal to combine with the ESL-Expert stream, not carry accuracy unaided (Table 2 reports the 1D  $P_{\text{native}}$ -only baseline isolating this stream’s standalone contribution), and it must keep the full system CPU-deployable in under 2 seconds per essay (Appendix B), a hard constraint given our target of under-resourced schools. Whether a larger model would improve  $P_{\text{native}}$ ’s power enough to justify its cost is untested.

$$P_{\text{native}} = \exp\left(-\frac{1}{t} \sum_{i=1}^t \log p_{\theta}(x_i | x_{<i})\right). \quad (2)$$

Lower  $P_{\text{native}}$  indicates text aligned with standard native English.

#### 3.3 Stream 2: ESL-Expert (Character-Level LSTM)

A custom character-level LSTM (Hochreiter and Schmidhuber, 1997) in PyTorch (Paszke et al., 2019), trained on the full ELLIPSE corpus ( $\approx 15\text{M}$  characters): one-hot encoding over 104 characters; two stacked LSTM layers, 256 hidden units (dropout 0.3); Adam optimizer ( $\text{lr} = 0.001$ ); 30

<sup>1</sup>At the time of experimentation, these represented frontier-class LLMs. Evaluating Meridian against newer adversaries is an important direction for future work.

epochs on 60-character chunks.

$$P_{\text{esl}} = \exp\left(-\frac{1}{n} \sum_{i=1}^n \log p_{\psi}(c_i | c_{i-60}^{i-1})\right). \quad (3)$$

Here  $c_{i-60}^{i-1}$  denotes the 60 characters preceding  $c_i$ . The LSTM is chosen for three reasons: (a) its recurrent bias naturally captures sequential morphological patterns from one-hot character input; (b) the resulting model is  $\approx 5$  MB, running in  $< 2$  s per essay on CPU with no API required; and (c) the 60-character context provides a compact window in which local morphological and orthographic patterns associated with interlanguage may occur, though the specific window size is a design choice rather than one derived from either [Selinker \(1972\)](#) or [Ellis \(1994\)](#).

### 3.4 Geometric Classification in Bivariate Feature Space

Each essay produces a two-dimensional feature vector  $\mathbf{v}_i = (P_{\text{esl}}^{(i)}, P_{\text{native}}^{(i)})$ . A 2D Logistic Regression classifier is trained on this space (70/30 train/test split, seed 42). The linear classifier is an intentional design choice: Figure 1 shows classes are largely linearly separable, so nonlinear boundaries would add complexity without geometric justification. Interpretability is a hard requirement for high-stakes educational deployment: a linear boundary allows instructors to understand *why* an essay was flagged ([Bender et al., 2021](#)). Learned coefficients are in Appendix D.

**Note on feature independence.** Phase 1 is the only phase with ELLIPSE overlap (the LSTM’s training corpus); Phases 2 and 3 use entirely disjoint corpora (COREFL and TOEFL11). The improving FPR trend (9.7%  $\rightarrow$  8.3%  $\rightarrow$  3.9%) confirms the overlap does not inflate results.

**Relationship to native-vs-ESL classification.** Meridian’s target task is authorship (human vs. AI), not proficiency or native-language identification, and it is not a native-vs-ESL classifier.  $P_{\text{esl}}$  is a population-conditioned reference signal, not a label prediction: it asks how consistent a text is with authentic ESL writing patterns, and that question is answered for every essay regardless of who wrote it, exactly as  $P_{\text{native}}$  asks how consistent a text is with fluent native English for every essay. The two tasks are related only in that solving authorship detection fairly, in our view, requires a reference distribution matched to the writer’s population rather than a

single default one; it does not require first solving, or even attempting, native-vs-ESL classification as a separate step.

## 4 Experiments and Results

### 4.1 Red-Team Audit of Existing Detectors and Baseline

As a primary contribution (and a prerequisite for motivating Meridian’s design), we conducted a systematic fairness audit of five publicly available AI text detectors plus a TF-IDF lexical baseline on 900 essays (300 IMDB native, 300 ELLIPSE ESL, 300 GPT-4o-mini AI-ESL mimic). This is a *red-team evaluation* against a controlled fairness benchmark: we explicitly test for the harmful behavior of disparate impact on ESL writers while simultaneously testing robustness to adversarial mimicry. Results are in Table 1.

Most audited detectors exhibit a clear fairness-robustness tension: systems achieving low ESL FPR miss adversarial mimics, and vice versa. The RoBERTa variants and RADAR flag the majority of authentic ESL writing while detecting well under half of adversarial mimics (RADAR inverts the intended behavior almost completely, at 98.7% ESL FPR), and DeBERTa-v3-base achieves a low 1.0% ESL FPR only by missing a third of mimics. This pattern is consistent with the theoretical impossibility results of [Sadasivan et al. \(2023\)](#).

Two systems, however, do not fit this pattern, and we state this plainly rather than treat the tension as universal. TF-IDF + LR’s 0.0% ESL FPR / 98.6% TPR is a domain artifact: IMDB reviews are lexically trivial to separate from student essays, so the classifier separates corpora, not authorship; Phases 1–3 avoid this confound via same-domain evaluation throughout. ChatGPT-Detector is the more substantive exception (8.0% ESL FPR, 96.7% AI-ESL TPR), tempered by two caveats: the disparity still holds in relative terms (8.0% vs. 2.0% Native FPR is a fourfold higher false-accusation rate for ESL writers on the same system), and since our mimics are GPT-4o-mini while this detector targets ChatGPT-family output, its high TPR here may reflect an in-distribution advantage a cross-generator evaluation would erode; this is untested, so it should be read as a hypothesis, not a limitation we have demonstrated. Our claim is therefore the narrower one: the fairness-robustness tension is systematic among perplexity- and likelihood-based detectors that anchor to a single native-English ref-

Table 1: Multi-detector audit. ESL FPR and Native FPR are false positive rates on genuine human writing; AI-ESL TPR is detection rate on adversarial mimics.

| Detector                              | ESL FPR | Native FPR | AI-ESL TPR |
|---------------------------------------|---------|------------|------------|
| RoBERTa-Base (Solaiman et al., 2019)  | 72.7%   | 94.4%      | 37.5%      |
| RoBERTa-Large (Solaiman et al., 2019) | 96.7%   | 97.3%      | 53.0%      |
| ChatGPT-Detector                      | 8.0%    | 2.0%       | 96.7%      |
| RADAR (Hu et al., 2023)               | 98.7%   | 86.3%      | 100%       |
| DeBERTa-v3-base                       | 1.0%    | 5.0%       | 66.7%      |
| TF-IDF + LR                           | 0.0%    | 0.0%       | 98.6%      |

erence, not that no deployed classifier can score acceptably on both axes of a single benchmark.

We additionally evaluated Fast-DetectGPT (Bao et al., 2024), a state-of-the-art zero-shot method, on the Phase 2 hard-case benchmark. Fast-DetectGPT achieves 54.5% accuracy, **85.0% ESL FPR**, 94.0% AI-ESL TPR, and  $AUC = 0.313$ , *inversely calibrated* on ESL text, demonstrating that the single-axis failure mode can persist even in a substantially more sophisticated zero-shot likelihood-based detector.

#### 4.2 Phase 1: The 1D Breakdown

Under  $P_{\text{native}}$  alone, Human ESL and AI distributions overlap heavily, replicating prior findings (Liang et al., 2023): as Table 2 shows,  $P_{\text{native}}$  ranks the two classes reasonably well in isolation ( $AUC = 0.95$ ), but its default decision threshold collapses toward the AI class, producing near-100% ESL FPR at  $\sim 50\%$  accuracy – a reminder that ranking quality (AUC) and operating-point accuracy are distinct properties, and that  $P_{\text{native}}$  alone is not usable as a fair standalone classifier even though it is informative. Under  $P_{\text{esl}}$  alone, the AI-ESL Mimic distribution overlaps almost perfectly with Human ESL (both  $\approx 2.0\text{--}2.6$ ). The 1D  $P_{\text{esl}}$  classifier achieved 75.0% accuracy,  $FPR = 29.0\%$ , and  $ROC\text{-}AUC = 0.73$ , consistent with the theoretical prediction of Sadasivan et al. (2023).

#### 4.3 Phase 1: Geometric Separation via Dual-Stream Architecture

When essays are mapped into the bivariate space  $[P_{\text{esl}}, P_{\text{native}}]$ , geometric structure emerges (Figure 1): Human ESL occupies the *upper-left* (high  $P_{\text{native}}$ , low  $P_{\text{esl}}$ : L1-transfer errors render subword sequences unpredictable to the Transformer); AI-ESL Mimics cluster *lower-left* (low on both axes: the LLM’s token selection remains anchored to high-probability standard English despite surface errors (Radford et al., 2019; Fraser et al., 2025)).

The 2D classifier achieved 93.3% accuracy (+18.3 pp over 1D), FPR of 9.7% (3 false positives; 66.7% relative reduction), 1 false negative

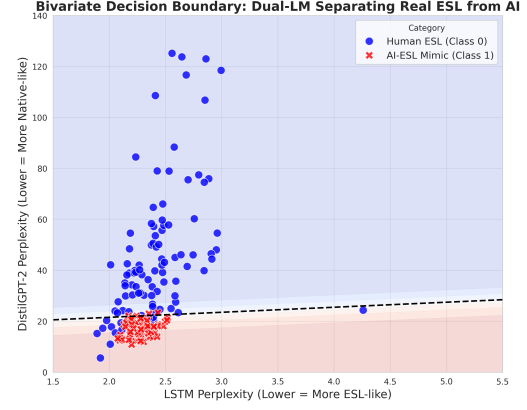


Figure 1: Bivariate decision boundary in Phase 1. Human ESL (blue circles) occupies upper-left; AI-ESL Mimics (red triangles) cluster lower-left. The two axes separate cases that neither alone distinguishes. Dashed line: logistic regression boundary (93.3% accuracy).

Table 2: Performance summary. Primary task: Human ESL vs. AI-ESL Mimic. Fast-DetectGPT (Bao et al., 2024) evaluated zero-shot on Phase 2;  $AUC < 0.5$  indicates its curvature signal is inversely calibrated on ESL text.

| Method                            | Dataset       | Acc.         | ESL FPR      | AI TPR         | AUC          |
|-----------------------------------|---------------|--------------|--------------|----------------|--------------|
| RoBERTa (Solaiman et al., 2019)   | 300 (ELLIPSE) | <50%         | 72.7%        | 37.5%          | –            |
| 1D $P_{\text{native}}$            | 300 (ELLIPSE) | $\sim 50\%$  | $\sim 100\%$ | –              | 0.95         |
| 1D $P_{\text{esl}}$               | 300 (ELLIPSE) | 75.0%        | 29.0%        | 79.3%          | 0.73         |
| Fast-DetectGPT (Bao et al., 2024) | 600 (COREFL)  | 54.5%        | 85.0%        | 94.0%          | 0.313        |
| Binoculars (Hans et al., 2024)    | 600 (COREFL)  | 93.8%        | 1.7%         | 89.3%          | –            |
| <b>2D Ph. 1</b>                   | 300 (ELLIPSE) | <b>93.3%</b> | <b>9.7%</b>  | <b>96.6%</b>   | <b>0.96</b>  |
| <b>2D CV Ph. 2</b>                | 600 (COREFL)  | <b>94.5%</b> | <b>8.3%</b>  | <b>&gt;99%</b> | <b>0.962</b> |
| <b>Zero-shot Ph. 3</b>            | 330 (TOEFL11) | –            | <b>3.9%</b>  | <b>98.3%</b>   | –            |

(vs. 6 for 1D), and  $ROC\text{-}AUC$  of 0.96 (vs. 0.73). *In plain terms: adding the second axis cuts wrongful flags from 9 to 3 and drops missed AI essays from 6 to 1.*

#### 4.4 Phase 2: Cross-Model and Cross-Corpus Generalization

AI-ESL perplexity curves from GPT-4.1, Gemini, and Claude 3.5 Sonnet collapse into the same tight  $P_{\text{esl}}$  band ( $\approx 2.1\text{--}2.6$ ) regardless of model family, supporting the hypothesis that these three tested LLM families share a generation signature under this style-mimicry setting. McFadden’s pseudo- $R^2$  confirms additive discriminative value:  $P_{\text{esl}}$  alone achieves 0.47,  $P_{\text{native}}$  alone 0.65, and the combination 0.73. Five-fold cross-validation yields  $94.5 \pm 0.8\%$  accuracy,  $8.3 \pm 2.8\%$  FPR, and  $ROC\text{-}AUC$  of 0.962. *In plain terms: the geometry holds across three different AI companies and a corpus the system never trained on.*

Table 2 consolidates results across all phases.

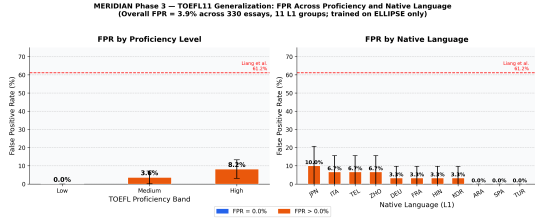


Figure 2: Phase 3 TOEFL11 false positive rates. *Left:* FPR by proficiency band. Low-proficiency writers achieve 0.0% FPR — the inverse of prior detector failure modes. *Right:* FPR by native language (11 L1 groups; trained on ELIPE only); no group exceeds 10%. Error bars: 95% Clopper-Pearson CIs.

#### 4.5 Phase 3: Zero-Shot TOEFL11 Fairness Evaluation

Applying the frozen Phase 2 classifier to 330 TOEFL11 essays yields **overall FPR 3.9%** (13/330), more than  $15\times$  lower than the 61.2% reported by Liang et al. (2023), with *zero TOEFL training data*. Figure 2 visualizes FPR by proficiency band and native language; Table 4 (Appendix F) reports exact per-group values with 95% confidence intervals. No L1 group exceeds 10% FPR, and low-proficiency writers achieve **0.0% FPR** (95% Clopper-Pearson CI: [0.0%, 3.3%],  $n=110$ ), the inverse of the failure mode reported for prior detectors (Liang et al., 2023).

On 60 AI-ESL mimic essays (TOEFL-style prompts, same three LLM families), the frozen classifier detected 59/60 mimics (**TPR 98.3%**; 95% CI: [91.1%, 99.7%]): 20/20 for GPT-4.1, 20/20 for Claude 3.5 Sonnet, 19/20 for Gemini 2.5 Flash.

*In plain terms:* of 330 authentic student essays from 11 language backgrounds, Meridian wrongly flagged only 13, and none of those were from the lowest-proficiency writers most harmed by existing tools. Of 60 AI-mimicked essays, it caught 59.

## 5 Discussion

### 5.1 Why the Geometry Works

The geometric separation admits a mechanistic explanation grounded in SLA theory. The ESL-Expert LSTM asks: “Does this text follow the character-level patterns of a developing learner?” The AI-ESL Mimic can answer “yes”; it mimics average ESL character patterns well enough to fool a 1D detector (Sadasivan et al., 2023; Cheng et al., 2025). The Native-Expert Transformer asks a different question: “Is the sub-word token sequence surprising relative to fluent standard English?” When a human ESL student makes a grammati-

cal error, the specific token choices arise from idiosyncratic L1 transfer (Corder, 1967; Ellis, 1994), producing high and variable  $P_{\text{native}}$ . An LLM generating ESL-style text faces a fundamental tension: maintaining coherent meaning requires anchoring token-level generation to high-probability sequences (Radford et al., 2019). Style-mimicry perturbations cannot simultaneously satisfy both constraints, as evidenced by within-class correlations of  $r=0.21$  (Human ESL) vs.  $r=0.47$  (AI-ESL) and consistent geometric separation across three LLM families.

**A causal account of  $P_{\text{native}}$  elevation.** We propose the following causal chain underlying Meridian’s geometric separation:

L1 background  $\rightarrow$  interlanguage error patterns  
 $\rightarrow$  elevated  $P_{\text{native}}$   $\rightarrow$  reduced FPR

Specifically: a learner’s native language determines the transfer errors they systematically make in English (Corder, 1967; Ellis, 1994); these errors manifest as character- and morpheme-level patterns that the character LSTM associates with authentic ESL writing (low  $P_{\text{esl}}$ ), while simultaneously producing token sequences that DistilGPT-2 finds surprising relative to fluent English (high  $P_{\text{native}}$ ) because the specific token choices arise from idiosyncratic L1 grammar rather than any distribution the Transformer was trained on. The within-class correlations and the Phase 3 error analysis (§5.1) are consistent with this causal pathway, though it is not formally identified; see Limitations.

### Qualitative analysis: what the geometry is capturing.

To test whether the geometric separation reflects genuine interlanguage patterns rather than surface error rate alone, we tagged Phase 1 sentences for five documented L1-transfer error types and compared sentence-level  $P_{\text{native}}$  (full analysis in Appendix E, Table 3). Critically, the gap persists when both classes produce the *same* surface error type: on error-tagged sentences, Human ESL averages  $P_{\text{native}} = 146.4$  versus 82.5 for AI-ESL Mimics ( $\Delta = +63.9$ ,  $p < 0.0001$ ), showing that matched surface error category alone does not explain the observed gap. A matched pair makes the mechanism concrete: the human article-omission sentence “*There in a river finded the frog behind the in trunk*” scores  $P_{\text{native}} = 1,728$ , while a matched AI-ESL sentence carrying the same error tag scores 10.7. The two streams thus capture complementary, non-redundant signals: the

LSTM identifies learner-like character patterns, while DistilGPT-2 identifies the idiosyncratic token choices that genuine L1 transfer produces but AI mimicry cannot replicate without sacrificing fluency.

**Error analysis.** Of the 13 TOEFL11 false positives, 9 (69%) are high-proficiency and 4 (31%) medium-proficiency; *zero are low-proficiency*. False positive essays exhibit markedly lower  $P_{\text{native}}$  (mean 19.7 vs. 66.5 overall;  $t = -4.31$ ,  $p < 0.001$ ), indicating that residual errors concentrate where the human–AI boundary is genuinely ambiguous, not among the vulnerable writers most harmed by existing detectors.

**Cross-L1 validation of the causal account.** A cross-L1 analysis reveals a broader pattern that provides population-level evidence consistent with the proposed mechanism: mean  $P_{\text{native}}$  across L1 groups is significantly negatively correlated with FPR (Spearman  $r = -0.721$ ,  $p = 0.012$ ,  $n = 11$  L1 groups). The three groups with 0.0% FPR (TUR, SPA, ARA) have the highest mean  $P_{\text{native}}$  (76.1, 79.6, 81.7 respectively), while the groups with elevated FPR (ITA: 6.7%, JPN: 10.0%) have the lowest (56.5, 58.6). This pattern is visible and systematic: L1 backgrounds that produce more idiosyncratic token-level errors generate higher  $P_{\text{native}}$ , pushing essays further into the human region of the bivariate space. The elevated FPR for JPN writers in particular concentrates among high-proficiency writers, whose mean  $P_{\text{native}}$  falls to 33.1 (approaching the AI-generated range) and warrants targeted investigation.

**Comparison with baselines.** Table 2 consolidates all methods on the Phase 2 hard-case benchmark (300 Human ESL vs. 300 AI-ESL Mimic, COREFL), a controlled evaluation specifically designed to stress-test the fairness-robustness tradeoff where human ESL and AI-ESL mimics are maximally confounded for single-axis detectors.

Fast-DetectGPT (Bao et al., 2024) achieves 85.0% ESL FPR with AUC=0.313, performing worse than random on this task: its log-probability curvature signal is systematically *more* sensitive to authentic ESL writing than to AI mimicry, confirming that methodological sophistication cannot compensate for a native-English perplexity anchor.

Binoculars (Hans et al., 2024) (ICML 2024) achieves 1.7% ESL FPR but only 89.3% AI-ESL TPR, missing 32 of 300 adversarial mimics that

Meridian catches, a practically significant gap when students can prompt LLMs to mimic ESL patterns. Meridian achieves  $8.3 \pm 2.8\%$  ESL FPR with  $>99\%$  TPR at its own default operating point. We report both methods at their published default thresholds rather than at a matched FPR or via a full ROC comparison, since we do not have access to Binoculars’ per-essay scores on this benchmark to recompute one; the comparison therefore shows that both regimes are achievable, not that Meridian strictly dominates Binoculars across all thresholds. Combined with Fast-DetectGPT’s failure above, no single-axis method we tested reaches Meridian’s combination of low ESL FPR and high AI-ESL TPR at its own default threshold, which is the most we can claim without a matched-threshold analysis (see Limitations).

## 5.2 Implications for Educational Equity and Positive Impact

At the 61.2% FPR documented by Liang et al. (2023), an institution processing 1,000 authentic ESL submissions would expect approximately 612 wrongful flags. At Meridian’s 3.9% zero-shot FPR, this falls to approximately 39, a  $15.7\times$  **reduction in population-scale harm**. This magnitude is comparable to intervention effects in domains such as automated content moderation and algorithmic fairness (Blodgett et al., 2020), where similar-scale reductions in disparate impact are considered policy-relevant achievements. Bender et al. (2021) warned that NLP systems risk encoding existing social inequities; our results demonstrate that principled multi-stream feature design can substantially mitigate this for the specific case of perplexity-based ESL detection. In the framing of NLP for positive impact, the intervention here is direct: a documented educational inequity, an identifiable affected population, and a measurable order-of-magnitude reduction in harm.

Bassett et al. (2026) argue that AI detection should not be used in education at all; we engage this position in Appendix A and note here only that Meridian occupies a harm-reduction stance fully compatible with it: institutions are deploying detection tools at scale regardless of academic argument, and a  $15.7\times$  reduction in wrongful ESL flags is a concrete welfare improvement within that reality.

**Deployment.** Meridian is a screening tool, not a verdict system. Institutions can tune the operating point to their risk tolerance: at threshold 0.50,

overall TOEFL11 FPR is 3.9% (low-proficiency: 0.0%); raising to 0.80 reduces overall FPR to 0.9% while maintaining 94.7% TPR. The DistilGPT-2 stream is  $\approx 82$  MB; the LSTM is  $\approx 5$  MB; inference runs in under 2 seconds per essay on CPU with no internet required, enabling deployment in under-resourced educational contexts where commercial detection subscriptions and GPU infrastructure are unavailable.

**Responsible deployment.** Full deployment guidelines are in Appendix B. In a preliminary consultation, an ELD (English Language Development) educator at a US secondary school confirmed that AI detection flags on ESL students is a recognized practitioner concern. Code, trained weights, and an inference script are available at [github.com/pranilraichura/MERIDIAN-EMNLP-Research](https://github.com/pranilraichura/MERIDIAN-EMNLP-Research).

## 6 Conclusion

Meridian introduces a dual-stream perplexity architecture that substantially reduces false positives on held-out ESL writing while preserving strong detection of adversarial AI mimicry. On zero-shot TOEFL11 (11 L1 backgrounds, 3 proficiency bands, no retraining) overall FPR falls to 3.9% with 0.0% for low-proficiency writers (95% CI: [0.0%, 3.3%]) and 98.3% TPR on AI-ESL mimics from three LLM families. The reversed bias profile (residual errors concentrated among high-proficiency writers rather than the most vulnerable low-proficiency group) is the inverse of the failure mode reported for prior detectors.

The bivariate geometry is grounded in SLA interlanguage theory: the two perplexity axes capture complementary signals that neither axis alone provides, and the resulting linear decision boundary is fully interpretable. Many current AI text detectors can encode native-English-centric assumptions; Meridian demonstrates that equitable detection is achievable through principled architectural choices. More broadly, this work illustrates how NLP methods can be directed at measurable social inequities in education: the harm is quantifiable, the affected population is identifiable, and a principled architectural intervention reduces it by an order of magnitude. Broader adversarial scope, additional baselines, and user studies with ESL educators remain important future work.

## Limitations

**Data and overlap.** All three learner corpora cover persuasive and descriptive essays; generalization to STEM writing is untested. The ESL-Expert LSTM was trained on ELLIPSE character sequences, which also supplies the Phase 1 human ESL essays, the only evaluation overlap. Phase 2 uses 300 COREFL essays entirely disjoint from LSTM training, and Phase 3 uses TOEFL11, which is also completely held out. The improving FPR trend across phases (9.7%  $\rightarrow$  8.3%  $\rightarrow$  3.9%) confirms that the Phase 1 overlap does not inflate downstream results; nonetheless, a cleaner protocol for Phase 1 would exclude evaluation essays from LSTM training.

**Causal identification.** The causal account in §5.1 is consistent with the observed geometry but is not formally identified. Essay prompt topic is a plausible confounder for  $P_{\text{native}}$  elevation that has not been controlled for across matched prompts: if ESL and AI-ESL Mimic essays were generated on systematically different prompts, topic-level differences rather than L1 transfer could explain part of the geometric separation. Controlling for prompt topic across matched human and AI-generated essays would strengthen the causal interpretation.

**Baselines and evaluation scope.** Phase 3 TPR uses 60 AI-ESL mimics; larger matched samples are needed. Per-L1 sample sizes ( $n=30$ ) yield wide confidence intervals (e.g., 0/30 corresponds to 95% CI [0.0%, 11.6%]); larger-scale evaluation is needed to confirm equitable performance across all language backgrounds. Additional same-domain lexical baselines would further strengthen the comparison.

A more consequential gap concerns the ESL-Expert stream specifically. Every baseline we compare against is anchored to native English; the character-level LSTM is the only component in our evaluation trained on ESL data at all. This design cleanly demonstrates that adding an ESL-conditioned reference helps, but it does not isolate *which* property of that reference is responsible, and two distinct attributions remain confounded: that the reference is population-matched, and that it operates at the character level rather than over sub-word tokens. Distinguishing them requires baselines we did not run: a sub-word model fine-tuned on the same ESL corpus (for instance a RoBERTa or small decoder adapted with LoRA at comparable

parameter cost), substituted for the LSTM within the same bivariate architecture. If such a model matched the LSTM’s contribution, the operative factor would be population-matched training data rather than character-level granularity, and our appeal to interlanguage morphology in §5.1 would be weaker than stated. We regard this as the most important open question about the architecture, and we note that our motivation for the character-level choice rests on SLA theory and the qualitative error analysis in Appendix E, not on a controlled comparison against a tokenization-matched alternative.

Our comparison against Binoculars in §5.1 uses each method’s published default threshold rather than a matched-FPR or full ROC comparison, since we do not have access to Binoculars’ per-essay scores on our benchmark to recompute one. The comparison shows that Meridian’s default operating point achieves a combination of ESL FPR and AI-ESL TPR that Binoculars’ default operating point does not, not that Meridian dominates Binoculars at every possible threshold.

**Adversarial scope.** Our AI-ESL mimics are generated with a single generic B1-level prompt per model family (Appendix C); we did not test L1-specific prompting, few-shot exemplars drawn from real learner errors, or iterative/adaptive attacks that adjust to Meridian’s own feedback. A generation strategy targeting a specific L1 background’s documented transfer errors is a plausible stronger adversary than the generic B1 prompts we evaluate, and our TPR numbers should be read as an upper bound on Meridian’s robustness rather than a worst-case estimate. Separately, Meridian was not tested against fine-tuned adversaries (LoRA-adapted models on ESL corpora) or token-level perturbation attacks (Gu et al., 2026; Cheng et al., 2025). We conjecture the bivariate constraint creates structural resistance to the latter (such an adversary would need to simultaneously generate idiosyncratic L1 morphological errors and maintain token-level coherence, requirements that pull in opposite directions), but this remains empirically open, and we make no corresponding conjecture about L1-targeted prompting, which we consider the more immediately testable gap of the two.

**Interpretability validation.** The interpretability advantage is argued on structural grounds but has not been validated with educators. Future directions include user studies with ESL educators, syntactic surprisal as additional geometric dimensions,

multi-L1 LSTM training, mixed-authorship detection, and exploration of formative feedback to learners via error-type  $P_{\text{native}}$  profiling.

## Ethics Statement

All datasets (ELLIPSE, COREFL, TOEFL11) are publicly available and contain no personally identifiable information. No human subjects were recruited; all essays are pre-existing and de-identified. No automated detection system, including Meridian, should be used as the sole basis for academic disciplinary proceedings. A responsible deployment framework is provided in Appendix B.

## Broader Impact Statement

This work directly addresses a documented population-scale harm in educational AI deployment: the systematic misidentification of English-as-a-Second-Language student writing as AI-generated by commercially deployed text detectors. At baseline false positive rates of 61–72%, ESL students in English-medium institutions (the populations represented in our corpora, and by extension students in similar English-medium academic settings globally) face wrongful academic integrity accusations from tools they have no power to contest. This harm is not distributed uniformly: it falls most heavily on students who are already linguistically, economically, and institutionally disadvantaged, consistent with the broader pattern of WEIRD bias in AI systems (Bender et al., 2021; Blodgett et al., 2020).

Meridian demonstrates that this harm can be substantially reduced ( $15.7\times$  reduction in false flags on held-out TOEFL11) through principled architectural design. We explicitly designed for interpretability (linear decision boundary, CPU deployability) to enable adoption in under-resourced educational contexts.

We emphasize that Meridian is a screening tool, not a verdict system. No automated detection output, regardless of accuracy, should substitute for human judgment in academic integrity proceedings. The deployment guidelines in Appendix B specify mandatory human review, threshold selection guidance, and student transparency requirements. The most significant societal risk of this work is misuse as an autonomous arbiter rather than as a triage assistant; we actively discourage such deployment through explicit framing and guidelines.

## Acknowledgments

The author thanks Prof. Mayank Kejriwal (University of Southern California) for his mentorship and guidance throughout this project.

## References

- Adnan Al Ali, Jindřich Helcl, and Jindřich Libovický. 2026. Different time, different language: Revisiting the bias against non-native speakers in GPT detectors. *arXiv preprint arXiv:2602.05769*.
- Anthropic. 2024. [Claude sonnet 3.5: Technical overview](#). Technical Report.
- Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2024. Fast-DetectGPT: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*.
- Advik Raj Basani and Pin-Yu Chen. 2025. Diversity boosts AI-generated text detection. *arXiv preprint arXiv:2509.18880*.
- M. A. Bassett, W. Bradshaw, H. Bornsztejn, A. Hogg, K. Murdoch, B. Pearce, and C. Webber. 2026. [Heads we win, tails you lose: AI detectors in education](#). *Journal of Higher Education Policy and Management*. Published online 29 January 2026.
- Priyam Basu, Yunfeng Zhang, and Vipul Raheja. 2025. [BAID: A benchmark for bias assessment of AI detectors](#). *Preprint*, arXiv:2512.11505. Accepted at AAAI 2026 Workshop on Agentic AI Benchmarks and Applications for Enterprise Tasks.
- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 610–623.
- Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of “bias” in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5454–5476.
- Yize Cheng, Vinu Sankar Sadasivan, Mehrdad Saberi, Shoumik Saha, and Soheil Feizi. 2025. Adversarial paraphrasing: A universal attack for humanizing AI-generated text. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38. ArXiv:2506.07001.
- S. Pit Corder. 1967. The significance of learners’ errors. *IRAL: International Review of Applied Linguistics in Language Teaching*, 5(1–4):161–170.
- Scott Crossley, Yu Tian, Perpetual Baffour, Alex Franklin, Youngmeen Kim, Wesley Morris, Meg Benner, Aigner Picou, and Ulrich Boser. 2023. The EL-LIPSE corpus: A large-scale, longitudinal, learner corpus of written English. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*.
- Ana Díaz-Negrillo, Cristóbal Lozano, and Marcus Callies. 2019. Introducing the Corpus of English as a Foreign Language (COREFL): A bimodal, multi-task corpus for SLA research. Paper presented at the 5th Learner Corpus Research Conference, University of Warsaw.
- 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. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 29069–29113.
- Rod Ellis. 1994. *The Study of Second Language Acquisition*. Oxford University Press, Oxford.
- Virginia Eubanks. 2018. *Automating Inequality*. St. Martin’s Press.
- Kathleen C. Fraser, Hillary Dawkins, and Svetlana Kiritchenko. 2025. [Detecting AI-generated text: Factors influencing detectability with current methods](#). *Journal of Artificial Intelligence Research*, 82:2233–2278. ArXiv:2406.15583.
- Lukas Gehring and Benjamin Paaßen. 2025. Assessing LLM text detection in educational contexts: Does human contribution affect detection? *arXiv preprint arXiv:2508.08096*.
- Google DeepMind. 2025. [Gemini 2.5 flash: A multi-modal AI model](#). Technical Report.
- Alex Graves. 2013. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*.
- Yongtong Gu, Songze Li, and Xia Hu. 2026. MASH: Evading black-box AI-generated text detectors via style humanization. *arXiv preprint arXiv:2601.08564*.
- 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*.
- Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Spotting LLMs with binoculars: Zero-shot detection of machine-generated text. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 17519–17537. PMLR.

- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.
- 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 (NeurIPS)*, volume 36.
- R. Paul Lege. 2025. Auditing the fairness of AI-detection tools: A comparative study of ESL, published, and AI-generated texts and their misclassification risks. *International Journal of Teaching, Learning and Education*. Available online 18 October 2025.
- Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. 2023. GPT detectors are biased against non-native English writers. *Patterns*, 4(7):100779.
- 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 (ICML)*, volume 202, pages 24950–24962.
- Safiya Umoja Noble. 2018. *Algorithms of Oppression*. New York University Press.
- OpenAI. 2025. [Introducing GPT-4.1 in the API](#). Technical Report.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32.
- Ahmad R. Pratama. 2025. [The accuracy-bias trade-offs in AI text detection tools and their impact on fairness in scholarly publication](#). *PeerJ Computer Science*, 11:e2953.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).
- Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. Can AI-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.
- Larry Selinker. 1972. Interlanguage. *IRAL: International Review of Applied Linguistics in Language Teaching*, 10(1–4):209–232.
- Evelyn Shue, Neil K. Jairath, and Gangqing Hu. 2025. [Response to “ChatGPT-polished scientific writing and artificial intelligence detection: Cohorts, baselines, fairness”](#). *JAAD International*.
- Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, and Jasmine Wang. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.
- Zhenpeng Su, Xing Wu, Wei Zhou, Guangyuan Ma, and Songlin Hu. 2023. HC3 Plus: A semantic-invariant human ChatGPT comparison corpus. *arXiv preprint arXiv:2309.02731*.
- Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, Alham Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024. M4GT-Bench: Evaluation benchmark for black-box machine-generated text detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3964–3992. Association for Computational Linguistics.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP)*, pages 38–45.
- 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*, 51(1):275–338.
- Junchao Wu, Runzhe Zhan, Derek F. Wong, Shu Yang, Xinyi Yang, Yulin Yuan, and Lidia S. Chao. 2024. [DetectRL: Benchmarking LLM-generated text detection in real-world scenarios](#). *Preprint*, arXiv:2410.23746.
- Ying Zhou, Ben He, and Le Sun. 2024. Humanizing machine-generated content: Evading AI-text detection through adversarial attack. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8427–8437, Torino, Italia.

## A AI Detection as Civic Infrastructure

Academic integrity systems are a form of institutional civic infrastructure: they adjudicate disputes

over authorship, govern access to academic credentials, and shape the legitimacy of educational institutions in the eyes of their communities (Eubanks, 2018; Noble, 2018). When these systems systematically produce biased outcomes for a specific population, issuing wrongful academic misconduct findings at rates exceeding random chance, they do not merely produce technical errors; they undermine the legitimacy of institutional deliberation itself (Eubanks, 2018). Meridian’s design philosophy is that fair AI detection is a prerequisite for AI systems that can genuinely support, rather than corrupt, legitimate institutional processes. An AI text detector that wrongly accuses ESL students at 61% rates is not a tool for civic integrity; it is a source of systematic institutional injustice (Noble, 2018). This framing connects biased AI detection to the broader project of socially responsible NLP: technical systems deployed in institutional settings inherit institutional responsibilities, and correcting their disparate impacts is a direct form of positive social impact.

**Engaging the “no-detection” position.** Bassett et al. (2026) argue that AI detection should not be used in education at all, because detectors produce unverifiable probabilistic outputs violating procedural fairness norms. This critique is correct on its own terms, and Meridian does not refute it. Meridian occupies a harm-reduction position: institutions are deploying these tools at scale regardless of academic argument, and a 15.7× reduction in wrongful ESL flags is a concrete welfare improvement within that reality. The two positions are compatible: if institutions must use detection tools, use a fair one; and separately, the community should debate whether to use them at all.

## B Deployment Guidelines for Educators

Meridian is designed as a screening and triage tool, not a verdict system. The following guidelines are intended for program directors, academic integrity officers, and ESL instructors considering deployment:

**(a) Mandatory human review for all flagged essays.** No automated detection output should initiate disciplinary proceedings without review by a qualified educator familiar with the student’s language background. Meridian’s output is a probability score, not a finding of academic dishonesty.

**(b) Operating threshold selection.** At the default threshold (0.50), overall TOEFL11 FPR is 3.9%; raising the threshold to 0.80 reduces FPR to 0.9% (low: 0.0%, medium: 0.9%, high: 1.8%) while maintaining 94.7% TPR. Institutions with low tolerance for false accusations should select a conservative threshold.

**(c) High-proficiency ESL writers require additional scrutiny.** Phase 3 results show FPR increases from 0.0% (low proficiency) to 8.2% (high proficiency) on TOEFL11. Students approaching native-like fluency are most at risk of false flagging and should receive elevated human review.

**(d) Institutional transparency.** Students should be informed when AI detection tools are used to evaluate their submissions (Bender et al., 2021).

## C Generation Prompts

All AI essays were generated using the prompt templates below.

**System prompt (GPT-4.1 only).** GPT-4.1 was called with a system message; Gemini 2.5 Flash and Claude 3.5 Sonnet received the full instruction as a user message.

“You are a helpful writing tutor who writes persuasive high school essays following the user’s instructions.”

### AI-Native prompt prefix.

“Write a 400-word persuasive essay responding to the following prompt as a fluent native English-speaking high school student in an honors or AP class. Use richer vocabulary, more complex sentences (with subordinate clauses), and varied transitions, while staying clear and well-organized, and generate only the essay itself: [PROMPT]”

### AI-ESL Mimic prompt prefix.

“Write a 400-word persuasive essay responding to the following prompt as a mid-level ESL high school student (roughly B1). Use everyday vocabulary, some repeated phrases, and mostly simple sentences. Include a few grammar and punctuation mistakes (for example with articles, prepositions, or verb tense), but keep the overall meaning clear, and generate only the essay itself: [PROMPT]”

All three models used temperature 0.8. Each model generated 100 AI-Native and 100 AI-ESL Mimic essays across 5 prompts (20 per prompt per style).

## D Logistic Regression Coefficients

The Phase 2 logistic regression classifier has learned coefficients  $w_{\text{esl}} = -1.8166$ ,  $w_{\text{native}} = -0.2831$ , and intercept  $b = 12.0569$ . The decision boundary is:

$$-1.8166 \cdot P_{\text{esl}} - 0.2831 \cdot P_{\text{native}} + 12.0569 = 0. \quad (4)$$

The negative  $w_{\text{native}}$  confirms that higher sub-word perplexity (more surprising to the Transformer) shifts classification toward “human,” consistent with the interpretable account in Section 5.1.

## E Qualitative Sentence-Level Analysis

We split each Phase 1 essay into sentences, computed DistilGPT-2 perplexity per sentence, and tagged sentences for five documented L1-transfer error types: article omission, subject-verb agreement, tense errors, preposition errors, and unnecessary article insertion (Corder, 1967; Ellis, 1994).

Across all sentences, Human ESL achieves mean  $P_{\text{native}} = 160.8$  versus 97.1 for AI-ESL Mimics ( $\Delta = +63.7$ ). This gap persists and slightly grows when restricted to sentences that contain tagged grammatical errors: Human ESL 146.4 vs. AI-ESL 82.5 ( $\Delta = +63.9$ ).

Table 3 shows  $P_{\text{native}}$  by error type. Four of five categories are significant ( $p < 0.05$ ); only tense errors are not, likely due to small sample size ( $n = 28$  vs. 13). The matched examples reveal the mechanism directly: a human ESL article-omission sentence such as “*There in a river finded the frog behind the in trunk*” ( $P_{\text{native}} = 1,728$ ) produces a wildly surprising token sequence because *finded* and *the in trunk* arise from specific L1 transfer constructions no Transformer distribution anticipates, while a matched AI-ESL article-omission sentence such as “*If I am too busy, I don’t have time to relax or spend time with my family*” ( $P_{\text{native}} = 10.7$ ) remains anchored to high-probability token sequences despite the surface-error tag firing. The same asymmetry holds across all error types: human SV-agreement errors (“*He leave his frog to them and they wave goodbye*”,  $P_{\text{native}} = 1,331$ ) versus AI SV-agreement errors (“*It show us what we did wrong*”,  $P_{\text{native}} = 23.8$ ).

Table 3: Sentence-level  $P_{\text{native}}$  by grammatical error type. Human ESL produces substantially higher  $P_{\text{native}}$  on every error type, even when AI-ESL Mimics produce the same surface error pattern. Tense error is non-significant due to small  $n$ .

| Error type                 | Human<br>$P_{\text{native}}$ | AI-ESL<br>$P_{\text{native}}$ | $\Delta$     | $p$               |
|----------------------------|------------------------------|-------------------------------|--------------|-------------------|
| Article omission           | 145.4                        | 82.3                          | +63.1        | <0.0001           |
| S-V agreement              | 198.6                        | 121.3                         | +77.2        | 0.016             |
| Tense error                | 123.4                        | 70.9                          | +52.5        | 0.194             |
| Preposition error          | 133.9                        | 64.6                          | +69.3        | <0.0001           |
| Unnecessary article        | 123.0                        | 68.7                          | +54.3        | <0.0001           |
| <b>All error sentences</b> | <b>146.4</b>                 | <b>82.5</b>                   | <b>+63.9</b> | <b>&lt;0.0001</b> |

## F Confidence Intervals

Table 4: 95% Clopper–Pearson confidence intervals for Phase 3 FPR values.

| Group      | $n$        | FP        | FPR         | 95% CI              |
|------------|------------|-----------|-------------|---------------------|
| ARA        | 30         | 0         | 0.0%        | [0.0%, 11.6%]       |
| DEU        | 30         | 1         | 3.3%        | [0.1%, 17.2%]       |
| FRA        | 30         | 1         | 3.3%        | [0.1%, 17.2%]       |
| HIN        | 30         | 1         | 3.3%        | [0.1%, 17.2%]       |
| ITA        | 30         | 2         | 6.7%        | [0.8%, 22.1%]       |
| JPN        | 30         | 3         | 10.0%       | [2.1%, 26.5%]       |
| KOR        | 30         | 1         | 3.3%        | [0.1%, 17.2%]       |
| SPA        | 30         | 0         | 0.0%        | [0.0%, 11.6%]       |
| TEL        | 30         | 2         | 6.7%        | [0.8%, 22.1%]       |
| TUR        | 30         | 0         | 0.0%        | [0.0%, 11.6%]       |
| ZHO        | 30         | 2         | 6.7%        | [0.8%, 22.1%]       |
| <b>All</b> | <b>330</b> | <b>13</b> | <b>3.9%</b> | <b>[2.1%, 6.6%]</b> |
| Low        | 110        | 0         | 0.0%        | [0.0%, 3.3%]        |
| Medium     | 110        | 4         | 3.6%        | [1.0%, 9.0%]        |
| High       | 110        | 9         | 8.2%        | [3.8%, 15.0%]       |