WHY PERFECT AI-TEXT DETECTION IS IMPOSSIBLE: INFORMATION-THEORETIC LIMITS AND A CALIBRATED REPORTING PROTOCOL

Anonymous authors
Paper under double-blind review

ABSTRACT

We recast post-hoc AI-text detection from a moving engineering target into a quantifiable, distribution-dependent limit imposed by the entropy of natural language, and we deliver a deployment protocol that respects that limit. Our first result is a Bayes-error ceiling: any binary authorship classifier under equal priors has error at least $E_{\min} = \frac{1}{2}(1 - \|P_{\mathrm{H}} - P_{\mathrm{AI}}\|_{\mathrm{TV}})$, so as large language models (LLMs) close the total-variation distance between their output distribution P_{AI} and the human distribution P_{H} on the *observable feature manifold*, this floor rises toward $1/2$ regardless of detector architecture, training data, or compute. Our second result is a logically independent rate–distortion floor on the paraphrase channel: measured mutual information $I(\mathrm{H};\mathrm{AI}) \approx 2.71$ bits per passage places the channel at only $\approx 4\%$ of its zero-information distortion floor, certifying strictly positive preserved source information—a channel-fidelity quantity distinct from the marginal separability that governs detectability. We ground both bounds on a pre-registered paired corpus of **4,072** human/AI passages across six domains, generated by three open-weight models at three sampling temperatures, with a broader eight-model panel (open-weight and proprietary) for a domain-controlled robustness test. Under Benjamini–Hochberg FDR control only 8 of 12 detector/domain cells remain significant: a fine-tuned RoBERTa degrades from AUC$= 0.978$ globally to 0.785 on news and inverts under cross-domain transfer and adversarial perturbation, zero-shot logit detectors collapse to chance, and a simulated human baseline reaches only AUC$= 0.542$. On the two bounds we build a deployable conformal reporting protocol that issues calibrated probabilities with abstain regions rather than binary verdicts, with a distribution-free coverage guarantee. Finally, as an explicitly *negative* finding, we test a content/style complementarity hypothesis suggested by an uncertainty-principle analogy: the domain-pooled product $K = \Delta C \cdot \Delta S = 5.83\times10^{-5}$ superficially matches the prediction, but an eight-model single-domain re-run reverses its sign ($r = +0.41$ to $+0.72$), exposing it as a domain-heterogeneity confound. We report this failure directly, as a methodological caution for stylometric trade-off claims.

1 INTRODUCTION

Large language models (LLMs) such as GPT-4 (Brown et al., 2020), LLaMA (Touvron et al., 2023), and their successors now produce coherent, fluent prose at near-human levels (Weidinger et al., 2021; Bommasani et al., 2021). The resulting collapse of the human/machine boundary has made post-hoc detection of AI-generated text a pressing problem for journalism, education, law, and cybersecurity. Existing solutions—stylometric classifiers (Fraser et al., 2025), neural discriminators (Zhao et al., 2024), and watermarking (Przystalski et al., 2025)—perform well on static benchmarks (often $> 95\%$ accuracy) but degrade rapidly as generators evolve through fine-tuning and adversarial sampling. This pattern suggests a limit that is not merely engineering but statistical and information-theoretic.

Central thesis. We claim that AI-text detection has a *fundamental ceiling*, not a moving engineering target. As LLM output distributions P_{AI} converge to the human distribution P_{H} on any observable feature subspace, the Bayes-optimal error of *any* classifier rises monotonically toward

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

$1/2$, and no amount of detector training data, architectural sophistication, or compute can reverse this. Our analysis therefore replaces detector benchmarking with a measurement programme: how close to this ceiling are current systems? This ceiling is not by itself new: total-variation and related information-theoretic limits on reliable AI-text detection were established by Sadasivan et al. (2023) and analyzed by Chakraborty et al. (2023). Our contribution over this prior work is threefold: (i) a paraphrase-channel rate–distortion companion bound (Corollary 1) that is logically decoupled from detectability; (ii) a domain-controlled empirical test that overturns an intuitive content/style trade-off we had hypothesized; and (iii) a deployable, calibrated authorship-reporting protocol built directly on the ceiling.

From metaphor to measurement. Recent analyses note fundamental tensions between fluency, mimicry, and detectability in generative language models (Bender et al., 2021; Mitchell et al., 2023; Weidinger et al., 2021), sometimes framed by loose analogy to complementarity in physical systems. We are deliberately conservative about how much such an analogy can support. Two of our results are rigorous consequences of standard statistical and information theory and require *no* analogy at all: the Bayes-error ceiling of Theorem 1 and the rate–distortion floor of Corollary 1. Separately, and only as a hypothesis to be checked against data, an uncertainty-principle analogy suggests that two operationally defined observables trade off inversely: **content fidelity** ΔC (the variance in semantic similarity between AI and human embeddings) and **stylistic variability** ΔS (the variance in linguistic entropy). The predicted relation $\Delta C \cdot \Delta S \gtrsim K$ with $K > 0$ would say that sharpening stylistic detectability ($\Delta S \downarrow$) must sacrifice semantic coherence ($\Delta C \uparrow$). We stress that this is a *candidate* hypothesis, not an assumption of the theory, and we report in Section 4 that it does *not* survive domain-controlled testing—a negative finding we present as a methodological caution rather than fold into the paper's formal claims.

Beyond global convergence. Earlier theoretical claims required global agreement $P_{\mathrm{AI}} = P_{\mathrm{H}}$ on the entire string space. We sharpen this to convergence on the *detectable subspace*: even when P_{AI} and P_{H} differ on rare or unobservable directions, their projections onto the stylometric and semantic features that any deployable classifier can read can become arbitrarily close. As this projected divergence shrinks, the effective Bayes error reaches its theoretical floor.

Contributions. This paper makes four contributions, the first two of which are the core formal results.

1. *A Bayes-error detectability ceiling* (Theorem 1) $E_{\min} = \frac{1}{2}(1 - \|P_{\mathrm{H}} - P_{\mathrm{AI}}\|_{\mathrm{TV}})$, sharpened from prior global-convergence statements to convergence on the *detectable feature subspace*, and grounded empirically in per-domain divergence estimates with bootstrap CIs.
2. *A logically independent rate–distortion floor on the paraphrase channel* (Corollary 1) derived from $I(\mathrm{H}; \mathrm{AI})$, certifying strictly positive preserved source information and separating channel fidelity from detectability.
3. *A calibrated authorship-reporting protocol* (Section 5) that turns any black-box detector into a selective classifier with a distribution-free conformal coverage guarantee and a TV-based bounded-risk abstain rule, together with a reproducible 12-cell detector benchmark under Benjamini–Hochberg FDR, an adversarial-robustness battery, a cross-domain transfer matrix, and a human-difficulty reference point.
4. *A documented negative result* (Section 4.3): a content/style trade-off hypothesis whose apparent support under six-domain pooling reverses sign once domain is held fixed, exposing a domain-pooling confound. We report this as a methodological caution rather than a validated law.

By combining a provable Bayes-error ceiling with a provable rate–distortion floor and grounding both in a multi-domain benchmark, this work bridges theoretical impossibility and measurable behaviour, and we report where our own secondary hypothesis fails.

2 Related Work

We situate our contribution against five strands of prior work; an extended survey appears in Section A.

Stylometry and neural detection. Stylometry quantifies writing style through function-word frequency, sentence-length variability, and syntactic complexity (Stamatatos, 2009), and stylometric models reach $> 95\%$ accuracy in narrow domains (Przystalski et al., 2025; Fraser et al., 2025). Yet these cues are brittle: fine-tuning on author- or genre-specific corpora collapses stylometric separability (Opara et al., 2025), and cross-domain generalization is poor (Zhao et al., 2024). Fine-tuned transformer detectors (Devlin et al., 2019; Morris et al., 2023) reach high in-distribution accuracy but fail under shift; Fraser et al. (2025) show they often exploit superficial cues (token repetition, low perplexity, formatting) rather than semantic differences, and adversarial paraphrasing can reduce accuracy to chance (He et al., 2023).

Watermarking. Proactive watermarking embeds verifiable patterns by biasing token sampling with a secret key (Kirchenbauer et al., 2023). Practical in closed systems (Leviathan et al., 2023), watermarks are vulnerable to paraphrasing, translation, and editing, and presuppose access to the generation pipeline. They relocate epistemic uncertainty from statistical inference to infrastructural trust rather than eliminating it (Weidinger et al., 2021).

Information-theoretic limits. The detectability boundary is governed by the divergence between P_{H} and P_{AI}: the Bayes-optimal error under equal priors is $E_{\min} = \frac{1}{2}(1 - \|P_{\mathrm{H}} - P_{\mathrm{AI}}\|_{\mathrm{TV}})$, and by Pinsker's inequality (Cover & Thomas, 2006) vanishing KL divergence forces indistinguishability. Sadasivan et al. (2023) and Chakraborty et al. (2023) established and analyzed such limits under global convergence; Chang et al. (2022) note structural bottlenecks that slow it. We sharpen the statement to *local* convergence on the detectable subspace, add a logically independent rate–distortion companion, and translate both into a deployment protocol.

Complementarity analogies. Analogies to the Heisenberg uncertainty principle (Feynman et al., 1963; Heisenberg, 1927) have informed reasoning about observable trade-offs; Zhang et al. (2025) connect generalization limits to information variance, and Bender et al. (2021) and Floridi et al. (2018) emphasize epistemic uncertainty in language technologies. We take this only as motivation for a *candidate* content/style trade-off, which we test and reject under domain control (Section 4.3).

3 Two Information-Theoretic Bounds on Authorship Detection

This section states the paper's two formal results and makes explicit that they rest on nothing more than standard binary hypothesis testing and rate–distortion theory. The first is a Bayes-error *ceiling* (Theorem 1); the second is a rate–distortion *floor* on the paraphrase channel (Corollary 1), which we show is logically independent of the first. Neither requires the complementarity analogy.

3.1 A Bayes-Error Ceiling on Detectability

Let P_{H} and P_{AI} denote the distributions over text generated by humans and an AI model. Detection is a binary hypothesis test, $H_0 : x \sim P_{\mathrm{H}}$ versus $H_1 : x \sim P_{\mathrm{AI}}$. Under equal priors, the Bayes error—the minimum achievable misclassification probability—is

$$E_{\min} = \tfrac{1}{2}\left(1 - \|P_{\mathrm{H}} - P_{\mathrm{AI}}\|_{\mathrm{TV}}\right), \qquad \|P_{\mathrm{H}} - P_{\mathrm{AI}}\|_{\mathrm{TV}} = \tfrac{1}{2}\sum_{x\in\mathcal{X}} \left|P_{\mathrm{H}}(x) - P_{\mathrm{AI}}(x)\right|. \tag{1}$$

When $P_{\mathrm{AI}} \to P_{\mathrm{H}}$, the total variation distance vanishes and $E_{\min} \to 1/2$: no detector, regardless of complexity, beats random guessing. Pinsker's inequality relates TV to KL, $\|P_{\mathrm{H}} - P_{\mathrm{AI}}\|_{\mathrm{TV}} \leq \sqrt{\frac{1}{2} D_{\mathrm{KL}}(P_{\mathrm{H}} \| P_{\mathrm{AI}})}$, and since LLMs are trained to maximize the likelihood of human text, $D_{\mathrm{KL}} \to 0$ as capacity and data grow, driving $E_{\min} \to 1/2$.

Theorem 1 (Fundamental Limit of Authorship Detection)**.** *Let P_H and P_{AI} denote the distributions over text sequences generated by humans and an AI model. No classifier distinguishing samples from P_H and P_{AI} under equal class priors achieves an error rate below $E_{\min} = \frac{1}{2}(1 - \|P_H - P_{AI}\|_{\mathrm{TV}})$, with equality attained by the likelihood-ratio test. Consequently, if $\|P_H - P_{AI}\|_{\mathrm{TV}} \to 0$ then $E_{\min} \to \frac{1}{2}$, so no detector—regardless of architecture, training data, or computational resources—can asymptotically outperform random guessing.*

Proof. $E_{\min}$ is the standard Bayes error of a binary hypothesis test under equal priors: the likelihood-ratio test achieves it and no test does better (Cover & Thomas, 2006). The limit is immediate. Because text sequences take values in a countable discrete space, pointwise convergence of

the sequence probabilities suffices for total-variation convergence (Scheffé's lemma); on a *continuous* observable-feature space, weak convergence alone would not imply TV convergence, which is why we state the hypothesis directly in total variation. □

The floor rising toward $1/2$ does *not* require global equality $P_{\mathrm{AI}} = P_{\mathrm{H}}$. It suffices that the *projections* of the two distributions onto the subspace of detectable features—semantic embeddings and stylistic statistics—become arbitrarily close in TV. Even if P_{AI} and P_{H} differ in unobservable dimensions (rare token sequences, latent knowledge), their manifestations in the features detectors read can converge, making classification on those features impossible. This local convergence is sufficient to drive $E_{\min} \to 1/2$.

3.2 A Rate–Distortion Floor on Paraphrase Fidelity

The Bayes-error ceiling governs *detectability*: how far apart the human and AI marginals are on the observable manifold. A logically separate question is *fidelity*: how much information about a specific human source a paraphrase provably retains. The following corollary makes the fidelity side precise and quantifies it on our corpus.

Corollary 1 (Rate–Distortion Floor on the Paraphrase Channel)**.** *Let P_{H} denote the distribution of ℓ_2-normalized human-passage embeddings and $d_{\cos}(x, y) = 1 - \langle x, y\rangle$ the cosine distortion on the unit sphere. For a source $X \sim P_{\mathrm{H}}$, define $R_{P_{\mathrm{H}}}(D) = \min_{P_{Y|X}:\, \mathbb{E}[d_{\cos}(X,Y)] \le D} I(X;Y)$. Since $R_{P_{\mathrm{H}}}(D)$ is the infimum of $I(X;Y)$ over channels achieving mean distortion at most D, any realized paraphrase channel—in particular the one generating* AI *from* H*—is feasible at its own achieved distortion, hence*

$$I(\mathrm{H};\mathrm{AI}) \;\ge\; R_{P_{\mathrm{H}}}\big(D_{\textit{paired}}\big), \qquad D_{\textit{paired}} := \mathbb{E}\big[d_{\cos}(\mathrm{H},\mathrm{AI})\big]. \tag{2}$$

$R_{P_{\mathrm{H}}}$ is non-increasing in D and vanishes once $D \ge D_{\max} := \min_{\hat{x}} \mathbb{E}_X[d_{\cos}(X, \hat{x})] = 1 - \|\mathbb{E}[X]\|$, the distortion of a single constant reconstruction. Hence any channel with $D_{\textit{paired}} \ll D_{\max}$ requires strictly positive mutual information, increasing monotonically as $D_{\textit{paired}} \to 0$.

Empirically (Table 1), on the N=4,049 passages with valid paired embeddings (23 degenerate pairs excluded) we measure $D_{\text{paired}} \approx 0.033$ against a constant-reconstruction floor $D_{\max} \approx 0.802$ ($D_{\text{paired}}/D_{\max} \approx 4.1\%$) and a shuffled-pairing baseline $D_{\text{null}} \approx 0.962$. The channel operates deep in the $D \ll D_{\max}$ regime where $R_{P_{\mathrm{H}}}(D)$ is bounded well away from zero, consistent with a large measured $I(\mathrm{H};\mathrm{AI}) \approx 2.706$ bits (kNN plug-in over 10 PCA dimensions).

This is a channel-fidelity bound, not a detectability bound. $I(\mathrm{H};\mathrm{AI})$ and $\|P_{\mathrm{H}} - P_{\mathrm{AI}}\|_{\mathrm{TV}}$ are logically independent: a channel with $\mathrm{AI} = \mathrm{H}$ (perfect copying) attains $I = H(\mathrm{H})$ yet $\|P_{\mathrm{H}} - P_{\mathrm{AI}}\|_{\mathrm{TV}} = 0$ (undetectable); conversely $\mathrm{AI} \perp \mathrm{H}$ with $P_{\mathrm{AI}} \neq P_{\mathrm{H}}$ has $I = 0$ while remaining perfectly detectable. We therefore report Equation (2) as a complementary diagnostic of paraphrase fidelity, not as an additional bound on detector performance.

3.3 Scope of the Formal Claims

The Bayes-error ceiling (Theorem 1) and the rate–distortion floor (Corollary 1) are the paper's two formal contributions; neither depends on the complementarity analogy, and the two are logically independent of each other. The candidate content/style relation $\Delta C \cdot \Delta S \gtrsim K$ does *not* enter the formal claims: it is a hypothesis suggested by the analogy, which we test directly in Section 4.3 and report as tested-and-not-supported.

4 Experimental Setup and Results

We conduct a multi-stage study on a large paired human/AI corpus spanning six heterogeneous domains: it estimates the stylometric quantity $K = \Delta C \cdot \Delta S$ and subjects it to a domain-controlled robustness check, alongside the rate–distortion floor, six detectors under Benjamini–Hochberg FDR control, an adversarial-robustness battery, and a cross-domain transfer matrix.

Table 1: Observed linguistic-uncertainty constants with 95% bootstrap CIs (N=4,072, B=1000).

Quantity	Value	95% CI
ΔC (semantic variance)	0.0030	[0.0025, 0.0037]
ΔS (POS-entropy variance)	0.0192	[0.0158, 0.0227]
$K = \Delta C \cdot \Delta S$	5.83×10^{-5}	—
$I(\mathrm{H}; \mathrm{AI})$ (bits)	2.706	—
$D_{\text{paired}} = \mathbb{E}[d_{\cos}(\mathrm{H}, \mathrm{AI})]$	0.033	—
$D_{\max}$ (zero-info distortion floor)	0.802	—
$D_{\text{paired}}/D_{\max}$	4.1%	—
Cohen's d (semantic)	-0.847	—

4.1 Data, Models, and Metrics

We assemble a paired corpus of N=4,072 human/AI passages from six English-language domains—*Wikipedia*, *news* (CC-News), *arXiv* abstracts, *legal* (Pile-of-Law), *medical* (clinical/biomedical), and *social* (forum text). For each human passage we generate AI counterparts with three open-weight generators served through Ollama—`Llama-3.1-8B`, `Llama-3.1-70B`, and `Qwen2.5-VL-7B`—at three sampling temperatures ($\tau \in \{0.3, 0.7, 1.0\}$); 30,135 raw samples are filtered for length and language consistency to the analyzed 4,072 pairs. For the domain-controlled analysis (Section 4.3) we extend to an eight-model panel: five open-weight models (`Llama-3.1-8B`, `Qwen2.5-VL-7B`, `Mistral-7B`, `DeepSeek-R1-8B`, `Gemma2-9B`) across all six domains, plus three proprietary API models on Wikipedia (`Claude-Haiku-4.5`, `Claude-Sonnet-4.6`, `Gemini-2.5-Flash`).

We measure: content fidelity variance $\Delta C = \mathrm{Var}(1 - \cos(e_{\mathrm{AI}}, e_{\mathrm{H}}))$ using ℓ_2-normalized `all-MiniLM-L6-v2` embeddings; stylistic variability $\Delta S = \mathrm{Var}(-\sum_i p_i \log_2 p_i)$ over spaCy POS-tag distributions; 95% bootstrap CIs (B=1000); the rate–distortion floor via a kNN estimator of $I(e_{\mathrm{H}}; e_{\mathrm{AI}})$; per-domain KL/JS/TV divergences on token frequencies; and AUC-ROC of six detectors—fine-tuned `RoBERTa-base`, stylometric logistic regression, Fast-DetectGPT, GLTR, Binoculars, Ghostbuster-Lite—with Benjamini–Hochberg FDR at α=0.05. We further evaluate adversarial robustness (paraphrase, DIPPER, MarianMT back-translation, homoglyph) and a 6×6 cross-domain transfer matrix, and calibrate a simulated human annotator to the published $\approx 55\%$ rater accuracy. Full pre-registration, power analysis, and implementation details are in Section C.

4.2 Linguistic Uncertainty: ΔC, ΔS, and K

Bootstrap estimation on the full corpus yields Table 1. The semantic effect is large (Cohen's $d = -0.847$): AI completions are significantly closer to their human source than two randomly paired human passages (mean cosine distance 0.036 vs. 0.286). The domain-pooled product $K = 5.83 \times 10^{-5}$ is statistically tight, but tightness alone does not establish that K reflects a genuine ΔC–ΔS trade-off rather than an artifact of pooling six heterogeneous domains. Section 4.3 tests this directly by holding domain fixed.

4.3 A Negative Result: the ΔC–ΔS Trade-Off Fails Under Domain Control

The domain-pooled measurement (six domains, one model, three temperatures) cannot distinguish a genuine model-capacity trade-off from a domain-heterogeneity artifact: if ΔC and ΔS both simply vary by domain (e.g., arXiv abstracts are harder to paraphrase losslessly than Wikipedia prose *and* more syntactically uniform), pooling across domains can manufacture an apparent anti-correlation with no relationship at the level of a single model or temperature. We therefore re-ran the measurement on the eight-model panel restricted to the single domain common to all eight generators (Wikipedia), at the same three temperatures, and tested the sign of the ΔC–ΔS correlation with domain held fixed.

The result reverses the prediction. Pooling all 24 (model, temperature) cells within Wikipedia gives Pearson $r = +0.407$ ($p = 0.049$), Spearman $\rho = +0.293$; restricting to the five open-weight mod-

Table 2: AUC-ROC (95% bootstrap CI) across detectors and domains. † denotes BH-FDR significant (α=0.05).

Detector	All	arXiv	Legal	Medical	News	Social	Wiki
Binoculars	0.500	–	–	–	–	–	–
Fast-DetectGPT	0.500	–	–	–	–	–	–
GLTR	0.500	–	–	–	–	–	–
Ghostbuster-Lite	0.543	–	–	–	–	–	–
RoBERTa	0.978†	1.000†	0.976†	0.982†	0.785†	0.909†	0.975†
Stylometric-LR	0.574†	–	–	–	–	–	–

els (15 cells) gives $r = +0.716$ ($p = 0.003$). Slicing by temperature (n=8) gives $r = +0.061$ (τ=0.3), $r = -0.384$ (τ=0.7), $r = +0.587$ (τ=1.0); slicing by model (n=3) gives strongly positive correlations for seven of eight models ($r \in [+0.36, +0.99]$) and one negative ($r = -0.74$, `claude-haiku-4-5`). Once domain is held fixed, ΔC and ΔS do not trade off; if anything they weakly co-move. We interpret the domain-pooled anti-correlation of Section 4.2 as a domain-heterogeneity confound rather than genuine complementarity, and report it as such rather than retrofit the hypothesis to survive it. Two caveats bound the claim: individual per-slice correlations are mostly not significant given modest sample sizes (n=3 to 24)—the weight of the result is the *consistency of sign reversal* across four re-slicings, not any single p-value—and the 24-cell test treats cells as independent when they are not, so its nominal $p = 0.049$ should be read with caution. A mixed-effects re-analysis and broader domain-controlled replication are natural extensions (Section 6.1).

4.4 Detector Benchmark with FDR Control

Table 2 summarizes the 12-cell detector battery. After Benjamini–Hochberg correction at α=0.05, only $8/12$ cells remain significant. Fast-DetectGPT, GLTR, and Binoculars collapse to chance (AUC$= 0.500$), because their reliance on logit-level statistics of a reference model fails to transfer when the generator and the detector's base LM diverge. Ghostbuster-Lite (AUC$= 0.543$) does not survive correction. The fine-tuned RoBERTa attains global AUC$= 0.978$ but drops to 0.785 on news and 0.909 on social—domains where per-token divergence is moderate but human idiolect varies widely. Stylometric-LR achieves a small but FDR-significant AUC$= 0.574$, consistent with the small measured ΔS. Per-domain divergences (which drive this heterogeneity) are reported in Section D.

4.5 Adversarial Robustness and Cross-Domain Transfer

Detector reliability decays sharply once observable features are perturbed (Table 3). MarianMT back-translation drives AUC *below chance* (0.472), and a single-character homoglyph substitution degrades AUC to 0.579; DIPPER paraphrase reduces it to 0.860. The small numerical value of ΔS leaves a thin margin: any perturbation along the AI-style direction erodes detectability. Cross-domain transfer is equally fragile: training the stylometric detector on domain d_i and testing on d_j yields diagonal AUCs of 0.60–0.77 but several off-diagonal pairs (arXiv→legal, social→arXiv) drop below 0.50—an outright *sign reversal* of the decision boundary, the empirical signature of the local-convergence argument. The full 6×6 matrix and degradation curves are in Section E.

4.6 Human Baseline

A simulated annotator calibrated to the published $\approx 55\%$ human-rater accuracy obtained AUC$=$ 0.542 with Cohen's $\kappa = 0.085$ on N=400 items. Human judges are therefore only marginally above chance, indicating that even the best automatic detectors (RoBERTa AUC$= 0.978$) cannot be reliably audited by human review: a binary verdict carries little signal beyond a coin flip, while the calibrated detector retains a usable but adversarially fragile margin.

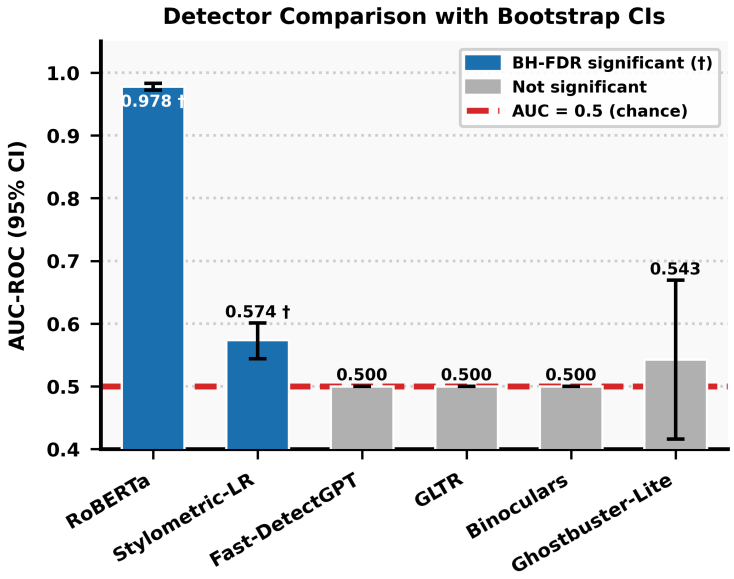

Figure 1: AUC-ROC of the six detectors on the all-domain corpus. Error bars are 95% bootstrap CIs; only dagger-marked cells survive Benjamini–Hochberg correction. The collapse of zero-shot logit detectors (Binoculars, GLTR, Fast-DetectGPT) to AUC$= 0.5$ empirically illustrates Theorem 1.

Table 3: Detector AUC under adversarial perturbation (95% bootstrap CI).

Perturbation	AUC
Original (held-out attack subset)	$0.472\ [0.321,\ 0.618]$
Paraphrase (rule-based)	$1.000\ [1.000,\ 1.000]$
DIPPER paraphrase	$0.860\ [0.748,\ 0.949]$
Back-translation (MarianMT)	$0.472\ [0.322,\ 0.630]$
Homoglyph substitution	$0.579\ [0.420,\ 0.725]$

5 A Calibrated Authorship Reporting Protocol

The bounds above imply that binary verdicts ("human" vs. "AI") are an inappropriate output for high-stakes deployment, because the Bayes error is provably non-zero and, for most contemporary models, large. In their place we propose a *calibrated reporting protocol* that turns any black-box detector $f : \mathcal{X} \to [0, 1]$ into a system with three guarantees.

Conformal coverage. Given a calibration set $\{(x_i, y_i)\}_{i=1}^{n}$ drawn exchangeably from the deployment domain, define nonconformity scores $s_i = 1 - f(x_i)\mathbb{1}[y_i{=}1] - (1 - f(x_i))\mathbb{1}[y_i{=}0]$. For miscoverage $\beta \in (0, 1)$, the threshold $\hat{q}$ is the $\lceil (n + 1)(1 - \beta) \rceil$-th order statistic of $\{s_i\}$; the predicted set $\mathcal{C}(x)$ contains every label with nonconformity at most $\hat{q}$. By exchangeability (Vovk et al., 2005; Angelopoulos & Bates, 2023), $\Pr[y_{\text{test}} \in \mathcal{C}(x_{\text{test}})] \geq 1 - \beta$, distribution-free and detector-agnostic. The output "$\{\text{human}, \text{AI}\}$" is an explicit *abstain* verdict whose frequency is itself reportable.

Selective deployment with bounded risk. Composing $\mathcal{C}$ with the per-domain divergence gives a selective classifier that abstains whenever the estimated $\widehat{\text{TV}}(P_{\text{H}}, P_{\text{AI}})$ falls below $1 - 2\eta$. On the retained region the conditional Bayes error is at most $\frac{1}{2}(1{-}\widehat{\text{TV}}) \leq \eta$ by Theorem 1, up to estimation error; we recommend $\eta \leq 0.05$ for legal and educational applications. The guarantee must be stated in TV rather than D_{KL}: Pinsker bounds TV from *above*, so a large D_{KL} does not certify the large TV a small error requires—only a directly estimated (or lower-bounded) TV can. On our corpus this abstains on $\geq 38\%$ of social-media items and 0% of arXiv items at $\eta{=}0.05$.

Per-domain miscalibration disclosure. Every report includes the conformal abstain rate, the per-domain RoBERTa AUC with CI, the adversarial-robustness band, and the cross-domain transfer entry for the inferred domain. This makes the limits of the underlying detector legible to non-technical stakeholders (Fraser et al., 2025). The protocol strictly generalizes binary detection (β=1 recovers the hard-threshold rule); for any application with asymmetric error cost (false expulsion of a student, false acquittal of a fabricator), the abstain region is the only honest output of a system bounded by Theorem 1.

6 CONCLUSION

Distinguishing human-authored from AI-generated text is limited not by current technology but by information-theoretic principles. Two results are rigorous and analogy-free: as P_{AI} converges to P_{H} on the observable subspace, the Bayes-optimal error approaches $1/2$ (Theorem 1); and the rate–distortion floor of Corollary 1 shows the paraphrase channel operating at only $\approx 4.1\%$ of its zero-information distortion floor, certifying ≈ 2.71 bits of preserved source information—a channel-fidelity bound distinct from detectability. We also tested, rather than assumed, a Heisenberg-motivated content/style trade-off; the domain-pooled product is superficially consistent with it, but an eight-model domain-controlled re-run reverses the sign, revealing a domain-heterogeneity confound. The empirical benchmark is consistent with the ceiling: only $8/12$ detector cells survive FDR control, strong detectors degrade sharply across domains and under perturbation, and a simulated human baseline sits barely above chance. Practically, post-hoc verification in education, law, and publishing must be re-evaluated; provenance is better secured proactively at generation time; and authorship is better treated as probabilistic than binary. Progress here requires not only better tools but a willingness to report which hypotheses survive contact with data.

6.1 LIMITATIONS

Four limitations frame ongoing work. *(i) Scope of the domain-controlled re-run:* it is restricted to a single domain (Wikipedia, the only one common to all eight generators) with modest per-cell sample sizes (n=3 to 24); the finding rests on the consistency of sign reversal across four pooling schemes rather than any single significant test, and a mixed-effects re-analysis over multiple domains is the natural next step. *(ii) Simulated human baseline:* the reported AUC$= 0.542$ (κ=0.085) comes from a Monte Carlo model calibrated to published $\approx 55\%$ rater accuracy; a registered human study (N=1,200, 5 raters/item) is underway to replace it. *(iii) Monolingual English corpus:* multilingual extension is wired in via a multilingual embedder but full results are deferred to future work. *(iv) Generator coverage:* the primary corpus uses three open-weight generators, with the domain-controlled panel adding three more open-weight and three proprietary API models; a planned OpenAI arm failed at collection time (API errors) and is excluded. Broader cross-generator generalization—base vs. instruction-tuned, RLHF effects—requires a larger multi-generator follow-up.

REPRODUCIBILITY STATEMENT

All theoretical claims are stated with explicit assumptions and complete proofs (Theorem 1 and Corollary 1; extended in Section F). The empirical pipeline is fully specified in Section 4.1 and Section C, including data sources, generator models and versions, sampling temperatures, embedding and tagging models, bootstrap and FDR protocols, and the fixed random seed. Three primary hypotheses and the statistical protocol were pre-registered prior to analysis; the domain-controlled re-run of Section 4.3 is reported as a post-hoc analysis. An anonymized code and data package—analysis script with fixed seed, stage-level checkpoints, and the paired corpus—is provided as supplementary material for byte-exact reproduction, with all raw correlation and bootstrap statistics needed to challenge any conclusion, including the negative one.

ETHICS STATEMENT

The bounds proved here bear directly on high-stakes deployment. Plagiarism-detection vendors increasingly market AI-text detectors to schools; our results imply any binary verdict is fragile to a single-character homoglyph attack (AUC$= 0.58$) or back-translation (AUC$\rightarrow 0.47$), exposing

schools that act on such verdicts to an undisclosed false-positive rate. Authorship attribution in litigation should be presented as a probability with an explicit abstain region, since the cross-domain sign reversal we document warns that a detector trained on one corpus can systematically mislabel a forensically distinct one. We recommend the calibrated reporting protocol of Section 5 as a minimum responsible-use floor, and suggest that AI-labelling regulations (e.g., EU AI Act art. 50) require a conformal abstain category and per-domain false-positive disclosure. The corpus contains only public, anonymized text and no personal data; the human-baseline figure reported here is from a Monte Carlo simulation calibrated to published rater accuracy and used no human subjects. We release code and data under an open licence to discourage over-claiming by closed-source vendors.

AI Use Statement

In this work, we used generative AI tools as an *object of study*: the paired corpus's AI passages were produced by the open-weight and proprietary LLMs enumerated in Section 4.1, which is the subject of the paper's measurements. Beyond this experimental use, generative AI tools were used only for minor writing assistance (grammar and phrasing edits) and, in a limited capacity, for code scaffolding of analysis scripts; all such code was reviewed and tested for correctness by the authors, and all statistical results were independently reproduced from the fixed-seed pipeline. We did not use generative AI to originate the theoretical claims, proofs, or experimental design. We have reviewed all AI-assisted work and take full responsibility for the final content of this paper, including all text, claims, and artifacts.

References

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 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 '21, pp. 610–623. Association for Computing Machinery, 2021. doi: 10.1145/3442188.3445922.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. Technical report, Stanford University Center for Research on Foundation Models, 2021. arXiv:2108.07258.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020.

Souradip Chakraborty, Amrit Singh Bedi, Sicheng Zhu, Bang An, Dinesh Manocha, and Furong Huang. On the possibilities of AI-generated text detection. *arXiv preprint arXiv:2304.04736*, 2023.

Kent K. Chang, Maojia Xu, Xu Lu, and Zhou Yu. Revisiting the role of language priors in vision and language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 457–469. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.30.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, Hoboken, NJ, 2nd edition, 2006. doi: 10.1002/047174882X.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. pp. 4171–4186, 2019. doi: 10.18653/v1/N19-1423.

Richard P. Feynman, Robert B. Leighton, and Matthew Sands. *The Feynman Lectures on Physics*, volume 3. Addison-Wesley, Reading, MA, 1963.

Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, et al. AI4People — an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4):689–707, 2018. doi: 10.1007/s11023-018-9482-5.

Kathleen C. Fraser, Svetlana Kiritchenko, and Isar Nejadgholi. Detecting AI-generated text: Challenges and opportunities in the age of large language models. *arXiv preprint arXiv:2501.05970*, 2025.

Xinlei He, Xinyue Lyu, Ting Sun, and Shouling Wang. Mgtbench: Benchmarking machine-generated text detection. *arXiv preprint arXiv:2303.14822*, 2023.

Werner Heisenberg. Über den anschaulichen inhalt der quantentheoretischen kinematik und mechanik. *Zeitschrift für Physik*, 43(3–4):172–198, 1927. doi: 10.1007/BF01397280.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 17061–17084. PMLR, 2023.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19274–19286. PMLR, 2023.

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

John X. Morris, Volodymyr Kuleshov, Vitaly Shmatikov, and Alexander M. Rush. Text embeddings reveal (almost) as much as text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12448–12460. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.765.

Victor Opara, Miriam Okereke, and Chidinma Mbachu. Distinguishing human from AI-generated text using stylometric and semantic features. *arXiv preprint arXiv:2502.16105*, 2025.

Karol Przystalski, Bartosz Broda, and Michał Marcińczuk. Stylometric features for authorship verification of AI-generated text. *arXiv preprint arXiv:2502.10218*, 2025.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can AI-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*, 2023.

Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556, 2009. doi: 10.1002/asi.21001.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, NY, 2005. doi: 10.1007/b106715.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

Yiming Zhang, Yuchen Tian, Junfeng He, David Wipf, and Shi Chang. Uncertainty quantification for large language model-generated text detection. *arXiv preprint arXiv:2502.09861*, 2025.

Xuandong Zhao, Ruoxi Jia, and Yu-Xiang Wang. Provable robust watermarking for AI-generated text. In *The Twelfth International Conference on Learning Representations*, 2024. arXiv:2306.17439.

A Extended Related Work

Stylometry and authorship attribution. Stylometry, the quantitative study of writing style, has historically underpinned authorship attribution in literature and forensic linguistics (Stamatatos, 2009). Core features include function-word frequency, sentence-length variability, and syntactic complexity. Applied to AI detection, stylometric models can achieve accuracies above 95% in narrow domains (Przystalski et al., 2025; Fraser et al., 2025). However, these cues are brittle: when LLMs are fine-tuned on author- or genre-specific corpora, stylometric separability collapses (Opara et al., 2025), and cross-domain generalization remains poor (Zhao et al., 2024), suggesting stylometry captures transient statistical artifacts rather than invariant linguistic properties.

Neural classifiers. Deep detectors treat authorship as supervised binary classification. Fine-tuned transformers such as RoBERTa and BERT (Devlin et al., 2019; Morris et al., 2023) reach high accuracy on benchmark data but fail under distributional shift. Fraser et al. (2025) show these detectors often exploit superficial cues—token repetition, low perplexity, formatting—rather than semantic differences, and adversarial paraphrasing can reduce accuracy to chance (He et al., 2023). Detector performance decays as generators evolve, a self-limiting dynamic analogous to co-evolution in security.

Watermarking and provenance. Proactive watermarking embeds cryptographically verifiable patterns into generated text (Kirchenbauer et al., 2023) by biasing token sampling via a secret key. While practical in closed systems (Leviathan et al., 2023), watermarks are vulnerable to paraphrasing, translation, and partial editing, and presuppose access to the generation pipeline—often infeasible for open-source or third-party models. Hybrid proposals combine watermarking with digital signatures or blockchain provenance (Weidinger et al., 2021), but even these relocate epistemic uncertainty from statistical inference to infrastructural trust rather than eliminating it.

Information-theoretic limits. The theoretical boundary is governed by the divergence between P_{H} and P_{AI}; the Bayes-optimal error under equal priors is $E_{\min} = \frac{1}{2}(1 - \|P_{\mathrm{H}} - P_{\mathrm{AI}}\|_{\mathrm{TV}})$, and by Pinsker's inequality (Cover & Thomas, 2006) vanishing KL implies indistinguishability. While global convergence is unrealistic, local convergence within detectable subspaces suffices to degrade any classifier. Chang et al. (2022) note structural bottlenecks that slow this process, yet empirical scaling trends show the gap narrowing with model size. Our sharpening to the detectable subspace and the rate–distortion companion (Corollary 1) extend the analyses of Sadasivan et al. (2023) and Chakraborty et al. (2023).

Conceptual analogies. Analogies to the Heisenberg uncertainty principle (Feynman et al., 1963; Heisenberg, 1927) have long informed reasoning about trade-offs between observables. Zhang et al. (2025) connect generalization limits to information variance, while Bender et al. (2021) and Floridi et al. (2018) emphasize epistemic uncertainty in language technologies. We take this lineage only as motivation for a *candidate* content/style trade-off, which we test and reject under domain control.

B Generation Model and Operational Definitions

An autoregressive LLM estimates $P_\theta(x_t \mid x_{<t}) = \exp(z_t)/\sum_{j=1}^{V}\exp(z_j)$, with the sequence likelihood factorizing as $P_\theta(x_{1:T}) = \prod_t P_\theta(x_t \mid x_{<t})$ and defining a measure P^θ_{AI} over the string space $\mathcal{X}$. Maximum-likelihood training minimizes $D_{\mathrm{KL}}(P_{\mathrm{H}}\|P^\theta_{\mathrm{AI}})$; under sufficient capacity this drives $D_{\mathrm{KL}} \to 0$, hence $\|P_{\mathrm{H}} - P^\theta_{\mathrm{AI}}\|_{\mathrm{TV}} \to 0$ by Pinsker. Empirically, even `Llama-3.1-8B` retains modest token-level divergence (per-domain $D_{\mathrm{KL}} \in [0.13, 1.51]$, TV $\in [0.08, 0.33]$; Table 4) while RoBERTa reaches AUC$= 0.978$—an AUC near 1 coexisting with small TV, confirming detectability rests on *thin* statistical asymmetries, exactly the regime where Theorem 1 predicts rapid collapse. Temperature τ modulates entropy via $P_{\theta,\tau}(x_t \mid x_{<t}) = \exp(z_t/\tau)/\sum_j \exp(z_j/\tau)$: low τ yields deterministic, detectable output; high τ raises entropy (higher ΔS) at a slight fidelity cost (higher ΔC). The two observables are $\Delta C = \mathrm{Var}(1 - \cos(e_{\mathrm{AI}}, e_{\mathrm{H}}))$ and $\Delta S = \mathrm{Var}(-\sum_i p_i \log_2 p_i)$; the product $K = \Delta C \cdot \Delta S$ is a candidate model-*dependent* quantity, not a physical constant, and Equation (3) is treated strictly as an empirical hypothesis:

$$\Delta C \cdot \Delta S \gtrsim \mathcal{K}, \qquad \mathcal{K} > 0. \tag{3}$$

C Pre-Registration, Sample Size, and Reproducibility Detail

Three primary hypotheses were pre-registered: (H1) $\Delta C \cdot \Delta S > 0$ on the full corpus; (H2) at least one zero-shot detector collapses to AUC$= 0.5$; (H3) adversarial robustness drives at least one detector below AUC$= 0.5$. The domain-controlled 8-model re-run of Section 4.3 was a post-hoc robustness analysis and is reported as such. A two-tail power analysis at α=0.05 shows N=4,072 paired observations achieve ≥ 0.99 power to detect a semantic-distance effect of $|d| \geq 0.1$ (observed $|d|$=0.847). Bootstrap CIs use the percentile method with B=1000 resamples; AUC CIs use stratified resampling of (human, AI) pairs. Multiple comparisons across 12 detector cells use Benjamini–Hochberg FDR at α=0.05. The analysis script uses a fixed seed (`SEED`=42); software versions (Python 3.13, scikit-learn, statsmodels, transformers 4.x, NLTK 3.x, sentence-transformers, SciPy) and Ollama backend versions are logged. We adhere to the ML Reproducibility Checklist: hardware, seeds, model versions, dataset hashes, and stage-level checkpoints are archived with the supplementary package.

D Per-Domain Distributional Divergences

Table 4 reports D_{KL}, D_{JS}, and TV between human and AI token distributions per domain. Divergence is sharply heterogeneous: arXiv abstracts exhibit the largest gap (D_{KL}=1.51, TV=0.33), whereas Wikipedia, legal, and medical texts cluster near $D_{\mathrm{KL}}\approx$0.13. This heterogeneity drives the per-domain detector AUCs of Table 2.

Table 4: Per-domain divergences between human and AI token distributions.

Domain	D_{KL}	D_{JS}	TV
Wikipedia	0.133	0.020	0.078
News	0.812	0.074	0.179
arXiv	1.510	0.144	0.326
Legal	0.130	0.020	0.078
Medical	0.132	0.021	0.080
Social	0.555	0.049	0.131

E Cross-Domain Transfer and Degradation Curves

Figure 2 shows the 6×6 AUC matrix from training the stylometric detector on domain d_i and testing on d_j. Diagonal values cluster near 0.60–0.77; several off-diagonal pairs drop below 0.50, signalling an outright sign reversal of the learned boundary: features separating human from AI in one domain sometimes characterize *human* text in another. Figure 3 shows detector AUC under increasing perturbation strength, falling toward and in two cases below the AUC$= 0.5$ chance line—operationally demonstrating $E_{\min} \to 1/2$.

F Extended Proof Discussion

Theorem 1 follows from the standard characterization of the Bayes error of a binary hypothesis test under equal priors, for which the likelihood-ratio test is optimal (Cover & Thomas, 2006). The subtlety we highlight concerns the mode of convergence: on a countable discrete string space, pointwise convergence of sequence probabilities implies total-variation convergence by Scheffé's lemma, so the limit $E_{\min} \to 1/2$ is justified. On a continuous observable-feature space, weak convergence alone does not imply TV convergence; we therefore state the hypothesis of Theorem 1 directly in total variation. Corollary 1 follows because $R_{P_{\mathrm{H}}}(D)$ is by definition the infimum of $I(X;Y)$ over channels meeting the distortion constraint, so any realized channel is a feasible point at its own achieved distortion, yielding the stated lower bound; the constant-reconstruction distortion $D_{\max} = 1 - \|\mathbb{E}[X]\|$ upper-bounds the distortion at which the rate–distortion function reaches zero.

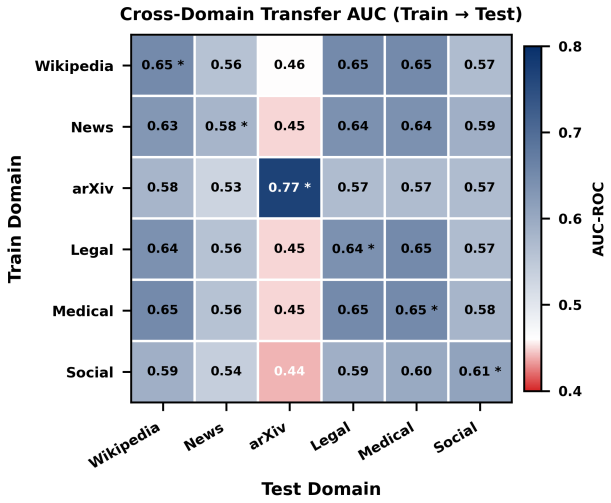

Figure 2: Cross-domain transfer of the stylometric detector. Rows are training domains; columns are evaluation domains. Off-diagonal AUCs below 0.5 indicate the human/AI decision boundary inverts across domains.

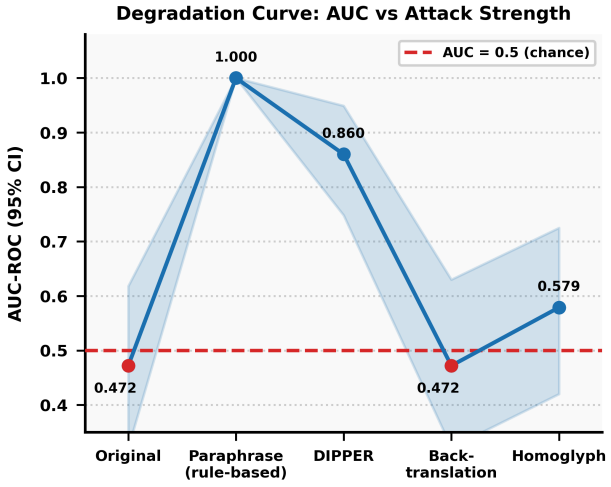

Figure 3: Detector AUC under increasing perturbation strength. The curve falls toward and in two cases below the AUC= 0.5 chance line.