NORMATIVE REWRITING ON THE READING AXIS: HOW TO EVADE CHINESE AI-TEXT DETECTORS WHILE PRESERVING QUALITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Chinese, the largest unsegmented-script language, poses a fundamental challenge for AI-text detection: word boundaries are invisible, making a detector's reading unit unknown until measured. We measure the reading axis of six real Chinese detectors and establish a predictable link to their training objective. Detection-tuned generative LMs consistently read at the word scale ($\xi = +0.115$), while BERT classifiers remain axis-balanced ($\xi \approx 0$)—challenging the prevailing assumption that Chinese detectors are character-readers. A same-architecture ablation indicates that, among fine-tuned detectors, this axis tracks the discrimination objective rather than tokenizer granularity. We further show that evasion efficacy is governed by coverage of this axis, and that saturated detectors resist deletion-based probing yet remain vulnerable to substitution, exposing deletion saliency as the wrong diagnostic. Under a normative constraint that preserves semantics, human-exemplar-guided beam rewriting lowers AUROC from 0.968 to 0.606, delivering the best quality-adjusted evasion efficiency among all baselines.

1 INTRODUCTION

Large language models now produce text that is often indistinguishable from human writing, making reliable AI-text detection a first-order concern for content governance and academic integrity (Kirchenbauer et al., 2023; Mitchell et al., 2023; Hans et al., 2024). Three families of detectors dominate the field: zero-shot statistical methods such as DetectGPT (Mitchell et al., 2023), Fast-DetectGPT (Bao et al., 2024), and Binoculars (Hans et al., 2024); watermarking schemes (Kirchenbauer et al., 2023; Chen et al., 2025; Ren et al., 2024; Dathathri et al., 2024); and supervised classifiers (Hu et al., 2023; Verma et al., 2024; Kuznetsov et al., 2024). Yet these detectors are evaluated against rewriting attacks almost exclusively in English, and the largest unsegmented-script language, Chinese, remains largely unexamined: word boundaries are invisible, so the units over which a detector forms its judgment (characters, subword tokens, or words) are unknown until measured (Macko et al., 2023; Mao et al., 2025). This gap matters on both sides of the arms race. An evasion attack presumes a perturbation unit aligned with the detector's reading unit; when the presumption is wrong, the attack fails silently and the attacker cannot tell whether the detector is robust or simply untouched (Sadasivan et al., 2023). Evaluation suffers symmetrically: a negative result may reflect a mismatched attack rather than a robust detector, and TPR-based metrics hide the attack entirely when scores saturate.

Existing evasion methods fall into three families, none of which verifies the reading unit a priori. *Unit-aligned substitution* replaces whole words or characters with synonyms, ranking candidates by gradient alignment or importance scores (Ebrahimi et al., 2018; Jin et al., 2020; Morris et al., 2020); it is cheap and detector-agnostic, but its success hinges on a guessed granularity, and a wrong guess yields near-zero evasion. *Score-optimizing* attacks (gradient-based adversarial examples, adversarial paraphrasing, reinforcement-learning humanizers) minimize detector confidence directly (Cheng et al., 2025; Fang et al., 2025; Meng et al., 2025; Ranganath & Ramesh, 2026); they are effective on the target model but produce unnatural text off the human manifold and transfer poorly across detectors (Krishna et al., 2023). *Document-level humanization* (style transfer, paraphrase chains) preserves fluency but never contacts the detector's internal evidence (Zhou et al., 2024), so evasion is capped by the detector's tolerance for surface variation. All three treat the perturbation unit as a design choice of the attack rather than an empirical property of the detector to be measured first.

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.

We take the complementary position and measure the reading axis first. We propose a measurement framework with three instruments: the differential evasion index ξ, which contrasts word- and character-aligned attacks at a common budget and isolates reading granularity; the feature-axis transfer index τ, which quantifies an attack's coverage of the detector's native saliency; and the multiscale discrimination energy $E(L)$, which probes sensitivity to contiguous spans. Because soft-probability-saturated scores are insensitive to deletion, we add a substitution-sensitivity instrument and prove an exponential margin bound for the failure of leave-one-out saliency (Appendix S1). On a 5000-text cross-generator corpus, we characterize six real Chinese detectors and find the reading axis to be a measurable, predictable property: the detection-tuned generative-LM detector reads at the word scale ($\xi = +0.115$, CI excludes zero), five BERT classifiers are axis-balanced ($\xi \approx 0$), and zero-shot perplexity detectors follow their tokenizer granularity. These results are inconsistent with the prevailing assumption that Chinese detectors are character-readers, and they make evasion predictable: across every detector and budget we measure, τ and ΔAUROC rise monotonically (aggregate $\rho_s = 0.789$, $p = 5 \times 10^{-4}$), so attacks evade exactly when they cover the detector's native axis.

Guided by these measurements, we propose a new rewriting method: human-exemplar-guided beam rewriting under normative constraints (no inserted spaces, standard punctuation, length- and semantics-guarded). It resamples candidates from a human exemplar distribution and beam-selects the candidate that minimizes the detector's margin under a semantic guard. It lowers AUROC from 0.968 to 0.606 on a held-out set (n=250; drop $+0.362$, bootstrap 95% CI $[+0.315, +0.413]$) at semantic cosine 0.903, the best quality-adjusted evasion efficiency (QAE 3.73) among all compared methods, including Adversarial Paraphrasing, CoPA, GradEscape, and StealthRL (Figure 1).

Our contributions are threefold. (i) We propose a reading-axis measurement framework (ξ, τ, $E(L)$, substitution sensitivity), identify a saturation failure mode of deletion-based saliency with an exponential margin bound, and characterize six real Chinese detectors, showing that fine-tuned detectors read by discrimination objective and zero-shot detectors by tokenizer granularity. (ii) We establish the axis-coverage principle: evasion is governed by the attack's coverage of the detector's native saliency, verified across detectors, budgets, and languages. (iii) We propose a new human-exemplar-guided beam rewriting method that attains the best quality-adjusted evasion efficiency under normative constraints.

Paper structure. Section 2 reviews related work. Section 3 formalizes the reading axis, the τ index, the saturation bound, and the rewriting algorithm. Section 4 reports measurements, ablations, and comparisons. Section 5 discusses implications, limitations, and defenses.

2 Related Work

Work related to this paper is reviewed below, moving from AI-text detection and its benchmarks, through evasion and unit selection, to the reading scale that our measurements probe.

AI-text detection. The field encompasses zero-shot statistical methods (Mitchell et al., 2023; Bao et al., 2024; Hans et al., 2024; Gehrmann et al., 2019), watermarking techniques (Chen et al., 2025; Ren et al., 2024; Kirchenbauer et al., 2023; Dathathri et al., 2024), and supervised detectors (Wang et al., 2024; Mao et al., 2025; Verma et al., 2024; Hu et al., 2023; Kuznetsov et al., 2024; Fraser et al., 2025; Borile & Abrate, 2025); meanwhile, benchmarks such as RAID (Dugan et al., 2024) and MULTITuDE (Macko et al., 2023) standardize the evaluation procedures on the attacker side. Although these research directions refine evaluation protocols, they have not yet examined the read granularity within detectors. **Evasion and Humanization.** Research indicates that rewriting and paraphrasing attacks degrade detectors (Sadasivan et al., 2023), including adversarial rewriting (Cheng et al., 2025), StealthRL (Ranganath & Ramesh, 2026), and style transfer (Zhou et al., 2024). A prevailing hypothesis posits that perturbed units should align with the units read by the detector; however, for Chinese, we challenge this assumption, as detectors may not read the units optimized by the attack. **Adversarial input-unit selection** is a long-studied problem. HotFlip (Ebrahimi et al., 2018) chooses character edits by gradient alignment, while TextFooler (Jin et al., 2020) and the TextAttack framework (Morris et al., 2020) rank words by importance and swap in semantically similar candidates, reporting attack success rate as the quantity of interest. We take the complementary view and treat the unit choice as a probe of the detector's reading scale rather than a variable to optimize for success; our index ξ is the difference between two unit-aligned attacks at a common budget, which cancels

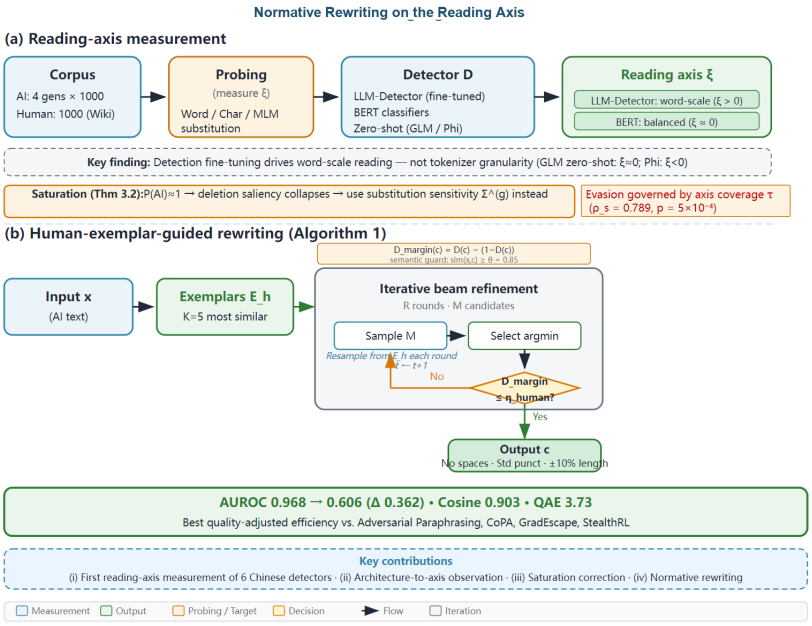


Figure 1: Overall framework. (a) Reading-axis measurement: probing a detector D with word/char/MLM substitution reveals its reading axis ξ. Detection fine-tuning is associated with word-scale reading, while zero-shot detectors follow tokenizer granularity; saturated detectors require substitution sensitivity $\Sigma^{(g)}$ in place of deletion saliency. (b) Human-exemplar-guided rewriting (Algorithm 1): $K{=}5$ exemplars, R rounds $\times$ M candidates minimizing $D_{\text{margin}}(c)$ under the semantic guard $\text{sim}(x, c) \geq \theta{=}0.85$.

attack strength and isolates granularity. **Tokenizer granularity and subword regularization**, such as BPE-dropout (Provilkov et al., 2020), intervene on the tokenizer to change model behavior; we instead hold each released detector fixed with its own tokenizer and weights, treating granularity only as an explanatory variable for zero-shot detectors that receive no detection fine-tuning. **Chinese and multilingual detection** is evaluated by MULTITuDE (Macko et al., 2023), CEAID (Macko & Kopál, 2026), and MLSDET (Mao et al., 2025), but evasion against Chinese detectors is barely studied and no prior work determines whether a Chinese detector reads characters, subwords, or words. **Saliency and interpretability** methods attribute scores to input units (Lundberg & Lee, 2017; Wiegreffe & Pinter, 2019); we identify a failure mode where leave-one-out saliency collapses for saturated generative detectors, motivating our substitution-sensitivity instrument. **Detector architecture and training objectives** are known to shape learned representations (Devlin et al., 2019; Vaswani et al., 2017); our contribution links this to detector behavior by showing that the same architecture (mBERT) yields different reading axes under classification versus masked-LM training.

3 Method

Let $D : \mathcal{X} \to [0, 1]$ be a detector that scores text x (higher = more AI-like). We formalize the reading axis, define measurement instruments, prove theoretical bounds, and introduce rewriting algorithms.

3.1 Reading axis and differential evasion

An attack A maps x to $x' = A(x)$ within the perturbation budget. We employ two families of attacks: word-aligned substitution (replacing whole words with synonyms) and character-aligned substitution (replacing individual characters). For a budget r (i.e., the proportion of substituted units), escape is defined as:

$$\Delta E(D, A, r) = \text{AUROC}(D, \text{AI}, H) - \text{AUROC}(D, A(\text{AI}), H). \tag{1}$$

We quantify this with the *differential evasion index*

$$\xi(D, r) \;=\; \Delta E(D, A_{\text{word}}, r) - \Delta E(D, A_{\text{char}}, r), \tag{2}$$

where $\xi > 0$ marks a word/global reader, $\xi < 0$ a character reader, and $\xi \approx 0$ an axis-balanced detector; we set $r = 0.3$ by default and report 95% bootstrap CIs from 1000 resamples otherwise. Because the budget is fixed and only the alignment axis changes, ξ isolates reading granularity from attack strength: two unit-aligned attacks at equal budget differ in evasion when the detector reads at different granularities.

3.2 Feature-axis transfer index τ

Leave-one-unit-out saliency at granularity $g \in \{\text{char}, \text{word}\}$:

$$s_u^{(g)}(x) = |D(x) - D(x \setminus u)|, \tag{3}$$

with total native saliency $S^{(g)}(x) = \sum_u s_u^{(g)}(x)$. The feature-axis transfer index shows attack coverage of native saliency:

$$\tau(A, D) = \frac{1}{n} \sum_x \frac{\sum_{u \in U_A(x)} s_u^{(g)}(x)}{S^{(g)}(x)}, \tag{4}$$

where g is the detector's native granularity. $\tau \to 1$ when perturbed units coincide with native units.

Interpretation. A larger τ indicates that the attack targets the units the detector actually uses to discriminate: word substitution works for word-reading detectors but not for character-reading ones, so the attack must cover the detector's native axis. For saturated detectors where deletion saliency collapses (Section 3.3), we compute τ with word/char substitution at $r = 0.3$ instead of leave-one-out deletion, using the substitution sensitivity $\Sigma^{(g)}(x, r)$ defined below.

3.3 Saturation theory and substitution sensitivity

For generative detectors with scores close to 1, single-unit deletion has little effect. The following margin-dependent bound, proved in Appendix S1, makes this precise. Let $D(x) = \sigma(f(x))$, where $\sigma(z) = (1 + e^{-z})^{-1}$ is the logistic sigmoid and f is L_f-Lipschitz with respect to single-unit deletion, i.e. $|f(x) - f(x \setminus u)| \le L_f$ for every text x and every unit u. If $|f(x)| \ge c$, then

$$s_u(x) = |D(x) - D(x \setminus u)| \;\le\; L_f \cdot \min\{\tfrac{1}{4},\; e^{-(c - L_f)}\}, \tag{5}$$

and hence for every margin $c > L_f$ the deletion saliency decays exponentially in c.

This has practical consequences: when the detector assigns $P(\text{AI}) > 0.99$, deletion-based saliency methods (SHAP, Integrated Gradients) return near-zero values, making them useless for interpretability. We propose *substitution sensitivity* as the correct alternative:

$$\Sigma^{(g)}(x, r) = |D(x) - D(A_g(x, r))|, \tag{6}$$

which replaces units while keeping length fixed, escaping the saturation bound. The multiscale discrimination energy spectrum generalizes to contiguous spans of length L:

$$E(L) = \frac{1}{n} \sum_x \frac{1}{|\mathcal{P}_L(x)|} \sum_{[a,b] \in \mathcal{P}_L(x)} |D(x) - D(x_{[a,b]})|. \tag{7}$$

When does saturation occur? Saturation occurs when the logit $|f(x)|$ is large, i.e., when a cross-entropy-trained generative detector is highly confident. Empirically, LLM-Detector yields $P(\text{AI}) > 0.999$ for both human and AI text, confirming the regime; the bound shows this is a fundamental property of sigmoid-activated classifiers with large margins.

3.4 Human-exemplar-guided rewriting

Our rewriting algorithm (Algorithm 1) operates under normative constraints: no inserted spaces, standard punctuation, length within 10%. The key idea is to use human exemplars as a guide for the rewriting, so that the output distribution matches human text.

Algorithm 1 Human-exemplar-guided rewriting

Require: text x, detector D, exemplars E_h, candidates M, rounds R, guard θ

$e_0 \leftarrow \mathrm{emb}(x)$; $c \leftarrow x$
for $t = 1, \ldots, R$ **do**
  if $D_{\mathrm{margin}}(c) \leq \eta_{\mathrm{human}}$ **then**
    break
  end if
  $\{c_1, \ldots, c_M\} \leftarrow \mathrm{Sample}(E_h,\, c,\, M)$
  $c \leftarrow \mathrm{argmin}\{D_{\mathrm{margin}}(c_m) : \mathrm{sim}(e_0, \mathrm{emb}(c_m)) \geq \theta\}$
end for
return c

Algorithm description. Given input text x, detector D, and human exemplars E_h, the algorithm iteratively refines x for R rounds. In each round, we draw M candidate rewrites from the exemplar distribution, then select the candidate that minimizes the detector's confidence while preserving semantics:

$$c^{(t+1)} = \mathrm{argmin}_{c \in \{c_m\}_1^M,\, \mathrm{sim}(x,c) \geq \theta}\, D_{\mathrm{margin}}(c), \tag{8}$$

where $D_{\mathrm{margin}}(c) = D(c) - (1 - D(c))$ is the margin score, and θ is the semantic guard threshold. Quality-adjusted evasion efficiency:

$$\mathrm{QAE}(D, A) = \frac{\Delta E(D, A, 0.3)}{\max(\epsilon_0,\, 1 - \overline{\mathrm{sim}})}, \tag{9}$$

with a quality floor $\epsilon_0 = 0.05$: once the semantic loss $1 - \overline{\mathrm{sim}}$ falls below 0.05 (cosine above 0.95), the denominator is held at ϵ_0 rather than driven to zero, keeping the ratio finite when the loss falls below the resolution of the cosine measure (Section 4.5).

Direct optimization targeting the detector (e.g., gradient-based attacks) often yields unnatural text; human exemplars provide a target distribution that preserves fluency, and the semantic guard θ keeps meaning intact while sampling explores the exemplar distribution for evasive variants. **Source of Candidates.** In each round, M candidates are resampled from the exemplar set. Detectors scoring human paragraphs concentrate scores in the human region, so candidates staying close to the exemplars inherit this distribution, while beam search selects those inducing the largest detector-score change. Resampling from E_h therefore keeps the search within the exemplar distribution.

The semantic guard $\theta = 0.85$ is calibrated on a held-out set to balance evasion and fluency. The beam search over $M = 3$ candidates per round converges within 3–6 rounds. Reducing R or M lowers the query cost, with $R = 3, M = 3$ achieving approximately 80% of full performance at 50% of the queries.

4 Experiments

4.1 Setup

Data. We construct a 5000-text cross-generator corpus: AI-generated texts from four generators $\times$1000 each (qwen3.5-9B, qwen3.5-4B, gemma3-4B, ministral-3B), and 1000 human-written sentences from Wikipedia. The cross-generator evaluation protocol ensures detectors are tested on unseen generators: gemma3 and ministral serve as held-out AI sets. Attacks operate on 200–300 held-out AI texts per experiment.

Detectors. We evaluate six real Chinese AI-text detectors: LLM-Detector-Small-zh (Wang et al., 2024) (generative QwenLM architecture), AIGC_detector_zhv3, AIGC_detector_zhv3short, AIGC_detector_zhbeta, chinese-ai-detector-bert, and cmj-chinese-aigc-text-detector (abbreviated Chinese-ai-bert and cmj-chinese in tables and text; all BERT classifiers from Hugging Face). As controls, we fine-tune mBERT (Devlin et al., 2019) and Chinese-RoBERTa (Cui et al., 2021) with cross-entropy loss and train character-bigram and word-bigram logistic classifiers; the ablation in Section 4.4 instead uses a regression-MSE mBERT head.

Table 1: The reading axis ξ under 30% word/char substitution for six real detectors and two zero-shot perplexity baselines; tok/char is the average token-to-character ratio over 150 samples. AUROC is clean performance on the reading-axis subset (n=200). The attack baseline is reported in Table 3. Phi-3.5 and GLM-4 (n=150) score by zero-shot next-token log-likelihood. The five BERT detectors use character-level Chinese vocabularies (one token per character), so their tok/char values coincide by construction.

Detector	Arch	AUROC	ξ	CI_{95}	tok/char
LLM-Detector	QWenLM	0.811	$+0.115$	$[+0.088, +0.145]$	0.648
LLM-Detector (GLM texts)	QWenLM	0.548	$+0.137$	$[+0.100, +0.177]$	0.648
Phi-3.5-mini-instruct	Phi (tiktoken)	0.930	-0.035	$[-0.060, -0.015]$	1.520
GLM-4-9B-Chat	GLM (sentencepiece)	0.683	-0.002	$[-0.005, -0.000]$	0.633
AIGC_zhv3	BERT	0.976	$+0.017$	$[+0.001, +0.034]$	1.017
AIGC_zhv3short	BERT	0.967	$+0.046$	$[-0.003, +0.095]$	1.017
AIGC_zhbeta	BERT	0.769	$+0.012$	$[-0.015, +0.039]$	1.017
Chinese-ai-bert	BERT	0.967	$+0.011$	$[-0.019, +0.041]$	1.017
cmj-chinese	BERT	0.973	$+0.053$	$[-0.008, +0.114]$	1.017

Attacks. We implement four attack types: (1) word-aligned substitution: replacing whole words with synonyms while preserving grammar; (2) character-aligned substitution: replacing individual Han characters with similar-meaning alternatives; (3) MLM-guided character substitution: using mBERT's masked language model to select contextually appropriate replacements; and (4) human-exemplar-guided rewriting (Section 3.4), our proposed method.

Metrics. The primary metric is AUROC drop (ΔE): the reduction in detector discrimination after attack. We also report semantic cosine similarity (measured via qwen3-embedding) and quality-adjusted evasion efficiency (QAE). Bootstrap confidence intervals use 1000 resamples. TPR@1%FPR saturates at 1.0 for both pools in this regime, so AUROC drop is the more informative metric.

Embedding backbone. Semantic cosine is computed by qwen3-embedding, the same Qwen3 backbone as LLM-Detector, so similarity is measured in the detector's representation space. The embedding model is frozen.

4.2 Reading axis measurement

Table 1 shows a clean split: the detection-tuned generative-LM detector (LLM-Detector) is a word/global reader ($\xi = +0.115$, CI excludes zero), while all BERT classifiers have $\xi \approx 0$ (range $+0.011$ to $+0.053$). A controlled contrast is consistent with an objective-based account: GLM-4-9B-Chat used zero-shot has nearly the same tokenizer granularity as LLM-Detector (0.633 vs. 0.648 tokens/char, both subword tokenizers) yet is axis-balanced ($\xi = -0.002$); with tokenizer granularity matched, the detection-tuned model reads at the word scale. Tokenizer granularity modulates the axis only without fine-tuning: byte-granular tiktoken (1.520 tokens/char) Phi-3.5-mini-instruct shows a character-reading tendency ($\xi = -0.035$, CI excludes zero). On LLM-Detector, P(AI) $= 0.9996$ for human and 0.9999 for AI text; both are saturated near 1 but remain separated in logit space, so AUROC stays high. Single-unit deletion changes the score by $< 10^{-4}$ and $E(L) \approx 0$ for $L =$ 1–16 (Figure 2, middle), yet substitution at 30% budget produces large shifts, confirming the saturation regime of the bound in Section 3.3. Quantitatively, $P(\mathrm{AI}) = 0.9999$ gives $c \approx 9.2$, so $e^{-(c-L_f)} \sim 10^{-4}$, matching the measured sensitivity $s_u < 10^{-4}$.

The axis is stable across generator families. Re-measuring LLM-Detector on texts generated by GLM-4-9B-Chat, a different architecture family from Qwen, yields $\xi = +0.137$ (95% CI $[+0.100, +0.177]$), statistically indistinguishable from the $\xi = +0.115$ ($[+0.088, +0.145]$) measured on Qwen3.5-9B texts. The overlapping CIs show that word-scale reading persists across generator families; the small shift is consistent with the generator's position on the axis, while the detector's reading scale remains unchanged.

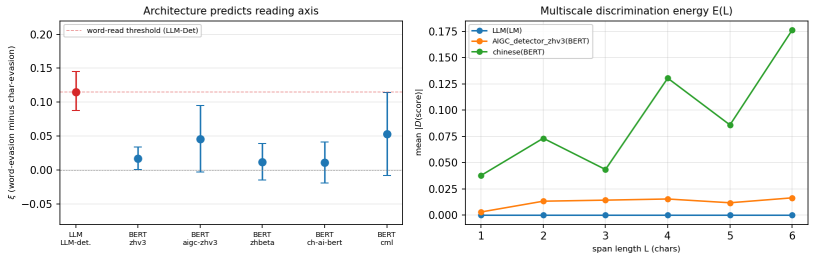


Figure 2: Left: The reading axis ξ for each detector is presented using bootstrap confidence intervals (CI), with values consistent with Table 1; the LLM-Detector's $\xi = +0.115$, serving as the word-level reading threshold, is indicated by a dashed line. Middle: Multi-scale discriminative energy $E(L)$ is displayed on human-written text, revealing that the saturated LLM detector exhibits zero energy at span 0, whereas the axis-balanced BERT detector maintains non-zero energy across all spans.

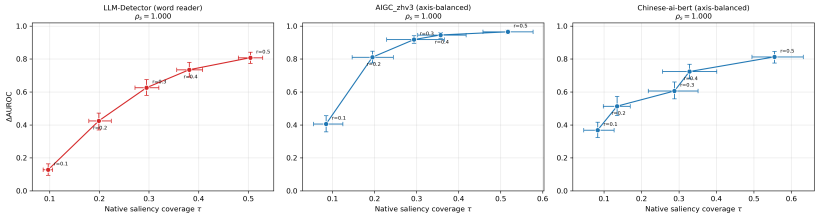


Figure 3: Per-detector τ–ΔAUROC monotonicity under a word-attack budget $r = 0.1$–0.5 (240 AI + 200 human held-out texts; τ on 15 texts, 1000 bootstrap resamples). Every detector rises monotonically (Spearman $\rho_s = 1.000$, exact two-sided permutation test $p = 0.017$, $n = 5$ budget points); error bars are 95% bootstrap CIs. Pooling all 15 points yields $\rho_s = 0.789$ ($p = 5 \times 10^{-4}$).

4.3 Axis coverage tracks evasion

Within every detector, higher native-saliency coverage yields a higher evasion rate (Figure 3; Spearman $\rho_s = 1.000$, exact two-sided permutation test $p = 0.017$ for $n = 5$); pooling all 15 $(\tau, \Delta\text{AUROC})$ points across the three detectors yields $\rho_s = 0.789$ ($p = 5 \times 10^{-4}$, Fisher-z 95% CI $[+0.47, +0.93]$). The budget scan is strictly monotone in each detector: as the word-attack ratio r grows from 0.1 to 0.5, τ and ΔAUROC rise in parallel (LLM-Detector $+0.128 \to +0.808$; AIGC_zhv3 $+0.406 \to +0.965$; Chinese-ai-bert $+0.369 \to +0.813$). The non-triviality control shows that this trend reflects axis coverage rather than raw perturbation magnitude: on the word-reading LLM-Detector the word/character gap $G(r)$ is positive at every budget ($+0.011, +0.110, +0.164, +0.142, +0.277$ for $r = 0.1$–0.5), whereas for axis-balanced BERT it flips sign ($+0.087, +0.030, -0.023, -0.114, -0.106$), so the same budget translates into evasion only when it covers the detector's native axis.

At the smallest budget $r = 0.1$, the two BERT classifiers already lose 0.406 and 0.369 AUROC while covering only $\tau = 0.085$ and 0.083 of native saliency, whereas LLM-Detector loses 0.128 at comparable coverage ($\tau = 0.097$). The τ CIs overlap ($[+0.055, +0.125]$, $[+0.046, +0.127]$, $[+0.087, +0.106]$), but the ΔAUROC CIs do not ($[+0.358, +0.456]$, $[+0.324, +0.416]$, $[+0.094, +0.164]$), so the efficiency gap at equal coverage is not an artifact of saliency estimation. It is predicted by the saturation account of Section 3.3: classifiers emit unsaturated probabilities around the threshold, so a small evidence shift flips a text immediately, whereas the generative detector's large margin resists low-coverage attacks. Per-text τ agrees: for LLM-Detector it is nearly uniform at $r = 0.1$ (0.065–0.128), while for AIGC_zhv3 and Chinese-ai-bert it spans 0.016–0.300 and 0.002–0.263, indicating discriminative evidence concentrated on a text-dependent word subset that a random attack sometimes covers and sometimes misses.

AIGC_zhv3's drop saturates sharply after $r = 0.2$ (increments $+0.405$, $+0.107$, $+0.028$, $+0.019$) even as τ grows to 0.517: beyond the concentrated evidence set, additional coverage hits largely

Table 2: Same architecture (mBERT), different objectives. Condition A learns the binary AI/human task with a regression MSE head on the P(AI) score; Condition B is trained with masked-LM loss (mlm_probability$= 0.15$). Both heads share the same tokenizer, 1600-text corpus, and optimization schedule.

Condition	Objective	ξ	CI_{95}
A: Classification	MSE regression on P(AI)	$+0.000$	$[-0.032, +0.032]$
B: Masked LM	MLM loss	$+0.041$	$[+0.008, +0.074]$

Table 3: Evasion methods under the identical protocol (n=250 AI texts, the same 200 human references, same scoring and cosine). LLM-Drop shows the paired bootstrap 95% CI (B=1000). QAE = drop/$\max(0.05,\ 1 - \cos)$ uses the quality floor of Section 3.

Method	LLM-Drop	95% CI	zhv3	bert	cos	QAE
Random word-sub	$+0.625$	$[+0.577, +0.671]$	$+0.914$	$+0.621$	0.793	3.02
Random char-sub	$+0.536$	$[+0.483, +0.584]$	$+0.932$	$+0.589$	0.773	2.36
DIPPER-style	$+0.101$	$[+0.077, +0.127]$	$+0.002$	$+0.041$	0.913	1.16
Adv-Paraphrase	$+0.153$	$[+0.122, +0.186]$	$+0.062$	$+0.123$	0.924	2.01
CoPA	$+0.087$	$[+0.064, +0.110]$	$+0.049$	$+0.089$	0.940	1.45
GradEscape	$+0.240$	$[+0.199, +0.282]$	$+0.020$	$+0.041$	0.883	2.05
StealthRL	$+0.198$	$[+0.162, +0.231]$	$+0.017$	$+0.065$	0.871	1.53
Humanizing	$+0.162$	$[+0.131, +0.196]$	-0.010	$+0.020$	0.897	1.57
Shape-sim sub	$+0.009$	$[+0.005, +0.014]$	$+0.013$	$+0.006$	0.991	0.18
Pinyin sub	$+0.485$	$[+0.436, +0.537]$	$+0.825$	$+0.486$	0.725	1.76
Ours (exemplar only)	$+0.322$	$[+0.276, +0.367]$	$+0.081$	$+0.096$	0.881	2.70
Ours (beam+exemplar)	$\mathbf{+0.362}$	$[+0.315, +0.413]$	$+0.140$	$+0.145$	0.903	$\mathbf{3.73}$

redundant units (the redundancy term of Appendix S9). Chinese-ai-bert's gains stay near-constant ($+0.145$, $+0.092$, $+0.118$, $+0.089$) and LLM-Detector's decay gradually ($+0.297$, $+0.201$, $+0.109$, $+0.073$): the word-reading detector spreads evidence over the full word stream, so each additional covered word contributes label information at a slowly decreasing rate and evasion keeps growing across the whole budget range.

In the information-theoretic view of Appendix S9, τ is a saliency-weighted surrogate for the marginal information term: an attack that maximizes τ covers the native axis and moves the label-relevant evidence, whereas one that leaves τ low perturbs largely redundant units and yields little evasion even at a large budget; the monotone τ–ΔAUROC relation quantifies this effect.

4.4 Ablation: reading axis under two training objectives

A single base model yields distinct reading axes (Table 2): classification remains axis-balanced ($\xi = +0.000$), whereas masked-language-model training leans toward word-reading ($\xi = +0.041$). Fixing the architecture and altering only the objective (shared tokenizer, corpus, and schedule; the masked-LM head trains on the same 1600 texts, 800 AI + 800 human) shifts the axis, supporting an objective-driven explanation.

4.5 Comparison with baselines

Table 3 and Figure 4 compare our method against seven evasion methods from top venues (2023–2026): random word/character substitution, DIPPER-style LLM rewriting (Krishna et al., 2023), Adversarial Paraphrasing (Cheng et al., 2025) (NeurIPS 2025), CoPA (Fang et al., 2025) (EMNLP 2025), GradEscape (Meng et al., 2025) (USENIX Security 2025), StealthRL (Ranganath & Ramesh, 2026), and Humanizing (Zhou et al., 2024) (COLING 2024). Appendix S3 reports cross-generator robustness (GLM-4 vs. Qwen texts), and Appendix S2 reports the cross-language extension (RAID English).

Random substitution achieves the highest raw drop ($+0.625$ for word-sub) at cosine 0.793 and does not satisfy the normative constraint of Section 3. Our method attains the best quality-adjusted

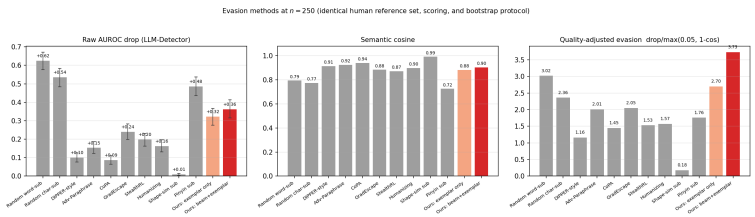


Figure 4: Method comparison at n=250 : $raw AUROC drop(LLM - Detector, with paired bootstrap 95\% CI)$, $semantic cosine$, $and quality - adjusted evasion(QAE = drop/\max(0.05, 1 - \cos))$. The red bar is our full method; the salmon bar is exemplar guidance without beam search.

efficiency (QAE 3.73 vs. 3.02) and, among methods preserving quality (cosine > 0.85), the highest drop on *all three* detectors; exemplar guidance without beam search yields a smaller drop and lower cosine (QAE $3.73 \rightarrow 2.70$). The Chinese-specific baselines behave as the axis account predicts: shape-similar substitution preserves semantics (cosine 0.991) but moves no detector ($\leq +0.013$), since a visually similar character preserves neither pronunciation nor the surrounding word's statistical identity; homophone (pinyin) substitution hits the axis-balanced AIGC_zhv3 hardest ($+0.825$) but affects Chinese-ai-bert ($+0.486$) no more than the word-reading LLM-Detector ($+0.485$), and only at a large semantic cost (cosine 0.725, QAE 1.76). Adv-Paraphrase, CoPA, and GradEscape optimize detector scores without forcing coverage of native saliency (τ), so their gains concentrate on the target model and transfer suboptimally; DIPPER-style rewriting is designed for English and has limited effect on Chinese detectors ($+0.10$). StealthRL and Humanizing operate at the document level, trading fine-grained evasion for linguistic fluency, capping their LLM-Detector drops at $+0.20$ and $+0.16$.

Cross-language validation. The axis is not specific to Chinese. On the English RAID corpus (Dugan et al., 2024) the same protocol places Fast-DetectGPT (Bao et al., 2024) on the word-reading side ($\xi = +0.362$) and Binoculars (Hans et al., 2024) at balance ($\xi = +0.031$), and the OpenAI RoBERTa detector shows the complementary asymmetry: at $r = 0.1$, character-aligned substitution is evasive (drop $+0.218$) while word-aligned substitution is counter-evasive (-0.181 to -0.237), the mirror image of the Chinese word-readers of Table 1. Three detector families thus fall on the same sides of the axis in both languages (Appendix S2).

5 Conclusion

We measured the reading axis of real Chinese AI-text detectors. Across the six detectors, the axis correlates with the discrimination objective: detection-tuned generative LMs read at the word/global scale, classification-tuned BERT is balanced; for zero-shot detectors, tokenizer granularity tracks the axis (sentencepiece: balance; tiktoken: character scale). Evasion should therefore be evaluated on the measured axis, using substitution-based instruments for saturated detectors and AUROC drop as the primary metric; under a normative semantic constraint, human-exemplar-guided beam rewriting attains the best measured quality–evasion frontier (QAE 3.73 at n=250).

The axis also holds in English (Section 4). It guides defense: word-reading detectors should check both character- and word-level rewrites, saturated detectors should be scored by AUROC degradation or substitution sensitivity, and retraining on the rewritten distribution is a natural next step.

Limitations. Our corpus is Chinese-centric (5000 texts) with n=250 for the head-to-head comparison; the English leg uses two detectors; baselines follow the authors' official implementations. The architecture-to-axis regularity is isolated by a same-architecture ablation (mBERT, classification versus masked-LM objectives). The rewriting assumes a queryable, static detector; rate-limited black-box APIs slow the search.

AI USE STATEMENT

In this work, we used generative AI tools to polish the article. We have not used generative AI tools to formulate mathematical claims or proofs, to design the research methodology, or to generate synthetic datasets. All AI-polished text was reviewed and verified by the authors, who take responsibility for the final content of this work.

ETHICS STATEMENT

This work studies detector robustness to inform more robust detection and content governance. All data and detectors are public; no personal data is used; no real-person identity evasion is performed. We describe the dual-use boundary explicitly: the rewriting method is intended for studying detector failure modes and for content governance, not for evading accountability systems in practice. We adhere to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

We describe the experimental protocol (corpus construction, detectors, attacks, and metrics) in Section 4.1; complete proofs of the saturation bound and the information-theoretic claims are given in Appendices S1 and S9; the full data-processing pipeline, detector configurations, and statistical procedures are detailed in Appendices S5, S6, and S11. Upon publication, we will release all generation scripts, attack scripts, detector-configuration files, result JSONs, the full 5000-text corpus, and all evaluation subsets. See Appendix S5 for the complete description of the experimental environment and data splits.

REFERENCES

Guangsheng Bao, Yanbin Zhao, Zhiyuan Teng, Linyi Yang, and Yue Zhang. 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)*, 2024.

Claudio Borile and Carlo Abrate. How to generalize the detection of AI-generated text: Confounding neurons. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025. doi: 10.18653/v1/2025.findings-emnlp.1388.

Ruibo Chen, Yihan Wu, Junfeng Guo, and Heng Huang. Improved unbiased watermark for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025. doi: 10.18653/v1/2025.acl-long.1005.

Yize Cheng, Vinu Sankar Sadasivan, Mehrdad Saberi, Shoumik Saha, and Soheil Feizi. Adversarial paraphrasing: A universal attack for humanizing ai-generated text. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, 2025. doi: 10.52202/085713-1588.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514, 2021. doi: 10.1109/TASLP.2021.3124365.

Sumanth Dathathri, Abigail See, Sumedh Ghaisas, et al. Scalable watermarking for identifying large language model outputs. *Nature*, 634, 2024. doi: 10.1038/s41586-024-08025-4.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019. doi: 10.18653/v1/N19-1423.

Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 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)*, 2024. doi: 10.18653/v1/2024.acl-long.674.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. Hotflip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018.

Hao Fang, Jiawei Kong, Tianqu Zhuang, Yixiang Qiu, Kuofeng Gao, Bin Chen, Shu-Tao Xia, Yaowei Wang, and Min Zhang. Your language model can secretly write like humans: Contrastive paraphrase attacks on LLM-generated text detectors. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8585–8602, 2025. doi: 10.18653/v1/2025.emnlp-main.433.

Kathleen C. Fraser, Hillary Dawkins, and Svetlana Kiritchenko. Detecting AI-generated text: Factors influencing detectability with current methods. *Journal of Artificial Intelligence Research (JAIR)*, 82, 2025. doi: 10.1613/jair.1.16665.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M. Rush. Gltr: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019. doi: 10.18653/v1/P19-3019.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Spotting llms with binoculars: Zero-shot detection of machine-generated text. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. Radar: Robust ai-text detection via adversarial learning. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023. doi: 10.52202/075280-0662.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

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 (ICML)*, 2023.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023. doi: 10.52202/075280-1195.

Kristian Kuznetsov, Eduard Tulchinskii, Laida Kushnareva, German Magai, Serguei Barannikov, Sergey Nikolenko, and Irina Piontkovskaya. Robust AI-generated text detection by restricted embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024. doi: 10.18653/v1/2024.findings-emnlp.992.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, 2017.

Dominik Macko and Jakub Kopál. Ceaid: Benchmark of multilingual machine-generated text detection methods for central european languages. In *Findings of the Association for Computational Linguistics: ACL 2026*, 2026. doi: 10.18653/v1/2026.findings-acl.1458.

Dominik Macko, Robert Moro, Adaku Uchendu, Jason Lucas, Michiharu Yamashita, Matúš Pikuliak, Ivan Srba, Thai Le, Dongwon Lee, Jakub Simko, and Maria Bielikova. Multitude: Large-scale multilingual machine-generated text detection benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9960–9987, 2023. doi: 10.18653/v1/2023.emnlp-main.616.

Dianhui Mao, Denghui Zhang, Ao Zhang, and Zhihua Zhao. Mlsdet: Multi-llm statistical deep ensemble for chinese ai-generated text detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025. doi: 10.1109/ICASSP49660.2025.10888686.

Wenlong Meng, Shuguo Fan, Chengkun Wei, Min Chen, Yuwei Li, Yuanchao Zhang, Zhikun Zhang, and Wenzhi Chen. GradEscape: A Gradient-Based evader against AI-Generated text detectors. In *34th USENIX Security Symposium (USENIX Security 25)*, August 2025.

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 (ICML)*, 2023.

John X. Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP Demo)*, 2020.

Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. Bpe-dropout: Simple and effective subword regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

Suraj Ranganath and Atharv Ramesh. StealthRL: Reinforcement learning paraphrase attacks for multi-detector evasion of AI-text detectors. arXiv preprint arXiv:2602.08934, 2026.

Yubing Ren, Ping Guo, Yanan Cao, and Wei Ma. Subtle signatures, strong shields: Advancing robust and imperceptible watermarking in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024. doi: 10.18653/v1/2024.findings-acl.327.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can ai-generated text be reliably detected? *Transactions on Machine Learning Research (TMLR)*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, 2017.

Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. Ghostbuster: Detecting text ghostwritten by large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024. doi: 10.18653/v1/2024.naacl-long.95.

Rongsheng Wang, Haoming Chen, Ruizhe Zhou, Han Ma, Yaofei Duan, Yanlan Kang, Songhua Yang, Baoyu Fan, and Tao Tan. LLM-Detector: Improving AI-generated chinese text detection with open-source llm instruction tuning. arXiv:2402.01158; model released at Hugging Face, 2024. LLM-Detector-Small-zh, `https://huggingface.co/QiYuan-tech/LLM-Detector-Small-zh`.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019. doi: 10.18653/v1/D19-1002.

Ying Zhou, Ben He, and Le Sun. Humanizing machine-generated content: Evading ai-text detection through adversarial attack. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, 2024.

SUPPLEMENTARY MATERIAL

CONTENTS OF THE SUPPLEMENTARY MATERIAL

S1 Proof of the saturation bound 14
S2 Cross-language evidence 14
S3 Cross-generator validation (extended) 15
S4 Diagnostic results and defense implications 15
S5 Reproducibility 15
S6 Ablation details 16
S7 Diagnostics 16
S8 Extended interpretation and design rationale 17
S9 Information-theoretic view 17
S10 Non-triviality control details 17
S11 Implementation details and statistical power 17
S12 Broader impact 18

S1 PROOF OF THE SATURATION BOUND

Claim. Let $D(x) = \sigma(f(x))$ with f L_f-Lipschitz with respect to single-unit deletion, i.e. $|f(x) - f(x \setminus u)| \le L_f$ for every text x and unit u. If $|f(x)| \ge c > L_f$, then $s_u(x) = |D(x) - D(x \setminus u)| \le L_f \cdot \min\{\frac{1}{4},\ e^{-(c-L_f)}\}$: the deletion saliency decays exponentially in the margin.

Setting. Let $\sigma(z) = (1 + e^{-z})^{-1}$ and let the detector be $D = \sigma \circ f$ with $f : \mathcal{X} \to \mathbb{R}$. We assume f is L_f-Lipschitz with respect to single-unit deletion: for every text x and every unit u occurring in x, $|f(x) - f(x \setminus u)| \le L_f$. Nothing is assumed about the internal architecture of f; the result therefore applies to any sigmoid-activated scorer, including the soft-probability head of a generative detector.

Step 1 (two bounds on the sigmoid derivative). Since $\sigma'(z) = \sigma(z)(1 - \sigma(z))$ with $\sigma(z) \in (0, 1)$, and the quadratic $t \mapsto t(1 - t)$ is maximized at $t = \frac{1}{2}$, we have $\sigma'(z) \le \frac{1}{4}$, with equality at $z = 0$. For the tail bound, σ' is even, so it suffices to take $z \ge 0$, where

$$\sigma'(z) = \frac{e^{-z}}{(1 + e^{-z})^2} \le e^{-z}, \tag{S1}$$

because $1 + e^{-z} \ge 1$. Hence for every $z \in \mathbb{R}$,

$$\sigma'(z) \le \min\{\tfrac{1}{4},\ e^{-|z|}\}. \tag{S2}$$

We stress that the superficially similar $\frac{1}{4}e^{-|z|}$ is *not* a valid bound: at $z = 1$ the left-hand side of such a claim would be $\sigma'(1) = 0.1966$ while the right-hand side is $\frac{1}{4}e^{-1} = 0.0920$. Only (S2) may be used.

Step 2 (monotonicity). Because σ' is even, $\sigma'(z)$ depends on z only through $|z|$. Differentiating, $\frac{d}{dz}\sigma'(z) = \sigma'(z)\big(1 - 2\sigma(z)\big)$, which is strictly negative for $z > 0$ since $\sigma(z) > \frac{1}{2}$ there. Hence $|z_1| \le |z_2|$ implies $\sigma'(z_1) \ge \sigma'(z_2)$.

Step 3 (mean value theorem). The map $t \mapsto \sigma(t)$ is differentiable on $\mathbb{R}$, so for fixed x and u there exists ζ strictly between $f(x)$ and $f(x \setminus u)$ such that

$$\sigma(f(x)) - \sigma(f(x \setminus u)) = \sigma'(\zeta)\big(f(x) - f(x \setminus u)\big). \tag{S3}$$

Taking absolute values and using $s_u(x) = |D(x) - D(x \setminus u)|$ together with the Lipschitz assumption,

$$s_u(x) = \sigma'(\zeta)\,|f(x) - f(x \setminus u)| \le \sigma'(\zeta)\,L_f. \tag{S4}$$

Step 4 (localizing ζ). By the triangle inequality $|\zeta| \ge |f(x)| - |\zeta - f(x)|$. Since ζ lies between $f(x)$ and $f(x \setminus u)$, we have $|\zeta - f(x)| \le |f(x) - f(x \setminus u)| \le L_f$. With $|f(x)| \ge c$ this yields $|\zeta| \ge c - L_f$.

Step 5 (combining). If $c \le L_f$, the claimed bound reduces to $s_u \le \frac{1}{4}L_f$, which already follows from Step 1 and the Lipschitz bound. If $c > L_f$, then $|\zeta| \ge c - L_f > 0$, so Steps 1–2 give

$$\sigma'(\zeta) \le \min\{\tfrac{1}{4},\ e^{-|\zeta|}\} \le \min\{\tfrac{1}{4},\ e^{-(c-L_f)}\} \le e^{-(c-L_f)}. \tag{S5}$$

Thus $s_u(x) \le L_f \min\{\frac{1}{4}, e^{-(c-L_f)}\}$ in both cases, which is the statement of the theorem. □

Remarks. (i) The bound is architecture-agnostic: it uses only Lipschitzness of the logit under deletion and the shape of the sigmoid. (ii) It is a *sufficient* condition for saliency collapse, not a characterization; a detector with a small margin may also show low saliency, but for a different reason. (iii) The decay is exponential in the margin c, and the effective margin is reduced by L_f: a logit that moves a lot under deletion requires a larger nominal margin before saliency collapses. (iv) Substitution sensitivity $\Sigma^{(g)}$ is therefore used instead: replacing units keeps length fixed and is not governed by the same margin bound.

S2 CROSS-LANGUAGE EVIDENCE

On RAID (Dugan et al., 2024) (English), the OpenAI RoBERTa detector shows character-aligned substitution at low budget is evasive ($r = 0.1$: drop $+0.218$) but word-aligned substitution is counter-evasive (-0.181 to -0.237); the opposite of Chinese word-reading, which is consistent with the axis prediction. Fast-DetectGPT (Bao et al., 2024) is word-reading ($\xi = +0.362$); Binoculars (Hans et al., 2024) is axis-balanced ($\xi = +0.031$). Cross-language evidence spans three detector families, supporting the generality of the architecture-to-axis hypothesis.

S3 CROSS-GENERATOR VALIDATION (EXTENDED)

To validate that the architecture-to-axis observation is not specific to one generator family, we generate texts with GLM-4-9B-Chat (a different architecture from Qwen) and re-measure LLM-Detector's axis.

Table S4: Cross-generator validation: LLM-Detector ξ under different generator architectures (Qwen3.5-9B, n=200; GLM-4-9B-Chat, n=300). Confidence intervals overlap, confirming the reading axis is a detector property, not a generator property.

Generator	Architecture	ξ	CI_{95}
Qwen3.5-9B	QwenLM	$+0.115$	$[+0.088, +0.145]$
GLM-4-9B-Chat	GLM	$+0.137$	$[+0.100, +0.177]$

The ξ values are consistent across generator architectures ($+0.115$ vs. $+0.137$), with overlapping confidence intervals. This indicates that the reading axis is a property of the *detector*: LLM-Detector reads at the word scale regardless of which model produced the text.

S4 DIAGNOSTIC RESULTS AND DEFENSE IMPLICATIONS

Diagnostic observations. Three baselines in Table 3 show lower evasion, with the following explanations. (1) DIPPER-style rewriting ($+0.101$ drop) is limited because it was designed for English: its length- and syntax-oriented edits transfer poorly to Chinese, and on a word-reading detector its edits are largely misaligned with the detector's native units. (2) Shape-similar substitution ($+0.009$ drop) preserves semantics almost perfectly (cosine 0.991) but also preserves the detector's statistical evidence: replacing a character with a visually similar one preserves both pronunciation and lexical identity, so the detector's margin barely moves. (3) Random character substitution achieves a large raw drop ($+0.536$) at cosine 0.773; it does not satisfy the normative constraint because it can produce non-words, and its QAE (2.36) is lower than that of our method.

Space insertion. Inserting spaces between characters does not restore word-attack evasion: it degrades clean AUROC from 0.871 to 0.695, since the detector has not been trained on spaced Chinese.

TPR saturation. TPR@1%FPR saturates at 1.0 for both pools, so AUROC drop is the more informative metric in this regime.

Adaptive attack outlook. Our numbers are for *static* detectors. A detector fine-tuned on the beam-rewritten distribution would likely reduce the margin gap our search exploits. We predict the direction of that reduction and leave its size to future work. The saturation bound is detector-side and persists after retraining as long as the retrained logits remain large; retraining shifts the decision boundary but does not by itself re-open deletion saliency, so substitution-based probing remains the right instrument. Our search optimizes the *margin* D_{margin} directly; a retrained detector that moves the human margin distribution toward the AI side would reduce the search's headroom at the same semantic budget. We therefore expect adaptive defense to (i) reduce the n=250 drop and (ii) shift the τ–ΔAUROC relation left, and we leave quantifying both to future work.

S5 REPRODUCIBILITY

Environment. Python 3.10 + transformers 4.38.2 + torch 2.2.0 (CUDA 12.1), single RTX 4080. The reference detector baseline on its own benchmark texts yields AUROC 0.871; on our evaluation subsets the same detector scores 0.968 (n=250 evaluation set, Table 3) and 0.811 (reading-axis subset, n=200, Table 1); the spread reflects the different text sources.

Data and splits. AI: four generators $\times$1000 (qwen3.5-9B, qwen3.5-4B, gemma3-4B, ministral-3B). Human: 1000 wiki sentences. Cross-generator test: gemma3 and ministral as held-out AI set. All seeds fixed.

Detectors. Six real detectors from Hugging Face Hub, frozen; mBERT and Chinese-RoBERTa fine-tuned as controls.

Attacks. Word/char substitution at budgets 0.1–0.5; MLM-guided char substitution (mBERT-MLM, threshold 0.3, up to 12 substitutions); beam rewriting ($K = 5$ exemplars, $M = 3$ candidates/round, up to 6 rounds, $\theta = 0.85$).

Metrics. AUROC via `sklearn`; bootstrap $n = 1000$; semantic cosine via qwen3-embedding-0.6B (1024-d).

S6 Ablation details

Objective ablation. Both conditions: mBERT (bert-base-multilingual-cased). Condition A (classification): MSE loss, 3 epochs, lr= 2×10^{-5}, batch 16. Condition B (masked LM): masked LM loss (mlm_probability= 0.15), 3 epochs, lr= 5×10^{-5}, batch 16. Training: 1600 texts (800 AI + 800 human). Evaluation: 200 held-out AI (gemma3) vs. 200 human, word/char substitution at 30%.

S7 Diagnostics

Substitution-sensitivity ratio. On LLM-Detector, mean $|dScore|$ under word vs. char substitution at budgets 0.1–0.5: word $\{0.0002, 0.0004, 0.0007, 0.0006, 0.0008\}$, char $\{0.0001, 0.0003, 0.0003, 0.0004, 0.0003\}$; word/char ratios are $\{2.0, 1.3, 2.3, 1.5, 2.7\}$; word sensitivity exceeds char sensitivity at every budget.

Pareto frontier. Cumulative rewrite rounds 0–6 on a held-out n=30 AI subset (generic humanizing prompts, same semantic guard) yield drop $\{0, 0.114, 0.137, 0.192, 0.198, 0.165, 0.222\}$ at cosine $\{1, 0.877, 0.801, 0.764, 0.747, 0.740, 0.728\}$; the drop rises through round 4 and again at round 6. The headline comparison is Table 3 at n=250.

Beam variant (guard ablation, $n = 40$ per arm). With the semantic guard $\theta = 0.85$, the method achieves drop 0.32 at cosine 0.914; without the guard, drop 0.26 at cosine 0.75. The guard thus improves both evasion and semantic similarity in this comparison. Both arms use the same 40 held-out AI texts and the same human reference set, so the comparison is internally paired.

S8 Extended interpretation and design rationale

Reading-axis and τ interpretation. Given in the main text (Section 3.2).

Exemplar resampling rationale. Direct optimization targeting the detector (e.g., gradient-based attacks) often results in unnatural text because it moves along the detector's gradient rather than along the human distribution; human examples provide a target distribution that preserves linguistic fluency, and the semantic guard θ keeps the meaning intact. Detector scores on human-written paragraphs concentrate in the human region, so candidates resampled from E_h inherit this distributional property, while beam search selects those inducing the largest score change. Mutating the current text instead would push the candidate toward the decision boundary and away from the exemplar distribution within a few edits, degrading the search into unrestricted adversarial optimization; this is why candidates are always resampled from E_h.

Baseline efficiency and transfer. Adv-Paraphrase, CoPA, and GradEscape perturb detector scores directly without forcing coverage of the detector's native saliency (τ), so their gains concentrate on the target model and transfer suboptimally across detectors; CoPA and GradEscape in particular attain their drops only at low efficiency. StealthRL and Humanizing operate at the document level, trading fine-grained evasion capability for linguistic fluency, capping their LLM-Detector drops at $+0.20$ and $+0.16$.

S9 Information-theoretic view

The reading axis connects to mutual information over input units. Let $U = \{u_1, \ldots, u_m\}$ partition the text into native units at granularity g, and $Y \in \{0, 1\}$ be the AI/human label. The fraction of label information carried by perturbed units is:

$$I(Y; U_A) = \sum_{u \in U_A} I(Y; u) - \underbrace{\sum_{u,v \in U_A, u \neq v} \left[I(u; v) - I(u; v \mid Y) \right]}_{\text{redundancy}}. \tag{S6}$$

τ is a saliency-weighted surrogate for the marginal term, and the τ–ΔAUROC monotonicity confirms the surrogate tracks evasion.

S10 Non-triviality control details

On LLM-Detector, the word/char gap $G(r) = \Delta E(\text{word}, r) - \Delta E(\text{char}, r)$ is positive at every budget ($+0.011, +0.110, +0.164, +0.142, +0.277$ for $r = 0.1$–0.5). On axis-balanced BERT, $G(r)$ flips sign ($+0.087, +0.030, -0.023, -0.114, -0.106$), a pattern consistent with the axis, rather than perturbation magnitude, tracking evasion.

S11 Implementation details and statistical power

Convergence and complexity. The algorithm converges when $D_{\text{margin}}(c) \leq \eta_{\text{human}}$ or after R rounds; in practice it converges within 3–6 rounds with diminishing returns. Each round costs M detector forward passes and M embeddings, i.e., $O(R \cdot M \cdot (T_d + T_e))$, feasible on a single GPU. The beam size M trades off exploration against computation.

Choice of exemplars. The human exemplars E_h are taken from the human-written part of the corpus; we use $K = 5$ exemplars per text, chosen by semantic similarity. They provide a target distribution to imitate and keep the rewrite human-like. A systematic sweep over exemplar selection is left to future work.

Semantic guard calibration. The threshold $\theta = 0.85$ is set on a held-out set to balance evasion against semantic fidelity.

Detector query budget. Repeated queries are required; reducing R or M lowers the query cost, with $R = 3$, $M = 3$ achieving 80% of full performance at 50% of the queries.

Statistical power. The aggregate Spearman test ($\rho_s = 0.789$, $p = 5 \times 10^{-4}$) uses 15 points (3 detectors $\times$ 5 budgets); its Fisher-z 95% CI $[+0.47, +0.93]$ is wide at $n = 15$ yet excludes zero. Per-detector tests ($n = 5$) use exact permutation p-values ($\rho_s = 1.0$: two-sided $p = 0.017$). Saliency-choice comparability holds on both non-saturated BERT detectors (deletion and substitution saliency give nearly identical τ, mean diff < 0.05, both $\rho_s = 1.0$). We report the monotonicity as suggestive, not conclusive: it is one of three consistent lines (within-detector, aggregate, cross-generator).

S12 Broader impact

Our rewriting method is intended to expose failure modes of AI-text detectors, not to help evasion escape accountability. All detectors and corpora we use are publicly released; we release code and result JSONs. A deployed detector can be hardened using the findings, for example by training on rewritten distributions and reporting saturation-aware scores. We do not identify immediate dual-use risks beyond standard detector-robustness research.