# ANCHORWRITE: DETECTOR-QUERY-FREE REWRITING OF LLM-GENERATED TEXT VIA HUMAN-REFERENCE ALIGNMENT

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Large language models (LLMs) generate fluent text that can remain statistically distinguishable from human writing. Existing post-generation rewriting approaches are often evaluated primarily through detector reduction, making it difficult to separate detector-specific optimization from movement toward an independently defined human linguistic reference. We introduce **AnchorWrite**, a detector-query-free framework that treats rewriting as constrained alignment to a fixed human reference. AnchorWrite constructs cross-order language-model statistics from human-authored text, calibrates them against a length-conditioned human reference, and uses token-level evidence to localize regions of linguistic deviation. A pretrained language model proposes local rewrites, which are accepted only when they reduce human-reference distance while satisfying semantic, linguistic, and edit constraints. Across six domains, mean human-reference distance reduction increases from 0.068 at 5.60% realized editing to 0.363 at 24.32%. On frozen outputs, AUROC decreases by 0.057–0.089 across six external AI-text detection methods, despite no detector feedback during rewriting. Matched-edit comparisons with instruction-based humanization and neural paraphrasing show that detector responses remain transformation-dependent. At the strongest regime, BERTScore remains 0.968 while perplexity increases from 8.88 to 10.71. These results position human-reference alignment as an interpretable control signal for constrained rewriting, with reduced detector separability emerging as an external transfer effect rather than an optimization target.

## 1 INTRODUCTION

Large language models (LLMs) can generate fluent text across academic, journalistic, and creative domains, creating challenges for provenance and attribution. AI-generated text can remain statistically distinguishable from human writing, yet post-generation rewriting can substantially alter detector signals. Existing detection approaches include likelihood- and rank-based statistics, perturbation and analytical methods, and supervised classifiers (Gehrmann et al., 2019; Solaiman et al., 2019; Mitchell et al., 2023; Bao et al., 2024; Hans et al., 2024), whose performance can degrade after paraphrasing or rewriting (Krishna et al., 2023; Sadasivan et al., 2023).

Post-generation humanization typically relies on generic transformations or detector-guided optimization. Generic paraphrasing and instruction-based rewriting alter surface form without an explicit human-derived target, whereas detector-guided approaches optimize against detector or surrogate feedback. Although either strategy may reduce detector confidence, detector reduction alone does not characterize whether rewriting follows an independently defined linguistic objective.

We therefore ask whether a fixed statistical reference derived from human writing can guide controlled rewriting without querying the downstream detectors used for evaluation. We introduce **AnchorWrite**, a detector-query-free framework that formulates rewriting as constrained human-reference alignment. Cross-order language-model statistics identify token-level anomalies, a pretrained language model proposes local rewrites, and a length-conditioned document-level reference distance determines which valid candidate to retain. After each accepted edit, the document is rescored and localization is recomputed, producing an iterative *localize–rewrite–validate* procedure under semantic, linguistic,



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.

and edit-budget constraints. Human-reference alignment refers throughout to movement toward this fixed corpus-derived reference statistic, rather than to a universal measure of human-likeness. Figure 1 summarizes the framework.

We evaluate AnchorWrite on 900 AI-generated documents spanning six domains under four predefined operating regimes, L1–L4. We compare against instruction-based LLM humanization and Pegasus paraphrasing at closely matched realized edit ratios, and evaluate frozen outputs using six external AI-text detectors. The study further examines automatic quality metrics and blinded human evaluation, domain-level behavior, adaptive stopping, cross-backbone transfer, and bootstrap uncertainty where per-sample detector scores are available.

Our contributions are:

- We introduce **AnchorWrite**, a detector-query-free framework that uses an independently constructed human-text reference as the control objective for constrained post-generation rewriting.
- We develop an iterative rewriting procedure that combines cross-order anomaly localization, local LLM-based proposal generation, and length-conditioned document-level acceptance under preservation and edit constraints.
- We provide a six-domain evaluation across six external detectors, matched-edit rewriting baselines, automatic and blinded human quality assessment, domain and stopping analyses, cross-backbone transfer, and bootstrap uncertainty, demonstrating consistent optimization of the fixed reference objective together with detector-dependent external transfer and high perceived rewrite quality.

## 2 Related Work

### 2.1 AI-generated text detection

AI-generated text detection includes statistical, analytical, perturbation-based, and supervised methods. Token probability and rank signals underpin GLTR and related zero-shot approaches (Gehrmann et al., 2019; Solaiman et al., 2019; Su et al., 2023); DetectGPT and Fast-DetectGPT use probability-curvature or discrepancy signals (Mitchell et al., 2023; Bao et al., 2024), while Binoculars and LastDE exploit model-distribution and token-sequence information (Hans et al., 2024; Xu et al., 2025). Supervised classifiers form another major family but remain sensitive to distribution shift (Zellers et al., 2019; Uchendu et al., 2020). Recent work targets machine-revised text through stylistic-preference alignment (Chen et al., 2025; Dai et al., 2026), multiscale detection (Zhou et al., 2026), creator–editor modeling (Li et al., 2026), and multilingual/refinement robustness (Wu et al., 2026).

### 2.2 Paraphrasing and Humanization

Paraphrasing is a well-established challenge for AI-text detection (Krishna et al., 2023; Sadasivan et al., 2023). PEGASUS, BART, and T5 provide standard rewriting mechanisms (Zhang et al., 2020; Lewis et al., 2020; Raffel et al., 2020). Recent work explores preference-based style shifting (Pedrotti et al., 2025), multi-stage black-box humanization in MASH (Gu et al., 2026), and detection of LLM-paraphrased text in SearchLLM (Nguyen-Son et al., 2026). AnchorWrite differs from detector-guided approaches by using no detector feedback for localization or rewrite selection.

### 2.3 Human-reference statistical modeling

Count-based language models provide transparent lexical and phrase-level statistics (Chen & Goodman, 1999; Jurafsky & Martin, 2023; Heafield, 2011), while human-reference signals support stylometry and authorship analysis (Stamatatos, 2009; Potthast et al., 2018). Authorship-obfuscation methods similarly use explicit stylometric signals or interpretable style perturbations (Bevendorff et al., 2019; Fisher et al., 2024), and revised-text detectors model stylistic preferences through machine-style alignment in ImBD (Chen et al., 2025) and human-language preference alignment in HLPD (Dai et al., 2026). AnchorWrite does not claim n-gram modeling itself as novel; it uses cross-order human-reference statistics for token localization and a length-calibrated document-level objective for constrained rewriting without external-detector feedback.

# 3 ANCHORWRITE: HUMAN-REFERENCE GUIDED REWRITING

## 3.1 PROBLEM FORMULATION AND DETECTOR-ACCESS SETTING

Let $x = (w_1, \ldots, w_L)$ be an LLM-generated document of $L$ whitespace-delimited words. AnchorWrite seeks a local rewrite $x'$ that reduces distance to a fixed statistical reference derived from human-authored text while preserving meaning under a bounded edit budget. No external AI-text detector feedback is available during rewriting. We formulate

$$x^\star = \arg\min_{x' \in \mathcal{R}(x)} D(x') \quad \text{s.t.} \quad r(x, x') \le b, \qquad Q(x, x') = 1, \tag{1}$$

where $\mathcal{R}(x)$ is the set of reachable local rewrites, $D(\cdot)$ is the length-conditioned reference distance, $r(\cdot, \cdot)$ is realized edit ratio, $b$ is the edit ceiling, and $Q$ denotes the frozen preservation constraints. Thus $D$, rather than a downstream-detector score, provides the corpus-relative optimization signal.

## 3.2 CROSS-ORDER HUMAN-REFERENCE STATISTICS

For an order-$m$ human-reference language model,

$$H_m(x) = -\frac{1}{L} \log_{10} p_m(w_{1:L}, \langle \text{eos} \rangle \mid \langle \text{bos} \rangle). \tag{2}$$

Our implementation uses fixed word-level 2-gram and 4-gram KenLM models and the normalized cross-order statistic

$$A_{2,4}(x) = \frac{H_2(x) - H_4(x)}{|H_2(x)| + \epsilon}, \qquad \epsilon = 10^{-12}. \tag{3}$$

For localization, token-level surprisal and cross-order contribution are

$$\ell_{m,i} = -\log_{10} p_m(e_i \mid h_i^{(m)}), \qquad d_i^{(2,4)} = \ell_{2,i} - \ell_{4,i}, \tag{4}$$

with EOS included in scoring but never used as a rewrite target.

A fixed UMBC calibration table provides a length-conditioned center $\widehat{\mu}(L)$ and scale $\widehat{\sigma}(L)$ for $A_{2,4}$:

$$z_{2,4}(x) = \frac{A_{2,4}(x) - \widehat{\mu}(L)}{\widehat{\sigma}(L) + \epsilon}, \qquad D(x) = |z_{2,4}(x)|. \tag{5}$$

Thus, $D(x)$ is the absolute standardized displacement from the fixed reference center, while the sign of $z_{2,4}$ records its side. Appendix B gives the exact scoring conventions, document–token decomposition, UMBC reference construction, and length-conditioned calibration.

## 3.3 CROSS-ORDER ANOMALY LOCALIZATION

At each rewriting round, the current document is rescored and eligible token positions are ranked by robust cross-order anomaly:

$$s_i(x) = \frac{\left| d_i^{(2,4)} - \text{median}\big(d^{(2,4)}\big) \right|}{\text{RobustScale}\big(d^{(2,4)}\big)} \, w_{\text{POS}}(i). \tag{6}$$

Here $w_{\text{POS}}(i)$ is a fixed editability weight. Higher $s_i$ identifies editable positions whose cross-order behavior is atypical relative to the current document. Localization determines only *where* rewriting is attempted; movement toward the reference is determined by candidate acceptance through $D$. Exact robust scaling, eligibility rules, POS weights, ranking tie-breaks, and sentence mapping are given in Appendix C.1.

## 3.4 CONSTRAINED LOCAL REWRITING

For each selected sentence, the rewriting model receives the preceding, target, and following sentences and rewrites only the target. A candidate $\widetilde{x}$ must satisfy the current edit ceiling and the frozen semantic, linguistic, and structural preservation constraints, and must improve the reference objective:

$$\Delta_{\text{human}} = D(x_{\text{current}}) - D(\widetilde{x}) > 0. \tag{7}$$

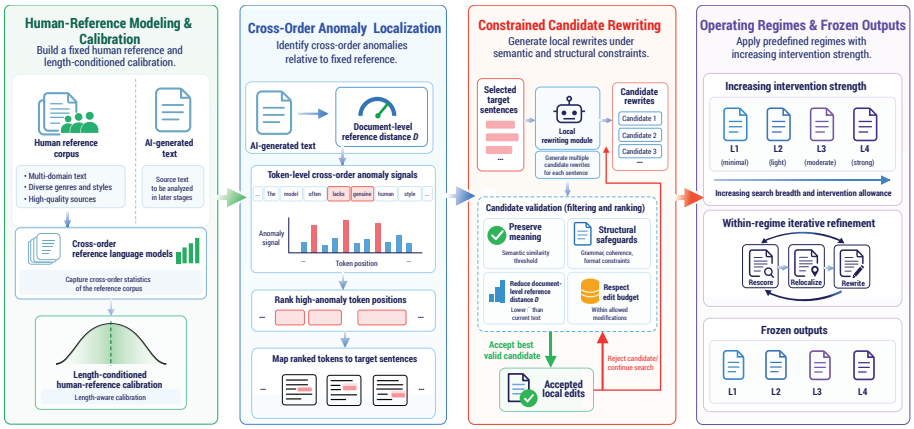



Figure 1: Overview of AnchorWrite. Cross-order statistics from a fixed human-derived reference localize anomalous regions, a pretrained language model proposes constrained rewrites, and document-level reference distance governs candidate acceptance. L1–L4 define progressively stronger operating regimes; external detectors are evaluated only on frozen outputs.

Among valid proposals, AnchorWrite retains the candidate with the largest $\Delta_{\text{human}}$, then rescores and relocalizes the accepted document. Appendix C specifies the exact prompt, proposal decoding, candidate filtering, tie-breaking, and full rewriting loop; Appendix D lists the regime-specific preservation thresholds.

### 3.5 Edit Accounting and Operating Regimes

Configured edit budgets are ceilings rather than prescribed rewrite rates. Normalized whitespace-token sequences are aligned with a sequence matcher; for each non-equal block containing $a_o$ source and $b_o$ rewritten words,

$$C(x, x') = \sum_{o \in \mathcal{O}_{\neq}} \max(a_o, b_o), \qquad r(x, x') = \frac{C(x, x')}{\max(1, L)}. \tag{8}$$

Nested stages carry the accepted text forward and stop at the current ceiling or round limit, or when no valid reference-improving candidate is found among the sampled proposals. Realized edit ratio is therefore the intervention measure.

AnchorWrite defines four fixed operating regimes, L1–L4, with progressively broader search, larger edit ceilings, and relaxed preservation thresholds. Because several factors vary jointly, these are operating regimes rather than single-factor ablations. Appendix D gives the complete configurations.

## 4 Experimental Setup

### 4.1 Six-domain benchmark

We evaluate 900 LLM-generated English documents from six domains, with 150 documents per domain. Social Sciences, STEM, and Humanities contain academic-style Moonshot generations, while Essay, Reuters, and WritingPrompts (WP) use ChatGPT-generated samples from MGTBench (He et al., 2024). Detector evaluation uses 900 human controls (150 per domain). MGTBench domains use exact human counterparts where available; academic-domain controls are matched by domain, source type, and nearest available length rather than treated as paired source documents. These controls are distinct from the UMBC corpus used to construct the human reference. Appendix A provides full provenance and pairing conventions.

### 4.2 HUMAN REFERENCE AND CALIBRATION

The fixed reference is constructed from the UMBC WebBase corpus (Han et al., 2013) using word-level 2-gram and 4-gram KenLM models (Heafield, 2011) with modified Kneser–Ney smoothing (Chen & Goodman, 1999). A 10,000-text UMBC calibration table defines the length-conditioned reference and contains no AI-generated examples or evaluation-benchmark documents. Appendix B gives the complete preprocessing, model-construction, scoring, and calibration procedure.

### 4.3 REWRITING BACKBONE AND OPERATING REGIMES

Qwen3-8B (Yang et al., 2025) is the primary rewriting backbone. Thinking output is disabled and residual thinking tags are removed before candidate parsing. One frozen output is retained for every document under each L1–L4 regime. Because the regimes jointly vary search breadth, decoding, edit ceilings, and preservation constraints, comparisons use realized edit ratio as the intervention measure. Exact configurations are reported in Appendix D.

A supporting cross-backbone study evaluates Llama-3.2-3B-Instruct, Mistral-7B-Instruct-v0.3, and Phi-4-mini-instruct on a fixed 300-document subset (50 per domain) under a separate 60%-budget stress endpoint. It is therefore treated as transfer evidence rather than a direct numerical comparison with the main Qwen3-8B L1–L4 study; full details are provided in Appendix K.

### 4.4 EXTERNAL DETECTOR SUITE

We evaluate six external AI-text detection methods: Fast-DetectGPT (FastDetect) (Bao et al., 2024), LastDE++ (Xu et al., 2025), Binoculars (Hans et al., 2024), `roberta-base-openai-detector`, likelihood-to-log-rank ratio (LRR), and LogRank (Su et al., 2023). LRR and LogRank use GPT-Neo-2.7B as the scoring language model. Detector-score orientation is fixed globally so that larger scores indicate stronger AI evidence and is never selected separately by condition. Additional statistics and operating-point results are reported in Appendix F.

### 4.5 MATCHED-EDIT BASELINES

The main comparison uses two matched-edit baselines. **Qwen Humanize** uses the same Qwen3-8B backbone with an explicit humanization instruction, while **Pegasus** (Zhang et al., 2020) provides a conventional neural paraphrasing baseline. For each source and L1–L4 operating point, the baselines target AnchorWrite's realized edit ratio; because discrete transformations cannot always reproduce it exactly, the achieved ratio is retained and reported. The comparison therefore matches realized intervention magnitude rather than proposal count or decoding compute. Appendix E gives the complete baseline definitions.

Additional Qwen Rewrite and English→Chinese→English back-translation controls are reported in Appendix H. No baseline receives AnchorWrite's reference scores, localization signals, or candidate-selection information.

### 4.6 QUALITY AND PRESERVATION METRICS

Semantic preservation is measured primarily with raw BERTScore F1, and fluency with GPT-Neo-2.7B per-document perplexity. We additionally report TTR, Distinct-2, average document length, and CoLA acceptability. BERTScore and PPL are independent post-hoc measurements, whereas CoLA and bidirectional NLI participate in candidate filtering and are therefore treated as constraint-compliance diagnostics. Appendix G reports complete quality results and evaluator settings.

### 4.7 HUMAN EVALUATION

We conduct a blinded human evaluation at L4 using the same fixed 50-document sample for AnchorWrite, Qwen Humanize, and Pegasus. Six native-English-speaking graduate-student annotators rate semantic preservation, factual/content consistency, fluency, and naturalness on five-point scales. Each

Table 1: Alignment to the fixed UMBC-derived reference at the final operating point of each AnchorWrite regime. $\Delta D = D_{\text{before}} - D_{\text{after}}$; higher values indicate greater reduction in reference distance. CIs are 95% bootstrap intervals over 900 documents.

| Regime | Edit (%) | Final $D$ | $\Delta D$ [95% CI] | Improved (%) |
|---|---|---|---|---|
| L1 | 5.60 | 5.745 | 0.068 [0.063, 0.073] | 76.1 |
| L2 | 11.53 | 5.666 | 0.147 [0.140, 0.155] | 93.6 |
| L3 | 20.40 | 5.524 | 0.289 [0.277, 0.301] | 98.9 |
| L4 | 24.32 | 5.450 | 0.363 [0.350, 0.376] | 99.3 |

of the 150 source–method pairs is rated by three annotators, yielding 450 judgments under a balanced incomplete assignment.

Inference uses the 50 paired source-level means per method and dimension. We report 95% confidence intervals from 10,000 source-level bootstrap resamples and test method differences using the Friedman test followed by two-sided paired Wilcoxon signed-rank tests with Holm correction. Appendix L provides the sampling procedure, blinding and assignment design, rating rubric, descriptive statistics, statistical tests, and agreement analysis.

### 4.8 Evaluation and uncertainty protocol

Original–rewrite comparisons are aligned by $(\text{domain}, \text{canonical sample id})$, and every main detector condition is evaluated against the same 900 human controls. All rewritten outputs are frozen before external-detector evaluation.

For FastDetect, RoBERTa, LogRank, and LRR, retained per-sample scores permit paired bootstrap analysis with 2,000 replicates. Aligned Original–transformed AI samples are resampled jointly and the human-control pool independently; canonical AUROC reproduction is verified before intervals are accepted. LastDE++ and Binoculars are reported with frozen point estimates because equivalent per-sample scores were not retained. The intervals quantify document-sampling uncertainty for one frozen rewrite per document. Appendix J gives the complete protocol and confidence intervals.

## 5 Results

### 5.1 Human-reference alignment

We first evaluate whether AnchorWrite attains its fixed-reference objective. Across the 900 documents, the mean initial reference distance is 5.813 and decreases to 5.745, 5.666, 5.524, and 5.450 for L1–L4, corresponding to mean reductions of 0.068, 0.147, 0.289, and 0.363 (Table 1). The reduction increases with realized intervention, and the direction is consistent across domains: at L4, mean $\Delta D$ ranges from 0.274 for Humanities to 0.484 for Essay, with positive $\Delta D$ for 98–100% of documents in every domain (Figure 2; Appendix I).

Because positive reference-distance reduction is an acceptance condition, these results primarily verify attainment of AnchorWrite's internal fixed-reference objective rather than external transfer. A separate audit using the same 900 held-out human controls used for detector evaluation shows that they have a larger mean $D$ than the Original AI documents under the UMBC-derived calibration, with the same ordering across all six domains. This indicates reference-corpus shift: $D$ measures proximity to the fixed UMBC-derived statistic rather than to the empirical center of arbitrary human text. Appendix B.4 reports the aggregate, domain-level, and scoring-compatibility analyses.

### 5.2 External detector transfer

Table 2 reports AUROC for six external AI-text detection methods, where lower values indicate reduced separability between human controls and rewritten AI text. For AnchorWrite, AUROC decreases monotonically from L1 to L4 across all six methods. At L4, AUROC decreases from 0.9677 to 0.9036 for FastDetect, 0.9577 to 0.8936 for LastDE++, 0.8958 to 0.8388 for Binoculars, 0.7980 to

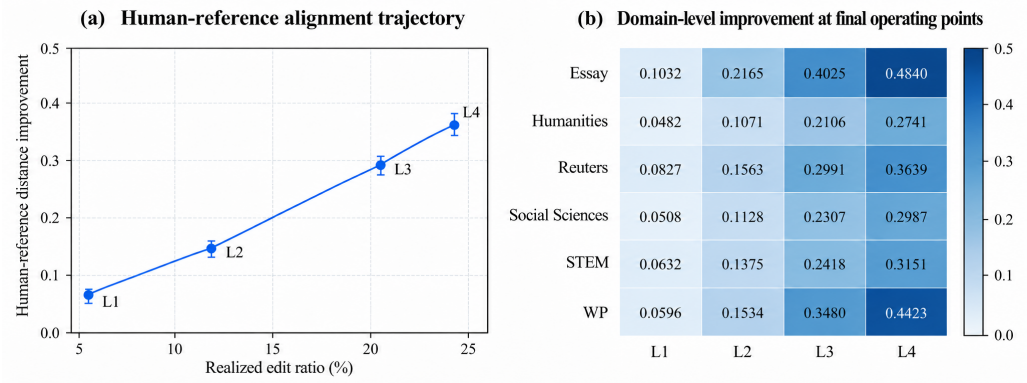



Figure 2: Alignment to the fixed UMBC-derived reference across AnchorWrite regimes. Panel (a) shows mean reference-distance reduction versus realized edit ratio, with 95% confidence intervals. Panel (b) reports the corresponding mean reduction for each of the six domains.

Table 2: Matched-edit comparison across L1–L4. Detector columns report AUROC (↓), Edit is realized word-edit ratio, and BERT is raw BERTScore F1 (↑). Among the three rewritten methods within each regime, the best result is shown in **bold** and the second-best is dotted-underlined. Original is excluded from ranking.

| Regime | Method | FastD | LastDE++ | Binoc. | RoBERTa | LRR | LogRank | Edit (%) | BERT ↑ |
|---|---|---|---|---|---|---|---|---|---|
| – | Original | 0.9677 | 0.9577 | 0.8958 | 0.7980 | 0.8101 | 0.8009 | 0.00 | 1.0000 |
| L1 | Qwen Humanize | 0.9452 | 0.9340 | 0.8923 | **0.7624** | 0.7815 | 0.7709 | 5.82 | 0.9815 |
| | Pegasus | 0.9473 | 0.9339 | 0.9017 | 0.7736 | 0.7803 | **0.7679** | 6.04 | 0.9796 |
| | Ours | **0.9446** | **0.9325** | **0.8886** | 0.7644 | **0.7796** | 0.7719 | 5.60 | **0.9852** |
| L2 | Qwen Humanize | 0.9349 | 0.9251 | 0.8874 | 0.7492 | 0.7670 | 0.7589 | 11.63 | 0.9773 |
| | Pegasus | 0.9392 | 0.9208 | 0.9006 | 0.7846 | 0.7681 | 0.7535 | 11.57 | 0.9735 |
| | Ours | **0.9290** | **0.9138** | **0.8686** | **0.7308** | **0.7533** | **0.7505** | 11.53 | **0.9780** |
| L3 | Qwen Humanize | 0.9199 | 0.9092 | 0.8753 | 0.7331 | 0.7415 | 0.7394 | 20.22 | **0.9712** |
| | Pegasus | 0.9302 | 0.9041 | 0.9035 | 0.7910 | 0.7418 | **0.7287** | 20.18 | 0.9644 |
| | Ours | **0.9126** | **0.9000** | **0.8457** | **0.7230** | **0.7303** | 0.7317 | 20.40 | 0.9709 |
| L4 | Qwen Humanize | 0.9103 | 0.8997 | 0.8672 | 0.7259 | 0.7303 | 0.7301 | 24.05 | **0.9685** |
| | Pegasus | 0.9244 | 0.8969 | 0.9046 | 0.8017 | 0.7319 | **0.7180** | 23.89 | 0.9603 |
| | Ours | **0.9036** | **0.8936** | **0.8388** | **0.7192** | **0.7214** | 0.7242 | 24.32 | 0.9678 |

0.7192 for RoBERTa, 0.8101 to 0.7214 for LRR, and 0.8009 to 0.7242 for LogRank, corresponding to absolute reductions of 0.057–0.089.

For the four detectors with retained per-sample scores, the L4 Original-to-AnchorWrite reductions are supported by paired bootstrap intervals: ΔAUROC is 0.0641 [0.0552, 0.0729] for FastDetect, 0.0788 [0.0650, 0.0924] for RoBERTa, 0.0767 [0.0703, 0.0831] for LogRank, and 0.0887 [0.0807, 0.0973] for LRR. Appendix J reports the complete L1–L4 intervals. LastDE++ and Binoculars are reported as frozen point estimates because equivalent per-sample scores were not retained.

### 5.3 Quality Preservation

Table 3 summarizes the quality trajectory of AnchorWrite. BERTScore remains high as realized edit ratio increases, decreasing from 0.9852 at L1 to 0.9678 at L4. Perplexity increases from 8.88 on Original text to 9.62, 10.12, 10.54, and 10.71 across L1–L4, showing a measurable fluency cost with stronger intervention. CoLA remains near 0.85, TTR increases modestly, and Distinct-2 remains close to 0.60.

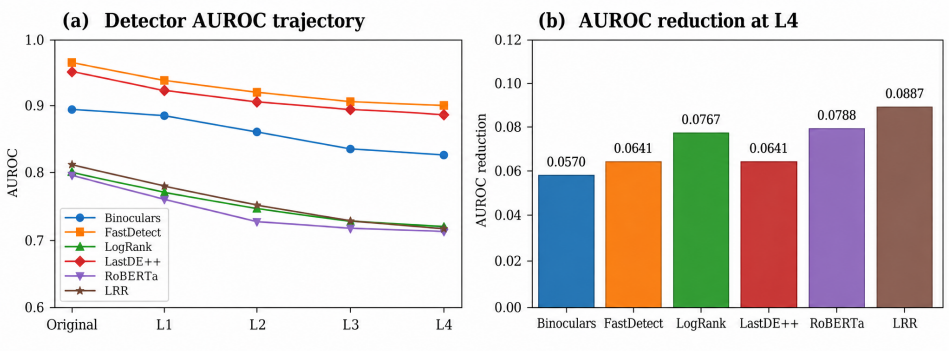



Figure 3: External detector transfer across AnchorWrite regimes. Panel (a) shows AUROC from Original to L1–L4 for six external AI-text detection methods. Panel (b) reports the absolute AUROC reduction from Original to L4. All detector evaluations are post-hoc on frozen outputs.

Table 3: Quality trajectory of AnchorWrite. BERTScore and PPL provide the primary preservation and fluency measurements; CoLA is a gate-linked diagnostic.

| Condition | Edit (%) | BERT ↑ | PPL ↓ | CoLA | TTR | Dist.-2 |
|---|---|---|---|---|---|---|
| Original | 0.00 | 1.0000 | 8.88 | 0.8452 | 0.5543 | 0.5962 |
| L1 | 5.60 | 0.9852 | 9.62 | 0.8536 | 0.5599 | 0.6009 |
| L2 | 11.53 | 0.9780 | 10.12 | 0.8551 | 0.5653 | 0.6036 |
| L3 | 20.40 | 0.9709 | 10.54 | 0.8568 | 0.5704 | 0.6025 |
| L4 | 24.32 | 0.9678 | 10.71 | 0.8558 | 0.5711 | 0.6002 |

Average document length changes from 288.5 words for Original to 297.5 at L4 (approximately 3.1%), while the broader lexical metrics remain comparatively stable (Appendix G). These results show a preservation–intervention trade-off: stronger reference optimization requires more rewriting and incurs higher PPL, while semantic similarity and lexical-diversity measures remain comparatively stable.

### 5.4 Human Evaluation

Table 4 reports the blinded L4 evaluation on the fixed 50-document subset. All methods receive high ratings: AnchorWrite scores 4.84, 4.89, 4.87, and 4.87 for semantic preservation, factual consistency, fluency, and naturalness, respectively; Qwen Humanize closely matches semantic preservation (4.85), while Pegasus scores 4.77–4.79.

Paired tests show no overall difference for semantic preservation ($p = 0.224$) or fluency ($p = 0.166$). Factual consistency differs overall ($p = 0.0209$), with AnchorWrite exceeding Pegasus after Holm correction ($p_{\text{Holm}} = 0.0267$), while AnchorWrite–Qwen Humanize is not significant ($p_{\text{Holm}} = 0.190$). Naturalness shows an omnibus difference ($p = 0.0461$), but no pairwise comparison survives correction. Given the strong upper-end concentration of ratings, these results support high perceived rewrite quality rather than large between-method differences; full statistics and agreement analysis are in Appendix L.

### 5.5 Matched-edit baseline comparison

The baselines closely match AnchorWrite's realized intervention: at L4, edit ratios are 24.32%, 24.05%, and 23.89% for AnchorWrite, Qwen Humanize, and Pegasus, respectively. AnchorWrite yields lower AUROC than Qwen Humanize on all six detectors and lower AUROC than Pegasus on five, while Pegasus is slightly lower on LogRank (0.7180 versus 0.7242). These comparisons match realized intervention rather than decoding compute; Table 2 reports the matched-edit point estimates, Appendix J reports available Original-to-AnchorWrite bootstrap uncertainty, and Appendix H provides additional transformation controls.

Table 4: Blinded human evaluation at L4. Values are source-level means [95% bootstrap CI] on five-point scales (↑). Bold marks the highest descriptive mean only.

| Method | Semantic | Factual | Fluency | Naturalness |
|---|---|---|---|---|
| AnchorWrite | 4.84 [4.78, 4.89] | **4.89 [4.84, 4.94]** | **4.87 [4.81, 4.92]** | **4.87 [4.81, 4.93]** |
| Qwen Humanize | **4.85 [4.79, 4.90]** | 4.83 [4.77, 4.88] | 4.79 [4.71, 4.85] | 4.81 [4.73, 4.87] |
| Pegasus | 4.79 [4.73, 4.85] | 4.78 [4.72, 4.84] | 4.77 [4.71, 4.84] | 4.79 [4.73, 4.84] |

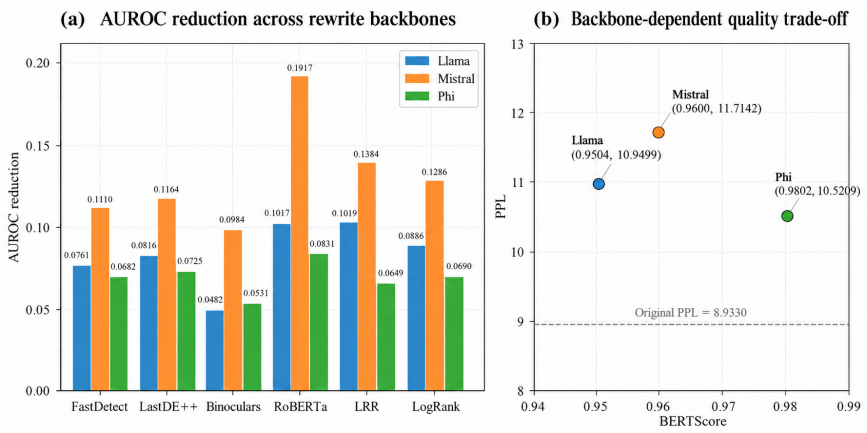



Figure 4: Cross-backbone transfer on the fixed 300-document subset. (a) AUROC reduction for six detectors under three rewriting backbones. (b) Corresponding BERTScore–PPL trade-off. Full values are reported in Appendix K.

### 5.6 Domain consistency and adaptive stopping

Mean $\Delta D$ increases from L1 to L4 in every domain; at L4 it ranges from 0.274 for Humanities to 0.484 for Essay, with WP at 0.442, Reuters at 0.364, STEM at 0.315, and Social Sciences at 0.299. Only 1.56%, 5.89%, 0.89%, and 0.33% of samples terminate by reaching the edit ceiling for L1–L4, respectively; most instead stop when no valid sampled candidate further improves the reference objective while satisfying the preservation constraints. Thus, the ceilings primarily act as upper bounds on intervention. Full results are provided in Appendix I.

### 5.7 Cross-backbone transfer

We evaluate the same fixed 300 documents with Llama-3.2-3B-Instruct, Mistral-7B-Instruct-v0.3, and Phi-4-mini-instruct under the separate 60%-budget stress endpoint. Figure 4 shows AUROC reductions across all six detectors for every backbone, ranging from 0.0482–0.1019 for Llama, 0.0984–0.1917 for Mistral, and 0.0531–0.0831 for Phi. Mean BERTScore remains at least 0.9504; Phi retains the highest BERTScore (0.9802) and lowest rewritten PPL (10.5209), whereas Mistral yields the largest detector reductions but highest PPL (11.7142). Thus, transfer extends beyond Qwen3-8B with a backbone-dependent detectability–quality trade-off; full results are in Appendix K.

## 6 Conclusion

We presented **AnchorWrite**, a detector-query-free framework that uses cross-order human-reference statistics to localize and accept constrained LLM rewrites. Across six domains, reference-distance reduction increases from 0.068 to 0.363 over 5.60–24.32% realized editing, while frozen outputs yield AUROC reductions of 0.057–0.089 across six external detectors. At L4, BERTScore remains 0.968 and blinded human ratings remain high (4.84–4.89/5), with PPL increasing from 8.88 to 10.71. Matched-edit and cross-backbone evaluations further show transformation- and backbone-dependent transfer. Overall, AnchorWrite demonstrates that a fixed corpus-derived reference can provide an interpretable rewriting objective without downstream-detector feedback.

### AI USE STATEMENT

Generative AI tools were used to provide feedback on research methodology and experimental design, assist with implementation and debugging, support analysis and interpretation of experimental outputs, and assist with drafting, editing, and structuring the manuscript and scientific visualizations. All experimental configurations, generated code, analyses, claims, and manuscript text were reviewed by the authors, who take responsibility for the final content.

### ETHICS STATEMENT

This work studies controlled rewriting of LLM-generated text and therefore has dual-use implications: the same techniques used to evaluate detector robustness may also reduce the detectability of machine-generated content. Our experiments are designed to characterize this limitation under controlled research conditions rather than to support deceptive deployment. We explicitly separate rewrite guidance from external-detector feedback, report quality costs and detector disagreement, and discuss limitations relevant to provenance systems.

## REPRODUCIBILITY STATEMENT

The manuscript and appendices specify the human-reference construction, preprocessing, Anchor-Write rewriting procedure, operating regimes, detector score orientation, quality metrics, matched-edit baselines, and statistical protocol. The implementation code, experiment configurations, and processed evaluation data used in this study will be released upon publication.

## REFERENCES

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In *International Conference on Learning Representations*, volume 2024, pp. 24814–24836, 2024.

Janek Bevendorff, Martin Potthast, Matthias Hagen, and Benno Stein. Heuristic authorship obfuscation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1098–1108, 2019.

Jiaqi Chen, Xiaoye Zhu, Tianyang Liu, Ying Chen, Chen Xinhui, Yiwen Yuan, Chak Tou Leong, Zuchao Li, Long Tang, Lei Zhang, et al. Imitate before detect: Aligning machine stylistic preference for machine-revised text detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 23559–23567, 2025.

Stanley F Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. *Computer speech & language*, 13(4):359–394, 1999.

Fangqi Dai, Xingjian Jiang, and Zizhuang Deng. Hlpd: Aligning llms to human language preference for machine-revised text detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 30440–30448, 2026.

Jillian Fisher, Skyler Hallinan, Ximing Lu, Mitchell L Gordon, Zaid Harchaoui, and Yejin Choi. Styleremix: Interpretable authorship obfuscation via distillation and perturbation of style elements. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4172–4206, 2024.

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: system demonstrations*, pp. 111–116, 2019.

Yongtong Gu, Songze Li, and Xia Hu. Mash: Evading black-box ai-generated text detectors via style humanization. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 29749–29769, 2026.

Lushan Han, Abhay L Kashyap, Tim Finin, James Mayfield, and Jonathan Weese. Umbc_ebiquity-core: Semantic textual similarity systems. In *Second joint conference on lexical and computational semantics (* SEM), volume 1: Proceedings of the main conference and the shared task: Semantic textual similarity*, pp. 44–52, 2013.

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*, pp. 17519–17537, 2024.

Xinlei He, Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. Mgtbench: Benchmarking machine-generated text detection. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pp. 2251–2265, 2024.

Kenneth Heafield. Kenlm: Faster and smaller language model queries. In *Proceedings of the sixth workshop on statistical machine translation*, pp. 187–197, 2011.

Daniel Jurafsky and James H. Martin. *Speech and Language Processing*. Prentice Hall, 3rd edition, 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. volume 36, pp. 27469–27500, 2023.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. pp. 7871–7880, 2020.

Yang Li, Qiang Sheng, Zhengjia Wang, Yehan Yang, Danding Wang, and Juan Cao. Beyond the final actor: Modeling the dual roles of creator and editor for fine-grained llm-generated text detection. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5188–5203, 2026.

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

Hoang-Quoc Nguyen-Son, Minh-Son Dao, and Koji Zettsu. Searchllm: Detecting llm paraphrased text by measuring the similarity with regeneration of the candidate source via search engine. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1755–1772, 2026.

Andrea Pedrotti, Michele Papucci, Cristiano Ciaccio, Alessio Miaschi, Giovanni Puccetti, Felice Dell'Orletta, and Andrea Esuli. Stress-testing machine generated text detection: Shifting language models writing style to fool detectors. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 3010–3031, 2025.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 231–240, 2018.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

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.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*, 2019.

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

Jinyan Su, Terry Zhuo, Di Wang, and Preslav Nakov. Detectllm: Leveraging log rank information for zero-shot detection of machine-generated text. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 12395–12412, 2023.

Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. Authorship attribution for neural text generation. pp. 8384–8395, 2020.

Junchao Wu, Yefeng Liu, Chenyu Zhu, Hao Zhang, Zeyu Wu, Tianqi Shi, Yichao Du, Longyue Wang, Weihua Luo, Jinsong Su, et al. Detectrl-x: Towards reliable multilingual and real-world llm-generated text detection. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 38247–38294, 2026.

Yihuai Xu, Yongwei Wang, Yifei Bi, Huangsen Cao, Zhouhan Lin, Yu Zhao, and Fei Wu. Training-free llm-generated text detection by mining token probability sequences. In *International Conference on Learning Representations*, volume 2025, pp. 19072–19098, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. *Advances in neural information processing systems*, 32, 2019.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International conference on machine learning*, pp. 11328–11339. PMLR, 2020.

Bing Zhou, Zhe Huang, Shilei Tan, Kai Zhao, and Zhou Yongcheng. Lamcl: A length-aware momentum contrastive learning framework for multiscale machine-revised text detection. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24366–24380, 2026.

## A Dataset and Domain Construction

The benchmark contains six domains, with 150 AI-generated documents and 150 human controls per domain. Table 5 distinguishes the AI source from the human-control construction because the academic controls are matched documents rather than paired source texts.

Table 5: Benchmark provenance. "Matched control" denotes a human document selected by domain, source type, and nearest available length; it is not assumed to be the source document from which the corresponding AI text was generated.

| Domain | AI source | Style/source | Human-control relation | $n$/class |
|---|---|---|---|---|
| Social Sciences | Moonshot | Academic | matched control, not paired | 150 |
| STEM | Moonshot | Academic | matched control, not paired | 150 |
| Humanities | Moonshot | Academic | matched control, not paired | 150 |
| Essay | ChatGPT | MGTBench Essay | exact counterpart where available | 150 |
| Reuters | ChatGPT | MGTBench Reuters | exact counterpart where available | 150 |
| WP | ChatGPT | MGTBench WritingPrompts | exact counterpart where available | 150 |

**Pairing semantics.** Every AnchorWrite output is paired with its corresponding original AI document by canonical sample identifier. This pairing is used for BERTScore, realized edit ratio, reference-distance reduction, and paired bootstrap analyses. Detector evaluation instead contrasts each AI condition against the same fixed human-control pool. Detector AUROC is therefore a class-discrimination statistic and does not require every AI document to have a paired human source.

**Separation from the human reference.** The detector human controls are distinct from the UMBC examples used to construct the 2-gram and 4-gram human-reference models and the 10,000-text calibration table. The evaluation benchmark is therefore separate from human-reference construction and calibration.

**Source-generation provenance.** For the three academic domains, the surviving dataset metadata identifies the generator as Moonshot and records sampling seed 42. The exact Moonshot model/version, original API model string, generation prompt, and decoding settings are not recoverable from the surviving records and are therefore not inferred.

**Scope.** The main benchmark is English-only and contains two AI source families across six domains. Conclusions from the main experiment are limited to the evaluated generators, domains, Qwen3-8B operating regimes, detection methods, and quality evaluators. A separate cross-backbone analysis evaluates alternative rewriting backbones on a fixed 300-document subset under a distinct endpoint protocol.

# B Human-Reference Construction and Cross-Order Statistics

## B.1 UMBC reference models

We construct the fixed human reference from the UMBC WebBase corpus (Han et al., 2013). The frozen cleaned corpus contains 12,862,924 text lines and occupies approximately 5.0 GB. Null characters and line breaks are replaced with spaces, repeated whitespace is collapsed, and leading and trailing whitespace are removed, while case, punctuation, and numbers are preserved. KenLM therefore operates on whitespace-delimited word tokens.

We train fixed word-level 2-gram and 4-gram models using KenLM commit `4cb443e` (Heafield, 2011) with modified Kneser–Ney smoothing (Chen & Goodman, 1999). No externally restricted vocabulary is imposed; unseen words use KenLM's `<unk>` and backoff behavior. The 2-gram model captures shorter-range regularity, whereas the 4-gram model incorporates broader local context. Their difference therefore measures how additional context changes predictability under the fixed human-derived reference.

The canonical build is equivalent to:

```
lmplz -o 2 < umbc_human_clean.txt > umbc_2gram.arpa
build_binary probing umbc_2gram.arpa umbc_2gram.binary
lmplz -o 4 < umbc_human_clean.txt > umbc_4gram.arpa
build_binary trie umbc_4gram.arpa umbc_4gram.binary
```

The model order, smoothing, preprocessing, and binary construction are fixed across all experiments.

## B.2 Reference language-model scores

Let $x = (w_1, \ldots, w_L)$ contain $L$ observed whitespace-delimited words. For model order $m$, define the event surprisal as

$$\ell_{m,i} = -\log_{10} p_m(e_i \mid h_i^{(m)}), \tag{9}$$

where $e_i = w_i$ for $i \leq L$ and $e_{L+1} = \langle \text{eos} \rangle$. BOS/EOS scoring is enabled, and

$$H_m(x) = \frac{1}{L} \sum_{i=1}^{L+1} \ell_{m,i}. \tag{10}$$

Thus, $L+1$ scoring events contribute to the numerator while normalization remains by the $L$ observed words.

The token-level 2–4 cross-order contribution is

$$d_i^{(2,4)} = \ell_{2,i} - \ell_{4,i}. \tag{11}$$

These contributions recover the document-level cross-order difference:

$$H_2(x) - H_4(x) = \frac{1}{L} \sum_{i=1}^{L+1} d_i^{(2,4)}. \tag{12}$$

The sign of $d_i^{(2,4)}$ gives the local cross-order direction, while its magnitude reflects disagreement between the two reference-model orders.

At the document level, we use

$$A_{2,4}(x) = \frac{H_2(x) - H_4(x)}{|H_2(x)| + 10^{-12}}. \tag{13}$$

This normalizes the raw cross-order difference by the magnitude of the lower-order score and is interpreted relative to the fixed UMBC-derived reference for texts of comparable length.

### B.3 Length-conditioned human-reference calibration

The calibration table contains $M = 10{,}000$ UMBC human texts represented by pairs $(L_j, A_j)$, where $A_j = A_{2,4}(x_j)$. For a candidate document of length $L$, the preferred calibration neighborhood is

$$\mathcal{N}_\rho(L) = \{j : \max(1, (1-\rho)L) \le L_j \le (1+\rho)L\}, \qquad \rho = 0.25. \tag{14}$$

If fewer than $k = 100$ calibration texts fall within this window, we instead select the 100 examples with the smallest log-length distance

$$|\log(1 + L_j) - \log(1 + L)|. \tag{15}$$

The length-conditioned reference center $\widehat{\mu}(L)$ is the median of the neighborhood $A_j$ values. Let $\mathrm{SD}_0$ denote their population standard deviation and let $\mathrm{IQR} = Q_{0.75} - Q_{0.25}$. The corresponding scale is

$$\widehat{\sigma}(L) = \max\left(\frac{\mathrm{IQR}}{1.349}, \mathrm{SD}_0\right), \tag{16}$$

with fallback value 1 if the resulting scale is invalid or degenerate.

The signed standardized position and reference distance are

$$z_{2,4}(x) = \frac{A_{2,4}(x) - \widehat{\mu}(L)}{\widehat{\sigma}(L) + 10^{-12}}, \qquad D(x) = |z_{2,4}(x)|. \tag{17}$$

The sign of $z_{2,4}(x)$ identifies the side of the length-conditioned reference center, whereas $D(x)$ measures its standardized displacement. AnchorWrite minimizes this fixed corpus-derived distance rather than imposing a universal increase or decrease in $A_{2,4}$.

The calibration table contains no AI-generated examples, generator identities, detector outputs, or evaluation labels.

**Calibration reuse and role separation.** The calibration is frozen and reused for every current and candidate document; only observed document length determines the corresponding calibration neighborhood. No parameter is re-estimated from AI-generated text, rewrite candidates, or detector outputs. Token-level $d_i^{(2,4)}$ values support localization, whereas $D(x)$ is the document-level candidate-acceptance objective. Consequently, changes in $D$ measure movement relative to the fixed length-conditioned UMBC reference rather than adaptation of the reference itself.

### B.4 Held-Out Human Reference Audit

To quantify reference-corpus dependence, we separately score the same 900 held-out human controls used in external detector evaluation together with the Original and AnchorWrite L1–L4 conditions. This audit is performed after the rewriting outputs are frozen and does not affect localization, candidate generation, or acceptance.

**Aggregate reference distance.** Table 6 reports the resulting reference distance $D$. The held-out human controls have mean $D = 6.1435$, compared with $D = 5.8129$ for the Original AI documents. AnchorWrite progressively reduces the fixed-reference distance from 5.7450 at L1 to 5.4499 at L4. Thus, movement toward the UMBC-derived reference center does not imply movement toward the empirical center of this held-out human-control population.

Table 6: Held-out human-reference audit. $D$ is standardized distance to the fixed length-conditioned UMBC-derived cross-order reference. Intervals are 95% document-level confidence intervals.

| Condition | $n$ | Mean $D$ | Median $D$ | 95% CI |
|---|---|---|---|---|
| Held-out Human | 900 | 6.1435 | 6.1760 | [6.1211, 6.1643] |
| Original AI | 900 | 5.8129 | 5.8713 | [5.7898, 5.8377] |
| L1 | 900 | 5.7450 | 5.8047 | [5.7183, 5.7694] |
| L2 | 900 | 5.6657 | 5.7131 | [5.6406, 5.6920] |
| L3 | 900 | 5.5242 | 5.5554 | [5.4990, 5.5480] |
| L4 | 900 | 5.4499 | 5.4868 | [5.4254, 5.4749] |

**Domain-level reference shift.** The same held-out-human versus Original-AI ordering appears in every evaluated domain (Table 7). Its magnitude varies by domain, indicating distributional shift between the fixed UMBC-derived reference and the heterogeneous held-out human-control collection rather than an effect confined to one benchmark component.

Table 7: Mean reference distance for held-out human controls and Original AI documents by domain. Larger values indicate greater standardized displacement from the fixed UMBC-derived reference.

| Domain | Held-out Human $D$ | Original AI $D$ |
|---|---|---|
| Essay | 6.1109 | 5.6631 |
| Humanities | 6.1381 | 5.9724 |
| Reuters | 6.0884 | 5.8252 |
| Social Sciences | 6.1020 | 5.7419 |
| STEM | 6.2781 | 5.9445 |
| WritingPrompts | 6.1437 | 5.7306 |

**Scoring compatibility check.** To verify that the observed shift is not caused by inconsistent scoring, we independently recompute $H_2$, $H_4$, and $A_{2,4}$ for 50 deterministically selected UMBC calibration texts. All 50 reproduce the frozen canonical values at the checked numerical precision, ruling out a discrepancy between the calibration and evaluation scoring paths.

**Interpretation.** The audit confirms that $D$ is relative to the fixed UMBC-derived cross-order reference and is sensitive to reference-corpus shift. Because the held-out human controls never enter rewriting, this result does not alter the optimization procedure or the external detector, automatic-quality, and blinded human-evaluation results.

## C Algorithmic Specification

Algorithm 1 summarizes the complete AnchorWrite rewriting loop under a fixed operating regime. Localization and candidate acceptance use distinct signals: the token-level score $s_i$ prioritizes proposal locations, whereas the document-level reference distance $D$ determines whether a valid proposal is retained.

**Algorithm 1** ANCHORWRITE: Human-reference guided constrained rewriting

**Require:** Original AI document $x_0$; operating regime $R$; human-reference models $p_2, p_4$; calibration table $\mathcal{C}$
**Ensure:** Frozen rewritten document $x^\star$

$x \leftarrow x_0$
**for** each nested edit ceiling $b$ in $R$ **do**
  $t \leftarrow 0$
  **while** $t < T_{\max}$ **and** $r(x_0, x) < b$ **do**
    $S \leftarrow \mathrm{Localize}(x; R)$ using Eq. 6
    $\mathcal{V} \leftarrow \emptyset$
    **for** each target sentence $s \in S$ **do**
      **for** each $\widetilde{x} \in \mathrm{Generate}(s; x, R)$ **do**
        **if** $r(x_0, \widetilde{x}) \le b$ **and** $\mathrm{Preserve}(x_0, x, \widetilde{x}; R)$ **then**
          $\Delta_{\mathrm{human}} \leftarrow D(x) - D(\widetilde{x})$
          **if** $\Delta_{\mathrm{human}} > 0$ **then**
            $\mathcal{V} \leftarrow \mathcal{V} \cup \{(\widetilde{x}, \Delta_{\mathrm{human}})\}$
          **end if**
        **end if**
      **end for**
    **end for**
    **if** $\mathcal{V} = \emptyset$ **then**
      **break**
    **end if**
    $x \leftarrow \mathrm{Best}(\mathcal{V})$
    $t \leftarrow t + 1$
  **end while**
**end for**
**return** $x^\star \leftarrow x$

$\mathrm{Localize}(\cdot)$ applies the frozen lexical and POS eligibility rules, ranks token positions by $s_i$, maps retained positions to distinct target sentences, and enforces the regime-specific target limits. $\mathrm{Generate}(\cdot)$ uses the frozen local rewrite prompt and decoding configuration, while $\mathrm{Preserve}(\cdot)$ collects the deterministic, semantic-similarity, bidirectional-NLI, acceptability, structural, and edit constraints. $\mathrm{Best}(\cdot)$ first maximizes $\Delta_{\mathrm{human}}$; ties are resolved by document similarity, bidirectional entailment, sentence similarity, and changed-word count, in that order. For every candidate, $D(\widetilde{x})$ is recomputed using the same fixed reference models and the length-conditioned UMBC calibration corresponding to the candidate's length. After each accepted rewrite, the document is rescored and localization is recomputed.

### C.1 LOCALIZATION AND PROPOSAL IMPLEMENTATION

**Token-level localization.** For each non-EOS token position $i$, let $d_i^{(2,4)} = \ell_{2,i} - \ell_{4,i}$ denote its cross-order contribution. AnchorWrite first computes

$$m_d = \mathrm{median}_j \, d_j^{(2,4)} \tag{18}$$

over all finite non-EOS token-level contributions in the current document, before lexical eligibility filtering. The implementation uses the median absolute deviation

$$\mathrm{MAD}(d) = \mathrm{median}_j \left| d_j^{(2,4)} - m_d \right|, \tag{19}$$

with

$$\mathrm{RobustScale}(d) = \max\left(1.4826 \ \mathrm{MAD}(d), 10^{-6}\right). \tag{20}$$

If no finite token-level contributions are available, the implementation returns center $0$ and scale $1$. The token anomaly and localization score are

$$a_i = \frac{\left| d_i^{(2,4)} - m_d \right|}{\mathrm{RobustScale}(d)}, \qquad s_i = a_i \, w_{\mathrm{POS}}(i). \tag{21}$$

Eligible positions are sorted by $s_i$ in descending order, with $a_i$ and then $|d_i^{(2,4)}|$ used as descending tie-breakers.

**POS weighting and lexical eligibility.** After computing the token-level anomaly, AnchorWrite applies fixed coarse POS weights: $w_{\mathrm{POS}}(\mathrm{ADV}) = 1.40$, $w_{\mathrm{POS}}(\mathrm{ADJ}) = 1.30$, $w_{\mathrm{POS}}(\mathrm{VERB}) = 1.15$, and $w_{\mathrm{POS}}(\mathrm{NOUN}) = 0.45$; unrecognized categories receive zero weight. The evaluated implementation uses mandatory NLTK POS tagging.

For each whitespace-token position, a lexical core is extracted using `split_token_parts` and retained only if it matches `^[A-Za-z]+(?:['’-][A-Za-z]+)*$`. Eligibility is then determined by the frozen lexical and POS filters: the lowercase core must satisfy the minimum-length and content-POS requirements, and must not belong to the fixed stopword or auxiliary sets, contain a digit, form a possessive, or trigger the probable-named-entity filter. Repeated nouns are additionally excluded when the same lowercase core occurs at least twice in the current document. Frequencies are computed before eligibility filtering, and the final LLM-budget path uses no additional excluded-position set.

At most `llm_target_count` ranked token positions are retained.

**Mapping ranked positions to sentences.** The retained token positions are traversed in localization-score order and mapped to sentence spans using `sentence_bounds`. Duplicate sentence bounds are skipped, so the highest-ranked token represents each target sentence and the ordering among distinct sentences is preserved. Collection stops after `llm_target_sentences` distinct sentences are obtained or the ranked token list is exhausted. The immediately preceding and following sentences, when available, are supplied as local context to the proposal model.

Thus, `Targets` and `Sent./round` control different stages. For L4, `Targets=40` retains at most 40 ranked token positions before sentence deduplication, whereas `Sent./round=12` retains at most 12 distinct target sentences after mapping and deduplication. The `Cand.` parameter is separate: `llm_candidates` rewrite sequences are requested for each retained target sentence through `num_return_sequences`; hence `Cand.=40` in L4 denotes 40 requested proposals per target sentence, not 40 proposals for the entire round. Candidate cleanup and validation may reduce the number of usable proposals below this requested count.

**Rewrite prompt.** The exact system instruction is:

> You are a conservative local paraphraser. Rewrite ONLY the TARGET sentence. Preserve its factual meaning, entities, numbers, polarity/negation, tense, and logical relations. Do not summarize, add facts, delete facts, or explain your answer. Return only one rewritten sentence.

The exact user-template content is:

> `PREVIOUS CONTEXT:` {`previous_sentence or '[none]'`}
> `TARGET:` {`target_sentence`}
> `NEXT CONTEXT:` {`next_sentence or '[none]'`}
>
> Rewrite the TARGET using meaning-preserving changes in wording and local syntax. Aim to change about {`desired_changed_words`} words and do not intentionally change more than {`max_changed_words`} words. Keep the sentence natural. Output only the rewritten TARGET sentence.

Let $N$ denote the original document word count, $b$ the current edit budget, $C$ the number of words already changed relative to the original document, and $f$ the regime-specific desired-change fraction. The implementation computes

$$M = \max\left(1, \lfloor bN \rfloor - C\right), \tag{22}$$

$$R = \min\left(M,\ \max\left(2, \mathrm{round}(Nf)\right)\right), \tag{23}$$

and supplies

$$\texttt{max_changed_words} = \max(2, M), \qquad \texttt{desired_changed_words} = \max(2, R). \tag{24}$$

These quantities guide proposal generation only. The hard document-level edit ceiling is enforced separately through the realized edit-ratio check, so any candidate exceeding the applicable ceiling is rejected regardless of the requested prompt-level change count.

**Proposal decoding.** When a tokenizer chat template is available, the system and user messages are rendered with `apply_chat_template(..., tokenize=False, add_generation_prompt=True)`; otherwise, they are concatenated with two newlines and the literal suffix `REWRITE:`. The resulting prompt is tokenized with truncation enabled and a maximum input length of 1536 tokens.

For each retained target sentence, candidate generation uses sampling with the regime-specified temperature and top-$p$, together with `top_k=50`, `repetition_penalty=1.05`, `num_return_sequences` equal to the regime-specific `Cand.` value, the configured maximum new-token count, and the tokenizer's pad and EOS token IDs. Generated outputs are cleaned before validation; empty outputs, case-insensitively unchanged rewrites, and case-insensitive duplicate rewrites are discarded.

## D Operating-Regime Hyperparameters

Table 8 reports the frozen L1–L4 configurations used by AnchorWrite. Nested edit ceilings bound intervention, whereas the main paper reports the realized edit ratio actually attained.

Table 8: Frozen AnchorWrite operating-regime configurations.

| Regime | Nested edit ceilings | Cand. | Targets | Sent./round | Temp. | Max rounds | Target-change frac. | $\tau_{\text{sent}}$ | $\tau_{\text{doc}}$ | CoLA min |
|---|---|---|---|---|---|---|---|---|---|---|
| L1 | 5,10,15,20% | 10 | 12 | 6 | 0.70 | 8 | 0.035 | 0.86 | 0.90 | 0.50 |
| L2 | 5,10,15,20,25% | 20 | 20 | 8 | 0.70 | 12 | 0.070 | 0.80 | 0.85 | 0.40 |
| L3 | 20,30,40,50% | 30 | 30 | 10 | 0.80 | 15 | 0.200 | 0.72 | 0.78 | 0.30 |
| L4 | 30,40,50,60% | 40 | 40 | 12 | 0.85 | 20 | 0.300 | 0.68 | 0.74 | 0.25 |

The target-change fraction controls proposal-search aggressiveness and is distinct from the document-level edit ceiling. It defines a length-scaled desired change count that is capped by the remaining edit budget before being inserted into the rewrite prompt; Appendix C.1 gives the exact calculation. The target-change fraction is therefore neither a required modification rate nor the realized edit ratio.

The bidirectional NLI constraints use minimum entailment thresholds of 0.90, 0.85, 0.75, and 0.70 for L1–L4, with contradiction ceilings of 0.10, 0.15, 0.20, and 0.25, respectively. Maximum allowed CoLA drops are 0.25, 0.35, 0.40, and 0.45, while local-context similarity uses $\max(0.80, \tau_{\text{sent}} - 0.03)$. Top-$p$ is fixed at 0.95 and generation is capped at 256 new tokens across all regimes. Budget tolerances are 0.0125, 0.02, 0.03, and 0.04 for L1–L4, respectively. All settings are fixed before external-detector evaluation.

## E Baseline Definitions and Matching

All baselines are generated independently of the external detectors used for evaluation. They receive neither detector feedback nor AnchorWrite's human-reference statistics, localization signals, or candidate-selection information.

**Qwen Humanize.** Qwen Humanize uses the same Qwen3-8B rewriting backbone as AnchorWrite, but with an explicit instruction to make the text read more naturally and human-like while preserving its original content. It receives no human-reference scores, reference-distance improvements, localization signals, or detector feedback.

**Pegasus.** PEGASUS (Zhang et al., 2020) provides a conventional neural paraphrasing baseline. Sentence-level paraphrases are accumulated toward the matched realized edit ratio until the intervention target is reached or no additional valid transformation can be applied.

**Qwen Rewrite (appendix control).** Qwen Rewrite uses the same Qwen3-8B backbone with a generic rewriting instruction, without explicit humanization guidance or access to AnchorWrite's human-reference objective. This control characterizes the effect of general LLM rewriting and is therefore reported in the appendix rather than the main matched-edit comparison.

**Back-translation (appendix control).** English text is translated to Chinese and back to English using the frozen MarianMT `opus-mt-en-zh` and `opus-mt-zh-en` models with two-beam decoding and early stopping. This provides a non-LLM control based on round-trip translation.

**Matched intervention.** For each source document and L1–L4 operating point, the corresponding baseline targets the realized edit ratio produced by AnchorWrite on that same document. Because discrete sentence-level transformations cannot always reproduce the target exactly, the achieved edit ratio is retained and reported rather than replaced by the nominal target. All baseline outputs are frozen before evaluation and are processed using the same external-detector and post-hoc quality pipelines as the corresponding AnchorWrite outputs.

# F Additional Detector Results

## F.1 Direct zero-shot statistics

Table 9 reports additional GPT-Neo-2.7B zero-shot statistics for AnchorWrite. These are omitted from the main comparison to avoid overweighting multiple correlated statistics derived from the same scoring language model.

Table 9: Additional zero-shot AUROC (↓) for Original and AnchorWrite.

| Condition | LogRank | LogLikelihood | Entropy |
|---|---|---|---|
| Original | 0.8009 | 0.7959 | 0.6042 |
| L1 | 0.7719 | 0.7685 | 0.5823 |
| L2 | 0.7505 | 0.7486 | 0.5683 |
| L3 | 0.7317 | 0.7306 | 0.5622 |
| L4 | 0.7242 | 0.7235 | 0.5610 |

All three statistics decrease progressively from Original to L4. This provides additional evidence that the external transfer observed for AnchorWrite is not confined to a single likelihood- or rank-based statistic.

## F.2 Operating-point metrics at L4

For LastDE++, Binoculars, and LRR, the frozen detector outputs also retain AUPRC and TPR at fixed false-positive rates. Table 10 reports these metrics at L4. Under the globally fixed score orientation, lower values indicate reduced separation of rewritten AI text from the human-control condition.

The operating-point results broadly agree with the AUROC analysis while preserving detector-specific differences. At L4, AnchorWrite yields the lowest AUPRC among the rewritten methods for all three reported detector families and the lowest fixed-FPR TPRs for Binoculars and LRR. For LastDE++, differences among the rewritten conditions are smaller and vary with the selected FPR operating point. These results therefore support broad external transfer without implying uniform behavior across detector families.

**Score orientation.** All detector statistics use a single globally fixed orientation in which larger scores indicate stronger AI evidence. For FastDetect, RoBERTa, LogRank, and LRR, AUROC recomputed from the retained raw scores reproduces the corresponding frozen canonical value within the preset numerical tolerance before any uncertainty estimate is accepted.

# G Complete Quality and Intervention Results

Table 11 reports the complete quality and intervention trajectory of AnchorWrite. BERTScore and PPL serve as the primary measures of semantic preservation and fluency, respectively. CoLA is a gate-linked diagnostic, while TTR, MTLD, Distinct-2, and average document length are reported descriptively.

Table 10: L4 AUPRC and TPR at fixed FPRs for detector families with retained operating-point outputs. Lower values indicate reduced detector separability on rewritten text.

| Detector | Method | AUPRC | TPR@1%FPR | TPR@5%FPR | TPR@10%FPR |
|---|---|---|---|---|---|
| LastDE++ | Original | 0.9566 | 0.6233 | 0.8533 | 0.8956 |
| | Qwen Humanize | 0.9028 | 0.3867 | 0.6567 | 0.7389 |
| | Pegasus | 0.9015 | 0.4111 | 0.6478 | 0.7256 |
| | Ours | 0.8980 | 0.3833 | 0.6533 | 0.7278 |
| Binoculars | Original | 0.8908 | 0.2811 | 0.5389 | 0.7133 |
| | Qwen Humanize | 0.8588 | 0.2133 | 0.4556 | 0.6556 |
| | Pegasus | 0.8957 | 0.2744 | 0.5456 | 0.7533 |
| | Ours | 0.8325 | 0.1589 | 0.3967 | 0.5744 |
| LRR | Original | 0.7979 | 0.0522 | 0.3622 | 0.5344 |
| | Qwen Humanize | 0.6988 | 0.0089 | 0.1767 | 0.3633 |
| | Pegasus | 0.7100 | 0.0089 | 0.2000 | 0.3922 |
| | Ours | 0.6895 | 0.0056 | 0.1567 | 0.3511 |

Table 11: Quality and intervention trajectory for Original and AnchorWrite.

| Condition | BERT ↑ | PPL ↓ | CoLA ↑ | TTR | MTLD | Dist.-2 | Avg. length | Edit ratio |
|---|---|---|---|---|---|---|---|---|
| Original | 1.0000 | 8.8813 | 0.8452 | 0.5543 | 99.6387 | 0.5962 | 288.4744 | 0.0000 |
| L1 | 0.9852 | 9.6204 | 0.8536 | 0.5599 | 102.4284 | 0.6009 | 288.8422 | 0.0560 |
| L2 | 0.9780 | 10.1185 | 0.8551 | 0.5653 | 105.7702 | 0.6036 | 290.4933 | 0.1153 |
| L3 | 0.9709 | 10.5399 | 0.8568 | 0.5704 | 109.9482 | 0.6025 | 294.8067 | 0.2040 |
| L4 | 0.9678 | 10.7066 | 0.8558 | 0.5711 | 109.9561 | 0.6002 | 297.5089 | 0.2432 |

Across L1–L4, BERTScore remains high while PPL increases with intervention strength. Lexical-diversity measures vary modestly, and average document length increases from 288.47 words for Original to 297.51 at L4. These results characterize the preservation–intervention trade-off rather than implying uniform improvement across all quality dimensions.

**Evaluator details.** BERTScore is computed using RoBERTa-large layer 17 with `idf=False` and `rescale_with_baseline=False`. Perplexity is computed with GPT-Neo-2.7B using a maximum sequence length of 512 and is reported as the mean of per-document $\exp(\text{loss})$. CoLA uses the frozen RoBERTa-base acceptability evaluator and is treated as a constraint-compliance diagnostic because acceptability also participates in candidate gating. Bidirectional NLI is likewise gate-linked and is therefore not treated as independent semantic-preservation evidence in the headline results.

## H Additional Transformation Controls

The main paper limits the matched-edit comparison to Qwen Humanize and Pegasus. Here we report two additional generic transformation controls—Qwen Rewrite and back-translation—to characterize how detector responses depend on the transformation itself. These controls provide context for detector sensitivity and are not used to claim mechanism-specific superiority.

The additional controls further illustrate detector-specific variation. Qwen Rewrite yields lower AUROC than AnchorWrite for several detection methods, whereas AnchorWrite yields lower Binoculars AUROC at L2–L4. Back-translation produces particularly large reductions for FastDetect, LastDE++, LRR, and LogRank, but its RoBERTa AUROC increases at stronger intervention levels, reaching 0.8258 at L4 compared with 0.7980 for Original. These divergent responses show that a generic transformation can strongly affect individual detection methods without producing a consistent detector profile.

Table 12: AUROC (↓) for additional transformation controls at matched L1–L4 intervention levels.

| Regime | Method | FastD | LastDE++ | Binoc. | RoBERTa | LRR | LogRank | Edit (%) |
|---|---|---|---|---|---|---|---|---|
| L1 | Qwen Rewrite | 0.9397 | 0.9262 | 0.8863 | 0.7421 | 0.7669 | 0.7607 | 5.97 |
| | Back-translation | 0.9206 | 0.9047 | 0.8685 | 0.7723 | 0.7566 | 0.7405 | 6.01 |
| | Ours | 0.9446 | 0.9325 | 0.8886 | 0.7644 | 0.7796 | 0.7719 | 5.60 |
| L2 | Qwen Rewrite | 0.9237 | 0.9100 | 0.8764 | 0.7190 | 0.7387 | 0.7382 | 11.63 |
| | Back-translation | 0.8863 | 0.8629 | 0.8361 | 0.7836 | 0.7193 | 0.6980 | 11.64 |
| | Ours | 0.9290 | 0.9138 | 0.8686 | 0.7308 | 0.7533 | 0.7505 | 11.53 |
| L3 | Qwen Rewrite | 0.8951 | 0.8781 | 0.8590 | 0.6825 | 0.6921 | 0.7006 | 20.40 |
| | Back-translation | 0.8294 | 0.7974 | 0.7987 | 0.8066 | 0.6700 | 0.6352 | 20.34 |
| | Ours | 0.9126 | 0.9000 | 0.8457 | 0.7230 | 0.7303 | 0.7317 | 20.40 |
| L4 | Qwen Rewrite | 0.8777 | 0.8627 | 0.8496 | 0.6705 | 0.6680 | 0.6820 | 24.16 |
| | Back-translation | 0.8013 | 0.7643 | 0.7857 | 0.8258 | 0.6460 | 0.6039 | 24.21 |
| | Ours | 0.9036 | 0.8936 | 0.8388 | 0.7192 | 0.7214 | 0.7242 | 24.32 |

Table 13: Quality characteristics of the additional transformation controls. PPL is omitted because comparable frozen PPL outputs are unavailable for these transformations.

| Regime | Method | BERT ↑ | CoLA ↑ | TTR | MTLD | Dist.-2 | Avg. length |
|---|---|---|---|---|---|---|---|
| L1 | Qwen Rewrite | 0.9812 | 0.8553 | 0.5599 | 102.10 | 0.6019 | 288.76 |
| | Back-translation | 0.9795 | 0.8310 | 0.5579 | 99.91 | 0.5971 | 288.11 |
| L2 | Qwen Rewrite | 0.9769 | 0.8520 | 0.5653 | 106.07 | 0.6068 | 288.85 |
| | Back-translation | 0.9736 | 0.8097 | 0.5606 | 100.47 | 0.5975 | 287.46 |
| L3 | Qwen Rewrite | 0.9706 | 0.8471 | 0.5714 | 109.35 | 0.6105 | 289.22 |
| | Back-translation | 0.9650 | 0.7823 | 0.5627 | 98.79 | 0.5953 | 286.46 |
| L4 | Qwen Rewrite | 0.9680 | 0.8452 | 0.5744 | 112.06 | 0.6120 | 289.27 |
| | Back-translation | 0.9610 | 0.7711 | 0.5625 | 98.35 | 0.5930 | 286.18 |

The quality results provide complementary context for the detector changes. Qwen Rewrite maintains high BERTScore and relatively stable CoLA across intervention levels. Back-translation also retains high BERTScore, but its CoLA score decreases from 0.8310 at L1 to 0.7711 at L4. For comparison, AnchorWrite retains BERTScore 0.9678 and CoLA 0.8558 at L4 (Table 11). These results reinforce the main interpretation: lower detector AUROC alone is insufficient to characterize controlled rewriting because different transformations exhibit distinct detector and preservation profiles.

# I Domain Alignment and Adaptive Stopping

## I.1 Domain-level human-reference improvement

Table 14 reports mean reference-distance reduction by domain and operating regime. Mean $\Delta D$ increases from L1 to L4 in all six domains, although the magnitude of improvement varies by domain.

At L4, the fraction of documents with positive $\Delta D$ is 99.3% for Essay, 100% for Humanities, 98.7% for Reuters, 100% for Social Sciences, 100% for STEM, and 98.0% for WP. Thus, the aggregate alignment trajectory is observed across all evaluated domains rather than being driven by a single domain.

## I.2 Stopping reasons

Table 15 reports the terminal condition reached by each document at the final operating point of each regime.

Table 14: Mean human-reference distance reduction $\Delta D$ by domain and operating regime. Higher values indicate greater reduction in calibrated reference distance.

| Domain | L1 | L2 | L3 | L4 |
|---|---|---|---|---|
| Essay | 0.1032 | 0.2165 | 0.4025 | 0.4840 |
| Humanities | 0.0482 | 0.1071 | 0.2106 | 0.2741 |
| Reuters | 0.0827 | 0.1563 | 0.2991 | 0.3639 |
| Social Sciences | 0.0508 | 0.1128 | 0.2307 | 0.2987 |
| STEM | 0.0632 | 0.1375 | 0.2418 | 0.3151 |
| WP | 0.0596 | 0.1534 | 0.3480 | 0.4423 |

Table 15: Stopping reasons at the final operating point for the 900 documents in each regime.

| Regime | Edit ceiling reached | No valid sampled candidate found |
|---|---|---|
| L1 | 14 (1.56%) | 886 (98.44%) |
| L2 | 53 (5.89%) | 847 (94.11%) |
| L3 | 8 (0.89%) | 892 (99.11%) |
| L4 | 3 (0.33%) | 897 (99.67%) |

Across all regimes, most documents stop because no valid sampled candidate is found that both improves the fixed-reference objective and satisfies the frozen preservation constraints. Only 0.33–5.89% terminate by reaching the applicable edit ceiling. The configured ceilings therefore act primarily as upper bounds on intervention rather than prescribed modification levels.

# J Statistical Uncertainty

## J.1 Bootstrap protocol

For each detector with retained per-sample scores, Original and transformed AI outputs are aligned by canonical sample identity. In each of 2,000 bootstrap replicates, aligned Original–transformed AI pairs are resampled jointly, while the common pool of 900 human controls is resampled independently. We define

$$\Delta\mathrm{AUROC} = \mathrm{AUROC}_{\mathrm{Original}} - \mathrm{AUROC}_{\mathrm{Condition}},$$

so positive values indicate reduced separability between human controls and the transformed AI condition. The reported 95% confidence intervals are percentile intervals of the bootstrap differences.

Before an interval is accepted, AUROC recomputed from the retained raw scores must reproduce the corresponding frozen canonical value within the preset numerical tolerance. This audit passes all 48 tested cells: three rewritten methods across four operating regimes for FastDetect, RoBERTa, LogRank, and LRR. Table 16 reports the resulting Original-to-AnchorWrite intervals.

Table 16: Original-to-AnchorWrite AUROC reduction with 95% bootstrap confidence intervals. Positive values indicate reduced detector separability after rewriting.

| Detector | L1 | L2 | L3 | L4 |
|---|---|---|---|---|
| FastDetect | 0.0231 [0.0190, 0.0272] | 0.0387 [0.0327, 0.0451] | 0.0550 [0.0471, 0.0629] | 0.0641 [0.0552, 0.0729] |
| RoBERTa | 0.0336 [0.0257, 0.0414] | 0.0672 [0.0553, 0.0792] | 0.0749 [0.0624, 0.0884] | 0.0788 [0.0650, 0.0924] |
| LogRank | 0.0290 [0.0261, 0.0319] | 0.0504 [0.0459, 0.0549] | 0.0692 [0.0637, 0.0752] | 0.0767 [0.0703, 0.0831] |
| LRR | 0.0305 [0.0270, 0.0340] | 0.0568 [0.0515, 0.0622] | 0.0798 [0.0728, 0.0870] | 0.0887 [0.0807, 0.0973] |

Human-reference alignment is assessed separately using 2,000 bootstrap replicates over the 900 document-level $\Delta D$ values; the corresponding intervals are reported in Table 1. LastDE++ and Binoculars are reported with frozen point estimates because equivalent per-sample scores were not retained. These intervals quantify document-sampling uncertainty conditional on the frozen rewrite outputs and therefore do not include variability from stochastic candidate generation.

# K Cross-Backbone Transfer Analysis

## K.1 Protocol

The cross-backbone study evaluates Llama-3.2-3B-Instruct, Mistral-7B-Instruct-v0.3, and Phi-4-mini-instruct on the same fixed 300-document subset used in the supporting transfer experiment, comprising 50 documents from each of the six domains. Each backbone rewrites the same source documents independently under the separate 60%-budget stress endpoint. This protocol is distinct from the primary Qwen3-8B L1–L4 experiment and is used to assess whether the observed external detector-transfer effect persists across rewriting backbones rather than to provide a direct numerical comparison with the main operating regimes. The resulting outputs are frozen before external detector and post-hoc quality evaluation.

## K.2 Detector transfer

Table 17 reports the complete cross-backbone detector results underlying Figure 4. For each detector,

$$\Delta\mathrm{AUROC} = \mathrm{AUROC}_{\mathrm{Original}} - \mathrm{AUROC}_{\mathrm{Rewrite}},$$

so positive values indicate reduced separability between the human controls and rewritten AI documents. The Original values in this table correspond to the fixed 300-document cross-backbone subset and therefore need not equal the Original AUROCs reported for the 900-document main experiment.

Table 17: Cross-backbone detector transfer on the fixed 300-document subset. Values report AUROC (↓) and absolute reduction $\Delta = \mathrm{AUROC}_{\mathrm{Original}} - \mathrm{AUROC}_{\mathrm{Rewrite}}$. Higher positive $\Delta$ indicates greater reduction in detector separability. AUROC and $\Delta$ values are independently rounded from the corresponding full-precision estimates.

| Detector | Original | Llama | Δ | Mistral | Δ | Phi | Δ |
|---|---|---|---|---|---|---|---|
| FastDetect | 0.9654 | 0.8893 | 0.0761 | 0.8544 | 0.1110 | 0.8971 | 0.0682 |
| LastDE++ | 0.9569 | 0.8753 | 0.0816 | 0.8405 | 0.1164 | 0.8844 | 0.0725 |
| Binoculars | 0.8822 | 0.8340 | 0.0482 | 0.7838 | 0.0984 | 0.8291 | 0.0531 |
| RoBERTa | 0.8036 | 0.7019 | 0.1017 | 0.6119 | 0.1917 | 0.7205 | 0.0831 |
| LRR | 0.7938 | 0.6920 | 0.1019 | 0.6555 | 0.1384 | 0.7289 | 0.0649 |
| LogRank | 0.7630 | 0.6744 | 0.0886 | 0.6344 | 0.1286 | 0.6940 | 0.0690 |

AUROC decreases for all six detectors under all three rewriting backbones. The reductions range from 0.0482–0.1019 for Llama, 0.0984–0.1917 for Mistral, and 0.0531–0.0831 for Phi. Mistral produces the largest reduction for each detector in this evaluated configuration, while the magnitude of transfer varies across both detectors and rewriting backbones. These results support transfer beyond the primary Qwen3-8B backbone without implying backbone-invariant behavior.

## K.3 Quality preservation across rewriting backbones

Table 18 reports the corresponding semantic-similarity and fluency measurements. The same frozen evaluation pipeline used in the main experiment is applied to all three rewritten conditions.

Table 18: Detailed quality statistics for the cross-backbone experiment.

| Backbone | Mean BERT | Median BERT | Mean PPL | Median PPL |
|---|---|---|---|---|
| Llama | 0.9504 | 0.9497 | 10.9499 | 10.0291 |
| Mistral | 0.9600 | 0.9605 | 11.7142 | 10.7819 |
| Phi | 0.9802 | 0.9822 | 10.5209 | 9.3679 |

The corresponding Original documents have mean and median GPT-Neo-2.7B perplexities of 8.9330 and 7.6825, respectively. Across all three rewriting backbones, mean BERTScore remains at least 0.9504, while perplexity increases relative to Original. Phi retains the highest semantic similarity

and lowest rewritten perplexity, whereas Mistral exhibits the largest detector-AUROC reductions together with the highest mean rewritten perplexity. Llama lies between these behaviors. Together with Table 17, these results show a backbone-dependent detectability–quality trade-off rather than a uniform response across rewriting models.

# L Human Evaluation Protocol and Detailed Results

This appendix gives the complete protocol and statistical analysis for the blinded human evaluation reported in Sections 4.7 and 5.4. The study assesses perceived rewrite quality rather than authorship.

## L.1 Sampling and evaluated conditions

We sample 50 unique source documents from the frozen main evaluation population using deterministic seed 20260924, independently of detector scores, BERTScore, perplexity, human-reference distance, edit ratio, visual inspection, and human ratings. The sample contains nine documents each from Essay and STEM and eight each from Humanities, Reuters, Social Sciences, and WritingPrompts.

The same 50 canonical sources are evaluated under three previously frozen L4 conditions: AnchorWrite, Qwen Humanize, and Pegasus. No text is regenerated for the human study, yielding 150 unique source–method pairs.

## L.2 Annotators, blinding, and assignment

Six native-English-speaking graduate-student annotators (A01–A06) participate under full method blinding. Annotators see only the source text, one rewritten text, and an opaque item identifier; method names, detector information, human-reference scores, edit ratios, and revealing filenames are hidden.

Each of the 150 source–method pairs is rated by exactly three distinct annotators, yielding 450 judgments. Each annotator evaluates 75 items: 25 AnchorWrite, 25 Qwen Humanize, and 25 Pegasus outputs. The balanced incomplete assignment ensures that no annotator sees all three rewrites of the same source. For each annotator, 25 sources occur once and 25 occur twice; when two versions of one source are assigned, their booklet positions are separated by 50 items. Presentation order is independently randomized for each annotator.

## L.3 Rating criteria

For each source–rewrite pair, annotators provide four integer ratings from 1 to 5 using the anchors in Table 19.

Annotators are instructed to judge each item independently, use only integer scores from 1 to 5, and not attempt to identify the rewriting system.

## L.4 Statistical analysis

The 450 individual judgments are not treated as independent source observations. For each source, method, and quality dimension, the three ratings are averaged, yielding 50 paired source-level observations per method and dimension as the unit of inference.

Confidence intervals are percentile intervals from 10,000 nonparametric source-level bootstrap resamples using fixed seed 20260924, with the same sampled source indices used across methods within each replicate. For each quality dimension, an omnibus Friedman test compares the three methods. Pairwise comparisons use two-sided Wilcoxon signed-rank tests, with zero differences omitted and Holm correction applied to the three pairwise $p$-values within each dimension.

## L.5 Detailed descriptive results

Table 20 reports source-level descriptive statistics. Ratings are uniformly high, with medians between 4.667 and 5.000.

Table 19: Human-evaluation rating criteria.

| Criterion | Score | Definition |
|---|---|---|
| Semantic preservation | 1 | Meaning is substantially changed. |
| | 2 | Several important meaning changes. |
| | 3 | Main meaning is preserved, but noticeable differences exist. |
| | 4 | Meaning is preserved with only minor differences. |
| | 5 | Meaning is fully preserved. |
| Factual consistency | 1 | Major facts, entities, numbers, polarity, or claims are changed. |
| | 2 | Several important inconsistencies or omissions. |
| | 3 | Minor inconsistency or omission. |
| | 4 | Content is preserved with negligible issues. |
| | 5 | No factual/content change is apparent. |
| Fluency | 1 | Very difficult to read or seriously ungrammatical. |
| | 2 | Multiple grammatical or readability problems. |
| | 3 | Understandable but somewhat awkward. |
| | 4 | Fluent with only minor awkwardness. |
| | 5 | Fully fluent and grammatical. |
| Naturalness | 1 | Very unnatural. |
| | 2 | Noticeably unnatural. |
| | 3 | Generally acceptable. |
| | 4 | Natural. |
| | 5 | Very natural. |

Table 20: Detailed source-level human-evaluation results. SD is the sample standard deviation across the 50 source-level mean ratings; CIs use 10,000 source-level bootstrap resamples.

| Method | Dimension | Mean | Median | SD | 95% CI |
|---|---|---|---|---|---|
| AnchorWrite | Semantic | 4.840 | 5.000 | 0.215 | [4.780, 4.893] |
| | Factual | 4.893 | 5.000 | 0.196 | [4.840, 4.940] |
| | Fluency | 4.867 | 5.000 | 0.213 | [4.807, 4.920] |
| | Naturalness | 4.873 | 5.000 | 0.201 | [4.813, 4.927] |
| Qwen Humanize | Semantic | 4.847 | 5.000 | 0.193 | [4.793, 4.900] |
| | Factual | 4.827 | 5.000 | 0.215 | [4.767, 4.880] |
| | Fluency | 4.787 | 4.833 | 0.259 | [4.713, 4.853] |
| | Naturalness | 4.807 | 5.000 | 0.253 | [4.733, 4.873] |
| Pegasus | Semantic | 4.787 | 4.667 | 0.221 | [4.727, 4.847] |
| | Factual | 4.780 | 4.667 | 0.219 | [4.720, 4.840] |
| | Fluency | 4.773 | 4.667 | 0.247 | [4.707, 4.840] |
| | Naturalness | 4.787 | 4.667 | 0.188 | [4.733, 4.840] |

The individual-rating distributions are strongly concentrated near the upper end of the scale. Table 21 reports the fraction of the 150 ratings per method and dimension that equal 5 or are at least 4.

Table 21: Distribution of individual human ratings. Each method has 150 ratings per quality dimension.

| Method | Dimension | Score $= 5$ (%) | Score $\geq 4$ (%) |
|---|---|---|---|
| AnchorWrite | Semantic | 84.0 | 100.0 |
| | Factual | 89.3 | 100.0 |
| | Fluency | 86.7 | 100.0 |
| | Naturalness | 88.0 | 99.3 |
| Qwen Humanize | Semantic | 84.7 | 100.0 |
| | Factual | 82.7 | 100.0 |
| | Fluency | 80.0 | 98.7 |
| | Naturalness | 82.0 | 98.7 |
| Pegasus | Semantic | 78.7 | 100.0 |
| | Factual | 78.7 | 99.3 |
| | Fluency | 80.0 | 97.3 |
| | Naturalness | 78.7 | 100.0 |

### L.6 Paired significance tests

The omnibus tests in Table 22 show method differences for factual consistency and naturalness, but not for semantic preservation or fluency.

Table 22: Friedman tests over the 50 paired source-level observations.

| Dimension | $\chi^2_F(2)$ | $p$ |
|---|---|---|
| Semantic preservation | 2.992 | 0.2240 |
| Factual consistency | 7.740 | 0.0209 |
| Fluency | 3.590 | 0.1661 |
| Naturalness | 6.153 | 0.0461 |

Table 23 reports the corrected pairwise tests. The only significant comparison after Holm correction is higher factual consistency for AnchorWrite than Pegasus; the naturalness omnibus effect does not yield a significant corrected pairwise difference.

Table 23: Two-sided paired Wilcoxon signed-rank comparisons on source-level mean ratings. Mean difference is first method minus second method. Holm adjustment is performed separately within each quality dimension.

| Dimension | Comparison | Mean diff. | Raw $p$ | Holm $p$ |
|---|---|---|---|---|
| Semantic | AnchorWrite – Qwen Humanize | -0.0067 | 0.8474 | 0.8474 |
| | AnchorWrite – Pegasus | 0.0533 | 0.1551 | 0.3751 |
| | Qwen Humanize – Pegasus | 0.0600 | 0.1250 | 0.3751 |
| Factual | AnchorWrite – Qwen Humanize | 0.0667 | 0.0950 | 0.1899 |
| | AnchorWrite – Pegasus | 0.1133 | 0.0089 | **0.0267** |
| | Qwen Humanize – Pegasus | 0.0467 | 0.2694 | 0.2694 |
| Fluency | AnchorWrite – Qwen Humanize | 0.0800 | 0.1426 | 0.4277 |
| | AnchorWrite – Pegasus | 0.0933 | 0.1475 | 0.4277 |
| | Qwen Humanize – Pegasus | 0.0133 | 0.7500 | 0.7500 |
| Naturalness | AnchorWrite – Qwen Humanize | 0.0667 | 0.1664 | 0.3328 |
| | AnchorWrite – Pegasus | 0.0867 | 0.0616 | 0.1848 |
| | Qwen Humanize – Pegasus | 0.0200 | 0.9723 | 0.9723 |

### L.7 Rating agreement and ceiling effect

Because each source–method pair is rated by three of six annotators, agreement is evaluated using ordinal Krippendorff's $\alpha$, treating the 150 source–method pairs as observational units and the six annotators as raters with missing entries for unassigned items. We also report exact pairwise agreement among the three ratings available for each item.

Table 24: Human-rating agreement. Exact agreement is the fraction of within-item annotator pairs assigning the same integer rating.

| Dimension | Ordinal $\alpha$ | Exact agreement (%) |
|---|---|---|
| Semantic preservation | -0.042 | 69.8 |
| Factual consistency | -0.017 | 72.0 |
| Fluency | -0.004 | 70.4 |
| Naturalness | -0.049 | 70.2 |

Chance-corrected agreement is near zero or slightly negative, while exact pairwise agreement is approximately 70%. These values occur alongside strongly concentrated marginals: 97.3–100% of individual ratings across method–dimension combinations are at least 4. We therefore report both agreement measures transparently and do not interpret the study as showing strong inter-rater reliability. The pronounced ceiling concentration also limits sensitivity to small between-method differences and motivates the paired source-level analysis and conservative interpretation used in the main text.

### L.8 Qualitative examples

To complement the aggregate human-evaluation results, we report one qualitative AnchorWrite example from each of the six domains. To avoid cherry-picking, the six source identifiers were selected deterministically from the fixed 50-source human-evaluation subset using the smallest SHA-256 hash of `20260924|source_id` within each domain. Selection was therefore independent of model scores, human ratings, and qualitative inspection. For each selected source, we reproduce only the frozen Original and corresponding AnchorWrite L4 output. No example was regenerated or manually edited for presentation.

**Essay** `original_essay_112`

**Original.**
The Canadian government has a long history of passing legislation that impacts Indigenous rights and identity. One of the most significant pieces of legislation in this regard is Bill C-31, which was passed in 1985 to amend the Indian Act. The bill aimed to address longstanding gender discrimination within the Indian Act by restoring status to Indigenous women who had lost their status through marriage to non-Indigenous men. However, the bill had unintended consequences, including the creation of a new "second-class" category of Indigenous people and a further erosion of Indigenous identity and rights. The historical context in which Bill C-31 was passed is important to understanding its implications for Indigenous rights. Canada has a history of colonialism and forced assimilation of Indigenous people, including the residential school system, which forcibly removed Indigenous children from their families and communities to assimilate them into Euro-Canadian culture. The Indian Act, which has been in place since 1876, was another tool of assimilation, as it gave the government control over Indigenous people's lives, including their identity, land, and culture. Bill C-31 was a recognition of past injustices and discrimination against Indigenous women, but its impact has been complicated. Many Indigenous women who were denied status are now able to reclaim it, but their children

**AnchorWrite L4.**
The Canadian government has a long history of passing legislation that impacts Indigenous rights and identity. One of the most significant pieces of legislation in this regard is Bill C-31, which was passed in 1985 to amend the Indian Act. The bill aimed to address longstanding gender discrimination within the Indian Act by restoring status to Indigenous women who had lost their status through marriage to non-Indigenous men. However, the bill had unintended consequences, including the creation of a new "second-class" category of Indigenous people and a further erosion of Indigenous identity and rights. The

context behind the introduction of Bill C-31 is essential for understanding its impact on Indigenous rights. Canada has a long history of colonial rule and the imposition of assimilationist policies on Indigenous peoples, such as the residential school system, which systematically took Indigenous children away from their families and communities with the aim of integrating them into Euro-Canadian cultural norms. The Indian Act, which came into effect in 1876, served as another means of promoting assimilation, as it allowed the government to regulate various aspects of Indigenous people's lives, such as their identity, land, and cultural practices. Bill C-31 was a recognition of past injustices and discrimination against Indigenous women, but its impact has been complicated. Many Indigenous women who lost their status can now reclaim it, but their children

**Humanities** `original_humanities_124`

**Original.**
The concept of the good has been a subject of philosophical inquiry for centuries, with various interpretations across different cultures and time periods. Plato, the ancient Greek philosopher, proposed the existence of an eternal realm of forms or ideas, where the greatest of these ideas is the essence of being, known as "The Good." According to Plato, The Good is a perfect and eternal idea that exists independently of human perception. In this view, The Good represents the right relationship between all that exists, and it is believed to reside in the mind of the Divine or within a heavenly realm. It encompasses various aspects such as harmony in a just political community, love, friendship, the ordered human soul filled with virtues, and the appropriate relationship with the Divine and Nature. Plato's dialogues often highlight the virtues of a philosopher, or a lover of wisdom, who seeks to understand and embody The Good. The concept of the good is also closely related to theistic beliefs, where a person believes in the existence of a Supreme Being or gods (monotheism or polytheism). Theists may argue that the universe has a purpose and value according to the will of these creator(s), which may be partially beyond human understanding. One notable proponent of this view is Thomas Aquinas, a medieval philosopher and theologian, who attempted to prove the existence of God and establish the right relations that humans ought to have with the divine first cause. Monotheistic信仰者 may also hold the hope for infinite universal love, which is often referred to as "faith." In many religious doctrines, wisdom is defined as the knowledge and understanding of innate goodness. Additionally, the concepts of innocence, spiritual purity, and salvation are interconnected with the idea of the good. These concepts emphasize the importance of moral and spiritual development in achieving a harmonious relationship with the Divine and the natural world. In summary, the concept of the good is a multifaceted and complex idea that has been explored by philosophers, theologians, and believers across different cultures and time periods. It encompasses a wide range of values and virtues, including justice, love, friendship, and spiritual purity, all of which contribute to the understanding of the ultimate purpose and value of existence.

---

**AnchorWrite L4.**
The notion of the good has historically occupied a central place in philosophical discourse, encompassing a wide range of viewpoints across different cultures and time periods. Plato, the ancient Greek philosopher, introduced the concept of an eternal realm of forms or ideas, with the highest of these being the essence of being, referred to as "The Good." According to Plato, The Good is a perfect and eternal idea that exists independently of human perception. In this perspective, The Good symbolizes the correct relationship among all things that exist, and it is thought to dwell in the mind of the Divine or within a celestial realm. It includes diverse elements like harmony in a righteous political society, love, friendship, a well-ordered human soul rich in virtues, and the proper connection with the Divine and Nature. Plato's dialogues often highlight the virtues of a philosopher, or a lover of wisdom, who seeks to understand and embody The Good. The idea of the good is also intimately connected to religious beliefs, in which an individual holds faith in the presence of a Supreme Being or deities, whether singular or multiple. Theists may argue that the universe has a purpose and value according to the will of these creator(s), which may be partially beyond human understanding. One notable proponent of this view is Thomas Aquinas, a medieval philosopher and theologian, who attempted to prove the existence of God and establish the right relations that humans ought to have with the divine first cause. Monotheistic信仰者 may also hold the hope for infinite universal love, which is often referred to as "faith." In many religious doctrines, wisdom is defined as the knowledge and understanding of innate goodness. Moreover, the notions of innocence, spiritual purity, and salvation are closely linked to the concept of the good. These concepts emphasize the importance of moral and spiritual development in achieving a harmonious relationship with the Divine and the natural world. In conclusion, the notion of the good is a diverse and intricate concept that has been examined by philosophers, theologians, and individuals from various cultures and historical eras. It covers a broad spectrum of values and virtues, such as justice, love, friendship, and spiritual purity, all of which play a role in defining the fundamental aim and significance of life.

**Reuters** `original_reuters_22`

**Original.**
Abidjan, Ivory Coast: The West African nation's cocoa shipments are predicted to hit 120,000 tonnes by the end of November. This comes as welcome news to Ivory Coast cocoa farmers, who produce almost half of the world's cocoa beans. The cocoa industry is a significant contributor to Ivory Coast's economy, accounting for around a fifth of the nation's GDP and employing over 3 million people. The forecast increase in shipments signals a positive trend for Ivory Coast's cocoa industry. "We are optimistic that our farmers will be able to meet the shipment targets set by the government," said Aka Marcel, the head of the Ivorian Coffee and Cocoa Council. "The government's plan to improve our shipping infrastructure has played a crucial role in enabling us to meet these targets." The Ivorian government has invested heavily in improving the nation's cocoa supply chain, including road network upgrades and increasing storage facilities. These investments have been critical in enabling Ivory Coast to boost its cocoa production, helping it to regain its position as the world's leading cocoa producer. The increase in production has had a positive effect on Ivory Coast's economy. It has helped to fund public infrastructure, including hospitals, schools, and water supply systems, which have benefited thousands of Ivorians. The Ivory Coast government

**AnchorWrite L4.**
Abidjan, Ivory Coast: The West African nation's cocoa shipments are predicted to hit 120,000 tonnes by the end of November. This comes as welcome news to Ivory Coast cocoa farmers, who produce almost half of the world's cocoa beans. The cocoa sector plays a major role in Ivory Coast's economy, contributing roughly a fifth of the country's GDP and providing jobs for more than 3 million individuals. The expected rise in exports indicates an encouraging outlook for Ivory Coast's cocoa sector. "We are confident that our farmers will be capable of fulfilling the shipment goals established by the government," stated Aka Marcel, the leader of the Ivorian Coffee and Cocoa Council. "The government's plan to improve our shipping infrastructure has played a crucial role in enabling us to meet these targets." The Ivorian government has made significant investments in upgrading the country's cocoa supply chain, such as improvements to road networks and the construction of additional storage centers. These investments have been crucial in helping Ivory Coast boost its cocoa output, allowing it to once again take the lead as the world's largest cocoa producer. The increase in production has had a positive effect on Ivory Coast's economy. It has helped fund public infrastructure, including hospitals, schools, and water systems, which have enhanced the quality of life for thousands of Ivorians. The Ivory Coast government

**Social Sciences** `original_social_sciences_10`

**Original.**
In the legal profession, the term "practice of law" encompasses a variety of activities, including representation in tribunals and administrative agencies. Additionally, numerous jurisdictions have permitted law students and non-locally admitted individuals to represent indigent persons or others through clinical legal education programs. The more contentious issue revolves around out-of-court activities, such as drafting documents and providing advice, and whether these actions are considered unauthorized practice of law. Some states have expanded the definition of "practice of law" to include individuals who represent clients in arbitration or serve as arbitrators in disputes. There is an emerging conflict between the multijurisdictional practice of law in arbitration proceedings within the financial services industry and state regulation of lawyers. Generally, the rule is that an appearance at an arbitration does not constitute the practice of law, with a few exceptions. The United States Bankruptcy Court for the Eastern District of Tennessee has ruled that providing clients with explanations or definitions of legal terms is, by itself, considered giving legal advice. Similarly, the North Carolina State Bar has determined that defining lien law terms, issuing warnings regarding time requirements, and sending reminders about preliminary notices, combined with the preparation of legal documents, constitute providing legal advice. In Texas, the law generally prohibits a person who is not an attorney from representing a client in a personal capacity. It is essential to note that regulations and definitions may vary across jurisdictions, and understanding the specific rules in each area is crucial for legal practitioners and those seeking representation.

**AnchorWrite L4.**
In the legal field, the phrase "practice of law" refers to a range of activities, such as providing representation in courts and government agencies. Many jurisdictions also allow law students and non-resident attorneys to represent low-income clients or others via clinical legal education programs. The more controversial matter relates to activities outside the courtroom, like preparing documents and offering counsel, and whether these acts constitute the unauthorized practice of law. Some states have expanded the definition of "practice of law" to include individuals who represent clients in arbitration or serve as arbitrators in disputes. There is an emerging conflict between the multijurisdictional practice of law in arbitration proceedings within

the financial services industry and state regulation of lawyers. Typically, it is the case that participating in an arbitration proceeding does not amount to engaging in the practice of law, except in certain limited circumstances. The United States Bankruptcy Court for the Eastern District of Tennessee has ruled that providing clients with explanations or definitions of legal terms is, by itself, considered giving legal advice. Similarly, the North Carolina State Bar has concluded that explaining terms related to lien laws, informing clients about statute of limitations, sending reminders concerning preliminary notices, and drafting legal documents are all regarded as offering legal advice. In Texas, the law generally prohibits a person who is not an attorney from representing a client in a personal capacity. It is important to recognize that laws and interpretations can differ from one region to another, and being aware of the particular guidelines in each location is vital for legal professionals and individuals looking for legal assistance.

**STEM** `original_stem_101`

**Original.**
This section delves into the intricacies of defect density scaling during a slow quench in critical systems, particularly highlighting the distinction between quenches through a critical line versus a critical point. The authors propose a general conclusion derived from a $d$-dimensional model with $z= u=1$, showcasing how the defect density scales with the quench rate, $\tau$, and is influenced by the dimensionality of the critical surface traversed during the quench. The innovative aspect of this research is the formulation of a scaling law for defect density applicable to arbitrary critical systems, encapsulated in the equation \( n \sim 1/\tau^{m u/(z u+1)} \), which elegantly generalizes previous results and offers new insights into the behavior of defect densities in various dimensions and critical exponents. Furthermore, the paper explores the correlation of defects in a 2D system, introducing operators that measure the correlation between defects produced during the quench. The authors demonstrate that the correlation function can be calculated similarly to the 1D case, and provide an expression for the expectation value of the correlation operator in terms of the Majorana fermion operators. A key result is the derivation of the correlation function in the post-quench state, revealing the spatial anisotropy of the defect correlation function, even for equal couplings \( J_1 = J_2 \). The paper also presents a numerical evaluation of the correlation function, illustrating the anisotropic nature of defect correlations and their dependence on the quench rate. The language used in this section is clear and concise, facilitating an understanding of the complex phenomena under investigation. The results are presented in a manner that emphasizes the generalizability of the findings, making a significant contribution to the understanding of defect dynamics in critical systems. The figures included in the paper offer a visual representation of the theoretical predictions, aiding in the comprehension of the spatial anisotropy and the scaling behavior of defect correlations.

---

**AnchorWrite L4.**
This section explores the complexities of defect density scaling in the context of a slow cooling process within critical systems, with a focus on the difference between quenches across a critical line compared to those passing through a critical point. The authors propose a general conclusion derived from a $d$-dimensional model with $z= u=1$, showcasing how the defect density scales with the quench rate, $\tau$, and is influenced by the dimensionality of the critical surface traversed during the quench. The innovative aspect of this research is the formulation of a scaling law for defect density applicable to arbitrary critical systems, encapsulated in the equation \( n \sim 1/\tau^{m u/(z u+1)} \), which elegantly generalizes previous results and offers new insights into the behavior of defect densities in various dimensions and critical exponents. Furthermore, the paper explores the correlation of defects in a 2D system, introducing operators that measure the correlation between defects produced during the quench. The researchers show that the correlation function can be evaluated in a manner analogous to the 1D scenario, and they present an expression for the expected value of the correlation operator using Majorana fermion operators. A key result is the computation of the correlation function in the post-quench state, revealing the spatial anisotropy of the defect correlation function, despite equal couplings $ J_1 = J_2 $. The paper also presents a numerical evaluation of the correlation function, illustrating the anisotropic nature of defect correlations and their dependence on the quench rate. The terminology in this section is straightforward and to the point, aiding in the comprehension of the intricate processes being examined. The results are presented in a manner that emphasizes the generalizability of the findings, making a significant contribution to the understanding of defect dynamics in critical systems. The figures included in the paper offer a visual representation of the theoretical predictions, thereby aiding in the comprehension of spatial anisotropy and the scaling behavior of defect correlations.

**WritingPrompts / WP** `original_wp_82`

**Original.**
As Jack lay in bed, reflecting on the past day, he knew it had been the worst day of his life. His boss had scolded him, his car had broken down, and he had received some bad news from his doctor. As he closed his eyes, he wished he could go back and change it all. Suddenly, he was surrounded by three glowing orbs. A voice spoke to him, "You are given three options. Continue your life as is, restart the day, or restart your entire life. Choose wisely." Jack thought long and hard. He didn't want to relive the same day again, knowing how terrible it was. He also didn't want to start his entire life over, as there were many good moments he cherished. So, he chose to continue his life as it was, determined to make the best of every day. From that day on, Jack made conscious efforts to be positive and productive. He worked hard at his job, maintained a healthy lifestyle, and spent quality time with loved ones. The challenges he faced didn't disappear, but he learned to handle them with grace and resilience. Years went by, and Jack realized that he made the right choice that fateful night. He couldn't change the past, but he had

**AnchorWrite L4.**
As Jack lay in bed, thinking back on the previous day, he realized it had been the worst day of his life. His boss had scolded him, his car had broken down, and he had received some bad news from his doctor. As he closed his eyes, he wished he could go back and change it all. Suddenly, he was surrounded by three glowing orbs. A voice spoke to him, "You are given three options. Proceed with your current existence, begin anew for the day, or reset your whole life. Choose wisely." Jack thought long and hard. He didn't want to relive the same day again, knowing how terrible it was. He also wasn't keen on beginning anew, as he had many precious memories that were important to him. Therefore, he decided to keep living his life the way it was, resolute in his commitment to making the most of each day. From that point forward, Jack consciously aimed to stay positive and remain productive. He put in long hours at work, stayed active and balanced in his daily life, and dedicated time to being with family and friends. The challenges he faced didn't disappear, but he learned to handle them with grace and resilience. Years went by, and Jack realized that he made the right choice that fateful night. He couldn't change the past, but he had

**Qualitative observations.** These deterministically selected examples illustrate the lexical and structural changes produced by AnchorWrite at L4 rather than evidence of aggregate superiority. The rewrites primarily use local lexical substitution and sentence restructuring while retaining source-level entities, quantities, and domain-specific content. Source artifacts, awkward phrasing, and truncation are retained where present rather than manually corrected. The examples should therefore be interpreted alongside the aggregate automatic and human quality results.

## M Failure Modes and Interpretation Boundaries

The following limitations define the scope in which the reported results should be interpreted.

**Detector disagreement.** Rewriting does not affect all detectors uniformly. For example, Pegasus at stronger regimes reduces LogRank AUROC while RoBERTa and Binoculars do not follow the same monotonic trend; Qwen Rewrite and back-translation exhibit different profiles again. Detector reduction should therefore be interpreted across multiple detector families rather than as a transformation-independent property.

**Search saturation.** Most documents terminate because no valid sampled candidate is found that both improves the fixed-reference objective and satisfies the frozen preservation constraints, even when edit allowance remains. AnchorWrite therefore does not guarantee reaching a configured edit ceiling, and realized edit ratio is the appropriate measure of intervention strength (Appendix I).

**Reference-corpus dependence.** AnchorWrite optimizes distance to a fixed UMBC-derived reference rather than a universal notion of human-likeness. The held-out audit in Appendix B.4 shows that valid human text from the evaluation domains can have larger $D(x)$ than the Original AI documents. Accordingly, lower $D(x)$ should be interpreted only as movement toward the fixed corpus-derived statistic, and reduced detector separability should not be interpreted as evidence of human authorship.

**Cross-backbone scope.** The supporting transfer study evaluates three alternative rewriting backbones on a fixed 300-document subset under a single 60%-budget stress endpoint. AUROC decreases

for all six detectors under all three backbones, but both reduction magnitude and quality cost vary across models. These results support transfer beyond the primary Qwen3-8B backbone but do not establish backbone invariance or universal generalization. Confidence intervals were not computed for these cross-backbone detector comparisons.

**Computational cost.** We do not report wall-clock or GPU-hour comparisons, so the baseline study should not be interpreted as compute-matched. AnchorWrite incurs additional cost from repeated localization, candidate generation, validation, and document rescoring after accepted edits. This cost increases with the operating regime: L1–L4 request 10, 20, 30, and 40 candidates per retained target sentence, consider up to 6, 8, 10, and 12 target sentences per round; and allow up to 8, 12, 15, and 20 rounds, respectively. In contrast, the baseline comparison controls realized edit ratio rather than proposal count or inference cost. A dedicated runtime and GPU-efficiency comparison is therefore left for future work.

**Human-evaluation scope.** The blinded study covers 50 sources and six native-English-speaking graduate-student annotators. Ratings are strongly concentrated near the upper end of the five-point scales, limiting sensitivity to small between-method differences; chance-corrected agreement is also low despite approximately 70% exact pairwise agreement (Appendix L). We therefore interpret the study as evidence of high perceived rewrite quality within the evaluated sample, rather than as a fine-grained ranking of methods or evidence of human authorship.

# N Reproducibility Checklist

## N.1 Reproduction workflow

The reproduction workflow is ordered so that downstream detector evaluation cannot influence rewriting or baseline generation:

1. construct the fixed UMBC 2-gram and 4-gram human-reference models and the 10,000-text length-conditioned calibration table;
2. generate and freeze L1–L4 AnchorWrite outputs for all 900 main-study AI documents;
3. generate the matched-edit baseline outputs using the realized intervention targets of the frozen AnchorWrite outputs;
4. construct the canonical evaluation inputs and fixed human-control pool;
5. run the frozen post-hoc quality-evaluation pipeline;
6. evaluate Original, AnchorWrite, and baseline conditions against the same human controls using the external detectors; and
7. perform canonical-ID integrity checks, score-orientation audits, bootstrap uncertainty estimation, and final result aggregation.

The supporting cross-backbone experiment is run separately on its fixed 300-document subset (50 per domain) under the distinct 60%-budget stress endpoint described in Appendix K. Its outputs are likewise frozen before detector and quality evaluation.

## N.2 Integrity rules

Original–rewrite comparisons in the main study are joined by $(\text{domain}, \text{canonical sample id})$ and must satisfy one-to-one alignment. Each complete main-study AI condition contains 900 documents; the separate cross-backbone conditions contain the fixed 300-document subset defined in Appendix K. Detector-score orientation is fixed globally according to the detector or statistic definition before condition-level AUROC is examined. For detectors with retained per-sample scores, AUROC recomputed from the raw scores must reproduce the corresponding frozen canonical point estimate within the preset numerical tolerance before any bootstrap confidence interval is accepted.

### N.3 FROZEN EVALUATORS

The primary post-hoc quality evaluators are BERTScore with RoBERTa-large (layer 17, `idf=False`, `rescale_with_baseline=False`) and GPT-Neo-2.7B per-document perplexity with maximum sequence length 512. CoLA uses the frozen RoBERTa-base acceptability evaluator and is treated as a constraint-compliance diagnostic because acceptability also participates in candidate gating. Bidirectional DeBERTa NLI is likewise gate-linked and is not treated as independent semantic-preservation evidence. Deterministic lexical-diversity, length, and edit-analysis scripts are kept fixed across conditions.

External-detector implementations, checkpoints, preprocessing, and score orientation are fixed across Original, AnchorWrite, and baseline conditions. The same 900 human controls are used throughout the main detector evaluation; the cross-backbone study uses the corresponding fixed evaluation protocol described in Appendix K.

### N.4 HUMAN-REFERENCE ASSETS

Reproduction of the human-reference component requires the cleaned-UMBC preprocessing specification, the fixed word-level 2-gram and 4-gram KenLM configurations, and the 10,000-text human-reference calibration table. Appendix B.4 gives the scoring and calibration checks used to verify compatibility between the frozen reference assets and evaluation pipeline. Where redistribution of the underlying corpus is restricted, users must obtain the source corpus independently and apply the preprocessing and construction procedure specified above.

### N.5 SOFTWARE AND HARDWARE

The experiments use Python 3.10, PyTorch 2.5.1+cu121, Transformers 4.51.3, Sentence-Transformers 4.1, pandas 2.3.3, NumPy 2.2.6, scikit-learn 1.7.2, and KenLM. Generation and evaluation use NVIDIA RTX 3090- and 4090-class GPUs. Documents are processed independently across devices, with no cross-sample information sharing.