

Shortcut-Controlled Evaluation for AI-Generated Text Detection

Hee-Young Jung^{1*}, Jae-Eon Ahn^{2*}, Seong-Jin Park¹, Kang-Min Kim^{1†}

¹Department of Software Convergence

Kyung Hee University, Yongin, Republic of Korea

²Department of Artificial Intelligence

The Catholic University of Korea, Bucheon, Republic of Korea

{heeyoung37, sjpark011025, kangmin89}@khu.ac.kr, {cuk_13210067}@catholic.ac.kr

Abstract

Large language models (LLMs) are increasingly employed in writing-assistance for on-line text, making AI-generated text (AIGT) detection a critical challenge for managing the provenance of text. However, it remains unclear whether the high detection performance reported in existing AIGT benchmarks reflects a detector's ability to capture essential differences between human-written and AI-generated text. Rather, this high performance might be driven by easily measurable shortcuts in those benchmark settings. To address this question, we decompose the detection signal in AIGT and systematically analyze the role of shortcuts in existing benchmark settings. Our analysis reveals that a simple logistic regression classifier trained solely on surface shortcut features (e.g., formatting style and lexical usage) achieves strong classification performance, without relying on any semantic signals. Furthermore, the consistent performance degradation of existing detectors when such shortcuts are removed suggests that current AIGT evaluations are influenced by shortcut signals. Motivated by these findings, we present AIGT-SIGNALBENCH, a shortcut-controlled benchmark designed to reflect realistic AI-assisted writing scenarios.

 [GitHub Repository](#)

1 Introduction

Large language models (LLMs) have become a central part of how online text is produced and revised. Online texts (e.g., news articles, blog posts, social media posts, and technical documents) are increasingly generated by LLMs or written with LLM assistance (Lee et al., 2022; Padmakumar and He, 2024). This recent shift changes the composition of the web-scale text environment.

At the same time, recent LLMs can produce fluent and contextually appropriate text that is difficult

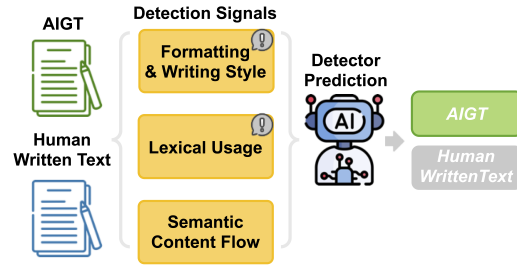


Figure 1: Problem definition. Easily measurable detection signals can act as shortcuts in distinguishing human-written and AI-generated texts.

to distinguish from human-written text—making the rapid growth of AI-generated text (AIGT) a matter of both social and technical concern. Socially, low-cost text generation can scale junk websites, spam, and coordinated political manipulation (Weidinger et al., 2022; Yang and Menczer, 2023). As AIGT accumulates online, future models risk being trained increasingly on AIGT rather than human-produced text, a feedback loop that can lead to model collapse and the gradual erosion of rare but important patterns in the original data distribution (Dodge et al., 2021; Shumailov et al., 2024). Thus, the spread of AIGT affects not only the integrity of online information, but also the reliability of future LLM training corpora.

These concerns have made it increasingly important to determine whether a given text was written by a human or generated by AI. In response, prior work has proposed a wide range of AIGT detection methods, including watermarking-based methods (Kirchenbauer et al., 2023), likelihood- and perplexity-based methods (Mitchell et al., 2023; Hans et al., 2024), neural detectors trained on human-written and AI-generated text, pre-trained language model (PLM) fine-tuning (Solaiman et al., 2019), and rewrite-based detectors (Zhou et al., 2026). Many of these methods report strong performance on existing AIGT benchmark setting.

Despite these strong results, studies of AIGT

*These authors contributed equally to this work.

†Corresponding author.

detectors have repeatedly reported unreliable behavior in real-world. For GPTZero (Tian and Cui, 2023), false-positive rates as high as 50% have been reported in evaluations of human-written and machine-translated texts, while other studies observed false-negative rates above 30% for AIGT (Weber-Wulff et al., 2023; Ibrahim et al., 2023). These results point to a gap between reported benchmark performance and real-world reliability.

This gap motivates our central question: Do existing AIGT benchmarks rely on easily measurable shortcuts? To address this question, we analyze human-written and AI-generated text in existing benchmark settings through a signal-decomposition of AIGT detection, illustrated in Figure 1. Our analysis shows that formatting and lexical signals (e.g., paragraph structure, transition words, cliché phrases) appear systematically in AIGTs. Because these signals are easy to detect and reduce without semantic-level analysis, we define them as shortcuts. A simple classifier trained only on these shortcut features achieves strong performance, and most detectors consistently degrade when these shortcuts are reduced.

This finding suggests that detectors should be evaluated under shortcut-controlled conditions, requiring them to rely on more essential signals. To this end, we present AIGT-SIGNALBENCH, which controls for shortcut signals and reflects realistic AI-assisted writing conditions, where AIGT may be further edited by an LLM or a human (Padmakumar and He, 2024). We evaluate existing detectors on this benchmark and analyze their performance.

In summary, our contributions are as follows:

- We propose a signal-decomposition of AIGT detection signals that distinguishes formatting and writing-style signals, lexical usage signals, and semantic-content flow signals.
- We provide an empirical diagnostic study showing that existing benchmark settings contain systematic and measurable shortcut signals between human-written and AI-generated texts, and that most baseline detectors degrade when easily controllable surface shortcuts are reduced.
- We present a shortcut-controlled benchmark AIGT-SIGNALBENCH, designed to evaluate AIGT detectors beyond surface-level artifacts under more realistic AI-assisted writing conditions.

Signal type	Fine-grained signals
Formatting & writing style	Paragraph count; average paragraph length; comma usage; em-dash usage; numeric expressions; sentence-length variance
Lexical usage	Transition words; vague intensifiers; cliché phrases
Semantic-content flow	Generic conclusions; semantic redundancy; formulaic openings; template-like progression; flow predictability

Table 1: Three types of AIGT detection signals. Details on how the fine-grained signals were selected and measured are provided in Appendix A.

2 A Signal-decomposition of AIGT Detection

AIGT detection is typically formulated as a binary provenance classification problem to predict whether an input text is written by a human or generated by an AI (Wu et al., 2025a). However, a simple binary label does not reveal what textual evidence the detector actually used. A single text contains multiple types of signals, ranging from easily measurable surface properties to deeper semantic-content flow signals. We therefore analyze AIGT through distinct types of detection signals.

2.1 Detection Signals in AIGT

Inspired by recent work that analyzes human- and LLM-generated texts across lexical, syntactic, and semantic dimensions (Reinhart et al., 2025; Zanutto and Aroyehun, 2025), we organize AIGT detection signals into three categories: formatting and writing-style, lexical usage, and semantic-content flow. Table 1 summarizes these three categories and their fine-grained signals.

2.2 Defining Shortcut Signals

We define a *shortcut* as a benchmark-contingent surface signal that can be quantified using simple counting- or frequency-based measures and readily controlled through prompt constraints or post-processing rules. We use this operational definition to examine whether detector performance remains robust when such easily modifiable surface regularities are reduced.

Shortcut Signals Formatting and writing-style signals include paragraph count, average paragraph length, punctuation usage, numeric expressions, and sentence-length regularity. Lexical usage signals include transition words, vague intensifiers, and cliché phrases. The corresponding word and

phrase lists are provided in Appendix A.1. We treat these two categories as shortcut signals in our analysis because they appear as systematic human–AI differences in benchmark data, can be measured using simple counting- or frequency-based methods, and can be explicitly reduced through prompt constraints or post-processing. If such signals remain strong in an evaluation set, detector performance may partly reflect sensitivity to benchmark-contingent surface regularities rather than robustness to their modification.

Non-shortcut Signals Semantic-content flow signals include generic conclusions, semantic redundancy, formulaic openings, template-like progression, and flow predictability. These signals are detection-relevant, but we do not treat them as shortcuts because they are less directly measurable and less easily controlled than formatting or lexical cues. Instead, they require semantic-level analysis, such as judging whether a text follows a formulaic progression or repeats the same content in different words. This distinction allows us to test how much detector performance is supported by easily controllable surface cues, while separately examining whether higher-level semantic-content patterns remain after such cues are reduced.

3 Analysis of Shortcut Signals in Existing AIGT Benchmarks

This section examines whether shortcut signals are present in conventional AIGT benchmark settings (Bao et al., 2024). We first use shortcut feature analysis to identify measurable human–AI gaps, then test their separability before and after shortcut control using logistic regression and PLM embedding-based classification. We also conduct LLM evaluation to examine semantic-level patterns.

3.1 Paired Human–AI Corpus

We use a paired corpus consisting of human-written source texts and corresponding AIGTs. The analysis includes 2,400 human–AI pairs. Human texts are collected from XSum (Narayan et al., 2018), SQuAD (Rajpurkar et al., 2016), WritingPrompts (Fan et al., 2018), and PubMedQA (Jin et al., 2019). AI-generated counterparts are produced by prompting each LLM with the first 30 tokens of the corresponding human-written text, following the generation protocol of FastDetectGPT (Bao et al., 2024). For PubMedQA, the associated question is used as the prompt instead. Source generator models are replaced with recent LLMs commonly used

for writing assistance, including GPT-5.5 (OpenAI, 2026), GPT-4 (Achiam et al., 2023), GPT-3.5-Turbo (OpenAI, 2023), Gemini 2.5 Pro (Comanici et al., 2025), Claude Sonnet 4.6 (Anthropic, 2026), and DeepSeek-V4-Pro (DeepSeek-AI, 2026).

3.2 Statistics-based Shortcut Feature Extraction

We extract nine shortcut features from each text to quantify formatting and writing style signals, as well as lexical usage signals. These features are designed to capture surface-level but systematic differences between human-written and AI-generated texts, each quantifiable through simple counting or frequency-based procedures. We focus on features that capture recurring patterns in AIGT and exhibit the largest absolute standardized differences between human-written and AI-generated texts. Details on how the nine features were selected are provided in Appendix A.2.

Feature Distribution We compare human-written and AI-generated text across the nine shortcut features using their mean, variance, standard deviation, Cohen’s d , and paired Cohen’s d_z . Cohen’s d measures the standardized mean difference between the two groups, and paired Cohen’s d_z captures how consistently each feature changes from the human source text to its AIGT counterpart (Cohen, 1988).

Table 2 and Figure 2 show that human-written and AI-generated texts differ in consistent surface-level patterns. AIGTs are more structurally segmented: they have shorter average paragraphs, more paragraphs, and more commas. At the same time, human-written texts contain more numerical expressions. These results show that formatting and lexical usage signals are easily observable in paired human–AI data.

3.3 Shortcut Signals Drive Linear Separability

To test whether the observed shortcut signals directly support human–AI classification, we evaluate linear separability before and after shortcut control. The controlled setting is constructed through prompt-based editing that reduces shortcuts while preserving the original content. Details of the shortcut-control prompts, post-control feature statistics, and content-preservation validation of semantic similarity, named entities, and numeric expressions are provided in Appendix B.

Feature	H mean	H var	H std	AI mean	AI var	AI std	AI-H	d	d_z
Avg. paragraph length	115.139	2664.474	51.619	71.018	2070.709	45.505	-44.122	-0.907	-0.693
Paragraph count	1.000	0.000	0.000	2.550	5.878	2.424	+1.550	0.904	0.639
Commas / 100w	4.616	10.316	3.212	6.481	8.846	2.974	+1.864	0.602	0.503
Numbers / 100w	1.667	6.022	2.454	0.875	2.300	1.517	-0.793	-0.389	-0.404
Transition words / 100w	0.148	0.226	0.475	0.275	0.370	0.608	+0.127	0.232	0.176
Em-dashes / 100w	0.053	0.090	0.301	0.140	0.231	0.480	+0.088	0.219	0.170
cliché phrases / 100w	0.003	0.004	0.066	0.028	0.025	0.158	+0.025	0.210	0.164
Sent. length variance CV	0.463	0.053	0.230	0.422	0.041	0.204	-0.042	-0.191	-0.152
Vague intensifiers / 100w	0.083	0.179	0.423	0.160	0.290	0.539	+0.077	0.159	0.117

Table 2: Descriptive statistics for nine shortcut features across human-written (H) and AI-generated (AI) texts. Features marked with “/100w” are normalized as counts per 100 words. AI-H denotes the raw mean difference; d and d_z report Cohen’s d and its variance-corrected counterpart, respectively.

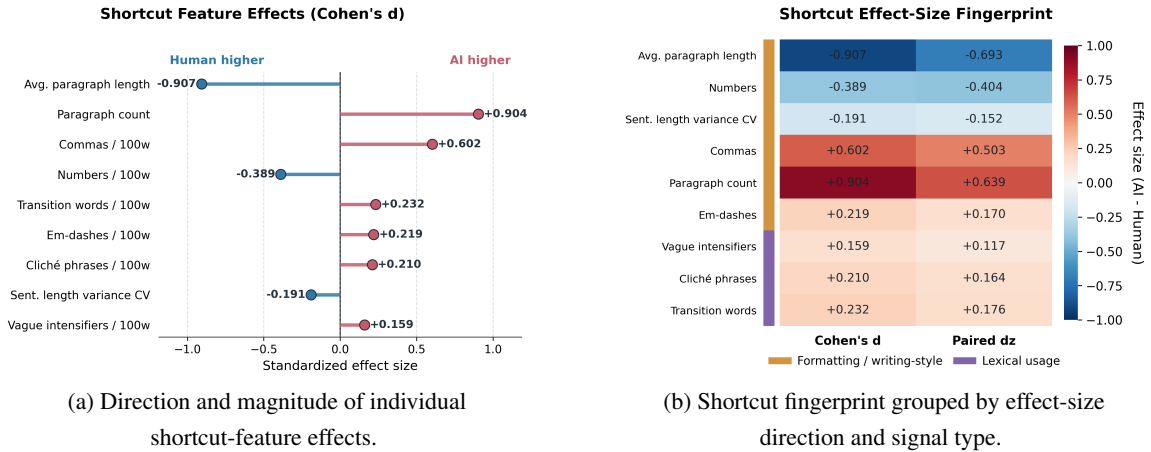


Figure 2: Visualization of shortcut feature effects. Positive effect sizes indicate higher values in AIGTs, while negative effect sizes indicate higher values in human-written texts.

Shortcut-only Logistic Regression We first train a logistic regression classifier (Cox, 1958) using only the nine shortcut features. Each input is represented as a nine-dimensional vector of counting- and frequency-based statistics. The AI-generated class is treated as positive label, and 5-fold cross-validation is used to prevent leakage.

Table 3 shows that these surface features alone provide strong discriminative signals in the conventional setting. Shortcut-only LR reaches AUROC 0.884 and AUPRC 0.902 overall (Fawcett, 2006; Davis and Goadrich, 2006), indicating that human-written and AI-generated texts can be separated without semantic signals. After shortcut control, performance drops sharply to AUROC 0.528 and AUPRC 0.530. The shortcut-only LR experiment is intended as a diagnostic probe rather than a standalone proof of detector shortcut reliance. Because the classifier is restricted to the nine controlled features, its post-control degradation is expected and primarily verifies that the intervention weakens the targeted measurable features.

Embedding-based Classification We next examine whether human-written and AI-generated texts are separable in PLM embedding space. Unlike the shortcut-only logistic regression classifier, PLM embeddings can encode broader textual information (e.g., lexical patterns, topic information, and semantic-content flow) (Devlin et al., 2019). We use sentence-transformers/all-mpnet-base-v2 (Reimers and Gurevych, 2019) to encode each text into a 768-dimensional embedding, and train an L2-regularized logistic regression classifier on the resulting representations. We refer to this embedding-based classification approach as PLM+LR.

Table 3 shows that PLM+LR also separates human-written texts from AIGTs in the original setting, achieving AUROC 0.910 and AUPRC 0.913 overall. After shortcut control, performance decreases to AUROC 0.794 and AUPRC 0.796. Dataset-wise, XSum and SQuAD exhibit clear degradation after control, whereas WritingPrompts and PubMedQA show only marginal changes. The

Model	Scope	F1			AUROC			AUPRC		
		Orig.	Ctrl.	Δ	Orig.	Ctrl.	Δ	Orig.	Ctrl.	Δ
Shortcut LR	Overall	0.808	0.488	-0.320	0.884	0.528	-0.356	0.902	0.530	-0.372
	XSum	0.948	0.501	-0.447	0.980	0.530	-0.450	0.985	0.535	-0.450
	SQuAD	0.705	0.500	-0.205	0.752	0.520	-0.232	0.757	0.505	-0.252
	WritingPrompts	0.950	0.572	-0.378	0.980	0.640	-0.340	0.986	0.666	-0.320
	PubMedQA	0.739	0.475	-0.264	0.826	0.510	-0.316	0.837	0.505	-0.332
PLM+LR	Overall	0.835	0.714	-0.121	0.910	0.794	-0.116	0.913	0.796	-0.117
	XSum	0.853	0.669	-0.184	0.929	0.738	-0.191	0.932	0.741	-0.191
	SQuAD	0.802	0.596	-0.207	0.883	0.647	-0.236	0.888	0.649	-0.239
	WritingPrompts	0.868	0.868	0.000	0.940	0.935	-0.004	0.941	0.933	-0.008
	PubMedQA	0.837	0.822	-0.015	0.916	0.892	-0.023	0.920	0.899	-0.021

Table 3: Linear separability of human-written and AI-generated texts before and after shortcut control. Orig. and Ctrl. denote the original and shortcut-controlled settings, respectively. Δ is computed as Ctrl. – Orig. Negative values indicate performance drops after shortcut control.

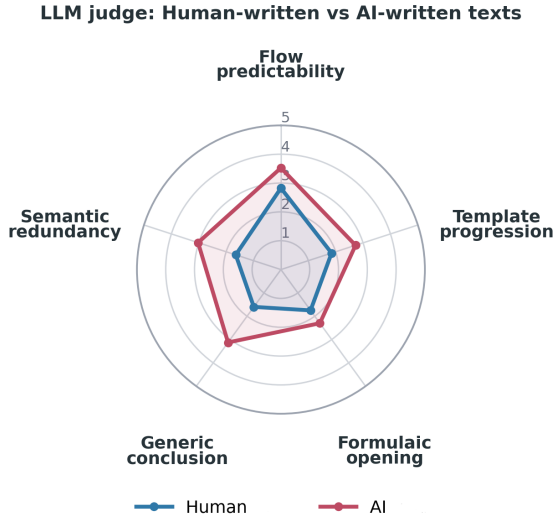


Figure 3: Rubric-based LLM-assisted evaluation of semantic-level signals. The radar chart shows mean scores for human-written and AI-generated texts across five semantic-level criteria.

latter result should not be interpreted as evidence that these datasets are shortcut-free. The stronger evidence comes from the PLM+LR and detector-level experiments, where models have access to broader textual information beyond the controlled surface features. The remaining PLM+LR performance after control further indicates that shortcut control reduces, but does not eliminate, other semantic, topical, or discourse-level signals.

3.4 Semantic-level Signals

LLM-assisted Evaluation The preceding analyses focus on surface features and embedding-based separability. To obtain diagnostics of semantic-content flow, we conduct rubric-based LLM evaluation using five semantic-level criteria (e.g., flow predictability, template progression, formulaic open-

ing, generic conclusion, and semantic redundancy). Each criterion is scored on a 1–5 scale, where higher scores indicate stronger presence of the corresponding pattern. We use Grok 4.3 (xAI, 2026) as the LLM evaluation model because it is not part of the generator model families used to produce AIGTs. Since LLM evaluation is not a substitute for human judgment, we also conduct a calibration study using a balanced sample of 80 texts scored by two human evaluators. The two evaluators show moderate-to-high agreement, and the LLM scores align well with their averaged ratings while reproducing AI–human gaps in the same direction across all five criteria. Detailed prompts and calibration results are provided in Appendix C.

Figure 3 shows that AIGTs receive higher mean scores than human-written texts across all five semantic-level criteria. The largest gaps appear in generic conclusion and semantic redundancy, suggesting that AIGTs more often end with broad, generic closing patterns and repeat similar meanings. These results indicate that human-written and AI-generated texts may differ not only in surface-level features, but also in semantic-level organization. This suggests that semantic-level differences constitutes an essential signal for AIGT detection, capturing structural tendencies beyond surface-level features. Detailed quantitative results and distributions are provided in Appendix C.2.

4 Results and Analysis: Detector Sensitivity to Shortcut Control

We evaluate the robustness of existing AIGT detectors to reductions in easily measurable shortcut signals. Using the paired corpus described in Section 3.1, we compare each detector under the orig-

Method	AUROC												AUPRC											
	XSum			SQuAD			WritingPrompts			PubMedQA			XSum			SQuAD			WritingPrompts			PubMedQA		
	Orig.	Ctrl.	Δ	Orig.	Ctrl.	Δ	Orig.	Ctrl.	Δ	Orig.	Ctrl.	Δ	Orig.	Ctrl.	Δ	Orig.	Ctrl.	Δ	Orig.	Ctrl.	Δ	Orig.	Ctrl.	Δ
<i>Zero-shot / Scoring-based detectors</i>																								
DetectGPT	0.670	0.361	-0.309	0.585	0.400	-0.185	0.842	0.646	-0.196	0.655	0.448	-0.207	0.646	0.414	-0.232	0.547	0.430	-0.117	0.854	0.645	-0.209	0.597	0.442	-0.155
DetectLLM-LRR	0.717	0.480	-0.237	0.655	0.390	-0.265	0.616	0.392	-0.225	0.671	0.452	-0.219	0.717	0.492	-0.225	0.657	0.434	-0.223	0.642	0.433	-0.209	0.644	0.473	-0.171
DetectLLM-NPR	0.667	0.398	-0.269	0.554	0.399	-0.156	0.824	0.601	-0.223	0.605	0.430	-0.175	0.639	0.439	-0.200	0.528	0.430	-0.098	0.829	0.598	-0.231	0.551	0.428	-0.123
FastDetectGPT	0.799	0.485	-0.314	0.752	0.452	-0.300	0.747	0.535	-0.212	0.675	0.371	-0.303	0.823	0.530	-0.293	0.785	0.494	-0.292	0.797	0.598	-0.199	0.665	0.427	-0.238
Binoculars	0.885	0.583	-0.302	0.798	0.549	-0.249	0.839	0.626	-0.213	0.786	0.470	-0.317	0.904	0.631	-0.273	0.801	0.576	-0.225	0.873	0.712	-0.161	0.808	0.506	-0.303
Avg.	0.748	0.461	-0.286	0.669	0.438	-0.231	0.774	0.560	-0.214	0.678	0.434	-0.244	0.746	0.501	-0.245	0.664	0.473	-0.191	0.799	0.597	-0.202	0.653	0.455	-0.198
<i>Training-based detectors</i>																								
RAIDAR	0.884	0.718	-0.167	0.889	0.701	-0.188	0.883	0.647	-0.236	0.763	0.639	-0.123	0.887	0.696	-0.192	0.888	0.725	-0.164	0.894	0.636	-0.258	0.748	0.605	-0.143
L2D	0.964	0.895	-0.069	0.857	0.842	-0.015	0.933	0.810	-0.122	0.859	0.792	-0.067	0.966	0.886	-0.079	0.865	0.824	-0.041	0.932	0.783	-0.149	0.866	0.787	-0.079
MoSEs-Logistic	0.931	0.782	-0.150	0.898	0.769	-0.130	0.960	0.845	-0.114	0.793	0.682	-0.111	0.945	0.789	-0.156	0.924	0.710	-0.214	0.962	0.877	-0.085	0.763	0.683	-0.080
MoSEs-XGBoost	0.964	0.810	-0.154	0.924	0.782	-0.141	0.988	0.922	-0.066	0.843	0.767	-0.076	0.970	0.819	-0.151	0.946	0.730	-0.217	0.989	0.925	-0.064	0.790	0.719	-0.071
BiScope	0.992	0.915	-0.077	0.948	0.878	-0.070	0.997	0.960	-0.037	0.939	0.832	-0.107	0.992	0.910	-0.081	0.948	0.881	-0.067	0.997	0.956	-0.041	0.933	0.838	-0.095
Avg.	0.947	0.824	-0.123	0.903	0.794	-0.109	0.952	0.837	-0.115	0.839	0.742	-0.097	0.952	0.820	-0.132	0.914	0.774	-0.141	0.955	0.835	-0.119	0.820	0.726	-0.094

Table 4: AUROC and AUPRC of baseline detectors under the original (Orig.) and shortcut-controlled (Ctrl.) settings. Δ is computed as Ctrl.—Orig. Detectors are grouped into zero-shot/scoring-based and training-based methods; Avg. rows report per-column averages within each detector group.

inal and shortcut-controlled benchmark settings. This comparison quantifies the change in detector performance when formatting and lexical shortcuts are explicitly weakened. To assess whether the observed trends generalize beyond this corpus, we repeat the analysis on a six-domain subset of RAID (Dugan et al., 2024). Detailed results are provided in Appendix D.

4.1 Experimental Setup

We evaluate representative detectors from two broad families. The first group consists of zero-shot or scoring-based detectors, including DetectGPT (Mitchell et al., 2023), DetectLLM-LRR, DetectLLM-NPR (Su et al., 2023), FastDetectGPT (Bao et al., 2024), and Binoculars (Hans et al., 2024). The second group consists of training-based detectors, including RAIDAR (Mao et al., 2024), L2D (Zhou et al., 2026), MoSEs-Logistic, MoSEs-XGBoost (Wu et al., 2025b), and BiScope (Guo et al., 2024). For each detector, we compare the original benchmark results with the shortcut-controlled results.

4.2 Detector-level Results and Analysis

Table 4 shows that shortcut control reduces detector performance across both detector families, but the magnitude of degradation differs substantially. Zero-shot and scoring-based detectors show consistently large average drops across all datasets, average AUROC drops range from -0.214 to -0.286. Training-based detectors show smaller but still consistently negative average drops, with AUROC decreases from -0.097 to -0.123. Thus, shortcut sensitivity is not confined to a single detector type, but its magnitude depends on how each detector extracts and combines evidence.

Zero-shot and Scoring-based Detectors Zero-shot and scoring-based detectors show larger performance drops because they rely on language-model scoring signals (e.g., token likelihoods, ranks, perturbation responses). These signals can be highly responsive to shortcut regularities (e.g., formulaic transitions, lexical smoothing). When shortcuts are reduced, the score distributions used by these methods become less separable.

Training-based Detectors RAIDAR, L2D, MoSEs-Logistic, and MoSEs-XGBoost show smaller degradation than scoring-based methods, with MoSEs-XGBoost showing some of the most moderate drops overall. BiScope achieves the strongest performance under both settings, yet still exhibits consistent negative Δ values after shortcut control. L2D is especially robust, because it learns an adaptive reconstruction distance rather than relying on fixed heuristic scores. However, the dominant trend is that performance tends to decline once formatting and lexical shortcuts are controlled, regardless of the detector’s underlying method. These results indicate that existing AIGT benchmark performance is partially supported by shortcut signals.

4.3 Dataset-level Results and Analysis

Furthermore, our analysis reveals that dataset-level patterns are heterogeneous. XSum shows large drops for most zero-shot detectors, suggesting that the original XSum setting contains shortcut signals that strongly support detection. SQuAD also shows substantial degradation for several methods, especially FastDetectGPT and DetectLLM-LRR. WritingPrompts and PubMedQA show more method-dependent behavior: some detectors de-

grade sharply, while training-based methods remain relatively stable. These results indicate that shortcut effects are not uniform across datasets, but interact with text type, domain structure, and generation setting.

4.4 Summary of Findings

The main experiment yields three findings. First, many existing AIGT detectors degrade when shortcut signals are reduced, indicating that original benchmark performance is partially driven by shortcut signals. Second, shortcut sensitivity varies by detector family: zero-shot detectors tend to show larger drops, while some training-based detectors are more stable. Third, shortcut sensitivity is dataset-dependent. These findings motivate benchmark designs that control shortcuts.

5 The AIGT-SIGNALBENCH

The analyses in Sections 3 and 4 show that conventional AIGT evaluations can be supported by measurable surface shortcuts. We introduce AIGT-SIGNALBENCH constructed to minimize avoidable shortcuts while preserving realistic writing tasks. The benchmark evaluates detectors whether detectors trained on conventional data remain reliable under shortcut controlled and realistic AI-assisted writing.

5.1 Benchmark Construction

Design principle AIGT-SIGNALBENCH is not designed to be merely a more difficult dataset. It is designed around two evaluation requirements. First, detectors should be evaluated under conditions where easily measurable shortcuts are reduced. Shortcuts can support high detection performance without requiring detectors to capture more robust AIGT signals. Second, detectors should be evaluated in realistic AI-assisted writing conditions. AIGT may be edited by an LLM or a human before use (Padmakumar and He, 2024; Durandard et al., 2025).

Dataset Composition AIGT-SIGNALBENCH covers seven writing settings: essay writing (Chillies, 2024), abstract writing (Clement et al., 2019), blog post writing (Kamaljp, 2023), product review writing (Hou et al., 2024), peer review writing (Kang et al., 2018), meeting summarization (Hu et al., 2023), and non-English-to-English translation (HuggingFaceFW, 2026).¹ For each domain–

model combination, we construct 30 sampled texts divided into three variants: `raw_ai`, `ai_edited`, and `human_edited`. The final dataset contains 1,260 sampled texts, each paired with a human-written reference. The detailed dataset composition is provided in Appendix E.1.

Shortcut-controlled Generation The key construction step is reference-matched generation. For each source instance, we extract structural information from the paired human reference (e.g., approximate length, paragraph structure). This information is incorporated into the generation prompt together with the task source. The goal is to prevent AIGT from being distinguishable simply because it is longer, more polished, or more formulaic than human-written text. Shortcut-controlled generation prompt is provided in Appendix E.2.

Realistic Editing Variants Each AIGT is constructed under three production conditions. The `raw_ai` variant contains text generated directly from the task source. The `ai_edited` variant models a scenario in which an LLM edits its own output. The `human_edited` variant models a more realistic human–AI writing workflow in which a human editor edits the AI-generated draft. The editing prompt is provided in Appendix E.2.

5.2 Evaluation and Results

Shortcut-control Validation We first verify whether AIGT-SIGNALBENCH reduces easy separability between human-written and AI-generated text. Using the same linear separability setup as in Section 3.3, we trained Shortcut-only LR and PLM+LR classifiers and evaluate each variant. On AIGT-SIGNALBENCH, both probes show lower separability than in the existing benchmark setting. The `human_edited` variant yields the lowest separability, indicating that human post-editing further weakens easily measurable shortcut signals. Detailed variant-level results are provided in Appendix E.4.

Baseline Detector Validation We next evaluate training-based detectors on AIGT-SIGNALBENCH. Table 5 shows that detector performance is generally lower than in the existing benchmark setting. L2D achieves the strongest overall AUROC, but its performance is lower than its results on the existing benchmark setting. BiScope remains competitive in AUPRC.

RAIDAR and the MoSEs variants show weaker performance. These results suggest that shortcut-

¹All datasets used in this work are publicly available for research purposes.

Baseline	AUROC				AUPRC			
	Overall	raw_ai	ai_edited	human_edited	Overall	raw_ai	ai_edited	human_edited
RAIDAR	0.759	0.655	0.752	0.682	0.719	0.608	0.744	0.646
L2D	0.883	0.878	0.901	0.873	0.863	0.861	0.901	0.849
MoSEs-Logistic	0.613	0.602	0.648	0.590	0.636	0.634	0.658	0.616
MoSEs-XGBoost	0.629	0.612	0.661	0.615	0.632	0.622	0.643	0.629
BiScope	0.802	0.803	0.808	0.795	0.864	0.862	0.869	0.862

Table 5: AUROC and AUPRC of baseline detectors on AIGT-SIGNALBENCH across editing conditions. Overall reports aggregate performance across all conditions; raw_ai, ai_edited, and human_edited denote the three benchmark variants.

controlled generation and editing variants reduce the usefulness of surface-oriented signals that detectors could exploit in easier benchmark settings.

Benchmark transfer Validation We further evaluate a transfer setting in which training-based detectors are trained on conventional benchmark data and tested on AIGT-SIGNALBENCH. This experiment asks whether detectors trained on raw-generation benchmarks generalize to shortcut-controlled and post-edited AI-assisted writing. The results show low performance across all baselines, with especially weak performance on the human_edited variant. Detailed results are provided in Appendix E.5.

Mixed-Training Validation We also conduct mixed-training experiments by replacing varying proportions of the RAID (Dugan et al., 2024) training set with AIGT-SIGNALBENCH examples. With a 30% mix, the average AUROC increases from 0.675 to 0.728 when evaluated on AIGT-SIGNALBENCH, while remaining high at 0.934 when evaluated on RAID. These results suggest that AIGT-SIGNALBENCH serves not only as a robustness evaluation benchmark but also as a source of complementary hard-case supervision for developing more robust detectors. Detailed experiment results are provided in Appendix E.6.

These results show that AIGT-SIGNALBENCH reduces easy surface separability and provides a complementary setting for evaluating detector behavior under edited AI-assisted writing conditions.

6 Related Work

6.1 Linguistic and Stylistic Signatures of AIGT

Prior studies have shown that human-written and AI-generated texts differ across multiple linguistic and stylistic dimensions (Zamaraeva et al., 2025). These differences appear at the levels of morphol-

ogy, syntax, semantics, and emotionality, with model-generated text often showing lower variability and stronger homogenization patterns (Zanotto and Aroyehun, 2025). LLM outputs also exhibit grammatical and rhetorical signatures such as nominalizations, present participial clauses, and phrasal coordination, with stronger effects in instruction-tuned models (Reinhart et al., 2025). O’Sullivan (2025) likewise shows that AI-generated creative writing forms denser clusters than human-authored writing, suggesting detectable stylistic signatures. These studies demonstrate that AIGT exhibits systematic linguistic and stylistic regularities.

However, their primary goal is to characterize human–AI text differences, not to examine whether such regularities function as shortcuts in detection tasks. Our work reinterprets these differences as potential shortcut signals and analyzes the extent to which detectors rely on shortcuts.

6.2 LLM-assisted Writing and Text Homogenization

Recent studies examine how LLM assistance shapes the writing process itself. LLM-assisted writing can reduce lexical and content diversity, making essays written by different users more homogeneous (Padmakumar and He, 2024). In QA settings, AI-generated answers show stronger semantic alignment with questions, while human-written answers exhibit stronger stylistic alignment (Durandard et al., 2025). These findings suggest that LLM involvement introduces systematic regularities in content, style, and semantic alignment. These regularities are directly relevant to AIGT detection, as real AIGT rarely appears as raw model output alone (Mysore et al., 2025). It may be revised by humans or produced through iterative human–AI collaboration.

However, many existing AIGT benchmarks are built around comparatively simplified generation

settings, leaving open the question of how detectors behave under more realistic AI-assisted writing conditions (Li et al., 2024; Wang et al., 2024; Dugan et al., 2024). Our work builds on this line of research by examining whether such regularities function as shortcut signals in detector evaluation, and by introducing AIGT-SIGNALBENCH, which covers `raw_ai`, `ai_edited`, and `human_edited` variants.

6.3 Detector Robustness under Stylistic and Editing Shifts

Recent studies have also examined the brittleness of AIGT detectors under stylistic variation and realistic task conditions. Detector performance can be highly sensitive to writing style, with some classifiers degrading to random-level performance under such variation (Doughman et al., 2025). Detectors trained on generic generation data can also struggle with task-specific Wikipedia editing practices and overfit superficial artifacts of machine generation (Quaremba et al., 2026). Our work builds on these observations by explicitly decomposing detection signals into formatting and writing-style, lexical usage, and semantic-content flow signals, quantifying the role of easily controllable surface shortcuts, and constructing a shortcut-controlled benchmark that distinguishes raw AI generation, AI-edited text, and human-edited AI text.

7 Conclusion

This study analyzes AIGT detection through the textual signals that distinguish human-written and AI-generated text. We show that human-written and AI-generated text differ not only in semantic-content flow but also in easily measurable shortcut signals, including paragraph structure and lexical markers. Our diagnostic experiments demonstrate that simple shortcut features alone can support strong classification performance. We further show that many existing AIGT detectors degrade under shortcut-controlled evaluation on existing benchmarks, suggesting that existing benchmark performance can be influenced by easily measurable shortcut signals. Motivated by these findings, we introduce AIGT-SIGNALBENCH, a shortcut-controlled benchmark designed to reflect realistic AI-assisted writing across diverse domains. Our work highlights the need to evaluate AIGT detectors beyond surface artifacts and provides a benchmark for studying more robust detection methods under realistic AI-assisted writing scenarios.

Limitations

This study has several limitations. First, although AIGT-SIGNALBENCH includes multiple domains and editing variants, its scale is smaller than large-scale AIGT benchmarks. Second, while we define and reduce several formatting and lexical shortcuts, we do not claim to remove all possible shortcuts. Topic distribution, source structure, prompt design, and domain-specific conventions may also affect detector performance. Third, our semantic-content flow analysis remains exploratory. LLM-assisted evaluation with human calibration suggests that AIGT may exhibit more predictable and template-like discourse patterns, but what deeper semantic or discourse-level signals detectors should rely on after surface shortcuts are controlled remains an open question. Future work should expand shortcut-controlled datasets and develop detection methods that move beyond surface artifacts.

Ethical Considerations

AIGT detection can support transparency, provenance tracking, and the curation of future training corpora. However, detector outputs should not be used as sole evidence for punitive decisions about individual writers. Our benchmark is intended for research on corpus-level and system-level evaluation rather than individual authorship accusation. We also emphasize that AI-assisted writing may involve legitimate human use, editing, and collaboration. Future releases of AIGT-SIGNALBENCH should document data sources, generation prompts, editing instructions, and known limitations to support responsible use.

Acknowledgements

We thank the anonymous reviewers for their helpful comments. This work was supported by the Technology Innovation Program (RS-2024-00398858, Development of AI-based urban flood damage risk prediction and evaluation technology for practical use) funded by the Ministry of the Interior and Safety (MOIS, Korea), and by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (RS-2026-25556931).

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Anthropic. 2026. [System card: Claude Sonnet 4.6](#). February 17, 2026. Accessed: 2026-05-24.
- Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2024. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In *The Twelfth International Conference on Learning Representations*.
- chillies. 2024. [IELTS-writing-task-2-evaluation](#). Hugging Face dataset.
- Colin B. Clement, Matthew Bierbaum, Kevin P. O’Keeffe, and Alexander A. Alemi. 2019. [On the use of ArXiv as a dataset](#). *Preprint*, arXiv:1905.00075.
- Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd edition. Lawrence Erlbaum Associates.
- Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Naveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.
- D. R. Cox. 1958. [The regression analysis of binary sequences](#). *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232.
- Jesse Davis and Mark Goadrich. 2006. [The relationship between Precision-Recall and ROC curves](#). In *Proceedings of the 23rd International Conference on Machine Learning*, pages 233–240. ACM.
- DeepSeek-AI. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. [Documenting large webtext corpora: A case study on the colossal clean crawled corpus](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Jad Doughman, Osama Mohammed Afzal, Hawau Olamide Toyin, Shady Shehata, Preslav Nakov, and Zeerak Talat. 2025. [Exploring the limitations of detecting machine-generated text](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4274–4281, Abu Dhabi, UAE. Association for Computational Linguistics.
- Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 2024. [RAID: A shared benchmark for robust evaluation of machine-generated text detectors](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12463–12492, Bangkok, Thailand. Association for Computational Linguistics.
- Noé Durandard, Saurabh Dhawan, and Thierry Poibeau. 2025. [LLMs stick to the point, humans to style: Semantic and stylistic alignment in human and LLM communication](#). In *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 206–213, Avignon, France. Association for Computational Linguistics.
- Angela Fan, Mike Lewis, and Yann Dauphin. 2018. [Hierarchical neural story generation](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.
- Tom Fawcett. 2006. [An introduction to ROC analysis](#). *Pattern Recognition Letters*, 27(8):861–874.
- Hanxi Guo, Siyuan Cheng, Xiaolong Jin, Zhuo Zhang, Kaiyuan Zhang, Guanhong Tao, Guangyu Shen, and Xiangyu Zhang. 2024. [BiScope: Ai-generated text detection by checking memorization of preceding tokens](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 104065–104090.
- Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. [Spotting LLMs with binoculars: Zero-shot detection of machine-generated text](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 17519–17537. PMLR.
- Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2024. [Bridging language and items for retrieval and recommendation](#). *arXiv preprint arXiv:2403.03952*.
- Yebowen Hu, Timothy Ganter, Hanieh Deilamsalehy, Franck Dernoncourt, Hassan Foroosh, and Fei Liu. 2023. [MeetingBank: A benchmark dataset for meeting summarization](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16409–16423, Toronto, Canada. Association for Computational Linguistics.

- HuggingFaceFW. 2026. [FineTranslations: The world's knowledge in 1+1t tokens of parallel text](#). Hugging Face dataset.
- Hazem Ibrahim, Fengyuan Liu, Rohail Asim, et al. 2023. [Perception, performance, and detectability of conversational artificial intelligence across 32 university courses](#). *Scientific Reports*, 13:12187.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. [PubMedQA: A dataset for biomedical research question answering](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.
- Kamaljp. 2023. [medium_articles](#). Hugging Face dataset.
- Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. [A dataset of peer reviews \(PeerRead\): Collection, insights and NLP applications](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1647–1661, New Orleans, Louisiana. Association for Computational Linguistics.
- John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. [A watermark for large language models](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 17061–17084. PMLR.
- Mina Lee, Percy Liang, and Qian Yang. 2022. [Coauthor: Designing a human-AI collaborative writing dataset for exploring language model capabilities](#). In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–19. Association for Computing Machinery.
- Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. 2024. [MAGE: Machine-generated text detection in the wild](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–53, Bangkok, Thailand. Association for Computational Linguistics.
- Chengzhi Mao, Carl Vondrick, Hao Wang, and Junfeng Yang. 2024. [Raidar: generative AI detection via rewriting](#). In *The Twelfth International Conference on Learning Representations*.
- Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. [Detectgpt: Zero-shot machine-generated text detection using probability curvature](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 24950–24962. PMLR.
- Sheshera Mysore, Debarati Das, Hancheng Cao, and Bahareh Sarrafzadeh. 2025. [Prototypical human-AI collaboration behaviors from LLM-assisted writing in the wild](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16819–16846, Suzhou, China. Association for Computational Linguistics.
- Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. [Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.
- OpenAI. 2023. [Introducing apis for GPT-3.5 Turbo and Whisper](#). Updated April 24, 2024. Accessed: 2026-05-24.
- OpenAI. 2026. [GPT-5.5 system card](#). Published April 23, 2026; updated April 24, 2026. Accessed: 2026-05-24.
- James O'Sullivan. 2025. [Stylometric comparisons of human versus AI-generated creative writing](#). *Humanities and Social Sciences Communications*, 12(1):1708.
- Vishakh Padmakumar and He He. 2024. [Does writing with language models reduce content diversity?](#) In *The Twelfth International Conference on Learning Representations*.
- Gerrit Quaremba, Elizabeth Black, Denny Vrandečić, and Elena Simperl. 2026. [TSM-Bench: Detecting LLM-generated text in real-world Wikipedia editing practices](#). *arXiv preprint arXiv:2605.31113*.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. [SQuAD: 100,000+ questions for machine comprehension of text](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-bert: Sentence embeddings using siamese bert-networks](#). In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992.
- Alex Reinhard, Ben Markey, Michael Laudenbach, Kachad Pantusen, Ronald Yurko, Gordon Weinberg, and David West Brown. 2025. [Do LLMs write like humans? variation in grammatical and rhetorical styles](#). *Proceedings of the National Academy of Sciences of the United States of America*, 122(8):e2422455122.
- Jenna Russell, Marzena Karpinska, and Mohit Iyyer. 2025. [People who frequently use chatgpt for writing tasks are accurate and robust detectors of ai-generated text](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5342–5373.

- Anastassia Shaitarova, Nikolaj Bauer, Jannis Vamvas, and Martin Volk. 2024. Tracing linguistic footprints of chatgpt across tasks, domains and personas in english and german. In *Proceedings of the 9th edition of the Swiss Text Analytics Conference*, pages 102–112.
- Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. [AI models collapse when trained on recursively generated data](#). *Nature*, 631(8022):755–759.
- Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, Miles McCain, Alex Newhouse, Jason Blazakis, Kris McGuffie, and Jasmine Wang. 2019. [Release strategies and the social impacts of language models](#). *arXiv preprint arXiv:1908.09203*.
- Jinyan Su, Terry Zhuo, Di Wang, and Preslav Nakov. 2023. [DetectLLM: Leveraging log rank information for zero-shot detection of machine-generated text](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12395–12412, Singapore. Association for Computational Linguistics.
- Edward Tian and Alexander Cui. 2023. [GPTZero: Towards detection of AI-generated text using zero-shot and supervised methods](#). Accessed: 2026-05-25.
- Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, Thomas Arnold, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024. [M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1369–1407, St. Julian’s, Malta. Association for Computational Linguistics.
- Debora Weber-Wulff, Alla Anohina-Naumeca, Sonja Bjelobaba, Tomáš Foltýnek, Jean Guerrero-Dib, Olu-mide Popoola, Petr Šigut, and Lorna Waddington. 2023. [Testing of detection tools for ai-generated text](#). *International Journal for Educational Integrity*, 19:26.
- Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Charlotte Biles, Sasha Brown, Zach Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, et al. 2022. [Taxonomy of risks posed by language models](#). In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 214–229. Association for Computing Machinery.
- Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. 2025a. [A survey on LLM-generated text detection: Necessity, methods, and future directions](#). *Computational Linguistics*, 51(1):275–338.
- Junxi Wu, Jinpeng Wang, Zheng Liu, Bin Chen, Dongjian Hu, Hao Wu, and Shu-Tao Xia. 2025b. [MoSEs: Uncertainty-aware AI-generated text detection via mixture of stylistics experts with conditional thresholds](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5786–5805, Suzhou, China. Association for Computational Linguistics.
- xAI. 2026. Grok 4.3. <https://docs.x.ai/developers/models/grok-4.3>. XAI developer documentation. Accessed: 2026-05-26.
- Kai-Cheng Yang and Filippo Menczer. 2023. [Anatomy of an AI-powered malicious social botnet](#). *arXiv preprint arXiv:2307.16336*.
- Olga Zamaraeva, Dan Flickinger, Francis Bond, and Carlos Gómez-Rodríguez. 2025. Comparing llm-generated and human-authored news text using formal syntactic theory. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9041–9060.
- Sergio E. Zanutto and Segun Aroyehun. 2025. [Linguistic and embedding-based profiling of texts generated by humans and large language models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22841–22858, Suzhou, China. Association for Computational Linguistics.
- Hongyi Zhou, Jin Zhu, Kai Ye, Ying Yang, Erhan Xu, and Chengchun Shi. 2026. [Learn-to-distance: Distance learning for detecting LLM-generated text](#). In *The Fourteenth International Conference on Learning Representations*.

A Fine-Grained Signal Selection and Measurement

A.1 Count- and Frequency-Based Surface Measurements

Table 7 reports the measurement procedure for each candidate surface feature. For the lexical features, the word and phrase lists draw on prior work on discourse connectives and recurring lexical and formulaic patterns in human-written and AI-generated text (Shaitarova et al., 2024; Russell et al., 2025).

A.2 Candidate Feature and Selection Criteria

Feature	H mean	H var	H std	AI mean	AI var	AI std	AI-H	d	d_z
Avg. paragraph length	115.139	2664.474	51.619	71.018	2070.709	45.505	-44.122	-0.907	-0.693
Paragraph count	1.000	0.000	0.000	2.550	5.878	2.424	+1.550	0.904	0.639
Commas / 100w	4.616	10.316	3.212	6.481	8.846	2.974	+1.864	0.602	0.503
Numbers / 100w	1.667	6.022	2.454	0.875	2.300	1.517	-0.793	-0.389	-0.404
Transition words / 100w	0.148	0.226	0.475	0.275	0.370	0.608	+0.127	0.232	0.176
Em-dashes / 100w	0.053	0.090	0.301	0.140	0.231	0.480	+0.088	0.219	0.170
Cliché phrases / 100w	0.003	0.004	0.066	0.028	0.025	0.158	+0.025	0.210	0.164
Sent. length variance CV	0.463	0.053	0.230	0.422	0.041	0.204	-0.042	-0.191	-0.152
Vague intensifiers / 100w	0.083	0.179	0.423	0.160	0.290	0.539	+0.077	0.159	0.117
Type-token ratio	0.726	0.010	0.101	0.723	0.015	0.124	-0.003	-0.022	-0.026
Hapax ratio	0.799	0.005	0.073	0.799	0.009	0.095	+0.000	0.004	0.003
Zipf deviation	2.177	3.026	1.740	2.265	4.396	2.097	+0.088	0.046	0.038
Modal verbs / 100w	1.095	1.934	1.391	0.988	2.075	1.441	-0.108	-0.076	-0.065
Semicolons / 100w	0.095	0.331	0.575	0.056	0.084	0.290	-0.039	-0.085	-0.068
Parentheses / 100w	0.344	1.115	1.056	0.323	1.089	1.044	-0.022	-0.021	-0.016

Table 6: Descriptive statistics for the full candidate set of count- and frequency-based surface features. The first nine features, ranked by the largest absolute standardized human–AI gaps, are used as the main surface shortcut signals in Table 1.

We first constructed a broader set of count- and frequency-based surface features to capture formatting, writing-style in human-written and AI-generated texts. Table 6 reports descriptive statistics for this full feature set. We did not use all candidate features as main detection signals. Instead, we selected the shortcut signals that showed the largest absolute standardized human–AI gaps in the paired corpus.

Specifically, the surface shortcut signals in Table 1 correspond to the top-ranked features in Table 6, from average paragraph length to vague intensifiers per 100 words. This selection keeps the main analysis focused on features that show clear empirical differences and can be measured with simple count- or frequency-based procedures. Features with smaller human–AI differences, such as type-token ratio, hapax ratio, Zipf deviation, modal verbs, semicolons, and parentheses, are reported in Table 6 for transparency but are not included in the main shortcut-signal set.

Beyond empirical separability, the selected surface signals also had to meet two practical requirements. First, they had to be directly observable from the text without semantic interpretation. Second, they had to be plausibly reducible through explicit prompt constraints or post-processing. These requirements align with our operational definition of surface shortcuts in Section 2.2.

A.3 Semantic-Content Flow Criteria

Semantic-content flow signals are not selected from the count-based feature pool in Table 6. Instead, they are defined as semantic-level patterns that require semantic or rhetorical judgment. These signals include generic conclusions, semantic redundancy, formulaic openings, template-like progression, and flow predictability.

Because these signals cannot be reliably captured by simple frequency counts, we evaluate them separately through the rubric-based LLM-assisted procedure described in Appendix C. Thus, Appendix A distinguishes two types of fine-grained signal definitions: shortcut signals, which are selected from measurable count- and frequency-based features, and semantic-content flow signals, which are defined as higher-level semantic criteria.

Feature	Measurement	Pattern list or definition
<i>Formatting and writing-style features</i>		
Paragraph count	Number of paragraphs in the text.	Paragraphs are segmented using line breaks or blank-line boundaries.
Avg. paragraph length	Mean number of word tokens per paragraph.	$\frac{1}{ \mathcal{P} } \sum_{p \in \mathcal{P}} p $, where $\mathcal{P}$ is the set of paragraphs.
Sent. length variance CV	Coefficient of variation of sentence lengths.	$\sigma(\ell_1, \dots, \ell_m) / \mu(\ell_1, \dots, \ell_m)$, where ℓ_i is the word length of sentence i .
Commas / 100w	Frequency of comma characters, normalized per 100 words.	Number of comma characters multiplied by 100 and divided by the number of word tokens.
Em-dashes / 100w	Frequency of em-dash characters, normalized per 100 words.	Number of em-dash characters multiplied by 100 and divided by the number of word tokens.
Numbers / 100w	Frequency of numeric expressions, normalized per 100 words.	Matches digit-based expressions such as years, counts, dates, scores, and measurements.
Semicolons / 100w	Frequency of semicolon characters, normalized per 100 words.	Number of semicolon characters multiplied by 100 and divided by the number of word tokens.
Parentheses / 100w	Frequency of parenthesis characters, normalized per 100 words.	Number of parenthesis characters multiplied by 100 and divided by the number of word tokens.
<i>Lexical-usage features</i>		
Transition words / 100w	Frequency of predefined discourse transition markers, normalized per 100 words.	Examples include <i>however, moreover, furthermore, additionally, therefore, overall, and in conclusion</i> .
Vague intensifiers / 100w	Frequency of predefined vague or underspecified intensifiers, normalized per 100 words.	Examples include <i>significantly, substantially, considerably, extensively, invaluable, groundbreaking, and potential</i> .
Cliché phrases / 100w	Frequency of predefined formulaic or cliché phrases, normalized per 100 words.	Examples include <i>shed light on, underscore the importance, highlight the importance, pave the way, serve as a testament, and far-reaching implications</i> .
Modal verbs / 100w	Frequency of modal verbs, normalized per 100 words.	Includes <i>would, could, may, might, should, must, will, and can</i> .
Type-token ratio	Ratio of unique word types to total word tokens.	$ \mathcal{V} /N$, where $\mathcal{V}$ is the vocabulary of the text.
Hapax ratio	Ratio of words appearing only once to the number of unique word types.	$ \{w \in \mathcal{V} : \text{count}(w) = 1\} / \mathcal{V} $.
Zipf deviation	Deviation of the observed word-frequency distribution from an ideal Zipfian distribution.	Computed from ranked word frequencies; lower values indicate a more Zipf-like distribution.

Table 7: Measurement procedures for the candidate surface features used in Table 6. Features marked with “/100w” are normalized per 100 word tokens.

B Full Prompt Specifications

B.1 Prompt-Based Shortcut Control Prompt

To evaluate shortcut removal, we use the following conditional prompt structure. The shared prompt template is combined with dataset-specific guidance for each source dataset.

Shared Prompt Template (Common Base)

```
## Shared Prompt Template
You perform minimal shortcut removal on AIGT. Preserve input content and follow constraints exactly.
Apply minimal edits to SOURCE TEXT to reduce AI-shortcut signals while preserving the source facts and genre.
Keep the meaning, factual content, and narrative or explanatory flow close to SOURCE TEXT.
Change only expressions that are necessary for shortcut removal; do not paraphrase for style or improve the writing.
1) Keep meaning/content and narrative flow as close as possible to SOURCE TEXT.
    Do not introduce, remove, or reorder substantive content except where required by shortcut control.
2) Preserve the source mix of short, medium, and long sentence units without making the text mechanical.
3) Do not use these cliché/formulaic expressions: shed light on, underscore the importance, highlight the importance, marks a significant, paves the way, serves as a reminder, serves as a testament, serves as a beacon, significant milestone, it is worth noting, it is important to note, it is crucial to, it is essential to, it is widely regarded, it is generally considered, echoes through, reverberates across, brings together, testament to, far-reaching implications, far-reaching impact, breathing new life.
4) Avoid vague or hollow intensifier words unless concrete evidence is attached. Avoid words like: significantly, substantially, considerably, immensely, tremendously, widespread, extensively, comprehensively, invaluable, indispensable, potential.
5) Do not end with a vague future-looking, essay-like, or moralizing closing sentence.
6) Output only the shortcut-removed text. No preface, no bullets, no explanations, text only.
```

Dataset-specific conditional guidance. The following conditional segments are appended directly to the Shared Prompt Template based on the target task domain.

Dataset Guidance: XSum

```
### XSum
Transition-control hint:
transition control means avoiding formulaic AI-style connectors, not banning every marker. Keep a natural connector only when the source sentence relationship already supports it. Do not add Overall/Furthermore-style essay framing.
Dataset preservation guidance:
keep a plain news-reporting voice. Preserve dates, scores, counts, named people, teams, places, and quoted attribution from the source. Do not add a broad summary-style ending. Keep enough natural commas for attribution, apposition, dates, ages, and score clauses; do not turn a comma-rich news sentence into a set of clipped simple sentences. Do not force AI-skewed transition markers when the source does not need them, but do not erase a natural news contrast marker when the sentence relationship already supports it.
Dataset metric guidance:
the main risk is overcorrecting commas and transition markers. Aim for plain news prose with natural comma punctuation. A single news-natural contrast marker such as however is acceptable only when the source already contains a real contrast.
```

Dataset Guidance: SQuAD

```
### SQuAD
Transition-control hint:
transition control means avoiding formulaic AI-style connectors, not banning every marker. Keep a natural connector only when it is needed to preserve the source explanation. Do not add Overall/Furthermore-style essay framing.
```

Dataset preservation guidance:

keep the explanatory answer style, but avoid textbook-like connective framing. Preserve named entities, dates, quantities, and cause/effect facts. Use local cause/effect wording rather than formulaic openers. One natural connector is acceptable when it keeps an existing explanation coherent. Avoid Unicode dashes unless they are necessary to preserve source meaning.

Dataset metric guidance:

reduce comma-heavy textbook phrasing while keeping the answer clear. Avoid repeated formulaic transition openers, but do not strip every connector from explanatory prose. Keep dash usage at zero.

Dataset Guidance: WritingPrompts

WritingPrompts

Transition-control hint:

transition control mainly means avoiding repeated formulaic markers while preserving coherence.

Dataset preservation guidance:

preserve narrative rhythm, short sentences, fragments, and dialogue cadence. Do not polish the passage into an essay arc or add a moralizing final sentence. Keep abrupt beats, fragments, and longer descriptive stretches when they already exist in the source.

Dataset metric guidance:

sentence-length rhythm matters most. Preserve short beats, fragments, dialogue-like cadence, and occasional longer sentences; do not normalize the passage into evenly sized sentences.

Dataset Guidance: PubMed

PubMed

Transition-control hint:

transition control mainly means avoiding repeated formulaic markers while preserving coherence.

Dataset preservation guidance:

Preserve measurements, conditions, abbreviations, numerical values, and comparison terms. Do not compress away technical clauses, study conditions, mechanisms, or qualifiers just to sound cleaner. Keep the output close to the target length. Preserve the source sentence-count pattern when possible, and keep modest sentence-length variation by mixing concise technical statements with condition/detail clauses.

Dataset metric guidance:

do not shorten the text. Preserve all technical detail from the source, including numbers, abbreviations, conditions, comparison terms, and mechanism clauses. Preserve the source sentence-length variation by avoiding both over-splitting and over-smoothing.

B.2 Effect of Prompt-Based Shortcut Control

To verify that the shortcut-controlled setting actually reduces the intended shortcut signals, we recompute the same nine shortcut features after applying the prompt-based shortcut control procedure described in Appendix B.1. Table 8 reports the resulting feature statistics.

Feature	H mean	H var	H std	AI mean	AI var	AI std	AI-H	d	d_z
Avg. paragraph length	110.965	5130.627	71.628	112.266	5151.703	71.775	+1.301	0.018	0.196
Paragraph count	1.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000
Commas / 100w	5.031	11.726	3.424	5.903	13.015	3.608	+0.872	0.248	0.232
Numbers / 100w	1.651	7.728	2.780	0.907	3.074	1.753	-0.744	-0.320	-0.295
Transition words / 100w	0.207	0.421	0.649	0.154	0.274	0.524	-0.053	-0.090	-0.068
Em-dashes / 100w	0.064	0.127	0.357	0.001	0.001	0.029	-0.063	-0.250	-0.176
cliché phrases / 100w	0.009	0.023	0.151	0.019	0.017	0.132	+0.010	0.068	0.048
Sent. length variance CV	0.444	0.067	0.259	0.426	0.056	0.237	-0.018	-0.074	-0.062
Vague intensifiers / 100w	0.089	0.226	0.475	0.096	0.203	0.450	+0.007	0.015	0.011

Table 8: Descriptive statistics for the nine shortcut features after prompt-based shortcut control. Compared with the original statistics in Table 2, most formatting and lexical human–AI differences are substantially reduced, indicating that the control procedure directly weakens the targeted shortcut signals.

Feature-level analysis. Table 8 shows that the prompt-based control procedure reduces the intended shortcut signals. Relative to the original statistics in Table 2, the largest formatting differences are largely removed. Paragraph count becomes identical between human-written and AI-generated texts, and the

average paragraph-length gap changes from a large human–AI difference of -44.122 words to a negligible difference of +1.301 words.

Lexical and punctuation-based markers also show weaker Human–AI differences. Transition words, cliché phrases, sentence-length variance, and vague intensifiers show near-zero effect sizes after control. Em-dash usage is strongly reduced in AIGTs as well. Some residual gaps remain, especially for numerical expressions and comma usage, likely because the prompts preserve factual details, domain-specific conventions, and natural punctuation.

These feature-level changes help interpret the detector degradation observed under shortcut-controlled evaluation. Since the intervention directly narrows the measurable formatting and lexical cues targeted in our analysis, the performance drops should not be viewed simply as an artifact of unconstrained rewriting-induced distribution shift. Rather, the drops suggest that detectors had benefited from shortcuts that were present in the original benchmark setting, and that their performance decreases when those shortcuts are reduced. The controlled texts are not shortcut-free, but they provide a cleaner setting for evaluating whether detectors can rely on more robust semantic-content flow signals. All analyses reported here exclude samples for which shortcut control failed.

B.3 Content-Preservation Validation

Scope	Semantic similarity			Named entity		Numeric expression	
	MPNet Cos.	BERTScore-R	BERTScore-F1	NER-R	NER-F1	Number-R	Number-F1
Overall	0.9410	0.9393	0.9479	0.9194	0.9419	0.9916	0.9920
XSum	0.9715	0.9463	0.9549	0.9250	0.9409	0.9945	0.9939
SQuAD	0.9657	0.9489	0.9565	0.9202	0.9341	0.9926	0.9930
WritingPrompts	0.9194	0.9321	0.9428	0.9089	0.9446	0.9731	0.9796
PubMedQA	0.9077	0.9301	0.9375	0.9221	0.9561	0.9965	0.9959

Table 9: Semantic similarity, named-entity preservation, and numeric-expression preservation across datasets.

Prompt-based shortcut control could introduce a confound if the editing procedure changes the text beyond the targeted surface signals. We therefore evaluate content preservation using the same original-AI and shortcut-controlled AI pairs used in the linear-classification, PLM, and detector-level experiments.

For semantic preservation, we compute cosine similarity using sentence-transformers/all-mpnet-base-v2 embeddings and BERTScore recall and F1 using roberta-large. For named-entity preservation, we compute recall and F1 over entities extracted with spaCy, focusing on concrete factual types including PERSON, ORG, GPE, LOC, FAC, NORP, PRODUCT, EVENT, WORK_OF_ART, LAW, and LANGUAGE. For numeric-expression preservation, we compute recall and F1 over digit-based expressions extracted with regular expressions, including years, dates, counts, decimals, and percentages. For the named-entity and numeric-expression metrics, the entities and expressions in the original AI text serve as the reference set against which the shortcut-controlled text is compared.

Table 9 reports the results. The shortcut-controlled texts remain highly similar to the original AI texts across all four datasets. MPNet cosine similarity ranges from 0.9077 to 0.9715, BERTScore F1 ranges from 0.9375 to 0.9565, named-entity F1 ranges from 0.9341 to 0.9561, and numeric-expression F1 ranges from 0.9796 to 0.9959. These results indicate that the control procedure largely preserves semantic and factual content while reducing the targeted shortcut features. They do not imply that prompt-based editing introduces no changes, but they show that the intervention is not an unconstrained rewrite or aggressive content deletion.

C Additional LLM-Assisted Evaluation Results

This appendix provides additional details for the LLM-assisted semantic-content flow evaluation. The evaluation uses five semantic-content flow criteria: flow predictability, template progression, formulaic opening, generic conclusion, and semantic redundancy. Each criterion is scored on a 1–5 scale, where higher scores indicate stronger presence of the corresponding pattern.

C.1 LLM Evaluation Prompt

We use the following rubric-based prompt for LLM-assisted evaluation (see Table 10).

Flow predictability: whether the text follows a highly predictable or mechanical discourse flow.

Template progression: whether the text follows a fixed template-like structure.

Formulaic opening: whether the text begins with a generic or formulaic opening.

Generic conclusion: whether the text ends with a broad, vague, or essay-like conclusion.

Semantic redundancy: whether the text repeats similar ideas without adding substantive new information.

Return only the five scores and a brief justification for each criterion.

Criteria	LLM Eval prompt
5	Evaluate the following text using five discourse-structure criteria. Scoring rule for every criterion: - 1 = very low presence of this pattern - 2 = low presence - 3 = moderate or mixed presence - 4 = high presence - 5 = very high presence Criteria: 1. flow_predictability <i>Definition:</i> Whether the next sentence or next idea is easy to anticipate because the logical flow is overly smooth, conventional, or predictable. Score high when the text proceeds in a highly expected sequence with little surprising semantic movement. 2. template_progression <i>Definition:</i> Whether the information follows a repeated template such as background → explanation → implication → balanced conclusion, or problem → solution → benefit → limitation. Score high when the text feels organized around a generic reusable template. 3. formulaic_opening <i>Definition:</i> Whether paragraphs or major sentences begin with formulaic discourse markers, broad general statements, or conventional transition phrases. Examples include: Additionally, Furthermore, In conclusion, Moreover, Overall, This suggests, It is important to note, In recent years. If the text has only one paragraph, evaluate sentence openings and major transitions instead. 4. generic_conclusion <i>Definition:</i> Whether the ending is generic, safe, broadly balanced, or uses a common wrap-up pattern. Score high for endings that summarize cautiously, call for further research, give a broad implication, or close with a polished but generic final statement. 5. semantic_redundancy <i>Definition:</i> Whether the same meaning is repeated in different words, or whether the text has more sentences than the amount of new information warrants. Score high when the text repeats ideas, restates earlier claims, or expands with low semantic novelty. Important instructions: - Do not infer or mention authorship. - Do not use knowledge of the source, model, dataset, or label. - Do not reward or punish factual accuracy. - Use the full 1-5 scale when appropriate. - Scores must be integers. - Keep rationales short and evidence-based.

Table 10: LLM evaluation prompt for semantic-content flow analysis. The prompt instructs Grok 4.3 to evaluate five semantic-level criteria on a 1–5 scale.

C.2 Score Distributions of LLM Evaluation

Figure 4 shows the score distributions for human-written and AI-generated texts across the five semantic-content flow criteria. AIGTs receive higher mean scores than human-written texts on all criteria: flow

predictability, template progression, formulaic opening, generic conclusion, and semantic redundancy. The largest gaps appear for generic conclusion and semantic redundancy, suggesting that AIGTs more often contain broad closing statements and repeated semantic content.

The distributions also show that human-written texts are more concentrated in lower score ranges for most criteria, especially template progression, formulaic opening, generic conclusion, and semantic redundancy. AIGTs shift more mass toward mid-to-high scores. Flow predictability is less sharply separated, since some human-written texts also follow predictable discourse patterns, but the AIGTs still show a higher mean score. These results support the claim that AIGTs exhibit stronger semantic-content flow regularities.

LLM-Eval score distributions by criterion and text role

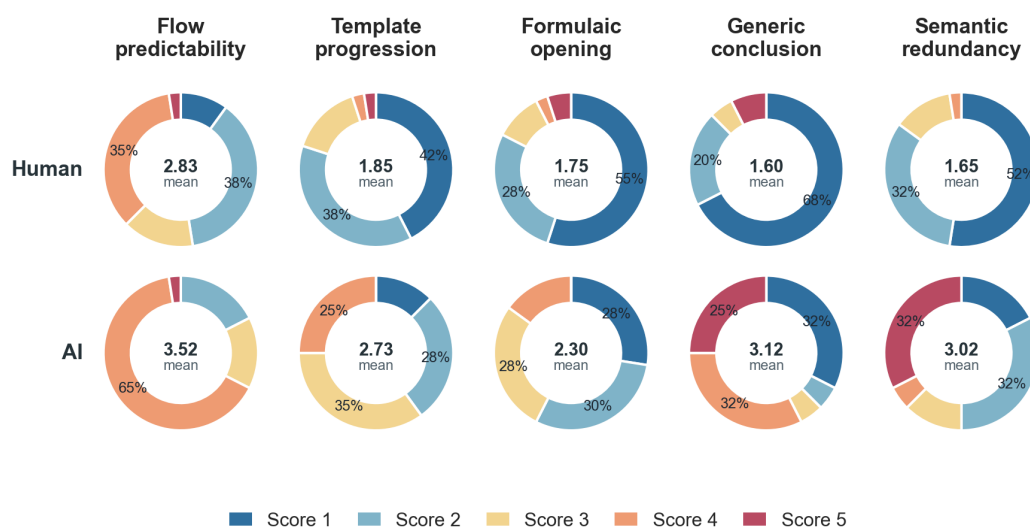


Figure 4: LLM-evaluated score distributions by criterion and text role.

C.3 Human Rater Agreement

To examine whether the rubric can be applied consistently by human evaluators, two human raters scored the same label-balanced calibration sample. Figure 5 reports agreement between the two raters using Pearson correlation, Spearman correlation, weighted kappa, and within-one agreement.

The two raters show moderate to high agreement across criteria. Spearman correlations range from 0.52 to 0.73, Pearson correlations from 0.53 to 0.76, and weighted kappa from 0.44 to 0.72. Within-one agreement is consistently high, ranging from 0.88 to 0.95, indicating that the two raters rarely differ by more than one point. Agreement is weakest for flow predictability, which is the most subjective criterion, but remains acceptable for calibration. These results support averaging the two human ratings for comparison with the LLM evaluator.

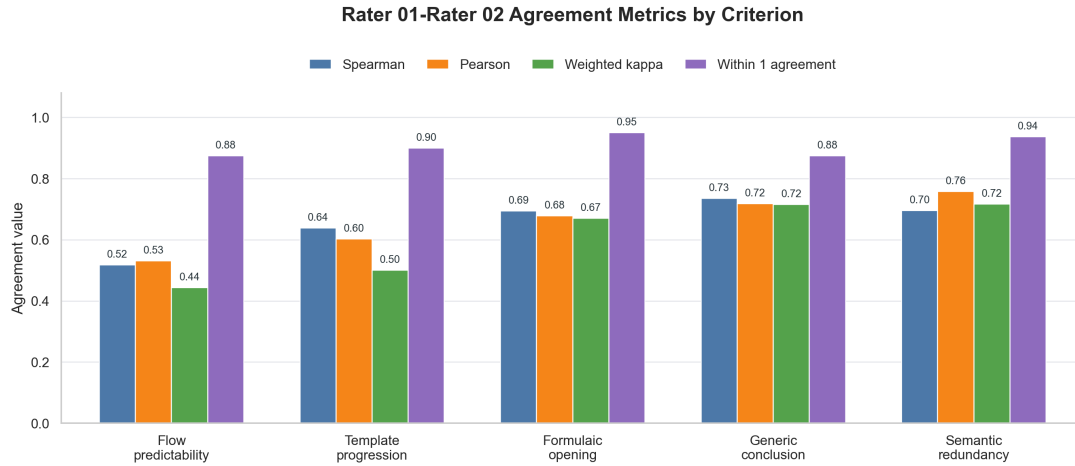


Figure 5: Agreement between the two human raters by criterion.

C.4 Calibration with Human Judgments

We next compare the LLM evaluator with the averaged human ratings. Figure 6 shows moderate to high alignment with human judgments across most criteria. Pearson correlations range from 0.67 to 0.82, and Spearman correlations range from 0.65 to 0.77. Alignment is strongest for generic conclusion and semantic redundancy, while template progression shows weaker exact agreement.

Although weighted kappa is lower for template progression, Pearson and Spearman correlations remain moderate, suggesting that the LLM evaluator captures the relative ordering better than exact score calibration for this criterion. Since weighted kappa is sensitive to exact ordinal-category agreement, we rely primarily on correlation and AI-human gap direction rather than exact agreement when interpreting this criterion.

Because the main purpose of this analysis is to identify whether AIGTs show stronger semantic-content flow patterns than human-written texts, we also examine directional consistency. Figure 7 compares the AI-human score gaps estimated by the LLM evaluator and by averaged human ratings. Both evaluations assign positive AI-human gaps for all five criteria, with similar magnitudes for flow predictability, template progression, generic conclusion, and semantic redundancy.

These calibration results suggest that the LLM-assisted evaluation is directionally aligned with human ratings. We therefore use it as a lightweight diagnostic for semantic-content flow patterns, rather than as a full substitute for human evaluation.

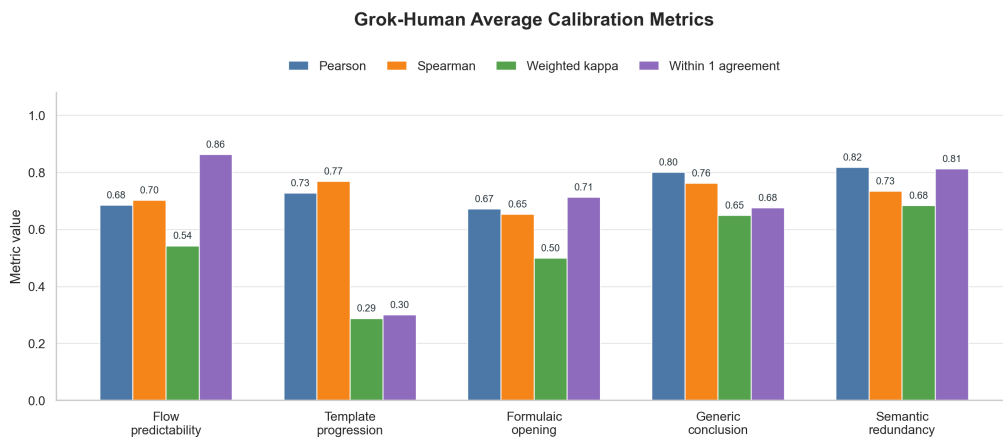


Figure 6: Calibration metrics between the LLM evaluator and averaged human ratings across semantic-content flow criteria.

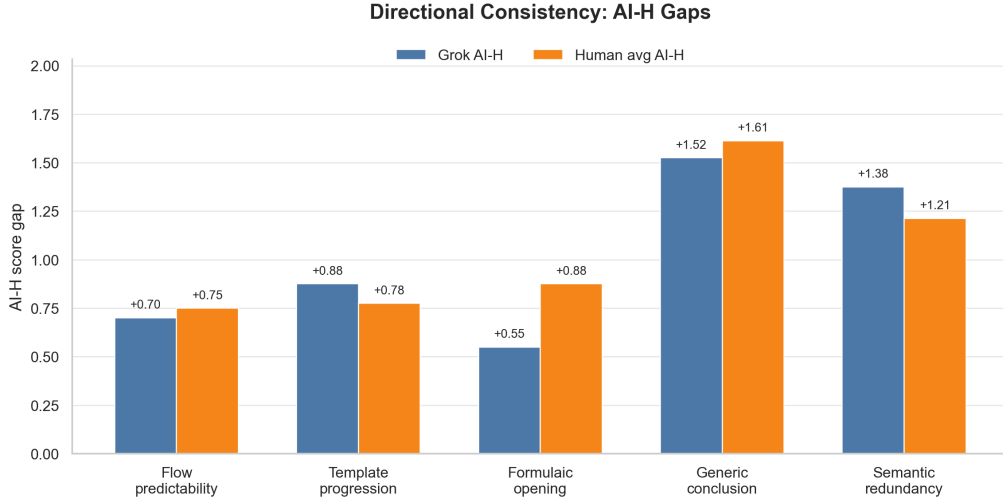


Figure 7: Directional consistency between LLM-evaluated and human-evaluated AI-human score gaps. Positive values indicate higher scores for AIGTs.

D Shortcut-Controlled Evaluation on RAID

Method	News			Wiki			Reddit			Reviews			Abstracts			Books		
	Orig.	Ctrl.	Δ	Orig.	Ctrl.	Δ	Orig.	Ctrl.	Δ	Orig.	Ctrl.	Δ	Orig.	Ctrl.	Δ	Orig.	Ctrl.	Δ
<i>Zero-shot / scoring-based detectors</i>																		
DetectGPT	0.912	0.789	-0.123	0.934	0.812	-0.122	0.901	0.776	-0.125	0.945	0.823	-0.122	0.889	0.754	-0.135	0.923	0.801	-0.122
DetectLLM-LRR	0.960	0.853	-0.107	0.949	0.814	-0.135	0.959	0.844	-0.115	0.964	0.848	-0.116	0.920	0.787	-0.133	0.967	0.824	-0.143
DetectLLM-NPR	0.945	0.832	-0.113	0.967	0.851	-0.116	0.938	0.819	-0.119	0.956	0.841	-0.115	0.912	0.789	-0.123	0.951	0.834	-0.117
FastDetectGPT	0.994	0.888	-0.106	0.998	0.816	-0.182	0.979	0.836	-0.143	0.986	0.819	-0.167	0.972	0.874	-0.098	0.980	0.846	-0.134
Binoculars	0.990	0.882	-0.108	0.999	0.868	-0.131	0.996	0.887	-0.109	0.994	0.894	-0.100	0.980	0.882	-0.098	0.992	0.880	-0.112
Avg.	0.960	0.849	-0.111	0.969	0.832	-0.137	0.955	0.832	-0.122	0.969	0.845	-0.124	0.935	0.817	-0.117	0.963	0.837	-0.126
<i>Training-based detectors</i>																		
RAIDAR	0.978	0.891	-0.087	0.985	0.902	-0.083	0.972	0.883	-0.089	0.981	0.895	-0.086	0.965	0.874	-0.091	0.979	0.893	-0.086
L2D	0.991	0.912	-0.079	0.994	0.921	-0.073	0.987	0.905	-0.082	0.993	0.915	-0.078	0.978	0.894	-0.084	0.989	0.911	-0.078
MoSEs-Logistic	0.990	0.894	-0.096	0.992	0.904	-0.088	0.990	0.902	-0.088	0.988	0.891	-0.097	0.984	0.900	-0.084	0.972	0.884	-0.088
MoSEs-XGBoost	0.988	0.902	-0.086	0.982	0.892	-0.090	0.970	0.892	-0.078	0.981	0.886	-0.095	0.976	0.898	-0.078	0.978	0.868	-0.110
BiScope	0.998	0.916	-0.082	0.992	0.898	-0.094	0.996	0.888	-0.108	0.992	0.901	-0.091	0.988	0.894	-0.094	0.990	0.902	-0.088
Avg.	0.989	0.903	-0.086	0.989	0.903	-0.086	0.983	0.894	-0.089	0.987	0.898	-0.089	0.978	0.892	-0.086	0.982	0.892	-0.090

Table 11: AUROC of baseline detectors across six RAID domains under the original (Orig.) and shortcut-controlled (Ctrl.) settings.

To examine whether the shortcut effects identified in our paired corpus extend to an established benchmark, we apply the same shortcut-controlled evaluation to six domains from RAID: News, Wiki, Reddit, Reviews, Abstracts, and Books. Table 11 compares detector performance under the original and shortcut-controlled settings.

Across all six domains, every evaluated detector shows lower AUROC after shortcut control. The same detector-family pattern observed in the main experiments also emerges: zero-shot/scoring-based detectors exhibit larger average decreases than training-based detectors. These consistent trends suggest that easily measurable formatting and lexical signals are not specific to our paired corpus but are also present in the RAID benchmark setting and partially support detector performance. The RAID results therefore provide cross-benchmark evidence for the shortcut sensitivity identified in our main analysis.

E Additional AIGT-SIGNALBENCH Results and Examples

E.1 Dataset Composition

Table 12 reports the full domain, task-type, model, and editing-variant composition of AIGT-SIGNALBENCH.

Domain	Task type	Models	Variants	Total sampled texts
Essay writing	Long-form writing	6	raw_ai / ai_edited / human_edited	180
Abstract writing	Academic writing	6	raw_ai / ai_edited / human_edited	180
Blog post writing	Informal writing	6	raw_ai / ai_edited / human_edited	180
Product review	Opinion writing	6	raw_ai / ai_edited / human_edited	180
Peer review	Evaluative writing	6	raw_ai / ai_edited / human_edited	180
Meeting summary	Summarization	6	raw_ai / ai_edited / human_edited	180
Translation	Non-English to English	6	raw_ai / ai_edited / human_edited	180
Total				1,260

Table 12: Composition of AIGT-SIGNALBENCH. Each sampled text is paired with a human-written reference.

E.2 Prompt Examples

Controlled generation prompt. The following prompt is used to generate shortcut-controlled AI text from a structural plan and a constrained fact source.

Controlled Generation Prompt

```
System:
You are a domain-specific writing model for a controlled AI-generated text detection benchmark.
Use STRUCTURE_CARD only as the layout plan.
Use FACT_CARD only as the content source.
Do not invent facts, names, dates, numbers, quotations, scenes, or outcomes.
User:
Write the target text for the following domain and task.
DOMAIN: {domain}
TASK: {task_instruction}
Shape summary:
- Required paragraphs: {paragraph_count}
- Target total words: about {target_words}
- Paragraph plan: {paragraph_plan}
- Style constraints: {style_constraints}
Generation rules:
1) Match the layout contract as closely as possible.
2) Follow the paragraph order and sentence-count hints from STRUCTURE_CARD.
3) Use FACT_CARD as the only content source.
4) Do not add unsupported facts, names, dates, numbers, quotes, claims, or outcomes.
5) Do not mention STRUCTURE_CARD, FACT_CARD, or the source passage.
6) Do not output JSON, markdown, bullets, labels, explanations, or metadata.
7) Separate paragraphs with exactly one blank line unless the domain naturally requires a single paragraph.
8) Prefer direct prose over stock transitions, clichés, vague intensifiers, and broad generic endings.
9) Keep punctuation natural. Avoid long comma-chained sentences, em dashes, and decorative formatting.
10) Use numbers and quotations only when they are explicitly supported by FACT_CARD.
STRUCTURE_CARD:
{structure_card}
FACT_CARD:
{fact_card}
Return only the generated text.
```

ai_edited prompt. The following prompt illustrates the ai_edited revision step used to reduce surface shortcuts while preserving the source content and domain-specific writing constraints.

ai_edited Prompt Example

```
## Role
You are an editor revising AI-generated draft text for a benchmark of AI-generated text detection. Your task is not to improve the draft broadly, but to make minimal edits that reduce easily measurable AI-shortcut signals while preserving the same content, genre, and task answer.
## Input fields
DOMAIN: {domain}
SOURCE / TASK CONTEXT: {source_context}
RAW AI DRAFT: {raw_ai_text}
REFERENCE STYLE PROFILE: approximate length, paragraph count, sentence-length rhythm, and domain conventions extracted from the paired human reference.
## Editing instructions
1) Preserve the meaning, factual content, stance, named entities, dates, quantities, and domain-specific terminology in RAW AI DRAFT.
```

2) Keep the output close to the reference style profile: similar length, paragraph structure, sentence rhythm, and level of detail.

3) Edit only expressions needed for shortcut control; do not rewrite from scratch, add new claims, remove substantive information, or reorder major content.

4) Reduce formulaic AI-style transitions and openers, including Overall, Furthermore, Moreover, In conclusion, It is important to note, and This highlights. Keep a connector only when it is locally necessary for coherence.

5) Avoid cliché or promotional phrases such as shed light on, underscore the importance, paves the way, serves as a testament, significant milestone, far-reaching implications, and breathing new life.

6) Avoid vague intensifiers unless concrete evidence supports them, including significantly, substantially, considerably, immensely, tremendously, widespread, comprehensively, invaluable, indispensable, and potential.

7) Do not add a vague future-looking, moralizing, or essay-like final sentence. Preserve abrupt or concise endings when the source genre supports them.

Domain control

News and summaries: preserve names, dates, counts, attribution, and natural comma punctuation.

QA and technical writing: preserve cause/effect relations, measurements, abbreviations, conditions, and qualifiers.

Narrative and opinion writing: preserve short beats, fragments, dialogue cadence, stance, and paragraph rhythm.

Output

Return only the revised ai_edited text. Do not include a preface, bullet list, explanation, score, or metadata.

E.3 Generated Dataset Examples

Table 13 reports the complete sampled texts for the representative examples selected from AIGT-SIGNALBENCH. The examples are organized as a single multi-page table so that all domains share one table number while preserving the full text for each generation/editing condition. The paired human-written references and full metadata are retained in the released dataset.

Example metadata	Full sampled text
Essay Writing raw_ai model: claude-sonnet-4-6 244w	Some governments spend considerable money renovating and restoring old buildings, yet many argue these funds could be better directed toward new housing and infrastructure. The capital invested in construction renovation is large compared with the cost of original housing development, which raises serious questions about value. This essay argues that renovation funds should be redirected to modern housing expansion and transportation routes in urban cities. More than half of the population in developing countries lacks a home, which means the need for new housing is urgent and cannot be ignored. When people cannot find adequate shelter, they are pushed into slums where they face floods, disease outbreaks, and insecurity on a daily basis. Building new lodges in undeveloped areas is therefore considered more advantageous to society than spending heavily on restoring structures that already exist. Urban cities worldwide are experiencing congestion due to overcrowding, and improving transportation routes alongside housing would address this pressure directly. Demolition of structural elements in existing buildings, if not handled by experts, can lead to significant loss and death, making renovation a risky choice for governments with limited skilled personnel. The risks involved in retrofitting can also cause costs to exceed the original budget, which weakens the financial case for restoration projects. Retrofitting can be cheaper than new construction for smaller projects, so investment decisions should be made with respect to project size and available expertise. Restructuring of existing housing should be considered only as a last resort for expanding the housing supply.
Essay Writing ai_edited model: gpt-5.5 241w	Some people believe that teenagers should take part in unpaid community work, and I agree as long as the work is not harmful. This idea is reasonable because volunteering can support the development of the community and may also help reduce unemployment among young people. However, the main condition is that teenagers must be treated fairly and not used as forced workers for tasks they do not want to do. Community progress is one clear benefit of unpaid youth work, especially when teenagers help with simple public needs. For example, they could clean public paths in their society and raise awareness about cleanliness, which would directly benefit local people. The public could also organize teenage groups for different tasks, such as cleaning, awareness activities, or education-related work. Youth unemployment is another important reason to support community work, because unemployed juveniles may have free time but no useful activity. Instead of staying inactive, they could help their association and make a small contribution to the country. This is better than spending every day on computers and video games, especially when volunteering supports charitable work and self-discipline. Teenagers should not be forced to volunteer, because forced work can reduce the value of community service. Their opinions should be considered before giving them community work, and the work should not disturb their education. For these reasons, I support unpaid community work for teenagers, but only when it is safe, voluntary, and balanced with study.

Example metadata	Full sampled text
Essay Writing human_edited model: deepseek-v4-pro 243w	Some people believe that the most effective method for teaching children how to cooperate is through team sports at school. I agree with this view because such activities place young people in situations where they must work closely with others to succeed. The experiences gained from playing sports like football and basketball can build essential teamwork habits that last a lifetime. Team sports expose children to working with different personalities toward a specific goal. In football and basketball, teams have more than four players with separate opinions, personalities, and skills. Without a plan and adherence to it, team play often results in a poor performance. So children quickly learn that cooperation is not optional but necessary. Another key benefit is that team sports help children learn conflict resolution skills. Differing ideas in a team can lead to arguments, and young players must find ways to settle these disagreements to keep functioning. A coach often serves as a mentor in handling disagreements and guides children through the process of listening and compromising until it becomes natural. I saw this process firsthand when I was part of a sports team at a younger age where a coach taught conflict management. Our team practiced conflict resolution skills, leading to quick resolution of quarrels that might otherwise have disrupted our matches. A team of footballers once attributed their success to teamwork, and I believe my own early lessons in cooperation came directly from those shared sporting challenges.
Research Abstract raw_ai model: gpt-5.5 154w	This abstract presents an application of ab initio wavefunction-based techniques to electron-correlation effects in oxide band structures, using MgO as a representative closed-shell ionic oxide and treating its electronic structure through a framework designed to connect finite local calculations with the extended crystal. The analysis is formulated in terms of a local Hamiltonian on finite pieces taken from the extended crystal, while localized Wannier functions and embedding potentials are supplied by earlier periodic Hartree-Fock calculations, providing a reference for assessing correlation-induced changes. The work examines how different correlation contributions shrink the Hartree-Fock band gap and alter band widths, including on-site charge relaxation, nearest-neighbor charge relaxation, and long-range polarization as distinct microscopic components of the correlated description. The results show that correlation effects are necessary for reliable band-gap values, while in MgO the Hartree-Fock band widths are changed less strongly than the gap, with the O 2p valence bands exhibiting an unexpected increase in width.
Research Abstract ai_edited model: gemini-2.5-pro 157w	This study investigates whether the blazar gamma-ray spectrum hardens with increasing flux using nine years of data from the Energetic Gamma Ray Experiment Telescope (EGRET). We conducted an exhaustive search of the EGRET archives, selecting all blazars observed more than twice. This yielded a sample of 67 blazars, including 18 flat-spectrum radio quasars and 6 low-frequency peaked BL Lac objects, representing the most comprehensive such study to date. Across the full sample, no clear pattern of variation in the spectral index was detected. Some blazars showed no statistical evidence of spectral variability, while in a few cases, the spectrum hardened as the gamma-ray flux increased. We also found evidence for a flux-hardness anticorrelation at low fluxes for five blazars. The most well-observed sources in the sample were 3C 279, 3C 273, PKS 0528+134, 1622-297, and 0208-512. For three of these—3C 279, PKS 0528+134, and 1622-297—we observed previously unreported spectral hysteresis on weekly timescales.
Research Abstract human_edited model: gemini-2.5-pro 145w	A complete thermodynamical analysis of the 3 dimensional massless Gross-Neveu model has been performed. The analysis utilized the optimized perturbation theory, which is a non-perturbative method for these calculations. It enables results beyond the known large-N findings, even at the lowest order. The study's results are valid for a finite number of fermion species, denoted as N. A tricritical point was found to exist in the temperature and chemical potential phase diagram of the model. The location of the tricritical point for the discrete chiral phase transition was precisely determined. The phase diagram was also studied in the pressure and inverse density plane. It revealed the existence of a liquid-gas phase in the model, which was previously unknown. N-dependent analytical expressions were derived for the fermionic mass and critical temperature. The study also derived N-dependent analytical expressions for the critical chemical potential of the system.

Example metadata	Full sampled text
Blog Post raw_ai model: claude-sonnet-4-6 181w	Code review is one of the most debated practices in software engineering, yet its core purpose is often misunderstood. Nick Bull wrote a guide on what code review is actually meant to accomplish. The piece appeared in Better Programming, a publication on Medium. It was published on October 23, 2020, and draws on best practices from 5 major tech companies. Those companies are Apple, Google, Microsoft, Amazon, and Facebook. Many engineers treat code review as a checklist exercise, focused on confirming that changes work, that style guides are followed, and that no regressions have been introduced. Bull argues that while those checks matter, they are not the point of the process. The real purpose, according to Bull, is to ensure the solution solves the problem in the best possible way, taking into account both code quality and product considerations. That framing shifts code review from a gatekeeping task into a collaborative problem-solving conversation. Checking for regressions and style guide adherence are considered necessary parts of the review, but treating them as the goal misses the broader function that code review serves.
Blog Post ai_edited model: gpt-4 155w	Joe Rogan recently took to Instagram to share his personal experience with the carnivore diet, a regimen primarily composed of protein. Known for its exclusionary nature, the carnivore diet places a heavy emphasis on protein intake. Given his flexible schedule and his own podcast, Rogan decided to give this diet a shot. However, the outcome wasn't entirely positive. Rogan faced some severe bowel-related side effects from the diet, an issue he candidly shared with his followers. "Carnivore diet update; the good and the bad. Let's start with the bad. There's really only one "bad" thing, and that thing is diarrhoea," Rogan revealed on Instagram. His experience highlights the potential risks of such a protein-centric diet. National dietary guidelines suggest that protein should constitute between ten to thirty-five percent of our diet. The carnivore diet, with its extreme protein focus, clearly deviates from these recommendations. This could potentially explain the digestive problems Rogan encountered.
Blog Post human_edited model: gpt-3.5-turbo 220w	Instagram recently updated its platform, sparking controversy among users and influencers like James Charles. The redesign placed essential features in the upper-right corner, leading to user dissatisfaction. Reels and Shop buttons are now prominently displayed in the bottom menu for easier access. However, the update made the camera and notifications buttons less accessible to users. Instagram's focus with this update is to prioritize Reels and e-commerce features over traditional posts. Some users may unintentionally tap on Reels or Shop due to muscle memory from the old layout. An Instagram blog post hints at the platform's awareness of potential user disappointment. In the words of Mariana Vargas, "Instagram is very explicit with its intentions on its blog: to prioritize Reels and its e-commerce platform over traditional posts." Overall, the update has stirred mixed reactions within the Instagram community. Users are adjusting to the new layout, but many find the changes disruptive to their usual browsing habits. Influencers like James Charles have openly criticized the update, expressing their frustrations with the platform's direction. The UX team at Instagram may need to address user feedback and consider adjustments to enhance user experience. As Instagram continues to evolve its platform, user feedback and adaptability will drive future updates. The controversy surrounding the recent Instagram update highlights the delicate balance between innovation and user satisfaction.
Product Review raw_ai model: gemini-2.5-pro 112w	I really love this mascara. It comes in a very stylish container. The formula itself is thin, and it applies without any clumping at all. This product provides more of a minimal look, so it is not the right choice if you are hoping to add significant volume. It lasts all day long and I have not had any issues with it smearing. At the end of the day, it is easy to wash off. The applicator also stays clean. Even though this mascara is inexpensive, it does not seem cheaply made in any way. I actually prefer this product to my usual, more expensive brand. I will be buying this again.
Product Review ai_edited model: gpt-4 112w	I recently purchased a foot bath and, to be honest, it's been a letdown. The noise it generates is pretty loud, which isn't ideal when you're trying to unwind with a soothing soak. What's more, it doesn't do a great job of heating the water - it barely manages to keep it lukewarm. I was really hoping for a comforting, warm foot bath, but this product just didn't deliver. The rollers, in particular, were a major issue for me. They're made of hard plastic and some parts are actually sharp. They were uncomfortable to the point of causing me pain. Even the fixtures that hold the rollers in place were sharp and hard.
Product Review human_edited model: deepseek-v4-pro 106w	I've used both over-the-counter and prescription retinol products before, so I know what to expect. I apply retinol only at night and stay out of direct sunlight afterward. The Face Oil Retinol Serum doesn't specifically warn users to avoid direct sun exposure, which feels like an oversight. What I do like is the dry oil texture that absorbs easily without feeling oily or sticky. It moisturizes my skin well and seems to help with firmness and overall tone. I have sensitive skin and experienced no redness, irritation, or sensitivity at all. My dermatologist would probably still want a clearer sun warning on the bottle.

Example metadata	Full sampled text
Peer Review raw_ai model: claude-sonnet-4-6 165w	This paper presents an interesting contribution to maximum entropy problems in deep learning. The core proposal is to learn a smooth invertible transformation that maps a simple distribution to a maximum entropy distribution, rather than optimizing over the continuous density directly, which is a meaningful reframing of the problem. The method combines stochastic optimization with the augmented Lagrangian method to solve a finite-dimensional constrained optimization, and this integration of normalizing flow networks with classical constrained optimization machinery is the technical centerpiece of the work. Applications to finance and computer vision are presented to demonstrate flexibility and accuracy, offering concrete evidence that the framework generalizes across problem domains. Reviewers did raise concerns about the novelty of the method relative to existing approaches, and those concerns deserve a direct and careful response in any revision. The paper is seen as connecting deep learning to the wider information theory community, which is a genuinely underserved bridge given how rarely maximum entropy problems are treated within deep learning contexts.
Peer Review ai_edited model: gpt-5.5 185w	Emergent Predication Structure in Vector Representations of Neural Readers pursues a worthwhile goal by examining recent reading comprehension work through the lens of logical or predication structure. I appreciate the effort to organize models such as Attentive Reader, Stanford Reader, Attention-Sum Reader, Gated-Attention Reader, and Attention-over-Attention Reader into the categories of “aggregation readers” and “explicit reference models.” The paper is generally well written, and section 3 is particularly helpful: the discussion of predication makes the framing clearer and more useful than a broad appeal to logical structure. The shift in terminology from logical structure to predication is appropriate, and the authors’ clarification is detailed and useful. That said, I still have some uncertainty about the paper’s contribution, since the current analysis does not fully show what the predication perspective adds beyond a descriptive organization of neural readers. The abstract’s reference to the Who-did-What dataset is relevant to the paper’s objective, but I remain unsure whether this dataset is well suited to that objective, especially in light of prior concerns about the CNN/DailyMail dataset associated with Chen et al. at ACL’16.
Peer Review human_edited model: claude-sonnet-4-6 195w	This paper proposes neural equivalence networks for learning representations of symbolic expressions and presents the motivation for this approach in a manner that is well articulated and reasonably easy to follow, which is a genuine strength of the submission. The evaluation is also reasonably extensive, covering a range of experimental conditions that support the core claims. That said, the primary concern centers on the choice of evaluation metric, which appears to be precision computed on a per-query basis and is non-standard for this type of retrieval or matching problem. Precision-recall curves or ROC analysis would be considerably more informative, and the current metric is expected to produce inflated results when equivalence classes are larger, because the likelihood of a random expression matching a query grows with class size and the denominator does not account for this structural bias. The motivation offered for the subexpforce loss is not sufficiently developed, and the reviewer recommends that the authors strengthen this justification before the work is reconsidered. A direct ablation comparing models trained with and without the subexpforce loss, while holding residual connections and normalization constant across conditions, would clarify whether the loss component contributes meaningfully to performance.
Meeting Summary raw_ai model: claude-sonnet-4-6 120w	Council Bill 18-1089, presented by Laura Brzezinski of the Office of Economic Development and reviewed by the Safety Committee, proposes increasing the minimum affordability period for city-subsidized housing projects from 20 years to 60 years, with an amendment introduced by Councilwoman Kenny limiting the change to contracts dated on or after February 1, 2019, to provide certainty for developers whose tax credit letters of intent to the Colorado Housing and Finance Authority are due before that deadline. The amendment serves as a compromise against a full postponement to February 2019, while the bill also clarifies definitions and notice periods in the preservation ordinance based on three years of implementation experience since the right of first refusal was added in 2015.
Meeting Summary ai_edited model: claude-sonnet-4-6 107w	Recommendation to declare city-owned properties at 240 Long Beach Boulevard and 322–356 East Third Street surplus and authorize their sale to the Art Exchange for \$2.3 million. The project would establish a visual arts center in the East Village Arts District featuring artist studios, galleries, residential uses, and adaptive reuse of the Acres of Books building. Vice Mayor Lowenthal, sponsor of the District 2 item, described the sale as more than 15 years in the making. Councilman Andrus conditioned his support on receipt of a staff report confirming that Long Beach residents would be hired for demolition and construction work, as promised by the developers.

Example metadata	Full sampled text
Meeting Summary human_edited model: gpt-5.5 120w	Resolution 1207, approving a \$1.4 million contract with Mental Health Center of Denver to expand the STAR support team assisted response program, an alternative to armed police response for people in crisis. The contract identified Mental Health Center of Denver as the main contractor, and supported a program said to have started in the summer of 2020, with five vans expected subject to supply chain delays, seven newly hired clinicians, six newly hired EMT and medical tech staff, and an estimated 10,000 contacts. The contract included Urban Institute as third-party evaluator and a community advisory committee for oversight, input, and feedback. Councilmember Canete discussed the item, with Councilwoman Gilmore, Councilmember Ortega, Councilmember Flynn, and the Council President identified.
Translation raw_ai model: gemini-2.5-pro 122w	The Aichi Prefectural Police has designated the 11th of every month as "Crosswalk Day," and through public awareness campaigns calling for the protection of crossing pedestrians, strengthened enforcement of obstruction violations, and the inspection and maintenance of road signs and markings, it aims to foster drivers' awareness of pedestrian protection and promote pedestrian use of crosswalks. On the first "Crosswalk Day" on July 11, the Nishi Police Station held an event at a crosswalk near a large shopping center, where they practiced safe crossing methods and used hand plates to call for pedestrian protection (pictured). Preschool children and elderly residents from the jurisdiction participated. Officers publicized "Crosswalk Day" while distributing awareness items and conducted enforcement against violations such as obstructing crossing pedestrians.
Translation ai_edited model: deepseek-v4-pro 112w	The Mexican U-20 Women's National Team lost 1-0 to Spain in the quarterfinals of the 2022 Costa Rica World Cup. Their tournament ended after a run that included advancing from the group stage and beating a powerhouse like Germany. Mexico put in a strong quarterfinal showing, pushing for a result until the final minutes despite Inma Gabarro's goal and forcing Spain to drop back defensively. Amid scandals and obstacles, the team still managed to compete with a global power. The result at the Estadio Nacional in San José, Costa Rica, echoed the negative outcome of their first meeting at the 2018 U-17 Women's World Cup, where Mexico lost 2-1.
Translation human_edited model: gpt-3.5-turbo 113w	The World Breastfeeding Week extends until August 7th, with this year's theme being "Protect breastfeeding, a shared responsibility." The aim is to view breastfeeding as a cultural and collective practice, involving society to ensure children's right to be breastfed. Breast milk is the best and the only food specifically designed for a baby in the first 6 months of life, remaining their primary source of nutrition up to the first year. UNICEF estimates that exclusive breastfeeding for six months could annually prevent the deaths of 1.3 million children under five years old. In Argentina, breastfeeding prevalence at birth is over 95%. However, by 6 months, exclusive breastfeeding sharply declines to around 40%.

Table 13: Complete representative generated examples from the sampled field of AIGT-SIGNALBENCH.

E.4 Variant-Level Linear Separability Results

Table 14 reports variant-level linear separability results on AIGT-SIGNALBENCH using the same experimental setup as the linear separability analysis in Section 3.3. We train classifiers with either the nine surface shortcut features or PLM embeddings and evaluate each benchmark variant separately against human-written references. Across both settings, the human_edited variant yields the lowest F1, AUROC, and AUPRC, indicating that human post-editing further weakens easily separable shortcut signals. This result supports the need to evaluate AIGT detectors in realistic AI-assisted writing scenarios where shortcuts are explicitly controlled, rather than relying only on raw-generation settings.

Model	Variant	Metric		
		F1	AUROC	AUPRC
Shortcut LR	raw_ai	0.633	0.677	0.660
	ai_edited	0.673	0.716	0.665
	human_edited	0.573	0.605	0.582
PLM+LR	raw_ai	0.657	0.701	0.696
	ai_edited	0.645	0.691	0.678
	human_edited	0.540	0.588	0.576

Table 14: Variant-level linear separability results on AIGT-SIGNALBENCH. Shortcut LR uses the nine shortcut features, while PLM+LR uses pretrained language model embeddings with logistic regression.

E.5 Transfer from Conventional Benchmarks to AIGT-SIGNALBENCH

Table 15 reports transfer results where training-based baseline detectors are trained on conventional benchmark data and evaluated on AIGT-SIGNALBENCH. AUROC and AUPRC are reported in separate columns for the full benchmark and for each editing variant.

Model	Overall		raw_ai		ai_edited		human_edited	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
MoSEs	0.568	0.563	0.569	0.577	0.569	0.556	0.565	0.558
BiScope	0.719	0.731	0.720	0.740	0.766	0.762	0.670	0.690
RAIDAR	0.666	0.592	0.676	0.607	0.676	0.603	0.649	0.572
L2D	0.595	0.579	0.564	0.555	0.609	0.585	0.613	0.605

Table 15: Transfer results from conventional benchmark training to AIGT-SIGNALBENCH. Models are trained on conventional benchmark data and evaluated on AIGT-SIGNALBENCH. AUROC and AUPRC are reported separately for each evaluation setting.

E.6 Mixed-Training Evaluation

Test Dataset	Method	0% Mix	10% Mix	20% Mix	30% Mix
RAID	RAIDAR	0.978	0.963	0.949	0.927
	L2D	0.991	0.975	0.958	0.953
	MoSEs-Logistic	0.990	0.967	0.944	0.919
	MoSEs-XGBoost	0.988	0.959	0.938	0.934
	BiScope	0.998	0.979	0.961	0.939
	Avg.	0.989	0.969	0.950	0.934
AIGT-SIGNALBENCH	RAIDAR	0.701	0.732	0.748	0.766
	L2D	0.812	0.845	0.862	0.859
	MoSEs-Logistic	0.548	0.575	0.593	0.615
	MoSEs-XGBoost	0.571	0.592	0.611	0.604
	BiScope	0.741	0.766	0.781	0.797
	Avg.	0.675	0.702	0.719	0.728

Table 16: Training-based detector performance on RAID and AIGT-SIGNALBENCH at different mixing ratios.

We examine whether AIGT-SIGNALBENCH can also provide useful training supervision by replacing 0%, 10%, 20%, and 30% of the RAID training set with AIGT-SIGNALBENCH examples. The five training-based detectors are trained under each mixing ratio and evaluated on both RAID and AIGT-SIGNALBENCH. Table 16 reports the resulting AUROC.

Increasing the proportion of AIGT-SIGNALBENCH examples consistently improves average performance on AIGT-SIGNALBENCH, from 0.675 with RAID-only training to 0.728 at the 30% mixing ratio. All five detectors improve over their RAID-only baselines. This improvement is accompanied by a decrease in average RAID AUROC from 0.989 to 0.934, revealing a trade-off between generalization to shortcut-controlled examples and performance on the original benchmark. These results suggest that AIGT-SIGNALBENCH can serve not only as a robustness evaluation benchmark but also as a source of complementary hard-case supervision for detector training.

F Baseline Detector Configurations

For all baseline detectors, we used the publicly released implementations and followed their default experimental configurations unless otherwise required for dataset formatting. For MoSEs, the official pipeline uses BAAI/bge-m3 as the embedding model and EleutherAI/gpt-neo-2.7B as both the scoring and reference language model. Its detection pipeline combines the released SAR and CTE components,

including the default sub-centroid configuration for SAR and the default nearest-neighbor/PCA-based CTE probability estimator.

For BiScope, we followed the released implementation, which extracts suffix-based contrastive features from a chosen detector backbone and trains a Random Forest classifier with 100 trees. The public code treats the detector backbone as a required argument rather than a fixed default, and applies the released feature extraction procedure with a fixed token clipping window.

For RAIDAR, we used the official rewriting-based detection pipeline. The public implementation uses gemma-9b-instruct as the default rewriting model, generates four rewrites per input, and trains the released MLP classifier on the resulting discrepancy features. Its generation configuration uses temperature sampling by default, while top- k and top- p filtering are only applied when the corresponding flags are enabled.

For L2D, we used the released AdaDist/L2D implementation with gemma-9b-instruct as the default base and rewriting model. The public code fine-tunes the detector using the released LoRA-based scoring model setup and generates four rewrites per input for likelihood-ratio-based detection. Our experiments follow the official public-code configurations, with only dataset paths adapted to our transfer-evaluation setting.

All experiments were run on GPU cluster using NVIDIA RTX A5000 24GB GPUs; the most memory-intensive L2D runs used two A5000 GPUs with model parallelism.

G Human Editing and Calibration Materials

This appendix provides supplementary screenshots for the human calibration and human editing procedures used in this work. We recruited two university undergraduate students as human participants for these procedures.

Human calibration materials Figures 8, 9, and 10 show the human calibration materials used to compare the rubric-based LLM evaluator with human judgments. These screenshots document the instructions given to participants, the calibration workflow, and the judgment interface used by human raters. Participants were instructed to evaluate each text according to the rubric criteria and assign scores consistently across the calibration examples.

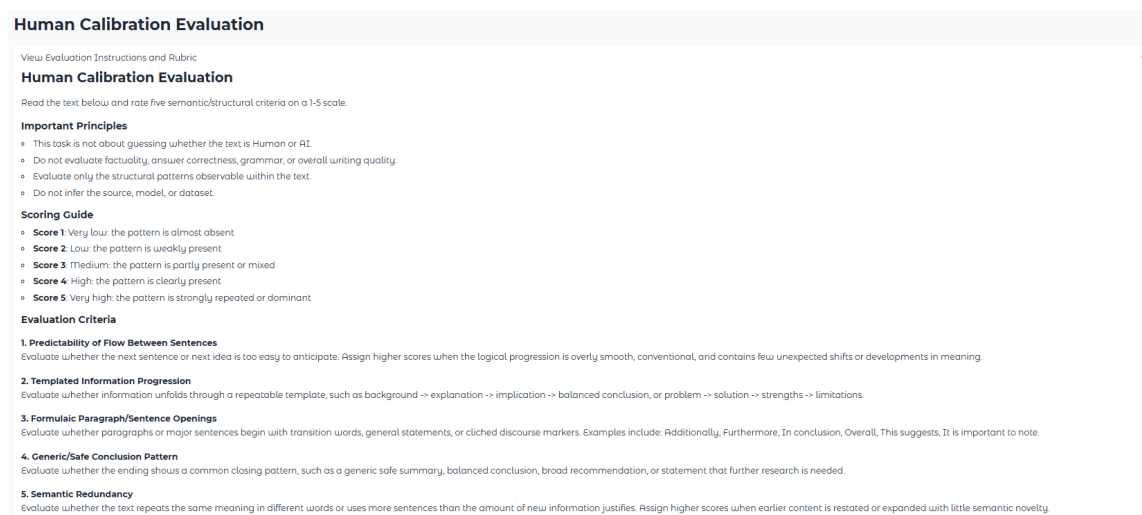


Figure 8: Human calibration material for rubric-based semantic-content flow evaluation.

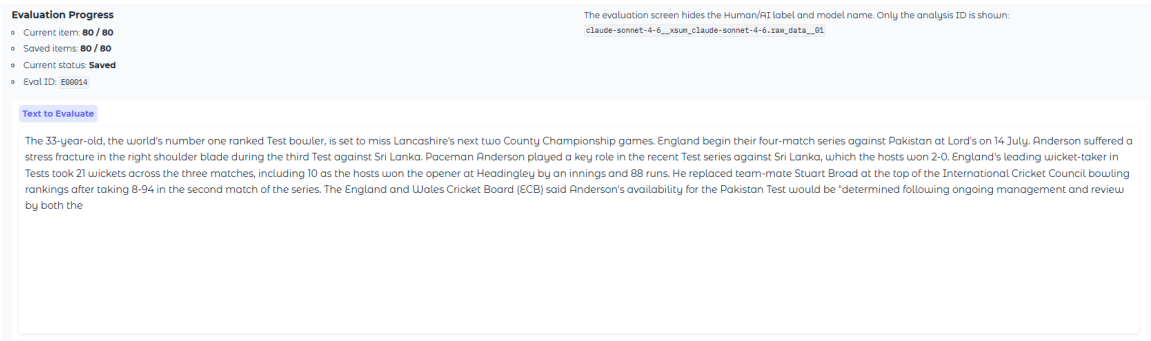


Figure 9: Human calibration interface for assigning rubric scores across semantic-content flow criteria.

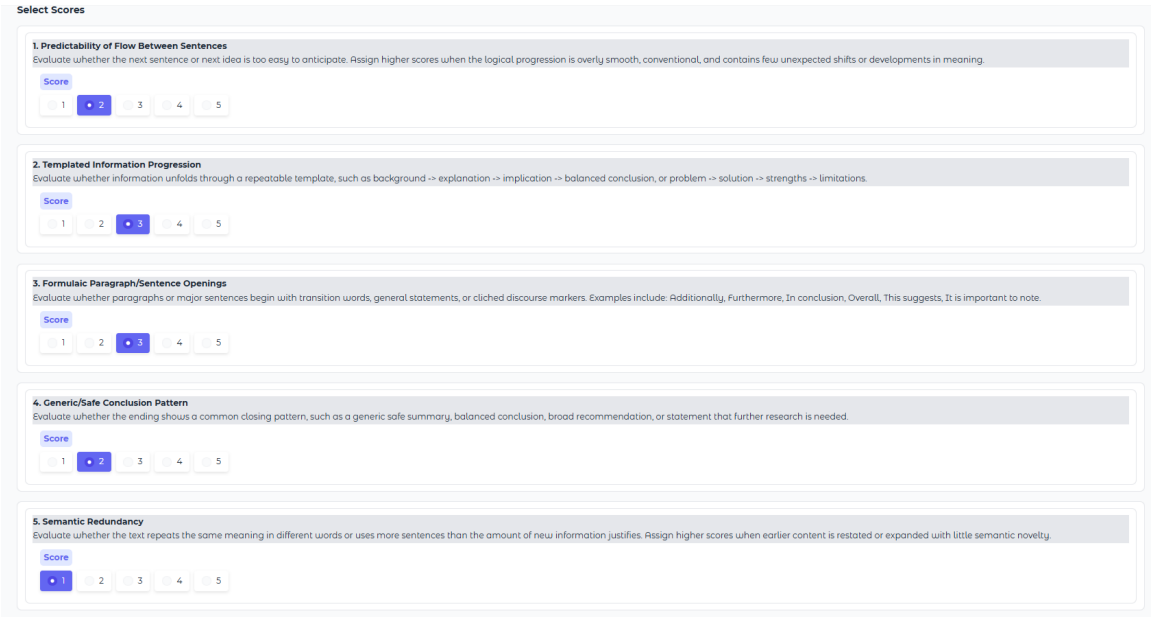


Figure 10: Additional human calibration material used for validating rubric-based semantic-content flow judgments.

Human editing materials Figure 11 shows the human editing interface used to construct the human_edited variant.

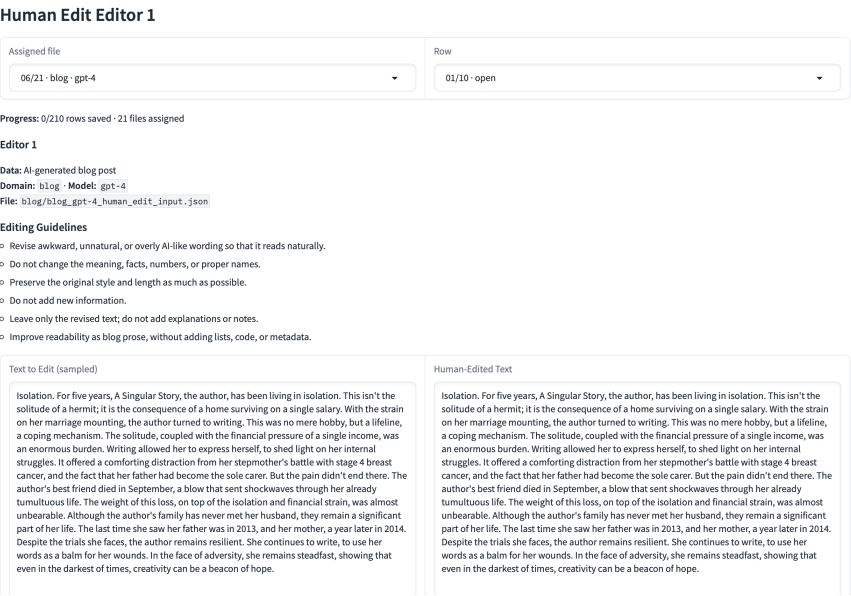


Figure 11: Human editing interface used to construct the human_edited variant. Participants improved naturalness and clarity while preserving the original content and avoiding unsupported information.