REVERSE TOBLEND: ANALYZING DETECTION SIGNALS UNDER TOKEN-LEVEL MULTI-MODEL BLENDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Detectors and attribution tools for AI-generated text usually assume one model per passage; token-level blending breaks this by drawing each short continuation from a different model in a pool. What a defender can recover from such text remains open, even with every pool model available: machine generation, blending, or the model behind each word. We study this case with likelihood statistics from every pool model, supervised text classifiers, and 600 new continuations whose generating model is logged at every step. Encouragingly, detection is reliable: a classifier over pool likelihoods reaches AUROC of at least 0.85 in all 36 pool, domain and generation settings, and a fine-tuned RoBERTa classifier is stronger on one of two pools. Both complement an off-the-shelf detector, RADAR, in all 18 settings where it applies. At the same time, how text was blended is hard to recover: on the logged continuations, blended and single-model text separate with AUROC of at most 0.58, and word attribution reaches 41.3% (25% chance) even with the logged decoding distributions. We examine two four-model pools across three English domains. Detection and source tracing thus appear to be distinct problems: a detection alone says little about which models wrote a passage.

1 INTRODUCTION

Machine-generated text motivates both detection and provenance analysis, with practical implications for the interpretation of online content (Sadasivan et al., 2025; Zhou et al., 2023); for example, ChatGPT-written explanations of implicit hate speech can reinforce lay readers' judgments of the posts (Huang et al., 2023). Detection methods include supervised classifiers (Solaiman et al., 2019; Fagni et al., 2021), watermarking (Kirchenbauer et al., 2023), and statistical tests based on likelihood, rank and sampling discrepancy (Gehrmann et al., 2019; Mitchell et al., 2023; Bao et al., 2024). Their behavior depends on the generating process, the scoring model, the domain and the representation presented to the detector (Bao et al., 2024; Dugan et al., 2024).

Token-level multi-model blending, proposed as an attack on AI-generated text detection by ToBlend (Huang et al., 2024), continues a human prefix by repeatedly drawing a model from a candidate pool. A draw can select the same model again, so a redraw does not always change the generating model. Unlike an edit applied after generation, blending changes the process that produces the continuation. This setting raises three separable questions: whether text is machine-generated, whether it was produced by blending rather than one fixed model, and whether individual spans can be assigned to their true generators. High performance on the first question does not answer the other two. They matter even after a supervised classifier such as RoBERTa has flagged a passage. Identifying which pool members produced it supports provenance and accountability, for example by showing which model providers to contact. Telling blending apart from ordinary single-model generation shows a defender whether countermeasures designed for one generator suffice or whether the whole pool must be addressed.

We study these questions from the defender's side and assume white-box access to every model in the pool, the setting most favorable to detection and attribution. We use two four-model pools, three English domains, supervised text baselines, outside-pool scorers, perturbation controls, and a new matched experiment in which blended and fixed-model continuations share prompts, generation loop

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

and stopping rule, and every draw is logged. Our analysis begins with the known Fast-DetectGPT formulation (Bao et al., 2024): when one model both samples and scores, its numerator combines log likelihood and predictive entropy. This identity is exact for scores; how each component's AUROC changes under blending is an empirical question that we test separately.

Detection with pool access is strong. A component reader, logistic regression on 17 pool likelihood statistics, reaches AUROC at least 0.85 in all 36 non-sentence evaluation cells (one per pool, domain and generation setting). A fine-tuned RoBERTa classifier is stronger still on the advanced pool, including on the same cleaned text, so the reader does not meet the planned detector rule, which required beating RoBERTa and a character n-gram classifier: it does so in six of 18 classic-pool cells and none of 18 advanced-pool cells, and four classic label-permutation AUROCs fall below the required band. Fused with RADAR (Hu et al., 2023), an off-the-shelf detector applied out of distribution, the reader improves AUROC in all 18 cells with RADAR scores, as do both supervised text classifiers, which shows that in-domain supervised detection complements an out-of-distribution detector. At the same time, recovering how text was blended, the second and third questions, is harder. Classifiers trained to separate blended from fixed-model text reach AUROC of at most 0.5791. The logged decoding distributions raise word attribution from 36.53% to 41.31% (25% chance), and boundary F1 stays below that of marking a model change at every word.

Our contributions are: (1) a provenance-logged matched benchmark of 600 continuations in 299 prompt clusters, extended by 600 each for the advanced pool at TL3 and the classic pool at TL1 (TLk: k-token chunks), that evaluates detection, blend specificity and word attribution as separate tasks against trivial, zero-shot and supervised baselines, with the reader frozen before generation (fitted once on the historical ToBlend corpus without the first 100 prompts per domain, never refit); (2) evidence that pool-access likelihood features detect blended text and, like supervised text classifiers, complement an out-of-distribution detector, while a fine-tuned text classifier remains stronger on the advanced pool on identical input; (3) bounds on what these signals recover about blending in the tested settings: trained blended-versus-fixed AUROC of at most 0.58 for the advanced pool at TL1, and word attribution of 41.3% at TL1 and TL3 even with the logged decoding distributions; and (4) controls that show which apparent blending signatures are shared with other perturbations, depend on domain or text view, or are not seen at logged model changes.

2 Related Work

Statistical and supervised detection. Likelihood and rank diagnostics characterize text under a scoring model (Gehrmann et al., 2019; Ippolito et al., 2020). DetectGPT (Mitchell et al., 2023) and Fast-DetectGPT (Bao et al., 2024) use probability curvature or sampling discrepancy, while DetectLLM (Su et al., 2023) develops log-rank signals. Supervised detectors learn from labeled human and generated text (Solaiman et al., 2019; Fagni et al., 2021); contrastive coherence methods and detector surveys provide complementary context (Liu et al., 2023; Tang et al., 2024).

Multiple scoring models. Combining language-model signals also has substantial precedent. Binoculars contrasts one model's perplexity with the cross-perplexity between two closely related models, a zero-shot use of two-model disagreement (Hans et al., 2024). SeqXGPT uses sequences of white-box log probabilities with convolution and self-attention for sentence-level detection (Wang et al., 2023). Ghostbuster combines probabilities from weaker language models through structured feature construction and a trained classifier (Verma et al., 2024). Building on this precedent, we use a logistic component reader as an analysis instrument for comparing features and controls under a specified blending process, not as a new detection method.

Attribution and mixed authorship. Source attribution asks which model produced a text. LLMDet records next-token probabilities of salient n-grams to compute a proxy perplexity for each candidate model (Wu et al., 2023), and SeqXGPT extends white-box log-probability features to labeling each sentence with the model that generated it (Wang et al., 2023). Text that mixes human and machine writing motivates fine-grained detection (Zhang et al., 2024) and detection of the boundary between human-written and generated parts (Zeng et al., 2024; Kushnareva et al., 2024). In token-level blending, boundaries can occur between machine generators at every step rather than between a human and a machine, and here the true generator of every step is logged.

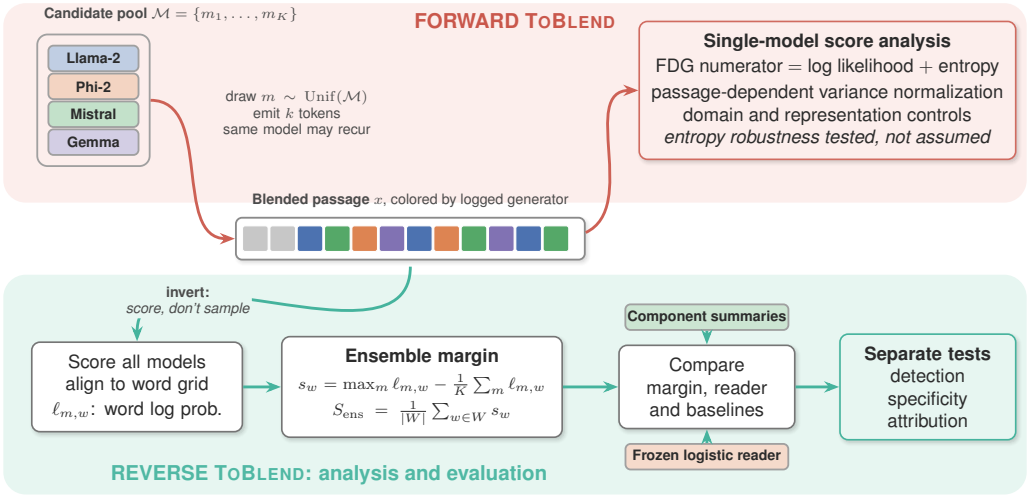

Figure 1: Reverse ToBlend study pipeline. Forward generation, as in ToBlend (Huang et al., 2024), redraws a model for each chunk of k tokens, counted as NLTK tokens (Bird et al., 2009); a redraw can repeat the same model. Reverse analysis scores the passage under every candidate, aligns model tokens onto words, and evaluates the ensemble margin S_{ens} (how much the best pool member's word log probability exceeds the pool mean, averaged over words) and the component reader. Detection, blend-versus-fixed specificity, and attribution against true logs are separate tests. Colors illustrate logged generators, not predicted provenance. A natural hypothesis from the score identity is that entropy stays discriminative under blending; Section 4 tests it as a sanity check.

Attacks and robustness. Paraphrasing and character- or word-level perturbations can alter detection performance (Sadasivan et al., 2025; Krishna et al., 2023; Wolff & Wolff, 2020), and benchmarks such as RAID evaluate detectors across many attacks and decoding strategies (Dugan et al., 2024). RADAR trains a detector adversarially against a paraphraser (Hu et al., 2023); OUTFOX uses adversarially generated examples in an in-context approach (Koike et al., 2024); rewriting-based approaches are another related defense (Mao et al., 2024; Hao et al., 2025). ToBlend (Huang et al., 2024) evaluates blending as an attack; we ask what a defender with pool access can recover.

3 Study Design

3.1 Corpora, Pools, and Views

The historical blending corpus covers SQuAD (Rajpurkar et al., 2016), XSum (Narayan et al., 2018), and WritingPrompts (Fan et al., 2018), with 300, 500 and 500 human/generated pairs per setting before filtering; this corpus and its detector battery come from ToBlend (Huang et al., 2024), whose battery results we reference rather than repeat (Appendix K.1). The classic pool contains GPT-2 XL (Radford et al., 2019), OPT-2.7B (Zhang et al., 2022), GPT-Neo-2.7B (Black et al., 2021) and GPT-J-6B (Wang & Komatsuzaki, 2021); the advanced pool contains Llama-2-7B (Touvron et al., 2023), Phi-2 (Javaheripi & Bubeck, 2023), Mistral-7B-v0.3 (Jiang et al., 2023; Mistral AI, 2024) and Gemma-7B (Gemma Team, 2024; Banks & Warkentin, 2024). Historical settings are token lengths TL1–TL5, random length in 1–5, and sentence-level generation; Appendix B.1 describes the blending procedure. Token length refers to the generation loop's NLTK (Bird et al., 2009) word-token units rather than a common model-token vocabulary. The archived advanced XSum TL4 setting actually uses $k = 3$; we retain the archived label with this correction. Sentence generation uses a different pipeline and is a separate sensitivity, not a pure change in granularity.

V0 is original text. V1 (normalized text) applies the same spacing normalization to both classes and re-scores that text; it is the primary normalized view for the component analyses in Section 4. V2-lite-mask (masked features) drops leaked special-token marker and quote words from cached features while keeping their original context in the remaining scores; it is the view on which the

detector rule is evaluated. V2-lite-text (cleaned text) strips markers and converts quotes in V1 text. The masked reader and text classifiers consequently share held-out rows and folds but read distinct representations. In the matched experiment, V2-lite scores are recomputed from cleaned text and are therefore distinct from masked cached scores.

3.2 Statistics and Component Reader

For scoring model m, let $\ell_{m,w}$ be the log probability summed over model tokens aligned to whitespace word w. The ensemble margin is

$$S_{\text{ens}}(x) = \frac{1}{|W|} \sum_{w \in W} \left(\max_m \ell_{m,w} - \frac{1}{K} \sum_m \ell_{m,w} \right).$$

By construction, this nonnegative statistic measures pool disagreement, not generator identity. It is large when one pool member explains a word much better than the others, but it compares uncalibrated log probabilities, so a member that assigns systematically higher probability can dominate it (Table 32); Section 5.3 also tests a calibrated variant. The maximizing model label is evaluated separately against logged provenance. Algorithm 1 specifies this scoring procedure. Figure 1 keeps detection, specificity and attribution distinct.

The component reader uses four summaries per member: token-weighted mean cross-entropy, token-weighted predictive entropy, standard deviation of word log probability, and mean predictive log-probability variance. Together with S_{ens}, the four-model pool supplies 17 features. The classifier is standardized L2 logistic regression with $C = 1$ and maximum 2,000 iterations, fixed without tuning. Passage length is excluded from the primary features; a length-only baseline and length-covariate sensitivity are reported separately. The reader therefore has white-box access to every candidate generator: its features contain the ingredients of each member's own Fast-DetectGPT statistic, so it can detect text from any single pool member, blended or not (Table 14).

We use six protocol families with 324 protocols in total (Appendix D), with prompt-grouped folds and training-only preprocessing (Appendix C). Same-prompt human/generated pairs pass a planned 10% relative word-length caliper, which retains 0.592 to 0.830 of the rows in the 36 evaluation (gate) cells (Appendix C).

3.3 Supervised Comparators and Decision Rules

A char 1–4-gram TF–IDF logistic classifier and a fine-tuned RoBERTa-base (Liu et al., 2019) text classifier provide supervised comparisons. RoBERTa uses pooled training per generating pool: two pools, three seeds and five shared prompt folds yield 30 fits and 109,200 held-out predictions. Training settings are in Appendix C. The char and RoBERTa classifiers read cleaned text, while the corresponding reader uses masked features. Comparisons use the same held-out rows and prompt folds, and Section 5 repeats the detector rule with the reader rescored on the same V2-lite text, so that all three detectors read identical input.

All decision rules and thresholds in Table 1 were written in an analysis plan dated 18 September 2026, before the final run of 20 September 2026 (hash in Appendix B); we refer to them as planned rules. The historical cells had already been analyzed before this plan was written. The planned null test uses one draw of each null control, a paired prompt-group label permutation and an exact character-aligned shared-prefix window, with a tolerance band that applies to their point AUROCs, not their confidence intervals. We keep this test and its verdict as planned, and report the 1,000-draw permutation p-value as a post-hoc calibration (Section 5). The plan left two thresholds unfixed: the material fraction in the boundary rule and the fragment-word exclusion rule; rather than choose them after seeing results, we apply neither the boundary-effect label nor the fragment exclusion.

AUROC intervals and paired differences generally use 2,000 prompt-cluster bootstrap replicates, retaining paired classes and repeated settings jointly. The per-cell RADAR fusion analysis uses 1,000 replicates and reports multiplicity-adjusted comparisons (calibration set and score orientation: Appendix C).

Table 1: Planned decision rules. Cells are the 18 non-sentence cells per pool; "majority" means at least 10 of 18.

Rule	Condition and outcome labels
Detector	Per pool, in V2-lite-mask: AUROC ≥ 0.85 in ≥ 14 cells; paired CI above 0 against both char and RoBERTa, and against the single-scorer reader, each in a majority of cells; both null point AUROCs in $[0.45, 0.55]$ in every cell. All hold: detector claim; otherwise no detector claim.
Specificity	Blended-vs-fixed S_{ens} AUROC ≥ 0.65 with CI lower bound > 0.5: pass; CI includes 0.5: fail; otherwise: reported without a specificity claim.
Mixing	Entropy AUROC gain of blended over fixed (each vs. human) ≥ 0.10 with CI above 0: mixing drives the rise; $[0.05, 0.10)$: partial; < 0.05: null; ≤ -0.05: reversed.
Attribution	Word-accuracy CI inside $[0.23, 0.27]$: chance; CI excludes 0.25 and point ≥ 0.30: weak attribution; CI excludes 0.25 and point < 0.30: small margin; otherwise inconclusive.
Tracking	Within-minus-across-chunk agreement CI includes 0: none; CI excludes 0 and effect < 0.05: slight; ≥ 0.05: weak tracking.
Boundary	$\lvert z\rvert < 2$: no detected effect; $\lvert z\rvert \geq 3$ and effect at least a material fraction of the global offset: boundary effect; otherwise inconclusive. The material fraction was never fixed, so the boundary-effect label is not applied.

3.4 Matched Generation and Frozen Evaluation

The new experiment uses the first 100 instances per domain and the advanced pool at TL1. For every human prompt, a blended arm redraws a model at each step and a fixed-model arm keeps one model. Both use the same loop, formatting, stochastic decoding configuration and stopping rule. Fixed-model assignments are balanced 25/25/25/25 within each domain. The loop copies 30 NLTK tokens, requests a short model continuation, appends the first k NLTK tokens, and preserves the historical pre-append 170-token stopping check, so the final append can cross that target. Each draw, seed and emitted character interval is logged. A second run adds matched arms for the advanced pool at TL3 and the classic pool at TL1 (100 prompts per domain and arm; Appendix J).

All 600 generations completed. Each scored view has 1,200 class rows (two arms, 300 instances each, with one human and one generated row per instance), and each comparison uses 300 paired instances in 299 actual-prompt clusters. Repeated prompts remain distinct instances within their shared bootstrap cluster. The 132 reader models were frozen before generation; training excludes every group touching the reserved first-100 indices, including 13 later Writing instances with repeated reserved prompts. The reader fitted on normalized text is applied to original-text scores, and the masked-feature reader to fresh cleaned-text scores, without refitting. These explicit view transfers are part of the interpretation. A separate 15-cell format-swap experiment renders historical single-model and human text in the same blend-like surface form (Appendix G).

4 Decomposition and Empirical Sanity Checks

When the reference and scoring models coincide, the numerator of the Fast-DetectGPT discrepancy (Bao et al., 2024) is the sum of log likelihood and predictive entropy (Appendix A). This identity constrains scores, not AUROCs. It suggests three natural hypotheses, which we test as sanity checks (details in Appendix C): that the entropy–likelihood AUROC difference predicts the margin's advantage over Fast-DetectGPT, that entropy stays discriminative under blending, since the original ToBlend evaluation found entropy-based detection and RADAR largely unaffected by blending (Huang et al., 2024), and that the likelihood-down, entropy-up sign pattern is specific to blending. The association falls from $r = 0.8622$ to a partial $r = 0.1940$ ($p = 0.2641$) after adjusting for the Fast-DetectGPT AUROC, and this design cannot separate it from confounding (Figure 2a; Appendix A.1). Entropy discrimination depends on domain and view: on normalized text, all 14 Writing entropy AUROCs are at most 0.55 (Figure 2b; Table 9). The sign pattern also holds with scorers outside the pool on SQuAD and XSum (96/96 scorer/cell combinations on original text; Table 10) and in all 16 planned perturbation comparisons (Figure 2c), so it is not specific to blending.

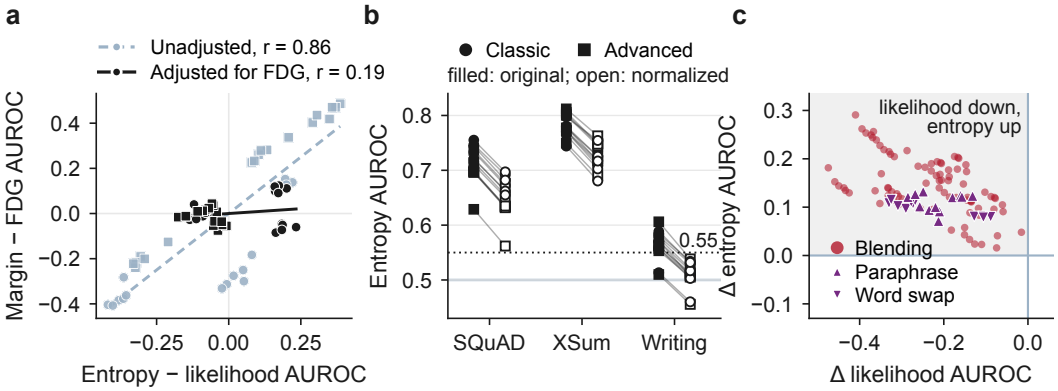

Figure 2: Three hypotheses suggested by the score identity describe general score behavior rather than blending. (a) Entropy–likelihood AUROC difference against the margin's advantage over Fast-DetectGPT (36 historical cells), centered (gray, $r = 0.86$) and adjusted for the Fast-DetectGPT AUROC (ink, partial $r = 0.19$). (b) Entropy AUROC per cell on original and normalized text. (c) Likelihood and entropy AUROC changes from blending (red) and from T5 paraphrase and word-swap perturbations (purple; 24 effects, 16 of them planned); all fall in the likelihood-down, entropy-up quadrant.

Table 2: Reader evaluation under V2-lite-mask: counts among 18 required non-sentence cells per pool. AUC: AUROC ≥ 0.85; Single: paired CI above the single-scorer reader; Supervised: paired CI above both char and RoBERTa comparators; Nulls: both null point AUROCs in $[0.45, 0.55]$. The planned detector rule fails.

Pool	AUC	Single	Supervised	Nulls
Classic	18/18	18/18	6/18	14/18
Advanced	18/18	18/18	0/18	18/18

5 Reader, Fusion, and Matched Results

5.1 Strong Discrimination and the Planned Detector Rule

In all 36 evaluation cells, the reader reaches AUROC of at least 0.85 and beats the single-scorer reader by paired CI. Six classic cells beat both supervised baselines by paired CI, and no advanced cell does (Table 2). The six favorable classic comparisons are SQuAD; the other 12 classic comparisons with RoBERTa are inconclusive. All 18 advanced cells favor RoBERTa by paired CI. At a nominal 5% false-positive threshold set on calibration prompts (realized test rates differ), the reader's median true-positive rate across gate cells is 0.947 (classic) and 0.774 (advanced), against 0.975 and 0.953 for RoBERTa and 0.517 and 0.611 for the char classifier (Table 21). Rescoring the reader on the same cleaned text that RoBERTa and char read leaves the verdict unchanged (Table 24): the reader beats RoBERTa by paired CI in 5 of 18 classic cells and none of 18 advanced cells, where RoBERTa is better in all 18.

The planned null test also fails in four classic cells: label-permutation AUROCs are 0.4317 for SQuAD TL3, and 0.4407, 0.4348 and 0.4480 for XSum TL1, TL2 and TL5. These are below the required point-estimate band; they are not evidence of a positive shared-prefix leakage effect. We retain the planned failure instead of relaxing the band after inspection. The final result is *no detector claim*. As a post-hoc calibration, we use the 1,000-draw permutation p-value as the calibrated null test: with 1,000 label permutations per gate cell, every reader AUROC has the smallest attainable permutation p (1/1001), and every recorded null point is reproduced. The calibration also shows that the band is narrow relative to null variation: a leakage-free reader would keep all 18 classic null points in the band with probability 0.13 if cells were independent (0.21 advanced; Table 20). The verdict, no detector claim, is unchanged, because the rule also fails on the supervised comparison.

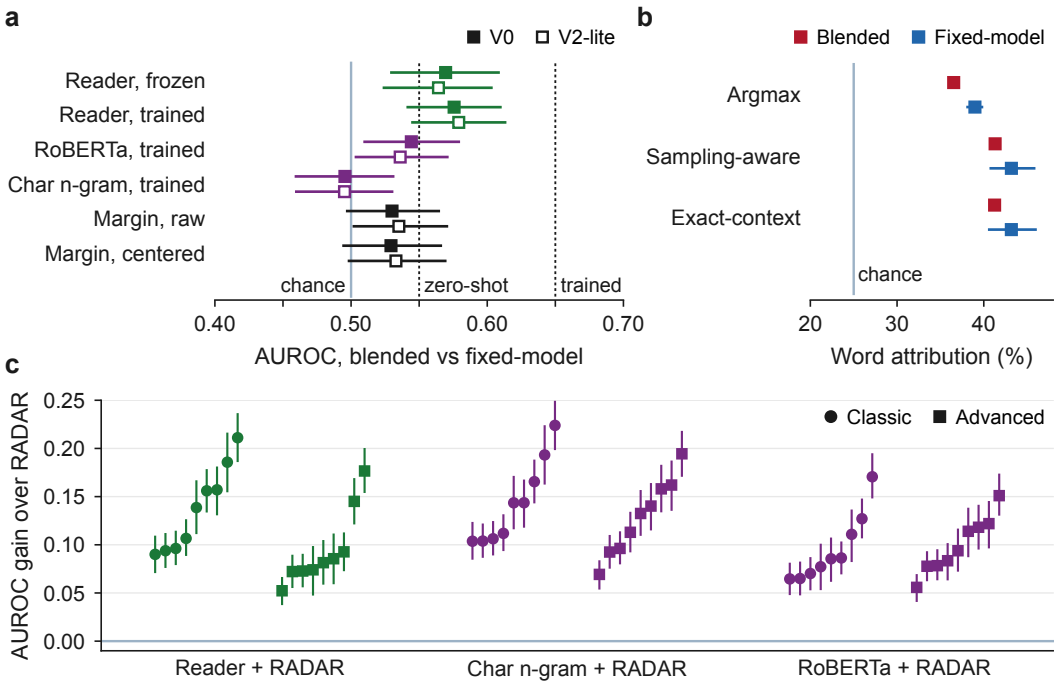

Figure 3: Matched generation bounds blend-specific signal: blended-versus-fixed AUROC stays at or below 0.58, word attribution reaches 41.3%, and every in-domain detector shares the RADAR fusion gain. (a) AUROC for blended versus fixed-model text on the 300 matched advanced-pool TL1 pairs, with 95% prompt-cluster intervals; filled squares, original text (V0); open squares, cleaned text (V2-lite). The frozen reader is transferred without refitting, trained detectors use five grouped prompt folds, and the margin is S_{ens}, raw or centered per model. Gray line, chance; dotted lines, the planned 95% CI lower bounds for a zero-shot (0.55) and a trained-detector (0.65) specificity claim. No lower bound reaches 0.55. (b) Word attribution accuracy against logged generators (V0) for the raw argmax, sampling-aware and exact-context attributors, on blended (red) and fixed-model (blue) text; gray line, 25% chance. (c) AUROC gain of RADAR fused with each in-domain supervised detector over RADAR alone in the 18 cells with RADAR scores (circles, classic; squares, advanced; nine cells per pool, sorted), with 95% intervals. All 54 gains are Holm-positive within their nine-cell families. Values are in Tables 4, 22, 23 and 26.

5.2 Complementarity with RADAR

RADAR scores exist for 18 of the 42 cells (TL1, random and sentence settings, nine per pool). In a nested per-cell analysis of original text over these cells, RADAR plus the reader improves AUROC over RADAR by 0.052 to 0.211, and all 18 gains survive Holm and Benjamini–Hochberg correction within their nine-cell families, whereas no margin gain does (Table 33). The same design with other in-domain supervised detectors gives the same result (Table 23; Figure 3c): RADAR plus a character 1–4-gram classifier gains 0.104 to 0.224 (classic) and 0.069 to 0.194 (advanced), and RADAR plus RoBERTa gains 0.065 to 0.171 and 0.056 to 0.151, each Holm-positive in all 18 cells. Each supervised detector alone already exceeds RADAR (the reader by 0.080 to 0.209 and 0.040 to 0.173). RADAR adds to the reader in 14 of 18 cells (5 of 9 classic, 9 of 9 advanced) by 0.004 to 0.031, to char in 2 of 18 and to RoBERTa in none. RADAR is a trained detector applied here out of distribution, whereas these detectors are trained on the same pool and domain, so the gain reflects complementarity between in-domain supervised detection and an out-of-distribution detector.

5.3 Matched Generation Bounds Blend Specificity

Table 4 (Appendix C) reports all six matched comparisons. For blended versus fixed-model text, S_{ens} gives AUROC 0.5301 [0.4962, 0.5653] on original text and 0.5351 [0.5011, 0.5713] on cleaned text. The original-text result fails the specificity rule because its interval includes chance; the

cleaned-text result is below the planned passing AUROC of 0.65 and is reported without a specificity claim.

The frozen reader separates blended from human text with AUROC 0.9506 on original and 0.9395 on cleaned text, and blended from fixed-model text with AUROC 0.5694 and 0.5642. It also separates fixed-model text from humans well. Because the frozen reader was trained on blended versus human text, we also train detectors directly for the blended-versus-fixed contrast with five grouped prompt folds (Figure 3a; Table 22). RoBERTa, the planned primary for this test, reaches 0.5443 [0.5090, 0.5800] on original and 0.5361 [0.5026, 0.5716] on cleaned text: above chance, but below the planned lower bound of 0.65, so we report it without a specificity claim. The 17-feature reader trained for the contrast reaches 0.5757 and 0.5791 and a character n-gram classifier stays at chance (0.4955); the frozen reader rescored on cleaned text gives 0.5670. No calibrated variant of S_{ens} (best 0.5329) and no zero-shot score, including Binoculars (0.5357) and Fast-DetectGPT under GPT-Neo (0.5404) or OLMo-2-7B (0.5179), has a CI lower bound above the planned 0.55 (Table 27), so the specificity bound is robust to calibration. On the new arms, S_{ens} gives 0.5434 [0.5031, 0.5854] for the advanced pool at TL3 and 0.5166 [0.4808, 0.5556] for the classic pool at TL1 (Table 29). The reader thus captures features for generated-versus-human discrimination, with at most a small blend-specific component.

Blending increases entropy's generated-versus-human AUROC relative to the fixed-model arm by 0.0516 [0.0158, 0.0884] on original and 0.0385 [0.0023, 0.0762] on cleaned text. Both are small positive effects: the first lies in the planned partial band (0.05 to 0.10), and the second lies below the 0.05 threshold, which the planned mixing rule labels "null" (Table 12).

5.4 Logged Attribution Reaches 41.3%

Word attribution accuracy of the raw argmax (the highest-likelihood member per word) is 36.53% [35.98%, 37.12%] on original and 35.79% [35.24%, 36.34%] on cleaned text (Table 11), above 25% uniform guessing, and 38.98% on fixed-model text (Table 16). A sampling-aware attributor scores every word under each member's logged decoding transform (temperature, top-k and top-p; removed tokens receive probability 10^{-6}), and an exact-context attributor scores every logged chunk under each member in the context the loop encoded, which for a uniform draw is the maximum-likelihood chunk attributor up to NLTK re-spacing (0.74% of steps). They reach 41.31% [40.70%, 41.88%] and 41.25% [40.65%, 41.84%] on original text (Table 22; Figure 3b), so under the planned rule the logged decoding distributions bound recoverable word provenance at 41.3%. Accuracy is 78.8% for Llama-2, the only member sampled at temperature 0.6 with top-p 0.9, and 25.9% to 32.5% for the three members sampled at temperature 1 (Table 26). Accuracy is similar for the advanced pool at TL3 (41.33%), and on the classic pool at TL1 the transform leaves it at 38.19%. As with the argmax, attribution is higher on fixed-model text (43.18%).

Exact adjacent-word boundary F1 is 0.7407 [0.7360, 0.7454] for the raw argmax and 0.7029 [0.6960, 0.7093] for the sampling-aware attributor. Uniform four-way redraw changes the generator at 74.8% of adjacent word pairs, so marking every pair as a boundary gives F1 of 0.8555 (a truth-independent predictor with the argmax's boundary density gives 0.7343). Both attributors stay below the always-boundary baseline (sampling-aware minus always-boundary -0.1527 [-0.1590, -0.1465]), so we make no boundary-recovery claim.

5.5 Logged Model Changes Behave like Same-Model Redraws

At approximate period-k positions in TL2–TL5, blended text shows lower word discrepancy at switch positions than at off-switch positions (-0.0179 classic, -0.0176 advanced, original text; $z = -6.27$ and -6.15 against random positions), with human-control CIs including zero; the effect persists in the mapped normalized grid (-0.0125 and -0.0146; Figure 4a) but is absent in the unmapped one (Table 34). At logged positions, chunk starts where the model changes do not differ from those where the same model is redrawn: 0.0066 [-0.0155, 0.0299] on original and 0.0049 [-0.0195, 0.0296] on cleaned text at TL1 (permutation $p = 0.5920$ and 0.7313; Figure 4b), -0.0109 [-0.0474, 0.0269] for the advanced pool at TL3 (10,824 changes, 3,513 redraws), and 0.0120 [-0.0091, 0.0333] for the classic pool at TL1 (Table 29). Following the planned rule, we attribute the historical effect to the approximate period-k grid rather than to model changes. Because the material fraction was never fixed, the boundary-effect label is not applied.

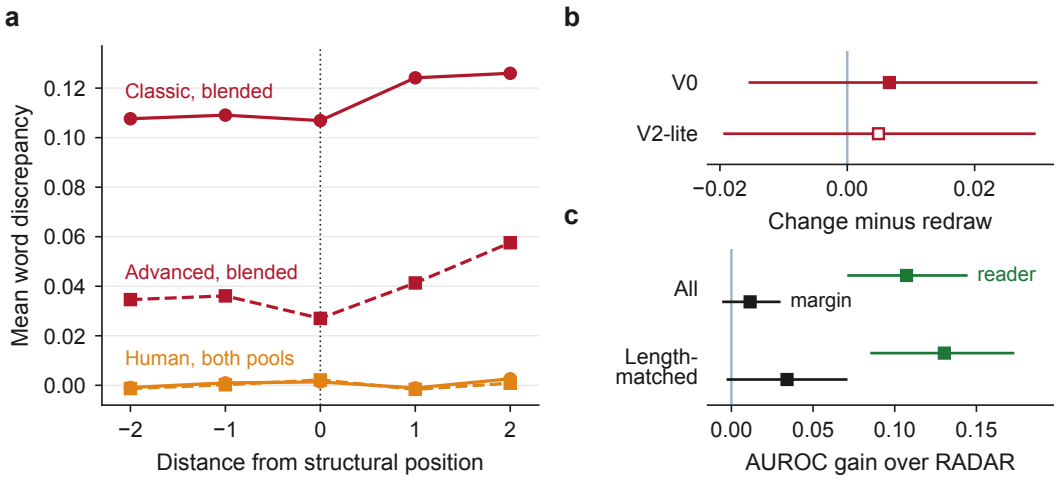

Figure 4: Switch positions show a historical grid effect, while logged model changes and same-model redraws behave alike. (a) Historical V1-mapped word-grid profiles around approximate period-k positions, pooled over TL2–TL5 (circles and solid lines, classic; squares and dashed lines, advanced). (b) Logged change-minus-redraw contrasts in matched advanced TL1 generation with 95% cluster intervals; both include zero. (c) Advanced V0 geometry-subset gains over RADAR with 95% intervals, before and after the length caliper.

6 Discussion and Conclusion

The known Fast-DetectGPT decomposition remains algebraically valid, and our sanity checks show that the three hypotheses it suggests describe general score behavior rather than blending: the adjusted association is weak, entropy discrimination depends on domain and view, and the likelihood-down, entropy-up pattern also appears under other perturbations. With white-box pool access, the reader detects blended text in all 36 evaluation cells and serves as a useful analysis instrument; a fine-tuned text classifier is stronger on the advanced pool, also on identical input, and the reader, a character n-gram classifier and RoBERTa each complement RADAR in every cell. At the same time, matched generation bounds the blend-specific signal: every trained classifier, calibrated margin and zero-shot score separates blended from fixed-model text at AUROC of at most 0.58, word attribution reaches 41.3% with the logged decoding distributions and is lower on blended than on fixed-model text, and logged model changes behave like same-model redraws. Reverse ToBlend therefore provides an evaluation framework that tests detection, blend specificity and attribution separately, and the reader's gains come from its full feature set rather than the scalar ensemble margin (further discussion in Appendix C). Detection and source tracing thus appear to be distinct problems: a detection alone says little about which models wrote a passage, so attribution claims need separate evidence.

7 Limitations

The study assumes white-box access to every pool model, a setting favorable to the defender, across three English domains and two fixed four-model pools of open models of nominal size 7B or less, released between 2019 and 2024; pool membership is confounded with model family and size. The matched experiment applies every analysis to the advanced pool at TL1 and extends only the margin, attribution and switch analyses to the advanced pool at TL3 and the classic pool at TL1, two added settings sampled in micro-batches. Because historical pipelines and some format controls differ and length matching can sharply reduce coverage, historical controls complement rather than replace the new matched generation. Word attribution is scored on words with one logged generator, excluding prompt, mixed-provenance and deleted-character words, and treats words or chunks independently; sequence models of the redraw process are a natural next step. Candidate-pool scoring needs access to each model, and its cost scales with pool size.

Ethical Statement

AI-text detection and attribution can be misused when uncertain scores are treated as proof of authorship. The failed detector rule, weak specificity and incomplete attribution coverage make such use particularly inappropriate here. The reported analyses do not validate a system for adjudicating misconduct or assigning responsibility to an author. Evaluation is limited to the stated English corpora and does not establish fairness across dialects or populations. Studying a detector-evasion process also has dual-use implications; the scientific purpose of these controls is to make the scope and failure modes of detection claims inspectable.

Code and Data Availability

The 1,800 matched generations with per-step model, seed and character-interval logs, the score caches, the frozen readers, the analysis plan and decision rules, and the analysis, reanalysis and table-generation code will be released publicly upon publication.

Acknowledgments

We thank the maintainers of the datasets, language models and detection implementations used in this study.

AI Use Statement

We used large language model assistants for the grammar checks and code debugs. The synthetic corpora are generated by the open language models; these are the object of study. The authors designed the study, fixed the decision rules, reviewed every analysis and interpretation, and take full responsibility for all the content and claims presented in the manuscript.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. The Falcon series of open language models. *arXiv preprint arXiv:2311.16867*, 2023. URL https://arxiv.org/abs/2311.16867.

Jeanine Banks and Tris Warkentin. Gemma: Introducing new state-of-the-art open models. Google Blog, February 2024. URL https://blog.google/innovation-and-ai/technology/developers-tools/gemma-open-models/.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. Fast-DetectGPT: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Bpcgcr8E8Z.

Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Sebastopol, CA, 2009. ISBN 978-0-596-51649-9.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. GPT-Neo: Large scale autoregressive language modeling with Mesh-Tensorflow. Zenodo, version 1.0, March 2021. URL https://doi.org/10.5281/zenodo.5297715.

Liam Dugan, Alyssa Hwang, Filip Trhlík, Josh Magnus Ludan, Andrew Zhu, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. RAID: A shared benchmark for robust evaluation of machine-generated text detectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12463–12492, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.674. URL https://aclanthology.org/2024.acl-long.674/.

Tiziano Fagni, Fabrizio Falchi, Margherita Gambini, Antonio Martella, and Maurizio Tesconi. TweepFake: About detecting deepfake tweets. *PLOS ONE*, 16(5):e0251415, 2021. doi: 10.1371/journal.pone.0251415. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0251415.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 889–898, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1082. URL https://aclanthology.org/P18-1082.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M. Rush. GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 111–116, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-3019. URL https://aclanthology.org/P19-3019/.

Gemma Team. Gemma: Open models based on Gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024. URL https://arxiv.org/abs/2403.08295.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Spotting LLMs with Binoculars: Zero-shot detection of machine-generated text. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 17519–17537. PMLR, 2024. URL https://proceedings.mlr.press/v235/hans24a.html.

Wei Hao, Ran Li, Weiliang Zhao, Junfeng Yang, and Chengzhi Mao. Learning to rewrite: Generalized LLM-generated text detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6421–6434, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.322. URL https://aclanthology.org/2025.acl-long.322/.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. RADAR: Robust AI-text detection via adversarial learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 15077–15095. Curran Associates, Inc., 2023. doi: 10.52202/075280-0662. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/30e15e5941ae0cdab7ef58cc8d59a4ca-Abstract-Conference.html.

Fan Huang, Haewoon Kwak, and Jisun An. Is ChatGPT better than human annotators? Potential and limitations of ChatGPT in explaining implicit hate speech. In *Companion Proceedings of the ACM Web Conference 2023*, pp. 294–297, Austin, TX, USA, 2023. ACM. doi: 10.1145/3543873.3587368. URL https://doi.org/10.1145/3543873.3587368.

Fan Huang, Haewoon Kwak, and Jisun An. ToBlend: Token-level blending with an ensemble of LLMs to attack AI-generated text detection. *arXiv preprint arXiv:2402.11167*, 2024. URL https://arxiv.org/abs/2402.11167.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. Automatic detection of generated text is easiest when humans are fooled. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1808–1822, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.164. URL https://aclanthology.org/2020.acl-main.164.

Mojan Javaheripi and Sébastien Bubeck. Phi-2: The surprising power of small language models. Microsoft Research Blog, December 2023. URL https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas

Wang, Timothée Lacroix, and William El Sayed. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023. URL https://arxiv.org/abs/2310.06825.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 17061–17084. PMLR, 2023. URL https://proceedings.mlr.press/v202/kirchenbauer23a.html.

Ryuto Koike, Masahiro Kaneko, and Naoaki Okazaki. OUTFOX: LLM-generated essay detection through in-context learning with adversarially generated examples. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(19):21258–21266, 2024. doi: 10.1609/aaai.v38i19.30120. URL https://ojs.aaai.org/index.php/AAAI/article/view/30120.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. In *Advances in Neural Information Processing Systems*, volume 36, pp. 27469–27500. Curran Associates, Inc., 2023. doi: 10.52202/075280-1195. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/575c450013d0e99e4b0ecf82bd1afaa4-Abstract-Conference.html.

Laida Kushnareva, Tatiana Gaintseva, German Magai, Serguei Barannikov, Dmitry Abulkhanov, Kristian Kuznetsov, Eduard Tulchinskii, Irina Piontkovskaya, and Sergey Nikolenko. AI-generated text boundary detection with RoFT. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=kzzwTrt04Z.

Xiaoming Liu, Zhaohan Zhang, Yichen Wang, Hang Pu, Yu Lan, and Chao Shen. CoCo: Coherence-enhanced machine-generated text detection under low resource with contrastive learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 16167–16188, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.1005. URL https://aclanthology.org/2023.emnlp-main.1005/.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. URL https://arxiv.org/abs/1907.11692.

Chengzhi Mao, Carl Vondrick, Hao Wang, and Junfeng Yang. Raidar: geneRative AI detection viA rewriting. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=bQWE2UqXmf.

Mistral AI. Mistral-7B-v0.3, 2024. URL https://huggingface.co/mistralai/Mistral-7B-v0.3. Hugging Face model card.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 24950–24962. PMLR, 2023. URL https://proceedings.mlr.press/v202/mitchell23a.html.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 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*, pp. 1797–1807, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1206. URL https://aclanthology.org/D18-1206/.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL https://aclanthology.org/D16-1264/.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can AI-generated text be reliably detected? Stress testing AI text detectors under various attacks. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=0OgsAZdFOt.

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. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*, 2019. URL https://arxiv.org/abs/1908.09203.

Jinyan Su, Terry Yue Zhuo, Di Wang, and Preslav Nakov. DetectLLM: Leveraging log rank information for zero-shot detection of machine-generated text. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 12395–12412, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.827. URL https://aclanthology.org/2023.findings-emnlp.827/.

Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. The science of detecting LLM-generated text. *Communications of the ACM*, 67(4):50–59, 2024. doi: 10.1145/3624725. URL https://doi.org/10.1145/3624725.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. URL https://arxiv.org/abs/2307.09288.

Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. Ghostbuster: Detecting text ghostwritten by large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1702–1717, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.95. URL https://aclanthology.org/2024.naacl-long.95/.

Evan Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, Michal Guerquin, David Heineman, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James Validad Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Jake Poznanski, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 OLMo 2 Furious (COLM's Version). In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=2ezugTT9kU.

Ben Wang and Aran Komatsuzaki. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax, May 2021.

Pengyu Wang, Linyang Li, Ke Ren, Botian Jiang, Dong Zhang, and Xipeng Qiu. SeqXGPT: Sentence-level AI-generated text detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 1144–1156, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.73. URL https://aclanthology.org/2023.emnlp-main.73/.

Max Wolff and Stuart Wolff. Attacking neural text detectors. *arXiv preprint arXiv:2002.11768*, 2020. URL https://arxiv.org/abs/2002.11768.

Kangxi Wu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. LLMDet: A third party large language models generated text detection tool. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 2113–2133, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.139. URL https://aclanthology.org/2023.findings-emnlp.139/.

Zijie Zeng, Lele Sha, Yuheng Li, Kaixun Yang, Dragan Gašević, and Guanliang Chen. Towards automatic boundary detection for human-AI collaborative hybrid essay in education. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20):22502–22510, 2024. doi: 10.1609/aaai.v38i20.30258. URL https://ojs.aaai.org/index.php/AAAI/article/view/30258.

Qihui Zhang, Chujie Gao, Dongping Chen, Yue Huang, Yixin Huang, Zhenyang Sun, Shilin Zhang, Weiye Li, Zhengyan Fu, Yao Wan, and Lichao Sun. LLM-as-a-coauthor: Can mixed human-written and machine-generated text be detected? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 409–436, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.29. URL https://aclanthology.org/2024.findings-naacl.29/.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. URL https://arxiv.org/abs/2205.01068.

Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G. Parker, and Munmun De Choudhury. Synthetic lies: Understanding AI-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, pp. 1–20, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3581318. URL https://doi.org/10.1145/3544548.3581318.

A Derivation and Interpretation of the Identity

From the Fast-DetectGPT formulation of Bao et al. (2024), under reference model equal to scoring model S,

$$d(x) = \frac{\sum_t [\log p_S(x_t) - \mu_t]}{\sqrt{\sum_t \sigma_t^2}},$$
$$\mu_t = \sum_v p_S(v) \log p_S(v) = -H_t.$$

Thus the numerator equals $\sum_t [\log p_S(x_t) + H_t]$. The variance normalization also depends on the passage.

For distinct scoring and reference models, the Fast-DetectGPT analytic sampling discrepancy (Bao et al., 2024) is

$$d(x) = \frac{\sum_t (\log p_S(x_t) - \mu_{R,t})}{\sqrt{\sum_t \sigma_{R,t}^2}}, \qquad \mu_{R,t} = \sum_v p_R(v) \log p_S(v).$$

Only when $R = S$ does $\mu_{R,t} = -H_t$. With $M_{2,t} = \sum_v p_S(v)[\log p_S(v)]^2$, the corresponding variance is $M_{2,t} - H_t^2$. Thus the numerator is exactly likelihood plus entropy, and the denominator remains $\sqrt{\sum_t (M_{2,t} - H_t^2)}$. A correlation between the unnormalized sum and discrepancy is not a test of equality of normalized scores, and the identity says nothing about additive AUROCs. Original non-sentence V0 discrepancy values retain the detector-file route. Normalized and sentence cache analyses use the declared word-grid variance approximation; those computations are not asserted to be exact model-token reconstructions. The adjustment reported in the main text concerns a descriptive association between cell-level AUROCs, with shared quantities and domain confounds; it is not a causal estimate.

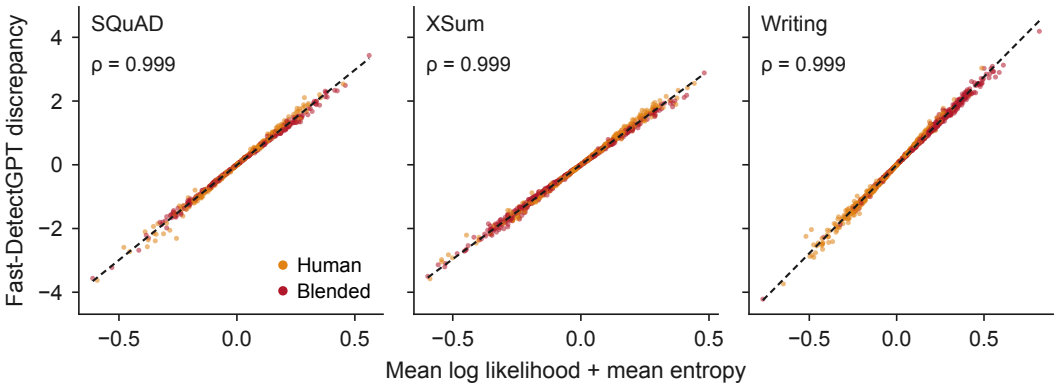

Figure 5: The Fast-DetectGPT discrepancy tracks mean log likelihood plus mean entropy almost exactly (Spearman $\rho \geq 0.998$ in each domain). Historical advanced-pool TL1 passages, human (orange) and blended (red). The correlation is a numerical association; it proves neither equality of the normalized scores nor additivity of AUROCs.

A.1 Adjusted Association between Component AUROCs

Let $\Delta = \mathrm{AUROC}_H - \mathrm{AUROC}_L$ and let the margin advantage be $\mathrm{AUROC}_{S_{\mathrm{ens}}} - \mathrm{AUROC}_{\mathrm{FDG}}$. The unadjusted correlation across 36 historical non-sentence cells is $r = 0.8622$. After adjusting for the Fast-DetectGPT AUROC, the partial correlation is $r = 0.1940$ ($p = 0.2641$). Within pools it has opposite signs: classic $r = -0.8872$ and advanced $r = 0.4789$. Excluding the mislabeled advanced XSum TL4 cell gives overall $r = 0.1857$ ($p = 0.2931$). Shared quantities, domain differences and scoring-pool membership confound the unadjusted association. The margin advantage is almost collinear with the Fast-DetectGPT AUROC ($r = -0.981$), so partialling it out leaves little variance, and the six settings within each pool/domain give nearly identical AUROCs, so the 36 cells are not independent; the six pool/domain means give $r = 0.192$ ($p = 0.757$). We therefore do not treat the association as a predictive regularity and retain the algebraic decomposition as an analytic reference.

B Additional Reproducibility Details

The completed core comprises eight CPU experiments and GPU stages for matched generation, supervised text comparison, normalized cross-pool scoring and format controls, followed by analysis using frozen reader models. GPU scoring completed 12,000 matched score files, eight normalized cross-pool caches with 145,600 passage scores, and nine format caches with 44,600 passage scores. RoBERTa's 30 complete three-epoch fits produced 109,200 held-out predictions. Reader refresh loaded the frozen coefficients and did not refit them. Complete execution does not imply that any scientific evaluation rule passed.

The main GPU work used four NVIDIA A100 80 GB cards. New causal-model generation and scoring used FP16, whereas the historical 2024 reference used FP32; we do not claim bitwise reproduction. RoBERTa training used float32 (Appendix C). Anchor maximum absolute log-probability differences were 0 for Phi-2, 0.0277405 for Llama, 0.0353785 for Mistral and 0.434619 for Gemma. Rescoring Gemma-7B in FP32 changes sampling-aware attribution accuracy only from 41.31% to 41.28% (Appendix J). Word grids and null positions matched. The analysis plan specifies reporting these differences without a numeric acceptance tolerance, so they are not relabeled as passing the separate replication tolerance used for the historical corpus.

Matched TL1 decoding used stochastic sampling, top-$k = 50$, repetition penalty 1, one return sequence, and at most four new model tokens per step before retaining one NLTK token. The shipped Llama configuration used temperature 0.6 and top-$p = 0.9$; Phi-2, Mistral and Gemma used temperature 1 and top-$p = 1$. Model-specific settings were retained identically across the two arms. Decoding kept special tokens and the tokenizer's cleanup default. Instance seeds were derived from SHA-1 hashes of arm, fixed-model identity (or advanced-pool label), domain, instance index and granularity; step seeds additionally identify step and selected model. They initialize PyTorch CPU and CUDA sampling, and model draws use a deterministic Python random generator. Exact effective configurations and seeds are retained with each generation artifact.

The audit retains actual-prompt folds, paired instance identities, output/text hashes and generation provenance. All 188 frozen/input hashes matched the pre-generation snapshot. Historical outputs from ToBlend (Huang et al., 2024) were reused for declared comparisons and were not modified.

The analysis plan containing the decision rules of Table 1 is dated 18 September 2026 and has SHA-256 `290508e8270f17c61b9ff31925f9e428d863796a3d8e67a18b422961c0313d9a`; the final run started on 20 September 2026. The plan states the null condition as both null AUROCs within $[0.45, 0.55]$; the implementation applies it in every gate cell. The plan also lists two formatting-failure conditions (V2-lite-mask AUROC more than 0.10 below V0 in a majority of cells, or a shared-prefix null above 0.55), neither of which was met. A claim-to-artifact map and the generators that render every numerical table from saved artifacts, without fitting models or modifying experimental outputs, will be released with the code. The reanalyses in Appendix I were run on 23 September 2026 on saved outputs, with their decision thresholds written before execution. The reanalyses and new matched arms in Appendix J were run on 24 September 2026 under decision rules written on 23 and 24 September 2026, before their jobs ran.

B.1 Blending Procedure

The historical corpus was generated with the token-level blending procedure of ToBlend (Huang et al., 2024), and the matched experiment reuses the same loop with logging added. Generation starts from the first 30 NLTK word tokens of a human passage, joined by spaces. At each step, one member of the four-model pool is drawn uniformly at random with replacement, so consecutive steps can use the same model. The drawn model receives the current text and samples a continuation of at most $k + 3$ new model tokens; the continuation is split into NLTK word tokens and its first k tokens are appended. The three extra tokens are a buffer that Huang et al. (2024) add to improve generation quality when k is small. In the TL-k settings, k is fixed at a value from 1 to 5 (the archived advanced XSum TL4 setting uses $k = 3$); in the random-length setting, k is drawn at random from 1 to 5 at each step. Before each append, the loop checks whether the text already has at least 170 NLTK tokens; if it does, that chunk is the last one appended, so the final append can cross 170 tokens. In the sentence setting, each draw continues the passage by a whole sentence rather than by

Table 3: Fusion minus RADAR on the advanced-pool V0 geometry subset: 240 human and 240 blended rows, with 81 distinct blended prompt indices. The length sensitivity retains 66 human and 101 blended rows. These pooled differences are separate from the 18-cell per-cell analysis.

Subset	Added feature	ΔAUROC [95% CI]
All	S_{ens}	0.0115 [-0.0056, 0.0301]
All	Reader	0.1072 [0.0710, 0.1446]
Length matched	S_{ens}	0.0341 [-0.0030, 0.0711]
Length matched	Reader	0.1304 [0.0850, 0.1732]

k tokens, through a separate generation pipeline. The classic pool contains GPT-2 XL, OPT-2.7B, GPT-Neo-2.7B and GPT-J-6B, and the advanced pool contains Llama-2-7B, Phi-2, Mistral-7B-v0.3 and Gemma-7B. The matched experiment runs this loop at $k = 1$ for both pools and at $k = 3$ for the advanced pool, adds a fixed-model arm that uses one member at every step, and logs each draw, seed and emitted character interval, with the decoding settings given above.

C Relocated Discussion

The decomposition offers a way to describe scores, but its exactness is insufficient to predict discrimination or establish causation. The adjusted association, Writing normalization results and perturbation controls expose three distinct limits: correlation can reflect a shared component, an apparent signal can depend on domain and representation, and a pattern can occur under several generation or editing processes.

The transformed-view fusion battery contains 162 computed rows and 90 unavailable F-variant rows because aligned rank/log-rank inputs are missing. We do not replace those comparisons with surrogates or infer their results. Historical approximate period-k analysis yields 59 slight-tracking and 13 no-tracking readings among 72 eligible non-TL1 cases, without a fragment exclusion rule, which was never fixed and is not applied (Appendix H).

Decomposition and sanity checks. *The identity does not imply AUROC additivity.* Under reference model equal to scoring model, the numerator of the Fast-DetectGPT discrepancy of Bao et al. (2024) equals $\sum_t [\log p_S(x_t) + H_t]$, while its variance normalization depends on the passage (Appendix A). This identity constrains numerical scores; neither the AUROC of a sum nor the ordering induced by its varying denominator is the sum of component AUROCs. It does not guarantee that blending raises entropy, lowers likelihood or produces a useful detector.

The adjusted association is weak. The unadjusted correlation between the entropy–likelihood AUROC difference and the margin's AUROC advantage over Fast-DetectGPT is $r = 0.8622$ across 36 historical non-sentence cells, but after adjusting for the Fast-DetectGPT AUROC the partial correlation is $r = 0.1940$ ($p = 0.2641$), with opposite signs within the two pools. Because the margin advantage is almost collinear with the Fast-DetectGPT AUROC ($r = -0.981$) and the cells are dependent, this design cannot separate a predictive regularity from confounding; we therefore do not use the association as evidence for one (Appendix A.1).

Entropy is domain and view dependent. The complete view analysis contains 168 view/cell rows. In V1, every one of the 14 Writing entropy AUROCs is at most 0.55 (Table 9). Selected normalized TL1 likelihood, entropy and Fast-DetectGPT results appear in Table 8. These results do not support the hypothesis that entropy discrimination holds under blending across all domains and representations.

Outside-pool scoring shows that the likelihood-down, entropy-up sign pattern does not require a pool member as scorer on SQuAD and XSum (96/96 V0 and 93/96 V1 scorer/cell combinations) but is weaker on Writing (Table 10).

The same tested sign pattern also appears in all 16 planned perturbation/generator/domain comparisons. It is therefore not unique to blending.

Table 4: Matched advanced-pool TL1 generation. Each comparison uses 300 paired instances in 299 prompt clusters; intervals use 2,000 cluster bootstrap replicates. Both generated arms share the generation loop and prompt set. Reader transfer is V1→V0 and V2-lite-mask→fresh V2-lite scores, with no refit. The first class in each comparison is positive.

View	Comparison	S_{ens} AUROC [95% CI]	Frozen reader AUROC [95% CI]
V0	Fixed vs. human	0.5230 [0.4830, 0.5621]	0.8678 [0.8374, 0.8979]
V0	Blended vs. human	0.5604 [0.5216, 0.5988]	0.9506 [0.9321, 0.9662]
V0	Blended vs. fixed	0.5301 [0.4962, 0.5653]	0.5694 [0.5287, 0.6093]
V2-lite	Fixed vs. human	0.4638 [0.4237, 0.5026]	0.8613 [0.8307, 0.8925]
V2-lite	Blended vs. human	0.5011 [0.4643, 0.5394]	0.9395 [0.9193, 0.9567]
V2-lite	Blended vs. fixed	0.5351 [0.5011, 0.5713]	0.5642 [0.5232, 0.6040]

RADAR subset. On a fixed advanced-pool subset of 240 human and 240 blended passages (81 distinct blended prompts, all disjoint from the frozen reader's training prompts), RADAR plus the frozen reader gains 0.1072 [0.0710, 0.1446] AUROC over RADAR, whereas adding S_{ens} gives 0.0115 [-0.0056, 0.0301] (Table 3; details below). Positive reader fusion neither establishes margin-specific added value nor overrides the failed reader detector rule.

Calibration set and score orientation (from Section 3). A separate fixed 30% prompt calibration set supports operating-point sensitivities; cells with fewer than 60 calibration humans are reported as non-estimable. We keep score orientation explicit and do not turn every below-chance AUROC into a favorable post-hoc score.

Perturbation controls (from Section 2). Our perturbation controls ask whether the tested likelihood/entropy pattern is specific to blending. The observed pattern appears in other tested perturbations, so a generative-process difference alone is insufficient evidence of a unique statistical signature.

Deployment requirements (from the Discussion). A deployment claim would require evidence across unseen pools and domains, current and black-box generators, calibrated operating points and attributors that model the redraw sequence. The present findings instead identify which comparisons are necessary before making those claims.

Protocol details (from Section 3). We use six protocol families: in-cell five-fold grouped evaluation, pooled-domain training across settings, TL1-to-other-granularity transfer, leave-one-domain-out evaluation, cross-pool transfer, and historical single-model controls. The completed run contains 324 protocols, including the normalized cross-pool refresh (Appendix D). Actual identical first-30-token prompts define groups across repeated instances, classes, settings and pools. Scalers, vectorizers and score-direction choices use training data only. Transfer evaluation separates source-training and target-test prompts. Same-prompt human/generated pairs pass a planned 10% relative word-length caliper; we report retained coverage rather than matching unrelated examples. In the 36 V2-lite-mask gate cells the caliper retains 0.592 to 0.830 of the original rows.

Format-swap design (from Section 3.4). A separate 15-cell format-swap experiment examines historical single-model and human text both rendered in blend-like surface form with the same NLTK token join. It is distinct from both V1 normalization and the newly matched generation. Length and prompt-identity exclusions are reported in Appendix G.

Outside-scorer analysis (from Section 4). Outside-pool scoring tests whether the observed likelihood/entropy sign pattern requires a generating-pool member as scorer. In the planned non-sentence SQuAD/XSum scope, the pattern passes in 96/96 V0 scorer/cell combinations, before and after length matching; the normalized V1 counts are 93/96 and 92/96. Writing is weaker: V1 passes 12/24 classic and 0/24 advanced outside-scorer combinations, before and after matching (Table 10). All 168 normalized scorer/cell results, including sentence sensitivities, were computed. GPT-Neo is inside the classic pool and outside the advanced pool. Scorer membership is still confounded with

family and size, so this establishes persistence under the tested outside scorers rather than an isolated causal effect of membership.

RADAR subset details (from Section 5). A fixed advanced-pool subset with RADAR scores, disjoint from the frozen reader's training prompts, shows the same reader gain as the per-cell analysis. On a fixed advanced-pool original-view subset, RADAR plus the frozen reader improves pooled AUROC over RADAR by 0.1072 [0.0710, 0.1446], or 0.1304 [0.0850, 0.1732] after the length caliper. The subset contains the 240 human and 240 blended passages of an earlier detector-geometry analysis: 80 human passages per domain (prompts 0 to 79) and 80 blended passages per domain drawn in rotation from TL1, random and sentence settings (prompts 0 to 26), hence only 81 distinct blended prompt indices. The frozen reader excludes every prompt group with index below 100 from training, so it never saw these prompts. The matched subset contains 66 human and 101 blended rows. By comparison, adding S_{ens} gives 0.0115 [-0.0056, 0.0301], or 0.0341 [-0.0030, 0.0711] after matching; neither interval satisfies the positive-lower-bound rule. The optional 42-cell extension was not run. Positive reader fusion neither establishes margin-specific added value nor overrides the failed reader detector rule.

Reader discussion (from the Discussion). The reader results distinguish favorable discrimination from validated comparative performance. Pool-specific probability features improve on a single-scorer reader and complement RADAR in every per-cell comparison, as supervised text classifiers also do (Section 5). Yet the stronger text baseline defeats the reader in all required advanced-pool cells, including at a 5% false-positive rate, and the null rule fails in four classic cells. These results support retaining the reader as an analysis instrument, with its supervised nature and access to candidate models explicit. They do not support presenting Reverse ToBlend as a validated detector or attributing the reader's gains to the scalar ensemble margin.

Supervised joins (from Section 5). The detector-rule failure is an observed comparative and null-control failure, not missing RoBERTa execution: all 42 numerical supervised joins, including sentence sensitivities, are complete. Complete reader values and RoBERTa comparison intervals appear in Tables 5 and 6; Table 7 provides all character-classifier differences.

Historical aligned-null count (from Section 5). The historical aligned-null experiment completed 204 tests.

Limitations on coverage (from the Limitations). Missing aligned rank/log-rank inputs leave 90 fusion rows unavailable. RADAR comparison coverage is 18 V0 cells and a separate geometry subset; the optional 42-cell extension was not run. In particular, advanced format controls retain only 16–24 pairs per domain; 12 classic SQuAD/OPT candidates fail reader prompt-identity checks and are excluded.

Concluding sentence (from the Conclusion). The resulting contribution is a controlled analysis of what these signals measure and where their interpretation remains limited.

Credit for the identity (from Section 2). We credit the analytic sampling-discrepancy formulation to Bao et al. (2024), who also note that when one model both samples and scores, its numerator is the sum of the likelihood and entropy baselines; the identity is therefore not a new detector statistic.

RoBERTa training settings (from Section 3). Each fit runs three epochs with float32 weights, batch size 8, maximum length 512, AdamW learning rate 2×10^{-5}, weight decay 0.01 and gradient clipping 1.

Specificity scope (from Section 5.3). We do not infer specificity merely because a score behaves differently on the older unmatched controls.

Attribution metrics (from the attribution results in Section 5). Neither the approximate structure analysis (Appendix H) nor boundary F1 should be substituted for identity accuracy. At TL3,

where 25.3% of adjacent pairs are generator changes, F1 is 0.3855 against 0.4032 for the always-boundary baseline (Table 28).

Boundary interpretation (from Section 5). Secondary TL1 contrasts against random positions are -0.0107 and -0.0148 with nominal $p = 0.0498$ and 0.0348 from 200 permutations, too few to separate them from 0.05 (Table 13). The approximate-grid signature thus depends on the grid and does not reproduce at logged TL1 positions, and because the material fraction was never fixed, the boundary-effect label is not applied.

Detector rule (from Section 3). The detector conditions are conjunctive; high AUROC alone cannot pass the rule.

Introduction (from Section 1). We report these limits together with the favorable results.

D Reader Protocols and Complete Supervised Comparisons

The final reader analysis contains 324 computed protocols: 126 in-cell, 18 pooled-domain, 72 granularity-transfer, 18 leave-one-domain-out, 84 cross-pool (42 original-view and 42 normalized refresh), and six historical single-model-control evaluations. Tables 5 and 6 report all 42 in-cell supervised joins. The planned gate uses TL1–TL5 and random; sentence rows are additional sensitivity results. Counts are prompt clusters after the same-prompt 10% length caliper. The masked reader retains original cached context, whereas char and RoBERTa consume V2-lite text; the held-out rows and folds match. A negative reader-minus-RoBERTa difference favors RoBERTa. Label-permutation and exact-shared-prefix entries are point AUROCs, and the required interval $[0.45, 0.55]$ is a tolerance band on those points, not a confidence interval.

Table 5: Classic pool: complete in-cell V2-lite-mask reader evaluation. Four required permutation-null points fall below 0.45 (SQuAD TL3; XSum TL1, TL2, TL5); this is an independent failure of the prescribed null rule.

Domain	Setting	n	Reader [95% CI]	ΔRoBERTa [95% CI]	Perm.	Prefix
SQuAD	TL1	244	0.9952 [0.9917, 0.9980]	0.0103 [0.0026, 0.0199]	0.4528	0.5000
SQuAD	TL2	243	0.9955 [0.9919, 0.9982]	0.0124 [0.0032, 0.0223]	0.4521	0.4964
SQuAD	TL3	248	0.9952 [0.9916, 0.9981]	0.0108 [0.0028, 0.0199]	0.4317	0.5000
SQuAD	TL4	249	0.9934 [0.9885, 0.9974]	0.0098 [0.0015, 0.0193]	0.5067	0.4999
SQuAD	TL5	244	0.9932 [0.9885, 0.9969]	0.0110 [0.0015, 0.0214]	0.4972	0.5000
SQuAD	random	242	0.9944 [0.9904, 0.9979]	0.0110 [0.0024, 0.0215]	0.4807	0.5001
SQuAD	sentence	94	0.9925 [0.9835, 0.9988]	0.0042 [-0.0082, 0.0191]	0.5396	0.5007
XSum	TL1	402	0.9878 [0.9826, 0.9924]	0.0008 [-0.0074, 0.0098]	0.4407	0.5000
XSum	TL2	396	0.9832 [0.9749, 0.9904]	-0.0044 [-0.0149, 0.0057]	0.4348	0.4995
XSum	TL3	404	0.9841 [0.9774, 0.9901]	-0.0057 [-0.0140, 0.0027]	0.4850	0.5000
XSum	TL4	388	0.9853 [0.9777, 0.9919]	-0.0006 [-0.0108, 0.0097]	0.4812	0.5001
XSum	TL5	383	0.9853 [0.9774, 0.9917]	0.0021 [-0.0084, 0.0125]	0.4480	0.5000
XSum	random	396	0.9835 [0.9769, 0.9893]	0.0008 [-0.0084, 0.0109]	0.5245	0.5000
XSum	sentence	134	0.9762 [0.9577, 0.9910]	-0.0110 [-0.0314, 0.0088]	0.5370	0.5005
Writing	TL1	288	0.9868 [0.9795, 0.9932]	-0.0056 [-0.0141, 0.0027]	0.4869	0.5011
Writing	TL2	285	0.9874 [0.9805, 0.9928]	-0.0054 [-0.0128, 0.0017]	0.4915	0.5012
Writing	TL3	301	0.9862 [0.9772, 0.9936]	-0.0079 [-0.0182, 0.0014]	0.5185	0.5007
Writing	TL4	301	0.9888 [0.9832, 0.9934]	-0.0042 [-0.0113, 0.0036]	0.5125	0.5012
Writing	TL5	314	0.9901 [0.9843, 0.9949]	-0.0004 [-0.0080, 0.0070]	0.4568	0.5012
Writing	random	291	0.9899 [0.9836, 0.9954]	-0.0019 [-0.0107, 0.0067]	0.4932	0.5012
Writing	sentence	265	0.9908 [0.9850, 0.9954]	-0.0024 [-0.0094, 0.0044]	0.4817	0.5007

Table 6: Advanced pool: complete in-cell V2-lite-mask reader evaluation. †The archived advanced XSum TL4 label has actual $k = 3$.

Domain	Setting	n	Reader [95% CI]	ΔRoBERTa [95% CI]	Perm.	Prefix
SQuAD	TL1	236	0.8589 [0.8285, 0.8883]	-0.1318 [-0.1627, -0.1022]	0.4639	0.4996
SQuAD	TL2	241	0.8778 [0.8498, 0.9042]	-0.1109 [-0.1396, -0.0838]	0.4737	0.5006
SQuAD	TL3	243	0.8909 [0.8661, 0.9157]	-0.0948 [-0.1196, -0.0697]	0.4765	0.5015
SQuAD	TL4	246	0.9031 [0.8799, 0.9254]	-0.0852 [-0.1080, -0.0628]	0.4940	0.5007
SQuAD	TL5	247	0.9053 [0.8813, 0.9278]	-0.0770 [-0.1024, -0.0536]	0.5135	0.5020
SQuAD	random	237	0.8751 [0.8462, 0.9037]	-0.1069 [-0.1366, -0.0778]	0.4676	0.4990
SQuAD	sentence	105	0.8809 [0.8383, 0.9192]	-0.1054 [-0.1463, -0.0681]	0.5197	0.5049
XSum	TL1	395	0.9463 [0.9310, 0.9606]	-0.0441 [-0.0603, -0.0281]	0.4889	0.5009
XSum	TL2	402	0.9589 [0.9460, 0.9707]	-0.0297 [-0.0439, -0.0160]	0.4708	0.5011
XSum	TL3	406	0.9625 [0.9507, 0.9728]	-0.0240 [-0.0374, -0.0113]	0.4805	0.5015
XSum	TL4†	397	0.9595 [0.9461, 0.9713]	-0.0276 [-0.0429, -0.0145]	0.4848	0.5023
XSum	TL5	380	0.9630 [0.9501, 0.9745]	-0.0266 [-0.0413, -0.0130]	0.4791	0.5027
XSum	random	390	0.9578 [0.9445, 0.9703]	-0.0333 [-0.0474, -0.0198]	0.5039	0.5020
XSum	sentence	146	0.9420 [0.9133, 0.9686]	-0.0549 [-0.0839, -0.0288]	0.5181	0.5054
Writing	TL1	288	0.9293 [0.9090, 0.9481]	-0.0532 [-0.0758, -0.0318]	0.5069	0.5015
Writing	TL2	315	0.9481 [0.9315, 0.9641]	-0.0383 [-0.0557, -0.0207]	0.4515	0.5010
Writing	TL3	300	0.9586 [0.9441, 0.9716]	-0.0301 [-0.0451, -0.0148]	0.4631	0.5004
Writing	TL4	328	0.9648 [0.9519, 0.9763]	-0.0231 [-0.0379, -0.0098]	0.5140	0.5006
Writing	TL5	322	0.9666 [0.9548, 0.9773]	-0.0162 [-0.0298, -0.0024]	0.4581	0.5009
Writing	random	328	0.9657 [0.9506, 0.9780]	-0.0221 [-0.0387, -0.0081]	0.5078	0.5018
Writing	sentence	257	0.9730 [0.9612, 0.9833]	-0.0182 [-0.0306, -0.0058]	0.5225	0.4997

Table 7: All 42 paired reader-minus-character-classifier comparisons, completing the supervised comparison alongside Tables 5 and 6. Both classifiers use identical held-out identities and folds; their declared representations differ. Advanced XSum TL4 retains the archived label for actual $k = 3$.

Domain	Setting	Classic Δchar [95% CI]	Advanced Δchar [95% CI]
SQuAD	TL1	0.1549 [0.1253, 0.1839]	0.0230 [-0.0225, 0.0680]
SQuAD	TL2	0.1526 [0.1226, 0.1846]	0.0443 [0.0063, 0.0802]
SQuAD	TL3	0.1817 [0.1475, 0.2116]	0.0881 [0.0528, 0.1270]
SQuAD	TL4	0.1782 [0.1470, 0.2104]	0.1070 [0.0671, 0.1463]
SQuAD	TL5	0.1949 [0.1623, 0.2284]	0.0947 [0.0583, 0.1318]
SQuAD	random	0.2020 [0.1686, 0.2371]	0.0728 [0.0315, 0.1163]
SQuAD	sentence	0.2703 [0.2057, 0.3372]	0.1461 [0.0803, 0.2113]
XSum	TL1	0.0874 [0.0681, 0.1070]	0.0492 [0.0254, 0.0744]
XSum	TL2	0.0970 [0.0752, 0.1190]	0.0752 [0.0478, 0.1030]
XSum	TL3	0.0888 [0.0691, 0.1088]	0.0833 [0.0596, 0.1063]
XSum	TL4	0.1167 [0.0948, 0.1402]	0.0671 [0.0426, 0.0935]
XSum	TL5	0.1181 [0.0946, 0.1425]	0.0856 [0.0603, 0.1122]
XSum	random	0.1246 [0.1022, 0.1478]	0.0858 [0.0592, 0.1114]
XSum	sentence	0.1953 [0.1470, 0.2478]	0.1544 [0.1010, 0.2048]
Writing	TL1	0.0875 [0.0641, 0.1104]	0.0305 [-0.0001, 0.0607]
Writing	TL2	0.0746 [0.0530, 0.0957]	0.0319 [0.0065, 0.0566]
Writing	TL3	0.0751 [0.0534, 0.0966]	0.0465 [0.0219, 0.0699]
Writing	TL4	0.0785 [0.0575, 0.0999]	0.0561 [0.0335, 0.0788]
Writing	TL5	0.0878 [0.0669, 0.1086]	0.0556 [0.0340, 0.0789]
Writing	random	0.0896 [0.0669, 0.1124]	0.0657 [0.0391, 0.0910]
Writing	sentence	0.1021 [0.0783, 0.1263]	0.0984 [0.0696, 0.1275]

E VIEW AND DOMAIN DEPENDENCE

V0 denotes original text; V1 denotes identically normalized text re-scored under the declared scorer. V2-lite-mask removes marked word contributions while preserving their original cached context.

It is therefore distinct from V2-lite text re-scoring. The following tables use the single GPT-Neo scorer for likelihood, entropy and Fast-DetectGPT, rather than substituting pool-averaged entropy. All reported AUROCs retain the fixed raw orientation; a value below 0.5 is not replaced with its complement.

Table 8: Normalized V1 TL1 results on the historical ToBlend corpus (Huang et al., 2024), with prompt-cluster bootstrap intervals. The identity concerns scores and does not make AUROCs additive.

Pool	Domain	Likelihood [95% CI]	Entropy [95% CI]	FDG [95% CI]
Classic	SQuAD	0.4952 [0.4565, 0.5328]	0.6918 [0.6608, 0.7227]	0.8008 [0.7672, 0.8348]
Classic	XSum	0.3837 [0.3543, 0.4131]	0.7544 [0.7305, 0.7776]	0.6911 [0.6617, 0.7201]
Classic	Writing	0.7416 [0.7148, 0.7683]	0.5078 [0.4803, 0.5371]	0.9006 [0.8820, 0.9194]
Advanced	SQuAD	0.3520 [0.3151, 0.3881]	0.6764 [0.6452, 0.7088]	0.5248 [0.4815, 0.5655]
Advanced	XSum	0.2229 [0.1987, 0.2494]	0.7628 [0.7390, 0.7868]	0.3701 [0.3387, 0.4027]
Advanced	Writing	0.6270 [0.5969, 0.6576]	0.5384 [0.5102, 0.5669]	0.7829 [0.7571, 0.8090]

Table 9: All 14 Writing V1 entropy cells on the historical ToBlend corpus have AUROC at most 0.55 (0.5100 to 0.6061 on original text). The original ToBlend evaluation found entropy-based detection largely unaffected by blending relative to its baselines (Huang et al., 2024); entropy discrimination depends on domain and view.

Pool	Writing setting	Entropy AUROC [95% CI]
Classic	TL1	0.5078 [0.4803, 0.5371]
Classic	TL2	0.5302 [0.5016, 0.5600]
Classic	TL3	0.5340 [0.5050, 0.5631]
Classic	TL4	0.5028 [0.4702, 0.5373]
Classic	TL5	0.5168 [0.4882, 0.5457]
Classic	random	0.5321 [0.5033, 0.5577]
Classic	sentence	0.4602 [0.4308, 0.4914]
Advanced	TL1	0.5384 [0.5102, 0.5669]
Advanced	TL2	0.5127 [0.4842, 0.5435]
Advanced	TL3	0.5076 [0.4788, 0.5384]
Advanced	TL4	0.5086 [0.4796, 0.5372]
Advanced	TL5	0.5054 [0.4763, 0.5345]
Advanced	random	0.5199 [0.4897, 0.5514]
Advanced	sentence	0.4556 [0.4261, 0.4862]

Table 10: Outside-scorer likelihood/entropy sign-pattern counts. Each denominator of 24 is four scorers times six non-sentence settings. The planned SQuAD/XSum scope passes 96/96 in V0 and 93/96 (92/96 matched) in V1; Writing is weaker. Membership remains confounded with model family and size.

View	Pool	Domain	All	Length matched
V0	Classic	SQuAD	24/24	24/24
V0	Classic	XSum	24/24	24/24
V0	Classic	Writing	16/24	16/24
V0	Advanced	SQuAD	24/24	24/24
V0	Advanced	XSum	24/24	24/24
V0	Advanced	Writing	0/24	0/24
V1	Classic	SQuAD	24/24	24/24
V1	Classic	XSum	24/24	24/24
V1	Classic	Writing	12/24	12/24
V1	Advanced	SQuAD	21/24	20/24
V1	Advanced	XSum	24/24	24/24
V1	Advanced	Writing	0/24	0/24

F Matched Attribution, Mixing, and Switch Contrasts

Table 11: Attribution against logged generating models on matched TL1 blended text. Prompt, mixed-provenance and deleted-character words are excluded according to the logged alignment. Both identity accuracies exceed 25% uniform guessing and meet the planned weak-attribution rule. F1 uses exact adjacent-word boundaries and is a descriptive metric.

View	Word accuracy, % [95% CI]	Boundary F1 [95% CI]	Evaluable	Excluded
V0	36.53 [35.98, 37.12]	0.7407 [0.7360, 0.7454]	34,895	12,129
V2-lite	35.79 [35.24, 36.34]	0.7409 [0.7360, 0.7455]	33,350	11,624

Table 12: Mixing contrast: blended-versus-human AUROC minus fixed-model-versus-human AUROC. V0 entropy satisfies the partial-effect rule. V2-lite entropy is below the planned 0.05 material band (operational label *null*), despite a positive interval; it is not an estimate of exactly zero effect.

View	Entropy difference [95% CI]	Likelihood difference [95% CI]
V0	0.0516 [0.0158, 0.0884]	0.0189 [-0.0228, 0.0589]
V2-lite	0.0385 [0.0023, 0.0762]	0.0414 [0.0008, 0.0792]

Table 13: Logged switch contrasts; 200 permutations per null. Primary change-minus-redraw intervals include zero in both views. Secondary random-position contrasts and nominal p-values are retained. The boundary materiality fraction was never fixed, so the offset fractions do not settle a material-boundary claim.

View	Change − redraw [95% CI]	p	Change − random	p	Offset fraction
V0	0.0066 [-0.0155, 0.0299]	0.5920	-0.0107	0.0498	6.39%
V2-lite	0.0049 [-0.0195, 0.0296]	0.7313	-0.0148	0.0348	4.43%

Table 15: Blended versus fixed-model text on the 300 matched pairs (AUROC [95% CI]; blended positive). Calibrated S_{ens} z-scores each model's word log probability with human-word statistics from training folds. The reader (17 features) and char 1–4-gram classifiers are trained for this contrast with five grouped prompt folds. The frozen reader of Table 4 gives 0.5694 and 0.5642.

View	S_{ens}	Calibrated S_{ens}	Trained reader	Trained char
V0	0.5301 [0.4962, 0.5653]	0.5261 [0.4904, 0.5635]	0.5757 [0.5407, 0.6106]	0.4816 [0.4455, 0.5183]
V2-lite	0.5351 [0.5011, 0.5713]	0.5293 [0.4921, 0.5673]	0.5791 [0.5441, 0.6141]	0.4821 [0.4459, 0.5184]

G Format-Swap Controls

These controls transfer both human and single-model text into the blend surface format using the same NLTK token join. They do not apply V1 normalization and do not replace the fresh matched generation in Table 4. All 15 pool/domain/generator cells were computed. The reader additionally requires a held-out first-100 index and verified archived-prompt identity. Advanced retention is only 24 SQuAD, 17 XSum and 16 Writing pairs; the intervals and exclusions are essential to interpreting the point estimates.

Table 14: Format-swap AUROC for formatted single-model text (positive) versus equivalently formatted human text, after the 10% relative-length caliper. n_S and n_R are retained pair counts for the margin and reader. Excl. counts failed reader prompt-identity checks, separate from length and training-index exclusions. Twelve SQuAD/OPT pairs fail this check and are excluded.

Pool	Domain	Generator	n_S	S_{ens} [95% CI]	n_R	Reader [95% CI]	Excl.
Classic	SQuAD	GPT-J	283	0.4448 [0.4069, 0.4836]	95	0.7825 [0.7264, 0.8389]	0
Classic	SQuAD	GPT-Neo	271	0.3551 [0.3188, 0.3909]	90	1.0000 [1.0000, 1.0000]	0
Classic	SQuAD	GPT-2	285	0.2842 [0.2516, 0.3167]	96	0.9998 [0.9990, 1.0000]	0
Classic	SQuAD	OPT	285	0.3399 [0.3022, 0.3785]	83	0.9868 [0.9730, 0.9964]	12
Classic	Writing	GPT-J	432	0.3484 [0.3156, 0.3814]	87	0.7539 [0.6904, 0.8168]	0
Classic	Writing	GPT-Neo	438	0.3044 [0.2727, 0.3344]	88	0.9994 [0.9976, 1.0000]	0
Classic	Writing	GPT-2	437	0.2912 [0.2632, 0.3202]	86	1.0000 [1.0000, 1.0000]	0
Classic	Writing	OPT	434	0.3768 [0.3439, 0.4072]	84	0.9552 [0.9254, 0.9810]	0
Classic	XSum	GPT-J	473	0.5444 [0.5170, 0.5707]	98	0.6840 [0.6093, 0.7540]	0
Classic	XSum	GPT-Neo	480	0.5130 [0.4830, 0.5414]	97	1.0000 [1.0000, 1.0000]	0
Classic	XSum	GPT-2	486	0.4121 [0.3863, 0.4378]	100	0.9989 [0.9968, 1.0000]	0
Classic	XSum	OPT	483	0.5145 [0.4866, 0.5430]	96	0.9391 [0.9032, 0.9681]	0
Advanced	SQuAD	Control	24	0.4514 [0.3108, 0.6042]	24	0.8229 [0.7257, 0.9184]	0
Advanced	Writing	Control	16	0.3281 [0.1250, 0.5822]	16	0.8984 [0.7656, 1.0000]	0
Advanced	XSum	Control	17	0.4498 [0.3391, 0.5675]	17	0.8304 [0.6609, 0.9585]	0

H Structural and Perturbation Controls

The historical structural analysis contains 72 eligible readings: two pools $\times$ three domains $\times$ four settings (TL2–TL5) $\times$ three variants (raw, punctuation-excluded, V1-mapped). Fifty-nine are labeled slight tracking and 13 no tracking under the analysis rule. TL1 has no within-chunk word pairs and is non-estimable. The original period-k proxy is approximate, and the archived advanced XSum TL4 cell uses actual $k = 3$. A split-subword fragment exclusion rule was never fixed and is not applied; punctuation exclusion alone is not that rule. These counts concern approximate chunk structure, whereas Table 11 evaluates true logged identity.

The perturbation comparison passes the tested likelihood/entropy sign pattern in all 16 planned attack/generator/domain cases. It supports a general signature across these tested perturbations, not uniqueness to multi-model blending. Differences in attack and blending generation pipelines make cross-method magnitudes descriptive. No claim of a largest likelihood drop is inferred from unnormalized direction counts.

I LOCAL CPU REANALYSES

This appendix reports CPU reanalyses of the saved outputs, run after the main experiments with decision rules written before execution (released with the code). They refit no frozen model and generate no text. Bootstrap intervals resample canonical prompt clusters (2,000 replicates); because this reanalysis draws its own resamples, some interval bounds differ from those of Table 11 (by at most 0.03 percentage points for accuracy and 0.0004 for boundary F1), while all point estimates agree. Classifiers and calibrated margins here use a different assignment of prompt clusters to the five folds than the GPU run (Appendix J), so quantities reported in both differ slightly, for example the character classifier (0.4816 here, 0.4955 there); both runs lead to the same conclusions.

Table 16: Argmax word attribution against logged generators and exact adjacent-word boundary F1 on matched TL1 text. Uniform four-model chance is 25%. *Always*: F1 of predicting a boundary at every adjacent word pair; *Independent*: expected F1 of a truth-independent predictor with the observed boundary density. Fixed acc. applies the same argmax to the fixed-model arm.

View	Blended acc. (%)	Fixed acc. (%)	Boundary F1	Always	Independent
V0	36.53 [35.98, 37.12]	38.98 [38.03, 39.96]	0.7407 [0.7356, 0.7457]	0.8555	0.7343
V2-lite	35.79 [35.27, 36.35]	38.25 [37.26, 39.27]	0.7409 [0.7361, 0.7456]	0.8561	0.7340

Table 17: Per-generator argmax attribution accuracy (%) by true generator. The argmax is more accurate on fixed-model text than on blended text for every generator.

View	Arm	Llama-2	Phi-2	Mistral	Gemma	Words
V0	blended	31.9	42.2	33.9	38.2	34,895
V0	fixed	32.1	44.8	37.3	41.7	38,005
V2-lite	blended	31.5	42.5	32.7	36.5	33,350
V2-lite	fixed	32.0	44.8	35.8	40.4	36,709

Table 18: Boundary densities among evaluable adjacent word pairs and F1 differences from the two baselines. Always $= 2q/(1+q)$; independent $= 2pq/(p+q)$.

View	True rate q	Predicted rate p	F1 − always	F1 − independent
V0	0.7475	0.7216	-0.1148 [-0.1188, -0.1107]	0.0064 [0.0035, 0.0096]
V2-lite	0.7483	0.7202	-0.1151 [-0.1192, -0.1110]	0.0069 [0.0039, 0.0100]

Table 19: Raw and calibrated ensemble margins on matched text. Calibration statistics come from human words in training folds; orientation is fixed a priori (higher means generated or blended).

View	S_{ens} variant	Blended vs. fixed	Blended vs. human	Fixed vs. human
V0	Raw	0.5301 [0.4962, 0.5653]	0.5604 [0.5216, 0.5988]	0.5230 [0.4830, 0.5621]
V0	Standardized	0.5261 [0.4904, 0.5635]	0.6062 [0.5681, 0.6454]	0.5644 [0.5216, 0.6072]
V0	Centered	0.5302 [0.4939, 0.5681]	0.6402 [0.6046, 0.6779]	0.5913 [0.5505, 0.6323]
V2-lite	Raw	0.5351 [0.5011, 0.5713]	0.5011 [0.4643, 0.5394]	0.4638 [0.4237, 0.5026]
V2-lite	Standardized	0.5293 [0.4921, 0.5673]	0.5561 [0.5179, 0.5954]	0.5218 [0.4795, 0.5639]
V2-lite	Centered	0.5332 [0.4973, 0.5704]	0.5912 [0.5542, 0.6283]	0.5496 [0.5080, 0.5895]

Table 22: Matched generation. (a) Blend specificity on the 300 advanced-pool TL1 pairs (blended positive); trained detectors use five grouped prompt folds, RoBERTa is the planned primary, and centered S_{ens} subtracts the mean human-word log probability of each model from training folds (further variants and zero-shot scores: Table 27). (b) Word attribution against logged generators on the blended arm (V0, 300 passages per setting, chance 25%). Sampling-aware: argmax of word log probability under the logged decoding transform of each member (floor 10^{-6}); exact-context: argmax of the probability of each logged chunk given the context the loop encoded; *Always*: F1 of marking every adjacent word pair as a boundary.

(a) Detector	V0 AUROC [95% CI]	V2-lite AUROC [95% CI]
Reader (17 features), trained	0.5757 [0.5407, 0.6106]	0.5791 [0.5441, 0.6141]
RoBERTa, trained	0.5443 [0.5090, 0.5800]	0.5361 [0.5026, 0.5716]
Char 1–4-gram, trained	0.4955 [0.4586, 0.5319]	0.4952 [0.4588, 0.5311]
S_{ens}, raw	0.5301 [0.4962, 0.5653]	0.5351 [0.5011, 0.5713]
S_{ens}, centered per model	0.5293 [0.4935, 0.5668]	0.5329 [0.4975, 0.5701]
Reader frozen on V2-lite text	–	0.5670 [0.5266, 0.6067]

(b) Setting	Attributor	Accuracy, % [95% CI]	Boundary F1	Always
Adv. TL1	Argmax, raw	36.53 [35.94, 37.12]	0.7407	0.8555
Adv. TL1	Sampling-aware	41.31 [40.70, 41.88]	0.7029	0.8555
Adv. TL1	Exact-context	41.25 [40.65, 41.84]	0.7014	0.8558
Adv. TL3	Sampling-aware	41.33 [40.73, 41.99]	0.3798	0.4032
Cls. TL1	Sampling-aware	38.19 [37.60, 38.79]	0.7568	0.8585

Table 20: Calibrated permutation null for the reader in the 36 gate cells: 1,000 random prompt-group label flips per cell, refitting the reader in the frozen folds. *Single draw outside* counts cells whose recorded single-realization null point lies outside the central 95% of that cell's permutation distribution. p is the one-sided permutation p-value of the observed reader AUROC. Treating cells as independent, a leakage-free reader keeps all 18 null points inside $[0.45, 0.55]$ with probability 0.13 (classic) and 0.21 (advanced); every recorded single-draw null point is reproduced exactly.

Pool	Cells	Null SD	Null 2.5–97.5% range	Single draw outside	Max. p
Classic	18	0.0259–0.0351	0.4260–0.5674	2	0.0010
Advanced	18	0.0242–0.0336	0.4288–0.5615	0	0.0010

Table 21: TPR at a 5% FPR threshold across the 18 gate cells per pool, from out-of-fold predictions on identical held-out rows. The threshold is the 95th percentile of calibration-prompt human scores; TPR and FPR are measured on the other prompts. TPR at 1% FPR is not estimated: calibration sets hold 66 to 123 human texts per cell.

Pool	Detector	TPR min	TPR median	TPR max	Test FPR
Classic	Reader	0.837	0.947	0.978	0.014–0.084
Classic	RoBERTa	0.915	0.975	0.995	0.022–0.119
Classic	Char	0.316	0.517	0.665	0.013–0.108
Advanced	Reader	0.497	0.774	0.885	0.020–0.142
Advanced	RoBERTa	0.872	0.953	0.978	0.018–0.087
Advanced	Char	0.272	0.611	0.736	0.030–0.088

J GPU Reanalyses and New Matched Arms

This appendix reports the second analysis run (57 jobs on four A100 80 GB cards, completed 24 September 2026). Its decision rules were written before any job ran (file dated 23 September 2026,

SHA-256 prefix `220dd9582df6c3b3`; the RADAR-control and identical-input rules were added on 24 September 2026, also before their jobs ran). All intervals are 95% prompt-cluster bootstrap intervals. The second outside scorer is OLMo-2-7B (`allenai/OLMo-2-1124-7B`) (Walsh et al., 2025), used because Qwen2.5-7B was not available; Binoculars uses Falcon-7B and Falcon-7B-Instruct (Almazrouei et al., 2023; Hans et al., 2024). The new matched arms (advanced pool at TL3, classic pool at TL1; 100 prompts per domain and arm) follow the design of Section 3.4 but sample in micro-batches, so a sampled chunk depends on its batch and per-chunk seeds are batch seeds. The raw-argmax row of Table 22 is recomputed in this run; its point estimate equals Table 11, and its interval differs only through resampling.

Table 23: RADAR fusion controls on the 18 cells with RADAR scores (SQuAD, XSum and Writing at TL1, random and sentence; nested per-cell V0 design of Table 33, 1,000 paired draws, Holm within each nine-cell family). Ranges are over the nine cells; Holm+ counts positive Holm-significant differences. The reader+RADAR row reproduces the published gains. RoBERTa reads V2-lite text; the char classifier reads V0 text. The stacked RoBERTa variant is a sensitivity analysis that can leak outer labels into features.

Pool	Detector	Fused − RADAR	Holm+	Alone − RADAR	Fused − alone	Holm+
Classic	Reader	0.090 to 0.211	9/9	0.080 to 0.209	0.001 to 0.010	5/9
Classic	Char 1–4-gram	0.104 to 0.224	9/9	0.104 to 0.224	0.000 to 0.003	1/9
Classic	RoBERTa (nested)	0.065 to 0.171	9/9	0.061 to 0.201	-0.036 to 0.005	0/9
Classic	RoBERTa (stacked)	0.068 to 0.176	9/9	0.061 to 0.201	-0.035 to 0.007	0/9
Advanced	Reader	0.052 to 0.177	9/9	0.040 to 0.173	0.004 to 0.031	9/9
Advanced	Char 1–4-gram	0.069 to 0.194	9/9	0.069 to 0.194	0.000 to 0.002	1/9
Advanced	RoBERTa (nested)	0.056 to 0.151	9/9	0.058 to 0.177	-0.026 to -0.003	0/9
Advanced	RoBERTa (stacked)	0.055 to 0.140	9/9	0.058 to 0.177	-0.042 to -0.003	0/9

Table 24: The detector rule with the reader rescored on the same V2-lite text that RoBERTa and char read (18 gate cells per pool; verdict: no detector claim). *Beats RoBERTa* and *RoBERTa better* count paired CIs of reader minus RoBERTa above and below zero. No cell loses more than 0.10 AUROC from V0 (0 of 36).

Pool	Reader AUROC	AUROC ≥ 0.85	Beats RoBERTa	RoBERTa better	Beats single	Nulls in band
Classic	0.983 to 0.995	18/18	5/18	1/18	18/18	15/18
Advanced	0.851 to 0.964	18/18	0/18	18/18	18/18	17/18

Table 25: Attribution on the blended arm of the matched advanced-pool TL1 generation (300 passages, 299 prompt clusters). The floor is the probability given to tokens that a member's decoding transform removes. *Independent*: truth-independent predictor with the observed predicted boundary density; *Random*: uniformly random labels.

View	Attributor, floor	Accuracy, %	Boundary F1	Always	Independent	Random	F1 − always
V0	Argmax, raw	36.53 [35.94, 37.12]	0.7407 [0.7361, 0.7455]	0.8555	0.7343	0.7488	-0.1148 [-0.1188, -0.1107]
V0	Sampling-aware, 10^{-4}	41.25 [40.64, 41.83]	0.7024 [0.6955, 0.7089]	0.8555	0.6863	0.7488	-0.1531 [-0.1593, -0.1471]
V0	Sampling-aware, 10^{-6}	41.31 [40.70, 41.88]	0.7029 [0.6960, 0.7093]	0.8555	0.6865	0.7488	-0.1527 [-0.1590, -0.1465]
V0	Sampling-aware, 10^{-8}	41.31 [40.70, 41.89]	0.7029 [0.6960, 0.7093]	0.8555	0.6864	0.7488	-0.1527 [-0.1590, -0.1465]
V0	Sampling-aware, Gemma FP32	41.28 [40.68, 41.86]	0.7029 [0.6961, 0.7093]	0.8555	0.6862	0.7488	-0.1526 [-0.1588, -0.1465]
V2-lite	Argmax, raw	35.79 [35.23, 36.34]	0.7409 [0.7364, 0.7458]	0.8561	0.7340	0.7492	-0.1151 [-0.1190, -0.1110]
V2-lite	Sampling-aware, 10^{-4}	40.09 [39.46, 40.65]	0.7033 [0.6971, 0.7093]	0.8561	0.6891	0.7492	-0.1527 [-0.1588, -0.1469]
V2-lite	Sampling-aware, 10^{-6}	40.13 [39.50, 40.70]	0.7033 [0.6971, 0.7093]	0.8561	0.6890	0.7492	-0.1527 [-0.1588, -0.1468]
V2-lite	Sampling-aware, 10^{-8}	40.13 [39.50, 40.70]	0.7034 [0.6971, 0.7093]	0.8561	0.6890	0.7492	-0.1527 [-0.1588, -0.1468]
V2-lite	Sampling-aware, Gemma FP32	40.15 [39.52, 40.71]	0.7034 [0.6972, 0.7093]	0.8561	0.6893	0.7492	-0.1526 [-0.1587, -0.1465]
V0	Exact-context, 10^{-4}	41.15 [40.56, 41.75]	0.7010 [0.6940, 0.7078]	0.8558	0.6852	0.7490	-0.1548 [-0.1614, -0.1484]
V0	Exact-context, 10^{-6}	41.25 [40.65, 41.84]	0.7014 [0.6944, 0.7081]	0.8558	0.6853	0.7490	-0.1544 [-0.1610, -0.1480]
V0	Exact-context, 10^{-8}	41.25 [40.65, 41.84]	0.7014 [0.6943, 0.7081]	0.8559	0.6852	0.7490	-0.1545 [-0.1612, -0.1481]

Table 26: Overall and per-generator attribution accuracy (%, V0, floor 10^{-6}). Llama-2 is the only member sampled at temperature 0.6 with top-p 0.9; the others use temperature 1 and top-p 1 (all top-k 50).

Arm	Attributor	Accuracy, %	Llama-2	Phi-2	Mistral	Gemma
Blended	Argmax, raw	36.53 [35.94, 37.12]	31.9	42.2	33.9	38.2
Blended	Sampling-aware	41.31 [40.70, 41.88]	78.8	32.5	25.9	27.8
Blended	Exact-context	41.25 [40.65, 41.84]	78.7	32.4	25.7	27.8
Fixed	Argmax, raw	38.98 [37.98, 39.94]	32.1	44.8	37.3	41.7
Fixed	Sampling-aware	43.18 [40.66, 45.96]	82.7	32.9	28.1	29.1
Fixed	Exact-context	43.18 [40.48, 46.12]	84.1	32.1	26.9	28.4

The sampling-aware scores treat each word independently. The exact-context attributor scores every logged chunk under each member given exactly the text the loop encoded at that step; with a uniform per-step draw, its argmax is the maximum-likelihood chunk attributor under the logged decoding distributions, up to NLTK re-spacing: the retained chunk's tokenization extends the prompt's tokenization for all four members in 99.26% of steps. Its chunk-level accuracy is 40.50% on the blended arm and 42.10% on the fixed arm. Rescoring Gemma-7B in FP32 on 2,400 matched texts changes word log probabilities by up to 5.34 nats (median over texts of the per-text maximum 3.13), but sampling-aware accuracy only from 41.31% to 41.28%.

Table 27: Blend specificity on the 300 matched advanced-pool TL1 pairs. Trained detectors use five grouped prompt folds (RoBERTa: three seeds, 15 fits per view). Calibrated S_{ens} variants use human-word statistics from training folds only. The planned rule requires a CI lower bound of at least 0.65 for a specificity claim from a trained detector and above 0.55 for a zero-shot signal; no entry meets either.

Family	Detector	Blended vs. fixed AUROC [95% CI]
Trained	RoBERTa, V0	0.5443 [0.5090, 0.5800]
Trained	RoBERTa, V2-lite	0.5361 [0.5026, 0.5716]
Trained	Char 1–4-gram, V0	0.4955 [0.4586, 0.5319]
Trained	Char 1–4-gram, V2-lite	0.4952 [0.4588, 0.5311]
S_{ens}	raw, raw word scores, V0	0.5301 [0.4962, 0.5653]
S_{ens}	centered, raw word scores, V0	0.5293 [0.4935, 0.5668]
S_{ens}	standardized, raw word scores, V0	0.5248 [0.4889, 0.5621]
S_{ens}	raw, raw word scores, V2-lite	0.5351 [0.5011, 0.5713]
S_{ens}	centered, raw word scores, V2-lite	0.5329 [0.4975, 0.5701]
S_{ens}	standardized, raw word scores, V2-lite	0.5285 [0.4919, 0.5663]
S_{ens}	raw, decoding-transformed scores, V0	0.4560 [0.4158, 0.4980]
S_{ens}	centered, decoding-transformed scores, V0	0.5271 [0.4917, 0.5611]
S_{ens}	standardized, decoding-transformed scores, V0	0.5064 [0.4671, 0.5445]
S_{ens}	raw, decoding-transformed scores, V2-lite	0.4715 [0.4306, 0.5138]
S_{ens}	centered, decoding-transformed scores, V2-lite	0.5324 [0.4951, 0.5685]
S_{ens}	standardized, decoding-transformed scores, V2-lite	0.5106 [0.4709, 0.5490]
Zero-shot	Binoculars (Falcon-7B pair), V0	0.5357 [0.4913, 0.5768]
Zero-shot	Falcon-7B observed-analytic score, V0	0.5391 [0.4976, 0.5804]
Zero-shot	Fast-DetectGPT, GPT-Neo, V0	0.5404 [0.4992, 0.5787]
Zero-shot	Fast-DetectGPT, OLMo-2-7B, V0	0.5179 [0.4803, 0.5563]

Table 28: Attribution on the new matched arms (blended arm, V0, 300 passages each; fixed accuracy: raw / sampling-aware; boundary F1: raw / sampling-aware). *True rate* is the share of adjacent word pairs with a generator change.

Arm	Argmax acc., %	Sampling-aware acc., %	Fixed acc., %	Boundary F1	Always	True rate
Advanced, TL3	37.44 [36.86, 38.02]	41.33 [40.73, 41.99]	37.37 / 41.75	0.3855 / 0.3798	0.4032	0.2525
Classic, TL1	38.20 [37.61, 38.80]	38.19 [37.60, 38.79]	39.51 / 39.30	0.7556 / 0.7568	0.8585	0.7521

Table 29: Blend specificity of the margin and the logged switch contrast (word discrepancy at chunk starts where the model changes minus chunk starts where the same model is redrawn) on the new matched arms.

Arm	S_{ens} blended vs. fixed	Change − redraw	Changes	Redraws
Advanced, TL3	0.5434 [0.5031, 0.5854]	-0.0109 [-0.0474, 0.0269]	10,824	3,513
Classic, TL1	0.5166 [0.4808, 0.5556]	0.0120 [-0.0091, 0.0333]	31,788	10,385

K Scoring Algorithm and Historical Context

The historical scoring experiments characterize detection, score composition, and predicted spans. Historical per-cell results use 300 human and 300 blended passages for SQuAD and 500 of each for XSum and Writing, unless a table states otherwise. They do not replace the matched-generation experiment or the final prompt-grouped reader comparisons. Archived advanced XSum TL4 means actual $k = 3$ throughout.

Algorithm 1 Reverse ToBlend word-aligned scoring and separate evaluation

Input: passage x; candidate models $\mathcal{M}$; chosen view and word alignment; frozen reader f when applicable

Output: margin S_{ens}; reader score; predicted word labels and runs

for each model $m \in \mathcal{M}$ **do**
  Score the full passage under m and retain token log probabilities and moments.
  Align model tokens to the shared whitespace words, summing log probabilities into $\ell_{m,w}$.
end for
for each retained word $w \in W$ **do**
  $a_w \leftarrow \arg\max_m \ell_{m,w}$; $s_w \leftarrow \max_m \ell_{m,w} - K^{-1} \sum_m \ell_{m,w}$
end for
$S_{\text{ens}} \leftarrow |W|^{-1} \sum_{w \in W} s_w$
Form the four declared component summaries per model and append S_{ens}.
Apply the training-fitted scaler and frozen reader f when the protocol calls for it.
Form maximal constant-a_w runs as predicted spans, not verified provenance.
Evaluate human/generated detection, blend/fixed specificity, and log-aligned attribution separately.

Word-level attribution is evaluated here only against logged generators; this paper evaluates no LLM-based judge and makes no explanation-faithfulness claim.

K.1 Historical Detection Battery

The historical blended corpus comes from ToBlend (Huang et al., 2024), which proposes token-level blending as an attack on AI-generated text detection. That work reports the detector battery on this corpus (a single-model replication of Fast-DetectGPT and the seven-detector battery for both pools). This paper does not reproduce the battery tables. It uses the historical corpus only as input to its own reader, margin, decomposition and switch analyses, all reported in this paper's own tables.

K.2 Score Decomposition and Numerical Correlation

Figure 5 plots the aligned historical detector records for individual passages in each domain. Dashed lines are fitted trends, not equality lines. In cross-entropy notation the numerator contains $H - \mathrm{CE}$, since log likelihood is negative cross-entropy; it is not $\mathrm{CE} + H$. The varying variance denominator is still essential. The raw pool/domain summary in Table 30 and the complete per-cell table report historical descriptive measurements, while the normalized analysis and adjusted regularity test in the main text bound its interpretation.

Table 30: Historical original-view mean AUROC over the six non-sentence granularities. The columns are three distinct discrimination measurements; they are not additive terms in an AUROC identity. The adjusted regularity and normalized-view results in the main text supersede a causal or universal-survivor interpretation.

Domain	Pool	likelihood	entropy	curvature
SQuAD	classic	0.48	0.74	0.85
SQuAD	advanced	0.35	0.71	0.58
XSum	classic	0.36	0.78	0.71
XSum	advanced	0.22	0.80	0.43
Writing	classic	0.72	0.58	0.94
Writing	advanced	0.63	0.57	0.83

Table 31: Historical original-view per-cell likelihood, entropy, and FDG AUROCs, with Spearman correlation of per-instance FDG and likelihood plus entropy. Near-one correlation is not equality of normalized scores. The dagger denotes advanced XSum TL4 with actual $k = 3$.

Pool	Cell	Lik.	Ent.	FDG	ρ
classic	SQuAD TL1	0.435	0.747	0.805	0.996
classic	SQuAD TL2	0.471	0.755	0.864	0.995
classic	SQuAD TL3	0.496	0.721	0.858	0.994
classic	SQuAD TL4	0.485	0.733	0.856	0.994
classic	SQuAD TL5	0.512	0.721	0.869	0.993
classic	SQuAD random	0.462	0.744	0.828	0.996
classic	XSum TL1	0.344	0.797	0.709	0.997
classic	XSum TL2	0.353	0.780	0.700	0.998
classic	XSum TL3	0.370	0.771	0.717	0.998
classic	XSum TL4	0.372	0.771	0.729	0.997
classic	XSum TL5	0.374	0.764	0.725	0.997
classic	XSum random	0.347	0.779	0.703	0.998
classic	Writing TL1	0.703	0.571	0.913	0.993
classic	Writing TL2	0.729	0.579	0.945	0.992
classic	Writing TL3	0.718	0.584	0.938	0.994
classic	Writing TL4	0.744	0.557	0.944	0.994
classic	Writing TL5	0.741	0.571	0.948	0.993
classic	Writing random	0.712	0.588	0.938	0.993
advanced	SQuAD TL1	0.297	0.736	0.532	0.999
advanced	SQuAD TL2	0.345	0.710	0.580	0.998
advanced	SQuAD TL3	0.364	0.699	0.582	0.999
advanced	SQuAD TL4	0.384	0.696	0.618	0.998
advanced	SQuAD TL5	0.376	0.697	0.602	0.998
advanced	SQuAD random	0.354	0.697	0.568	0.998
advanced	XSum TL1	0.194	0.812	0.397	0.999
advanced	XSum TL2	0.219	0.800	0.448	0.999
advanced	XSum TL3	0.204	0.806	0.420	0.999
advanced	XSum TL4†	0.249	0.772	0.458	0.998
advanced	XSum TL5	0.240	0.777	0.450	0.999
advanced	XSum random	0.212	0.802	0.428	0.999
advanced	Writing TL1	0.584	0.606	0.804	0.998
advanced	Writing TL2	0.635	0.562	0.827	0.997
advanced	Writing TL3	0.646	0.557	0.840	0.996
advanced	Writing TL4	0.654	0.563	0.857	0.998
advanced	Writing TL5	0.652	0.553	0.845	0.997
advanced	Writing random	0.636	0.577	0.835	0.997

K.3 Historical Pool Ablation

The historical advanced-pool ablation below reports the full-pool raw S_{ens}, each leave-one-model-out score, and the effect of skipping the copied prefix. It is exploratory and predates the final prompt-grouped evaluation. Raw score orientation is retained; a below-chance score is not automatically flipped.

Table 32: Historical advanced pool composition and prefix sensitivity: full-pool raw S_{ens}, each leave-one-model-out score, and skipping the copied 30-word prefix. Effects depend on model and domain; advanced XSum prefix exclusion raises all six cells by 0.0266 to 0.0396, so a uniformly negligible prefix effect is not claimed. These descriptive score ablations do not prove provenance recovery.

Cell	full	−gemma	−llama	−mistral	−phi	skip30
SQuAD TL1	0.70	0.65	0.72	0.70	0.70	0.70
SQuAD TL2	0.69	0.62	0.71	0.69	0.70	0.69
SQuAD TL3	0.67	0.59	0.70	0.67	0.68	0.67
SQuAD TL4	0.67	0.60	0.70	0.67	0.68	0.67
SQuAD TL5	0.67	0.58	0.69	0.67	0.68	0.66
SQuAD random	0.68	0.60	0.70	0.69	0.69	0.68
XSum TL1	0.71	0.65	0.66	0.67	0.82	0.74
XSum TL2	0.70	0.62	0.64	0.65	0.80	0.73
XSum TL3	0.72	0.64	0.66	0.68	0.83	0.75
XSum TL4	0.69	0.62	0.63	0.64	0.80	0.72
XSum TL5	0.72	0.63	0.65	0.67	0.83	0.75
XSum random	0.73	0.65	0.67	0.68	0.83	0.77
Writing TL1	0.51	0.42	0.50	0.49	0.59	0.49
Writing TL2	0.47	0.37	0.46	0.45	0.54	0.45
Writing TL3	0.45	0.36	0.45	0.44	0.53	0.44
Writing TL4	0.45	0.35	0.44	0.43	0.52	0.43
Writing TL5	0.45	0.37	0.45	0.44	0.52	0.44
Writing random	0.46	0.37	0.46	0.45	0.55	0.45

K.4 Aligned Boundary and Fusion Evidence

L Pooled RADAR Fusion and Switch Contrasts

Table 33: All 54 per-cell RADAR fusion comparisons. Each pool/method is a nine-cell multiplicity family, with 1,000 paired bootstrap replicates. The nested reader uses all complete-feature prompts with training separated from held-out folds. The first 100 arm is a disjoint-holdout sensitivity and does not replace the full per-cell reader evaluation.

Pool	Added feature	n	Δ range	BH	Holm
C	Margin	9	-0.0253 – 0.0054	0	0
C	First 100 reader	9	0.0702 – 0.2195	9	9
C	Nested reader	9	0.0901 – 0.2111	9	9
A	Margin	9	-0.0105 – 0.0292	0	0
A	First 100 reader	9	0.0305 – 0.1805	9	9
A	Nested reader	9	0.0523 – 0.1766	9	9

Table 34: Pooled TL2–TL5 aligned switch-minus-off-switch effects, 5,200 passages per row. Intervals bootstrap passages. Separately, z and p compare the switch mean to count-matched post-prompt random positions, with the two-sided 200 permutation plus-one correction. Negative mapped blend effects are detectable, but the material fraction was not fixed and no material-boundary conclusion is licensed. Humans provide separate null controls. Unmapped normalization is retained as a grid sensitivity, not the declared aligned result.

View	Pool	Class	Effect [CI]	z	p
V0	C	Blend	-0.0179 [-0.0244, -0.0118]	-6.27	0.0050
V0	C	Human	-0.0012 [-0.0061, 0.0036]	-0.46	0.7264
V1 unmap.	C	Blend	0.0009 [-0.0057, 0.0071]	0.47	0.6965
V1 unmap.	C	Human	-0.0014 [-0.0064, 0.0033]	-0.59	0.5771
V1 mapped	C	Blend	-0.0125 [-0.0193, -0.0059]	-4.12	0.0050
V1 mapped	C	Human	0.0014 [-0.0032, 0.0060]	0.42	0.7214
V0	A	Blend	-0.0176 [-0.0235, -0.0118]	-6.15	0.0050
V0	A	Human	-0.0008 [-0.0058, 0.0038]	-0.75	0.4677
V1 unmap.	A	Blend	-0.0028 [-0.0086, 0.0029]	0.05	0.9652
V1 unmap.	A	Human	-0.0015 [-0.0066, 0.0032]	-1.11	0.2537
V1 mapped	A	Blend	-0.0146 [-0.0202, -0.0085]	-4.05	0.0050
V1 mapped	A	Human	0.0026 [-0.0023, 0.0072]	0.88	0.4129