TOKEN-LEVEL MODEL BLENDING AND AI-GENERATED TEXT DETECTION: SEPARATING MODEL SWITCHING FROM THE GENERATION LOOP

Anonymous authors
Paper under double-blind review

ABSTRACT

Mixing several language models within one passage is a proposed way to evade detectors of AI-generated text. Token-level blending, which draws a model at random for each short chunk, lowers zero-shot detection scores, yet it also changes the generation procedure, leaving unclear whether detectors are vulnerable to model mixing at all. We separate the two by running every pool member alone through the identical procedure. Encouragingly for detection, switching itself adds no measurable evasion: four older open models run alone are as hard to detect as their blends (differences of at most 0.0071). When blends and single models are generated together, switching even raises detectability, by 0.0411 to 0.1049 for four newer models and 0.0271 to 0.0655 for four instruction-tuned models. At the same time, the chunked procedure itself creates a large gap: on news articles, Fast-DetectGPT AUROC is 0.7088 for the older models' blends and 0.9283 for their standard text. For the newer models, switching lowers detectability only relative to single models regenerated separately with the original procedure. We examine these patterns across three English domains and several detector families. The results suggest that model switching alone is not a route around current detectors and that evasion claims need same-pipeline single-model baselines.

1 INTRODUCTION

Large language models (OpenAI, 2022; Touvron et al., 2023) produce text that human readers often cannot tell apart from human writing (Ippolito et al., 2020; Clark et al., 2021); readers even rated ChatGPT's explanations of implicit hate speech as clearer than human-written ones (Huang et al., 2023). Detecting machine-generated text remains difficult when the generation process differs from the detector's reference model or training data (Sadasivan et al., 2025; Mitchell et al., 2023; Bao et al., 2024; Dugan et al., 2024). Existing attacks mostly modify single-model outputs by paraphrasing or lexical edits (Krishna et al., 2023; Sadasivan et al., 2025; Wolff & Wolff, 2020). We study a different construction: a passage assembled through repeated continuations from multiple models.

TOBLEND draws one member of a fixed model pool at each generation step, appends a short decoded continuation, and repeats (Figure 1). It uses no training or detector feedback. A detection gap between such text and standard single-model text can have two sources: switching between generators, whose chunks no single reference model matches, or the chunked loop itself, which re-tokenizes the growing text after every chunk and, in its original form, feeds each model NLTK-re-spaced text. Separating the two requires single models run through the same loop.

We therefore run every pool member alone through the identical loop (pool-of-one), in the original loop and in a native loop that keeps each model's own output text, and ask four questions. **RQ1:** How does blended text compare with standard single-model text under a fixed scorer, after formatting controls and against paraphrase attacks? **RQ2:** With the loop fixed, does model switching change detectability? **RQ3:** Is the reduction larger for the newer, advanced pool than for the older, classic pool, across scorers, references, and loops? **RQ4:** What quality and time costs accompany blending?

Our contributions are: (i) a controlled measurement design for multi-model generation, with pool-of-one controls through the identical loop, the original and native loops, three generation seeds, three Fast-DetectGPT scorers, Binoculars and supervised detectors, and DIPPER and instruction-

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.

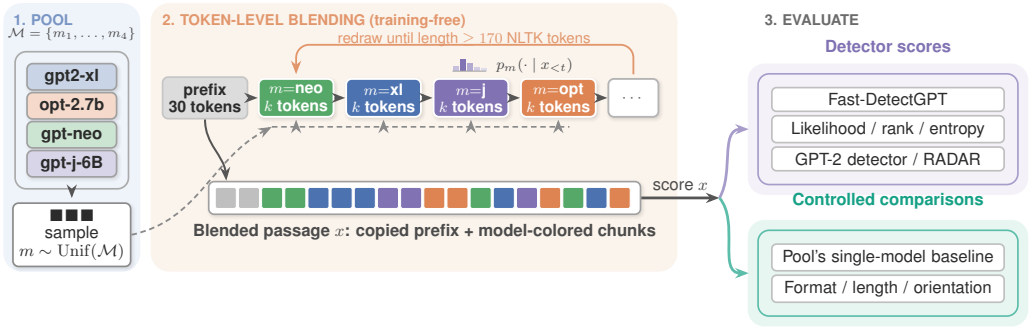


Figure 1: The buffered TOBLEND pipeline (classic pool illustrated). A human prefix of 30 NLTK (Bird et al., 2009) tokens seeds the passage; all counts in the figure are NLTK tokens. Each step draws one of four models uniformly, generates $k + 3$ model tokens, decodes and retains k NLTK tokens, then re-tokenizes the appended text for the next model. The stop rule counts at least 170 NLTK tokens including the prefix. A redraw repeats the same model with probability $1/4$. Detector scores are compared with standard single-model text and with each pool member run alone through the same loop (pool-of-one). Sentence-level and unbuffered protocols differ (appendix).

tuned paraphrasing; (ii) the finding that the classic-pool detection gap comes from the loop, not from switching, whereas, in the matched native loop, switching raises detectability in the advanced and instruction-tuned pools; (iii) the finding that the advanced pool's larger reduction depends on the scorer, the reference, and the loop; and (iv) measured quality and generation-time costs. A concurrent anonymous submission (Anonymous, 2027) analyzes the detection signals that remain under blending; every result reported here is computed from this paper's own data.

2 Related Work

Detectors. Zero-shot detectors score text with a language model: likelihood, rank, and entropy statistics (Gehrmann et al., 2019; Ippolito et al., 2020), probability curvature (Mitchell et al., 2023; Bao et al., 2024), log-rank ratios (Su et al., 2023), regeneration divergence (Yang et al., 2024), and cross-perplexity between two models (Hans et al., 2024). Supervised detectors train classifiers on human and machine text (Solaiman et al., 2019; Guo et al., 2023; Liu et al., 2023; Verma et al., 2024), with adversarial training or adversarial in-context examples for robustness (Hu et al., 2023; Koike et al., 2024), and representation-based detectors use a surrogate model's hidden states (Chen et al., 2025). Watermarking instead embeds a signal at generation time (Kirchenbauer et al., 2023); surveys cover these families (Tang et al., 2024). Benchmarks show that detector rankings change across generators, domains, decoding strategies, and attacks (Dugan et al., 2024; Wu et al., 2024).

Attacks and mixed authorship. Paraphrasing evades many detectors (Krishna et al., 2023; Sadasivan et al., 2025), as do character- and word-level edits (Wolff & Wolff, 2020); rewriting-based detectors respond to such edits (Mao et al., 2024; Hao et al., 2025). Text that mixes human and machine writing is harder to detect than purely machine text (Zhang et al., 2024). These settings change authorship after generation. TOBLEND instead mixes machine generators during generation.

Multi-model decoding. Collaborative decoding interleaves tokens from several models to improve task performance, for example by learning when an assistant model should generate (Shen et al., 2024). TOBLEND samples one model per chunk uniformly at random, without a learned router, a quality objective, or mixing of next-token distributions. We ask how this model switching affects detectors whose reference model or training data assume a single generator per passage.

3 Method and Evaluation

Generation protocol. The classic pool contains GPT-2-XL (1.5B), OPT-2.7B, GPT-Neo-2.7B, and GPT-J-6B (Radford et al., 2019; Zhang et al., 2022; Black et al., 2021; Wang & Komatsuzaki,

Algorithm 1 Buffered ToBlend generation for one prompt.

Require: pool $\mathcal{M} = \{m_1, \ldots, m_4\}$, human prefix x of 30 NLTK tokens, chunk size k, buffer $b = 3$
1: **while** x has fewer than 170 NLTK tokens **do**
2: draw $m \sim \mathrm{Uniform}(\mathcal{M})$; draw $k \sim \mathrm{Uniform}\{1, \ldots, 5\}$ in the random setting
3: tokenize x with m's tokenizer and sample $k + b$ new model tokens from m
4: decode the continuation, split it into NLTK tokens, and append the first k to x
5: **end while**
6: **return** x (NLTK tokens joined by single spaces)

2021). The advanced pool contains Llama-2-7B, Phi-2 (2.7B), Mistral-7B-v0.3, and Gemma-7B (Touvron et al., 2023; Javaheripi & Bubeck, 2023; Jiang et al., 2023; Gemma Team et al., 2024a). "Advanced" denotes this fixed pool of newer models, not a measured ranking of model capability or diversity. An instruction-tuned pool, Llama-3.1-8B-Instruct, Qwen2.5-7B-Instruct, Mistral-7B-Instruct-v0.3, and Gemma-2-9B-it (Grattafiori et al., 2024; Qwen Team et al., 2024; Jiang et al., 2023; Gemma Team et al., 2024b), runs without a chat template in the native loop only.

Algorithm 1 gives the buffered procedure; we write TLk for a token-level setting that retains k NLTK tokens per draw. The complete string is re-tokenized for the next model. We test $k = 1, 2, 3, 4, 5$ and a uniformly random $k \in \{1, \ldots, 5\}$. The nominal stop condition is at least 170 NLTK tokens *including* the 30-token prompt: saved token-level passages contain 170–180 NLTK tokens, 140–150 newly generated tokens. Sentence-level generation requests at most 50 model tokens and cuts at the first period, without a buffer. The advanced XSum TL4 corpus retained $k = 3$ per draw; a corpus regenerated with $k = 4$ has Fast-DetectGPT AUROC 0.6688 [0.6376, 0.6983] against 0.4575 [0.4254, 0.4892] for the $k = 3$ corpus and replaces it for any TL4 statement (Appendix A); the 2024 battery tables keep the $k = 3$ cell, marked.

Each generation notebook calls `set_seed(42)` once, with no per-passage seed. Models run in fp32 with sampling enabled and no explicit temperature, top-k, or top-p arguments. Recorded public defaults are temperature 1.0, top-k 50, and top-p 1.0; the Llama-2 generation configuration used in 2024 was not recorded. All scores use the saved texts; these gaps limit exact regeneration.

Loop variants and pool-of-one controls. The 2024 blends come from the *original loop*: each step re-encodes the whole NLTK-re-spaced text, decodes the new span with special tokens kept, and appends k NLTK tokens joined by spaces, so every model conditions on re-spaced text. A *native loop* changes only the string handling: it appends the prefix of the model's own decoded string that ends at the k-th NLTK token, so its text needs no cleaning. A *pool-of-one* arm runs one fixed pool member through the identical loop, which separates model switching from the loop. Original-loop pool-of-one arms were regenerated in half precision with batched sampling, whereas the 2024 blends were generated in fp32, so that comparison is matched in code but not in time; native-loop blends (three seeds) and pool-of-one arms were generated together and are matched in both. A native-loop sweep varies the per-step redraw probability $p \in \{0, 0.25, 0.5, 0.75, 1\}$ on 100 prompts per domain for $p < 1$.

Data and single-model controls. We use 500 XSum news passages (Narayan et al., 2018), 300 SQuAD encyclopedia passages (Rajpurkar et al., 2016), and 500 WritingPrompts stories (Fan et al., 2018), following the dataset construction of Bao et al. (2024). Raw human texts average 172.26, 177.16, and 186.09 NLTK tokens, respectively. The standard single-model arm uses the Fast-DetectGPT data-builder protocol (seed 0): a 30-token prompt, a 150–200-token output constraint, and pair-trimming against the human text. The advanced single-model arm contains all four models, totaling 5,200 generated passages. It is not generated by the ToBlend loop; its pooled machine length averages 145.7 whitespace words, versus 155.4 for classic singles, 156.7 for advanced TL1, and 157.1 for classic TL1. The paired human lengths also differ (142.8, 152.6, and 154.4 words for advanced singles, classic singles, and ToBlend TL1). We report length matching separately rather than assuming these differences cancel.

Detectors and scorer identity. We evaluate likelihood, rank, log-rank, predictive entropy, Fast-DetectGPT (Bao et al., 2024), DetectGPT (Mitchell et al., 2023), DetectLLM LRR/NPR (Su et al., 2023), the GPT-2/OpenAI RoBERTa detector (Solaiman et al., 2019), RADAR (Hu et al., 2023), and

Table 1: Fast-DetectGPT with GPT-Neo-2.7B as reference and scorer throughout. Singles are four-generator means within each pool; $n = 500/300/500$ prompts per generator and ToBlend setting. DiD is (advanced single minus advanced TL1) minus (classic single minus classic TL1), with a 2,000-resample document-clustered bootstrap. These are raw-text signed AUROCs; the two generation pipelines have different length/trim rules. The reduction relative to data-builder singles persists after V2 cleaning of both arms in all six cells (Appendix A); Section 5 shows that the classic pool members run alone through the same loop are, on average, as detectable as the classic blends.

Data	Classic single	Classic TL1	Advanced single	Advanced TL1	DiD [95% CI]
XSum	0.9283	0.7088	0.7395	0.3968	0.1232 [0.0847, 0.1596]
SQuAD	0.9474	0.8047	0.8105	0.5317	0.1361 [0.0987, 0.1765]
Writing	0.9714	0.9128	0.9217	0.8036	0.0594 [0.0356, 0.0818]

the Hello-SimpleAI RoBERTa detector (Guo et al., 2023) as a formatting diagnostic. Probability-based comparisons use GPT-Neo-2.7B as the scoring model; Fast-DetectGPT also uses it as the reference model. Self-scoring values (0.9771 to 0.9990), in which each generator scores its own text, are reported in Appendix B and are not the baseline for this fixed-scorer comparison. The same-scorer classic-pool four-generator means are in Table 1; a secondary mean excludes GPT-Neo because it is also the scorer. Supervised AI-label indices are selected on calibration data and then fixed; held-out calibration AUROCs are 0.9400, 0.7777, and 0.5988 for GPT-2 detector, RADAR, and Hello-SimpleAI. Fast-DetectGPT is also computed with two scorers from outside both pools, Falcon-7B (Almazrouei et al., 2023) and OLMo-2-7B (allenai/OLMo-2-1124-7B) (Walsh et al., 2025); OLMo-2-7B replaced the planned Qwen2.5-7B base model, which was unavailable on the server. Binoculars (Hans et al., 2024) uses Falcon-7B as observer and Falcon-7B-Instruct as performer. DIPPER (lexical 60, order 60) and Qwen2.5-7B-Instruct supply paraphrase attacks.

Text views. V0 is the raw stored text. V1 reverses the generation pipeline's NLTK spacing. V1-S excludes words overlapping the canonical leaked-marker mask from the normalized per-word score sums; it does not re-run the model on the edited sequence. V1-T truncates both members of a normalized human/generated pair to the shorter word length. V2 applies one cleaner to both classes, which reverses the NLTK spacing, removes leaked end-of-text and sentence markers, and restores tokenized quotation marks, and then performs an actual detector forward pass. We treat V1-S and V2 as distinct views and keep normalized results out of raw-text tables. Supervised V2 coverage is TL1, random, and sentence across both pools and three datasets (18 cells per detector); the other four settings report point estimates.

Metrics and uncertainty. AUROC uses the declared score orientation. Values below 0.5 indicate *inverted separation*: the classes separate in reverse. We therefore also report the orientation-free detectability $D(A) = \max(A, 1 - A)$ and a held-out oriented AUROC, which fits the score direction on a fixed 20% of prompts and evaluates the other 80%; choosing direction from labeled blends is an adaptation rather than zero-shot detection. We report percentile bootstrap confidence intervals (2,000 resamples) where passage arrays exist. Cached formatting and orientation-audit intervals use prompt-paired resampling. For the supervised V0/V2 cell tables, individual AUROC intervals resample human and generated classes independently, while V2-minus-V0 differences resample paired prompts. Pool comparisons use a document-clustered bootstrap with the same resampled documents in all arms, and equal weight for the three domain means; pool-of-one contrasts and pool differences in Sections 5 and 6 use $D(A)$ with held-out orientation and 95% prompt-cluster bootstrap intervals. Grouped-probe intervals use 1,000 group-bootstrap resamples. Operating-point analyses choose a 95th-percentile human threshold on a fixed 30% split and report TPR and realized FPR on the remaining 70%, unless marked as an in-sample diagnostic. Intervals quantify prompt sampling within the three domains; three native-loop seeds measure seed variation.

4 RQ1: Blended Versus Standard Single-Model Text

Classic-pool TL1 lowers signed Fast-DetectGPT AUROC relative to the same-scorer data-builder baseline in all three datasets (Table 1). All three classic blend AUROCs are above 0.5, so these

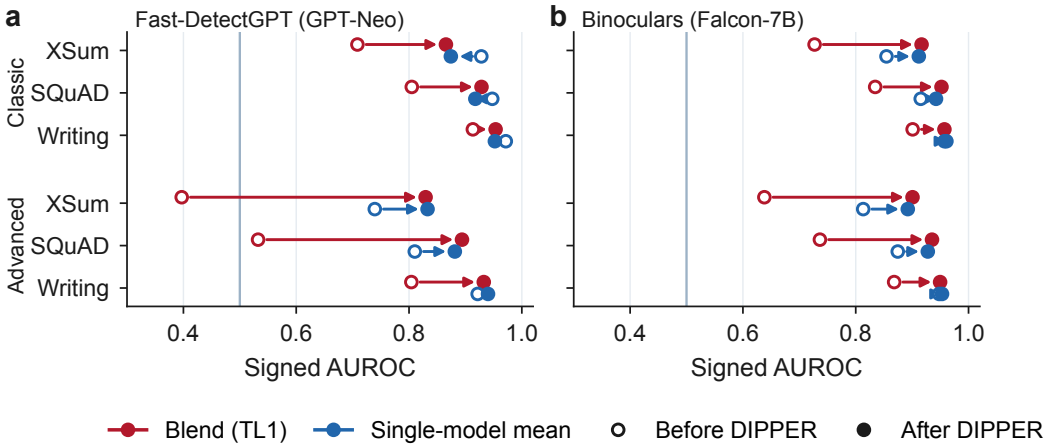


Figure 2: Paraphrasing removes the blend's detection gap. Signed AUROC before (open) and after (filled) DIPPER paraphrasing (lexical 60, order 60) for the 2024 TL1 blends (red) and the mean over each pool's four data-builder single-model sets (blue), under (a) Fast-DetectGPT with the GPT-Neo scorer and (b) Binoculars with the Falcon-7B pair. The gray line marks chance. Per-set values with 95% intervals are in Appendix A.

Table 2: Analytic Fast-DetectGPT recomputed from aligned per-word caches, TL1; 2,000 prompt-paired bootstrap resamples. V1-S applies the marker mask to spacing-normalized cached terms, and V1-T trims each normalized human/blended pair to its shorter word length. Neither is a new V2 detector forward pass. Small differences from Table 1 reflect analytic/cache versus sampled harness computation. A new Fast-DetectGPT pass on V2-cleaned blends, data-builder singles, and their human partners also keeps the reduction above zero in every cell (Appendix A).

Pool	Data	Raw V0 [95% CI]	Marker-masked V1-S [95% CI]	Length-matched V1-T [95% CI]
Classic	XSum	0.7077 [0.6766, 0.7372]	0.6911 [0.6596, 0.7211]	0.6777 [0.6463, 0.7073]
Classic	SQuAD	0.8093 [0.7761, 0.8404]	0.8008 [0.7672, 0.8335]	0.7962 [0.7632, 0.8282]
Classic	Writing	0.9139 [0.8954, 0.9310]	0.9006 [0.8812, 0.9197]	0.8990 [0.8799, 0.9177]
Advanced	XSum	0.3959 [0.3643, 0.4276]	0.3703 [0.3391, 0.4010]	0.3718 [0.3404, 0.4021]
Advanced	SQuAD	0.5397 [0.4954, 0.5816]	0.5248 [0.4798, 0.5668]	0.5213 [0.4775, 0.5625]
Advanced	Writing	0.8049 [0.7806, 0.8285]	0.7834 [0.7578, 0.8094]	0.7784 [0.7528, 0.8043]

are reductions in detectability under either orientation convention. The reduction persists when the V2 cleaner is applied to blends, single-model text, and their human partners: its interval stays above zero in all six pool/domain cells, for example 0.1396 [0.1119, 0.1687] on classic XSum in $D(A)$ (Appendix A). For XSum, the lowest of the seven observed settings is TL2: 0.7004 versus 0.9283, a reduction of 0.2279 [0.2044, 0.2534]. TL1 is the primary setting; minima over settings are descriptive. Full raw battery tables are in Appendix B.

Paraphrase attacks. DIPPER (Krishna et al., 2023) raises the Fast-DetectGPT AUROC of the blends, for example from 0.3970 to 0.8294 on advanced XSum (Figure 2). After DIPPER is applied to both arms, blends and the same pool's data-builder texts differ by at most 0.0127 AUROC, and every interval includes zero. An instruction-tuned paraphrase with Qwen2.5-7B-Instruct (Qwen Team et al., 2024), applied to the blends only, lowers classic-blend AUROC by 0.1445 to 0.2318 and has mixed effects on advanced blends. Against the saved T5 paraphrase (Raffel et al., 2020) and adjacent-word swaps of classic single-model text, the advanced blends have lower signed AUROC in all three domains, but not in absolute separation from chance against word swaps on XSum (-0.0261 [-0.0784, 0.0233]); these attacks are not generator-matched (Appendix E).

Formatting controls and supervised detectors. Fast-DetectGPT reductions also persist after spacing normalization and marker masking of the cached scores (Table 2): spacing normalization

Table 3: Supervised detectors against each pool's own single-model text, TL1, calibrated AI orientation. Single is the four-generator data-builder mean of the same pool; Δ is single minus blend AUROC, so a negative Δ means that blended text is detected more readily. Raw uses stored text. Cleaned applies the V2 cleaner to the blend and its human partners; the classic single-model text is also cleaned, and the advanced single-model text has no NLTK spacing and is used as stored. 95% CIs: 2,000 document-clustered bootstrap replicates shared across arms.

Detector	Pool	Data	Single (raw)	Blend (raw)	Δ raw [95% CI]	Δ cleaned [95% CI]
GPT-2 detector	Classic	XSum	0.9416	0.7211	0.2205 [0.1938, 0.2471]	0.0668 [0.0488, 0.0851]
GPT-2 detector	Classic	SQuAD	0.9335	0.7208	0.2127 [0.1785, 0.2482]	0.1081 [0.0783, 0.1383]
GPT-2 detector	Classic	WritingPrompts	0.9365	0.7434	0.1932 [0.1677, 0.2209]	0.0505 [0.0342, 0.0678]
GPT-2 detector	Advanced	XSum	0.8287	0.5622	0.2665 [0.2346, 0.2957]	0.1600 [0.1318, 0.1884]
GPT-2 detector	Advanced	SQuAD	0.8120	0.5221	0.2899 [0.2532, 0.3283]	0.1813 [0.1449, 0.2171]
GPT-2 detector	Advanced	WritingPrompts	0.8665	0.6484	0.2181 [0.1910, 0.2454]	0.0960 [0.0734, 0.1183]
RADAR	Classic	XSum	0.8904	0.8958	-0.0054 [-0.0205, 0.0093]	-0.0798 [-0.0983, -0.0612]
RADAR	Classic	SQuAD	0.8263	0.8395	-0.0132 [-0.0360, 0.0098]	-0.0469 [-0.0727, -0.0210]
RADAR	Classic	WritingPrompts	0.7855	0.8846	-0.0991 [-0.1159, -0.0818]	-0.1173 [-0.1382, -0.0966]
RADAR	Advanced	XSum	0.7994	0.9027	-0.1033 [-0.1202, -0.0871]	-0.0817 [-0.1005, -0.0637]
RADAR	Advanced	SQuAD	0.7675	0.8474	-0.0799 [-0.1042, -0.0562]	-0.0314 [-0.0576, -0.0057]
RADAR	Advanced	WritingPrompts	0.7179	0.8844	-0.1665 [-0.1872, -0.1455]	-0.1093 [-0.1317, -0.0867]

Table 4: Model switching with the generation loop held fixed: detectability of TL1 blends minus the mean detectability of the same pool's four members, each run alone through the identical loop (pool-of-one), with 95% prompt-cluster bootstrap CIs. Detectability is $D(A) = \max(A, 1-A)$ with the orientation fitted on 20% of prompts and evaluated on the other 80%. Negative values mean that switching lowers detectability. Original loop: the 2024 blends against pool-of-one text regenerated with the same loop; native loop: blends and pool-of-one text generated together by the loop that keeps model-native strings. Binoculars uses the Falcon-7B pair. Per-member values and the other detectors are in Appendix A.

Pool	Data	FDG, original	FDG, native	Binoculars, original	Binoculars, native
Classic	XSum	0.0071 [−0.0220, 0.0355]	0.0034 [−0.0111, 0.0177]	−0.0044 [−0.0296, 0.0192]	−0.0079 [−0.0265, 0.0112]
	SQuAD	0.0008 [−0.0276, 0.0295]	0.0016 [−0.0168, 0.0180]	−0.0099 [−0.0359, 0.0160]	−0.0119 [−0.0319, 0.0069]
	Writing	−0.0046 [−0.0204, 0.0090]	0.0050 [−0.0036, 0.0135]	−0.0144 [−0.0300, −0.0003]	−0.0051 [−0.0159, 0.0051]
Advanced	XSum	−0.0724 [−0.1078, −0.0406]	0.1049 [0.0852, 0.1228]	−0.1160 [−0.1438, −0.0898]	0.0873 [0.0731, 0.1017]
	SQuAD	−0.1721 [−0.2162, −0.1302]	0.0499 [0.0294, 0.0697]	−0.0819 [−0.1133, −0.0534]	0.0434 [0.0283, 0.0579]
	Writing	−0.0418 [−0.0642, −0.0209]	0.0411 [0.0302, 0.0516]	−0.0331 [−0.0501, −0.0176]	0.0326 [0.0254, 0.0396]

changes AUROC by -0.0334 to -0.0041 across all 42 cells, and marker masking by -0.0306 to $+0.0050$. For the GPT-2 detector, however, much of the raw reduction is formatting: cleaning raises its AUROC on blended text by 0.0918 to 0.2346 across the 18 re-scored cells. Against data-builder advanced text, RADAR detects native-loop advanced blends more readily (single minus blend -0.0624 to -0.1099); Section 5 shows that this difference comes from the generation pipeline. Table 3 gives the own-pool comparison for the 2024 blends: the GPT-2 detector's raw reductions (0.1932 to 0.2899) shrink after cleaning (0.0505 to 0.1813), and after cleaning RADAR scores the blends as more machine-like than the pool's data-builder single-model text in all six cells. Further supervised and entropy results are in Appendices C and D.

5 RQ2: Does Model Switching Lower Detectability?

With the loop held fixed (Table 4, Figure 3a), switching has no measurable Fast-DetectGPT effect in the classic pool: blend minus pool-of-one detectability is 0.0071, 0.0008, and -0.0046 in the original loop and 0.0016 to 0.0050 in the native loop, with every interval including zero. The classic-pool reduction in Table 1 therefore comes from the chunked, re-tokenized loop, not from switching between models. Individual classic pool-of-one arms in the original loop range from AUROC 0.6085 (OPT-2.7B, XSum) to 0.9808 (GPT-Neo, WritingPrompts).

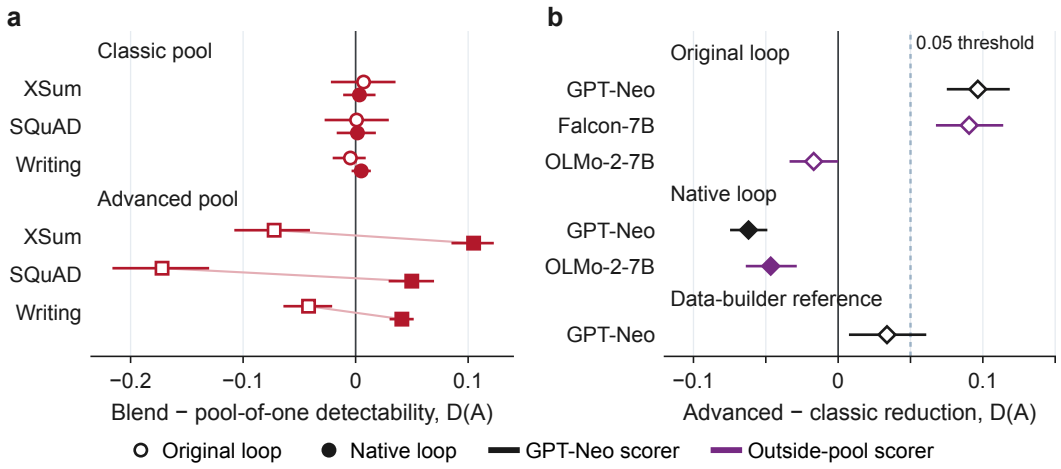


Figure 3: Model switching and the generation loop affect detectability differently (Fast-DetectGPT at TL1; 95% prompt-cluster bootstrap CIs). (a) Detectability $D(A)$ of blends minus the mean of the same pool's four members, each run alone through the identical loop (pool-of-one), with the score direction fitted on 20% of prompts; negative values mean that switching lowers detectability. Open markers: original loop; filled markers: native loop. Circles: classic pool; squares: advanced pool; GPT-Neo scorer throughout. (b) Advanced-pool minus classic-pool reduction, with reduction $= D(\text{reference}) - D(\text{blend})$ averaged over domains, for each scorer, loop and reference; purple marks scorers from outside both pools, and the dashed line is the pre-specified 0.05 threshold.

In the advanced pool, by contrast, the sign of the switching effect depends on the loop. Against original-loop pool-of-one text, regenerated separately with the same code, the 2024 blends are less detectable (-0.0724, -0.1721, -0.0418). When blends and pool-of-one arms are generated together by the native loop, the blends are more detectable (0.1049, 0.0499, 0.0411), with every interval excluding zero. Binoculars shows the same reversal: reductions in four of six original-loop cells (classic WritingPrompts and all three advanced domains) and increases in all three advanced native-loop cells. Across redraw probabilities p in the native loop, any switching raises advanced-pool AUROC above a single model kept for the whole passage ($p = 0$), without a monotone trend: on XSum, 0.7649 [0.7038, 0.8257] at $p = 0$ against 0.8907 [0.8729, 0.9083] at $p = 1$, with separated intervals also on WritingPrompts (Figure 4). In the original loop, advanced blends have lower per-token log-likelihood under GPT-Neo than their pool-of-one arms (by 0.3089 to 0.3981 nats) and a smaller drop in the scorer's expected log-likelihood (0.1456 to 0.2806), which lowers the analytic curvature by 0.7455 to 1.0254; classic differences are at most 0.0288 nats and 0.1017 in curvature.

In both pools, RADAR scores native-loop blends and single-model text from the same loop alike. Against native pool-of-one text, its AUROC on native blends differs by at most 0.0098; the blend is slightly higher in five of six cells, and no interval excludes zero. The larger RADAR difference against data-builder text therefore reflects the generation pipeline, not model switching.

6 RQ3: Is the Reduction Larger for the Advanced Pool?

Table 5 and Figure 3b compare the two pools' reductions. Under the primary specification, with GPT-Neo as scorer and original-loop pool-of-one text as each pool's reference, the advanced reduction exceeds the classic one by 0.0965 [0.0751, 0.1186], above the 0.05 threshold fixed before these runs, and the outside-pool scorer Falcon-7B gives 0.0905 [0.0676, 0.1141]. Yet the difference depends on the scorer and the loop. With OLMo-2-7B as scorer it is -0.0169 [-0.0336, -0.0001], and in the matched native loop it is -0.0619 [-0.0747, -0.0489] with GPT-Neo and -0.0466 [-0.0638, -0.0286] with OLMo-2-7B. Against data-builder single-model text it is 0.0337 [0.0075, 0.0609], consistent with the orientation-free estimates from the saved scores, 0.0374 [0.0105, 0.0622] for $D(A)$ and 0.0321 [0.0047, 0.0604] for held-out orientation (Appendix F); the signed-AUROC difference of 0.1062 [0.0869, 0.1266] reflects inverted scores on advanced XSum

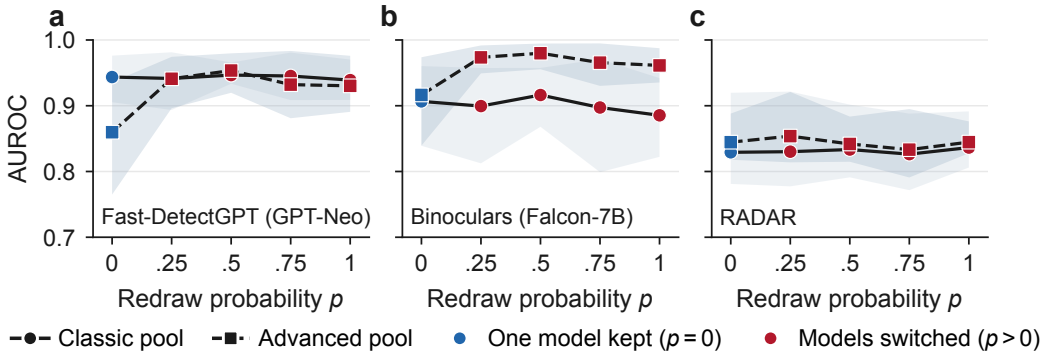


Figure 4: Switching models raises zero-shot detectability in the advanced pool but not in the classic pool (native loop, TL1). AUROC against the per-step redraw probability p for (a) Fast-DetectGPT (GPT-Neo scorer), (b) Binoculars (Falcon-7B pair) and (c) RADAR; $p = 0$ keeps one model for the whole passage (blue), $p > 0$ switches models (red). Lines: mean over domains per pool (classic, solid circles; advanced, dashed squares); bands: range across domains; per-domain 95% intervals in Table 22.

Table 5: Pool comparison (RQ3): advanced-pool reduction minus classic-pool reduction, each relative to the stated single-model reference, Fast-DetectGPT at TL1. Reduction $= D(\text{reference}) - D(\text{blend})$; a positive value means a larger advanced-pool reduction. The first row is the primary specification; the threshold for a supported pool difference is 0.05. OLMo-2-7B (allenai/OLMo-2-1124-7B) and Falcon-7B are scorers from outside both pools. Mean over domains with a 95% prompt-cluster bootstrap CI.

Scorer	Reference	Blends	Scale	XSum	SQuAD	Writing	Mean [95% CI]
GPT-Neo	Pool-of-one, original	2024	D	0.0795	0.1729	0.0372	0.0965 [0.0751, 0.1186]
GPT-Neo	Pool-of-one, original	2024	logit	0.6053	1.0167	0.4945	0.7055 [0.5731, 0.8565]
Falcon-7B	Pool-of-one, original	2024	D	0.1426	0.1017	0.0272	0.0905 [0.0676, 0.1141]
OLMo-2-7B	Pool-of-one, original	2024	D	0.0435	−0.1224	0.0282	−0.0169 [−0.0336, −0.0001]
GPT-Neo	Data-builder singles	2024	D	−0.0897	0.1346	0.0564	0.0337 [0.0075, 0.0609]
GPT-Neo	Pool-of-one, native	Native	D	−0.1015	−0.0483	−0.0361	−0.0619 [−0.0747, −0.0489]
OLMo-2-7B	Pool-of-one, native	Native	D	−0.1249	0.0428	−0.0578	−0.0466 [−0.0638, −0.0286]

Table 6: Instruction-tuned pool, native loop (TL1, seed 0; Fast-DetectGPT, GPT-Neo scorer): blend D and blend minus the mean D of the pool's four data-builder or pool-of-one arms [95% CI]; positive values mean a more detectable blend. Advanced pool: its native blends against its data-builder text, the pre-specified comparison. All detectors: Table 11.

	Instruction-tuned pool			Advanced pool	
Data	Blend D	vs. data-builder	vs. pool-of-one	Blend D	vs. data-builder
XSum	0.9706	0.1090 [0.0943, 0.1238]	0.0439 [0.0331, 0.0551]	0.8977	0.1516 [0.1296, 0.1725]
SQuAD	0.9844	0.1320 [0.1124, 0.1522]	0.0655 [0.0529, 0.0796]	0.9266	0.1120 [0.0897, 0.1329]
Writing	0.9864	0.0407 [0.0316, 0.0492]	0.0271 [0.0201, 0.0339]	0.9760	0.0539 [0.0421, 0.0658]

(AUROC 0.3968). The difference is stable across generation seeds: across three native-loop seeds it has a between-seed SD of 0.00097, against a mean CI half-width of 0.0132 (Table 10). Removing GPT-Neo, the scorer, from the classic pool leaves a native-loop reduction relative to its three pool-of-one arms on XSum only (−0.0310 [−0.0518, −0.0121]). The advanced pool's larger reduction is therefore specific to the GPT-Neo and Falcon-7B scorers and to the original loop, so the evidence ties its size to these choices rather than to generator capability.

Instruction-tuned pool. As for the advanced pool in the native loop, switching raises the detectability of the instruction-tuned pool (Table 6): its blends exceed its pool-of-one arms by 0.0439, 0.0655, and 0.0271 on XSum, SQuAD, and WritingPrompts and its data-builder text by 0.1090, 0.1320, and 0.0407, with every interval excluding zero; the blends themselves are highly detectable (D of 0.9706 to 0.9864), and pool-of-one arms range from AUROC 0.8245 (Mistral, XSum) to 0.9971 (Llama-3.1, WritingPrompts). RADAR and the GPT-2 detector agree in sign against data-builder text, and, unlike in the other pools, RADAR also rates the blends above pool-of-one text on XSum and SQuAD (Table 11). The increase over data-builder text, a contrast that also differs in sampling settings (Table 12), has the advanced pool's sign in every domain but is smaller on XSum, with non-overlapping intervals (0.1090 [0.0943, 0.1238] against 0.1516 [0.1296, 0.1725]). Under the rule fixed before the runs, the effect is thus pool-dependent; we claim no trend across pools.

7 RQ4: Quality and Cost

Three hired, experienced human raters, blind to text source and generation settings, scored coherence and fluency from 1 to 7 on five instances per dataset (Table 47, Appendix H). Of 42 pool/domain/setting cells, 16 meet or exceed GPT-2 on both rating metrics, 23 fall below it on both, and three have mixed directions; none reaches the human reference. Round-2 ordinal Krippendorff α (Krippendorff, 2019) is 0.502/0.508 for coherence/fluency, and ICC(2,3) (Shrout & Fleiss, 1979) is 0.790/0.794. Automatic quality proxies agree with the ratings in some comparisons and diverge in others (Appendix H.2). Generation is training-free, and its time cost varies: logged fp32 runs of the original loop have median per-passage times of 20.30 to 126.14 seconds for classic TL1 and 28.61 to 38.24 seconds for advanced TL1 (Appendix I).

8 Discussion and Conclusion

Holding the generation loop fixed changes how token-level blending should be read. For the classic pool, the Fast-DetectGPT gap relative to standard single-model text comes from the original loop, which re-tokenizes after every chunk and feeds each model NLTK-re-spaced text: the pool members run alone through that loop are, on average, as hard to detect as the blend. For the advanced pool, switching raises detectability when blends and pool-of-one arms come from the same native loop and lowers it only against original-loop pool-of-one text; Binoculars follows the same pattern, and RADAR scores native-loop blends and same-loop single-model text alike. The advanced pool's larger reduction depends on the scorer, the reference, and the loop. An instruction-tuned pool run in the native loop also becomes more detectable with switching, by an amount that differs from the advanced pool's on XSum. Exploratory probe and ensemble-reference curvature results leave open whether a signal is blend-specific (Appendix K).

Measurements of multi-model generation should therefore compare blends with single models run through the identical loop, use orientation-free measures, and report more than one scorer and detector family. With these controls, the classic pool's evasion traces to the generation loop, and matched generation shows switching raising detectability in the other two pools, which suggests that model switching alone is not a route around current detectors.

Limitations

Our conclusions cover three fixed pools of open models up to 9B parameters and three English domains. The instruction-tuned pool was run in the matched native loop with one seed and default sampling settings, and outside scorers covered part of its arms. Generation settings differ between the blending and data-builder controls, which qualifies comparisons with data-builder text, and the 2024 sampling configurations are partly recorded; the switching conclusions therefore rest on same-loop pool-of-one contrasts: the primary original loop is matched in code only, and the native loop in code and time, with blends over three seeds. Score directions and false-positive thresholds are fitted within each corpus and may not transfer to others. Supervised TL2–TL5 rows report point estimates; the human ratings, five items per cell over the 2024 blends, describe quality differences rather than equivalence, and automatic metrics and ratings diverge in several comparisons. The appendix mechanism and defense analyses are exploratory.

ETHICS STATEMENT

Token-level blending can be used to evade automated content checks and can retain model biases or generate misleading text. We study these behaviors to evaluate detection reliability. Detector outputs should not be treated as definitive evidence of authorship, particularly where a false accusation would have consequences. The authors' Institutional Review Board approved the design for collecting the human ratings; the approval number will be disclosed in the camera-ready version. Raters received information on mental-health hotlines and clinics, because generated text may contain uncomfortable content such as social bias, and they could stop or quit at any time. Raters were paid for their recorded working hours. The numerical reanalysis uses these existing text and human-rating records and adds no new annotation collection.

REPRODUCIBILITY STATEMENT

Saved raw corpora, cleaning definitions, per-passage scores, bootstrap settings, and dependency coverage are retained in the experiment package, and every table is generated from saved result files. CPU analyses and GPU scoring ran on four A100 80GB cards with Python 3.10.12, PyTorch 2.6.0, and Transformers 4.57.1; GPU scoring took 77.35 minutes, distinct from generation throughput. The controlled-loop experiments (57 jobs, same cards and software) passed a scorer anchor check against saved scores before running; nine further jobs generated and scored the instruction-tuned pool with the same software. Every table in Appendix A is generated from their saved analysis. Score-reproduction checks and their tolerances are described in Appendix L.

AI USE STATEMENT

We used large language model assistants for the grammar checks and code debugs. The blended and single-model corpora are generated by the language models listed in Section 3, plus GPT-3.5-turbo and GPT-4 for the appendix comparison; these are the object of study, not writing aids. 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. Every reported number is traceable to saved experiment outputs.

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.

Anonymous. Reverse ToBlend: Analyzing detection signals under token-level multi-model blending, 2027. Concurrent submission to ICLR 2027.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 675–718, Nusa Dua, Bali, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.ijcnlp-main.45. URL https://aclanthology.org/2023.ijcnlp-main.45/.

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.

Regina Barzilay and Mirella Lapata. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34, 2008. doi: 10.1162/coli.2008.34.1.1. URL https://aclanthology.org/J08-1001/.

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. URL https://www.nltk.org/book_1ed/.

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.

Xin Chen, Junchao Wu, Shu Yang, Runzhe Zhan, Zeyu Wu, Ziyang Luo, Di Wang, Min Yang, Lidia S. Chao, and Derek F. Wong. RepreGuard: Detecting LLM-generated text by revealing hidden representation patterns. *Transactions of the Association for Computational Linguistics*, 13:1812–1831, 2025. doi: 10.1162/tacl.a.61. URL https://aclanthology.org/2025.tacl-1.81/.

Avishek Choudhury and Hamid Shamszare. Investigating the impact of user trust on the adoption and use of ChatGPT: Survey analysis. *Journal of Medical Internet Research*, 25:e47184, 2023. doi: 10.2196/47184. URL https://www.jmir.org/2023/1/e47184.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 7282–7296, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.565. URL https://aclanthology.org/2021.acl-long.565/.

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, 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 *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, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-3019. URL https://aclanthology.org/P19-3019/.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund,

Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu-hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. Gemma: Open models based on Gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024a. URL https://arxiv.org/abs/2403.08295.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024b. URL https://arxiv.org/abs/2408.00118.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120, 2023. doi: 10.1073/pnas.2305016120. URL https://www.pnas.org/doi/10.1073/pnas.2305016120.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*, 2023. URL https://arxiv.org/abs/2301.07597.

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, 21–27 Jul 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, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.322. URL https://aclanthology.org/2025.acl-long.322/.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

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, Kunwoo Park, and Jisun An. ChatGPT rates natural language explanation quality like humans: But on which scales? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*

(LREC-COLING 2024), pp. 3111–3132, Torino, Italia, 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.277/.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. Automatic detection of generated text is easiest when humans are fooled. In *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, December 2023. URL https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/. Microsoft Research Blog.

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.

Katharina Kann, Sascha Rothe, and Katja Filippova. Sentence-level fluency evaluation: References help, but can be spared! In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pp. 313–323, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/K18-1031. URL https://aclanthology.org/K18-1031.

Antonia Karamolegkou, Jiaang Li, Li Zhou, and Anders Søgaard. Copyright violations and large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7403–7412, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.458. URL https://aclanthology.org/2023.emnlp-main.458.

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, 23–29 Jul 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.

Klaus Krippendorff. *Content Analysis: An Introduction to Its Methodology*. SAGE Publications, Thousand Oaks, CA, 4th edition, 2019. ISBN 978-1-5063-9566-1. doi: 10.4135/9781071878781.

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.

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, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.1005. URL https://aclanthology.org/2023.emnlp-main.1005/.

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.

Marianna J. Martindale and Marine Carpuat. Fluency over adequacy: A pilot study in measuring user trust in imperfect MT. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pp. 13–25, Boston, MA, 2018. Association for Machine Translation in the Americas. URL https://aclanthology.org/W18-1803/.

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, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/mitchell23a.html.

Sandra Mitrović, Davide Andreoletti, and Omran Ayoub. ChatGPT or human? Detect and explain. Explaining decisions of machine learning model for detecting short ChatGPT-generated text. *arXiv preprint arXiv:2301.13852*, 2023. URL https://arxiv.org/abs/2301.13852.

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/.

OpenAI. Introducing ChatGPT, November 2022. URL https://openai.com/index/chatgpt/. OpenAI Blog.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. Is ChatGPT a general-purpose natural language processing task solver? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 1339–1384, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.85. URL https://aclanthology.org/2023.emnlp-main.85/.

Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL https://jmlr.org/papers/v21/20-074.html.

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, 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=OOgsAZdFOt.

Shannon Zejiang Shen, Hunter Lang, Bailin Wang, Yoon Kim, and David Sontag. Learning to decode collaboratively with multiple language models. In *Proceedings of the 62nd Annual Meeting*

of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 12974–12990, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.701. URL https://aclanthology.org/2024.acl-long.701/.

Patrick E. Shrout and Joseph L. Fleiss. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428, 1979. doi: 10.1037/0033-2909.86.2.420.

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, 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://dl.acm.org/doi/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, 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.

Max Wolff and Stuart Wolff. Attacking neural text detectors. *arXiv preprint arXiv:2002.11768*, 2020. URL https://arxiv.org/abs/2002.11768. Accepted at the ICLR 2020 Workshop Towards Trustworthy ML: Rethinking Security and Privacy for ML.

Junchao Wu, Runzhe Zhan, Derek F. Wong, Shu Yang, Xinyi Yang, Yulin Yuan, and Lidia S. Chao. DetectRL: Benchmarking LLM-generated text detection in real-world scenarios. In *Advances in Neural Information Processing Systems*, volume 37, pp. 100369–100401. Curran Associates, Inc., 2024. doi: 10.52202/079017-3186. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/b61bdf7e9f64c04ec75a26e781e2ad51-Abstract-Datasets_and_Benchmarks_Track.html. Datasets and Benchmarks Track.

Xianjun Yang, Wei Cheng, Yue Wu, Linda Ruth Petzold, William Yang Wang, and Haifeng Chen. DNA-GPT: Divergent n-gram analysis for training-free detection of GPT-generated text. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Xlayxj2fWp.

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, 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://dl.acm.org/doi/10.1145/3544548.3581318.

A CONTROLLED GENERATION-LOOP EXPERIMENTS

This appendix reports the complete results of the controlled experiments summarized in Sections 4–6. Every value is read from the saved analysis of that run; intervals are 95% prompt-cluster bootstrap intervals. Detectability $D(A) = \max(A, 1 - A)$ fixes each arm's orientation on 20% of prompts and evaluates the other 80%. The pool-of-one arms run one fixed model through the identical loop; the original-loop arms were regenerated in fp16 with batched sampling, whereas the 2024 blends were generated in fp32 by the original notebooks, so the original-loop comparison is matched in code but not in time. The native-loop blends and pool-of-one arms were generated together by one code path. The scorer labeled OLMo-2-7B is allenai/OLMo-2-1124-7B, used because the planned Qwen2.5-7B base model was not available on the server.

Table 7: Blend minus pool-of-one detectability for every detector. Blend and P1 are D of the TL1 blend and the mean D of the four pool-of-one arms. Negative differences mean that switching lowers detectability.

Detector	Pool	Data	Blend (orig.)	P1 (orig.)	Orig. diff. [CI]	Blend (native)	P1 (native)	Native diff. [CI]
Fast-DetectGPT (GPT-Neo, sampled)	Classic	XSum	0.7114	0.7043	0.0071 [−0.0220, 0.0355]	0.9145	0.9111	0.0034 [−0.0111, 0.0177]
		SQuAD	0.8034	0.8025	0.0008 [−0.0276, 0.0295]	0.9361	0.9345	0.0016 [−0.0168, 0.0180]
		Writing	0.9096	0.9142	−0.0046 [−0.0204, 0.0090]	0.9679	0.9629	0.0050 [−0.0036, 0.0135]
	Advanced	XSum	0.6169	0.6893	−0.0724 [−0.1078, −0.0406]	0.8977	0.7928	0.1049 [0.0852, 0.1228]
		SQuAD	0.5350	0.7070	−0.1721 [−0.2162, −0.1302]	0.9266	0.8767	0.0499 [0.0294, 0.0697]
		Writing	0.8033	0.8451	−0.0418 [−0.0642, −0.0209]	0.9760	0.9349	0.0411 [0.0302, 0.0516]
Fast-DetectGPT (GPT-Neo, analytic)	Classic	XSum	0.7114	0.7044	0.0071 [−0.0219, 0.0354]	0.9150	0.9114	0.0036 [−0.0108, 0.0178]
		SQuAD	0.8034	0.8026	0.0008 [−0.0275, 0.0296]	0.9362	0.9344	0.0017 [−0.0168, 0.0180]
		Writing	0.9093	0.9141	−0.0047 [−0.0205, 0.0089]	0.9677	0.9628	0.0049 [−0.0038, 0.0134]
	Advanced	XSum	0.6169	0.6894	−0.0725 [−0.1080, −0.0407]	0.8977	0.7929	0.1049 [0.0849, 0.1229]
		SQuAD	0.5346	0.7069	−0.1724 [−0.2165, −0.1303]	0.9265	0.8769	0.0496 [0.0291, 0.0692]
		Writing	0.8030	0.8450	−0.0420 [−0.0645, −0.0210]	0.9758	0.9347	0.0411 [0.0301, 0.0516]
Fast-DetectGPT (OLMo-2-7B)	Classic	XSum	0.7212	0.7110	0.0103 [−0.0147, 0.0360]	0.4975	0.5325	−0.0350 [−0.0604, −0.0082]
		SQuAD	0.9812	0.9785	0.0026 [−0.0041, 0.0085]	0.9709	0.9593	0.0116 [0.0021, 0.0208]
		Writing	0.7922	0.8160	−0.0238 [−0.0470, −0.0034]	0.8916	0.9163	−0.0247 [−0.0400, −0.0102]
	Advanced	XSum	0.7433	0.7765	−0.0333 [−0.0600, −0.0073]	0.7787	0.6888	0.0899 [0.0650, 0.1131]
		SQuAD	0.9799	0.8549	0.1251 [0.1110, 0.1393]	0.8594	0.8906	−0.0312 [−0.0615, −0.0014]
		Writing	0.7967	0.8487	−0.0520 [−0.0744, −0.0301]	0.9800	0.9470	0.0331 [0.0239, 0.0422]
Fast-DetectGPT (Falcon-7B)	Classic	XSum	0.5179	0.5129	0.0050 [−0.0436, 0.0541]	0.6079	0.6335	−0.0255 [−0.0510, 0.0002]
		SQuAD	0.6348	0.6595	−0.0247 [−0.0598, 0.0111]	0.7139	0.7536	−0.0397 [−0.0742, −0.0076]
		Writing	0.8286	0.8474	−0.0188 [−0.0383, −0.0006]	0.8818	0.9018	−0.0200 [−0.0365, −0.0044]
	Advanced	XSum	0.5193	0.6569	−0.1376 [−0.1812, −0.0948]	0.8760	0.7785	0.0974 [0.0771, 0.1162]
		SQuAD	0.6297	0.7561	−0.1265 [−0.1618, −0.0931]	0.9357	0.8764	0.0593 [0.0391, 0.0789]
		Writing	0.8350	0.8810	−0.0460 [−0.0659, −0.0264]	0.9766	0.9419	0.0347 [0.0247, 0.0438]
Binoculars (Falcon-7B pair)	Classic	XSum	0.7316	0.7360	−0.0044 [−0.0296, 0.0192]	0.8235	0.8314	−0.0079 [−0.0265, 0.0112]
		SQuAD	0.8367	0.8466	−0.0099 [−0.0359, 0.0160]	0.8932	0.9051	−0.0119 [−0.0319, 0.0069]
		Writing	0.8988	0.9132	−0.0144 [−0.0300, −0.0003]	0.9410	0.9461	−0.0051 [−0.0159, 0.0051]
	Advanced	XSum	0.6378	0.7539	−0.1160 [−0.1438, −0.0898]	0.9384	0.8510	0.0873 [0.0731, 0.1017]
		SQuAD	0.7453	0.8271	−0.0819 [−0.1133, −0.0534]	0.9635	0.9201	0.0434 [0.0283, 0.0579]
		Writing	0.8713	0.9045	−0.0331 [−0.0501, −0.0176]	0.9881	0.9555	0.0326 [0.0254, 0.0396]
RADAR	Classic	XSum	0.9055	0.8983	0.0072 [−0.0072, 0.0212]	0.8968	0.8939	0.0029 [−0.0121, 0.0164]
		SQuAD	0.8432	0.8404	0.0028 [−0.0215, 0.0265]	0.8060	0.8120	−0.0060 [−0.0318, 0.0178]
		Writing	0.8777	0.8738	0.0039 [−0.0093, 0.0182]	0.8046	0.7996	0.0050 [−0.0129, 0.0227]
	Advanced	XSum	0.9059	0.9031	0.0028 [−0.0123, 0.0186]	0.8839	0.8727	0.0112 [−0.0052, 0.0267]
		SQuAD	0.8465	0.8669	−0.0204 [−0.0426, 0.0000]	0.8242	0.8255	−0.0013 [−0.0270, 0.0227]
		Writing	0.8866	0.8958	−0.0092 [−0.0244, 0.0057]	0.8230	0.8162	0.0068 [−0.0112, 0.0245]
GPT-2 detector	Classic	XSum	0.7268	0.6986	0.0282 [0.0015, 0.0548]	0.9206	0.9221	−0.0015 [−0.0158, 0.0114]
		SQuAD	0.7200	0.7298	−0.0098 [−0.0477, 0.0252]	0.9102	0.9273	−0.0171 [−0.0403, 0.0028]
		Writing	0.7399	0.7294	0.0105 [−0.0166, 0.0360]	0.9239	0.9182	0.0057 [−0.0077, 0.0188]
	Advanced	XSum	0.5602	0.6166	−0.0564 [−0.0916, −0.0204]	0.8786	0.8572	0.0214 [0.0018, 0.0405]
		SQuAD	0.4591	0.6333	−0.1742 [−0.2291, −0.1129]	0.8810	0.8599	0.0211 [−0.0007, 0.0420]
		Writing	0.6388	0.7069	−0.0681 [−0.0982, −0.0415]	0.9250	0.8757	0.0493 [0.0333, 0.0645]
Hello-SimpleAI	Classic	XSum	0.9997	0.9980	0.0016 [0.0001, 0.0043]	0.5291	0.5486	−0.0195 [−0.0498, 0.0102]
		SQuAD	0.9912	0.9962	−0.0050 [−0.0135, 0.0013]	0.5371	0.5588	−0.0217 [−0.0605, 0.0164]
		Writing	0.9995	0.9981	0.0014 [0.0000, 0.0036]	0.5924	0.5965	−0.0041 [−0.0282, 0.0187]
	Advanced	XSum	0.9988	0.9967	0.0021 [0.0002, 0.0045]	0.5810	0.6201	−0.0392 [−0.0768, −0.0016]
		SQuAD	0.9931	0.9950	−0.0019 [−0.0108, 0.0048]	0.6015	0.6040	−0.0026 [−0.0437, 0.0355]
		Writing	0.9973	0.9982	−0.0009 [−0.0034, 0.0005]	0.6677	0.6492	0.0185 [−0.0036, 0.0401]

Table 8: Signed Fast-DetectGPT AUROC (GPT-Neo scorer) of each pool-of-one arm and of the TL1 blend, original and native loop.

Pool	Data	Arm	Original loop	Native loop
Classic	XSum	gpt2-xl	0.6787	0.8944
Classic	XSum	opt-2.7b	0.6085	0.8857
Classic	XSum	gpt-neo-2.7B	0.9060	0.9816
Classic	XSum	gpt-j-6B	0.6224	0.8645
Classic	XSum	TL1 blend	0.7085	0.9089
Classic	SQuAD	gpt2-xl	0.7743	0.9311
Classic	SQuAD	opt-2.7b	0.7455	0.9168
Classic	SQuAD	gpt-neo-2.7B	0.9454	0.9901
Classic	SQuAD	gpt-j-6B	0.7259	0.9102
Classic	SQuAD	TL1 blend	0.8049	0.9379
Classic	Writing	gpt2-xl	0.9084	0.9644
Classic	Writing	opt-2.7b	0.8750	0.9299
Classic	Writing	gpt-neo-2.7B	0.9808	0.9954
Classic	Writing	gpt-j-6B	0.8980	0.9719
Classic	Writing	TL1 blend	0.9132	0.9702
Advanced	XSum	llama-2-7b	0.9731	0.9979
Advanced	XSum	phi-2	0.5891	0.7868
Advanced	XSum	mistral-7b-v0.3	0.3442	0.6780
Advanced	XSum	gemma-7b	0.4573	0.6924
Advanced	XSum	TL1 blend	0.3970	0.8907
Advanced	SQuAD	llama-2-7b	0.9810	0.9973
Advanced	SQuAD	phi-2	0.7108	0.9034
Advanced	SQuAD	mistral-7b-v0.3	0.4534	0.7862
Advanced	SQuAD	gemma-7b	0.5940	0.7987
Advanced	SQuAD	TL1 blend	0.5324	0.9242
Advanced	Writing	llama-2-7b	0.9869	0.9879
Advanced	Writing	phi-2	0.8718	0.9518
Advanced	Writing	mistral-7b-v0.3	0.7385	0.8991
Advanced	Writing	gemma-7b	0.7903	0.9100
Advanced	Writing	TL1 blend	0.8041	0.9760

Table 9: Supervised detectors: single-model minus blend signed AUROC. A negative value means that the blend is detected more readily. Pre-committed outcome for RADAR on native text: point estimate negative in 5 of 6 cells, CI below zero in 0 (a consistent direction without significance).

Comparison	Pool	Data	Blend AUROC	Single mean	Single − blend [CI]
RADAR, native blends vs native pool-of-one	Classic	XSum	0.8914	0.8877	−0.0037 [−0.0162, 0.0093]
		SQuAD	0.8117	0.8129	0.0012 [−0.0190, 0.0218]
		Writing	0.8060	0.8023	−0.0037 [−0.0196, 0.0133]
	Advanced	XSum	0.8762	0.8673	−0.0089 [−0.0228, 0.0057]
		SQuAD	0.8300	0.8298	−0.0002 [−0.0216, 0.0229]
		Writing	0.8278	0.8180	−0.0098 [−0.0260, 0.0066]
RADAR, native blends vs data-builder singles	Classic	XSum	0.8914	0.8904	−0.0011 [−0.0159, 0.0144]
		SQuAD	0.8117	0.8263	0.0146 [−0.0076, 0.0370]
		Writing	0.8060	0.7855	−0.0205 [−0.0381, −0.0021]
	Advanced	XSum	0.8762	0.7994	−0.0768 [−0.0943, −0.0594]
		SQuAD	0.8300	0.7675	−0.0624 [−0.0876, −0.0376]
		Writing	0.8278	0.7179	−0.1099 [−0.1301, −0.0888]
RADAR, 2024 blends vs original-loop pool-of-one	Classic	XSum	0.8958	0.8921	−0.0036 [−0.0174, 0.0098]
		SQuAD	0.8395	0.8449	0.0054 [−0.0156, 0.0275]
		Writing	0.8846	0.8774	−0.0072 [−0.0197, 0.0048]
	Advanced	XSum	0.9028	0.9006	−0.0022 [−0.0160, 0.0119]
		SQuAD	0.8474	0.8702	0.0227 [0.0043, 0.0423]
		Writing	0.8844	0.8983	0.0140 [0.0003, 0.0283]
GPT-2 detector, native vs native pool-of-one	Classic	XSum	0.9254	0.9210	−0.0043 [−0.0158, 0.0081]
		SQuAD	0.9092	0.9300	0.0209 [0.0028, 0.0418]
		Writing	0.9203	0.9201	−0.0002 [−0.0124, 0.0127]
	Advanced	XSum	0.8814	0.8566	−0.0248 [−0.0418, −0.0081]
		SQuAD	0.8714	0.8594	−0.0121 [−0.0324, 0.0103]
		Writing	0.9219	0.8771	−0.0448 [−0.0586, −0.0298]
Hello-SimpleAI, native vs native pool-of-one	Classic	XSum	0.5238	0.5475	0.0238 [−0.0040, 0.0507]
		SQuAD	0.5397	0.5564	0.0167 [−0.0142, 0.0494]
		Writing	0.5899	0.5918	0.0019 [−0.0200, 0.0238]
	Advanced	XSum	0.5748	0.5800	0.0052 [−0.0220, 0.0337]
		SQuAD	0.5967	0.6022	0.0055 [−0.0257, 0.0401]
		Writing	0.6646	0.6463	−0.0183 [−0.0385, 0.0012]

Table 10: Generation-seed variation in the native loop: blend D per seed, between-seed SD and seed-0 CI half-width; the last row gives the pool DiD against native pool-of-one text per seed, its SD and the mean CI half-width.

Pool	Data	Seed 0	Seed 1	Seed 2	SD	CI half-width
Classic	XSum	0.9145	0.8990	0.9003	0.0086	0.0181
Classic	SQuAD	0.9361	0.9265	0.9429	0.0082	0.0200
Classic	Writing	0.9679	0.9702	0.9684	0.0012	0.0104
Advanced	XSum	0.8977	0.8683	0.8860	0.0148	0.0194
Advanced	SQuAD	0.9266	0.9382	0.9274	0.0064	0.0213
Advanced	Writing	0.9760	0.9763	0.9806	0.0026	0.0091
DiD	mean	−0.0619 [−0.0747, −0.0489]	−0.0637 [−0.0760, −0.0498]	−0.0621 [−0.0755, −0.0484]	0.00097	0.0132

Table 11: Instruction-tuned pool (Llama-3.1-8B-Instruct, Qwen2.5-7B-Instruct, Mistral-7B-Instruct-v0.3, Gemma-2-9B-it), native loop, TL1, seed 0, every detector: D of the blend and blend minus the mean D of the four data-builder or pool-of-one arms [95% CI]; the last two columns give the advanced pool's native blend against its data-builder text. Positive differences mean that the blend is more detectable. The outside scorers were applied to only some instruction-tuned arms; cells without complete arms are marked –. Hello-SimpleAI is a formatting diagnostic.

Detector	Data	Blend D	vs. data-builder [CI]	vs. pool-of-one [CI]	Adv. blend D	Adv. vs. data-builder [CI]
Fast-DetectGPT (GPT-Neo, sampled)	XSum	0.9706	0.1090 [0.0943, 0.1238]	0.0439 [0.0331, 0.0551]	0.8977	0.1516 [0.1296, 0.1725]
	SQuAD	0.9844	0.1320 [0.1124, 0.1522]	0.0655 [0.0529, 0.0796]	0.9266	0.1120 [0.0897, 0.1329]
	Writing	0.9864	0.0407 [0.0316, 0.0492]	0.0271 [0.0201, 0.0339]	0.9760	0.0539 [0.0421, 0.0658]
Fast-DetectGPT (GPT-Neo, analytic)	XSum	0.9707	0.1091 [0.0945, 0.1239]	0.0437 [0.0329, 0.0548]	0.8977	0.1515 [0.1294, 0.1723]
	SQuAD	0.9844	0.1318 [0.1124, 0.1521]	0.0657 [0.0529, 0.0798]	0.9265	0.1115 [0.0892, 0.1324]
	Writing	0.9862	0.0404 [0.0313, 0.0491]	0.0270 [0.0200, 0.0338]	0.9758	0.0537 [0.0418, 0.0656]
Fast-DetectGPT (OLMo-2-7B)	XSum	0.9586	0.1564 [0.1372, 0.1757]	–	0.7787	0.1802 [0.1527, 0.2053]
	SQuAD	–	–	–	0.8594	−0.0746 [−0.1004, −0.0504]
	Writing	–	–	–	0.9800	0.0368 [0.0268, 0.0464]
Fast-DetectGPT (Falcon-7B)	XSum	0.9811	0.0905 [0.0778, 0.1028]	–	0.8760	0.1400 [0.1181, 0.1618]
	SQuAD	–	–	–	0.9357	0.1120 [0.0893, 0.1337]
	Writing	–	–	–	0.9766	0.0370 [0.0265, 0.0472]
Binoculars (Falcon-7B pair)	XSum	0.9908	0.0574 [0.0476, 0.0674]	–	0.9384	0.1216 [0.1041, 0.1398]
	SQuAD	–	–	–	0.9635	0.0868 [0.0684, 0.1053]
	Writing	–	–	–	0.9881	0.0381 [0.0301, 0.0465]
RADAR	XSum	0.9607	0.0901 [0.0770, 0.1031]	0.0311 [0.0232, 0.0390]	0.8839	0.0792 [0.0591, 0.0985]
	SQuAD	0.9233	0.0801 [0.0628, 0.0974]	0.0643 [0.0464, 0.0811]	0.8242	0.0561 [0.0290, 0.0846]
	Writing	0.8811	0.0769 [0.0591, 0.0937]	0.0101 [−0.0024, 0.0225]	0.8230	0.1056 [0.0825, 0.1282]
GPT-2 detector	XSum	0.8886	0.0685 [0.0477, 0.0908]	0.0104 [−0.0093, 0.0298]	0.8786	0.0477 [0.0258, 0.0683]
	SQuAD	0.8743	0.1461 [0.1170, 0.1766]	0.0545 [0.0275, 0.0811]	0.8810	0.0664 [0.0415, 0.0907]
	Writing	0.8216	0.0603 [0.0372, 0.0828]	0.0381 [0.0157, 0.0601]	0.9250	0.0594 [0.0428, 0.0746]
Hello-SimpleAI	XSum	0.5322	−0.1319 [−0.1720, −0.0926]	−0.1846 [−0.2211, −0.1465]	0.5810	0.0403 [0.0044, 0.0760]
	SQuAD	0.5825	−0.0173 [−0.0617, 0.0237]	−0.1144 [−0.1586, −0.0721]	0.6015	0.0534 [0.0102, 0.0924]
	Writing	0.6013	−0.0380 [−0.0677, −0.0089]	−0.0532 [−0.0820, −0.0234]	0.6677	0.0225 [−0.0003, 0.0475]

Table 12: Instruction-tuned pool: signed Fast-DetectGPT AUROC (GPT-Neo scorer) of each model's data-builder text (DB) and pool-of-one arm (P1) in the native loop, and of the TL1 blend. Pool-of-one passages that did not reach the 170-token stop criterion are excluded (at least 294 passages per arm and domain remain). Loop arms sample with each model's default configuration (Llama-3.1: temperature 0.6, top-p 0.9; Qwen2.5: temperature 0.7, top-k 20, top-p 0.8, repetition penalty 1.05; Mistral and Gemma-2: temperature 1.0, top-k 50), whereas data-builder text uses temperature 1.0 and top-k 50 for every model. All arms run in bfloat16 from the raw 30-token prefix, without a chat template.

Arm	XSum DB	XSum P1	SQuAD DB	SQuAD P1	Writing DB	Writing P1
Llama-3.1-8B-Instruct	0.8247	0.9827	0.8201	0.9955	0.9421	0.9971
Qwen2.5-7B-Instruct	0.8228	0.9935	0.7645	0.9205	0.9625	0.9913
Mistral-7B-Instruct-v0.3	0.8540	0.8245	0.8682	0.8669	0.9633	0.9547
Gemma-2-9B-it	0.9329	0.8971	0.9395	0.8933	0.9303	0.9028
TL1 blend	0.9669		0.9820		0.9884	

Table 13: Fast-DetectGPT reduction relative to data-builder single-model text, $D(\text{single}) - D(\text{blend})$, on raw (V0) text and after the V2 cleaner is applied to blends, singles and their human partners. The reduction keeps a CI above zero in all six cells after cleaning.

Pool	Data	Single V0	Blend V0	Reduction V0 [CI]	Single V2	Blend V2	Reduction V2 [CI]
Classic	XSum	0.9304	0.7114	0.2190 [0.1906, 0.2498]	0.8670	0.7274	0.1396 [0.1119, 0.1687]
Classic	SQuAD	0.9485	0.8034	0.1451 [0.1121, 0.1776]	0.9192	0.8103	0.1089 [0.0771, 0.1408]
Classic	Writing	0.9720	0.9096	0.0624 [0.0462, 0.0800]	0.9621	0.9066	0.0555 [0.0405, 0.0733]
Advanced	XSum	0.7461	0.6169	0.1293 [0.0846, 0.1794]	0.6819	0.6049	0.0769 [0.0307, 0.1265]
Advanced	SQuAD	0.8147	0.5350	0.2797 [0.2394, 0.3202]	0.7752	0.5405	0.2347 [0.1942, 0.2763]
Advanced	Writing	0.9222	0.8033	0.1189 [0.0961, 0.1435]	0.9052	0.7990	0.1062 [0.0845, 0.1293]

Table 14: Advanced data-builder single-model text: Fast-DetectGPT AUROC after V2 cleaning minus raw AUROC.

Generator	XSum	SQuAD	Writing
llama-2-7b	−0.0031 [−0.0060, −0.0004]	−0.0007 [−0.0036, 0.0024]	0.0000 [−0.0018, 0.0018]
phi-2	−0.0306 [−0.0389, −0.0230]	−0.0449 [−0.0565, −0.0340]	−0.0111 [−0.0152, −0.0071]
mistral-7b-v0.3	−0.1184 [−0.1292, −0.1082]	−0.0679 [−0.0789, −0.0579]	−0.0425 [−0.0495, −0.0357]
gemma-7b	−0.1010 [−0.1104, −0.0922]	−0.0489 [−0.0594, −0.0397]	−0.0233 [−0.0289, −0.0182]

Table 15: Advanced XSum TL4: the 2024 corpus, which retained $k = 3$ tokens per draw, and a corpus regenerated with $k = 4$ in the original loop (signed AUROC, $n = 500$). The $k = 4$ corpus replaces the $k = 3$ cell for any TL4 statement.

Detector	TL4 corpus with $k = 3$	Regenerated $k = 4$	$k = 4$ minus $k = 3$
Fast-DetectGPT (GPT-Neo, sampled)	0.4575 [0.4254, 0.4892]	0.6688 [0.6376, 0.6983]	0.2113 [0.1773, 0.2453]
Fast-DetectGPT (GPT-Neo, analytic)	0.4578 [0.4258, 0.4897]	0.6691 [0.6378, 0.6984]	0.2112 [0.1771, 0.2456]
Fast-DetectGPT (OLMo-2-7B)	0.3341 [0.3033, 0.3641]	0.5405 [0.5069, 0.5729]	0.2064 [0.1707, 0.2408]
Fast-DetectGPT (Falcon-7B)	0.6537 [0.6226, 0.6846]	0.8453 [0.8215, 0.8676]	0.1916 [0.1634, 0.2193]
Binoculars (Falcon-7B pair)	0.7919 [0.7653, 0.8167]	0.9172 [0.9008, 0.9324]	0.1253 [0.1029, 0.1472]
RADAR	0.9391 [0.9255, 0.9527]	0.9488 [0.9353, 0.9614]	0.0096 [−0.0030, 0.0211]
GPT-2 detector	0.5757 [0.5425, 0.6089]	0.5546 [0.5197, 0.5886]	−0.0211 [−0.0561, 0.0125]
Hello-SimpleAI	0.0000 [0.0000, 0.0000]	0.0000 [0.0000, 0.0000]	0.0000 [0.0000, 0.0000]

Table 16: Classic pool without GPT-Neo (GPT-2-XL, OPT-2.7B, GPT-J-6B), native loop, GPT-Neo scorer: blend AUROC and blend minus reference detectability against the three native pool-of-one arms and against the three data-builder singles.

Data	Blend AUROC	vs pool-of-one [CI]	vs data-builder [CI]
XSum	0.8500	−0.0310 [−0.0518, −0.0121]	−0.0541 [−0.0745, −0.0348]
SQuAD	0.8964	−0.0176 [−0.0421, 0.0048]	−0.0346 [−0.0613, −0.0122]
Writing	0.9552	0.0022 [−0.0086, 0.0124]	−0.0087 [−0.0195, 0.0011]

Table 17: Paraphrase attacks, Fast-DetectGPT (GPT-Neo, sampled): signed AUROC before and after paraphrasing, and the change. DIPPER (lexical 60, order 60) was applied to the 2024 TL1 blends and to every data-builder single-model set; the instruction-tuned paraphrase (Qwen2.5-7B-Instruct) was applied to the blends only.

Detector	Paraphraser	Source	Data	Before	After	After − before [CI]
Fast-DetectGPT (GPT-Neo, sampled)	DIPPER	2024 blend, advanced	SQuAD	0.5324	0.8939	0.3615 [0.3250, 0.3976]
	DIPPER	2024 blend, advanced	Writing	0.8041	0.9325	0.1284 [0.1091, 0.1493]
	DIPPER	2024 blend, advanced	XSum	0.3970	0.8294	0.4324 [0.4006, 0.4620]
	DIPPER	2024 blend, classic	SQuAD	0.8049	0.9285	0.1236 [0.0974, 0.1499]
	DIPPER	2024 blend, classic	Writing	0.9132	0.9535	0.0403 [0.0272, 0.0551]
	DIPPER	2024 blend, classic	XSum	0.7085	0.8654	0.1569 [0.1325, 0.1822]
	DIPPER	single, advanced-gemma-7b	SQuAD	0.8173	0.8945	0.0772 [0.0498, 0.1045]
	DIPPER	single, advanced-gemma-7b	Writing	0.9319	0.9436	0.0118 [−0.0005, 0.0237]
	DIPPER	single, advanced-gemma-7b	XSum	0.7775	0.8465	0.0690 [0.0474, 0.0900]
	DIPPER	single, advanced-llama-2-7b	SQuAD	0.7430	0.8583	0.1153 [0.0864, 0.1433]
	DIPPER	single, advanced-llama-2-7b	Writing	0.8847	0.9282	0.0435 [0.0279, 0.0590]
	DIPPER	single, advanced-llama-2-7b	XSum	0.6473	0.8121	0.1648 [0.1385, 0.1924]
	DIPPER	single, advanced-mistral-7b-v0.3	SQuAD	0.7696	0.8694	0.0998 [0.0694, 0.1329]
	DIPPER	single, advanced-mistral-7b-v0.3	Writing	0.9134	0.9391	0.0258 [0.0114, 0.0396]
	DIPPER	single, advanced-mistral-7b-v0.3	XSum	0.7207	0.8126	0.0919 [0.0670, 0.1187]
	DIPPER	single, advanced-phi-2	SQuAD	0.9109	0.9026	−0.0083 [−0.0320, 0.0118]
	DIPPER	single, advanced-phi-2	Writing	0.9568	0.9493	−0.0075 [−0.0177, 0.0028]
	DIPPER	single, advanced-phi-2	XSum	0.8117	0.8602	0.0486 [0.0260, 0.0718]
	DIPPER	single, classic-gpt-j-6B	SQuAD	0.9286	0.9136	−0.0150 [−0.0356, 0.0049]
	DIPPER	single, classic-gpt-j-6B	Writing	0.9745	0.9520	−0.0226 [−0.0327, −0.0130]
	DIPPER	single, classic-gpt-j-6B	XSum	0.8869	0.8563	−0.0306 [−0.0510, −0.0107]
	DIPPER	single, classic-gpt-neo-2.7B	SQuAD	0.9957	0.9269	−0.0688 [−0.0887, −0.0509]
	DIPPER	single, classic-gpt-neo-2.7B	Writing	0.9982	0.9584	−0.0398 [−0.0502, −0.0298]
	DIPPER	single, classic-gpt-neo-2.7B	XSum	0.9882	0.8994	−0.0888 [−0.1054, −0.0724]
	DIPPER	single, classic-gpt2-xl	SQuAD	0.9348	0.9220	−0.0128 [−0.0306, 0.0042]
	DIPPER	single, classic-gpt2-xl	Writing	0.9794	0.9519	−0.0275 [−0.0368, −0.0185]
	DIPPER	single, classic-gpt2-xl	XSum	0.9302	0.8698	−0.0605 [−0.0801, −0.0431]
	DIPPER	single, classic-opt-2.7b	SQuAD	0.9314	0.9077	−0.0237 [−0.0448, −0.0043]
	DIPPER	single, classic-opt-2.7b	Writing	0.9342	0.9466	0.0123 [−0.0006, 0.0248]
	DIPPER	single, classic-opt-2.7b	XSum	0.9073	0.8710	−0.0362 [−0.0547, −0.0191]
	Qwen2.5-7B-Instruct	2024 blend, advanced	SQuAD	0.5324	0.6123	0.0800 [0.0441, 0.1154]
	Qwen2.5-7B-Instruct	2024 blend, advanced	Writing	0.8041	0.7361	−0.0680 [−0.0932, −0.0447]
	Qwen2.5-7B-Instruct	2024 blend, advanced	XSum	0.3970	0.4203	0.0233 [−0.0100, 0.0550]
	Qwen2.5-7B-Instruct	2024 blend, classic	SQuAD	0.8049	0.6604	−0.1445 [−0.1797, −0.1119]
	Qwen2.5-7B-Instruct	2024 blend, classic	Writing	0.9132	0.7415	−0.1717 [−0.1956, −0.1480]
	Qwen2.5-7B-Instruct	2024 blend, classic	XSum	0.7085	0.4767	−0.2318 [−0.2634, −0.2013]

Table 18: Paraphrase attacks, Binoculars (Falcon-7B pair): signed AUROC before and after paraphrasing, and the change. DIPPER (lexical 60, order 60) was applied to the 2024 TL1 blends and to every data-builder single-model set; the instruction-tuned paraphrase (Qwen2.5-7B-Instruct) was applied to the blends only.

Detector	Paraphraser	Source	Data	Before	After	After − before [CI]
Binoculars (Falcon-7B pair)	DIPPER	2024 blend, advanced	SQuAD	0.7365	0.9352	0.1987 [0.1677, 0.2289]
	DIPPER	2024 blend, advanced	Writing	0.8681	0.9496	0.0815 [0.0662, 0.0977]
	DIPPER	2024 blend, advanced	XSum	0.6380	0.9007	0.2627 [0.2349, 0.2889]
	DIPPER	2024 blend, classic	SQuAD	0.8345	0.9520	0.1176 [0.0926, 0.1422]
	DIPPER	2024 blend, classic	Writing	0.9009	0.9572	0.0563 [0.0429, 0.0708]
	DIPPER	2024 blend, classic	XSum	0.7270	0.9167	0.1897 [0.1665, 0.2131]
	DIPPER	single, advanced-gemma-7b	SQuAD	0.8826	0.9367	0.0541 [0.0314, 0.0774]
	DIPPER	single, advanced-gemma-7b	Writing	0.9496	0.9558	0.0063 [−0.0052, 0.0171]
	DIPPER	single, advanced-gemma-7b	XSum	0.8260	0.9121	0.0861 [0.0680, 0.1053]
	DIPPER	single, advanced-llama-2-7b	SQuAD	0.8761	0.9186	0.0425 [0.0181, 0.0662]
	DIPPER	single, advanced-llama-2-7b	Writing	0.9424	0.9395	−0.0028 [−0.0138, 0.0079]
	DIPPER	single, advanced-llama-2-7b	XSum	0.8006	0.8776	0.0769 [0.0556, 0.0987]
	DIPPER	single, advanced-mistral-7b-v0.3	SQuAD	0.8289	0.9162	0.0874 [0.0623, 0.1150]
	DIPPER	single, advanced-mistral-7b-v0.3	Writing	0.9331	0.9529	0.0198 [0.0096, 0.0303]
	DIPPER	single, advanced-mistral-7b-v0.3	XSum	0.7770	0.8782	0.1012 [0.0785, 0.1241]
	DIPPER	single, advanced-phi-2	SQuAD	0.9104	0.9395	0.0291 [0.0111, 0.0479]
	DIPPER	single, advanced-phi-2	Writing	0.9668	0.9638	−0.0030 [−0.0118, 0.0064]
	DIPPER	single, advanced-phi-2	XSum	0.8494	0.9013	0.0520 [0.0313, 0.0725]
	DIPPER	single, classic-gpt-j-6B	SQuAD	0.9140	0.9403	0.0263 [0.0036, 0.0481]
	DIPPER	single, classic-gpt-j-6B	Writing	0.9645	0.9622	−0.0023 [−0.0113, 0.0065]
	DIPPER	single, classic-gpt-j-6B	XSum	0.8498	0.9126	0.0628 [0.0452, 0.0805]
	DIPPER	single, classic-gpt-neo-2.7B	SQuAD	0.9001	0.9410	0.0409 [0.0222, 0.0593]
	DIPPER	single, classic-gpt-neo-2.7B	Writing	0.9485	0.9572	0.0087 [−0.0009, 0.0189]
	DIPPER	single, classic-gpt-neo-2.7B	XSum	0.8384	0.9074	0.0690 [0.0505, 0.0885]
	DIPPER	single, classic-gpt2-xl	SQuAD	0.9286	0.9448	0.0162 [−0.0008, 0.0344]
	DIPPER	single, classic-gpt2-xl	Writing	0.9605	0.9648	0.0043 [−0.0034, 0.0122]
	DIPPER	single, classic-gpt2-xl	XSum	0.8612	0.9117	0.0505 [0.0315, 0.0686]
	DIPPER	single, classic-opt-2.7b	SQuAD	0.9182	0.9432	0.0250 [0.0088, 0.0412]
	DIPPER	single, classic-opt-2.7b	Writing	0.9559	0.9589	0.0030 [−0.0076, 0.0136]
	DIPPER	single, classic-opt-2.7b	XSum	0.8687	0.9156	0.0470 [0.0292, 0.0652]
	Qwen2.5-7B-Instruct	2024 blend, advanced	SQuAD	0.7365	0.7238	−0.0127 [−0.0451, 0.0201]
	Qwen2.5-7B-Instruct	2024 blend, advanced	Writing	0.8681	0.8210	−0.0471 [−0.0680, −0.0267]
	Qwen2.5-7B-Instruct	2024 blend, advanced	XSum	0.6380	0.5152	−0.1228 [−0.1533, −0.0926]
	Qwen2.5-7B-Instruct	2024 blend, classic	SQuAD	0.8345	0.7208	−0.1137 [−0.1465, −0.0828]
	Qwen2.5-7B-Instruct	2024 blend, classic	Writing	0.9009	0.7894	−0.1115 [−0.1340, −0.0889]
	Qwen2.5-7B-Instruct	2024 blend, classic	XSum	0.7270	0.5079	−0.2191 [−0.2508, −0.1909]

Table 19: Paraphrase attacks, RADAR: signed AUROC before and after paraphrasing, and the change. DIPPER (lexical 60, order 60) was applied to the 2024 TL1 blends and to every data-builder single-model set; the instruction-tuned paraphrase (Qwen2.5-7B-Instruct) was applied to the blends only.

Detector	Paraphraser	Source	Data	Before	After	After − before [CI]
RADAR	DIPPER	2024 blend, advanced	SQuAD	0.8474	0.7965	−0.0509 [−0.0737, −0.0276]
	DIPPER	2024 blend, advanced	Writing	0.8844	0.7884	−0.0960 [−0.1162, −0.0754]
	DIPPER	2024 blend, advanced	XSum	0.9028	0.9293	0.0266 [0.0118, 0.0414]
	DIPPER	2024 blend, classic	SQuAD	0.8395	0.8204	−0.0191 [−0.0406, 0.0010]
	DIPPER	2024 blend, classic	Writing	0.8846	0.7709	−0.1138 [−0.1343, −0.0929]
	DIPPER	2024 blend, classic	XSum	0.8958	0.9399	0.0441 [0.0318, 0.0567]
	DIPPER	single, advanced-gemma-7b	SQuAD	0.8203	0.7778	−0.0426 [−0.0680, −0.0164]
	DIPPER	single, advanced-gemma-7b	Writing	0.7334	0.7057	−0.0277 [−0.0502, −0.0047]
	DIPPER	single, advanced-gemma-7b	XSum	0.8516	0.9050	0.0534 [0.0361, 0.0704]
	DIPPER	single, advanced-llama-2-7b	SQuAD	0.6544	0.7687	0.1143 [0.0891, 0.1405]
	DIPPER	single, advanced-llama-2-7b	Writing	0.6141	0.6978	0.0837 [0.0617, 0.1063]
	DIPPER	single, advanced-llama-2-7b	XSum	0.6906	0.9013	0.2107 [0.1871, 0.2360]
	DIPPER	single, advanced-mistral-7b-v0.3	SQuAD	0.7828	0.7854	0.0026 [−0.0238, 0.0291]
	DIPPER	single, advanced-mistral-7b-v0.3	Writing	0.7108	0.6812	−0.0295 [−0.0532, −0.0052]
	DIPPER	single, advanced-mistral-7b-v0.3	XSum	0.7975	0.8875	0.0900 [0.0710, 0.1097]
	DIPPER	single, advanced-phi-2	SQuAD	0.8125	0.7792	−0.0333 [−0.0599, −0.0073]
	DIPPER	single, advanced-phi-2	Writing	0.8134	0.7620	−0.0514 [−0.0729, −0.0310]
	DIPPER	single, advanced-phi-2	XSum	0.8579	0.9215	0.0635 [0.0484, 0.0791]
	DIPPER	single, classic-gpt-j-6B	SQuAD	0.7833	0.7653	−0.0179 [−0.0439, 0.0082]
	DIPPER	single, classic-gpt-j-6B	Writing	0.7873	0.6749	−0.1124 [−0.1373, −0.0880]
	DIPPER	single, classic-gpt-j-6B	XSum	0.8873	0.9136	0.0263 [0.0120, 0.0405]
	DIPPER	single, classic-gpt-neo-2.7B	SQuAD	0.8217	0.7946	−0.0271 [−0.0531, −0.0014]
	DIPPER	single, classic-gpt-neo-2.7B	Writing	0.7972	0.7264	−0.0709 [−0.0921, −0.0502]
	DIPPER	single, classic-gpt-neo-2.7B	XSum	0.9212	0.9396	0.0184 [0.0068, 0.0307]
	DIPPER	single, classic-gpt2-xl	SQuAD	0.8602	0.8165	−0.0437 [−0.0664, −0.0207]
	DIPPER	single, classic-gpt2-xl	Writing	0.7940	0.7094	−0.0846 [−0.1068, −0.0625]
	DIPPER	single, classic-gpt2-xl	XSum	0.9029	0.9270	0.0241 [0.0121, 0.0369]
	DIPPER	single, classic-opt-2.7b	SQuAD	0.8400	0.7808	−0.0593 [−0.0835, −0.0357]
	DIPPER	single, classic-opt-2.7b	Writing	0.7636	0.7370	−0.0266 [−0.0491, −0.0045]
	DIPPER	single, classic-opt-2.7b	XSum	0.8500	0.9273	0.0773 [0.0612, 0.0938]
	Qwen2.5-7B-Instruct	2024 blend, advanced	SQuAD	0.8474	0.8083	−0.0392 [−0.0567, −0.0204]
	Qwen2.5-7B-Instruct	2024 blend, advanced	Writing	0.8844	0.8718	−0.0125 [−0.0258, 0.0010]
	Qwen2.5-7B-Instruct	2024 blend, advanced	XSum	0.9028	0.9411	0.0384 [0.0268, 0.0503]
	Qwen2.5-7B-Instruct	2024 blend, classic	SQuAD	0.8395	0.8344	−0.0051 [−0.0239, 0.0126]
	Qwen2.5-7B-Instruct	2024 blend, classic	Writing	0.8846	0.8799	−0.0047 [−0.0192, 0.0091]
	Qwen2.5-7B-Instruct	2024 blend, classic	XSum	0.8958	0.9425	0.0467 [0.0363, 0.0580]

Table 20: After DIPPER paraphrasing of both arms: signed AUROC of the paraphrased blend minus the mean over the paraphrased single-model sets of the same pool.

Detector	Pool	Data	Single+DIPPER mean	Blend − single [CI]
Fast-DetectGPT (GPT-Neo, sampled)	Classic	XSum	0.8741	−0.0087 [−0.0266, 0.0092]
		SQuAD	0.9175	0.0109 [−0.0048, 0.0258]
		Writing	0.9522	0.0013 [−0.0074, 0.0100]
	Advanced	XSum	0.8329	−0.0035 [−0.0243, 0.0174]
		SQuAD	0.8812	0.0127 [−0.0068, 0.0312]
		Writing	0.9401	−0.0075 [−0.0202, 0.0053]
Fast-DetectGPT (GPT-Neo, analytic)	Classic	XSum	0.8741	−0.0083 [−0.0264, 0.0098]
		SQuAD	0.9174	0.0110 [−0.0049, 0.0256]
		Writing	0.9522	0.0011 [−0.0076, 0.0099]
	Advanced	XSum	0.8330	−0.0034 [−0.0242, 0.0178]
		SQuAD	0.8813	0.0126 [−0.0067, 0.0309]
		Writing	0.9401	−0.0076 [−0.0202, 0.0050]
Fast-DetectGPT (OLMo-2-7B)	Classic	XSum	0.7351	−0.0031 [−0.0265, 0.0203]
		SQuAD	0.1198	0.0101 [−0.0078, 0.0290]
		Writing	0.9628	−0.0022 [−0.0100, 0.0054]
	Advanced	XSum	0.7252	0.0093 [−0.0137, 0.0300]
		SQuAD	0.1399	−0.0082 [−0.0292, 0.0135]
		Writing	0.9609	−0.0077 [−0.0178, 0.0021]
Fast-DetectGPT (Falcon-7B)	Classic	XSum	0.7932	0.0026 [−0.0202, 0.0239]
		SQuAD	0.8375	0.0045 [−0.0208, 0.0282]
		Writing	0.9319	−0.0112 [−0.0223, −0.0002]
	Advanced	XSum	0.7949	0.0039 [−0.0182, 0.0256]
		SQuAD	0.8459	0.0083 [−0.0177, 0.0333]
		Writing	0.9312	−0.0141 [−0.0263, −0.0016]
Binoculars (Falcon-7B pair)	Classic	XSum	0.9118	0.0049 [−0.0083, 0.0181]
		SQuAD	0.9423	0.0097 [−0.0050, 0.0229]
		Writing	0.9608	−0.0035 [−0.0115, 0.0047]
	Advanced	XSum	0.8923	0.0084 [−0.0062, 0.0233]
		SQuAD	0.9278	0.0075 [−0.0109, 0.0255]
		Writing	0.9530	−0.0034 [−0.0135, 0.0071]
RADAR	Classic	XSum	0.9269	0.0130 [0.0022, 0.0236]
		SQuAD	0.7893	0.0311 [0.0068, 0.0537]
		Writing	0.7119	0.0590 [0.0357, 0.0815]
	Advanced	XSum	0.9038	0.0255 [0.0115, 0.0391]
		SQuAD	0.7778	0.0187 [−0.0081, 0.0441]
		Writing	0.7117	0.0767 [0.0542, 0.1000]
GPT-2 detector	Classic	XSum	0.9676	0.0047 [−0.0008, 0.0096]
		SQuAD	0.9670	0.0082 [−0.0011, 0.0167]
		Writing	0.9674	0.0003 [−0.0069, 0.0076]
	Advanced	XSum	0.9406	0.0101 [0.0012, 0.0192]
		SQuAD	0.9404	0.0085 [−0.0045, 0.0216]
		Writing	0.9439	0.0054 [−0.0044, 0.0154]
Hello-SimpleAI	Classic	XSum	0.7207	−0.0303 [−0.0598, 0.0007]
		SQuAD	0.7386	−0.0053 [−0.0362, 0.0257]
		Writing	0.7855	−0.0273 [−0.0522, −0.0032]
	Advanced	XSum	0.7093	−0.0566 [−0.0878, −0.0247]
		SQuAD	0.7350	−0.0130 [−0.0452, 0.0200]
		Writing	0.7824	−0.0285 [−0.0574, −0.0033]

Table 21: Decomposition of the GPT-Neo Fast-DetectGPT statistic, 2024 TL1 blends minus the mean of the original-loop pool-of-one arms, per token (nats): observed log-likelihood, the scorer's expected log-likelihood under its own sampling distribution, the standard deviation, and the resulting analytic curvature.

Pool	Data	Log-likelihood	Expected log-likelihood	SD	Curvature
Classic	XSum	−0.0287 [−0.0530, −0.0034]	−0.0288 [−0.0529, −0.0048]	0.0121 [0.0037, 0.0206]	−0.0055 [−0.0924, 0.0790]
Classic	SQuAD	−0.0228 [−0.0526, 0.0088]	−0.0145 [−0.0402, 0.0114]	0.0032 [−0.0066, 0.0128]	−0.0465 [−0.1530, 0.0647]
Classic	Writing	−0.0077 [−0.0369, 0.0193]	0.0105 [−0.0184, 0.0389]	0.0001 [−0.0098, 0.0097]	−0.1017 [−0.1846, −0.0244]
Advanced	XSum	−0.3415 [−0.3701, −0.3117]	−0.1728 [−0.1981, −0.1482]	0.0668 [0.0568, 0.0766]	−1.0254 [−1.1138, −0.9323]
Advanced	SQuAD	−0.3089 [−0.3415, −0.2776]	−0.1456 [−0.1755, −0.1182]	0.0582 [0.0479, 0.0688]	−1.0005 [−1.1171, −0.8831]
Advanced	Writing	−0.3981 [−0.4292, −0.3676]	−0.2806 [−0.3089, −0.2523]	0.0834 [0.0741, 0.0930]	−0.7455 [−0.8370, −0.6620]

Table 22: Switch rate in the native loop: signed AUROC [95% CI] against the per-step redraw probability p (100 prompts per domain for $p < 1$; $p = 1$ is the full native blend corpus). $p = 0$ keeps one model for the whole passage.

Detector	Pool	Data	$p = 0$	$p = 0.25$	$p = 0.5$	$p = 0.75$	$p = 1$
Fast-DetectGPT (GPT-Neo, sampled)	Classic	XSum	0.905 [0.863, 0.943]	0.894 [0.853, 0.936]	0.941 [0.908, 0.969]	0.909 [0.869, 0.944]	0.909 [0.892, 0.925]
		SQuAD	0.949 [0.923, 0.973]	0.950 [0.921, 0.974]	0.933 [0.895, 0.966]	0.946 [0.914, 0.973]	0.938 [0.920, 0.954]
		Writing	0.976 [0.958, 0.991]	0.981 [0.965, 0.993]	0.966 [0.941, 0.985]	0.981 [0.965, 0.993]	0.970 [0.962, 0.979]
	Advanced	XSum	0.765 [0.704, 0.826]	0.896 [0.856, 0.934]	0.920 [0.882, 0.953]	0.881 [0.834, 0.924]	0.891 [0.873, 0.908]
		SQuAD	0.878 [0.835, 0.921]	0.953 [0.924, 0.977]	0.961 [0.937, 0.981]	0.932 [0.894, 0.964]	0.924 [0.903, 0.942]
		Writing	0.936 [0.902, 0.964]	0.974 [0.953, 0.992]	0.980 [0.962, 0.994]	0.983 [0.967, 0.996]	0.976 [0.968, 0.984]
Binoculars (Falcon-7B pair)	Classic	XSum	0.839 [0.785, 0.889]	0.812 [0.761, 0.868]	0.868 [0.820, 0.911]	0.799 [0.745, 0.852]	0.822 [0.799, 0.845]
		SQuAD	0.920 [0.885, 0.952]	0.929 [0.894, 0.960]	0.923 [0.886, 0.955]	0.922 [0.888, 0.953]	0.892 [0.869, 0.915]
		Writing	0.960 [0.935, 0.981]	0.957 [0.928, 0.981]	0.958 [0.935, 0.977]	0.971 [0.952, 0.987]	0.942 [0.929, 0.955]
	Advanced	XSum	0.839 [0.787, 0.889]	0.949 [0.923, 0.973]	0.956 [0.930, 0.978]	0.930 [0.895, 0.963]	0.935 [0.922, 0.948]
		SQuAD	0.936 [0.906, 0.964]	0.980 [0.959, 0.994]	0.990 [0.978, 0.997]	0.972 [0.946, 0.991]	0.961 [0.947, 0.974]
		Writing	0.974 [0.956, 0.988]	0.992 [0.984, 0.998]	0.995 [0.988, 0.999]	0.995 [0.987, 0.999]	0.987 [0.982, 0.992]
RADAR	Classic	XSum	0.920 [0.886, 0.951]	0.921 [0.887, 0.952]	0.902 [0.862, 0.935]	0.888 [0.842, 0.927]	0.891 [0.872, 0.910]
		SQuAD	0.787 [0.734, 0.836]	0.791 [0.750, 0.839]	0.807 [0.752, 0.858]	0.820 [0.770, 0.867]	0.812 [0.783, 0.840]
		Writing	0.781 [0.720, 0.840]	0.777 [0.720, 0.833]	0.791 [0.728, 0.851]	0.772 [0.714, 0.830]	0.806 [0.781, 0.830]
	Advanced	XSum	0.888 [0.846, 0.924]	0.921 [0.888, 0.951]	0.883 [0.841, 0.923]	0.895 [0.856, 0.930]	0.876 [0.856, 0.895]
		SQuAD	0.828 [0.781, 0.873]	0.814 [0.762, 0.865]	0.827 [0.777, 0.870]	0.791 [0.743, 0.838]	0.830 [0.801, 0.857]
		Writing	0.818 [0.758, 0.870]	0.827 [0.776, 0.875]	0.815 [0.763, 0.870]	0.814 [0.767, 0.862]	0.828 [0.805, 0.849]

Table 23: Signed Fast-DetectGPT AUROC of blend corpora under three scorers. OLMo-2-7B assigns inverted scores to all SQuAD corpora, which the orientation-free D absorbs.

Corpus	Data	GPT-Neo	Falcon-7B	OLMo-2-7B
2024 classic TL1	XSum	0.7085 [0.6790, 0.7379]	0.4848 [0.4535, 0.5145]	0.2869 [0.2585, 0.3149]
	SQuAD	0.8049 [0.7733, 0.8376]	0.6292 [0.5903, 0.6699]	0.0187 [0.0116, 0.0274]
	Writing	0.9132 [0.8947, 0.9300]	0.8317 [0.8078, 0.8544]	0.7899 [0.7644, 0.8151]
2024 advanced TL1	XSum	0.3970 [0.3661, 0.4308]	0.4822 [0.4497, 0.5175]	0.2647 [0.2354, 0.2947]
	SQuAD	0.5324 [0.4901, 0.5720]	0.6158 [0.5775, 0.6548]	0.0197 [0.0116, 0.0294]
	Writing	0.8041 [0.7781, 0.8292]	0.8318 [0.8091, 0.8539]	0.7913 [0.7667, 0.8169]
Native classic	XSum	0.9089 [0.8921, 0.9251]	0.6122 [0.5819, 0.6445]	0.5042 [0.4733, 0.5349]
	SQuAD	0.9379 [0.9199, 0.9541]	0.7063 [0.6688, 0.7428]	0.0343 [0.0232, 0.0471]
	Writing	0.9702 [0.9619, 0.9791]	0.8864 [0.8676, 0.9050]	0.8972 [0.8791, 0.9148]
Native advanced	XSum	0.8907 [0.8729, 0.9083]	0.8779 [0.8586, 0.8968]	0.7802 [0.7538, 0.8054]
	SQuAD	0.9242 [0.9032, 0.9425]	0.9316 [0.9098, 0.9498]	0.1436 [0.1167, 0.1708]
	Writing	0.9760 [0.9679, 0.9839]	0.9768 [0.9686, 0.9845]	0.9769 [0.9695, 0.9842]
Native classic, no GPT-Neo	XSum	0.8500 [0.8289, 0.8716]	0.6356 [0.6041, 0.6672]	0.5438 [0.5115, 0.5755]
	SQuAD	0.8964 [0.8711, 0.9192]	0.7570 [0.7184, 0.7926]	0.0376 [0.0248, 0.0516]
	Writing	0.9552 [0.9431, 0.9661]	0.9001 [0.8818, 0.9174]	0.9099 [0.8927, 0.9259]

Table 24: Outcomes of the decision rules written before these jobs ran (dated 2026-09-23; round-2 rule G3 dated 2026-09-24).

Decision	Outcome
D1 pool difference (primary)	a
D2 switching (original loop, primary)	c
D2 switching (native loop)	c
D3 RADAR direction	b
D4 seed variation	seed-variation-small
D5 Binoculars	reduction
D5 instruction-tuned pool	pool-dependent
G3 cleaned text	persists
D6 title	measurement-title
D7 probe and FDG-Ens	appendix

B AUDITED HISTORICAL RAW BENCHMARK

Tables 26 and 27 preserve the broad raw-text battery after the cell audit. Each reference column now uses its own pool: the classic table uses the four classic generators, and the advanced table uses the new full-sample advanced arm. Table 1 highlights the resulting Fast-DetectGPT contrast. Raw supervised scores are retained for provenance, while cleaned results are the headline. Historical white-box Fast-DetectGPT single-model averages were 0.9845/0.9937/0.9977, but these used each generator as its own scorer and cannot quantify a drop against blended text scored by GPT-Neo. The tables distinguish corpora: the classic sentence likelihood-family rows use the older shorter unbuffered corpus, while buffered TL1–TL5/random and the later sentence detector corpora follow separate paths. They should not be interpreted as a perfectly isolated sweep of one parameter. For XSum TL2, the secondary baseline excluding the scorer's own generator is 0.9083, giving a reduction of 0.2079 [0.1850, 0.2334].

Table 25: Historical Fast-DetectGPT self-scorer replication. Each generator also serves as its own scorer/reference; these values are not the fixed-GPT-Neo attack baseline in Table 1. The final row is a published three-domain mean from Table 2 of Bao et al. (2024), independently checked against the ICLR 2024 proceedings. The publication uses 500 samples per dataset, whereas our stored SQuAD set has 300; this row is a published comparison, not an exact matched rerun.

Dataset	GPT-2-XL	OPT	GPT-Neo	GPT-J
XSum	0.9922	0.9806	0.9881	0.9771
SQuAD	0.9990	0.9949	0.9956	0.9854
Writing	0.9982	0.9972	0.9981	0.9974
Avg.	0.9965	0.9909	0.9939	0.9866
Bao et al. (2024), Table 2	0.9967	0.9908	0.9940	0.9866

Table 26: Audited historical raw-text AUROC, classic pool, $n = 500/300/500$ for XSum/SQuAD/Writing. Classic ref. is the four-classic-generator mean for every row (Fast-DetectGPT updated to the fixed GPT-Neo scorer); it is not an advanced-pool baseline. Lower signed AUROC is not necessarily weaker separability. GPT-2 detector and RADAR point estimates at TL2–TL5 have no recovered per-passage arrays; the new cleaned results are in Table 30. TL1–TL5 and random are buffered corpora. Sent. uses unbuffered sentence corpora; classic likelihood/rank/logrank/entropy sentence cells come from the older shorter corpus, unlike the Fast-DetectGPT sentence cells. Advanced XSum TL4 is actually $k = 3$. Bolding marks the lowest printed setting strictly below the reference, with ties retained; references are never bold.

Datasets	Detection Method	Classic ref.	TL=1	TL=2	TL=3	TL=4	TL=5	Rand.	Sent.
XSum	Likelihood	0.7837	**0.3444**	0.3526	0.3696	0.3716	0.3738	0.3474	0.4492
	Rank	0.7502	**0.4845**	0.4984	0.4956	0.5027	0.5066	0.5130	0.5797
	LogRank	0.8117	0.4201	**0.4189**	0.4355	0.4380	0.4339	0.4190	0.5191
	Entropy	0.5300	0.7971	0.7803	0.7714	0.7711	0.7644	0.7792	0.7435
	GPT-2 Detector	0.9416	0.7211	0.6924	0.7079	0.6922	0.6948	**0.6776**	0.7025
	RADAR	0.8904	0.8958	0.9225	0.9080	0.9130	0.9142	0.8937	0.8963
	Fast-DetectGPT	0.9283	0.7088	**0.7004**	0.7172	0.7291	0.7250	0.7026	0.8062
SQuAD	Likelihood	0.7573	**0.4350**	0.4707	0.4961	0.4847	0.5124	0.4624	0.5493
	Rank	0.7836	**0.5485**	0.5758	0.5728	0.5801	0.5790	0.5672	0.6224
	LogRank	0.8090	**0.5281**	0.5609	0.5850	0.5711	0.5993	0.5508	0.6357
	Entropy	0.5617	0.7469	0.7551	0.7212	0.7333	0.7208	0.7441	0.7152
	GPT-2 Detector	0.9335	0.7208	0.7027	**0.6922**	0.6963	0.7085	0.6991	0.7107
	RADAR	0.8263	0.8395	0.8641	0.8670	0.8812	0.8532	0.8564	**0.8067**
	Fast-DetectGPT	0.9474	**0.8047**	0.8643	0.8581	0.8557	0.8685	0.8282	0.9062
Writing	Likelihood	0.8904	**0.7029**	0.7293	0.7176	0.7444	0.7407	0.7123	0.7866
	Rank	0.8186	**0.6558**	0.6672	0.6822	0.6826	0.6867	0.6762	0.7145
	LogRank	0.9158	**0.7585**	0.7823	0.7727	0.7949	0.7904	0.7669	0.8305
	Entropy	0.3752	0.5708	0.5795	0.5840	0.5573	0.5706	0.5885	0.5135
	GPT-2 Detector	0.9365	0.7434	0.7316	0.7474	0.7405	**0.7303**	0.7444	0.7437
	RADAR	0.7855	0.8846	0.8347	0.8305	0.8275	0.8203	0.8345	**0.7761**
	Fast-DetectGPT	0.9714	**0.9128**	0.9451	0.9375	0.9439	0.9480	0.9379	0.9581

Table 27: Audited historical raw-text AUROC, advanced pool, $n = 500/300/500$ for XSum/SQuAD/Writing. Advanced ref. is the four-advanced-generator mean from the new full-sample data-builder arm, using the same GPT-Neo scorer or calibrated supervised detector as its row; generation and length protocols still differ from ToBlend. Lower signed AUROC is not necessarily weaker separability. GPT-2 detector and RADAR point estimates at TL2–TL5 have no recovered per-passage arrays; the new cleaned results are in Table 30. TL1–TL5 and random are buffered corpora. Sent. uses unbuffered sentence corpora; classic likelihood/rank/logrank/entropy sentence cells come from the older shorter corpus, unlike the Fast-DetectGPT sentence cells. Advanced XSum TL4 is actually $k = 3$. Bolding marks the lowest printed setting strictly below the reference, with ties retained; references are never bold.

Datasets	Detection Method	Adv. ref.	TL=1	TL=2	TL=3	TL=4	TL=5	Rand.	Sent.
XSum	Likelihood	0.6603	**0.1938**	0.2190	0.2036	0.2493	0.2396	0.2122	0.3023
	Rank	0.6689	**0.3538**	0.4036	0.3723	0.4246	0.3896	0.3794	0.4642
	LogRank	0.6820	**0.2426**	0.2606	0.2498	0.2917	0.2878	0.2506	0.3389
	Entropy	0.4774	0.8121	0.8002	0.8057	0.7723	0.7774	0.8025	0.7495
	GPT-2 Detector	0.8287	0.5622	**0.5411**	0.5723	0.5757	0.5546	0.5414	0.5520
	RADAR	0.7994	0.9027	0.9412	0.9423	0.9391	0.9419	0.9307	0.9075
	Fast-DetectGPT	0.7395	**0.3968**	0.4479	0.4203	0.4580	0.4499	0.4285	0.5382
SQuAD	Likelihood	0.6921	**0.2968**	0.3451	0.3638	0.3844	0.3762	0.3541	0.4800
	Rank	0.6992	**0.4016**	0.4297	0.4344	0.4730	0.4627	0.4472	0.5286
	LogRank	0.7301	**0.3600**	0.4041	0.4267	0.4545	0.4422	0.4140	0.5360
	Entropy	0.4844	0.7363	0.7096	0.6990	0.6961	0.6968	0.6974	0.6290
	GPT-2 Detector	0.8120	0.5221	0.5097	**0.5052**	0.5367	0.5131	0.5056	0.5368
	RADAR	0.7675	0.8474	0.8820	0.8785	0.8846	0.8862	0.8675	0.8380
	Fast-DetectGPT	0.8105	**0.5317**	0.5803	0.5816	0.6178	0.6024	0.5684	0.7046
Writing	Likelihood	0.8567	**0.5845**	0.6354	0.6461	0.6543	0.6520	0.6362	0.7126
	Rank	0.7698	**0.5952**	0.6101	0.6220	0.6358	0.6348	0.6197	0.6641
	LogRank	0.8727	**0.6358**	0.6807	0.6903	0.6980	0.6966	0.6828	0.7487
	Entropy	0.3296	0.6061	0.5619	0.5575	0.5626	0.5532	0.5766	0.5102
	GPT-2 Detector	0.8665	0.6484	0.6436	0.6280	0.6303	0.6381	0.6473	**0.6230**
	RADAR	0.7179	0.8844	0.8294	0.8233	0.8261	0.8407	0.8419	0.8056
	Fast-DetectGPT	0.9217	**0.8036**	0.8272	0.8398	0.8572	0.8448	0.8348	0.8846

C Complete Supervised Cleaning Results

V2 applies the same cleaner to human and generated texts before scoring. Its 69 canonical-cleaner self-tests pass. Calibrated AI-label orientation is used throughout; raw and cleaned results remain separate. New interval coverage is exactly the 18 TL1/random/sentence cells per detector. The raw GPT-2 detector and RADAR reproductions pass their anchors, and their calibration passes on held-out examples. The orientation audit of the broader historical battery identifies 26 inverted-separation cells and nine difference intervals that include zero, so neither all low scores nor all point-estimate drops warrant an evasion claim.

For Hello-SimpleAI, Table 32 gives the AI-positive orientation; for a fixed reversed score the AUROC is $1 - A$, with CI $[1 - U, 1 - L]$ when the displayed CI is $[L, U]$. For example, advanced XSum TL1 has raw AI-positive AUROC 0.0010, 0.9990 after reversing direction; its cleaned AI-positive AUROC is 0.2629, or 0.7371 reversed. Selecting a reversal after seeing labeled blended text is not a zero-shot evaluation. Human-only format swapping yields AUROC 0.999992, supporting the formatting diagnosis rather than an evasion interpretation. Table 28 gives the raw and cleaned TL1 AUROCs of RADAR and the GPT-2 detector. The GPT-2 detector's AUROC increases by at least 0.10 in 17 of the 18 re-scored cells; changes range from 0.0918 to 0.2346. Its cleaned advanced TL1 AUROCs are 0.6686, 0.6307, and 0.7705, still below the corresponding single-model references (Table 3). RADAR is not stable under cleaning: 7 of 18 changes exceed 0.05 in magnitude, with a maximum of 0.0628, and its cleaned TL1 AUROC is 0.7849–0.8811 across the six pool/domain cells. Own-pool comparisons with document-clustered intervals, raw and cleaned, are in Table 3.

Own-pool comparison against data-builder single-model text. Table 3 compares TL1 blends with the same pool's single-model text. For the GPT-2 detector, much of the raw reduction is formatting: cleaning raises its AUROC on blended text by 0.0918 to 0.2346 across the 18 re-scored cells (Appendix C), and the classic-pool reduction falls from 0.1932–0.2205 on raw text to 0.0505–0.1081 on cleaned text. A smaller reduction remains, largest for the advanced pool (0.0960–0.1813).

Table 28: Supervised-detector TL1 results after actual V2 re-scoring, with paired human and generated text cleaned by the same function ($n = 500/300/500$). Calibrated AI-label orientation is fixed before the analysis. Individual AUROC CIs use 2,000 resamples drawn independently within the human and generated classes. Full TL1/random/sentence coverage and within-view single-model references appear in Appendix C.

Detector	Pool	Data	Raw V0	Cleaned V2 [95% CI]
RADAR	Classic	XSum	0.8958	0.8721 [0.8499, 0.8933]
RADAR	Classic	SQuAD	0.8395	0.7849 [0.7489, 0.8191]
RADAR	Classic	Writing	0.8846	0.8219 [0.7966, 0.8464]
RADAR	Advanced	XSum	0.9027	0.8811 [0.8599, 0.9011]
RADAR	Advanced	SQuAD	0.8474	0.7990 [0.7630, 0.8335]
RADAR	Advanced	Writing	0.8844	0.8272 [0.8023, 0.8521]
GPT-2 Detector	Classic	XSum	0.7211	0.8473 [0.8224, 0.8712]
GPT-2 Detector	Classic	SQuAD	0.7208	0.8126 [0.7762, 0.8467]
GPT-2 Detector	Classic	Writing	0.7434	0.8760 [0.8541, 0.8966]
GPT-2 Detector	Advanced	XSum	0.5622	0.6686 [0.6347, 0.7025]
GPT-2 Detector	Advanced	SQuAD	0.5221	0.6307 [0.5874, 0.6793]
GPT-2 Detector	Advanced	Writing	0.6484	0.7705 [0.7397, 0.7992]

RADAR moves the other way. On raw text, advanced-pool blends are detected more readily than advanced single-model text by 0.0799 to 0.1665 AUROC, with intervals excluding zero, while the classic pool differs clearly only on WritingPrompts. RADAR is itself sensitive to NLTK spacing: formatting human text alone yields AUROC 0.846, 0.724, and 0.603 against the original human text (Table 29). On cleaned text, the blend AUROC exceeds the single-model AUROC in all six cells, by 0.0314 to 0.1173, with every interval excluding zero. Against single-model text generated by the same loop, this RADAR difference disappears (Section 5 and Appendix A).

Table 29: All human-only format-swap AUROCs (95% CIs); only Hello uses the reversed diagnostic direction.

Detector	Data	Calibrated AI direction [CI]	Blend-inversion direction [CI]
RADAR	XSum	0.846 [0.818, 0.872]	0.846 [0.818, 0.872]
RADAR	SQuAD	0.724 [0.681, 0.765]	0.724 [0.681, 0.765]
RADAR	Writing	0.603 [0.568, 0.639]	0.603 [0.568, 0.639]
RADAR	Pooled	0.692 [0.671, 0.713]	0.692 [0.671, 0.713]
GPT-2 det.	XSum	0.367 [0.333, 0.403]	0.367 [0.333, 0.403]
GPT-2 det.	SQuAD	0.270 [0.232, 0.311]	0.270 [0.232, 0.311]
GPT-2 det.	Writing	0.312 [0.279, 0.345]	0.312 [0.279, 0.345]
GPT-2 det.	Pooled	0.320 [0.299, 0.340]	0.320 [0.299, 0.340]
Hello	XSum	0.000 [0.000, 0.000]	1.000 [1.000, 1.000]
Hello	SQuAD	0.000 [0.000, 0.000]	1.000 [1.000, 1.000]
Hello	Writing	0.000 [0.000, 0.000]	1.000 [1.000, 1.000]
Hello	Pooled	0.000 [0.000, 0.000]	1.000 [1.000, 1.000]

Table 30: Complete GPT-2 detector results on the 18 scored cells, calibrated AI orientation. V0 is raw; V2 is an actual cleaned-text forward pass. Individual AUROC CIs independently resample each class; differences use prompt-paired resampling (2,000 samples each). TL2–TL5 were not re-scored.

Pool	Data	Setting	V0 [95% CI]	V2 [95% CI]	$V2 - V0$ [95% CI]
Classic	XSum	TL=1	0.7211 [0.6891, 0.7530]	0.8473 [0.8224, 0.8712]	0.1262 [0.1087, 0.1436]
Classic	XSum	Rand.	0.6776 [0.6433, 0.7097]	0.8753 [0.8526, 0.8958]	0.1978 [0.1774, 0.2187]
Classic	XSum	Sent.	0.7025 [0.6697, 0.7347]	0.8961 [0.8755, 0.9143]	0.1936 [0.1713, 0.2167]
Classic	SQuAD	TL=1	0.7208 [0.6777, 0.7620]	0.8126 [0.7762, 0.8467]	0.0918 [0.0756, 0.1086]
Classic	SQuAD	Rand.	0.6991 [0.6564, 0.7427]	0.8872 [0.8613, 0.9134]	0.1882 [0.1637, 0.2137]
Classic	SQuAD	Sent.	0.7107 [0.6699, 0.7494]	0.9093 [0.8852, 0.9323]	0.1986 [0.1719, 0.2266]
Classic	Writing	TL=1	0.7434 [0.7122, 0.7731]	0.8760 [0.8541, 0.8966]	0.1327 [0.1166, 0.1488]
Classic	Writing	Rand.	0.7444 [0.7134, 0.7758]	0.9008 [0.8814, 0.9197]	0.1564 [0.1388, 0.1744]
Classic	Writing	Sent.	0.7437 [0.7123, 0.7734]	0.9141 [0.8955, 0.9312]	0.1705 [0.1513, 0.1901]
Advanced	XSum	TL=1	0.5622 [0.5264, 0.5972]	0.6686 [0.6347, 0.7025]	0.1064 [0.0893, 0.1242]
Advanced	XSum	Rand.	0.5414 [0.5058, 0.5758]	0.7194 [0.6881, 0.7499]	0.1780 [0.1565, 0.2001]
Advanced	XSum	Sent.	0.5520 [0.5176, 0.5871]	0.7540 [0.7244, 0.7824]	0.2019 [0.1784, 0.2254]
Advanced	SQuAD	TL=1	0.5221 [0.4763, 0.5696]	0.6307 [0.5874, 0.6793]	0.1086 [0.0922, 0.1247]
Advanced	SQuAD	Rand.	0.5056 [0.4603, 0.5493]	0.6999 [0.6571, 0.7396]	0.1943 [0.1689, 0.2198]
Advanced	SQuAD	Sent.	0.5368 [0.4893, 0.5825]	0.7714 [0.7328, 0.8082]	0.2346 [0.2094, 0.2604]
Advanced	Writing	TL=1	0.6484 [0.6135, 0.6817]	0.7705 [0.7397, 0.7992]	0.1221 [0.1079, 0.1373]
Advanced	Writing	Rand.	0.6473 [0.6112, 0.6822]	0.8264 [0.7995, 0.8512]	0.1792 [0.1603, 0.1992]
Advanced	Writing	Sent.	0.6230 [0.5886, 0.6573]	0.8108 [0.7828, 0.8367]	0.1878 [0.1678, 0.2085]

Table 31: Complete RADAR results on the 18 scored cells, calibrated AI orientation. V0 is raw; V2 is an actual cleaned-text forward pass. Individual AUROC CIs independently resample each class; differences use prompt-paired resampling (2,000 samples each). TL2–TL5 were not re-scored.

Pool	Data	Setting	V0 [95% CI]	V2 [95% CI]	$V2 - V0$ [95% CI]
Classic	XSum	TL=1	0.8958 [0.8763, 0.9147]	0.8721 [0.8499, 0.8933]	-0.0236 [-0.0299, -0.0172]
Classic	XSum	Rand.	0.8937 [0.8726, 0.9126]	0.8637 [0.8398, 0.8841]	-0.0300 [-0.0401, -0.0202]
Classic	XSum	Sent.	0.8963 [0.8767, 0.9146]	0.8499 [0.8253, 0.8708]	-0.0464 [-0.0608, -0.0342]
Classic	SQuAD	TL=1	0.8395 [0.8078, 0.8690]	0.7849 [0.7489, 0.8191]	-0.0547 [-0.0646, -0.0453]
Classic	SQuAD	Rand.	0.8564 [0.8251, 0.8855]	0.8019 [0.7648, 0.8365]	-0.0546 [-0.0734, -0.0374]
Classic	SQuAD	Sent.	0.8067 [0.7731, 0.8388]	0.7498 [0.7130, 0.7858]	-0.0569 [-0.0763, -0.0376]
Classic	Writing	TL=1	0.8846 [0.8650, 0.9046]	0.8219 [0.7966, 0.8464]	-0.0628 [-0.0726, -0.0535]
Classic	Writing	Rand.	0.8345 [0.8088, 0.8586]	0.8131 [0.7870, 0.8377]	-0.0214 [-0.0311, -0.0118]
Classic	Writing	Sent.	0.7761 [0.7476, 0.8040]	0.7414 [0.7107, 0.7721]	-0.0347 [-0.0458, -0.0234]
Advanced	XSum	TL=1	0.9027 [0.8832, 0.9209]	0.8811 [0.8599, 0.9011]	-0.0216 [-0.0274, -0.0161]
Advanced	XSum	Rand.	0.9307 [0.9138, 0.9461]	0.8867 [0.8661, 0.9054]	-0.0440 [-0.0567, -0.0329]
Advanced	XSum	Sent.	0.9075 [0.8872, 0.9245]	0.8591 [0.8349, 0.8814]	-0.0485 [-0.0624, -0.0356]
Advanced	SQuAD	TL=1	0.8474 [0.8162, 0.8767]	0.7990 [0.7630, 0.8335]	-0.0485 [-0.0581, -0.0394]
Advanced	SQuAD	Rand.	0.8675 [0.8392, 0.8950]	0.8127 [0.7801, 0.8446]	-0.0548 [-0.0735, -0.0373]
Advanced	SQuAD	Sent.	0.8380 [0.8056, 0.8685]	0.7816 [0.7468, 0.8165]	-0.0564 [-0.0788, -0.0361]
Advanced	Writing	TL=1	0.8844 [0.8632, 0.9050]	0.8272 [0.8023, 0.8521]	-0.0572 [-0.0652, -0.0488]
Advanced	Writing	Rand.	0.8420 [0.8170, 0.8665]	0.8239 [0.7985, 0.8487]	-0.0180 [-0.0278, -0.0083]
Advanced	Writing	Sent.	0.8056 [0.7795, 0.8308]	0.7699 [0.7407, 0.7974]	-0.0356 [-0.0474, -0.0235]

Table 32: Complete Hello-SimpleAI RoBERTa results on the 18 scored cells, calibrated AI orientation. V0 is raw; V2 is an actual cleaned-text forward pass. Individual AUROC CIs independently resample each class; differences use prompt-paired resampling (2,000 samples each). TL2–TL5 were not re-scored.

Pool	Data	Setting	V0 [95% CI]	V2 [95% CI]	$V2 - V0$ [95% CI]
Classic	XSum	TL=1	0.0003 [0.0000, 0.0008]	0.3689 [0.3344, 0.4031]	0.3686 [0.3342, 0.4028]
Classic	XSum	Rand.	0.0000 [0.0000, 0.0000]	0.4148 [0.3783, 0.4501]	0.4148 [0.3783, 0.4501]
Classic	XSum	Sent.	0.0000 [0.0000, 0.0000]	0.4526 [0.4144, 0.4891]	0.4526 [0.4144, 0.4891]
Classic	SQuAD	TL=1	0.0068 [0.0006, 0.0158]	0.4911 [0.4466, 0.5387]	0.4842 [0.4391, 0.5322]
Classic	SQuAD	Rand.	0.0000 [0.0000, 0.0000]	0.4815 [0.4369, 0.5293]	0.4815 [0.4369, 0.5293]
Classic	SQuAD	Sent.	0.0000 [0.0000, 0.0000]	0.5731 [0.5282, 0.6215]	0.5731 [0.5282, 0.6215]
Classic	Writing	TL=1	0.0006 [0.0002, 0.0012]	0.5618 [0.5251, 0.5962]	0.5612 [0.5243, 0.5954]
Classic	Writing	Rand.	0.0000 [0.0000, 0.0000]	0.5888 [0.5529, 0.6232]	0.5888 [0.5529, 0.6232]
Classic	Writing	Sent.	0.0000 [0.0000, 0.0000]	0.6487 [0.6141, 0.6836]	0.6487 [0.6141, 0.6836]
Advanced	XSum	TL=1	0.0010 [0.0000, 0.0029]	0.2629 [0.2324, 0.2943]	0.2619 [0.2317, 0.2933]
Advanced	XSum	Rand.	0.0000 [0.0000, 0.0000]	0.3019 [0.2705, 0.3366]	0.3019 [0.2705, 0.3366]
Advanced	XSum	Sent.	0.0000 [0.0000, 0.0000]	0.3810 [0.3445, 0.4168]	0.3810 [0.3445, 0.4168]
Advanced	SQuAD	TL=1	0.0055 [0.0001, 0.0138]	0.4403 [0.3956, 0.4868]	0.4349 [0.3899, 0.4803]
Advanced	SQuAD	Rand.	0.0000 [0.0000, 0.0000]	0.4397 [0.3979, 0.4870]	0.4397 [0.3979, 0.4870]
Advanced	SQuAD	Sent.	0.0000 [0.0000, 0.0000]	0.5022 [0.4542, 0.5498]	0.5022 [0.4542, 0.5498]
Advanced	Writing	TL=1	0.0026 [0.0002, 0.0060]	0.5536 [0.5179, 0.5907]	0.5510 [0.5154, 0.5881]
Advanced	Writing	Rand.	0.0000 [0.0000, 0.0000]	0.5702 [0.5346, 0.6044]	0.5702 [0.5346, 0.6044]
Advanced	Writing	Sent.	0.0000 [0.0000, 0.0000]	0.6002 [0.5645, 0.6378]	0.6002 [0.5645, 0.6378]

Table 33: Single-arm references scored with the cleaning pipeline of this appendix. Classic rows average four generators. The advanced rows use the 100-prompt pilot control, not the full four-model advanced arm; the own-pool comparisons in Table 3 use the full arm as the advanced reference.

Detector	Pool	Data	View	Mean AUROC
RADAR	Classic	XSum	V0	0.8904
RADAR	Advanced	XSum	V0	0.7303
RADAR	Classic	SQuAD	V0	0.8263
RADAR	Advanced	SQuAD	V0	0.6484
RADAR	Classic	Writing	V0	0.7855
RADAR	Advanced	Writing	V0	0.7075
RADAR	Classic	XSum	V2	0.7923
RADAR	Advanced	XSum	V2	0.7348
RADAR	Classic	SQuAD	V2	0.7379
RADAR	Advanced	SQuAD	V2	0.6485
RADAR	Classic	Writing	V2	0.7045
RADAR	Advanced	Writing	V2	0.7174
GPT-2 detector	Classic	XSum	V0	0.9416
GPT-2 detector	Advanced	XSum	V0	0.8412
GPT-2 detector	Classic	SQuAD	V0	0.9335
GPT-2 detector	Advanced	SQuAD	V0	0.8028
GPT-2 detector	Classic	Writing	V0	0.9365
GPT-2 detector	Advanced	Writing	V0	0.8994
GPT-2 detector	Classic	XSum	V2	0.9141
GPT-2 detector	Advanced	XSum	V2	0.8281
GPT-2 detector	Classic	SQuAD	V2	0.9207
GPT-2 detector	Advanced	SQuAD	V2	0.7985
GPT-2 detector	Classic	Writing	V2	0.9265
GPT-2 detector	Advanced	Writing	V2	0.8969

D FORMATTING AND LENGTH SENSITIVITIES

All 42 combinations have cached V0/V1/V1-S/V1-T statistics. The tables below report V0 and the two controlled derivatives; plain V1 values are omitted for length. The largest difference between the cached analytic and the sampled-harness raw AUROC is advanced SQuAD TL1, 0.5397 versus 0.5317, within the 0.01 cache tolerance. V1-S drops marker-overlapping cached words, retaining the original conditioning of all remaining words, while V2 reconditions on an edited string. V1-T uses a common prefix length per pair and is exact as a truncation for a causal scorer. The raw whitespace-word means span 150.8–157.9 for human texts and 155.5–205.3 for blended cells, but raw whitespace counts are inflated by detached punctuation: 32 of 42 cells exceed a 10% difference under that count, versus only four under spacing-independent alphanumeric-word counts (XSum/SQuAD sentence cells). Token-chunk blended/human ratios then range from 0.937 to 1.036, and sentence ratios from 1.084 to 1.203. The choice of length measure is therefore material. Fast-DetectGPT reductions persist after spacing normalization and marker masking (Table 2). Across all 42 cells, spacing normalization changes its analytic AUROC by -0.0334 to -0.0041 relative to raw caches; additional marker masking yields changes of -0.0306 to $+0.0050$ from raw, lower in 39 of 42 cells. Thus the measured Fast-DetectGPT reduction is not removed by these formatting controls. This is narrower than claiming independence from quality, length, or all text-processing choices. The tables below report both views.

Predictive entropy also does not uniformly survive: spacing normalization lowers its AUROC in all 42 cells, by 0.0423 to 0.0677, and the 14 normalized WritingPrompts cells lie between 0.4556 and 0.5384, with 8 of 14 intervals including 0.5; a concurrent submission analyzes these entropy cells in detail (Anonymous, 2027). We therefore describe entropy as domain- and view-dependent. Hello-SimpleAI's near-zero raw AUROC is a formatting-driven inversion: applying NLTK formatting to human text alone gives pooled format-swap AUROC 0.999992. Its two orientations and cleaned results belong in Appendix C, not in an evasion claim.

Table 34: Classic XSum formatting/length sensitivity, analytic cached scores. V1-S and V1-T have distinct definitions and are not V2 rescoring. Entropy uses the signed $+H$ convention; an AUROC below 0.5 is inverted separation.

Data	Setting	Statistic	V0	V1-S	V1-T
XSum	TL1	FDG	0.7077	0.6911	0.6777
XSum	TL1	Likelihood	0.3425	0.3837	0.3811
XSum	TL1	Entropy	0.7967	0.7544	0.7513
XSum	TL2	FDG	0.7019	0.6790	0.6630
XSum	TL2	Likelihood	0.3508	0.3999	0.3924
XSum	TL2	Entropy	0.7804	0.7269	0.7280
XSum	TL3	FDG	0.7175	0.7010	0.6737
XSum	TL3	Likelihood	0.3677	0.4239	0.4107
XSum	TL3	Entropy	0.7714	0.7182	0.7188
XSum	TL4	FDG	0.7297	0.7200	0.6856
XSum	TL4	Likelihood	0.3700	0.4294	0.4085
XSum	TL4	Entropy	0.7710	0.7146	0.7170
XSum	TL5	FDG	0.7263	0.7238	0.6862
XSum	TL5	Likelihood	0.3725	0.4417	0.4134
XSum	TL5	Entropy	0.7642	0.7033	0.7160
XSum	random	FDG	0.7032	0.6988	0.6714
XSum	random	Likelihood	0.3455	0.4130	0.3962
XSum	random	Entropy	0.7792	0.7217	0.7238
XSum	sentence	FDG	0.8069	0.8120	0.7393
XSum	sentence	Likelihood	0.4464	0.5165	0.4494
XSum	sentence	Entropy	0.7440	0.6829	0.7146

Table 35: Classic SQuAD formatting/length sensitivity, analytic cached scores. V1-S and V1-T have distinct definitions and are not V2 rescoring. Entropy uses the signed $+H$ convention; an AUROC below 0.5 is inverted separation.

Data	Setting	Statistic	V0	V1-S	V1-T
SQuAD	TL1	FDG	0.8093	0.8008	0.7962
SQuAD	TL1	Likelihood	0.4366	0.4952	0.4985
SQuAD	TL1	Entropy	0.7467	0.6918	0.6853
SQuAD	TL2	FDG	0.8692	0.8475	0.8374
SQuAD	TL2	Likelihood	0.4716	0.5282	0.5257
SQuAD	TL2	Entropy	0.7553	0.6968	0.6949
SQuAD	TL3	FDG	0.8624	0.8427	0.8289
SQuAD	TL3	Likelihood	0.4977	0.5510	0.5501
SQuAD	TL3	Entropy	0.7216	0.6656	0.6573
SQuAD	TL4	FDG	0.8612	0.8441	0.8346
SQuAD	TL4	Likelihood	0.4861	0.5429	0.5410
SQuAD	TL4	Entropy	0.7334	0.6743	0.6707
SQuAD	TL5	FDG	0.8741	0.8550	0.8370
SQuAD	TL5	Likelihood	0.5131	0.5661	0.5645
SQuAD	TL5	Entropy	0.7210	0.6691	0.6576
SQuAD	random	FDG	0.8336	0.8219	0.8104
SQuAD	random	Likelihood	0.4635	0.5277	0.5273
SQuAD	random	Entropy	0.7442	0.6835	0.6746
SQuAD	sentence	FDG	0.9093	0.8951	0.8677
SQuAD	sentence	Likelihood	0.5495	0.6104	0.5713
SQuAD	sentence	Entropy	0.7150	0.6535	0.6775

Table 36: Classic Writing formatting/length sensitivity, analytic cached scores. V1-S and V1-T have distinct definitions and are not V2 rescoring. Entropy uses the signed $+H$ convention; an AUROC below 0.5 is inverted separation.

Data	Setting	Statistic	V0	V1-S	V1-T
Writing	TL1	FDG	0.9139	0.9006	0.8990
Writing	TL1	Likelihood	0.7035	0.7416	0.7447
Writing	TL1	Entropy	0.5705	0.5078	0.5002
Writing	TL2	FDG	0.9469	0.9295	0.9242
Writing	TL2	Likelihood	0.7297	0.7443	0.7431
Writing	TL2	Entropy	0.5795	0.5318	0.5256
Writing	TL3	FDG	0.9394	0.9201	0.9131
Writing	TL3	Likelihood	0.7180	0.7375	0.7348
Writing	TL3	Entropy	0.5841	0.5350	0.5284
Writing	TL4	FDG	0.9457	0.9281	0.9207
Writing	TL4	Likelihood	0.7447	0.7668	0.7614
Writing	TL4	Entropy	0.5575	0.5030	0.4995
Writing	TL5	FDG	0.9496	0.9337	0.9278
Writing	TL5	Likelihood	0.7408	0.7627	0.7586
Writing	TL5	Entropy	0.5709	0.5219	0.5158
Writing	random	FDG	0.9394	0.9221	0.9136
Writing	random	Likelihood	0.7124	0.7465	0.7417
Writing	random	Entropy	0.5883	0.5332	0.5280
Writing	sentence	FDG	0.9600	0.9475	0.9364
Writing	sentence	Likelihood	0.7871	0.8101	0.7802
Writing	sentence	Entropy	0.5134	0.4635	0.4904

Table 37: Advanced XSum formatting/length sensitivity, analytic cached scores. V1-S and V1-T have distinct definitions and are not V2 rescoring. Entropy uses the signed $+H$ convention; an AUROC below 0.5 is inverted separation.

Data	Setting	Statistic	V0	V1-S	V1-T
XSum	TL1	FDG	0.3959	0.3703	0.3718
XSum	TL1	Likelihood	0.1927	0.2229	0.2247
XSum	TL1	Entropy	0.8125	0.7628	0.7608
XSum	TL2	FDG	0.4476	0.4291	0.4172
XSum	TL2	Likelihood	0.2164	0.2601	0.2542
XSum	TL2	Entropy	0.8007	0.7441	0.7448
XSum	TL3	FDG	0.4206	0.4148	0.3997
XSum	TL3	Likelihood	0.2023	0.2506	0.2412
XSum	TL3	Entropy	0.8066	0.7487	0.7509
XSum	TL4	FDG	0.4560	0.4415	0.4276
XSum	TL4	Likelihood	0.2469	0.2937	0.2802
XSum	TL4	Entropy	0.7736	0.7164	0.7256
XSum	TL5	FDG	0.4475	0.4378	0.4109
XSum	TL5	Likelihood	0.2375	0.2809	0.2656
XSum	TL5	Entropy	0.7782	0.7263	0.7296
XSum	random	FDG	0.4262	0.4168	0.4047
XSum	random	Likelihood	0.2093	0.2548	0.2472
XSum	random	Entropy	0.8036	0.7460	0.7501
XSum	sentence	FDG	0.5358	0.5373	0.4773
XSum	sentence	Likelihood	0.3003	0.3595	0.3115
XSum	sentence	Entropy	0.7494	0.6880	0.7154

Table 38: Advanced SQuAD formatting/length sensitivity, analytic cached scores. V1-S and V1-T have distinct definitions and are not V2 rescoring. Entropy uses the signed $+H$ convention; an AUROC below 0.5 is inverted separation.

Data	Setting	Statistic	V0	V1-S	V1-T
SQuAD	TL1	FDG	0.5397	0.5248	0.5213
SQuAD	TL1	Likelihood	0.2992	0.3520	0.3585
SQuAD	TL1	Entropy	0.7364	0.6764	0.6671
SQuAD	TL2	FDG	0.5864	0.5656	0.5583
SQuAD	TL2	Likelihood	0.3458	0.3980	0.3996
SQuAD	TL2	Entropy	0.7112	0.6505	0.6461
SQuAD	TL3	FDG	0.5854	0.5670	0.5571
SQuAD	TL3	Likelihood	0.3637	0.4200	0.4190
SQuAD	TL3	Entropy	0.6997	0.6323	0.6292
SQuAD	TL4	FDG	0.6230	0.6113	0.5965
SQuAD	TL4	Likelihood	0.3849	0.4438	0.4397
SQuAD	TL4	Entropy	0.6969	0.6322	0.6301
SQuAD	TL5	FDG	0.6034	0.5841	0.5796
SQuAD	TL5	Likelihood	0.3758	0.4293	0.4271
SQuAD	TL5	Entropy	0.6973	0.6330	0.6314
SQuAD	random	FDG	0.5714	0.5763	0.5624
SQuAD	random	Likelihood	0.3526	0.4206	0.4176
SQuAD	random	Entropy	0.6994	0.6341	0.6298
SQuAD	sentence	FDG	0.7074	0.6997	0.6625
SQuAD	sentence	Likelihood	0.4801	0.5470	0.5020
SQuAD	sentence	Entropy	0.6289	0.5581	0.5928

Table 39: Advanced Writing formatting/length sensitivity, analytic cached scores. V1-S and V1-T have distinct definitions and are not V2 rescoring. Entropy uses the signed $+H$ convention; an AUROC below 0.5 is inverted separation.

Data	Setting	Statistic	V0	V1-S	V1-T
Writing	TL1	FDG	0.8049	0.7834	0.7784
Writing	TL1	Likelihood	0.5851	0.6271	0.6317
Writing	TL1	Entropy	0.6061	0.5384	0.5298
Writing	TL2	FDG	0.8288	0.7982	0.7929
Writing	TL2	Likelihood	0.6347	0.6554	0.6565
Writing	TL2	Entropy	0.5622	0.5145	0.5056
Writing	TL3	FDG	0.8423	0.8127	0.8089
Writing	TL3	Likelihood	0.6465	0.6648	0.6670
Writing	TL3	Entropy	0.5578	0.5111	0.5061
Writing	TL4	FDG	0.8600	0.8386	0.8336
Writing	TL4	Likelihood	0.6549	0.6850	0.6823
Writing	TL4	Entropy	0.5622	0.5078	0.5079
Writing	TL5	FDG	0.8485	0.8342	0.8245
Writing	TL5	Likelihood	0.6530	0.6832	0.6726
Writing	TL5	Entropy	0.5535	0.5041	0.5080
Writing	random	FDG	0.8356	0.8246	0.8122
Writing	random	Likelihood	0.6360	0.6769	0.6673
Writing	random	Entropy	0.5767	0.5190	0.5191
Writing	sentence	FDG	0.8856	0.8727	0.8513
Writing	sentence	Likelihood	0.7124	0.7461	0.7073
Writing	sentence	Entropy	0.5100	0.4547	0.4851

E Single-Model Controls and Simple Attacks

We compare TL1 against T5 paraphrasing (the public `Vamsi/T5_Paraphrase_Paws` checkpoint (Raffel et al., 2020)) and adjacent-word swaps applied to classic single-model outputs, using a mean over GPT-2-XL, OPT, and GPT-J to avoid including the scorer's own generator. For the *advanced* pool, the bootstrap CI of attack minus ToBlend signed AUROC lies above zero for both attacks in all three domains. Under absolute separation from chance, the advantage holds in SQuAD and WritingPrompts but not against word swaps on XSum (-0.0261 [-0.0784, 0.0233]), where the advanced blend is inverted. The attacks alter classic-pool text only, so these comparisons are not generator-matched. The classic pool is not stronger than word swaps on XSum or SQuAD, or than paraphrasing on SQuAD or WritingPrompts; the XSum paraphrase and WritingPrompts word-swap contrasts are not separated by their CIs. Likelihood-family results also contain exceptions, including SQuAD rank where word swapping is stronger than classic blending. The comparisons are given below.

Table 40: Advanced single-model references using the data-builder protocol, all scored by GPT-Neo-2.7B; $n = 500/300/500$ per generator. TPR is an in-sample operating point; held-out threshold estimates elsewhere use an explicit split.

Data	Generator	AUROC [95% CI]	TPR at 5% FPR
XSum	llama-2-7b	0.6471 [0.6171, 0.6770]	0.0940
XSum	phi-2	0.8121 [0.7872, 0.8372]	0.3620
XSum	mistral-7b-v0.3	0.7210 [0.6922, 0.7490]	0.1980
XSum	gemma-7b	0.7778 [0.7521, 0.8034]	0.2820
SQuAD	llama-2-7b	0.7437 [0.7091, 0.7787]	0.3167
SQuAD	phi-2	0.9112 [0.8908, 0.9319]	0.6600
SQuAD	mistral-7b-v0.3	0.7702 [0.7366, 0.8031]	0.3500
SQuAD	gemma-7b	0.8171 [0.7859, 0.8466]	0.4233
Writing	llama-2-7b	0.8849 [0.8644, 0.9035]	0.6100
Writing	phi-2	0.9567 [0.9459, 0.9664]	0.7820
Writing	mistral-7b-v0.3	0.9135 [0.8960, 0.9295]	0.6720
Writing	gemma-7b	0.9316 [0.9175, 0.9451]	0.6780

The simple attacks alter classic single-model generations only; the advanced ToBlend arm uses a different generator pool, and attack pairs are trimmed while ToBlend pairs are not. Common human-score equality establishes the GPT-Neo scorer for the likelihood family. For the classic-pool Fast-DetectGPT comparison, attack-minus-TL1 AUROC differences are: XSum word swap -0.1317 $[-0.1583, -0.1047]$ and paraphrase -0.0158 $[-0.0406, 0.0100]$; SQuAD word swap -0.1094 $[-0.1405, -0.0790]$ and paraphrase -0.0337 $[-0.0632, -0.0050]$; Writing word swap -0.0135 $[-0.0288, 0.0020]$ and paraphrase -0.0509 $[-0.0678, -0.0349]$. Negative values favor the simple attack in the declared direction. The advanced TL1 comparison has positive intervals for all six corresponding comparisons; these do not establish generator-matched causal superiority.

The orientation audit illustrates why signed AUROC alone is insufficient. Advanced XSum TL1 has raw AUROC 0.3968; fitting a reversed direction on a disjoint 20% split yields held-out AUROC 0.6170 [0.5815, 0.6498]. Under the original direction its held-out TPR is 0.0486 at a realized FPR of 0.0657. This is evidence of score inversion and weak original-direction detection, not information erasure.

Table 41: Buffered TL1 orientation and operating-point audit for Fast-DetectGPT. Raw AUROC uses all prompts. Oriented AUROC fixes direction on 20% of labeled prompt pairs and evaluates the remaining 80%, with prompt-paired bootstrap CIs; a reversed direction requires labeled blended data. Raw TPR uses the original direction and a target 5% FPR threshold fitted on 30% of human prompts, then evaluated on the remaining 70%; the realized FPR need not equal 5%. These are different splits and estimands, not directly paired AUROC improvements.

Pool	Data	Raw AUROC	Oriented held-out AUROC [95% CI]	Direction	Raw TPR	Realized FPR
Classic	XSum	0.7088	0.7114 [0.6783, 0.7422]	original	0.2257	0.0657
Classic	SQuAD	0.8047	0.8030 [0.7677, 0.8386]	original	0.3190	0.0476
Classic	Writing	0.9128	0.9091 [0.8883, 0.9286]	original	0.6800	0.0571
Advanced	XSum	0.3968	0.6170 [0.5815, 0.6498]	reversed	0.0486	0.0657
Advanced	SQuAD	0.5317	0.5345 [0.4909, 0.5836]	original	0.0714	0.0476
Advanced	Writing	0.8036	0.8027 [0.7743, 0.8288]	original	0.4229	0.0571

F Pool Comparison Against Data-Builder Single-Model Text

This section gives the pool comparison against data-builder single-model text computed from the saved scores; Section 6 compares it with the pool-of-one references.

Let $\Delta_p = A_{\text{single},p} - A_{\text{blend},p}$ for pool p under the same scorer. The primary TL1 comparison is $\Delta_{\text{advanced}} - \Delta_{\text{classic}}$ for Fast-DetectGPT, averaged equally over domains. On signed AUROC,

the advanced single-model mean is 0.8239; the classic and advanced reductions are 0.1403 [0.1266, 0.1544] and 0.2465 [0.2291, 0.2635], and their difference is 0.1062 [0.0869, 0.1266].

This signed difference is mostly score inversion. Advanced XSum TL1 has AUROC 0.3968, below chance; with the direction fitted on 20% of prompts, its held-out AUROC is 0.6170 (Appendix E). Replacing every AUROC by $D(A)$ gives a difference of 0.0374 [0.0105, 0.0622], and held-out orientation gives 0.0321 [0.0047, 0.0604] (Table 42). The XSum term changes sign (-0.0832 under $D(A)$); SQuAD (0.1361) and WritingPrompts (0.0594) are unchanged because no blend there is inverted. After truncating every text to 100 words, the signed difference is 0.0591 [0.0399, 0.0788] and the $D(A)$ difference is -0.0367 [-0.0707, -0.0088]. Excluding GPT-Neo, which is also the scorer, from the classic single-model reference gives $D(A)$ differences of 0.0524 [0.0258, 0.0769] at full length and -0.0155 [-0.0483, 0.0116] after length matching; the classic blends still contain GPT-Neo chunks. Pipeline and length effects also cancel in a difference-in-differences only if they are pool-independent: the data-builder stop uses model tokens, SentencePiece and GPT-2 BPE yield different word counts, and the data builder trims human partners whereas ToBlend does not. The evidence supports a larger signed reduction for the advanced pool, driven by inversion on XSum, but not a larger loss of detectability, and not a claim that stronger future models strengthen blending.

Table 42: Pool difference-in-differences against data-builder singles (advanced minus classic own-pool reduction), Fast-DetectGPT at TL1, under three estimands, two classic reference sets and two lengths. Per-domain columns are point estimates; the mean has a 95% document-clustered CI (2,000 replicates).

Length	Estimand	Classic ref.	XSum	SQuAD	Writing	Mean DiD [CI]
Full length	Signed AUROC	Four	0.1232	0.1361	0.0594	0.1062 [0.0869, 0.1266]
Full length	Signed AUROC	No GPT-Neo	0.1432	0.1521	0.0683	0.1212 [0.1017, 0.1415]
Full length	$D(A) = \max(A, 1-A)$	Four	-0.0832	0.1361	0.0594	0.0374 [0.0105, 0.0622]
Full length	$D(A) = \max(A, 1-A)$	No GPT-Neo	-0.0632	0.1521	0.0683	0.0524 [0.0258, 0.0769]
Full length	Held-out oriented	Four	-0.1006	0.1403	0.0566	0.0321 [0.0047, 0.0604]
Full length	Held-out oriented	No GPT-Neo	-0.0799	0.1573	0.0652	0.0475 [0.0206, 0.0755]
100 words	Signed AUROC	Four	0.0483	0.0908	0.0383	0.0591 [0.0399, 0.0788]
100 words	Signed AUROC	No GPT-Neo	0.0717	0.1156	0.0536	0.0803 [0.0611, 0.0999]
100 words	$D(A) = \max(A, 1-A)$	Four	-0.1978	0.0494	0.0383	-0.0367 [-0.0707, -0.0088]
100 words	$D(A) = \max(A, 1-A)$	No GPT-Neo	-0.1744	0.0742	0.0536	-0.0155 [-0.0483, 0.0116]
100 words	Held-out oriented	Four	-0.2202	0.0902	0.0319	-0.0327 [-0.0601, -0.0049]
100 words	Held-out oriented	No GPT-Neo	-0.1966	0.1158	0.0470	-0.0113 [-0.0384, 0.0160]

G Full-Sample Perturbation Detectors

The full-sample results cover all 21 settings in each pool for DetectGPT and DetectLLM LRR/NPR, with $n = 500/300/500$ per cell and 100 T5-3B (Raffel et al., 2020) perturbations per passage for DetectGPT and NPR. A 10%-subset run of the same detectors serves only as a sample-size sensitivity. The DetectGPT/NPR *single-model* references remain 10% subsets ($n = 50/30/50$), apart from one full XSum GPT-2-XL reference; LRR has a full-sample same-scorer baseline. These reused reference columns are means over the four *classic* generators in both pool tables; an advanced-pool DetectGPT/LRR/NPR single-model reference was not computed here. As a result, apparent baseline-drop comparisons are not all matched. Across settings, correlations with Fast-DetectGPT are positive but vary (Spearman $\rho = 0.393$–0.964); a uniform claim of larger drops is not supported. LRR is above DetectGPT in 11 of 21 classic and 15 of 21 advanced cells, not in every cell.

Table 43: Results for the classic LLMs set on the full datasets: AUROC of DetectGPT (Mitchell et al., 2023) and DetectLLM (LRR and NPR) (Su et al., 2023) under each TOBLEND setting, all scored by GPT-Neo-2.7B; DetectGPT and NPR use 100 T5-3B perturbations per passage (n = 500 / 300 / 500 prompts for XSum / SQuAD / Writing). Baselines marked * are 10% subsets of the single-model runs (mean of four classic generators, n = 50 / 30 / 50); the LRR baseline is the full-n same-scorer mean of the four classic generators. No full-n matched DetectGPT/NPR baseline exists: it would need 1,040,000 T5-3B perturbations. The italic row is the only full-n single-model DetectGPT/NPR run. In each row, the setting with the largest drop below the row's Baseline AUROC is in bold; no cell is bold when every setting is at or above the Baseline, and the Baseline cell is never bold. Candidate models are GPT-2, OPT, GPT-Neo, and GPT-J.

Datasets	Detection Method	Baseline	TL=1	TL=2	TL=3	TL=4	TL=5	Rand.	Sent.
XSum	DetectGPT	0.9079*	**0.7461**	0.8029	0.8268	0.8514	0.8276	0.8089	0.8603
	LRR	0.8219	0.6390	**0.6235**	0.6288	0.6304	0.6268	0.6286	0.6863
	NPR	0.9067*	**0.7940**	0.8263	0.8575	0.8696	0.8489	0.8376	0.8591
	Full-n single-model reference, GPT-2-XL (n = 500): DetectGPT 0.9116, LRR 0.7841, NPR 0.9059								
SQuAD	DetectGPT	0.8625*	**0.5982**	0.7455	0.7384	0.7177	0.7170	0.6866	0.7942
	LRR	0.8843	**0.7646**	0.7732	0.7839	0.7816	0.8020	0.7657	0.8132
	NPR	0.8897*	**0.6882**	0.7797	0.7983	0.7775	0.7774	0.7389	0.8029
Writing	DetectGPT	0.9202*	**0.7591**	0.8694	0.8774	0.8706	0.8832	0.8510	0.8786
	LRR	0.9397	**0.8517**	0.8650	0.8619	0.8718	0.8706	0.8553	0.8926
	NPR	0.9463*	**0.8320**	0.9092	0.9094	0.9041	0.9142	0.8907	0.8945

Table 44: Results for the advanced LLMs set on the full datasets: AUROC of DetectGPT (Mitchell et al., 2023) and DetectLLM (LRR and NPR) (Su et al., 2023) under each TOBLEND setting, all scored by GPT-Neo-2.7B; DetectGPT and NPR use 100 T5-3B perturbations per passage (n = 500 / 300 / 500 prompts for XSum / SQuAD / Writing). Baselines marked * are 10% subsets of the single-model runs (mean of four classic generators, n = 50 / 30 / 50); the LRR baseline is the full-n same-scorer mean of the four classic generators. No full-n matched DetectGPT/NPR baseline exists: it would need 1,040,000 T5-3B perturbations. The italic row is the only full-n single-model DetectGPT/NPR run. In each row, the setting with the largest drop below the row's Baseline AUROC is in bold; no cell is bold when every setting is at or above the Baseline, and the Baseline cell is never bold. Candidate models are Llama2, Phi-2, Mistral, and Gemma.

Datasets	Detection Method	Baseline	TL=1	TL=2	TL=3	TL=4	TL=5	Rand.	Sent.
XSum	DetectGPT	0.9079*	**0.4082**	0.5470	0.5633	0.5772	0.5788	0.5517	0.5916
	LRR	0.8219	0.4479	0.4444	0.4497	0.4723	0.4810	**0.4363**	0.4786
	NPR	0.9067*	**0.4893**	0.5867	0.6187	0.6144	0.6379	0.5996	0.6044
	Full-n single-model reference, GPT-2-XL (n = 500): DetectGPT 0.9116, LRR 0.7841, NPR 0.9059								
SQuAD	DetectGPT	0.8625*	**0.3412**	0.4467	0.5118	0.5003	0.4671	0.4448	0.5570
	LRR	0.8843	**0.5892**	0.5944	0.6280	0.6594	0.6468	0.6134	0.6841
	NPR	0.8897*	**0.4198**	0.4991	0.5856	0.5780	0.5323	0.5092	0.5678
Writing	DetectGPT	0.9202*	**0.5048**	0.6886	0.7158	0.7076	0.7120	0.6823	0.7392
	LRR	0.9397	**0.7472**	0.7670	0.7782	0.7818	0.7782	0.7790	0.8055
	NPR	0.9463*	**0.5918**	0.7342	0.7681	0.7564	0.7574	0.7340	0.7654

Table 45: 95% confidence intervals for Table 43 (percentile bootstrap, 2,000 resamples of prompts, each keeping its human and machine passage; a baseline interval resamples prompts within each generator's run and averages the four AUROCs).

Datasets	Method	Baseline	TL=1	TL=2	TL=3	TL=4	TL=5	Rand.	Sent.
XSum	DetectGPT	[0.883, 0.934]	[0.718, 0.773]	[0.780, 0.826]	[0.804, 0.848]	[0.829, 0.873]	[0.805, 0.848]	[0.785, 0.832]	[0.840, 0.878]
	LRR	[0.811, 0.833]	[0.612, 0.665]	[0.592, 0.652]	[0.599, 0.656]	[0.603, 0.659]	[0.597, 0.655]	[0.599, 0.656]	[0.658, 0.713]
	NPR	[0.882, 0.931]	[0.769, 0.818]	[0.805, 0.848]	[0.836, 0.878]	[0.850, 0.889]	[0.828, 0.869]	[0.816, 0.859]	[0.839, 0.879]
SQuAD	DetectGPT	[0.817, 0.906]	[0.558, 0.637]	[0.711, 0.779]	[0.701, 0.775]	[0.680, 0.755]	[0.680, 0.751]	[0.648, 0.722]	[0.759, 0.826]
	LRR	[0.872, 0.896]	[0.732, 0.795]	[0.739, 0.805]	[0.752, 0.816]	[0.751, 0.811]	[0.772, 0.832]	[0.733, 0.797]	[0.783, 0.842]
	NPR	[0.848, 0.931]	[0.648, 0.726]	[0.746, 0.811]	[0.763, 0.829]	[0.741, 0.812]	[0.743, 0.808]	[0.702, 0.774]	[0.768, 0.834]
Writing	DetectGPT	[0.892, 0.944]	[0.732, 0.787]	[0.851, 0.889]	[0.858, 0.896]	[0.850, 0.890]	[0.865, 0.902]	[0.830, 0.872]	[0.859, 0.897]
	LRR	[0.933, 0.946]	[0.830, 0.873]	[0.844, 0.884]	[0.842, 0.882]	[0.852, 0.891]	[0.850, 0.890]	[0.835, 0.876]	[0.875, 0.911]
	NPR	[0.923, 0.966]	[0.808, 0.855]	[0.893, 0.925]	[0.893, 0.925]	[0.885, 0.921]	[0.898, 0.929]	[0.873, 0.909]	[0.876, 0.911]

Table 46: 95% confidence intervals for Table 44 (percentile bootstrap, 2,000 resamples of prompts, each keeping its human and machine passage; a baseline interval resamples prompts within each generator's run and averages the four AUROCs).

Datasets	Method	Baseline	TL=1	TL=2	TL=3	TL=4	TL=5	Rand.	Sent.
XSum	DetectGPT	[0.883, 0.934]	[0.376, 0.439]	[0.514, 0.578]	[0.530, 0.596]	[0.545, 0.610]	[0.549, 0.608]	[0.522, 0.582]	[0.561, 0.621]
	LRR	[0.811, 0.833]	[0.418, 0.477]	[0.415, 0.473]	[0.422, 0.477]	[0.441, 0.504]	[0.450, 0.511]	[0.405, 0.466]	[0.447, 0.508]
	NPR	[0.882, 0.931]	[0.456, 0.521]	[0.554, 0.618]	[0.587, 0.649]	[0.583, 0.647]	[0.610, 0.666]	[0.569, 0.631]	[0.574, 0.634]
SQuAD	DetectGPT	[0.817, 0.906]	[0.302, 0.380]	[0.406, 0.484]	[0.469, 0.551]	[0.459, 0.543]	[0.426, 0.507]	[0.404, 0.484]	[0.516, 0.595]
	LRR	[0.872, 0.896]	[0.551, 0.625]	[0.559, 0.629]	[0.591, 0.665]	[0.623, 0.694]	[0.611, 0.683]	[0.577, 0.649]	[0.647, 0.717]
	NPR	[0.848, 0.931]	[0.378, 0.462]	[0.458, 0.537]	[0.542, 0.628]	[0.538, 0.617]	[0.493, 0.572]	[0.469, 0.551]	[0.525, 0.606]
Writing	DetectGPT	[0.892, 0.944]	[0.472, 0.536]	[0.659, 0.719]	[0.687, 0.744]	[0.676, 0.736]	[0.684, 0.740]	[0.654, 0.712]	[0.713, 0.765]
	LRR	[0.933, 0.946]	[0.720, 0.774]	[0.742, 0.792]	[0.751, 0.805]	[0.756, 0.807]	[0.752, 0.804]	[0.753, 0.805]	[0.782, 0.830]
	NPR	[0.923, 0.966]	[0.559, 0.622]	[0.707, 0.762]	[0.741, 0.794]	[0.728, 0.784]	[0.729, 0.783]	[0.705, 0.761]	[0.740, 0.791]

H Annotation Audit and Automatic Quality

Three raters scored coherence and fluency from 1 to 7 on five instances per dataset. The raters were hired, experienced human experts. We disclosed neither the source of each text nor the generation settings before annotation. The annotation took 14 working hours, and the raters were paid 280 US dollars in total. The full instructions appear in Table 48. The human reference column contains the first 200 NLTK tokens of each human source passage, which is the whole passage for 12 of the 15 rated passages; token-level blended passages contain 170–180 NLTK tokens and sentence-level passages 170–268. Round 2 contains 270 displayed condition cells but 240 unique texts after shared baselines are deduplicated, including 210 ToBlend texts. Round 1 contains 240 rated text presentations for the no-buffer study. Round-2 scores appear in Table 47.

Table 47: Round-2 mean coherence/fluency (1–7): five instances per dataset, three raters. Human is the first 200 NLTK tokens of each human source passage. The seven ToBlend conditions yield 210 unique texts after deduplicating repeated baselines; the rating-sheet audit is in Appendix L.

Data	Pool	GPT-2	Human	TL=1	TL=2	TL=3	TL=4	TL=5	Rand.	Sent.
XSum	Classic	5.1/5.0	6.3/6.7	4.3/4.3	3.9/4.3	3.9/4.1	2.8/2.9	5.1/5.5	5.7/6.0	4.7/4.9
	Advanced	5.1/5.0	6.3/6.7	4.1/4.4	4.1/4.6	4.2/4.5	4.1/4.3	4.8/5.1	5.6/6.1	4.2/4.3
SQuAD	Classic	5.4/5.4	6.1/6.5	4.3/4.4	4.1/4.3	4.1/4.2	4.7/4.7	3.9/4.3	5.3/5.7	4.1/4.9
	Advanced	5.4/5.4	6.1/6.5	4.6/4.8	4.7/4.9	4.1/4.4	4.3/4.5	4.6/4.9	5.1/5.2	4.8/4.9
Writing	Classic	3.8/4.2	5.7/6.1	4.5/4.7	3.9/4.1	4.3/4.7	4.1/4.4	4.6/4.9	5.4/5.5	4.6/4.9
	Advanced	3.8/4.2	5.7/6.1	4.3/4.5	4.5/4.6	4.9/5.0	4.3/5.0	4.2/4.5	4.7/4.9	4.3/4.7

Of 42 pool/domain/setting cells, 23 fall below GPT-2 on both rating metrics, 16 meet or exceed GPT-2 on both, and three have mixed directions. None reaches the human reference on either metric. There is therefore no general comparable-quality result. Round-2 ordinal Krippendorff α (Krippendorff, 2019) is 0.502/0.508 for coherence/fluency; ICC(2,1) (Shrout & Fleiss, 1979) is 0.556/0.563 and ICC(2,3) is 0.790/0.794. Agreement is lower in round 1 ($\alpha = 0.276/0.423$); repeated-rating and round effects are reported in Appendix H. Five examples per cell do not support precise population estimates or an equivalence claim.

Round 1's GPT-2 text was rated separately in each pool half, so the two baseline ratings must not be silently merged. Round-1 ICC(2,1) is 0.289/0.450 and ICC(2,3) is 0.549/0.710 for coherence/fluency (239 complete triples among 240 text presentations); round-2 agreement is given in the main text. Restricting round 2 to ToBlend yields ordinal $\alpha = 0.416/0.427$, showing that inclusion of the human ceiling affects agreement summaries.

The identical-GPT-2-text round effect depends on the round-1 convention: coherence rises by 0.133 ($p = 0.575$, advanced half), 0.467 ($p = 0.036$, classic half), or 0.300 ($p = 0.102$, mean of halves), each with 15 paired items. Fluency differences are 0.044, 0.000, and 0.022 ($p = 0.609, 0.875, 0.752$). The sentence-level repeated-text sensitivity has coherence difference 0.794 ($p = 0.083$) and fluency difference 0.460 ($p = 0.259$), with 21 pairs. Thus no single favorable convention identifies a buffer-caused quality improvement.

The 90 best-pick votes come only from the round-1 no-buffer study: GPT-2 received 23/90 votes (25.6%, against chance $1/8$ among eight choices), split 16/6/1 across raters. The other seven conditions received 8/6/8/15/7/11/12 votes. Round 2 collected no best-pick judgments. These votes cannot be attributed to the buffered study, or treated as independent proof of human equivalence. At five paired examples per cell, the minimum two-sided Wilcoxon p is 0.0625.

H.1 Original annotation task and round-1 scores

Coherence concerns topic consistency and the logical sequence of a passage (Barzilay & Lapata, 2008); fluency concerns natural phrasing and syntax (Kann et al., 2018), which shapes user trust (Martindale & Carpuat, 2018). These dimensions describe the rating task, rather than factual correctness. Table 48 gives the eight-candidate instructions. Its best-pick task belongs to round 1; it must not be attributed to the nine-condition round-2 rating sheet.

Table 48: Original eight-candidate annotation instructions, retained as a record of the round-1 task. The instructions' "around 200 words" description and "200 tokens" cutoff are reproduced as written; they do not override the measured token lengths or the actual generation stop rule.

Instructions: for each instance, you should:

1. Score from 1 to 7 (1 means poor, 7 means excellent) on the coherence score and fluency score for the column of 'a', 'b', 'c', 'd', 'e', 'f', 'g', 'h'.
* Coherence: The contextual information of the given short text (around 200 words) should logically make sense, i.e., maintaining topic consistency and logical sequence.
* Fluency: The text should read naturally, mimicking the style and syntax of human natural language.

2. From the column of 'a'–'h', select the best one that you think has the highest probability of being written by a real human and put the number in the column of 'best' (normally, the best one should be the one that possesses the highest coherence and fluency scores you annotated in the previous step).

Notes:
1. We have 6 files and 5 instances each. Completing all labeling for one instance should take less than 6 minutes; feel free to take more time if necessary.
2. All texts selected are only a slice of 200 tokens of their original source, so please do not consider a potentially incomplete sentence at the end of the text as one of your scoring criteria.

Table 49: Audited round-1 unbuffered coherence/fluency ratings (1–7, five instances per cell, three raters). GPT-2 was rated separately in each pool half. Primary means retain the two stored off-scale zero ratings, matching the archived averaged columns; exclusion is a sensitivity, not a silent data edit. This round collected the best-pick votes; round 2 did not.

Dataset	Pool	GPT-2	TL1	TL2	TL3	TL4	TL5	Rand.	Sent.
XSum	Classic	4.3/4.8	3.9/3.4	3.4/3.6	4.0/4.2	2.9/3.3	3.7/4.2	3.5/3.4	3.4/4.0
XSum	Advanced	4.3/4.7	2.0/1.5	3.9/3.4	3.3/3.1	3.7/3.6	4.2/3.7	3.6/3.7	3.9/4.3
SQuAD	Classic	5.2/5.7	4.0/3.8	5.1/5.3	4.1/4.9	4.1/4.6	4.3/4.9	5.4/5.5	4.7/5.6
SQuAD	Advanced	5.3/5.7	3.1/3.5	3.5/3.9	4.3/4.9	4.7/5.0	3.7/4.3	4.8/4.8	4.7/5.6
Writing	Classic	3.4/4.1	2.8/3.2	2.4/2.9	2.9/2.9	3.4/3.9	2.8/2.7	2.8/2.5	2.8/3.4
Writing	Advanced	4.3/4.0	3.4/3.4	4.2/4.0	3.4/3.1	3.3/3.9	2.2/2.5	3.3/2.8	3.2/3.2

H.2 Automatic metrics

Automatic quality metrics also do not provide a uniform favorable conclusion. Across 42 cells, the signed cross-entropy comparison with GPT-2 agrees with the average rating direction, but disagrees for WritingPrompts; the continuation comparison with human originals disagrees with both pooled rating directions. Distinct-n, marker, and detached-punctuation diagnostics contain further disagreements. Cross-entropy under one scorer is a proxy, not a human quality judgment, and measuring absolute rather than signed distance changes some conclusions. We report these disagreements instead of selecting the favorable metric (Appendix H.2).

The automatic battery includes scorer cross-entropy, distinct 1–3-grams, leaked-marker incidence, and detached punctuation across all 42 cells. Averaged rating differences from GPT-2 are -0.3238 coherence and -0.1619 fluency. Whole-passage ToBlend cross-entropy exceeds the GPT-2 reference by 0.2845 on average, agreeing with those aggregate directions. Within Writing, however, ratings favor ToBlend while the signed cross-entropy proxy disfavors it. Against human continuations, signed cross-entropy favors ToBlend on average by 0.0666, while ratings are lower by 1.5905/1.7175. This disagreement also appears in SQuAD and Writing separately. Reading cross-entropy as absolute distance from human text changes several comparisons and is retained as a sensitivity, not selected post hoc as the preferred quality definition. The full automatic table records both the proxy and surface diagnostics.

Table 50: Automatic quality proxies at full n. ΔCE: mean per-token cross-entropy under gpt-neo-2.7B of the machine text minus that of the human text for the same prompt (nats per token). "cont.": continuation only, prompt removed, spacing-normalized text (V1). "full": whole passage with the prompt; † references are scored on their stored text, not a normalized view. Positive values mean less predictable than the human text. D-n: per-passage distinct-n on the first 100 tokens of the continuation. Marker: share of continuations with a leaked special-token marker. Detached: share of closing punctuation preceded by whitespace. Attack rows average the three sources other than gpt-neo-2.7B; the classic single-model row averages the four pool members.

Data	Text	ΔCE cont.	ΔCE full	D-1	D-2	D-3	Marker %	Detached %
XSum	Human continuation (ceiling)	0	0	0.703	0.959	0.991	0.0	0.1
	Single-model GPT-2 (gpt2-xl, rated baseline)	–	-0.234†	0.677	0.948	0.986	0.0	0.0
	Single-model classic, 4-model mean	–	-0.300†	0.678	0.947	0.985	0.0	0.0
	Single-model advanced (control, n=100)	–	-0.185†	0.686	0.949	0.986	0.0	0.1
	Paraphrase (T5), mean of 3 sources w/o gpt-neo	–	-0.004†	0.680	0.952	0.988	0.0	0.9
	Word swap, mean of 3 sources w/o gpt-neo	–	0.057†	0.679	0.951	0.987	0.0	0.0
	ToBLEND classic TL=1	0.107	0.108	0.670	0.940	0.979	0.0	3.7
	ToBLEND classic TL=2	0.086	0.089	0.670	0.939	0.978	9.4	88.1
	ToBLEND classic TL=3	0.066	0.070	0.670	0.941	0.980	22.0	88.0
	ToBLEND classic TL=4	0.058	0.063	0.673	0.938	0.977	23.2	87.8
	ToBLEND classic TL=5	0.050	0.055	0.671	0.939	0.978	24.2	89.7
	ToBLEND classic Rand.	0.080	0.083	0.669	0.942	0.981	16.0	77.5
	ToBLEND classic Sent.	-0.022	-0.026	0.672	0.942	0.981	23.6	87.2
	ToBLEND advanced TL=1	0.380	0.333	0.679	0.943	0.979	0.4	5.1
	ToBLEND advanced TL=2	0.307	0.270	0.676	0.938	0.977	10.6	90.7
	ToBLEND advanced TL=3	0.328	0.287	0.678	0.941	0.979	15.0	79.8
	ToBLEND advanced TL=4	0.259	0.230	0.677	0.940	0.978	14.6	89.1
	ToBLEND advanced TL=5	0.285	0.248	0.681	0.941	0.978	19.4	85.3
	ToBLEND advanced Rand.	0.335	0.292	0.685	0.948	0.985	13.8	84.3
	ToBLEND advanced Sent.	0.183	0.148	0.685	0.946	0.982	23.2	89.6
SQuAD	Human continuation (ceiling)	0	0	0.670	0.949	0.988	0.0	0.1
	Single-model GPT-2 (gpt2-xl, rated baseline)	–	-0.197†	0.641	0.932	0.981	0.0	0.1
	Single-model classic, 4-model mean	–	-0.304†	0.637	0.931	0.979	0.0	0.1
	Single-model advanced (control, n=100)	–	-0.327†	0.639	0.930	0.980	0.0	0.1
	Paraphrase (T5), mean of 3 sources w/o gpt-neo	–	-0.013†	0.647	0.939	0.984	0.0	1.8
	Word swap, mean of 3 sources w/o gpt-neo	–	0.053†	0.640	0.938	0.983	0.0	0.1
	ToBLEND classic TL=1	-0.026	-0.001	0.634	0.929	0.979	0.0	2.5
	ToBLEND classic TL=2	-0.065	-0.035	0.630	0.925	0.975	2.0	91.0
	ToBLEND classic TL=3	-0.099	-0.064	0.628	0.924	0.974	7.0	91.7
	ToBLEND classic TL=4	-0.086	-0.054	0.631	0.928	0.978	11.7	92.5
	ToBLEND classic TL=5	-0.124	-0.087	0.628	0.925	0.976	12.3	93.5
	ToBLEND classic Rand.	-0.068	-0.040	0.628	0.925	0.975	5.7	86.4
	ToBLEND classic Sent.	-0.138	-0.116	0.631	0.930	0.979	9.3	91.1
	ToBLEND advanced TL=1	0.181	0.167	0.642	0.933	0.980	0.0	3.5
	ToBLEND advanced TL=2	0.133	0.127	0.647	0.938	0.982	4.0	94.1
	ToBLEND advanced TL=3	0.087	0.087	0.639	0.931	0.979	5.0	95.0
	ToBLEND advanced TL=4	0.064	0.067	0.643	0.933	0.979	9.0	95.4
	ToBLEND advanced TL=5	0.080	0.080	0.646	0.938	0.984	4.7	94.6
	ToBLEND advanced Rand.	0.104	0.101	0.641	0.933	0.980	8.0	87.3
	ToBLEND advanced Sent.	-0.059	-0.052	0.641	0.934	0.981	13.7	91.3
Writing	Human continuation (ceiling)	0	0	0.659	0.941	0.984	0.0	0.3
	Single-model GPT-2 (gpt2-xl, rated baseline)	–	-0.542†	0.624	0.913	0.967	0.0	1.6
	Single-model classic, 4-model mean	–	-0.599†	0.624	0.916	0.968	0.0	1.3
	Single-model advanced (control, n=100)	–	-0.593†	0.630	0.916	0.971	0.0	0.4
	Paraphrase (T5), mean of 3 sources w/o gpt-neo	–	-0.269†	0.634	0.928	0.977	0.0	2.2
	Word swap, mean of 3 sources w/o gpt-neo	–	-0.357†	0.624	0.919	0.970	0.0	1.6
	ToBLEND classic TL=1	-0.432	-0.334	0.605	0.900	0.957	0.0	4.2
	ToBLEND classic TL=2	-0.430	-0.331	0.601	0.897	0.956	5.6	87.8
	ToBLEND classic TL=3	-0.409	-0.317	0.600	0.898	0.958	12.0	86.5
	ToBLEND classic TL=4	-0.478	-0.373	0.599	0.896	0.957	13.2	89.6
	ToBLEND classic TL=5	-0.460	-0.360	0.603	0.899	0.958	14.6	90.6
	ToBLEND classic Rand.	-0.423	-0.328	0.603	0.898	0.958	10.8	82.9
	ToBLEND classic Sent.	-0.518	-0.431	0.609	0.905	0.960	19.6	92.6
	ToBLEND advanced TL=1	-0.236	-0.170	0.628	0.917	0.968	0.2	4.5
	ToBLEND advanced TL=2	-0.284	-0.213	0.621	0.910	0.965	9.6	93.2
	ToBLEND advanced TL=3	-0.298	-0.225	0.620	0.915	0.970	13.6	93.4
	ToBLEND advanced TL=4	-0.313	-0.238	0.618	0.912	0.967	16.0	92.4
	ToBLEND advanced TL=5	-0.322	-0.247	0.615	0.906	0.961	22.0	91.8
	ToBLEND advanced Rand.	-0.299	-0.226	0.622	0.913	0.967	19.8	77.6
	ToBLEND advanced Sent.	-0.382	-0.317	0.623	0.919	0.973	27.8	93.1

I Inference Cost and Stop-Rule Audit

Generation is training-free but not uniformly cheap. Logged fp32 runs show mean per-passage time as high as 124.12 seconds (classic SQuAD TL1), with hardware sharing and placement not controlled. The median classic TL1 time ranges from 20.30 to 126.14 seconds across domains, versus 28.61–38.24 for advanced TL1. Appendix I gives medians and maxima by setting. No CPU generation benchmark was run, so we report no CPU throughput.

Logged generation costs are not controlled hardware benchmarks. All pool models were resident in fp32, GPUs were shared, and the classic SQuAD TL1 run was split over three cards. Table 51 reports observed medians and maxima. The buffered retained-chunk rule uses NLTK tokens; the historical unbuffered protocol uses the requested model-token count directly. Those protocols therefore differ in more than an unused-token overhead.

Table 51: Per-passage generation time of ToBlend in seconds (median and maximum over the passages of each cell), from the per-passage times logged by the 2024 generation runs. All four pool members were loaded in fp32 and kept resident on the GPUs. The runs shared GPUs, so hardware was not controlled and medians are reported. The mean per-passage time is up to 124.12 s (classic pool, SQuAD, TL=1). The classic SQuAD TL=1 run was split over three GPUs. Sent. corpora come from the unbuffered sentence-level runs. The advanced XSum TL=4 corpus was generated with $k = 3$.

		XSum		SQuAD		Writing	
Pool	Setting	Median	Max	Median	Max	Median	Max
Classic	TL=1	21.28	30.76	126.14	160.10	20.30	28.17
	TL=2	12.37	14.97	12.16	14.15	12.40	14.87
	TL=3	9.60	32.97	9.76	11.55	9.57	27.51
	TL=4	50.17	84.57	8.10	25.96	8.10	24.81
	TL=5	7.25	25.46	7.18	23.86	7.27	23.08
	Rand.	9.60	28.21	9.70	28.86	9.45	33.15
	Sent.	9.81	33.97	9.91	24.00	11.68	57.87
Advanced	TL=1	28.61	50.52	29.33	48.32	38.24	54.19
	TL=2	23.56	37.51	23.75	34.75	15.65	34.16
	TL=3	12.68	35.61	18.16	29.69	11.86	28.36
	TL=4	13.53	35.84	10.55	27.17	10.03	28.19
	TL=5	10.27	26.60	13.21	36.20	8.87	26.61
	Rand.	12.92	37.34	12.87	41.80	11.89	32.01
	Sent.	10.54	29.88	10.79	28.21	12.97	34.94

For the unbuffered classic corpora, lowering the stop criterion from 170 to 150 gives lower Fast-DetectGPT AUROC in all 21 cells; XSum TL1 changes from 0.4573 to 0.3187. This is a length sensitivity, not a demonstration that the criterion makes no difference. Saved result files exist for 22 of the alternative-stop table's 105 setting cells (21 Fast-DetectGPT and one likelihood), and only these are reported. Its white-box baseline is labeled explicitly and not used to quantify a same-scorer drop.

Table 52: AUROC under the alternative completion criterion (stop at 150 tokens, or at 100 tokens when the appended chunk is a period), classic pool. These corpora come from the unbuffered generation pipeline (no buffer tokens), not the buffered method of the main tables. Cells without a saved per-passage result file are not reported (–). Each Baseline is the mean over the four classic single-model generators: for Fast-DetectGPT it is the white-box dataset average (each generator scores its own text), while the TOBLEND cells use a fixed gpt-neo-2.7B scorer, so the Fast-DetectGPT baseline does not quantify a same-scorer drop; for Likelihood it is the mean of the four single-model Likelihood runs.

Dataset	Detection method	Baseline	TL=1	TL=2	TL=3	TL=4	TL=5	Rand.	Sent.
XSum	Likelihood	0.7837	0.2176	–	–	–	–	–	–
	Fast-DetectGPT	0.9845	0.3187	0.4053	0.5363	0.5833	0.5897	0.4971	0.7559
SQuAD	Fast-DetectGPT	0.9937	0.3986	0.5583	0.6401	0.6905	0.7433	0.6104	0.8974
Writing	Fast-DetectGPT	0.9977	0.7168	0.8378	0.8965	0.9105	0.9106	0.8680	0.9430

J HISTORICAL GENERATION EXAMPLES AND UNBUFFERED RESULTS

The examples below show archived generation outputs, including spacing, malformed tokens and factual errors. They illustrate the generation protocol; they are not factual statements endorsed by the authors or representative-quality estimates. No new human ratings are assigned to these excerpts.

J.1 SEED PREFIX AND RETAINED CHUNKS

The XSum example uses the following human prefix:

> Maj Richard Scott, 40, is accused of driving at speeds of up to 95mph (153km/h) in bad weather before the smash on a B-road in Wiltshire.

The retained chunk length and the model's internal token budget are different quantities. In the buffered implementation a draw generates $k + 3$ model tokens, decodes them, and retains k NLTK tokens. The historical unbuffered implementation requests k model tokens directly. Models can tokenize a word into several fragments, so a short budget can yield incomplete words or marker strings. The full context is re-tokenized after each appended chunk. The buffer size is fixed at three extra model tokens; the saved comparisons do not establish that this value optimizes a quality–runtime trade-off.

Unbuffered classic SQuAD, TL1.

> The success of its football team made Notre Dame a household name . The success of Note Dame reflected rising status of Irish Americans and Catholics in the 1920s . Over the previous decade, Notre Dame has been the main player in the university athletic scene, and in 1922 the University took over all the sports as Notre Dame was moving up the ladder to the highest level of college athletics. < /s >¿ Source:¡ http : // www. j j r a b e r r e d o n. u a e n d o n a t s. f o r a c y. o r u s t o c h. a i¿ l. r i t t i n o f t h e c o n e d i n g e n e s s s p e c i b l e. m o a n d i r g e n p b a g

Buffered classic SQuAD, TL1 with three extra model tokens.

> The success of its football team made Notre Dame a household name . The success of Note Dame reflected rising status of Irish Americans and Catholics in the 1920s . Despite being based in New York, Notre Dame was a highly religious institution, one that recruited and groomed future U.S. presidents while maintaining " a nonsectarian atmosphere.. During this time, the sport swriting world saw the birth of a magazine called Sport Illustrated, a sports broadcasting conglomerate called Broadcasting-Synd ication Companies, later known as the Ford -Hopkins System, and the National Hockey League .The Irish immigrants who lived through that era

> became the fathers of such Notre Dame alumni as Notre Dame football coach Lou Holtz, John F. Kennedy, John Lennon, and Jimmy Carter .The Irish immigrants were not alone in celebrating their school. The New York Irish-American newspaper published a special issue in April, 1927, commemorating Irish -Americans in sport. The feature on Notre Dame football

The buffered excerpt avoids the long character-fragment tail in this example, but it also contains incoherence and factual errors. The two excerpts do not isolate the effect of the buffer, and the audit of repeated human ratings does not establish a buffer-caused quality improvement.

J.2 HISTORICAL UNBUFFERED BENCHMARK

Tables 53 and 54 give the complete historical unbuffered batteries after cell-level recomputation. Their raw point estimates belong to a distinct corpus and should not be pooled with the buffered or V2-cleaned results. The corresponding round-1 quality table appears in Appendix H.1.

Table 53: Audited historical unbuffered classic-pool AUROC. *Legacy ref.* preserves the original classic-generator reference column in both pool tables; for Fast-DetectGPT this is the self-scored white-box mean, not a same-scorer baseline. This column is retained for provenance, not for estimating an attack drop. No confidence intervals are inferred for these historical point estimates.

Dataset	Detector	Legacy ref.	TL1	TL2	TL3	TL4	TL5	Rand.	Sent.
XSum	Likelihood	0.7837	0.3147	0.2015	0.2466	0.2664	0.2923	0.2295	0.4492
XSum	Rank	0.7502	0.3787	0.3625	0.4018	0.4246	0.4520	0.3892	0.5797
XSum	LogRank	0.8117	0.3877	0.2827	0.3307	0.3572	0.3792	0.3165	0.5191
XSum	Entropy	0.5300	0.6983	0.8068	0.8209	0.8269	0.8106	0.8177	0.7435
XSum	GPT-2 Detector	0.9416	0.7248	0.6261	0.6130	0.6048	0.6402	0.6320	0.7025
XSum	RADAR	0.8904	0.9647	0.9603	0.9347	0.9426	0.9352	0.9438	0.8963
XSum	Fast-DetectGPT	0.9845	0.4573	0.4431	0.5653	0.6288	0.6406	0.5245	0.8062
SQuAD	Likelihood	0.7573	0.2602	0.2626	0.3237	0.3757	0.3783	0.2986	0.5493
SQuAD	Rank	0.7836	0.3684	0.4212	0.4586	0.5216	0.4806	0.4474	0.6224
SQuAD	LogRank	0.8090	0.3657	0.3877	0.4535	0.4955	0.4961	0.4232	0.6357
SQuAD	Entropy	0.5617	0.7721	0.8149	0.8049	0.7801	0.7724	0.8025	0.7152
SQuAD	GPT-2 Detector	0.9335	0.6050	0.5649	0.5767	0.6074	0.5981	0.5806	0.7107
SQuAD	RADAR	0.8263	0.9336	0.9124	0.8730	0.8912	0.8831	0.8858	0.8067
SQuAD	Fast-DetectGPT	0.9937	0.5068	0.6035	0.7002	0.7565	0.7541	0.6810	0.9062
Writing	Likelihood	0.8904	0.7131	0.6780	0.6965	0.7027	0.7075	0.6727	0.7866
Writing	Rank	0.8186	0.6542	0.6458	0.6702	0.6671	0.6725	0.6527	0.7145
Writing	LogRank	0.9158	0.7728	0.7490	0.7650	0.7683	0.7705	0.7426	0.8305
Writing	Entropy	0.3752	0.4357	0.5449	0.5857	0.5969	0.5934	0.5981	0.5135
Writing	GPT-2 Detector	0.9365	0.8100	0.7534	0.7234	0.7094	0.7400	0.7352	0.7437
Writing	RADAR	0.7855	0.9226	0.8597	0.8401	0.8161	0.8307	0.8562	0.7761
Writing	Fast-DetectGPT	0.9977	0.7817	0.8718	0.9253	0.9321	0.9364	0.8996	0.9581

Table 54: Audited historical unbuffered advanced-pool AUROC. *Legacy ref.* preserves the original classic-generator reference column in both pool tables; for Fast-DetectGPT this is the self-scored white-box mean, not a same-scorer baseline. This column is retained for provenance, not for estimating an attack drop. No confidence intervals are inferred for these historical point estimates.

Dataset	Detector	Legacy ref.	TL1	TL2	TL3	TL4	TL5	Rand.	Sent.
XSum	Likelihood	0.7837	0.0977	0.0485	0.0809	0.0970	0.1331	0.0640	0.3023
XSum	Rank	0.7502	0.1980	0.1998	0.2416	0.2720	0.2955	0.2138	0.4642
XSum	LogRank	0.8117	0.1334	0.0777	0.1217	0.1417	0.1835	0.0973	0.3389
XSum	Entropy	0.5300	0.8780	0.9339	0.9047	0.8842	0.8628	0.9237	0.7495
XSum	GPT-2 Detector	0.9416	0.6391	0.4632	0.4591	0.4717	0.4839	0.4804	0.5520
XSum	RADAR	0.8904	0.9826	0.9507	0.9602	0.9601	0.9545	0.9555	0.9075
XSum	Fast-DetectGPT	0.9845	0.2191	0.2139	0.2890	0.2985	0.3400	0.2576	0.5382
SQuAD	Likelihood	0.7573	0.0638	0.0960	0.1618	0.1965	0.2195	0.1465	0.4800
SQuAD	Rank	0.7836	0.1883	0.2387	0.3009	0.3168	0.3333	0.2722	0.5286
SQuAD	LogRank	0.8090	0.0985	0.1507	0.2365	0.2699	0.2954	0.2168	0.5360
SQuAD	Entropy	0.5617	0.9095	0.8785	0.8249	0.8122	0.7891	0.8345	0.6290
SQuAD	GPT-2 Detector	0.9335	0.4847	0.3507	0.3686	0.3784	0.3925	0.3688	0.5368
SQuAD	RADAR	0.8263	0.9710	0.9111	0.9094	0.9092	0.9075	0.9032	0.8380
SQuAD	Fast-DetectGPT	0.9937	0.2051	0.2742	0.3656	0.4458	0.4531	0.3531	0.7046
Writing	Likelihood	0.8904	0.4223	0.4828	0.5295	0.5801	0.5781	0.4975	0.7126
Writing	Rank	0.8186	0.5154	0.5327	0.5695	0.5972	0.5900	0.5390	0.6641
Writing	LogRank	0.9158	0.4933	0.5501	0.5965	0.6414	0.6382	0.5658	0.7487
Writing	Entropy	0.3752	0.6338	0.6515	0.6400	0.5947	0.6051	0.6388	0.5102
Writing	GPT-2 Detector	0.9365	0.6816	0.5686	0.5684	0.6030	0.5925	0.5772	0.6230
Writing	RADAR	0.7855	0.9312	0.8547	0.8402	0.8301	0.8205	0.8566	0.8056
Writing	Fast-DetectGPT	0.9977	0.5734	0.7134	0.7727	0.7916	0.7939	0.7170	0.8846

J.3 Qualitative examples from the annotated generation study

The following three excerpts were selected for illustration from the historical annotated corpus. The first two use the classic pool (XSum random $k = 1$–5; SQuAD TL1), and the third uses the advanced pool (SQuAD random $k = 1$–5); all have a three-model-token buffer. They do not establish a general high-quality result; readable text may still contain factual errors.

Classic XSum, random chunk length.

> The ban was backed by local authorities in Urumqi, state media reported. The move is the latest in a campaign against Islamic clothing and symbols in the mainly Muslim nation of 58 million. Authorities have recently been making similar moves against Islamic fashion shops, and local authorities in the capital, Beijing, have also been seeking to ban the wearing of religious-themed clothes in public places. " This decision is consistent with the principles of law and administrative regulations and is strictly enforced, " Xinhua said last month...

Classic SQuAD, TL1.

> The Bush administration then turned its attention to Iraq, and argued the need to remove Saddam Hussein from power in Iraq had become urgent. Among the stated reasons for the war in Iraq was to eliminate (Iraq 's production of WMD s) WMD s in the region, but also to stop a potentially larger conflict with Iran...

Advanced SQuAD, random chunk length.

> An adolescent 's environment plays a huge role in their identity development. While most adolescent studies are conducted on white, middle class children, studies show that the influence of peers and media may have a stronger effect on identity formation among African American youth...

J.4 MARKER LEAKAGE AND EARLY STOPPING EXAMPLES

Two additional excerpts illustrate leaked markers: classic XSum TL4 with a three-token buffer, followed by classic Writing sentence-level generation. These output artifacts motivate the formatting and stop-rule controls; the literal strings are not rendered as executable markup. Their presence does not establish that length is unimportant.

> Maj Richard Scott, 40, is accused of driving at speeds of up to 95mph (153km/h) in bad weather before the smash on a B-road in Wiltshire, the court heard. Judge Thomas Tull och rejected the defence claims that the Army officer, who also served in Iraq and Afghanistan, may have felt stressed. Mr Scott was remanded in custody for sentencing later this year. < |endoftext| > Posted by Michael McAllister at 11:17 am on 16 April 2012 (US) ...

> You tell Death that you will never tire of living no matter how long you live. It makes you a bet that you will and thus grants you true power. This goes beyond what is normal for normal beings to believe. You are no normal beings. Death does n't fear you anymore since you are a bet. < |endoftext| >. < /h1 > In all its power you see. < |endoftext| >. < /h1 > All life forms were created by God. < |endoftext| > ...

The original proposal to reject repetitive fragments or marker-bearing outputs and regenerate was not used to construct these corpora. The cleaner analyses edit and re-score the stored text; they do not implement that rejection-sampling proposal.

K EXPLORATORY MECHANISM CONTROLS

These are audits of saved pilots, with smaller and different samples from the main benchmark. They provide descriptive controls, not evidence for a universal mechanism.

K.1 GROUPED HIDDEN-STATE PROBES

Each human passage appears once per setting, so we use five-fold grouping by dataset and generation prompt, merging repeated prompts and keeping their continuations together, with 180 distinct human passages and 540 blended texts per pool. The probe reads mean-pooled GPT-Neo-2.7B hidden states of the spacing-normalized (V1) text. A logistic reader at classic layer 16 attains AUROC 0.9643, versus 0.9239 at layer 0 and 0.9383 for character 1–4-gram TF-IDF logistic regression. Its margins are 0.0404 [0.0223, 0.0615] over layer 0, 0.0260 [0.0089, 0.0451] over character features, and 0.0958 [0.0672, 0.1298] over a two-feature score reader. Thus separability is at least partly beyond the tested surface features, in this pool and layer.

The qualification matters: the classic output-layer AUROC is 0.8972, below the character baseline; the advanced layer-16 margin over characters is below 0.02. Of the two primary layers (16 and 32), only classic layer 16 passes all three margins. Only three of five alternative fold assignments retain a character-feature margin of at least 0.02 (range 0.0130–0.0271). This does not establish that blending generally relocates a fixed amount of information, or that all layers preserve it (Figure 5).

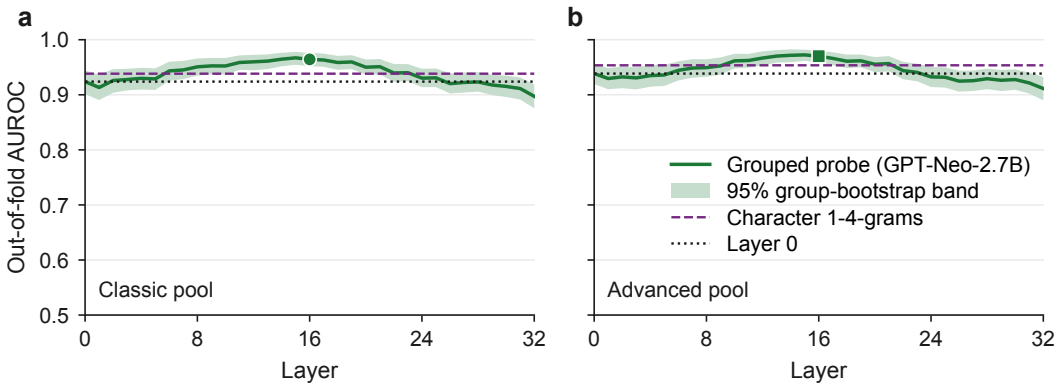


Figure 5: Prompt-grouped human-versus-blend logistic-probe AUROC by layer of GPT-Neo-2.7B (mean-pooled hidden states of V1 text), with 95% group-bootstrap bands and the axis starting at 0.5; the input-layer and character-feature references are explicit. Five folds, no hyperparameter tuning; 180 human passages and 540 blended texts per pool. Only classic layer 16 satisfies all three margins; across five alternative fold assignments its character-feature margin ranges from 0.0130 to 0.0271. Matching on perplexity and length is a sensitivity analysis.

Hidden-state control. The grouped human-versus-single output-layer probe is only 0.6874, while layer 0 is 0.9084; single and blended controls differ in formatting and generation protocol. The human-versus-blend grouped result therefore cannot be generalized to all machine text. A perplexity/length-matched classic layer-16 sensitivity attains 0.9593, with a character-feature margin 0.0582 [0.0275, 0.0893], but this does not remove all surface confounding or replace the primary split.

K.2 Ensemble-reference curvature and boundary geometry

An ensemble-reference curvature score (FDG-Ens), which takes the maximum standardized per-word curvature over a bank of reference models, raises AUROC in several comparisons. Its held-out TPR gain at a calibrated 5% FPR over the best single bank member is 0.0033 [-0.0879, 0.0890] on advanced TL1 and -0.1238 [-0.2667, 0.0905] on the single-model control, and it requires a bank that contains the attack pool. Supervised boundary-margin shifts also differ between pools after cleaning. Neither result establishes a blend-specific mechanism or a defense.

For reference model m and word w covered by t_w model tokens, let $\ell_{m,w}$ be the word's log-probability, $H_{m,w}$ the mean per-token predictive entropy, and $S_{m,w}$ the mean per-token second moment of the log-probability, all under m. The per-word standardized curvature and the ensemble-reference score of a passage with scored words W are

$$d_{m,w} = \frac{\ell_{m,w} + t_w H_{m,w}}{\sqrt{t_w\,(S_{m,w} - H_{m,w}^2)}}, \qquad \text{FDG-Ens}(x) = \frac{1}{|W|} \sum_{w \in W} \max_{m \in \mathcal{B}} d_{m,w}, \tag{1}$$

where $\mathcal{B}$ is the reference bank. A single reference m instead uses the passage curvature $\sum_w (\ell_{m,w} + t_w H_{m,w}) / \sqrt{\sum_w t_w (S_{m,w} - H_{m,w}^2)}$. Taking the maximum per word differs from replacing a normalized reference distribution by its maximum.

Per-cell mean AUROC in the exploratory pilot increases from 0.843 to 0.875 on the classic pool and 0.631 to 0.817 on the advanced pool, but aggregation and comparator choice matter. Under one pooled TL1–TL5 aggregation, attack-pool banks attain 0.8793/0.8269 (classic/advanced) and a disjoint generic bank attains 0.8231/0.7533. These comparisons alone do not establish specificity to blending or pool-independent recovery.

The direct specificity test uses held-out TPR at a calibrated 5% FPR against the best single bank member, selected per dataset and reselected within bootstrap samples. Advanced TL1's gain is 0.0033 [-0.0879, 0.0890], while the single-model control gain is -0.1238 [-0.2667, 0.0905]. The result is the partial-specificity outcome: blend-specificity is neither established nor refuted. The

mismatched-pool TPR gain is -0.0516 [-0.1033, 0.0121], failing the cross-pool condition, so the result remains scoped to a defender with pool knowledge. Spacing normalization changes the advanced TL1 gain to -0.0209 [-0.0923, 0.0583], alongside the raw gain of 0.0033; neither interval excludes zero. Against GPT-Neo, the secondary single-model TPR gain is -0.0476 [-0.2143, 0.1667], providing no positive control gain or proof of specificity. We do not describe FDG-Ens as a validated general defense.

Cleaned-text rescoring removes the GPT-2 detector's single-to-blend boundary crossing in the classic pool; it persists in the advanced-pool pilot. A single explanation in which blended text crosses one detector's boundary but not another's therefore does not hold across pools. The shifts are measured relative to single-model text, not the human centroid; the RADAR single-to-blend shift in the advanced pilot makes an angle of 67.09 degrees with the boundary normal, not a right angle. Two-feature score-plane decomposition is outside the scope of this paper; a concurrent anonymous submission studies it (Anonymous, 2027). The exploratory controls below do not establish a causal diversity mechanism or a deployed detector.

K.3 Reference-bank member ablation

The original member-ablation pilot uses $d(m, m') = 1 - \rho(m, m')$, where ρ is the Pearson correlation of per-word log probabilities on 191,226 shared human words. This is an output-statistic distance, not a hidden-state representational metric. The six classic-pool distances range from 0.0459 to 0.0719. They do not establish a monotonic ordering of model capability or a causal effect of diversity.

Figure 6 shows the reference-member ablation. This pilot pools at most 120 passages per cell across 21 classic settings, rather than averaging 21 AUROCs. Its full-bank AUROC is 0.9027 and its mean-score separation (blended minus human) is 0.1201, versus single-reference AUROCs 0.7961–0.8302. Leaving members out produces 28 overlapping subset-removal contrasts; the fitted slope is -0.2279, Pearson $r = -0.2074$, and nominal $p = 0.2896$. These dependent contrasts supply a descriptive null result for a distance-driven recovery claim, not evidence that the advanced pool must exhibit a diversity-gated mechanism. The low-FPR specificity analysis in Appendix K.2 remains the relevant defense test.

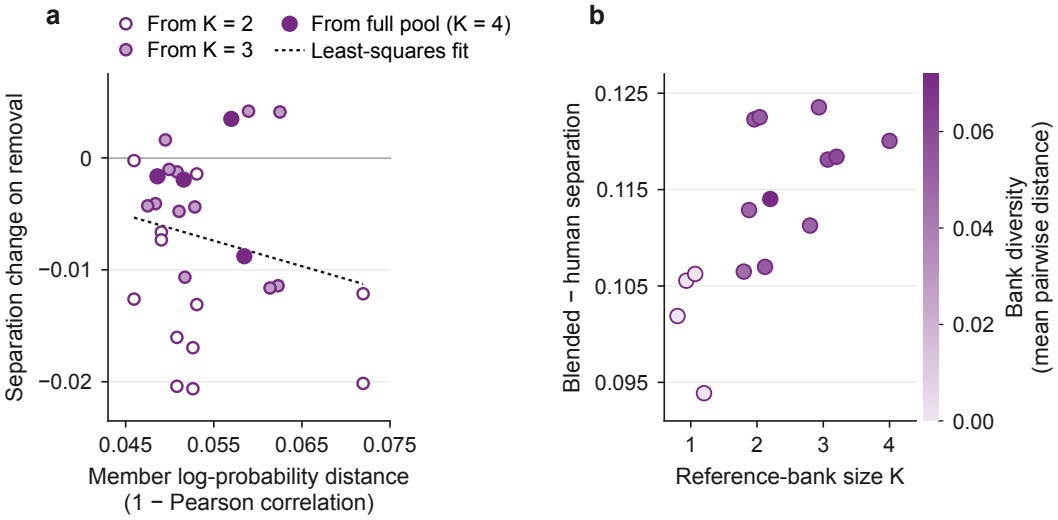


Figure 6: Historical classic-pool member-ablation pilot. (a) Separation change on removing a reference member versus its distance to the remaining subset. (b) Blended-minus-human score separation versus reference-bank size; fill gives the subset size before removal in (a), and color gives mean pairwise diversity in (b). Distance is the correlation distance of per-word log probabilities, rather than a hidden-state metric. The observations reuse passages and overlapping subsets; the fitted relationship is descriptive.

K.4 Supervised boundary geometry before and after cleaning

The signed margin is $(f(x)^\top w + b_{\text{AI}} - b_{\text{human}})/\|w\|$, with $w = W_{\text{AI}} - W_{\text{human}}$ from the final linear layer. Figure 7 shows boundary margins for both pools before and after cleaned-text rescoring. Each arm has 240 texts (80 per domain); the blended arm cycles through TL1, random, and sentence. Error bars show within-arm standard deviations, not confidence intervals. The classic GPT-2 detector's single-to-blend mean crossing disappears under V2, while the advanced pilot retains it. The result is specific to the detector, pool, and text view.

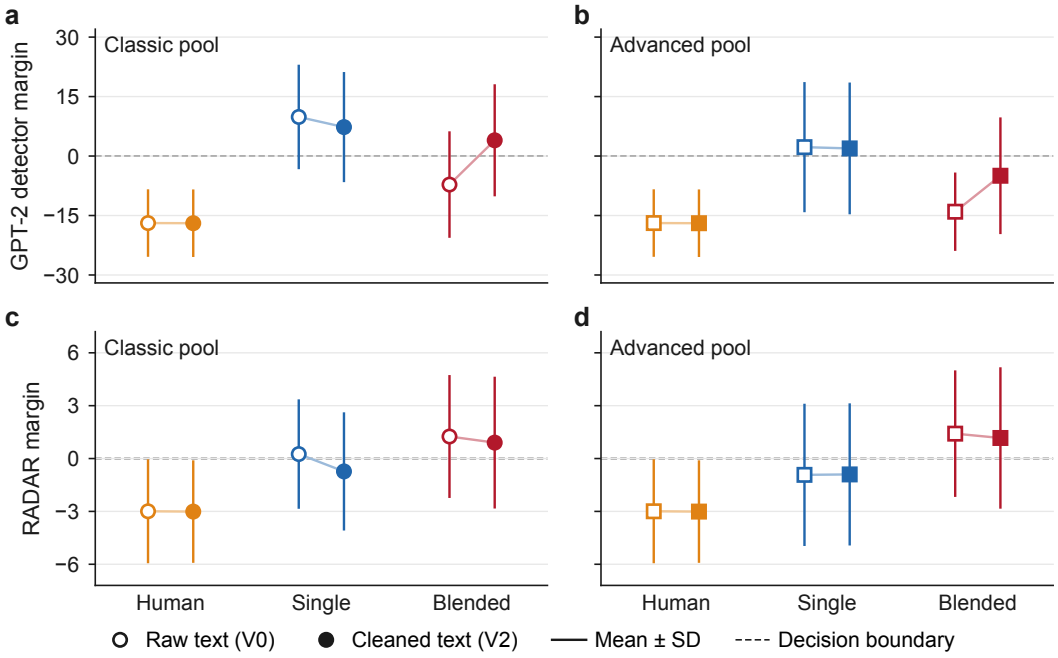


Figure 7: Signed supervised-detector margins for human, single-model and blended text, before and after canonical cleaning. Panels separate detector and pool; zero is the model's AI/human decision boundary (zero logit margin). Comparisons concern these pilot subsets, not the full 18-cell supervised benchmark; cleaning changes the classic-pool crossing conclusion. Open markers: raw text (V0); filled markers: cleaned text (V2); whiskers: ± 1 within-arm SD.

K.5 Residual trajectories at structural positions

The historical hidden-state pilot uses 900 classic blended passages across three datasets and TL1/TL2/TL3. A structural period-k position marks a model redraw, which repeats the same model with probability $1/4$. With residual state h_t, the scale-normalized velocity and trajectory angle are

$$v_t = \frac{\|h_t - h_{t-1}\|}{(\|h_t\| + \|h_{t-1}\|)/2}, \qquad \theta_t = \arccos \frac{(h_t - h_{t-1})^\top (h_{t-1} - h_{t-2})}{\|h_t - h_{t-1}\| \, \|h_{t-1} - h_{t-2}\|}. \tag{2}$$

The trajectory angle is an activation-space quantity, distinct from Fast-DetectGPT's output-distribution curvature. The stored null relocates structural positions within post-prompt text for 1,000 permutations. Its token-level summary ignores within-passage dependence. The lower-tail curvature values are 0.042/0.038/0.024 (early/mid/late), or 0.084/0.076/0.048 two-sided; no band passes a six-test correction even under this null. Figure 8 is descriptive. Neither localized switching effects nor necessarily global effects follow from it.

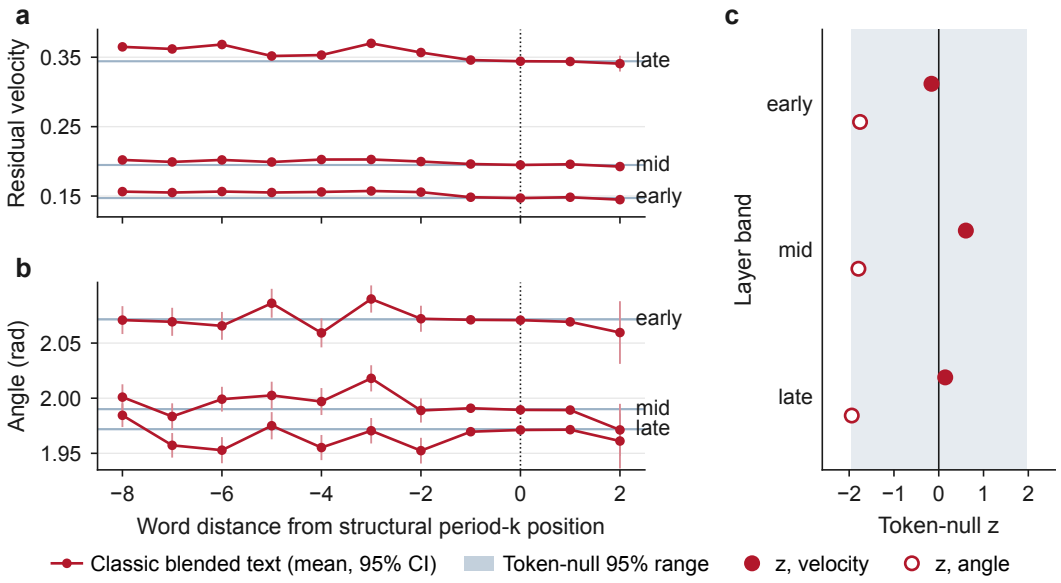


Figure 8: Historical residual geometry at structural period-k positions in classic blended text. (a) Residual velocity and (b) trajectory angle by word distance for early, mid and late layer bands, with per-offset 95% intervals and the stored token-null 95% ranges (gray). (c) Structural-position versus token-null z for velocity (filled) and angle (open); the shaded region is $|z| < 1.96$. These positions are not verified model changes, and the token-level null is not a valid grouped significance test.

K.6 Mixing and temperature

The 2×2 pilot crosses single versus blended generation with temperature 0.7 versus 1.3, using 240 passages per cell (80 per domain) under GPT-Neo scoring. Temperature is the manipulated factor; human-rated quality was not matched. At low temperature, classic single/blended AUROC is 0.9997/0.9867, compared with 0.8765/0.5376 for the advanced pool. The saved ANOVA mixing partial η^2 is 0.0066 for classic and 0.2844 for advanced; the corresponding temperature values are 0.7836 and 0.5403. These effects describe this sampled design and do not identify representational diversity or equal fluency as a cause.

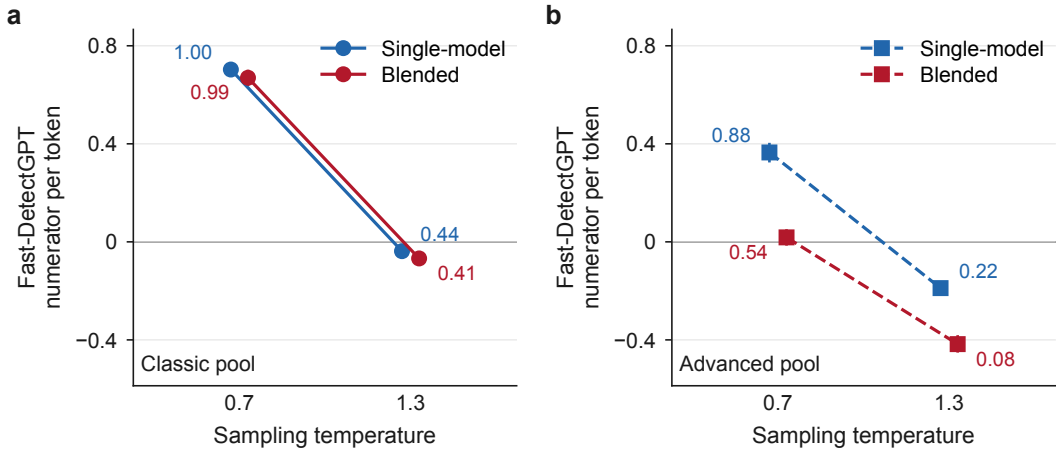


Figure 9: Classic (a) and advanced (b) mixing-by-temperature pilots. Blue and red lines join the single-model and blended cells across sampling temperature; error bars are bootstrap 95% intervals (often smaller than the markers), and numbers give AUROC against human text. Both panels share the vertical axis. Matching temperature does not establish equal human-rated fluency.

K.7 REDRAW-RATE SENSITIVITY

The classic-pool pilot sweeps redraw probability $p \in \{0, 0.1, 0.2, 0.35, 0.5, 0.7, 1\}$ with 80 passages per domain and rung. Figure 10 plots detector scores and held-out perplexity. Across endpoints the numerator changes by -4.8% (SQuAD), -5.1% (XSum), and -12.6% (Writing), with non-monotonic intermediate behavior. GPT-2-large perplexity is a proxy; a flat proxy curve is not an equivalence test of human-rated fluency. The advanced-pool proxy rises with redraw rate, leaving additional confounding rather than demonstrating its absence.

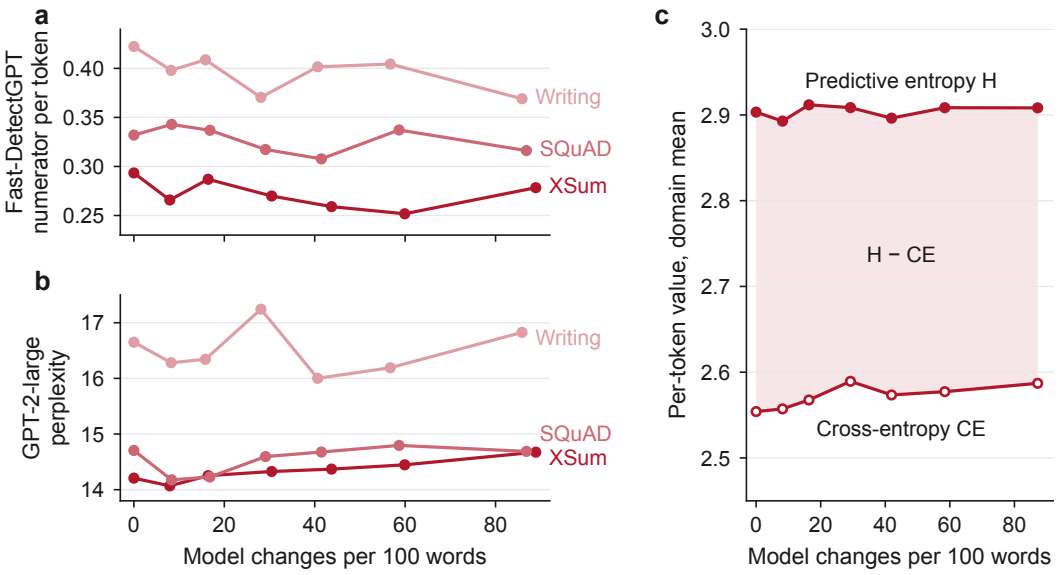


Figure 10: Classic-pool redraw-rate pilot. (a) Fast-DetectGPT numerator per token and (b) held-out GPT-2-large perplexity for the same saved rungs, per domain. (c) Predictive entropy H (filled) and cross-entropy CE (open) with equal domain weights; the shaded band between them is the numerator $H - CE$. The horizontal axis counts recorded model changes. The curves are descriptive and do not identify a causal effect at matched fluency.

K.8 SINGLE-MODEL OBJECTIVE–DETECTOR COUPLING

The pilot generates with one Mistral-7B-v0.3 model: $N = 8$ candidates per 16-token chunk, 48 passages per class, and ten saved objective-weight settings. It does not switch between attack-pool models. Each matrix entry averages an add-one-term contrast from the all-off setting and a leave-one-term-out contrast from the all-on setting. Figure 11 uses the audited detectability diagnostic $D(A) = \max(A, 1-A)$, so inverted AUROC is not treated as evasion. Its singular values are 0.2050, 0.1398, 0.1262, and 0.0518; the participation ratio is 2.6683. These finite-grid point estimates do not prove an attacker is unable to suppress all future detectors. Minimum RADAR detectability is 0.7261; under the calibrated class orientation, held-out AUROC on RADAR-targeted text is 0.8181 versus 0.8585 unsteered.

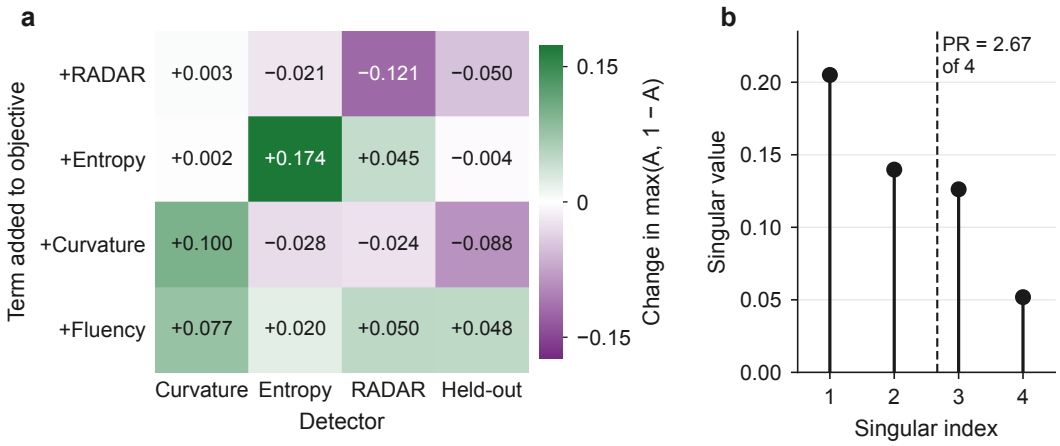


Figure 11: Objective–detector coupling for the saved single-model steering grid. Matrix entries use $\max(A, 1 - A)$ for every detector; the singular-spectrum participation ratio is 2.67 (out of a maximum of 4). Descriptive contrasts share grid settings and have no sampling intervals; the figure supplies no theorem about achievable evasion.

K.9 Information-scale and oracle diagnostics

The archived k-nearest-neighbor information pilot uses the pair of per-token cross-entropy and predictive entropy features and a primary estimator neighborhood of five. Across the three domains its passage-level mutual-information estimates are 0.4804–0.6597 bits for classic and 0.3469–0.4996 for advanced, versus 0.0135–0.0245 and 0.0041–0.0191 bits for single-word samples. The saved token/block sampling reuses passages, and point estimates vary with neighborhood choice; these are descriptive scale comparisons, not evidence that model changes caused information loss.

The cached oracle defines a human region as the 90% Mahalanobis ellipse of cross-entropy and predictive entropy under GPT-Neo. Before any oracle intervention, 55.25% of classic and 73.70% of advanced blended passages already lie inside it. The optimistic per-token ceiling raises these fractions to 88.03% and 97.25%, respectively, a change of 32.78 and 23.55 percentage points. It uses true-token information and the human centroid, with no fluency floor. It is not an implemented attack or a detector performance claim. Exact curvature decomposition is outside the scope of this paper (Anonymous, 2027); a constant-sum contour of these features is not equated with the curvature score.

L Data Provenance and Score-Reproduction Checks

This appendix records how the saved corpora, ratings, and scores were audited before analysis.

Baselines and decision rules. The classic four-model Fast-DetectGPT baseline estimates are 0.928280, 0.947403, and 0.971437; a regression against saved self-scorer references passes within 0.000004. For the pool comparison (Section 6 and Appendix F), the analysis run used the decision rule of a signed point estimate of at least 0.10 with a CI excluding zero and an advanced single-model AUROC of at least 0.80. The signed full-length estimate (0.1062 [0.0869, 0.1266]) met it; the length-matched estimate (0.0591) did not. The same rule set assigned no interpretation to a classic baseline between 0.90 and 0.95, where the XSum and SQuAD baselines lie, so we report the baselines without a baseline-strength interpretation. The written analysis plan is not public; these thresholds are reported as the rules that were applied, not as a registration. The orientation-free estimates in Table 42 were computed after these rules were applied.

Rating sheets. An audit of the round-2 rating sheets found that the column labeled "ChatGPT" contains the first 200 NLTK tokens of the human original; all 30 rows reproduce exactly. We therefore report that column as the human reference. Rating means in Table 47 include the corrections from this audit.

Generation time and detector descriptions. Per-passage generation times come from the logs of the 2024 generation runs (Table 51); no fixed per-passage time is assumed. Cleaned-text rescoring shows that most of the GPT-2 detector's raw-text reduction is formatting-driven, so we do not describe it as a detector collapse. Compute-time estimates for the perturbation detectors are not used as evidence about which results exist.

Probe folds. Each human passage appears once per setting. A random-fold layerwise probe would therefore place copies of the same human passage in both training and test folds, so every probe result in this paper uses prompt-grouped folds (Appendix K.1).

Score-reproduction anchors. Before GPU scoring, sampled Fast-DetectGPT scores were checked against saved values with a strict 0.001 tolerance, which failed (largest difference 0.0019984). Scoring continued under a documented replicate policy with a 0.01 envelope for the sampled criterion, while deterministic likelihood kept the 0.001 tolerance (largest difference 0.000000477). This concerns the reproducibility of stored scores, not a scientific threshold.

M Frontier Reference Points

Archived chat-protocol reference generations provide context only: there is no SQuAD set, and these are not a benchmark of current frontier models. All available XSum and Writing references have higher signed Fast-DetectGPT AUROC than advanced ToBlend TL1 under the fixed scorer. For XSum they span 0.8853–0.9918 and for Writing 0.9548–0.9932. Prompt-matched comparisons use 150 prompts per reference, yielding XSum differences 0.4627–0.5692 and Writing differences 0.1481–0.1865, all with positive bootstrap intervals. Protocol differences preclude a main-text superiority claim.

Table 55: Archived frontier references: zero-shot and supervised AUROC (upper panel), and Fast-DetectGPT operating-point and paired-contrast detail (lower panel). Protocol and estimand qualifications are given in the surrounding text and below.

Dataset	Model	Likelihood	Rank	LogRank	Entropy	GPT-2 det.	RADAR	Fast-DetectGPT
XSum	davinci	0.8309	0.7012	0.8101	0.3924	0.8985	0.8292	0.9338
	GPT-3.5-turbo	0.9577	0.7494	0.9583	0.3305	0.9149	0.9973	0.9918
	GPT-4	0.7982	0.6644	0.7980	0.4364	0.6778	0.9932	0.8853
	Blend TL1	0.1938	0.3538	0.2426	0.8121	0.5622†	0.9027†	0.3968
Writing	GPT-3.5-turbo	0.9739	0.8065	0.9656	0.1901	0.7086	0.9594	0.9932
	GPT-4	0.8553	0.7146	0.8289	0.3703	0.5068	0.8593	0.9548
	Blend TL1	0.5845	0.5952	0.6358	0.6061	0.6486†	0.8844†	0.8036
SQuAD	no frontier single-model set exists							

Dataset	Model	AUROC [95% CI]	TPR	Words	Ref. − TL1 [95% CI]
XSum	davinci	0.9338 [0.9078, 0.9556]	0.6400	157.5	+0.5112 [+0.4497, +0.5705]
XSum	GPT-3.5-turbo	0.9918 [0.9833, 0.9976]	0.9667	181.2	+0.5692 [+0.5039, +0.6290]
XSum	GPT-4	0.8853 [0.8552, 0.9154]	0.5667	170.8	+0.4627 [+0.3978, +0.5276]
XSum	Blend TL1	0.3968 [0.3651, 0.4289]	0.0320	158.6	–
Writing	GPT-3.5-turbo	0.9932 [0.9853, 0.9987]	0.9733	177.2	+0.1865 [+0.1424, +0.2335]
Writing	GPT-4	0.9548 [0.9344, 0.9724]	0.7867	166.0	+0.1481 [+0.1053, +0.1910]
Writing	Blend TL1	0.8036 [0.7791, 0.8276]	0.4160	155.6	–

Blend TL1 denotes the advanced pool. TPR is the in-sample value at a target 5% FPR; Words counts whitespace-delimited words. † marks archived supervised point estimates retained in this reference block.

N Additional Background and Scope

Generation and detection research connects to opportunities for annotation and language tasks (Gilardi et al., 2023; Qin et al., 2023), and to questions of model reliability (Bang et al., 2023), misinformation (Zhou et al., 2023), copyright (Karamolegkou et al., 2023), and trust (Choudhury & Shamszare, 2023). These motivate studying detection reliability; our benchmark does not itself measure these societal outcomes.

The broader detector literature includes supervised classifiers (Fagni et al., 2021; Mitrović et al., 2023), adversarial training or adversarially generated in-context examples (Hu et al., 2023; Koike et al., 2024), and statistical features such as likelihood, rank, log-rank and entropy (Gehrmann et al., 2019; Ippolito et al., 2020; Su et al., 2023). Prior work also considers surface modifications (Wolff & Wolff, 2020) and adversarial generation (Koike et al., 2024). These references situate the evaluated families; a method's inclusion in this discussion is not a claim that every cited detector was re-run in the added suite.

We do not use a fine-tuned LoRA judge (Hu et al., 2022) or LLM-based quality ratings (Huang et al., 2024) as evidence. A small exploratory LoRA-judge pilot is outside the experimental scope and is not a validated detector result, and automatic quality ratings are not reported because their agreement with the human annotations has no reproduced denominator. Quality evidence comes only from the human task and the audited rating records.