MELD: MULTI-TASK EQUILIBRATED LEARNING DETECTOR FOR AI-GENERATED TEXT

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models are widely used in everyday writing, making reliable AI-generated text detection crucial for academic integrity, content moderation, and provenance tracking. Yet high AUROC on clean, in-distribution benchmarks is not sufficient. Practical detectors must resist adversarial rewrites, generalize to unseen generators and writing domains, and maintain low false-positive rates (FPR). A pooled AI-versus-human objective does not explicitly require the model to distinguish among generator families, so it may fail to learn the generator-specific structure needed for generalization and attribution. We introduce MELD (Multi-Task Equilibrated Learning Detector), which instead trains on class-balanced, family-specific AI-versus-human tasks while sharing a common representation of human writing. MELD produces detection, generator-family, rewrite-task, and token-level predictions in a single forward pass. Format normalization further makes its scores invariant to the modeled reformatting operations. MELD ranks first among open-source submissions in the public RAID leaderboard and matches or exceeds supervised baselines on five of six held-out evaluation pools. To evaluate transfer to unseen generators, we introduce MELD-eval, a held-out test pool built from four frontier chat models. Without further fine-tuning, MELD achieves 99.7% TPR at 1% FPR on MELD-eval and 98% TPR at the same FPR on a held-out generator whose family is absent from training. Finally, in a case study of 5.9 million scientific texts from 2016–2026, MELD's prediction scores remain stable through 2022 and increase from 2023 onward, coinciding with the widespread adoption of LLM-based writing tools. The model and MELD-eval pool (https://huggingface.co/anon-review-meld-2026/meld) and code (https://anonymous.4open.science/r/MELD-4D74/README.md) are available anonymously.

1 INTRODUCTION

As large language models become common in everyday writing, reliable AI-text detection is increasingly important for academic integrity, content moderation, and provenance. In real-world use, overall accuracy alone is not enough. False positives can carry serious consequences for human authors, including accusations of academic misconduct and unfair penalties for non-native English writers (Liang et al., 2023). Simple paraphrasing and rewriting strategies have also been shown to evade or destabilize existing detectors (Krishna et al., 2023; Hu et al., 2023; Wu et al., 2024).

Current methods fall into three categories (Section 2): 1) training-free detectors use token-rank, likelihood-curvature, or cross-perplexity signals from reference language models (Gehrmann et al., 2019; Mitchell et al., 2023; Bao et al., 2024; Hans et al., 2024); 2) supervised encoder detectors learn binary classifiers from labeled examples through a single binary objective (Solaiman et al., 2019; Guo et al., 2023; Drayson et al., 2025); and 3) a newer line incorporates fine-grained authorship or generator structure through contrastive learning, auxiliary supervision, representation analysis, and out-of-distribution detection (Guo et al., 2024; Ta et al., 2026; Chen et al., 2025b; Zeng et al., 2025). These advances improve benchmark performance, but high overall AUROC can hide poor performance on rare, difficult, or unseen generators. A pooled AI-versus-human objective treats all generators as one "AI" class, providing no explicit supervision for generator- or rewrite-specific differences. As a result, the model may learn features sufficient for aggregate detection without capturing the finer structure needed for attribution and adaptation to new families. Strong performance

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.

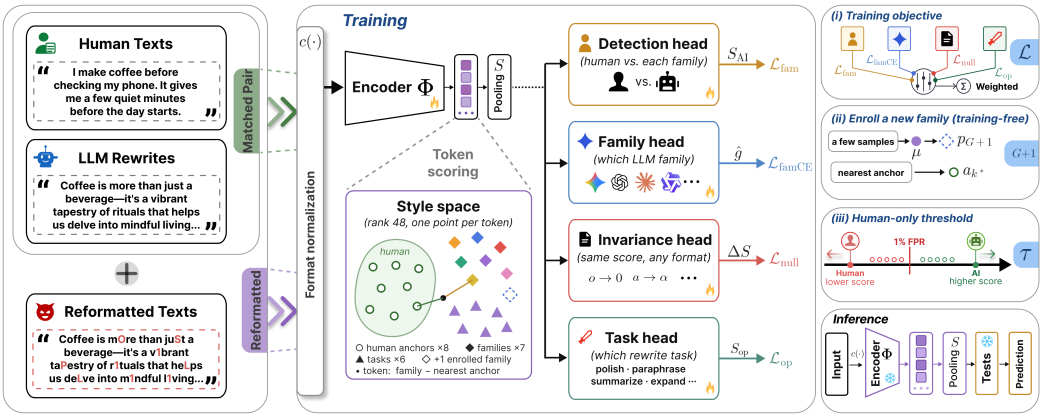

Figure 1: **Overview of MELD.** *Left:* each human text passage is paired with an AI rewrite of the same content and with reformatted copies, and every input undergoes format normalization first. *Middle:* a shared encoder maps tokens into a style space with eight human anchors and one center per generator family; the most AI-like tokens are pooled into the document score, and four heads are trained for family-balanced detection, family attribution, format invariance, and rewrite-task attribution. *Right:* the four losses form one objective, a new family is enrolled from a few samples without retraining, and the threshold is set from human text alone; at inference, one pass gives the detection score and the family, task, and span-level predictions.

on clean text also does not ensure reliable detection at low false-positive rates after rewriting or reformatting (Dugan et al., 2024). Three challenges therefore remain: combining detection with family and task attribution, maintaining robust detection at low false-positive rates under attack, and generalizing to unseen generators while enabling efficient enrollment of new families.

We propose MELD (Multi-Task Equilibrated Learning Detector), a detector that trains one class-balanced test per generator family instead of a single pooled AI-versus-human classifier. Each test distinguishes that family's text from human writing, and the family losses receive equal weight, making each family's detection task an explicit part of training. All tests share one model of human writing: a fine-tuned encoder maps each token into a small style space with several human anchors and one learned center per family. Each token receives a score based on its distances to a family center and the nearest human anchor, and a document score averages its most AI-like tokens (Figure 1). The same head is trained for generator-family and rewrite-task attribution and also provides predictions of AI-written spans. To support robust detection at low FPR, we train on AI rewrites, encourage consistent scores across formatting variants, and use format normalization to make scores invariant to different formatting changes. Detection thresholds are calibrated from representative human text alone. To encourage generalization, the training data include 1.25M AI rewrites of pre-2021 arXiv passages that match human passages in topic, field, length, and mathematical content, helping the detector learn how a text is written rather than what it is about. The family-center design also allows a new generator family to be enrolled from a few examples without retraining or changing the existing family tests. Our main contributions are as follows:

- **Multi-task family-aware detection.** MELD uses class-balanced, family-specific detection tasks over a shared model of human writing, while jointly supporting family attribution, rewrite-task prediction, and AI-span localization. Its scoring design also supports format invariance, human-only threshold calibration, and few-shot enrollment of new families.
- **MELD-eval.** We introduce a held-out evaluation pool built from four current-generation chat models, paired human text, and RAID-style domains and attacks to test zero-shot transfer to unseen generators.
- **Strong and robust performance.** MELD ranks first among open-source systems on the RAID leaderboard, achieves 99.7% TPR at 1% FPR on MELD-eval without further fine-tuning, and reaches 92% top-1 accuracy for seven-way family attribution. We further apply MELD to 5.9M scientific texts from 2016–2026.

Table 1: **RAID public leaderboard** (`https://raid-bench.xyz`, accessed 2026-09-23). AUROC and TPR at 5%/1% FPR ($\times 100$) on the official RAID test set. "All settings" includes RAID's attack suite; "No attack" is the clean subset. Commercial rows are public product submissions; open-source rows are leaderboard submissions with a paper and public model or code. "—": attacked scores unavailable. **Best**/second-best per column.

	All settings (with attack)			No attack		
Detector	**AUROC**	**TPR@5%**	**TPR@1%**	**AUROC**	**TPR@5%**	**TPR@1%**
Commercial detectors (closed-source)						
Grammarly (n.d.)	**99.87**	99.47	98.21	**99.97**	**99.91**	**99.69**
QuillBot (n.d.)	99.62	99.62	98.10	99.78	99.72	99.24
GPTZero (Adam et al., 2026)	99.25	97.24	92.93	99.53	98.37	95.72
Open-source detectors						
ModernBERT-Detect	97.65	94.14	88.23	99.12	99.11	98.18
Binoculars	—	—	—	84.40	78.98	69.54
RADAR	81.92	63.91	43.12	81.89	65.61	48.09
GLTR	70.90	51.48	36.48	72.68	59.69	45.57
LLMDet (Wu et al., 2023)	62.90	26.70	14.91	65.92	33.40	20.16
MELD (ours)	99.82	**99.78**	**99.24**	99.85	99.76	99.40

2 Related work

Training-free detectors. Training-free methods usually score text under one or more reference language models (LMs) and use token statistics, likelihood geometry, or cross-model discrepancies as evidence of generation. GLTR (Gehrmann et al., 2019) uses token-rank statistics. DetectGPT (Mitchell et al., 2023) and Fast-DetectGPT (Bao et al., 2024) rely on likelihood curvature. Binoculars (Hans et al., 2024) compares cross-perplexities from two LMs. These detectors are easy to deploy because they do not require detector-specific training, but their behavior is tied to the coverage and calibration of the reference models, making them sensitive to paraphrase and surface perturbations (Krishna et al., 2023; Dugan et al., 2024).

Supervised detectors. Supervised methods train discriminative models from labeled human and AI text. Early studies finetuned RoBERTa-style encoders (Solaiman et al., 2019; Guo et al., 2023). Subsequent work improved this recipe with structured features (Verma et al., 2024), adversarial paraphrasing (Hu et al., 2023), stronger encoder backbones (Warner et al., 2025; Drayson et al., 2025), representation-based detection (Chen et al., 2025b; Fu et al., 2025), and one-class objectives (Zeng et al., 2025). While these methods can perform well on in-distribution benchmarks, they are typically trained with a single binary head, so their pooled scores can mask poor performance on individual generator families.

Generator structure in detectors. Recent detectors use generator-aware contrastive learning (Guo et al., 2024), auxiliary tasks (Ta et al., 2026), few-shot style representations (Rivera Soto et al., 2024), family-aware mixtures of experts (Shi et al., 2025), semantic subspaces (Xiong et al., 2026), or prototypes for both classes under a pooled decision (Xu et al., 2026). MELD instead lets family identity define the tests, with a multi-anchor human model shared across them, and follows established practice for its detection-plus-attribution predictions (Shi et al., 2025; Ta et al., 2026). Group-DRO reweights weak groups (Sagawa et al., 2020), as do fairness objectives for visual deepfake detection (Ju et al., 2024); MELD keeps the group structure with uniform weights (Section 3.3).

Invariance and behavioral tests. Homoglyph attacks can change text-detector verdicts without changing appearance (Wolff & Wolff, 2020; Creo & Pudasaini, 2025). Learned normalization for equivariance (Kaba et al., 2023; Mondal et al., 2023), scale-space equivariance (Worrall & Welling, 2019), transformation-invariant pooling (Laptev et al., 2016), and robust encodings for typos (Jones et al., 2020) assume group-structured or enumerated transformation sets; MELD's format normalization absorbs the modeled formatting changes (Section 3.4). Counterfactual invariance formalizes stability under changes to spurious attributes (Veitch et al., 2021; Fawkes & Evans, 2023), and human originals paired with AI rewrites appear in ImBD (Chen et al., 2025a), rewrite-distance

detection (Hao et al., 2025), and matched-rewrite benchmarks (Perrone & Romano, 2026; Bsharat et al., 2026). Behavioral tests separate changes that should and should not affect predictions (Ribeiro et al., 2020; Goodfellow et al., 2009; Lovering et al., 2021; Tiwari et al., 2026); our two contrasts apply both kinds of change to the same document (Section 5.3).

3 MELD - Model, Method and Deployment

3.1 Family-specific detection with a shared human model

Let x denote a document with label $y \in \{0,1\}$, where $y = 1$ indicates AI-generated text. Each AI document has a generator-family label $g(x) \in \{1,\ldots,G\}$. A family groups related generators, such as GPT or Llama; we map the 78 training generator identifiers to G=7 families. MELD learns one binary test per family, distinguishing that family's text from human writing. All tests share one human model. At inference, their outputs form one detection score. We track each test during training and evaluation and report the worst-performing family alongside aggregate results.

3.2 Style representations and scoring

We use Ettin-1B (Weller et al., 2026), a ModernBERT-family encoder (Warner et al., 2025) with hidden size 1792 and an 8192-token context. A lightweight head (under 0.01% of the encoder's parameters) projects tokens into a 48-dimensional *style space*. We jointly train the projection and encoder to distinguish human and AI writing (Section 3.3).

Human writing has M=8 learned anchors in this space, and each generator family has one center. Affinities are negative squared distances scaled by a learned temperature, with a learned bias per family. A token's family score is its affinity to that family's center minus its affinity to the nearest human anchor, allowing different human styles to match different anchors. We clip these scores to $[-\kappa,\kappa]$, with κ=30, to bound individual token contributions (Huber, 1965). A log-sum-exp over the G family scores gives a combined AI score for each token.

Document scores average the highest $q = \lceil \rho n \rceil$ scores among n valid tokens, with ρ=0.25. Pooling family-specific scores gives $S_{\text{fam}}[g]$; pooling combined scores gives S_{AI}. This multiple-instance rule (Durand et al., 2016) reduces the dilution of localized AI content by surrounding human text and requires only document-level supervision. Affinities to six learned task centers (generate, polish, paraphrase, humanize, summarize, expand) use the same pooling to give S_{op}. An energy score S_{energy} over all valid tokens (Liu et al., 2020) is low when representations are far from the learned anchors and centers. Appendix A gives the complete definitions.

3.3 Training objective

We train each pooled family score $S_{\text{fam}}[g]$ with binary cross-entropy, using that family's AI documents as positives and all human documents in the batch as negatives. We average the two class means with equal weight, then average the family losses: $\mathcal{L}_{\text{fam}} = \frac{1}{G}\sum_{g=1}^{G}\mathcal{L}_g$. Softmax cross-entropy trains family attribution from family scores on AI examples ($\mathcal{L}_{\text{famCE}}$), while six-way cross-entropy trains task scores ($\mathcal{L}_{\text{op}}$). On 15% of training examples, a consistency loss $\mathcal{L}_{\text{null}}$ applies a Huber penalty to the score difference between two randomly formatted versions of the same text, encouraging stable raw scores before format normalization. The full objective is

$$\mathcal{L} = \mathcal{L}_{\text{fam}} + 0.15\,\mathcal{L}_{\text{op}} + 0.15\,\mathcal{L}_{\text{famCE}} + 0.10\,\mathcal{L}_{\text{null}}. \tag{1}$$

Human passages and paired rewrites enter as individual examples, without a pairwise loss (Section 4). Matching topic, field, section role, length band, and mathematical content across labels encourages learning differences in the writing process. Appendix J ablates each loss.

3.4 Deployment properties

Format invariance. Look-alikes, invisible characters, whitespace, and markdown can alter predictions without substantially changing displayed text (Wolff & Wolff, 2020; Creo & Pudasaini, 2025). MELD normalizes look-alikes, removes invisible characters and markup, and collapses

whitespace. For modeled operations, $R_k = \text{format}_k \circ \text{normalize}$ satisfies $R_k \circ R_j = R_k$ for all j,k. Reformatted versions share one normalized form and score. Unlike group-based approaches (Kaba et al., 2023; Mondal et al., 2023), no group is required. We observed no violations in 484,396 compositions; format normalization completely absorbs RAID's homoglyph and zero-width attacks. Robustness to wording changes such as paraphrasing is learned from training data.

Bounded token influence. For a fixed-length token-score vector, changing at most m entries changes the pooled score by at most $2\kappa m/\lceil \rho n \rceil$ (Proposition 1, Appendix D).

Human-only threshold calibration. Target-FPR thresholds are human-score quantiles from a held-in calibration set, stored with the checkpoint. We recalibrate on representative human text for new deployment populations (Section 6); no AI examples are needed.

New-family enrollment. We enroll a family without retraining by initializing its center from about 64 documents' mean style representation and adjusting it away from the nearest human anchor (Appendix E). The encoder, human anchors, and existing tests remain unchanged. If the new family's token scores stay sufficiently below the existing combined scores, the increase in combined score is exponentially small (Proposition 2, Appendix D).

4 Data and training

Training mixture. One of our main efforts is to curate a comprehensive training corpus and MELD-eval, a held-out transfer benchmark (Appendix H). The training mixture combines our rewrites with public datasets: 10.30M examples, one schema, and seven generator families (Appendix B, Table 5). Nine human registers, including scientific prose, peer reviews, and news, define calibration groups and evaluation subsets. Perturbations keep source family labels; sampling rebalances family counts. Validation has 50,128 examples across all families. Deduplication against a fixed evaluation registry removes 104,660 near-duplicates; arXiv identifiers exclude training papers from the case study (Section 6).

Paired scientific rewrites. We construct the scientific rewrite corpus from a fixed snapshot of 3.09M papers' LaTeX sources. Every recorded paper version must predate 1 January 2021. Four open-weight models use nine operations, mapped to six training tasks (Appendix G). The 3.07M examples comprise 1,476,558 source–rewrite pairs and 115,155 unpaired generation examples. Paired passages share topic, field, and section role, meet length constraints, and preserve all equations and citation keys.

Training procedure. Length-bucketed batches with a fixed token budget support up to 8192 tokens per document and are 50% human. Random cropping, line-wrap randomization, markdown removal/insertion, and light post-editing apply to both labels to reduce formatting and editing cues. Encoder and head are jointly fine-tuned for one 7.06M-example sampling epoch on eight L40 GPUs (Appendix F).

5 Results and Discussion

5.1 Validation and unseen-family tests

On the 50,128-example validation split, MELD achieves 99.78 AUROC and 97.5% TPR at 1% FPR. Worst-family AUROC is 99.68, only 0.10 points below the overall value. The stored nominal-1% threshold yields 1.6% FPR on validation (5.5% for texts of at most 192 words), 0.7% on paired arXiv, 0.3% on the DeepSeek evaluation set, and 0.0–1.9% across registers. Enumeration-heavy human text, including author lists and bibliographies, stays below 0.1% FPR.

DeepSeek-V4-Flash and -Pro, whose family is absent from training, both reach 99.9 AUROC, with 98.0% and 98.3% TPR at 1% FPR. The in-mixture gpt-oss-20b control reaches 99.82 AUROC on the same source passages and rewrite pipeline (Appendix G). Holding these fixed shows strong transfer to the held-out generators under this protocol. On Beemo, whose AI text is expert-edited, MELD reaches 81.77 AUROC versus 51–69 for three public baselines (Appendix J).

Table 2: **Held-out pool AUROC** ($\times 100$). Baselines use public checkpoints or official code on the same pools, all excluded from MELD training (Section 4). Entries give the estimate $\pm$ half-width of a 95% bootstrap CI. **Best**/second-best per column; RoBERTa-Large is OpenAI's detector.

Detector	HC3	MAGE	M4GT	Ghostbuster	DetectRL	MELD-eval
Zero-shot detectors (re-evaluated on the same pools as MELD)						
GLTR	99.5±0.1	64.4±0.5	73.4±1.7	88.2±1.0	71.4±0.5	42.4±0.6
Fast-DetectGPT	99.1±0.1	57.1±0.5	65.9±1.5	92.6±0.8	73.0±0.5	70.5±0.4
Binoculars	79.4±0.4	60.7±0.5	57.3±1.8	75.4±1.4	64.8±0.6	45.2±0.7
Supervised baselines (re-evaluated)						
RoBERTa-Large	95.5±0.1	77.3±0.4	65.4±2.1	47.2±1.7	63.1±0.6	71.7±0.6
RoBERTa-ChatGPT	**100.0±0.0**	56.7±0.5	53.5±2.2	73.5±1.4	75.8±0.5	38.6±0.6
RADAR	55.0±0.4	61.2±0.5	71.7±2.1	73.4±1.9	76.5±0.6	76.1±0.6
ModernBERT-Detect	98.6±0.1	97.9±0.1	76.9±1.8	94.8±0.8	87.9±0.4	99.7±0.1
Human-as-Outlier	99.3±0.1	98.0±0.1	76.1±1.8	94.0±0.8	87.7±0.4	65.8±0.6
RepreGuard	95.2±0.1	47.3±0.5	75.9±1.8	89.0±1.2	69.4±0.6	93.9±0.3
MELD (ours)	99.90±0.01	**99.1±0.1**	**77.5±1.3**	**99.4±0.1**	**99.8±0.0**	**99.90±0.04**

Family attribution reaches 91.9% top-1 across seven families on 8,400 validation examples sampled by writing category, with at least 84.3% in every register, indicating consistent performance across the evaluated registers.

5.2 Public benchmarks

Baselines use official code or public checkpoints where available. AUROC is on a 0–100 scale; TPR thresholds are calibrated on each pool's human scores, rather than shared across pools. Tables 2 and 3 report 95% percentile bootstrap CI half-widths (B=1,000; Appendix C).

RAID leaderboard. On RAID (Dugan et al., 2024), MELD has the strongest open-source results and is competitive with commercial detectors (Table 1). Across all settings, including attacks, its mean AUROC, TPR@5%FPR, and TPR@1%FPR exceeds the next-best open-source mean by 6.3 points and the highest commercial mean by 0.4.

Table 2 compares reference-LM, supervised, and representation-based detectors on HC3 (Guo et al., 2023), MAGE (Li et al., 2024), M4GT (Wang et al., 2024), Ghostbuster (Verma et al., 2024), and DetectRL (Wu et al., 2024), all excluded from training (Section 4). MELD has the highest AUROC point estimate on four of five pools. On HC3, which is closer to RoBERTa-ChatGPT's training distribution, that detector reaches 100.0 versus MELD's 99.90. On M4GT, MELD's 77.5 ± 1.3 and the next-best 76.9 ± 1.8 have overlapping intervals.

MELD-eval. GPT-5.4-Mini, Gemini-3-Flash, Claude-Haiku-4.5, and Qwen-3.6-Plus (OpenAI, 2026; Google DeepMind, 2025; Anthropic, 2025; Alibaba Cloud, 2026) generate text under one no-markdown instruction across eight RAID English domains, followed by RAID-style attacks. The constructed pool has 7,862 paired human texts, 31,448 clean AI examples, and 188,688 attacked examples (Appendix H).

Without further fine-tuning, MELD reaches at least 99.3% TPR at 1% FPR per generator and 99.7% overall (Table 3). ModernBERT-Detect is the strongest baseline at 95.5% overall under the same calibration protocol; most others are substantially lower. MELD is also more consistent across generators. Appendix I gives HC3, DetectRL, and MELD-eval examples.

5.3 Process and surface sensitivity

Evasion tests measure AI text escaping detection (Wolff & Wolff, 2020; Creo & Pudasaini, 2025; Dugan et al., 2024); they do not directly test human false positives under reformatting or compare the same edits across both classes. On held-out documents, the *process contrast* $\Delta_p = S(x^{\mathrm{ai}}) - S(x^{\mathrm{h}})$ compares content-matched AI rewrites and human sources. The *surface contrast* $\Delta_s = S(R_k x) - S(x)$ compares eight formats, including plain text, headings, bold, lists, fenced code, tables, and chat

Table 3: **MELD-eval TPR@1%FPR** ($\times 100$), per generator and overall, against paired human texts. All detectors are zero-shot for these generators. Entries give the estimate $\pm$ half-width of a 95% bootstrap CI. **Best**/second-best per column.

Detector	GPT-5.4-Mini	Gemini-3-Flash	Claude-Haiku-4.5	Qwen-3.6-Plus	Overall
Zero-shot reference-LM detectors					
GLTR	1.2±0.3	0.5±0.3	1.3±0.4	0.3±0.1	0.8±0.3
Fast-DetectGPT	9.2±1.7	19.8±2.1	14.3±2.0	24.7±2.3	17.0±2.0
Binoculars	0.2±0.1	0.6±0.1	1.1±0.2	0.7±0.2	0.6±0.1
Supervised baselines (re-evaluated)					
RoBERTa-Large	1.1±0.1	1.1±0.1	1.8±0.1	1.8±0.1	1.4±0.0
RoBERTa-ChatGPT	0.5±0.1	1.2±0.2	0.5±0.2	0.9±0.2	0.8±0.2
RADAR	0.4±0.2	1.4±0.5	2.2±0.8	1.2±0.5	1.3±0.5
Human-as-Outlier	2.7±1.2	1.6±0.9	0.8±0.4	1.1±0.8	1.6±0.8
ModernBERT-Detect	97.2±1.0	93.0±2.2	98.2±0.9	93.7±2.0	95.5±1.5
RepreGuard	23.0±1.3	27.2±1.4	41.2±1.8	45.4±2.0	34.2±1.6
MELD (ours)	**99.3**±0.3	**99.8**±0.4	**99.7**±0.5	**100.0**±0.2	**99.7**±0.3

Table 4: **Two contrasts: 400 held-out arXiv pairs $\times$ 8 formats.** $\pm$: 95% bootstrap CI half-width (B=1,000). Worst FPR (%) uses each detector's 1%-FPR threshold before/after format normalization. RoBERTa-Large is OpenAI's detector. **Best**/second-best: AUROC, SAR, FPR.

Detector	Pair AUROC↑	med. $\lvert\Delta_p\rvert$	med. $\lvert\Delta_s\rvert$	SAR↑	Raw FPR↓	Norm. FPR↓
RoBERTa-ChatGPT	63.85	0.89±0.19	0.35±0.03	2.5±0.6	4.3	**0.8**
RoBERTa-Large	63.56	0.97±0.12	0.54±0.02	1.8±0.2	**0.5**	**0.8**
ModernBERT-Detect	81.67	1.64±0.20	1.87±0.08	0.9±0.1	7.7	**0.8**
MELD (ours)	**99.97**	10.27±0.16	0.32±0.02	**31.7**±**2.3**	1.8	1.0

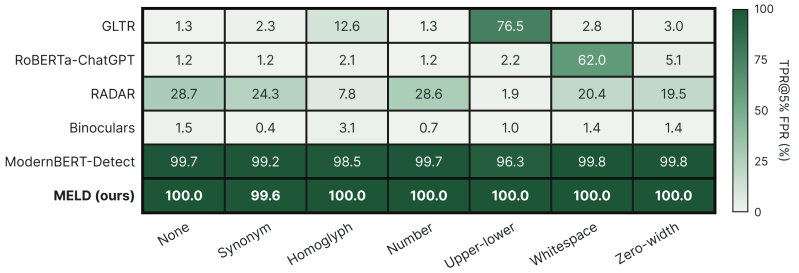

Figure 2: **Per-attack detection on MELD-eval.** TPR@5%FPR for the clean setting and each of the six attack types, for MELD and five baselines; every detector's threshold is set once on the paired human rows and held fixed across attacks.

layouts, with content and authorship fixed. Their dimensionless ratio, SAR $=$ median$|\Delta_p|$/median$|\Delta_s|$, applies to any scalar detector, including black boxes. Below one, median absolute score changes are larger for formatting than rewriting.

With format normalization disabled, Table 4 compares 400 held-out arXiv pairs across eight formats. MELD's SAR is 31.7, and every AI rewrite outscores its source. ModernBERT-Detect's SAR is 0.9, with worst-format human FPR of 7.7%: detection accuracy does not ensure format stability. Format normalization yields worst-format FPRs of 0.8–1.0% across all four detectors. MELD's deployed pipeline gives $\Delta_s = 0$, so SAR is undefined (Appendix C). Figure 2 complements these contrasts with per-attack detection under a fixed threshold.

5.4 Representation analysis

We analyze 144,000 embeddings per detector (Figures 3 and 4). MELD's between-/within-generator distance ratio is 2.2 versus the best baseline's 0.7; under attack, its within-/between-source ratio is 0.14 versus the nearest baseline's 0.62. The standardized within-/between-source distance gap

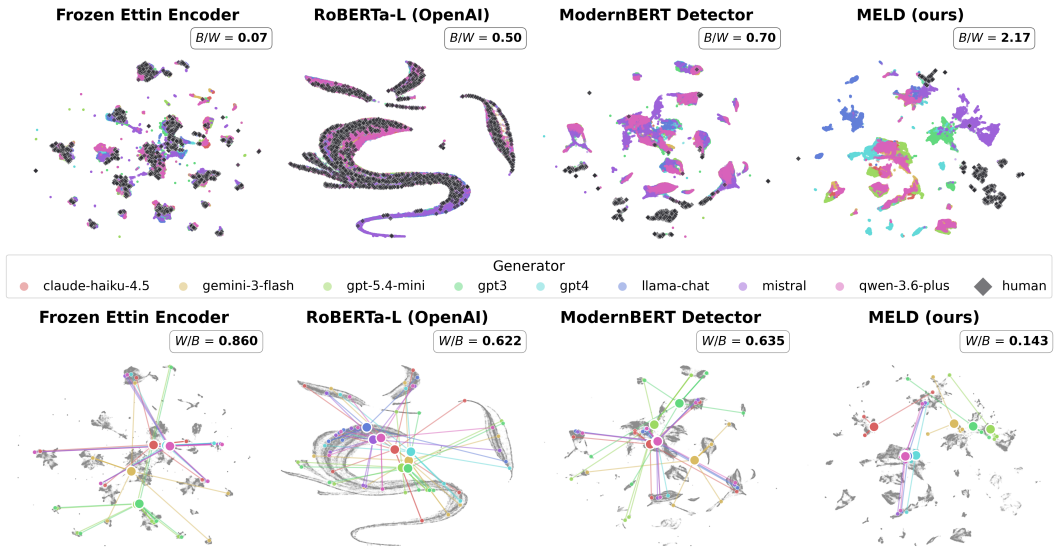

Figure 3: **Backbone geometry.** UMAP of 144,000 embeddings per panel. *Top:* generator colors, with human text in black; between-/within-generator distance B/W is 2.17 for MELD and at most 0.70 for baselines. *Bottom:* Original text joined to the attacked variants; within-/between-source distance W/B is 0.14 for MELD versus 0.62 for the nearest baseline.

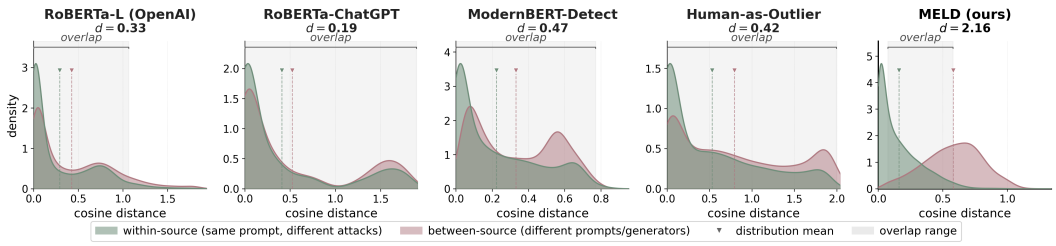

Figure 4: **Distance-space geometry.** Within-source and between-source cosine-distance distributions; d is their standardized gap (MELD 2.16, next ModernBERT-Detect 0.47).

is d=2.16 versus ModernBERT-Detect's 0.47. These results are consistent with greater generator separation and stability under the evaluated attacks.

5.5 Limitations

Evaluation covers English documents that are fully AI-generated or AI-edited; other languages and writing settings require corresponding data. TPR-at-FPR results use each pool's human-score thresholds, so a stored threshold need not retain its FPR across domains, registers, or lengths. Deployment requires calibration on representative human text (Section 6). Format normalization guarantees exact invariance to the modeled formatting transformations. Robustness to wording changes, including paraphrasing and humanization, is learned from training data and evaluated empirically on rewrite and attack benchmarks.

6 Case Study: Scientific Writing and Peer Review

We use MELD's scores and task predictions to track signs of AI assistance across years, fields, paper sections, and peer reviews, examining whether flagged text resembles AI editing or generation from scratch. We analyze 5.9 million texts: 5,682,241 section units from 1,344,610 arXiv papers (2016–2026), 92,174 accepted abstracts from nine conferences, and 143,685 review summaries. We keep the MELD checkpoint with nine operation labels fixed, exclude training papers by identifier, and match review fields across years (Appendix K). Separate 2016–2022 baselines for each corpus,

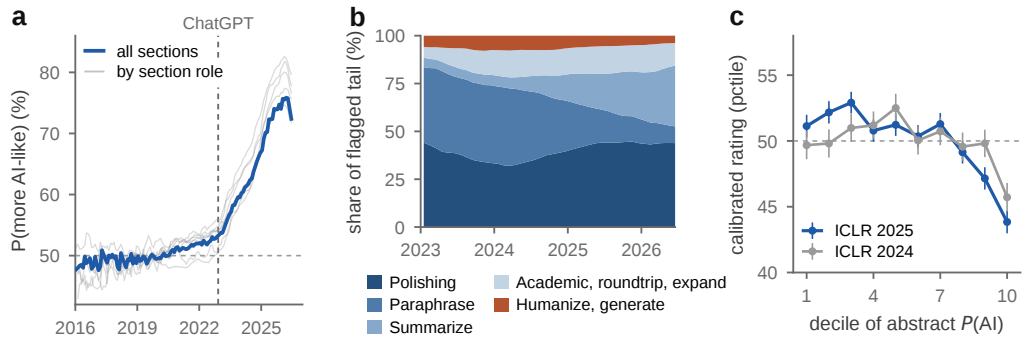

Figure 5: **Score trends, predicted writing tasks, and review ratings.** *(a)* Monthly probability that an arXiv section receives a higher MELD score than a baseline section of the same type. *(b)* Predicted writing tasks among flagged arXiv sections, shown as percentages over time. *(c)* Average ICLR review ratings for papers divided into ten groups by their abstracts' $P(\text{AI})$ scores, from lowest to highest. Ratings are converted to percentiles within each year, and error bars show one standard error.

section type, venue, and field define score percentiles; texts above the 99th percentile are *flagged*. $P(\text{AI}) = \sigma(S_{\text{AI}})$ rescales the score; it is not the probability that the text was written by AI.

Changes over time. Scores stay relatively stable through 2022 and rise from 2023 (Table 12; Figure 5a). The flagged arXiv share grows from 4.1% in 2023 to 25.4% in January–July 2026, about 24 times baseline, with a median shift of 0.72 baseline standard deviations (SD). Equal yearly samples and first submissions retain this pattern (Appendix K), suggesting that sample growth or revisions alone do not explain it. For most flagged texts, MELD predicts polishing or paraphrasing; fewer than 2% are assigned generation from scratch (Figure 5b). In some AI-related fields, scores plateau or fall in January–July 2026 as energy scores change, possibly reflecting changing generator styles rather than less AI use.

Sections and authors. By 2026, conclusions, discussions, and related work shift 1.6–3.4 baseline SD, versus less than one for methods, results, and abstracts. Timing is similar across fields and nine venues. Among 23,404 first authors publishing before 2023 and from 2023 onward, 77% have higher later score percentiles (Appendix K).

Peer review and outcomes. In 2025, median review-summary scores exceed those of 87–88% of matched baseline reviews; ICLR summaries are flagged 1.7 times as often as arXiv units using separate baselines. Flagged and unflagged reviews give similar rating percentiles within venue and year. Papers whose abstracts have the highest detector scores receive lower ratings (Figure 5c); acceptance is roughly flat across the first nine abstract-score deciles and lower in the highest (Appendix K).

7 Conclusion

We introduced MELD, a multi-task model that detects AI-generated text and predicts its generator family and rewrite task. It provides document- and token-level predictions in a single forward pass, remains invariant to the formatting changes, and allows new generator families to be enrolled without large retraining. We also introduced MELD-eval, a held-out benchmark covering four frontier chat models and different attacks. MELD ranks first among open-source submissions on the RAID leaderboard and remains competitive with commercial systems. It achieves 99.7% TPR at 1% FPR on MELD-eval without further fine-tuning and 98% TPR at the same FPR on a generator whose family is absent from training. Finally, a case study of 5.9 million scientific texts from 2016–2026 finds that detector scores remain relatively stable through 2022 and increase from 2023 onward.

AI USE STATEMENT

Generative models are the object of study and part of the method in this work: four open-weight instruction models generate the paired rewrite corpus (Appendix G), four API chat models generate the MELD-eval pool (Appendix H), and two held-out models generate the unseen-family test set (Section 5.1); prompts, decoding settings, and acceptance gates are documented in those appendices. Beyond these documented roles, we used AI assistants to polish the language of the manuscript. All AI-assisted text was reviewed by the authors, and every reported number was verified against the underlying experiment outputs. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses public corpora and AI text generated for the study; it involves no human-subjects experiments, and the paper and peer-review corpora are public documents analyzed at the population level. The main ethical risk of AI-text detection is the false positive. Section 1 motivates low-FPR operation, thresholds are population-specific quantiles of human scores, and deployments in critical settings such as academic integrity calibrate on their own population, since detectors can flag non-native English writers more often (Liang et al., 2023). The case study reports population-level statistics.

REPRODUCIBILITY STATEMENT

The architecture and objective are specified in Section 3, the training mixture and epoch plan in Section 4, and the full optimization settings in Appendix F. The rewrite-corpus and MELD-eval constructions, including prompts, decoding settings, and acceptance gates, are documented in Appendices G and H, and complete proofs of both propositions are in Appendix D. The trained checkpoint, the MELD-eval pool, a self-contained scoring snippet, and the source code are available anonymously through the links in the abstract; the stored human-quantile thresholds ship inside the checkpoint.

REFERENCES

George Alexandru Adam, Alexander Cui, Edwin Thomas, Emily Napier, Nazar Shmatko, Jacob Schnell, Jacob Junqi Tian, Alekhya Dronavalli, Edward Tian, and Dongwon Lee. GPTZero: Robust detection of LLM-generated texts. *arXiv preprint arXiv:2602.13042*, 2026. URL https://arxiv.org/abs/2602.13042.

Alibaba Cloud. Qwen 3.6 Plus, 2026. URL https://www.alibabacloud.com/help/en/model-studio/model-pricing. API model snapshot qwen3.6-plus-2026-04-02.

Anthropic. Claude Haiku 4.5, 2025. URL https://platform.claude.com/docs/en/about-claude/models/model-ids-and-versions. API model snapshot claude-haiku-4-5-20251001.

Ekaterina Artemova, Jason S. Lucas, Saranya Venkatraman, Jooyoung Lee, Sergei Tilga, Adaku Uchendu, and Vladislav Mikhailov. Beemo: Benchmark of expert-edited machine-generated outputs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6992–7018, 2025. doi: 10.18653/v1/2025.naacl-long.357.

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.

Sondos Mahmoud Bsharat, Jiacheng Liu, Xiaohan Zhao, Tianjun Yao, Xinyi Shang, Yi Tang, Jiacheng Cui, Ahmed Elhagry, Salwa K. Al Khatib, Hao Li, Salman Khan, and Zhiqiang Shen. Operation-guided progressive human-to-AI text transformation benchmark for multi-granularity AI-text detection. *arXiv preprint arXiv:2606.06481*, 2026. URL https://arxiv.org/abs/2606.06481.

Jiaqi Chen, Xiaoye Zhu, Tianyang Liu, Ying Chen, Chen Xinhui, Yiwen Yuan, Chak Tou Leong, Zuchao Li, Long Tang, Lei Zhang, Chenyu Yan, Guanghao Mei, Jie Zhang, and Lefei Zhang. Imitate before detect: Aligning machine stylistic preference for machine-revised text detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(22):23559–23567, 2025a. doi: 10.1609/aaai.v39i22.34525.

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, 2025b. doi: 10.1162/tacl.a.61.

Aldan Creo and Shushanta Pudasaini. SilverSpeak: Evading AI-generated text detectors using homoglyphs. In *Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect)*, pp. 1–46, 2025. URL https://aclanthology.org/2025.genaidetect-1.1/.

George Drayson, Emine Yilmaz, and Vasileios Lampos. Machine-generated text detection prevents language model collapse. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 29657–29673, 2025. doi: 10.18653/v1/2025.emnlp-main.1506.

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

Thibaut Durand, Nicolas Thome, and Matthieu Cord. WELDON: Weakly supervised learning of deep convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4743–4752, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/Durand_WELDON_Weakly_Supervised_CVPR_2016_paper.html.

Jake Fawkes and Robin J. Evans. Results on counterfactual invariance. *arXiv preprint arXiv:2307.08519*, 2023. URL https://arxiv.org/abs/2307.08519.

Jiachen Fu, Chun-Le Guo, and Chongyi Li. DetectAnyLLM: Towards generalizable and robust detection of machine-generated text across domains and models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 11229–11238, 2025. doi: 10.1145/3746027.3754862.

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

Ian Goodfellow, Quoc Le, Andrew Saxe, Honglak Lee, and Andrew Y. Ng. Measuring invariances in deep networks. In *Advances in Neural Information Processing Systems*, volume 22, 2009. URL https://papers.nips.cc/paper_files/paper/2009/hash/428fca9bc1921c25c5121f9da7815cde-Abstract.html.

Google DeepMind. Gemini 3 Flash, 2025. URL https://ai.google.dev/gemini-api/docs/changelog. API model gemini-3-flash-preview, released 2025-12-17.

Grammarly. Grammarly AI writing detector, n.d. URL https://www.grammarly.com/ai-detector. Commercial product; accessed 2026-09-13. Leaderboard scores are from RAID (https://raid-bench.xyz), as dated in Table 1.

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.

Xun Guo, Shan Zhang, Yongxin He, Ting Zhang, Wanquan Feng, Haibin Huang, and Chongyang Ma. DeTeCtive: Detecting AI-generated text via multi-level contrastive learning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 88320–88347, 2024. doi: 10.52202/079017-2802.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Spotting LLMs with Binoculars: Zero-shot detection of machine-generated text. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 17519–17537, 2024. 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, 2025. doi: 10.18653/v1/2025.acl-long.322.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. RADAR: Robust AI-text detection via adversarial learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 15077–15095, 2023. doi: 10.52202/075280-0662.

Peter J. Huber. A robust version of the probability ratio test. *The Annals of Mathematical Statistics*, 36(6):1753–1758, 1965. doi: 10.1214/aoms/1177699803.

Erik Jones, Robin Jia, Aditi Raghunathan, and Percy Liang. Robust encodings: A framework for combating adversarial typos. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2752–2765, 2020. doi: 10.18653/v1/2020.acl-main.245.

Yan Ju, Shu Hu, Shan Jia, George H. Chen, and Siwei Lyu. Improving fairness in deepfake detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4655–4665, 2024. URL https://openaccess.thecvf.com/content/WACV2024/html/Ju_Improving_Fairness_in_Deepfake_Detection_WACV_2024_paper.html.

Minseok Jung, Cynthia Fuertes-Panizo, Liam Dugan, Yi R. Fung, Pin-Yu Chen, and Paul Pu Liang. Group-adaptive threshold optimization for robust AI-generated text detection. *arXiv preprint arXiv:2502.04528*, 2025. URL https://arxiv.org/abs/2502.04528.

Sékou-Oumar Kaba, Arnab Kumar Mondal, Yan Zhang, Yoshua Bengio, and Siamak Ravanbakhsh. Equivariance with learned canonicalization functions. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 15546–15566, 2023. URL https://proceedings.mlr.press/v202/kaba23a.html.

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, 2023. doi: 10.52202/075280-1195.

Dmitry Laptev, Nikolay Savinov, Joachim M. Buhmann, and Marc Pollefeys. TI-POOLING: Transformation-invariant pooling for feature learning in convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 289–297, 2016. doi: 10.1109/cvpr.2016.38.

Haoran Li and Quan Wang. Continual origin tracing of LLM-generated text. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 479–489, 2025. doi: 10.1145/3726302.3729935.

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. MAGE: Machine-generated text detection in the wild. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 36–53, 2024. doi: 10.18653/v1/2024.acl-long.3.

Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. GPT detectors are biased against non-native English writers. *Patterns*, 4(7):100779, 2023. doi: 10.1016/j.patter.2023.100779.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, Daniel A. McFarland, and James Y. Zou. Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 29575–29620, 2024a. URL https://proceedings.mlr.press/v235/liang24b.html.

Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, Diyi Yang, Christopher Potts, Christopher D. Manning, and James Y. Zou. Mapping the increasing use of LLMs in scientific papers. In *First Conference on Language Modeling*, 2024b. URL https://openreview.net/forum?id=YX7QnhxESU.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21464–21475, 2020. URL https://papers.nips.cc/paper_files/paper/2020/hash/f5496252609c43eb8a3d147ab9b9c006-Abstract.html.

Charles Lovering, Rohan Jha, Tal Linzen, and Ellie Pavlick. Predicting inductive biases of pretrained models. In *The Ninth International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=mNtmhaDkAr.

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*, pp. 24950–24962, 2023. URL https://proceedings.mlr.press/v202/mitchell23a.html.

Arnab Kumar Mondal, Siba Smarak Panigrahi, Sékou-Oumar Kaba, Sai Rajeswar, and Siamak Ravanbakhsh. Equivariant adaptation of large pretrained models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 50293–50309, 2023. doi: 10.52202/075280-2187.

OpenAI. GPT-5.4 Mini, 2026. URL https://developers.openai.com/api/docs/models/gpt-5.4-mini. API model snapshot gpt-5.4-mini-2026-03-17.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems*, volume 37, pp. 30811–30849, 2024. doi: 10.52202/079017-0970.

Gaetano Perrone and Simon Pietro Romano. ARB: A matched authorship-rewriting benchmark dataset for AI-text detector evaluation. *arXiv preprint arXiv:2607.29539*, 2026. URL https://arxiv.org/abs/2607.29539.

QuillBot. QuillBot AI content detector, n.d. URL https://quillbot.com/ai-content-detector. Commercial product; accessed 2026-09-13. Leaderboard scores are from RAID (https://raid-bench.xyz), as dated in Table 1.

Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10(1–3):19–41, 2000. doi: 10.1006/dspr.1999.0361.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, 2020. doi: 10.18653/v1/2020.acl-main.442.

Rafael Rivera Soto, Kailin Koch, Aleem Khan, Barry Chen, Marcus Bishop, and Nicholas Andrews. Few-shot detection of machine-generated text using style representations. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=cWiEN1plhJ.

Giuseppe Russo, Manoel Horta Ribeiro, Tim Ruben Davidson, Veniamin Veselovsky, and Robert West. The AI review lottery: Widespread AI-assisted peer reviews boost paper scores and acceptance rates. *Proceedings of the ACM on Human-Computer Interaction*, 9(7):1–28, 2025. doi: 10.1145/3757667.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *The Eighth International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Yuhui Shi, Yehan Yang, Qiang Sheng, Hao Mi, Beizhe Hu, Chaoxi Xu, and Juan Cao. PhantomHunter: Detecting unseen privately-tuned LLM-generated text via family-aware learning. *arXiv preprint arXiv:2506.15683*, 2025. URL https://arxiv.org/abs/2506.15683.

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.

Zhen Sun, Yifan Liao, Zhicong Huang, Jiaheng Wei, Cheng Hong, Yutao Yue, and Xinlei He. When new generators arrive: Lifelong machine-generated text attribution via ridge feature transfer. *arXiv preprint arXiv:2606.05626*, 2026. URL https://arxiv.org/abs/2606.05626.

Minh Ngoc Ta, Dong Cao Van, Duc-Anh Hoang, Minh Le-Anh, Truong Nguyen, My Anh Tran Nguyen, Yuxia Wang, Preslav Nakov, and Dinh Viet Sang. FAID: Fine-grained AI-generated text detection using multi-task auxiliary and multi-level contrastive learning. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3275–3296, 2026. doi: 10.18653/v1/2026.eacl-long.151.

Rishabh Tiwari, Aditya Tomar, Udbhav Bamba, Monishwaran Maheswaran, Heng Yang, Michael W. Mahoney, Kurt Keutzer, and Amir Gholami. Reward under attack: Analyzing the robustness and hackability of process reward models. *arXiv preprint arXiv:2603.06621*, 2026. URL https://arxiv.org/abs/2603.06621.

Adaku Uchendu, Zeyu Ma, Thai Le, Rui Zhang, and Dongwon Lee. TURINGBENCH: A benchmark environment for Turing Test in the age of neural text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 2001–2016, 2021. doi: 10.18653/v1/2021.findings-emnlp.172.

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations: Why and how to pass stress tests. In *Advances in Neural Information Processing Systems*, volume 34, pp. 16196–16208, 2021. URL https://papers.nips.cc/paper_files/paper/2021/hash/8710ef761bbb29a6f9d12e4ef8e4379c-Abstract.html.

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, 2024. doi: 10.18653/v1/2024.naacl-long.95.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, Alham Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. M4GT-Bench: Evaluation benchmark for black-box machine-generated text detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3964–3992, 2024. doi: 10.18653/v1/2024.acl-long.218.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2526–2547, 2025. doi: 10.18653/v1/2025.acl-long.127.

Orion Weller, Kathryn Ricci, Marc Marone, Antoine Chaffin, Dawn Lawrie, and Benjamin Van Durme. Seq vs Seq: An open suite of paired encoders and decoders. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=z5Mn8Rxi3l.

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

Daniel E. Worrall and Max Welling. Deep scale-spaces: Equivariance over scale. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://papers.nips.cc/paper_files/paper/2019/hash/f04cd7399b2b0128970efb6d20b5c551-Abstract.html.

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, 2024. doi: 10.52202/079017-3186.

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

Shiyu Wu, Shuyan Li, Jing Li, Jing Liu, and Yequan Wang. Few-shot synthetic image attribution: Identifying unseen generators with limited samples. In *European Conference on Computer Vision (ECCV)*, 2026. URL https://arxiv.org/abs/2509.25682. Accepted paper; preprint arXiv:2509.25682.

Ruochong Xiong, Qien Li, Wangwang Lian, Yulong Wan, Hanlin Xue, Zhouxing Tan, Han Yang, Fengyu Lu, and Junfei Liu. Verifiable LLM-generated text detection via projected semantic-structural distributions. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14005–14042, 2026. doi: 10.18653/v1/2026.acl-long.638.

Yan Xu, Wenzhong Yang, Yabo Yin, Hongzhen Lv, Zhenhua Wang, Jingfeng He, Xiangyi Jia, and Xianfeng Wang. SAMP: Source-aware multi-prototype learning for machine-generated text detection. *Research Square preprint rs-9598516*, 2026. doi: 10.21203/rs.3.rs-9598516/v1.

Cong Zeng, Shengkun Tang, Yuanzhou Chen, Zhiqiang Shen, Wenchao Yu, Xujiang Zhao, Haifeng Chen, Wei Cheng, and Zhiqiang Xu. Human texts are outliers: Detecting LLM-generated texts via out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 38, pp. 181393–181423, 2025. doi: 10.52202/085713-5457.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. WildChat: 1M ChatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Bl8u7ZRlbM.

Xiaowei Zhu, Yubing Ren, Yanan Cao, Xixun Lin, Fang Fang, and Yangxi Li. Reliably bounding false positives: A zero-shot machine-generated text detection framework via multiscaled conformal prediction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12298–12319, 2025. doi: 10.18653/v1/2025.acl-long.601.

A SCORING GEOMETRY AND OBJECTIVE

Let x be a document. MELD combines G family-specific detection tests that share a human null hypothesis. For AI-generated text, $g(x) \in \{1,\dots,G\}$ denotes its generator family. The tests compare

$$\mathcal{H}_0 : x \text{ is human-written} \qquad \text{against} \qquad \mathcal{H}_1^{(g)} : x \text{ is generated by family } g, \quad g = 1,\dots,G.$$

Let $\Phi : \mathcal{X} \to \mathbb{R}^{L\times H}$ be a bidirectional encoder that produces token representations $h_\ell \in \mathbb{R}^H$. We use Ettin-1B (Weller et al., 2026), a ModernBERT-family encoder (Warner et al., 2025) with hidden size $H = 1792$ and a context length of 8192 tokens. A learned projection B maps each token representation into a style space of dimension $r = 48$:

$$u_\ell = \mathrm{LayerNorm}(B^\top h_\ell) \in \mathbb{R}^r.$$

The head contains 87,134 parameters, less than 0.01% of the backbone's parameter count.

Human writing is represented by $M = 8$ anchors a_k, and each generator family has a center p_g. Their affinities to a style representation u are

$$\varphi_k^{\mathrm{h}}(u) = -\tau\|u - a_k\|^2,$$
$$\varphi_g^{\mathrm{g}}(u) = -\tau\|u - p_g\|^2 + b_g,$$

where τ is a learned temperature and b_g is a learned family-specific bias. The human affinity is the affinity to the nearest human anchor:

$$\mathrm{hum}(u) = \max_{k\le M} \varphi_k^{\mathrm{h}}(u).$$

We define the family-specific and combined token scores as

$$\begin{aligned} s_{\ell g} &= \mathrm{clip}\big(\varphi_g^{\mathrm{g}}(u_\ell) - \mathrm{hum}(u_\ell), -\kappa, \kappa\big), \\ c_\ell &= \tau_{\mathrm{agg}}\,\mathrm{LSE}_{g\le G}\big(s_{\ell g}/\tau_{\mathrm{agg}}\big), \qquad \mathrm{LSE}_g\, z_g = \log\sum_g e^{z_g}. \end{aligned} \tag{2}$$

We use $\kappa = 30$ and $\tau_{\mathrm{agg}} = 1$. The margin $s_{\ell g}$ compares family g with the shared human model, while c_ℓ combines evidence across families. Clipping bounds each family-specific token margin (Huber, 1965). These affinity contrasts are a discriminatively trained analogue of the likelihood-ratio scores used with universal background models in speaker verification (Reynolds et al., 2000).

Document scores use top-quantile pooling (Durand et al., 2016). For a vector v of n valid token scores, let $I_q(v)$ index its $q = \lceil \rho n \rceil$ largest entries, where $\rho = 0.25$. Define

$$\mathrm{TQM}_\rho(v) = \frac{1}{q}\sum_{\ell\in I_q(v)} v_\ell.$$

The family score $S_{\mathrm{fam}}[g]$ applies this operation to the family-g token margins, and the detection score S_{AI} applies it to the combined token scores c_ℓ.

The same style space supports two additional outputs. Six task centers represent generate, polish, paraphrase, humanize, summarize, and expand. Their token affinities use the same pooling rule to produce $S_{\mathrm{op}} \in \mathbb{R}^6$. The energy score S_{energy} (Liu et al., 2020) averages log-sum-exp affinity over the n valid tokens:

$$S_{\mathrm{energy}} = \frac{1}{n}\sum_{\ell=1}^{n} \log\left(\sum_{g=1}^{G} e^{\varphi_g^{\mathrm{g}}(u_\ell)} + \sum_{k=1}^{M} e^{\varphi_k^{\mathrm{h}}(u_\ell)}\right).$$

Lower values indicate representations far from the learned human anchors and family centers.

For document i, the pooled family score $S_{\mathrm{fam}}[i,g]$ serves as the binary-classification logit. The loss for family g is

$$\begin{aligned} \mathcal{L}_g = \frac{1}{2}\Big[&\mathrm{mean}_{i\in P_g}\,\mathrm{BCE}(S_{\mathrm{fam}}[i,g], 1-\varepsilon) \\ &+ \mathrm{mean}_{i\in N}\,\mathrm{BCE}(S_{\mathrm{fam}}[i,g], 0)\Big], \end{aligned} \tag{3}$$

where P_g contains the batch's AI examples from family g and N contains its human examples. We use $\varepsilon = 0.02$ to smooth positive targets only. If a batch contains no examples from family g, the positive term is omitted. The family losses receive equal weight:

$$\mathcal{L}_{\text{fam}} = \frac{1}{G} \sum_{g=1}^{G} \mathcal{L}_g.$$

The term *equilibrated* refers to the class-balanced family losses. The auxiliary losses and their weights are given in Section 3.3.

Span and task predictions. When a token has zero margin for every family, its combined score is $\tau_{\text{agg}} \log G$. A two-state Viterbi decoder uses this neutral value to segment a mixed document into human and AI spans. Family and task scores provide the corresponding span attributions. No span-level labels are used during training. Appendix J evaluates mixed-document detection using pooled document scores. Writing $e_\ell = c_\ell - \tau_{\text{agg}} \log G$, the decoder uses emissions $-e_\ell$ for the human state and $+e_\ell$ for the AI state, with a cost of 4 for each state change.

Human-only thresholds. For a target false-positive rate, we select the corresponding quantile of human detection scores. The calibration pool contains human texts from nine writing categories (registers), including scientific papers, peer reviews, and news. We store the resulting thresholds in the checkpoint. No AI-generated examples are required. When the deployment population differs from the calibration pool, thresholds should be recomputed from representative local human text, as in Section 6.

The quantity $P(\text{AI}) = \sigma(S_{\text{AI}})$ is a monotone score transformation but not a calibrated authorship probability. Binary supervision is applied to S_{fam}, and the log-sum-exp aggregation can exceed the largest family margin by up to $\tau_{\text{agg}} \log G$ at each token. The measured expected calibration error is 6.5% under a balanced class prior.

B Training mixture and rewrite construction

Table 5 summarizes the training mixture by generator family. The raw corpus is imbalanced, and the epoch sampling plan reduces this imbalance. Rare families are oversampled toward a target of 100,000 draws, with total draws capped at four times the number of available examples in that family. Large families are capped at 1.1 million draws. Additional limits prevent any source, generator, or register from dominating a family's sampled examples. These limits can leave a concentrated family below the target, as for Claude in Table 5. Within each length bucket, human examples are balanced across the nine registers and make up half of each batch.

For the paired arXiv corpus, papers are divided into sections and then into prose windows of approximately 600 words. Equations and citation keys are temporarily replaced by numbered placeholders, such as `[MATH-1]` and `[CIT-1]`, before rewriting. The original equations and citation keys are restored exactly afterward. The nine operations range from light editing (*polish* and *academic*) to larger changes (*paraphrase*, *humanize*, *roundtrip*, *summarize*, and *expand*). The remaining operations, *generate from context* and *generate from abstract*, produce new section text using the paper's terminology.

Outputs must satisfy operation-specific length and copy-overlap limits, as well as checks for markdown, meta-commentary, repetition, and invented mathematics (Appendix G). After deduplication and removal of short-window copies, 1,246,160 source–rewrite pairs remain in the training index.

C Evaluation and ablation protocol

Surface-attribution ratio. SAR compares the median absolute process contrast with the median absolute surface contrast, as defined in Section 5.3. It is invariant to nonzero linear rescaling of the score, but not to arbitrary nonlinear transformations. No term in MELD's training objective directly optimizes the process contrast Δ_p.

Table 5: **Training mixture by generator family.** All family tests share the human examples. Public sources include RAID, MAGE, M4GT-Bench, DetectRL, HC-Var, WildChat (Zhao et al., 2024), and pre-2020 FineWeb (Penedo et al., 2024). Additional data include our newly constructed paired arXiv corpus, peer reviews, and imitations of student writing. HC3, Ghostbuster, and Beemo (Artemova et al., 2025) are held out. Family names are abbreviated.

Family test	Train rows	Val rows	Examples of raw tags	Epoch draws
Human (shared)	3,265,189	20,019	FineWeb, paired arXiv, reviews, essays	50%/batch
OpenAI GPT	2,565,898	2,877	gpt2 – gpt-4o, o1, gpt-oss-20b	681K
Other	1,753,988	7,796	mpt, cohere, unattributed, evasion tools	163K
Mistral	1,194,875	4,441	mistral, mistral-chat, ministral-3	788K
Gemini/Gemma	533,815	6,079	PaLM, gemini-3, gemma-4	780K
Qwen	491,549	3,116	qwen3.6-35b	711K
Llama	428,172	2,900	llama 1 – 3, llama-chat	352K
Claude	65,756	2,900	Claude-instant, claude-sonnet-4.6	56K
Total	10,299,242	50,128	78 raw generator tags	7.06M

Evaluation protocol. Unless stated otherwise, each evaluation uses the full held-out pool; MELD-eval is not subsampled. Thresholds for target false-positive rates are computed from the human examples in each pool. The reported operating points therefore measure separability under pool-specific calibration, rather than the transfer of a fixed threshold across populations. We emphasize low-FPR operating points because AUROC is already close to its maximum for several supervised detectors. Where reported, 95% percentile bootstrap confidence intervals use $B = 1{,}000$ resamples, and the $\pm$ values give their half-widths.

Ablation protocol. The ablation protocol removes one component at a time from the final configuration. The loss ablations remove family cross-entropy, format consistency, or task cross-entropy. Two data ablations compare removal of the pairing structure with a volume-matched reduction that retains complete pairs. Separate tests change the number of human anchors or the pooling rule and rescore the trained checkpoint without additional optimization. Appendix J describes the comparisons and reports the results.

D PROOFS

Proposition 1 (Bounded influence)**.** *Let $v, v' \in [c^-, c^+]^n$ be fixed-length token-score vectors, where $c^+ - c^- \le 2\kappa$, and let $q = \lceil \rho n \rceil$. If v and v' differ in at most m coordinates, then*

$$\left|\mathrm{TQM}_\rho(v') - \mathrm{TQM}_\rho(v)\right| \le \frac{m(c^+ - c^-)}{q} \le \frac{2\kappa m}{\lceil \rho n \rceil}.$$

Proof. Equation (2) gives $s_{\ell g} \in [-\kappa, \kappa]$ for every token and family. The combined token score therefore lies in

$$c_\ell \in [c^-, c^+], \qquad c^\pm = \pm\kappa + \tau_{\mathrm{agg}} \log G.$$

This interval has width 2κ.

Changing one coordinate by an amount of absolute size Δ changes the mean of the q largest coordinates by at most Δ/q. This holds whether the coordinate remains outside the selected set, remains inside it, or crosses its boundary. To move from v to v', change the differing coordinates one at a time. There are at most m such changes, each of size at most $c^+ - c^-$. Summing their effects gives

$$\left|\mathrm{TQM}_\rho(v') - \mathrm{TQM}_\rho(v)\right| \le \frac{m(c^+ - c^-)}{q} \le \frac{2\kappa m}{\lceil \rho n \rceil}.$$

□

Proposition 2 (Score stability under enrollment)**.** *Enrollment adds a new family center and bias while leaving all other parameters fixed. Consequently, the human affinity* $\mathrm{hum}(u_\ell)$ *and all existing family scores* $S_{\mathrm{fam}}[g]$*,* $g \le G$*, remain unchanged on every input. The combined token score can only increase, with*

$$c'_\ell - c_\ell = \tau_{\mathrm{agg}} \log\left(1 + \exp\left[\frac{s_{\ell,G+1} - c_\ell}{\tau_{\mathrm{agg}}}\right]\right) \le \tau_{\mathrm{agg}} \exp\left[\frac{s_{\ell,G+1} - c_\ell}{\tau_{\mathrm{agg}}}\right].$$

The document-score change satisfies

$$|S'_{\mathrm{AI}} - S_{\mathrm{AI}}| \le \max_\ell (c'_\ell - c_\ell).$$

In particular, if the new family's clipped margin satisfies $s_{\ell,G+1} \le c_\ell - \delta$ *at every token of a document, for* $\delta > 0$*, then*

$$|S'_{\mathrm{AI}} - S_{\mathrm{AI}}| \le \tau_{\mathrm{agg}} e^{-\delta/\tau_{\mathrm{agg}}}.$$

Proof. The enrollment rule in Equation (4) appends (p_{G+1}, b_{G+1}) without changing the encoder, projection, human anchors, or existing family parameters. Thus, u_ℓ, all human affinities, and all existing family affinities remain unchanged. The same is true of every existing token margin and pooled family score.

For the combined score, write

$$Z_\ell = \sum_{g \le G} e^{s_{\ell g}/\tau_{\mathrm{agg}}} = e^{c_\ell/\tau_{\mathrm{agg}}}.$$

After enrollment,

$$c'_\ell = \tau_{\mathrm{agg}} \log\left(Z_\ell + e^{s_{\ell,G+1}/\tau_{\mathrm{agg}}}\right) = c_\ell + \tau_{\mathrm{agg}} \log\left(1 + e^{(s_{\ell,G+1} - c_\ell)/\tau_{\mathrm{agg}}}\right).$$

The added term is nonnegative. Applying $\log(1+t) \le t$ for $t \ge 0$ gives the stated bound on each token-score change.

Let $\Delta_{\max} = \max_\ell (c'_\ell - c_\ell)$. Every coordinate increases by at most $\Delta_{\max}$. Top-quantile pooling is monotone, and adding the same constant to all coordinates adds that constant to the pooled score. It follows that $|S'_{\mathrm{AI}} - S_{\mathrm{AI}}| \le \Delta_{\max}$. Under the token-wise condition $s_{\ell,G+1} \le c_\ell - \delta$, we have $\Delta_{\max} \le \tau_{\mathrm{agg}} e^{-\delta/\tau_{\mathrm{agg}}}$. □

The enrollment adjustment in Appendix E is intended to keep the new family's margins below the existing combined scores on human text.

E NEW-FAMILY ENROLLMENT

We enroll a new family using $N_{\mathrm{e}} = 64$ documents from its generator, with a minimum of eight documents. Each document is encoded once, and its style representations are averaged over valid tokens. Let μ be the mean of these document vectors, and let ς be their mean distance from μ. The new center is

$$p_{G+1} = (1 - \lambda_{\mathrm{s}}) \left[\mu + \alpha \varsigma \frac{\mu - a_{k^*}}{\|\mu - a_{k^*}\|}\right] + \lambda_{\mathrm{s}} \bar{p}, \qquad k^* = \arg\min_k \|\mu - a_k\|, \tag{4}$$

where $\alpha = 1.5$, $\lambda_{\mathrm{s}} = 0.1$, and $\bar{p}$ is the mean of the existing family centers. The new bias is set to the median existing family bias. We then save the checkpoint with $G+1$ families.

The adjustment moves the sample centroid away from its nearest human anchor, with a displacement proportional to the sample scatter, and then shrinks it toward the mean family center. It is intended to address cases where the unadjusted centroid lies near human representations and provides little separation. The procedure requires no gradient updates, runs in seconds on a single GPU, and preserves the human model and all existing family tests (Proposition 2).

Related approaches address new generators through conformal or group-specific thresholds (Zhu et al., 2025; Jung et al., 2025), feature transfer with a frozen encoder (Sun et al., 2026), continual origin tracing (Li & Wang, 2025), and few-shot attribution of previously unseen image generators (Wu et al., 2026).

F TRAINING DETAILS

Checkpoint selection uses a metric that averages overall AUROC, worst-family AUROC, and both AUROC and fixed-threshold balanced accuracy on the five weakest supported family×task×register subsets. This metric increases throughout the epoch, so we select the final checkpoint. Human-score thresholds are computed from 22,190 calibration examples across nine registers and stored in the checkpoint. Unless otherwise stated, evaluation uses offline fp32 scoring with 2048-token windows.

Batching and augmentation. We train on documents of up to 8192 tokens using length-bucketed batches with a fixed token budget. Computation therefore depends on the length of each bucket rather than the maximum document length. Human examples make up 50% of each batch and provide the shared negatives in Equation (3). Source–rewrite pairs are sampled as individual examples rather than as paired training units.

Four data augmentations reduce formatting and minor editing differences that the model could use to distinguish human from AI training examples. For example, human arXiv passages often retain hard line breaks from TEX source, while AI rewrites usually do not. With probability 0.3, a document is replaced by a contiguous crop of 48–192 words; this operation is not applied to paired arXiv examples. Hard line breaks are removed or reintroduced with probability 0.75 in both human and AI examples. Line wrapping alone yields an AUROC of 0.74, motivating this augmentation.

Markdown structures, including fences, headings, bullets, and tables, are removed from examples that contain them with probability 0.75 and added to examples that lack them with probability 0.25. These rates are chosen to make markdown similarly common in human and AI examples. A light post-editing procedure introduces typos, omitted commas, and filler words in AI examples with probability 0.25. Its rate on human examples is chosen to make these edits similarly common in both groups.

Optimization. The encoder and head are fine-tuned jointly with AdamW. We use learning rates of 4.5×10^{-5} for the encoder and 6×10^{-4} for the head, weight decay of 0.01, and 600 warmup steps followed by cosine decay to 5% of the initial learning rate. Gradients are clipped at 1.0. Training uses bfloat16 mixed precision, with the scoring head in fp32, for one epoch of the 7.06-million-draw sampling plan on eight L40 GPUs.

Anchors and centers are initialized from $\mathcal{N}(0, 0.5^2)$, and biases are initialized to zero. The temperature is learned in log space, starting from $\log \tau = -1$. During training, we periodically evaluate a 12,000-example validation subset with the deployment scorer. We record AUROC and the formatting-sensitivity measures from Section 5.3: worst-format FPR, FPR after format normalization, and SAR.

G REWRITE CORPUS

Source and units. We construct the corpus from a fixed revision of a public mirror containing LATEX sources for 3,091,241 arXiv papers. Each paper is flattened, its macros are expanded, and its sections are extracted. References, acknowledgments, and front matter are discarded. Captions are retained as separate units, and body prose is divided at sentence boundaries into windows of approximately 600 words, with a preferred range of 400–850 words and a hard limit of 1000 words. For rewriting, equations and citation keys are replaced by temporary placeholders before generation and restored exactly afterward.

A unit is eligible only if every recorded version of its paper predates 1 January 2021. We sample papers in round-robin order, selecting at most three units per paper in each round. Captions are capped at 5% of the pool, and no arXiv category group may contribute more than 38%. This yields 8.0 million units from 1,735,926 papers. Train, validation, and test assignments are made at the

paper level, ensuring that the splits are paper-disjoint and that both members of a source–rewrite pair remain in the same split.

Generators. Four open-weight instruction models produce the corpus: Gemma-4-26B-A4B-it at temperature 0.65, Qwen3.6-35B-A3B at 0.70 with thinking disabled, Ministral-3-14B-Instruct-2512 at 0.55, and gpt-oss-20b at 0.70. All use top-p sampling with $p = 0.95$ and a maximum output length of 2400 tokens. Each source unit is assigned to one generator. Units with the greatest mathematical content are restricted to *polish* and *academic*, the two light-editing operations.

All four models receive the same system instruction, which prohibits preambles, markdown, lists, and statements about being an AI or about the changes made. Rejected outputs are retried at a slightly higher temperature, with a hint identifying the failed acceptance check. Rewrites from all four generators enter training.

Held-out generator test set. The unseen-family evaluation in Section 5.1 uses the same construction pipeline. DeepSeek-V4-Flash and DeepSeek-V4-Pro, both with thinking disabled, rewrite the human passages from gpt-oss-20b's held-out test shard. The prompts, placeholder replacement procedure, and acceptance checks are unchanged. The resulting test set contains 1,600 pairs per model across the nine operations. Neither DeepSeek model contributes training examples. Because the source passages are shared, the gpt-oss-20b subset provides a matched control for a generator represented in training.

Table 6: **Rewrite operations and acceptance criteria.** The nine operations map to six training task labels. *Length* gives the allowed ratio of output words to source words, except for the two generation operations, which use absolute word-count limits because they have no source passage. *Copy* is the maximum allowed 5-gram containment of the source in the output. Polish and academic allow greater overlap because they make smaller edits. *Rows* counts AI examples in the assembled corpus.

Operation	What the instruction asks for	Length	Copy	Rows
Polish	Light copy-editing while preserving the author's voice, terminology, and length	0.68–1.35	0.90	668K
Paraphrase	Change wording and sentence structure while preserving every claim, number, and logical step	0.62–1.40	0.68	203K
Academic	Use a formal journal style with precise terminology and appropriate hedging; preserve content	0.62–1.40	0.90	178K
Humanize	Use a natural, slightly informal first-person academic voice; preserve technical content	0.62–1.40	0.68	131K
Roundtrip	Rewrite as though translated into another language and then back into English by a fluent scientist	0.62–1.40	0.68	121K
Summarize	Reduce the passage to about half its length while preserving the main claims and reasoning	0.30–0.85	0.68	98K
Expand	Expand to approximately 1.35× the source length by explaining implicit steps, without adding unsupported content	1.08–2.00	0.68	78K
Generate from context	Write a new section from the title, abstract, and excerpts from the paper's other sections	130–900 w	0.68	65K
Generate from abstract	Write a new section from the title, abstract, and section heading alone	130–900 w	0.68	50K

Acceptance, pairing, and assembly. Table 6 gives the operation-specific length and copy-overlap limits. We additionally reject outputs that contain fewer than 15 words, markdown, meta-commentary, emoji, excessive repetition, or invented mathematics. Outputs that omit a placeholder or add one that was not in the source are also rejected. Near-verbatim outputs are excluded to avoid assigning an AI

label to a passage that is effectively unchanged from its human source. This check removes 22.6% of gpt-oss-20b outputs, 10.8% of Gemma outputs, 11.0% of Qwen outputs, and 0.1% of Ministral outputs.

We use approximately 400,000 outputs per model (391,000 for Qwen) and produce 3,068,271 examples. These comprise 1,476,558 source–rewrite pairs, each containing one human and one AI version of a section, and approximately 115,000 unpaired examples from the two *generate* operations. The latter produce new section text rather than rewrites of a source passage and are therefore included as ordinary AI examples. Equations and citation keys are preserved exactly for the seven rewriting operations, but not for the two generation operations.

Both members of each pair undergo format normalization during corpus construction. Without this step, the human source retains TEX line breaks while the model output is usually reflowed. Newline density alone distinguishes the two members in 93% of pairs; with the line-wrap augmentation in Appendix F, this falls to 56%. We do not normalize word length, which differs between the two classes (5.83 versus 6.09 characters on average).

The nine operations map to six training labels: *polish* and *academic* map to polish; *paraphrase* and *roundtrip* map to paraphrase; and both generation operations map to generate. Humanize, summarize, and expand retain their own labels. Every arXiv identifier used in training is recorded, and all contributing papers are excluded from the case study in Section 6.

H MELD-EVAL

MELD-eval evaluates transfer to four held-out generators under a common domain and attack protocol. It reuses RAID-style English domains, human seeds, and attacks to reduce differences unrelated to the generators.

Generators and domains. The pool covers eight English domains using two prompting modes. In *continuation* mode, used for books, news, and reddit, the model extends a short human seed. In *instruction* mode, used for abstracts, recipes, reviews, wiki, and poetry, the model generates text from a topic line. Table 7 lists the per-generator example counts, domain-specific seed limits, target lengths, and prompt inputs.

Table 7: **MELD-eval construction.** *Top:* clean and attacked AI examples per generator, with each generator's share of all AI examples. The 7,862 human seeds are shared across generators. *Bottom:* domain-specific prompting settings. Mode A continues a human seed; Mode B generates text from a topic line. The reviews domain has 862 seeds after English-language filtering.

Per-generator counts

Generator	Clean AI	Attacked AI	Ratio
GPT-5.4-Mini (OpenAI)	7.86K	47.17K	0.25
Gemini-3-Flash (Google)	7.86K	47.17K	0.25
Claude-Haiku-4.5 (Anthropic)	7.86K	47.17K	0.25
Qwen-3.6-Plus (Alibaba)	7.86K	47.17K	0.25
Total AI	31.45K	188.69K	1.00

Per-domain configuration

Domain	Mode	Seed cap	Target words	Seed input
books	A	1,000	500	first 80 words of human seed
news	A	1,000	500	first 80 words of human seed
reddit	A	1,000	300	first 60 words of human seed
abstracts	B	1,000	250	topic line
recipes	B	1,000	400	topic line
reviews	B	862	400	topic line
wiki	B	1,000	500	topic line
poetry	B	1,000	180	topic line

Prompting and decoding. All generators receive a shared system instruction requesting plain text without preambles, meta-commentary, chain-of-thought, or markdown. Continuation prompts ask the model to maintain the seed's voice and register. Instruction prompts provide only a topic line and request text in the target domain. We use temperature 0.7, top-p 0.95, and an output limit of 1024 tokens. Qwen-3.6-Plus is queried with thinking disabled to keep its outputs comparable in length and style. The code release contains the exact prompt templates and model identifiers.

Attacks and text normalization. Each of the 31,448 clean AI examples receives six RAID-style attacks: homoglyph substitution, number substitution, synonym substitution, case changes, whitespace modification, and zero-width-space insertion. We use RAID's per-token attack rates, producing 188,688 attacked examples. For reviews, recipes, and poetry, residual markdown is removed from both AI outputs and paired human text using the same procedure, reducing formatting differences as a detection cue.

I QUALITATIVE EXAMPLES

Table 8 compares MELD, Binoculars, ModernBERT-Detect, and RoBERTa-ChatGPT on selected cases where their predictions disagree. The examples come from HC3, DetectRL-test, and MELD-eval, illustrating clean classification, robustness to attacks, and transfer to held-out generators, respectively.

Table 8: **Per-text disagreements on HC3, DetectRL, and MELD-eval.** Each cell reports the standardized margin of a detector score from the pool-specific FPR=1% threshold: $\Phi^{-1}(\text{percentile-in-humans}) - \Phi^{-1}(0.99)$, in standard-deviation units. The threshold is 0 by construction, so ✓ marks correct classification (AI: margin > 0; human: margin ≤ 0) and ✗ marks an operating-point error. A score above every pool human saturates the transform, so $+1.51$ is the maximum expressible margin on MELD-eval. Rows 1–4 are HC3 clean examples (2 AI, 2 human); rows 5–8 are DetectRL attacked examples (3 AI, 1 human); rows 9–12 are clean MELD-eval AI examples, one per generator; rows 13–16 are attack-augmented MELD-eval AI examples, one per generator. These disagreement examples show cases where MELD remains on the correct side of the low-FPR threshold while other open-source detectors fail.

#	Source / attack	**Excerpt** (truncated)	**MELD**	**Bino**	**MBERT-D**	**RoB-CGPT**
HC3 (clean text)						
(1)	HC3 clean	*AI*: John Adams became President in 1797. He was the second President of the United States, serving one term from 1797 to 1801.	$+\mathbf{0.71}$ ✓	-1.32 ✗	-3.37 ✗	$+0.87$ ✓
(2)	HC3 clean	*AI*: Charcoal and regular coal are similar in that they are both made from carbon-rich materials and they can be used as fuels. However, there are some differences between the two that make burning charcoal less practical...	$+\mathbf{0.58}$ ✓	-1.52 ✗	$+0.42$ ✓	-6.60 ✗
(3)	HC3 clean	*Human*: * " I 'd imagine it has something to do with availability . " * True . We 'd probably have problems locking up women in cages too , though .	$-\mathbf{1.43}$ ✓	-0.65 ✓	$+1.70$ ✗	-1.15 ✓
(4)	HC3 clean	*Human*: Logical consequence (also entailment) is a fundamental concept in logic, which describes the relationship between statements that hold true when one statement logically follows from one or more statements. A valid logical argument...	$-\mathbf{4.35}$ ✓	-3.69 ✓	$+0.37$ ✗	-0.58 ✓

continued on next page

Table 8 (continued).

#	Source / attack	Excerpt (truncated)	MELD	Bino	MBERT-D	RoB-CGPT
DetectRL (attacked)						
(5)	DetectRL paraphrase	*AI*: Counting them up revealed steady progress one line at a time. The story was gradually taking shape sentence by sentence. Introducing details like checking my count kept the writing engaging. Finding creative ways to el...	+**0.08** ✓	−1.31 ✗	−2.69 ✗	−2.13 ✗
(6)	DetectRL perturbation	*AI*: I've also noticed that it has a peasant scent. It' not too strong or overpowering, but it's definitely noticeable. It's a nica bonus, especially since I enjoy using products that have a nice fragrance. The packagign i...	+**0.36** ✓	−0.94 ✗	−4.91 ✗	−1.38 ✗
(7)	DetectRL prompt	*AI*: The only thing that was edible was the soup. The service was also terrible. The waiter was rude and dismissive. He took our order and then disappeared for over 30 minutes. When he finally came back with our food, it w...	+**0.04** ✓	−0.85 ✗	−2.05 ✗	−0.75 ✗
(8)	DetectRL paraphrase	*Human*: In the hushed living room, my family surrounds me. The sole sound is a news reporter's voice droning from the TV. Dim light bathes the room in an eerie, pulsating blue glow that mimics the rhythm of my heartbeat. On t...	−**0.88** ✓	−1.19 ✓	+0.46 ✗	−1.03 ✓
MELD-eval clean						
(9)	MELD-eval clean (GPT-5.4-Mini, abstracts)	*AI*: Whitney's broken-circuit theorem gives a foundational expansion of the chromatic polynomial in terms of acyclic substructures of a graph and has played a central role in the combinatorial theory of graph colorings. In this paper...	+**1.23** ✓	−2.93 ✗	−2.10 ✗	−3.10 ✗
(10)	MELD-eval clean (Gemini-3-Flash, abstracts)	*AI*: The deployment of autonomous robotic systems in dynamic, unstructured environments necessitates perception pipelines capable of high-fidelity spatial reasoning under stringent temporal constraints. Reliable image segmentation...	+**0.48** ✓	−3.17 ✗	−0.56 ✗	−3.79 ✗
(11)	MELD-eval clean (Claude-Haiku-4.5, abstracts)	*AI*: The Riemann Hypothesis remains one of mathematics' most profound unsolved problems, asserting that all non-trivial zeros of the Riemann zeta function lie on the critical line Re(s) = 1/2. This review synthesizes...	+**0.60** ✓	−2.10 ✗	−1.50 ✗	−2.59 ✗
(12)	MELD-eval clean (Qwen-3.6-Plus, abstracts)	*AI*: Boolean satisfiability solving has evolved from a theoretically intractable problem into a cornerstone of modern electronic design automation, formal verification, and artificial intelligence. While early algorithms struggled...	+**0.87** ✓	−3.19 ✗	−1.75 ✗	−2.71 ✗

continued on next page

Table 8 (continued).

#	Source / attack	Excerpt (truncated)	MELD	Bino	MBERT-D	RoB-CGPT
MELD-eval attacked						
(13)	MELD-eval upper-lower (GPT-5.4-Mini, reddit)	*AI*: their gc and thEn whEn i asked whY i coUldn't join they Just Said "it's complicatEd" like Okay?? apparently it's fine for literallY everyone else except me. one of them wIll poSt pics of them all hanging out and i'm sit...	+**1.51** ✓	−2.65 ✗	−1.68 ✗	−3.28 ✗
(14)	MELD-eval upper-lower (Gemini-3-Flash, poetry)	*AI*: The clock face blurs in the heavy silence of Three a.m. The walls are thin, the shadows long, and I am the only pulse in tHis Room. I tiLt my chin toward the wiNdow, searching the blacK velvet of the void for a sign...	+**1.23** ✓	−1.67 ✗	−1.51 ✗	−2.59 ✗
(15)	MELD-eval homoglyph (Claude-Haiku-4.5, abstracts)	*AI*: Image-to-image translation, the task of converting images from one domain to another while preserving content structure, has become increasingly important for applications ranging from style transfer to medical imaging. Recent advances in generative adversarial networks...	+**1.51** ✓	−3.71 ✗	−0.94 ✗	−3.69 ✗
(16)	MELD-eval synonym (Qwen-3.6-Plus, reddit)	*AI*: I know moving is trying for everyone, but I didn't realize it would hit her this hard. She has always been a pretty chill cat, the kind that sleeps through vacuuming and doesn't even flinch when the doorbell rings. See...	+**1.51** ✓	−2.36 ✗	−3.59 ✗	−2.13 ✗

In the selected HC3 cases, baselines make errors at the chosen operating point despite their high aggregate AUROC, while MELD classifies all shown examples correctly. In DetectRL, MELD detects the displayed attacked AI examples and avoids the false positive that ModernBERT-Detect makes on the paraphrased human example. In MELD-eval, MELD detects the displayed clean and attacked examples from all four generators, whereas the baselines miss most of them. These selected cases illustrate individual errors; aggregate performance is reported in Tables 2 and 3.

J Ablations and attribution results

Unseen-family and additional benchmarks. The DeepSeek-V4-Flash and DeepSeek-V4-Pro test sets use held-out arXiv passages and the rewrite pipeline described in Appendix G. No DeepSeek family is present in training. The family head assigns most DeepSeek examples to OpenAI GPT (67–82%) or Qwen (15–26%), often with high margins. Requiring an attribution margin above 2 excludes 22–29% of these examples.

Beemo remains challenging for all evaluated detectors: ModernBERT-Detect, RoBERTa-ChatGPT, and RoBERTa-Large achieve AUROCs of 51–69, compared with approximately 82 for MELD. One possible explanation is the labeling convention under which lightly edited AI text can appear on the human side of the evaluation. On TuringBench, which contains 2019-era generated text, MELD achieves an AUROC of 80.1 (Uchendu et al., 2021).

Single-component ablations. Each loss ablation removes one component from the full configuration while keeping the corpus, seed, optimization schedule, and token budget fixed. The evaluation protocol uses the same 12,000 validation examples, 186 documents for formatting tests, and 150 held-out source–rewrite pairs for each configuration.

Two data ablations distinguish the effect of pairing from the effect of corpus size. The *unpaired* configuration removes one member of every pair, so a source passage no longer appears alongside

its rewrite under the opposite label. The *half-pairs* control removes complete pairs while retaining the pairing structure in the remaining examples. Both remove approximately 12% of the full corpus. Their comparison is intended to measure the contribution of retaining matched examples beyond the amount of training data.

Table 9: **Single-component ablation results.** The first panel reports AUROC, TPR at 1% FPR, family attribution accuracy, and coverage and accuracy at attribution margin > 1. The second reports raw worst-format FPR, FPR after format normalization, SAR, and task accuracy. AUROC, TPR, accuracy, and coverage are on a 0–100 scale. SAR uses the training-time evaluation pairs, which differ from those in Table 4. "—" indicates that no task head is trained.

Configuration	AUROC	TPR@1%	Family acc.	Coverage @ accuracy
MELD (full)	99.80	98.25	90.9	88.9 @ 95.4
no family CE	99.52	96.10	62.6	8.8 @ 92.2
no format consistency	99.61	96.84	89.7	86.2 @ 94.8
no task CE	99.70	97.45	89.9	87.5 @ 95.0
unpaired corpus	99.23	94.62	88.4	85.1 @ 94.1
half pairs (volume control)	99.58	96.71	89.8	87.3 @ 94.9

Configuration	Raw FPR	Normalized FPR	SAR	Task acc.
MELD (full)	1.61%	0.54%	144.8	95.6
no family CE	2.15%	0.54%	121.9	94.9
no format consistency	8.60%	0.54%	57.8	95.1
no task CE	1.61%	0.54%	131.2	—
unpaired corpus	2.15%	0.54%	88.4	92.7
half pairs (volume control)	2.15%	0.54%	118.6	93.8

We also examine the scoring geometry by modifying the trained head and rescoring without additional optimization. Replacing eight human anchors with one raises the median score on the human calibration pool from -5.5 to $+2.2$. The proportion of these documents with scores above zero rises from 0.8% to 100%. Replacing top-quantile pooling with mean pooling reduces AUROC by 1.3 points on documents containing 10% AI text.

Per-family detection. Table 10 evaluates each family test on its own AI examples against the shared human validation examples. AUROC ranges from 99.68 to 99.91, a spread of 0.23 points. Qwen and the heterogeneous Other family have the lowest AUROC, matching the worst-family value reported in Section 5.1.

Table 10: **Per-family detection AUROC on validation.** Each family test is evaluated on that family's AI examples against the shared human examples. AUROC is reported on a 0–100 scale.

Family test	AI rows	AUROC
Mistral	4,441	99.91
Claude	2,900	99.87
OpenAI GPT	2,877	99.81
Llama	2,900	99.80
Gemini/Gemma	6,079	99.80
Other	7,796	99.68
Qwen	3,116	99.68

Attribution accuracy. We evaluate the family and task heads separately from binary detection. The six-way task head achieves 95.6% top-1 accuracy on validation AI examples. Family attribution achieves 91.9% on the validation sample described in Section 5.1; Table 11 gives the per-family precision and recall. Claude has 99.2% recall and 99.7% precision, while the heterogeneous Other family has the lowest recall at 82.0%.

The attribution-margin threshold trades coverage for accuracy. At a margin greater than 1, the detector retains 90.1% of examples with 96.1% top-1 accuracy. Raising the threshold to 2 retains 81.4% of examples with 98.3% accuracy.

Table 11: **Task and family attribution.** Top-1 accuracy is reported for the six-way task head and seven-way family head. Family precision and recall use validation examples sampled by writing category, with 1,200 AI examples per family. All values are percentages. "—" indicates that only top-1 accuracy is reported for the task head.

Prediction	**Top-1 acc.**	**Macro prec.**	**Macro recall**
task (6-way)	95.6	—	—
family (7-way)	91.9	92.0	91.9
Per family		*Precision*	*Recall*
Claude		99.7	99.2
Llama		93.3	95.6
Mistral		87.4	92.6
Qwen		95.5	92.4
OpenAI GPT		86.5	91.8
Gemini/Gemma		91.0	89.8
Other		90.7	82.0

K CASE-STUDY METHODS

Previous studies have used distributional estimators to measure LLM-modified content in peer reviews (Liang et al., 2024a) and scientific abstracts (Liang et al., 2024b). Other work has reported an association between AI-assisted reviews and higher scores at one venue (Russo et al., 2025). Our case study in Section 6 analyzes eligible arXiv papers, abstracts from nine conferences, and peer reviews at section and monthly resolution. The analysis uses training-overlap controls, matched review fields, baseline-year negative controls, and threshold-free percentiles.

Each group has a separate threshold corresponding to a nominal 1% exceedance rate in its baseline population. The observed rate ranges from 0.81% to 1.47% across the seven baseline years, 2016–2022, which serve as negative controls. Threshold-free percentiles are the primary statistics.

Sections and authors. Table 13 reports changes by section role and primary arXiv category using P(more AI-like), defined as the probability that a text from a given year scores above a matched baseline text. This mean-rank statistic is lower than the median percentile in Table 12, reflecting the portion of sections that remain at low detector-score percentiles. To assess changes within authors, we also compare 23,404 first authors with at least three publications before 2023 and at least three from 2023 onward. Of these authors, 77% have higher later score percentiles, with a median change of $+10$ percentile points.

Peer review and outcomes. Because review forms change over time, the primary review series uses the summary field whose form is consistent across the included venue–year groups. In 2025, 38.6% of ICLR review summaries exceed their baseline threshold, compared with 23.0% of arXiv section units. Analyses of ratings and acceptance describe population-level associations rather than causal effects.

Table 12: **arXiv detector scores, 2016–2026.** Share above the 1% baseline threshold; median section's role-specific baseline percentile and shift in baseline SD. *2026: January–July.

Year (20xx)	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26***
Flagged (%)	1.3	1.5	1.0	1.0	0.8	0.8	0.9	4.1	10.5	23.0	25.4
Median pctile (%)	47.7	49.0	48.4	49.1	50.9	52.7	54.3	60.9	71.0	85.2	89.5
Shift (baseline SD)	0.00	0.00	0.00	0.00	0.00	0.00	+0.01	+0.02	+0.06	+0.35	+0.72
Section units (K)	13	17	26	56	186	727	759	868	1,034	1,233	762

Corpora. The arXiv corpus contains all 1,344,610 eligible papers from 2016–2026, divided into 5,682,241 units with seven section roles: abstract, introduction, related work, methods, results, discussion, and conclusion. Eligibility excludes all 1,700,807 paper identifiers recorded as contributing to MELD's training pool.

The conference corpus contains 92,174 accepted abstracts from CVPR, ICCV, ECCV, NeurIPS, ICML, ICLR, ACL, EMNLP, and NAACL, covering 2016–2025. The primary peer-review series contains 143,685 form-matched summaries, with reported lengths of 94–104 words across venue–year groups. A secondary series contains 170,720 full-prose reviews. Outcome analyses use 19,064 ICLR submissions with review ratings and acceptance decisions.

The most recent available data was until 11 July 2026. Accordingly, 2026 is a partial year throughout the analysis. Baseline references are balanced across years from 2016–2022. The measured baseline false-positive rate is 1.05% for arXiv and 1.02% for reviews.

Detector and reference distributions. The study uses a MELD checkpoint frozen before scoring, with a nine-operation task head matching the generation pipeline and the labels in Figure 5b (Appendix G). We report $P(\mathrm{AI}) = \sigma(S_{\mathrm{AI}})$ only as a monotone transformation of the detector score, not as a calibrated probability. All primary statistics are rank-based and therefore unchanged by this transformation. The reference distributions contain 1,784,201 arXiv section units, 409,743 arXiv paper abstracts, and 74,875 review units from years through 2022.

Training-overlap control. The training pool records arXiv identifiers for scientific-text examples. We exclude all 1,700,807 recorded paper IDs before selecting the study corpus. The final overlap between the scored corpus and the training set is zero among the 1,344,610 eligible papers.

Additional controls. *Leave-one-year-out calibration.* For each baseline year, we recompute the threshold without that year and evaluate the held-out examples. The resulting flagged rates range from 0.86% to 1.64%.

Equal sample sizes. Sampling 12,769 units from every year, equal to the size of the smallest annual group, reproduces the temporal pattern.

Paper versions. Restricting the analysis to first submissions (v1) gives a similar series. The flagged share in partial-year 2026 is 25.1%, compared with 25.4% in the full corpus. The pattern therefore persists after controlling for revision status in the frozen mirror.

Review length. The 2024–2025 increase remains within every review-length bucket, from 40–149 words to 900 or more words.

Review forms. The change in the ICLR review form between 2021 and 2022, both baseline years, produces score shifts whose direction and magnitude vary across detector checkpoints. We therefore compare only matched review fields across years. As a separate negative control, NeurIPS 2021 reviews are scored against the 2022 reference and have a median percentile of 48.5%, close to the baseline value of 50%.

Thresholds and corpus size. The temporal pattern persists at baseline exceedance levels of 1%, 5%, and 10%, and under the threshold-free statistics. Expanding the analysis from a 613,000-paper sample to the full eligible corpus changes no annual median percentile by more than 0.1 points.

Fields, venues, and replication. Score changes differ across primary arXiv categories (Table 13 and Figure 6). In partial-year 2026, cs.CV reaches $P(\text{more AI-like}) = 82.7\%$, with a shift of approximately $+3.6$ baseline standard deviations. For hep-th, the corresponding values are 66.0% and $+0.08$ standard deviations. In cs.AI, the monthly flagged share falls from 47% in January to 30% in June 2026, alongside a change in the energy score. A possible explanation is a change in generator writing styles.

With separate baseline references, the corresponding rank-based values for bioRxiv and medRxiv abstracts reach 77% and 85% in 2026. For accepted conference abstracts in 2025, venue-specific baseline percentiles range from the 89th to the 98th percentile. Flags also cluster within papers, with a conditional lift of $1.9\times$ in 2026. For papers with at least three scored units, this is the probability that another unit is flagged given a flagged unit, divided by the mean within-paper flagged fraction.

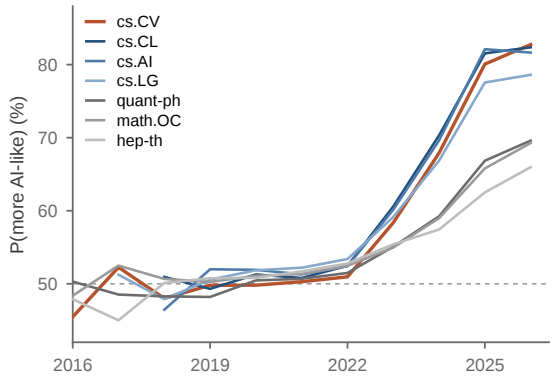

Figure 6: **Detector-score trends by primary arXiv category, 2016–2026.** Each curve gives the probability that a text from a given year scores above a 2016–2022 baseline text from the same category. The reference level is 50%.

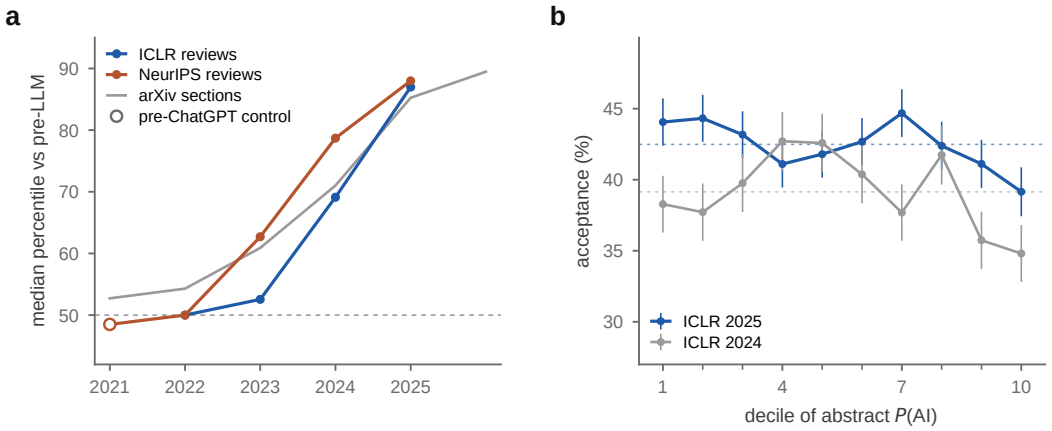

Figure 7: **Review-score trends and ICLR acceptance rates.** *(a)* Form-matched review summaries, with each venue compared against its own last baseline year. The concurrent arXiv series is shown for reference, and open markers denote negative controls. *(b)* Acceptance rates by within-year decile of abstract $P(\text{AI})$. Error bars show one standard error of the mean; dotted lines show the annual mean. Acceptance is approximately constant across the first nine deciles and lower in the highest decile.

Within-author analysis. For each of the 23,404 first authors, we compare section roles present in both periods using the same baseline reference. Changes form a continuous distribution. Using the study's change categories, 36.4% remain within baseline variation, 41.8% show modest changes, 20.2% show substantial changes, and 1.6% fall in the largest-change category.

Rating analyses. Table 14 and Figure 7a give the full form-matched review series. Rating scales differ across venues and years: ICLR 2020 uses $\{1,3,6,8\}$, and average NeurIPS ratings are approximately one point higher than ICLR ratings. Before making comparisons, we therefore convert each rating to its percentile within the same venue and year.

Flagged and unflagged reviews have similar calibrated ratings. The differences are -0.7 percentile points for ICLR 2025 and $+0.5$ points for ICLR 2024. Within calibrated-rating quintiles, the flagged share of a paper's reviews is associated with acceptance-rate differences ranging from -0.8 to $+5.5$ percentage points, with no consistent direction.

Papers with the highest abstract detector scores receive lower calibrated ratings overall. For ICLR 2024 and 2025, the reported differences are -6.8 and -5.6 percentile points, with $z = -4.5$ and $z = -7.9$, respectively. However, acceptance rates are approximately constant across the first nine

Table 13: **Score shifts by section role and arXiv category in 2026.** Each group is compared with its own baseline distribution. *Top:* section roles. Conclusions show the largest standardized shift (+3.39 baseline standard deviations). *Bottom:* selected primary categories. The listed computer-science categories show larger shifts than the listed physics and mathematics categories.

Section role	P(more AI-like) (%), 2026	Shift (baseline SD)
conclusion	81.4	+3.39
discussion	80.1	+2.03
related work	78.8	+1.56
introduction	76.8	+0.62
methods	74.9	+0.68
results	74.9	+0.89
abstract	73.5	+0.92

Category	P(more AI-like) (%), 2026	Shift (baseline SD)
cs.CV	82.7	+3.63
cs.CL	82.4	+2.70
cs.AI	81.6	+2.89
cs.LG	78.6	+1.65
quant-ph	69.6	+0.20
math.OC	69.3	+0.12
hep-th	66.0	+0.08

Table 14: **Form-matched review summaries by venue and year.** Each venue uses its own last baseline year as the reference. NeurIPS 2021 is a negative control evaluated against the 2022 reference; its median percentile is close to the reference value of 50%.

Venue	Role	Year	*n*	Median pctile (%)	Shift (baseline SD)
NeurIPS	control	2021	9,676	48.5	−0.03
NeurIPS	baseline	2022	9,536	50.0	0.00
NeurIPS		2023	14,060	62.7	+0.39
NeurIPS		2024	15,300	78.7	+1.26
NeurIPS		2025	21,278	88.0	+1.97
ICLR	baseline	2022	9,058	50.0	0.00
ICLR		2023	12,709	52.6	+0.06
ICLR		2024	20,069	69.1	+0.66
ICLR		2025	31,999	87.0	+1.86

abstract-score deciles and lowest in the highest decile (Figure 7b). The top-decile acceptance rates are 34.8% and 39.2% in 2024 and 2025, respectively, compared with corresponding mid-range rates of approximately 40% and 43%. Before 2024, flagged reviews tend to be unusually short. Flagged and unflagged reviews have similar median lengths: about 374 versus 371 words in 2024 and 384 versus 391 in 2025. The later comparisons therefore do not show the same length difference as the earlier flagged reviews.