TRUST, BUT VERIFY: A FRONTIER MODEL BENCHMARK AND A LEAKAGE-FREE PROTOCOL FOR AI-GENERATED TEXT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

AI-generated-text detectors increasingly inform academic-integrity and content-moderation decisions, yet reported accuracy is hard to trust. We audit three vulnerabilities. **(1) Evaluation leakage**: test-informed checkpoint selection inflates macro-F1 by 27.3 points in a controlled MAGE replication. Our COLING-2025 endpoints differ by 1.72 points (83.65 $\rightarrow$ 85.37), but incomplete candidate histories prevent attributing that difference solely to selection. **(2) Distribution-sensitive human-text flagging**: on domain-matched scientific abstracts, RoBERTa-large-OpenAI flags 45.5% of post-2023 text against 20.1% of pre-2015 text. This direction persists within both PubMed and arXiv, although source composition changes some detectors' aggregate conclusions. **(3) Heterogeneous frontier performance**: evaluated off-the-shelf detectors span macro-F1 58–85 on 2026 frontier text. The author-released DetectAnyLLM checkpoint achieves the highest external macro-F1, 84.60, compared with 76.67 for the COLING-2025 winner Advacheck, yet flags 48.78% of our post-2023 presumed-human abstract probe. We evaluate the author-released checkpoint, not a reproduction of the published result. Rankings change with the operating point, and Binoculars and Fast-DetectGPT recover only 14–19% of Claude Sonnet 4.6 and Grok-4.20-reasoning text. We release **FrontierBench-2026**, comprising 216,352 texts from 16 collected frontier LLM variants (15 evaluated), a provenance-verified pre-2015 human reference, a domain-matched contemporary probe, and an evaluation harness separating development-set selection from test reporting. Using a graph-fusion detector as an audit instrument, we find no stable advantage over the shared-task winner across three seeds (paired macro-F1 difference $+0.08 \pm 0.24$, mean $\pm$ s.d.). Frontier-aware training raises frontier macro-F1 from 78.4 to 92.8 on the shipped split and achieves 89.8 ± 0.2 on a content-controlled split with disjoint prompts and parent documents. Held-out-provider recall is 0.845 shipped and 0.764 content-controlled. Capability-tier transfer is provider-asymmetric, while short-form detection and genre-dependent contemporary-human false alarms remain unresolved.

1 INTRODUCTION

Detecting AI-generated text underpins academic-integrity tooling, platform moderation and content provenance, yet detectors report high benchmark accuracy and generalize poorly, and a growing body of work argues their numbers reflect dataset artifacts rather than machine-vs-human signal (Dugan et al., 2024; Li et al., 2024; Stowe & Patil, 2026). The problem is as much *how we evaluate* as the detectors themselves, so we treat evaluation as the object of study. Three gaps motivate this work.

(i) Frontier coverage. Public benchmarks (Dugan et al., 2024; Li et al., 2024; Wang et al., 2024a; 2025; 2024b) cap at 2023–2024 generators; none pairs 2025–2026 frontier models with matched reasoning-mode variants — the outputs deployed today (Table 1). **(ii) Evaluation integrity.** Results are often reported for the checkpoint or threshold scoring best *on the test set*; our retained COLING-2025 endpoints differ by $\sim$2 points (85.37 vs. a leakage-free 83.65), enough to reorder closely-spaced systems — though the candidate dev-score history behind that pair was not retained, so the

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.

controlled demonstration is our MAGE replication (Section 5). **(iii) Distribution-sensitive false positives.** Benchmarks pair human text of one era with machine text of another, so scores can track era-associated shift rather than authorship: on domain-matched abstracts a 2019-era detector flags 0.455 of post-2023 human text against 0.201 of pre-2015. But those populations also differ in source mixture, and the pooled direction is not stable across detectors once PubMed and arXiv are separated.

We address these with a resource, an audit and a method under one leakage-free protocol (Figure 1):

- **FrontierBench-2026**: the first benchmark to *jointly* combine broad 2026 frontier coverage across four providers (16 variants collected, 15 evaluated), reasoning/non-reasoning pairs, a provenance-verified pre-2015 human reference, a domain-matched post-2023 probe and a content-controlled split. We claim the combination, not sole 2026 coverage.
- **A selection-bias finding**: in the MAGE replication, where the full checkpoint sequence survives, test-informed *checkpoint* selection measurably inflates macro-F1, whereas *threshold* selection inflates zero-shot detectors far less (0.21–0.22). Our COLING-2025 endpoints (83.65 vs. 85.37) are an incomplete historical audit, not a measured effect.
- **A frontier and human-text audit** with bootstrap CIs and paired tests. Frontier performance is heterogeneous (58–85), the best external checkpoint also flags 48.78% of the post-2023 probe, and we bound every claim to the systems and operating points evaluated.
- **An interpretable detector as the instrument**, with no stable advantage over the shared-task winner under the clean protocol (83.16 ± 0.24 vs. 83.07).
- **Closing the frontier gap.** Frontier-aware training lifts frontier macro-F1 $78.4 \rightarrow 92.8$, and 89.8 ± 0.2 under content control, with held-out-provider recall 0.845 shipped / 0.764 content-controlled, and cuts pre-2015 human FPR to 0.0033 ± 0.0004 — though it still flags 0.119 ± 0.038 of matched post-2023 abstracts, genre-dependently. Capability-tier transfer is provider-asymmetric and its sign depends on the averaging rule. Residuals: reasoning and newest-flagship models off-the-shelf, short-form text for frontier-aware detection.

2 Related Work

Foundations. The task predates the LLM era (Zellers et al., 2019; Ippolito et al., 2020; Uchendu et al., 2020); recent surveys map its growth (Crothers et al., 2023; Tang et al., 2024; Wu et al., 2025). **Training-free detectors** divide by mechanism: *likelihood/curvature* methods score token probabilities under a proxy LM (Mitchell et al., 2023; Bao et al., 2024; Su et al., 2023); *perplexity-ratio* methods compare an observer and a performer model (Hans et al., 2024) — both assuming machine text is lower-perplexity, an assumption that weakens as generators grow fluent; *regeneration-based* methods compare a text with continuations of its own prefix (DNA-GPT; Yang et al., 2024); and *geometric* methods use embedding structure (Tulchinskii et al., 2023). **Supervised detectors** fine-tune encoders on human/machine labels (Solaiman et al., 2019) and add adversarial training (RADAR; Hu et al., 2023), in-context adversarial examples (Koike et al., 2024), fine-grained labels (Abassy et al., 2024), sentence-level sequence modeling (Wang et al., 2023) or token-relation modeling (Wu et al., 2026a); they are strong in-distribution but opaque and prone to overfitting their training generators. **Benchmarks** have driven progress: RAID (Dugan et al., 2024) contributes scale and adversarial attacks, MAGE (Li et al., 2024) out-of-distribution testbeds, M4/M4GT (Wang et al., 2024a) multilingual coverage, and SemEval-2024 (Wang et al., 2024b) and COLING-2025 (Wang et al., 2025; Dugan et al., 2025) shared-task leaderboards; recent efforts target real-world scenarios (Wu et al., 2024; 2026b), example-hardness (Ayoobi et al., 2025) and modern generators (Kempton et al., 2026), which we do not run as a baseline. **Robustness and generalization.** Detectors are fragile under paraphrase and adversarial edits (Zha et al., 2025; Rivera Soto et al., 2024; Bao et al., 2026; Song et al., 2026; Wang et al., 2026); rewriting-based detectors turn that asymmetry into a signal (Mao et al., 2024), theory bounds when detection remains possible as models approach human fluency (Chakraborty et al., 2024; Sadasivan et al., 2023), and audits of deployed education tools find them unreliable (Weber-Wulff et al., 2023; Stowe & Patil, 2026). **Interpretable and provenance approaches.** Feature-based detectors (Gehrmann et al., 2019; Verma et al., 2024) offer interpretability but target earlier generators; an orthogonal line embeds provenance at generation time via watermarking (Kirchenbauer et al., 2023), itself weakened by paraphrase (Krishna et al., 2023;

Kirchenbauer et al., 2024). We are complementary on three axes: a dedicated *frontier* benchmark with reasoning-mode pairs and temporally controlled human references; isolation of an underexamined source of inflated numbers, *test-informed model selection*; and a detector whose interpretable gate blends an encoder branch with theory-grounded linguistic features rather than perplexity.

3 FRONTIERBENCH-2026

Composition. FrontierBench-2026 is a balanced, English, binary detection benchmark of 216,352 texts (205,548 train / 5,402 dev / 5,402 test; 112,051 AI / 104,301 human). AI text is generated by **16 frontier LLM variants available in 2026** across four providers (Table 9), including explicit reasoning and non-reasoning modes—absent from the established public detection sets of Table 1. Fifteen of the sixteen are *evaluated*: Gemini-3-flash contributed only 14 documents, all in train/dev, so it carries no held-out test cell ("16 collected, 15 evaluated" throughout). Human text comes from **5 provenance-verified pre-2015 source families** (Appendix A), so it provably predates widespread LLM assistance. We also release a 1,105-document *post-2023 human* probe of arXiv and PubMed abstracts, dated 11.9% 2024, 13.2% 2025 and 74.9% 2026 — weighted towards 2026, not a 2024–2025 sample — matched to the pre-2015 abstracts *in domain but not in source composition* (85.2%/14.8% PubMed/arXiv vs. 47.4%/52.6%), which is why Section 5 reports source-standardised as well as pooled comparisons.

Two design limitations. The pre-2015 reference buys provenance at the cost of an *era separation between the class-labelled corpora* — human text is pre-2015, AI text 2026 — so era-correlated artefacts are confounded with the label; the post-2023 probe and the standardised analyses address this but do not remove it. And the AI test side is sparse mid-range: only 15 of 2,685 AI test documents fall between 200 and 1,000 characters. The near-even split (48%/52%) and per-cell length balancing remove class prior and coarse length as *straightforward* shortcuts, but with that band unpopulated we cannot claim length is uninformative across the whole range.

Generation and decontamination. Each provider is prompted with the same 1,860-entry prompt set (604 distinct texts) across 13 domains. **Standard-sampling models** are queried under a decoding grid (temperature $\in \{0.5, 0.7, 1.0\}$, top-p=0.95); **reasoning-mode models** are queried at their API-native settings, since those endpoints do not accept a non-default temperature or top-p. A length-extreme strategy targets very short ($<$ 25-word) and very long ($>$ 3000-word) outputs so length is not a trivially separating cue — a design *aim*, not a demonstration that length carries no signal — and reasoning models are queried in both modes where available, yielding matched pairs. Appendix B tabulates per-model decoding and one metadata caveat we later identified. The dev/test splits are *length-decontaminated* (balanced per domain×token-length bucket). A train/test audit (exact hashing, 100-word prefix hashing, 13-gram containment; script released) finds **0 exact duplicates** and 0 prefix collisions—both boundary-sensitive, hence the third test—and $>50\%$ 13-gram containment for 0.56% of *eligible* AI documents and 11.2% of *eligible* human documents, where eligible means long enough to yield a 13-gram (12/2,126 AI, 263/2,357 human; equivalently 0.45% of all 2,685 AI and 9.7% of all 2,717 human test documents). The human hits are almost entirely long-form sources chunked across the split (Appendix G).

What the shipped split does *not* control. The splits are decontaminated at the level of *text*, not *prompt*. One prompt pool is issued to every provider by design—that holds content fixed while the generator varies—but consequently **84.4**% of AI test documents answer a prompt also seen in training, and 289 distinct human parent-document titles are shared between train and test (Table 31). The AI figure is a *document*-level rate; the parent-document figure is a count of distinct *titles*, and the two are not interchangeable. We therefore also release a **content-controlled (grouped) split** where no prompt text or parent document crosses a boundary, and report every frontier-aware headline in both regimes (Table 5); the grouped split is the conservative reading for content-generalization claims. The corpus carries **13 canonical domain labels**, eight of them AI-only *in the training split* (code explanations, QA, reviews, tweets, definitions, one-liners, long essays, long stories — exactly 54,638 training documents). **AI-only status is split-scoped, not global**: six appear only in training, while *long essays* and *long stories* acquire human documents in dev (114, 278) and test (85, 259), which is why they are evaluated and why Table 17 reports human FPRs for them. Dev/test therefore span seven domains, of which *stories* (8/7 in test) is too small to tabulate (Table 24). These training-side AI-only domains cannot bias the domain-adversarial objective, because $\mathcal{L}_{\text{dom}}$ classifies legacy

Table 1: FrontierBench vs. prior benchmarks. "Front.": 2026 frontier generations; "Reas.": matched reasoning modes. *Kempton et al. (2026) regenerate RAID with GPT-5.4, Gemini 3.1 Pro and Claude Sonnet 4.6; no matched reasoning modes or pre-2015 human reference.

Benchmark	#Gen.	Front.	Reas.	Human
RAID	11	No	No	web
MAGE	27	No	No	recent
M4GT	9	No	No	recent
COLING-25	8	No	No	recent
DetectRL	4	No	No	recent
Modern RAID*	3	Yes	No	web
Ours	**16**	**Yes**	**Yes**	**$\leq$2015**

corpus provenance (HC3/M4GT/MAGE) rather than prompted domain: every FrontierBench record maps to `domain_id`$= -1$ and is masked out, so $\mathcal{L}_{\text{dom}}$ is identically zero in all frontier-aware runs (Appendix C). This is a property of the loss, not a mitigation of prompted-domain shortcuts, which remain possible through the shared representation. The 16 variants span four providers—Anthropic, OpenAI, Google, xAI—with matched reasoning modes where available (Appendix B, Table 9).

Recent 2026 work has also evaluated detection on modern frontier generators, including GPT-5.4 (Kempton et al., 2026), so we do not claim sole 2026 coverage. What we claim is the *combination*: 2026 frontier breadth across four providers, matched reasoning/non-reasoning pairs, a provenance-verified pre-2015 human reference, a matched post-2023 human probe, length decontamination, and a released content-controlled split (Table 1).

4 METHOD

Our detector (Figure 2, Appendix C) shares a single fine-tuned ModernBERT-large (Warner et al., 2024) backbone (hidden size 1,024, truncated from 28 layers to 18) between two branches. The **BERT branch** mean- and max-pools token states through the backbone head to produce $\text{logits}_{\text{BERT}}$. The **graph branch** pools token states into sentence nodes and applies a 2-layer sparse Graph Attention Network (Veličković et al., 2018) over a document graph whose edges combine local, semantic (cosine) and learned windowed connections, yielding $\text{logits}_{\text{graph}}$ and three graph-evidence signals (coverage, margin, density). A **linguistic fusion gate** computes a scalar β from six inputs — those three plus the backbone's decision margin, its entropy and normalised document length — and **65 stability-selected linguistic features** (the T1–T10 families), and blends the two branches:

$$\text{logits} = \beta \, \text{logits}_{\text{BERT}} + (1 - \beta) \, \text{logits}_{\text{graph}}. \tag{1}$$

The gate output is clamped to $[0.05, 0.95]$; a separate override routes documents with < 5 sentences to the BERT branch (β=1) without masking the graph-branch losses. Training minimizes label-smoothed cross-entropy on the fused logits with five regularizers: an auxiliary head predicting the 65 features, a domain-adversarial head with a gradient-reversal layer (Ganin et al., 2016), a symmetric-KL term between branches, a cosine alignment term on *pooled document* representations projected into a shared 128-dimensional space, and a one-sided gate floor at β=0.3. The 65 features are the subset of 186 numeric T1–T10 features stable across training corpora; selecting by *cross-dataset stability* rather than perplexity means no proxy LM is queried at inference. The backbone is warm-started from a short-document structural-augmentation (SDA) stage. Appendix C gives the full objective, hyperparameters and feature specification.

Why a linguistic layer? The likelihood-based detectors we evaluate and the older supervised checkpoints degrade sharply on frontier models (Table 4); a natural explanation is that fluent, well-calibrated generation erases the statistical signatures they exploit, though we measure the degradation, not that mechanism, and the newer trained baselines we evaluate do not degrade in the same way. The T1–T10 families instead capture *how* text is organized—negation scope, connective and discourse structure, hedging and stance, figurative usage, referential breadth, deixis, metadiscourse, surface/punctuation regularity—dimensions that differ systematically between human and machine authorship and that pre-training does not directly optimize. They are *interpretable* (each maps to a

named construct) and not tied to a particular proxy LM, avoiding one source of generator-dependent mismatch. We state the limit in advance: our multi-seed ablations (Appendix I) do not establish that the linguistic target content is the mechanism, so the features are an interpretable read-out rather than a validated causal mechanism. We make no causal claim for them. The auxiliary head recovers them faithfully (referential richness ρ=0.84, pragmatics ρ=0.81, surface & punctuation ρ=0.80), so the objective forces the shared representation to encode them. On legacy data the gate keeps the BERT branch in charge ($\beta > 0.9$) while the interpretable heads are still worth 3.5 macro-F1 (Table 14); under frontier-aware training the gate activates but that margin does not follow (Appendix I).

5 EXPERIMENTS

Protocol. Our own COLING-2025-trained variants select their checkpoint on a held-out COLING-2025 development set by macro-F1 and are evaluated *once* on the official test set (N=73,941; metric = macro-F1): backbone, augmentation stage and fusion checkpoint are each dev-selected, as is the choice between those variants. **This does not describe every trained model in the paper**: our FrontierBench-trained configurations are dev-selected on a balanced FrontierBench held-out split (Section 5.1), and third-party released checkpoints arrive pre-trained with selection histories we do not control and do not claim. Training-free detectors select only a threshold, on dev. We report bootstrap 95% CIs (2000 resamples) and McNemar paired tests.

Main comparison (Table 2). Our dev-selected run attains macro-F1 83.65 (95% CI [83.39, 83.92]) on the official COLING-2025 Task 1 test set; three fresh seeds of the identical configuration give $\mathbf{83.16 \pm 0.24}$. Leaderboard numbers come from the same leakage-free regime (blind submission, no test tuning), so the comparison is like-for-like, and our *own* best-reported 85.4 is the test-oracle endpoint rather than the leakage-free one. The released Advacheck checkpoint (Gritsai et al., 2025) reproduces its reported 83.07, so that comparison can be made *paired* per-sample. The reported RoBERTa baseline (73.42) has no predictions and is compared by interval only. We report paired McNemar tests ($p < 10^{-3}$) against the four baselines we ran end-to-end (Binoculars, Fast-DetectGPT, two RoBERTa-OpenAI detectors), where we hold aligned per-sample outputs.

We claim parity with the shared-task winner, not a ranking—and the reason is this paper's own thesis. On the single dev-selected run the paired difference against Advacheck is +0.58 (McNemar p=5×10^{-8}, CI [+0.33, +0.81]), which would read as a leaderboard win. Repeating the *identical* paired test at three seeds does not support that reading: +0.343 (p=2.9×10^{-4}), −0.139 (p=0.95) and +0.044 (p=0.18), a mean paired difference of $\mathbf{+0.08 \pm 0.24}$ whose *sign changes across seeds*. One run in three is significant; the ordering is not stable. We therefore state that our detector shows *no stable advantage* over Advacheck under this protocol and report parity within observed seed variability—a failure to establish a difference, not an equivalence test—and we give the single-run figure only as the dev-selected artefact it is. This is Finding 1 arriving in our own headline: a single run, selected without touching test, still yields an apparent ranking that multi-seed evaluation dissolves. A leaderboard claim resting on one seed is a weaker version of the same error this paper exists to document, which is why we do not make one.

Finding 1 — Evaluation leakage (Table 3). We distinguish leakage-free selection from test-informed selection. Reusing our detector as a case study, an identical architecture and recipe yields macro-F1 **85.4** when the backbone, augmentation stage, and fusion checkpoint are each chosen by test accuracy, but **83.65** under the leakage-free protocol. For zero-shot detectors—whose only free parameter is a threshold—inflation is smaller (0.21–0.22). Test-informed *checkpoint* selection, available only to trained detectors, dominates and is easy to introduce accidentally. +1.72 **is a historical test-score difference between retained endpoints, not a measured selection effect**: of 37 candidates only 14 retain test predictions and 3 retain dev scores (Table 15), so the dev-argmax of the surviving pool is itself unknown. The MAGE replication below, whose complete checkpoint sequence is retained, is the controlled demonstration; the ledger is the historical case study. Ablation shows feature *stability*, not quantity, drives generalization: the 65 stable features beat all 186 (83.65 vs. 82.74), with SDA supplying most of the rest (Table 14). On an independently-trained MAGE-supervised RoBERTa the same test-informed selection inflates FrontierBench macro-F1 by +27.3 points (44.56 → 71.87; Table 3) — the leak is a property of the protocol, not our architecture.

Finding 1b — the seed is a selection variable too. Across five component variants at three seeds each (15 runs, fixed in advance), the pooled seed main effect is $\mathbf{0.233}$ macro-F1—*larger than our*

Table 2: COLING-2025 Task 1 (English) test, leakage-free protocol (N=73,941). Shared-task leaders shown for context. Our primary figure is the three-seed mean; the dev-selected single run is listed beneath it for traceability. The paired difference against Advacheck is $+0.08 \pm 0.24$ over seeds and changes sign, so we claim parity, not a ranking. †See Table 16 (Appendix D).

System	macro-F1	95% CI
Ours: Graph-Fusion + SDA (3 seeds)	**83.16 ± 0.24**	—
dev-selected single run	*83.65*	[83.39, 83.92]
Advacheck (task winner)†	83.07	[82.80, 83.34]
Binoculars (official)	80.93	—
RoBERTa (shared task)	73.42	—
Fast-DetectGPT	71.04	[70.71, 71.38]
RoBERTa-large-OpenAI	56.84	[56.48, 57.21]
RoBERTa-base-OpenAI	53.72	[53.34, 54.06]

Table 3: Leakage-free vs. test-oracle selection (macro-F1). **The MAGE‡ row is the controlled demonstration** (full checkpoint sequence survives, so the oracle choice is reselectable); **the COLING-2025 row is an incomplete historical audit** — its candidate dev-score history was not retained, so $+1.72$ is a difference between retained endpoints, not an identified selection effect. ‡*Our own* RoBERTa-large trained on MAGE, a different model from the *released* MAGE checkpoint in Table 19 (67.66); the two are not comparable.

Detector	leak-free	oracle	Δ
Ours (ckpt/arch)	83.65	85.37	$+1.72$
Fast-DetectGPT (thr.)	71.04	71.26	$+0.22$
Binoculars (thr.)	80.93	81.14	$+0.21$
MAGE-RoBERTa (ours)‡	44.56	71.87	$\mathbf{+27.3}$

graph/linguistic branch's entire contribution (0.086) and comparable to the auxiliary objective's (0.226). So a single-run margin below the seed spread is not a stable ranking, including our own above. Nor is this confined to macro-F1: on the post-2023 probe the same configuration varies by $\mathbf{0.038}$ s.d. in FPR across seeds — enough for a single-run FPR comparison to invert its sign. 0.233 is not a field-wide threshold; the spread must be measured per configuration (Appendices I, H).

Finding 2 — Per-generator failure modes (Table 4). Every detector tested, the strongest included, has severe per-generator failure modes, concentrated on frontier flagship and reasoning models. Binoculars and Fast-DetectGPT recover only **14–19%** of Claude Sonnet 4.6 and Grok-4.20-reasoning text, while cheaper/flash models stay easy (0.78–0.94). *The reasoning-mode gap* is systematic: Binoculars recovers 47% of Grok-4.20-non-reasoning text but only 15% of the *reasoning* variant. On the small o3 cell (n=25) RoBERTa-large-OpenAI reaches 0.56 $[0.37, 0.73]$, DetectAnyLLM 0.52 $[0.33, 0.70]$, our legacy detector 0.36 $[0.20, 0.55]$ and Advacheck 0.12 $[0.04, 0.30]$ — intervals overlapping heavily, so no ranking; and RoBERTa-large-OpenAI's false alarms are catastrophic once the human era shifts (0.455 on post-2023 abstracts). The difficulty is shared across substantially different paradigms.

Finding 2b — Frontier performance is heterogeneous, not a uniform collapse (Tables 4, 19). RADAR (Hu et al., 2023) and the released MAGE checkpoint (Li et al., 2024) fall to macro-F1 66–68, RoBERTa-base-OpenAI to 58.03 and RoBERTa-large-OpenAI to 77.33 — but two recent systems complicate that picture. Advacheck (Gritsai et al., 2025), the COLING-2025 winner, reproduces its published COLING macro-F1 (83.08) and reaches 76.67 here; the author-released DetectAnyLLM checkpoint‡ (Fu et al., 2025) — a DDL-trained discrepancy detector, not training-free — reaches $\mathbf{84.60}$, the highest external result under the decision rules reported here and above our own legacy-trained detector (78.45). **That headline does not travel to contemporary human text: the same checkpoint flags 48.78% of the post-2023 presumed-human abstract probe** (Table 11). Off-the-shelf detection therefore does not uniformly collapse — and a high aggregate frontier score does not imply deployable behaviour.

Three qualifications travel with that. *Operating point*: at a nominal 1% dev-human FPR DetectAnyLLM recovers 0.3199 vs. Advacheck's 0.4719 and our frontier-aware model's 0.9423, at achieved test FPRs 0.0110/0.0129/0.0088 — close but not identical, so not a common-test-FPR comparison. *Generator*: recall runs from 0.9318 on o4-mini to 0.1948 on Claude-Opus-4.7. *Length*: 0/617 below 200 characters, 3/15 from 200–999, 1,942/2,053 at $\geq$ 1,000 (sum 2,685, none skipped). Our frontier-aware detector shares that short-form weakness; we note the coincidence without claiming a shared mechanism. **Scope.** We evaluated RADAR, the released MAGE checkpoint, both RoBERTa-OpenAI detectors, Advacheck and the pinned DetectAnyLLM release (a *trained* discrepancy detector, not zero-shot), plus proxy-LM estimators Binoculars and Fast-DetectGPT. **Not evaluated**: DNA-GPT (Yang et al., 2024), GPT-who and intrinsic-dimension methods (Tulchinskii et al., 2023); our conclusions do not extend to them.

Table 4: Per-generator recall (TPR), overall macro-F1 and human FPR (pre-2015 reference). ‡Author-released DetectAnyLLM checkpoint, *not* a reproduction of the published result (Appendix D). **Bold generator names** mark the reasoning / newest-flagship group only; no cell is bolded and shading carries no significance. Small cells (N=25–50) carry 95% intervals up to 36 points wide (o3, n=25: 0.56 [0.37, 0.73]). 15 of 16 collected variants appear; Gemini 3-flash has no held-out test cell.

Generator	N	Ours	Binoc.	Fast-DGPT	RoB-lg	Advach.	DetAnyLLM‡
Gemini-2.5-flash-lite	408	0.944	0.880	0.831	0.748	0.865	0.853
Gemini-2.5-flash	494	0.931	0.850	0.781	0.828	0.850	0.872
Gemini-2.5-pro	50	0.880	0.840	0.800	0.720	0.700	0.800
Grok-3	338	0.885	0.899	0.849	0.772	0.855	0.876
Grok-4-fast (non-reas.)	231	0.714	0.693	0.528	0.667	0.524	0.719
Claude Haiku 4.5	133	0.692	0.541	0.391	0.526	0.571	0.519
Gemini-3.1-pro	62	0.790	0.419	0.419	0.468	0.419	0.565
GPT-5.4-mini	366	0.678	0.579	0.522	0.754	0.598	0.724
Claude Opus 4.7	77	0.675	0.312	0.273	0.299	0.312	0.195
Grok-4.20 (non-reas.)	164	0.591	0.470	0.341	0.652	0.652	0.634
Claude Sonnet 4.6	105	0.724	0.162	0.190	0.448	0.610	0.343
GPT-5.5	44	0.727	0.295	0.205	0.341	0.273	0.318
Grok-4.20 (reasoning)	144	0.590	0.146	0.181	0.521	0.604	0.500
o4-mini	44	0.500	0.636	0.455	0.977	0.455	0.932
o3 (reasoning)	25	0.360	0.280	0.320	0.560	0.120	0.520
Overall macro-F1		*78.45*	*74.27*	*70.51*	*77.33*	*76.67*	***84.60***
Human FPR		*0.219*	*0.176*	*0.180*	*0.145*	*0.156*	*0.029*

Finding 3 — Human-text false positives depend jointly on era and source (Tables 11, 12). Comparing flagging rates on scientific abstracts from before 2015 (n=1,421) and after 2023 (n=1,105), RoBERTa-large-OpenAI ($0.2006 \to 0.4552$), Advacheck ($0.0929 \to 0.1439$), our frontier-aware detector ($0.0002 \to 0.1195$) and the released DetectAnyLLM checkpoint ($0.0394 \to \mathbf{0.4878}$) all flag more post-2023 text. **DetectAnyLLM's $+0.4484$ is the largest pooled difference in Table 11**, and it is the same checkpoint with the highest external frontier macro-F1: strong aggregate separation and heavy flagging of contemporary presumed-human abstracts coexist in one system. It survives standardisation ($+0.4640$ [$+0.433, +0.495$] equal; $+0.4662$ [$+0.434, +0.500$] combined) and holds within both sources, so it is not a mixture artefact — but the two populations differ in era *and* provenance status, so this is not evidence that era causally produces it. That pooled contrast is not purely temporal: the pre-2015 side is 85.2% PubMed and the post-2023 side 52.6% arXiv, and stratifying by source (Table 12) shows the pooled direction is stable for some detectors and not others. RoBERTa-large-OpenAI stays positive within both sources and under both standardisations; the legacy detector and Fast-DetectGPT are positive within PubMed and negative within arXiv, so composition *contributes* to their negative pooled values without explaining them away — equal-source standardisation moves the legacy difference to $+0.0050$ with an interval spanning zero, while Fast-DetectGPT stays negative under both fixed mixtures (-0.0353, -0.0040). Broad domain matching alone therefore does not isolate a source-independent temporal effect. Binoculars is reported pooled only ($0.1330 \to 0.1077$): no per-document post-2023 scores survive. The pre-2015 side is provenance-verified, the post-2023 side presumed human (Appendix A), so we report a flagging rate.

Frontier-aware training does not remove the effect, and it is genre-dependent. Across three seeds it flags 0.0002 ± 0.0004 of matched pre-2015 abstracts (0.0033 ± 0.0004 of all pre-2015 human test text) but $\mathbf{0.119 \pm 0.038}$ of the matched post-2023 probe, two orders of magnitude above its own pre-2015 rate, within both sources. Yet on the retained contemporary Wikinews evaluation (Table 13) the same detector flags $0/267$ while the legacy detector flags $89/267$: no single rate represents contemporary human text.

A design tension, examined. Pairing 2026 AI text with a *pre-2015* human reference could itself inflate the frontier gap. Swapping in the *matched post-2023* human split changes the human side and the class mixture, so we compare class-conditional rates rather than macro-F1: AI recall is unchanged by construction (same AI documents, same detector), and the human flagging rate on matched abstracts is *lower* on post-2023 text (0.168) than on pre-2015 (0.224; Table 11). The

Table 5: Content-and-provider transfer. The shipped split shares its prompt pool across providers; the grouped split shares nothing. Held-out provider recall is the N-weighted mean over four leave-one-provider runs. **Columns are seed-42 runs on different test partitions** (5,402 vs. 5,407): the grouped split partitions whole prompt/parent-document clusters, whose integer sizes cannot hit an identical total. So the comparison is same-seed but not paired; 0.844 is the seed-42 value, not the three-seed 0.845 of Table 27.

	shipped split	grouped split	Δ
In-distribution macro-F1	92.76	89.71	−3.05
In-distribution AI recall	0.859	0.797	−0.062
Held-out provider recall (N-weighted)	0.844	0.764	−0.080
Transfer gap (in-dist. − held-out)	**0.015**	**0.033**	+0.018
Human false-positive rate	0.0037	0.0015	−0.0022

pre-2015 reference therefore does not create the AI-side collapse. The era gap does leave an *orthographic* trace, but the frontier-aware headline does not rest on it: normalizing it away moves macro-F1 by -0.02 $[-0.06, +0.00]$ and flips 1 prediction in 5,402 (the *legacy* detector is the normalization-sensitive one, in the opposite direction to the confound this rules out; Appendix D).

Interpreting the detector. On legacy data the gate is BERT-dominant (β=0.958), yet ablating the graph/gate/aux heads drops macro-F1 from 83.65 to 80.19: the objective *regularizes* rather than hard-routes (Section 5.1). The detector reaches 99.6 on in-training HC3 (Guo et al., 2023) versus 83.65 on COLING-2025 and 78.45 on frontier text, and transfers to MAGE and RAID, so its signal is not COLING-specific (Appendix D).

5.1 Closing the Frontier Gap: Frontier-Aware Training and Cross-Provider Generalization

The degradation above is measured *legacy-to-frontier*: our detector was trained only on legacy (2023-era) data and evaluated with zero exposure to FrontierBench generations. That is not true of every baseline — Advacheck and the released DetectAnyLLM checkpoint have undocumented or later training data — so the comparison is like-for-like only for the legacy-trained systems. It is a supervised detector and is never counted as zero-shot. Does the gap close if the detector is allowed to see frontier text? We retrain the *same* architecture on FrontierBench's training split under the identical leakage-free protocol (checkpoint dev-selected on a balanced held-out split, test evaluated once).

(i) In-distribution recovery. Frontier-aware training lifts frontier macro-F1 from 78.45 to **92.76** — a paired gain of $+14.31$ $[+13.06, +15.62]$ — and, most consequentially for deployment, cuts the human false-positive rate from 0.219 to **0.004**, with AI recall rising to 0.859. On the content-controlled grouped split, where no prompt text and no parent document crosses a boundary, the same configuration reads **89.71** (Table 5), and 89.78 ± 0.16 across seeds 42/43/44 (Appendix G) — the conservative headline we would have a reader carry away. The shipped-split configuration is seed-stable in *aggregate* (92.82 ± 0.08), so Finding 1b's caution does not threaten the large adaptation gain itself — but it still applies to *component-level* comparisons here (Appendix I). The false-positive result is the cleanest in the paper (590 legacy false positives corrected against 6 introduced). Five of fifteen per-generator changes are significant after Holm correction, all gains; o3's apparent recovery ($0.360 \rightarrow 0.600$) is not among them at n=25 (Appendix E).

Content control (Table 5). We retrained the frontier-aware model and all four leave-one-provider holdouts on a re-split where nothing crosses a boundary—no prompt text, no parent document—verified before training. Absolute levels fall: $92.76 \rightarrow 89.71$ macro-F1 and $0.844 \rightarrow 0.764$ held-out provider recall (same-seed on different partitions, not paired; three-seed figures in Table 27 and Appendix G). **The transfer gap stays small in both regimes, but it is not unchanged**: 0.015 shipped against 0.033 grouped. It roughly doubles under content control, and we say so rather than round it away. What survives is the substantive claim: the N-weighted mean recall gap is 1.5 points shipped and 3.4 ± 0.3 points content-controlled, the human false-positive rate is *lower* under grouping (0.0015), and the shipped split scores higher on every level without changing whether transfer holds. We recommend the grouped split for any content-generalization claim.

(ii) Generalization to unseen models, providers, and capability tiers (Table 27). An in-distribution number alone could reflect memorizing the training generators, so we stress-test at three severities, each retrained from scratch with the held-out data removed: one model (siblings kept); an entire provider; and training only on older-tier 2026 models, testing on the newer tier. Each held-out figure is a three-seed mean within 3 points of its own ceiling — $\mathbf{0.856}/0.859$, $\mathbf{0.845}/0.859$, $\mathbf{0.725}/0.754$, seed spreads ≤ 0.006 — with human FPR ≤ 0.004 throughout. **The capability-tier figure is not uniform**: it is carried by Grok-4.20, and under equal provider weighting the tier detector falls *below* the legacy baseline (Appendix F).

(iii) The gate activates on frontier text—but activation is not contribution. In the *reported single runs*, mean β falls from 0.958 (legacy) to 0.634 and 28.5% of documents become graph/linguistic-dominant ($\beta < 0.5$), versus 0% on legacy data. Across the three ablation seeds the direction holds but the magnitude is seed-sensitive: mean β 0.663 ± 0.024 and $18.4\% \pm 9.1$ percentage points below 0.5 (Appendix I). The one-sided floor penalises only $\beta < 0.3$, so the shift is learned, not induced. **Our ablation refuses the natural reading that the linguistic layer carries frontier transfer**: removing the graph/gate branch costs 0.086 macro-F1, below the 0.233 seed spread, and a three-seed shuffled-target comparison establishes no content-specific mechanism. We report the shift and decline the mechanistic claim (Appendices C, I).

Residuals. Gains are not uniform: Google/Gemini was already easiest (0.925) and shows no gain, and several flagship/reasoning cells stay low in-distribution (o3 0.600, GPT-5.5 0.477, Gemini-3.1-pro 0.581, Claude-Opus-4.7 0.532); none is significant after Holm correction, so these are point estimates on small cells, not regressions.

The spread tracks *document length*, not generator difficulty: frontier-aware recall is 1.000 on every document $\geq$ 1,000 characters, for all fifteen generators, while documents under 200 characters are the residual (0.387 on this population). Generators differ in how much short text their sample contains, and that composition alone orders the column. **That long-form accuracy is bought by giving up short-form**: on the *same* independent E2 short slice (n=516, <200 chars) legacy recovers 0.822 and frontier-aware 0.140. That is a same-population trade-off; the 0.387 above is a different population (native FrontierBench test), and the two are not interchangeable (Appendix E).

The degradation finding rests on the pattern across generators *and* detectors, not any single small cell.

6 Discussion and Limitations

Three consequences follow. *A high aggregate score is not a deployment claim*: the checkpoint with the highest *external* frontier macro-F1 (84.60; our frontier-aware model reaches 92.76) also flags 48.78% of our post-2023 presumed-human probe. *Aggregation can change conclusions*: one detector's pooled era direction does not survive source standardisation and another's interval coverage depends on the weighting, while capability-tier transfer changes sign between document-weighted and provider-macro averaging — though several detectors do retain large positive standardised differences. *Clean accuracy is not robustness*: in our DIPPER (Krishna et al., 2023) experiment frontier-aware recall falls $0.82 \to 0.09$ ($329/400 \to 37/400$) — our own measurement, one attack, no watermarking evaluation. So the averaging rule must be stated with the result, and the primary deployment hazard is genre-dependent false alarms on contemporary human text: no detector here should be a sole arbiter.

Limitations. English-only; one 2026 snapshot; partly redistributable. Our architecture's novel components are small relative to seed variance, so the detector is an instrument for the audit, not this paper's contribution. The selection-bias ledger is partial, so $+1.72$ is an endpoint difference, not a measured selection effect (Table 15).

7 Conclusion

Reliable detection claims depend on how detection is evaluated. Frontier performance is heterogeneous, not uniformly collapsed; the best external checkpoint also flags nearly half our contemporary presumed-human probe; and aggregation choices can invert conclusions. We release FrontierBench-2026 dataset, code and harness.

REPRODUCIBILITY STATEMENT

We release the benchmark (with datasheet and decontamination audit) at an anonymized, non-tracking repository:

`https://osf.io/vka6s/?view_only=54d0b9a3f84f428ba23b3c1b77345a0e`

We provide the detector and baseline code, the leakage-free evaluation harness, and the bootstrap/McNemar scripts. **Per-sample predictions are released**, and every *benchmark* record carries per-sample provenance (generator or human source, provider, prompt/generation metadata, license, split), so *dataset-derived* tables can be recomputed. The exceptions below are not covered by that statement and are listed rather than implied.

Exceptions, stated explicitly. Per-record inputs, predictions and provenance are *not* available for every table, and we list where: *(a)* the 267-document Wikinews evaluation (Table 13) retains its predictions, but the input documents, the loaded checkpoint identity and the seed are not recoverable, so it is a retained evaluation and not independently rerunnable; *(b)* Binoculars retains matched *pre-2015* per-document official scores but only an *aggregate* post-2023 value, which is why it is absent from every source-stratified and standardised analysis, and why no within-source claim is made for it; *(c)* for DetectAnyLLM the correspondence between the pinned public release and the published experimental artifact is unresolved (our execution scored AUROC 0.9794 against a published 0.9370), and its paraphrase slice was **not run** because the DIPPER-paraphrased inputs were not retained — no adversarial number for it appears anywhere in this paper; *(d)* original benchmark-construction logs do not retain per-call attempted/empty/error counts for every generator, so **original generation failure rates cannot be reconstructed** and all reported per-generator rates are conditioned on surviving outputs (survivorship conditioning); later expansion-arm failure rates are not a substitute and are not presented as one. Ledgers and per-file status are in `ledgers/` and the release manifest. Section 5 and Appendix C specify the evaluation protocol, model selection, hyperparameters, and fixed random seeds; Appendix A and B document data sources, licenses, and the generation grid. For our leakage-free evaluations, our trained configurations use development-set checkpoint selection; third-party released checkpoints are evaluated as supplied. The operating-point convention used for every TPR-at-fixed-FPR figure is stated in Appendix C and ships as an explicit, logged parameter of the harness rather than an implicit default.

Two limits on regeneration are worth stating plainly, because both bear on what "reproducible" can mean for a benchmark of proprietary generators. First, **regeneration depends on provider API availability**: we release the prompts, decoding metadata and model snapshots needed to reissue any call, but the endpoints themselves change. When we re-generated a subset in August 2026, three of four generators would no longer accept the decoding parameters their own records carry (Appendix B), so exact regeneration was not possible for them at any effort. Second, and for the same reason, **recorded metadata should be verified**—including ours. Score-array orientation should be confirmed against a known-label subset before use, and a `decoding` block should be read as *intended* configuration unless it has been checked against provider behaviour. We give these warnings because we found both failure modes in our own artifacts during the audit above, not hypothetically.

ETHICS STATEMENT

AI-text detectors are increasingly used in high-stakes settings such as academic-integrity enforcement and content moderation, where false positives can unjustly penalize human authors. The finding that our legacy-trained detector mislabels 22% of genuine human text and misses 64% of a frontier reasoning model's output underscores that such tools must not be used as sole decision-makers. **The contemporary-text risk is the operative one.** On the matched post-2023 abstract probe our frontier-aware detector flags 11.95% (seed s.d. 3.83 percentage points across seeds 42/43/44 — a spread, not a confidence interval), and the released DetectAnyLLM checkpoint flags 48.78%. That probe is *presumed*-human, openly licensed arXiv and PubMed abstracts weighted towards 2026; undisclosed AI assistance in it cannot be ruled out, so these are flagging rates, not verified false-positive rates. They are also genre-specific: on the retained 267-article Wikinews evaluation the same frontier-aware detector flags $0/267$. That $0/267$ is a single retained evaluation on one genre

— it is not evidence of zero population risk, and it is not independently rerunnable (Table 13). We recommend human-in-the-loop use and, for governance, generation-time provenance such as watermarking — a recommendation based on the provenance literature, not on any watermarking experiment of ours. The human reference is strictly pre-2015 (release terms per source in Appendix A); the post-2023 probe consists of openly licensed arXiv and PubMed abstracts. Machine-generated text is released under the applicable provider terms. The benchmark is intended for detection research and evaluation, not for training evasive generators; releasing it lets defenders measure and close the gaps we document, which we judge to outweigh the residual risk.

AI USE STATEMENT

We disclose our use of AI assistants as follows. They were used (i) to suggest candidate experimental designs for several of the robustness and generalization analyses (strong-baseline, paraphrase, cross-benchmark, broadened-human-probe, and per-family experiments), (ii) to write and execute the corresponding experiment and analysis code, and (iii) to draft portions of the text and tables, and for routine copy-editing. Generative AI tools were also used to provide feedback on experimental design and to assist in interpreting results. They were not used to generate or label the human-side data, to formulate or prove mathematical claims, or to clean or relabel the released datasets.

All results were verified independently by the authors against a results ledger that records, for every number in this paper, the artifact it was read from and whether it has been independently recomputed. That verification used released per-sample prediction files *where they exist*; for the retained-only results listed in the Reproducibility Statement it rests on the surviving aggregate artifacts, and we say so rather than implying per-sample coverage everywhere. Where a question could not be settled from the artifacts, we report it as undetermined instead of resolving it in our own favour.

The authors reviewed all AI-assisted text and code, take full responsibility for the final content, and confirm that every claim is supported by the released artifacts. (Machine-generated text separately appears as *data* in FrontierBench-2026; its generation is documented in Appendix B.)

REFERENCES

Mervat Abassy, Kareem Elozeiri, Alexander Aziz, Minh Ngoc Ta, Raj Vardhan Tomar, Bimarsha Adhikari, Saad El Dine Ahmed, Yuxia Wang, Osama Mohammed Afzal, Zhuohan Xie, Jonibek Mansurov, Ekaterina Artemova, Vladislav Mikhailov, Rui Xing, Jiahui Geng, Hasan Iqbal, Zain Muhammad Mujahid, Tarek Mahmoud, Akim Tsvigun, Alham Fikri Aji, Artem Shelmanov, Nizar Habash, Iryna Gurevych, and Preslav Nakov. LLM-DetectAIve: A tool for fine-grained machine-generated text detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 336–343, Miami, Florida, USA, 2024.

Navid Ayoobi, Sadat Shahriar, and Arjun Mukherjee. Beyond easy wins: A text hardness-aware benchmark for LLM-generated text detection. *arXiv preprint arXiv:2507.15286*, 2025.

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 (ICLR)*, 2024.

Guangsheng Bao, Lihua Rong, Yanbin Zhao, Xiao Yu, Qiji Zhou, and Yue Zhang. Triospect: A three-dimensional framework for robust statistical AI-generated text detection against diverse attacks. *arXiv preprint arXiv:2606.31074*, 2026.

Souradip Chakraborty, Amrit Singh Bedi, Sicheng Zhu, Bang An, Dinesh Manocha, and Furong Huang. Position: On the possibilities of AI-generated text detection. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Evan N. Crothers, Nathalie Japkowicz, and Herna L. Viktor. Machine-generated text: A comprehensive survey of threat models and detection methods. *IEEE Access*, 11:70977–71002, 2023.

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, Bangkok, Thailand, 2024.

Liam Dugan, Andrew Zhu, Firoj Alam, Preslav Nakov, Marianna Apidianaki, and Chris Callison-Burch. GenAI content detection task 3: Cross-domain machine generated text detection challenge. In *Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect)*, pp. 377–388, Abu Dhabi, UAE, 2025. arXiv:2501.08913.

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 (MM '25)*, 2025.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

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

German Gritsai, Anastasia Voznyuk, Andrey Grabovoy, and Yury Chekhovich. Advacheck at GenAI detection task 1: AI detection powered by domain-aware multi-tasking. In *Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect)*, Abu Dhabi, UAE, 2025. arXiv:2411.11736.

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.

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 (ICML)*, 2024.

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

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1808–1822, 2020.

Tom Kempton, Viktor Drobnyi, Maeve Madigan, and Stuart Burrell. Log-likelihood, simpson's paradox, and the detection of machine-generated text. *arXiv preprint arXiv:2605.06294*, 2026.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Manli Shu, Khalid Saifullah, Kezhi Kong, Kasun Fernando, Aniruddha Saha, Micah Goldblum, and Tom Goldstein. On the reliability of watermarks for large language models. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Ryuto Koike, Masahiro Kaneko, and Naoaki Okazaki. OUTFOX: LLM-generated essay detection through in-context learning with adversarially generated examples. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 21258–21266, 2024.

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 (NeurIPS)*, volume 36, 2023.

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, Bangkok, Thailand, 2024.

Chengzhi Mao, Carl Vondrick, Hao Wang, and Junfeng Yang. Raidar: geneRative AI detection viA rewriting. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

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 (ICML)*, 2023.

Rafael Rivera Soto, Barry Chen, and Nicholas Andrews. Mitigating paraphrase attacks on machine-text detectors via paraphrase inversion. *arXiv preprint arXiv:2410.21637*, 2024.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can AI-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*, 2023.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, 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.

Anyang Song, Ying Cheng, Yiqian Xu, and Rui Feng. MAGA-Bench: Machine-augment-generated text via alignment detection benchmark. *arXiv preprint arXiv:2601.04633*, 2026.

Kevin Stowe and Kailash Patil. Spotlights and blindspots: Evaluating machine-generated text detection. In *Proceedings of the 2026 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC 2026)*, 2026.

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

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

Eduard Tulchinskii, Kristian Kuznetsov, Laida Kushnareva, Daniil Cherniavskii, Serguei Barannikov, Irina Piontkovskaya, Sergey Nikolenko, and Evgeny Burnaev. Intrinsic dimension estimation for robust detection of AI-generated texts. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. Authorship attribution for neural text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8384–8395, 2020.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations (ICLR)*, 2018.

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 (NAACL)*, 2024.

Mingyi Wang, Zhuoer Shen, Yuheng Bu, and Shaofeng Zou. Detector-evasive LLM paraphrasing via constrained policy optimization. *arXiv preprint arXiv:2606.00392*, 2026.

Pengyu Wang, Linyang Li, Ke Ren, Botian Jiang, Dong Zhang, and Xipeng Qiu. SeqXGPT: Sentence-level AI-generated text detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, Alham Fikri 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)*, Bangkok, Thailand, 2024a. arXiv:2402.11175.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, Chenxi Whitehouse, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. SemEval-2024 task 8: Multidomain, multimodel and multilingual machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pp. 2041–2063, Mexico City, Mexico, 2024b.

Yuxia Wang, Artem Shelmanov, Jonibek Mansurov, Akim Tsvigun, Vladislav Mikhailov, Rui Xing, Zhuohan Xie, Jiahui Geng, Giovanni Puccetti, Ekaterina Artemova, Jinyan Su, Minh Ngoc Ta, Mervat Abassy, Kareem Ashraf Elozeiri, Saad El Dine Ahmed El Etter, Maiya Goloburda, Tarek Mahmoud, Raj Vardhan Tomar, Nurkhan Laiyk, Osama Mohammed Afzal, Ryuto Koike, Masahiro Kaneko, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. GenAI content detection task 1: English and multilingual machine-generated text detection: AI vs. human. In *Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect)*, pp. 244–261, Abu Dhabi, UAE, 2025. arXiv:2501.11012.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*, 2024.

Debora Weber-Wulff, Alla Anohina-Naumeca, Sonja Bjelobaba, Tomáš Foltýnek, Jean Guerrero-Dib, Olumide Popoola, Petr Šigut, and Lorna Waddington. Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, 19(1):26, 2023.

Chenwang Wu, Yiu-ming Cheung, Bo Han, Shuhai Zhang, and Defu Lian. Multi-level contextual token relation modeling for machine-generated text detection. *arXiv preprint arXiv:2605.16107*, 2026a.

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 (NeurIPS), Datasets and Benchmarks Track*, volume 37, pp. 100369–100401, 2024.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia S. Chao, and Derek F. Wong. A survey on LLM-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, 51(1), 2025.

Junchao Wu, Yefeng Liu, Chenyu Zhu, Hao Zhang, Zeyu Wu, Tianqi Shi, Yichao Du, Longyue Wang, Weihua Luo, Jinsong Su, and Derek F. Wong. DetectRL-X: Towards reliable multilingual and real-world LLM-generated text detection. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 38247–38294, 2026b. arXiv:2605.15518.

Xianjun Yang, Wei Cheng, Yue Wu, Linda Petzold, William Yang Wang, and Haifeng Chen. DNA-GPT: Divergent n-gram analysis for training-free detection of GPT-generated text. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Yiwei Zha, Rui Min, and Shanu Sushmita. PADBen: A comprehensive benchmark for evaluating AI text detectors against paraphrase attacks. *arXiv preprint arXiv:2511.00416*, 2025.

Qihui Zhang, Chujie Gao, Dongping Chen, Yue Huang, Yixin Huang, Zhenyang Sun, Shilin Zhang, Weiye Li, Zhengyan Fu, Yao Wan, and Lichao Sun. LLM-as-a-coauthor: Can mixed human-written and machine-generated text be detected? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 409–436, Mexico City, Mexico, 2024. arXiv:2401.05952.

APPENDICES

OVERVIEW

Figure 1 summarises the problem, the approach and the result in one diagram. It is placed here rather than in the main text for space.

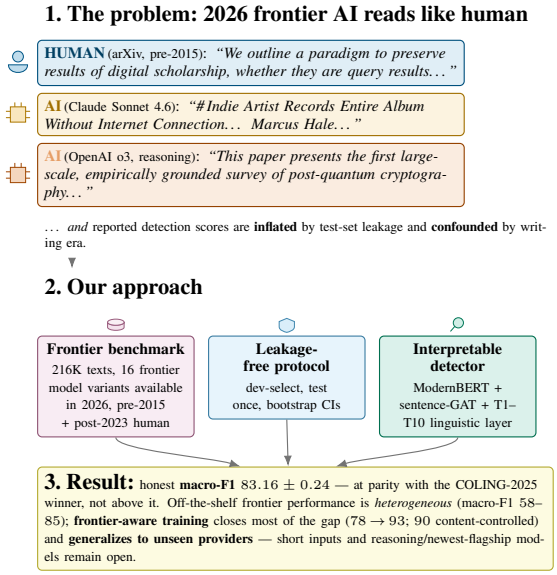

Figure 1: **Overview.** *Problem:* 2026 frontier text is fluent and human-like, and reported accuracy is sensitive to test-informed selection and confounded by writing era. *Approach:* a frontier benchmark, a leakage-free protocol, and an interpretable graph-fusion detector. *Result:* under honest evaluation our detector is at parity with the COLING-2025 shared-task winner rather than above it; off-the-shelf frontier performance is heterogeneous, while frontier-aware training closes most of the gap and generalizes to unseen providers.

TERMINOLOGY

Three terms recur throughout and are defined here once.

Legacy A detector trained only on pre-2024 data, or a corpus drawn from pre-2024 generators. Our *legacy* detector is the COLING-2025-trained model of Table 2; it has never seen a 2026 generation. "Legacy" describes the *training distribution*, not the architecture—the legacy and frontier-aware detectors in Table 21 are the same architecture trained on different data.

Training-free ("zero-shot") A detector that requires *no training on labelled AI/human pairs*. That is the defining property; it says nothing about mechanism. Training-free detection includes at least proxy-LM likelihood/curvature scoring, observer/performer perplexity ratios, regeneration-based comparison and geometric/embedding methods. **The training-free detectors we actually evaluate are the proxy-LM subset** — Binoculars (observer/performer ratio) and Fast-DetectGPT (curvature/sampling discrepancy) — whose only free parameter is a decision threshold; our training-free results should be read as statements about that subset, not the category. Two distinctions matter. The RoBERTa-OpenAI

detectors and Advacheck are *supervised* and never counted as zero-shot. DetectAnyLLM uses the Fast-DetectGPT scoring *framework* but is optimised with a task-oriented loss and ships trained weights, so it is a *trained discrepancy detector* — neither zero-shot nor a conventional supervised encoder. Regeneration-based (DNA-GPT) and geometric/intrinsic-dimension methods form further paradigms that we do not evaluate.

Frontier-aware The same architecture as our legacy detector, retrained with FrontierBench-2026 generations in the training split under the identical leakage-free protocol. Every held-out result in Appendix F removes the evaluated generators from both train and dev and retrains from scratch.

A HUMAN-SOURCE COMPOSITION, LICENSING, AND RELEASE

Table 6 lists the five provenance-verified pre-2015 human source families. Their release terms are source-specific and are reported in the table; all predate widespread LLM assistance. The domain-matched post-2023 human probe (1,105 documents) is drawn from openly licensed arXiv and PubMed abstracts dated 11.9% 2024, 13.2% 2025 and 74.9% 2026; unlike the pre-2015 corpus, which provably predates widespread LLM assistance, limited undisclosed AI assistance in this recent text cannot be ruled out — it is *presumed*, not verified, human. For the domain-generalization check in Finding 3 we additionally collect a non-scientific post-2023 slice of 267 *Wikinews* 2024–25 news articles (CC-BY 2.5), selected by category-timestamp window and decontaminated against the abstract probe by 200-character prefix hash. **This slice is an exception to the release statement**: its 267 *predictions* survive, but the input documents, the loaded checkpoint identity and the seed were not retained, so it does *not* ship with the same per-record provenance as the rest of the human side and is not independently rerunnable. Redistribution of machine-generated text follows each provider's usage terms; where terms preclude raw redistribution we release generation prompts and metadata sufficient to regenerate the text, with per-record provenance (generator, prompt, decoding, timestamp, license, split) for every record so that dataset-derived result tables can be recomputed from the released artifacts.

Table 6: Human-source composition of FrontierBench-2026 (all pre-2015).

Source	Docs	License
Project Gutenberg	40,823	Public domain
PubMed Central OA	19,983	CC-BY
Wikipedia (2014)	18,883	CC-BY-SA
Wikinews (2014)	14,951	CC-BY
arXiv (pre-2015)	9,661	non-exclusive
Total human	104,301	
Total AI (16 models)	112,051	provider ToS

Subset `frontierbench-2026-smalln-ext-v1`. *Purpose:* the four generators with the fewest native test samples (n=25–50) carry intervals too wide to support per-generator claims, so this subset adds 250 fresh samples each to narrow them and, separately, to test whether the native cells reproduce. *Generated* August 2026, in two arms, each with full per-record provenance (model snapshot string, prompt, prompt id, the exact decoding dict *sent*, the decoding dict the benchmark *records* for that cell, usage, finish reason, timestamp).

Arm 1 uses the providers' default reasoning effort with output budgets raised to $\geq$ 4000 tokens, because at the native short budgets reasoning tokens consume the entire budget before any visible text is emitted. *Arm 2* uses the benchmark's native budgets at low reasoning effort, and doubles as a **decoding-configuration sensitivity probe**: it is the arm whose sampling parameters match the recorded native grid, so it is the arm most fresh-vs-native comparisons come from. **"Matched decoding" applies only to the cells actually matched**: the other OpenAI arms are alternate-decoding *sensitivity probes*, not matched replications, and are labelled as such wherever they are reported. Retained request bodies and their verification dates are in the release manifest; no new paid generation calls were made for this revision.

Status taxonomy. Every record carries one of four values: `ok`, `empty_output` (the call returned no visible text, always with `finish_reason=length`), `api_refused` (a genuine parameter or model rejection), and `transient_failed` (reserved for rate-limit, 5xx, connection and timeout failures after exhausted retries—it did not occur in either arm). Rate-limit events are never recorded as refusals: infrastructure noise must not be attributable to model behaviour. Counts (`ok`/`empty_output`/`api_refused`/`transient_failed`) are $978/22/0/0$ for Arm 1 and $926/71/3/0$ for Arm 2.

Effective N and the empty-output rate. Records with no visible text cannot be scored as AI text and are excluded, with N reduced accordingly; they are **retained in the release with full configuration**, because "default reasoning effort exhausts even a raised budget on a minority of long-form calls" is itself documented API behaviour. We did not regenerate the failures: doing so at a larger budget would impose a length-conditioned selection effect, leaving a set that is no longer a sample from one generation process.

Table 7: `smalln-ext-v1`: empty-output rate and effective N per arm. Raising the output budget suppresses reasoning-token exhaustion more effectively than lowering reasoning effort does.

	Arm 1 (default effort, budget ≥ 4000)		Arm 2 (low effort, native budget)	
Generator	empty %	effective N	empty %	effective N
o3	5.6%	236	**21.4%**	195
o4-mini	2.4%	244	0.0%	250
GPT-5.5	0.8%	248	6.0%	234
Gemini-2.5-pro	0.0%	250	1.2%	247
total	*2.2%*	*978*	*7.1%*	*926*

Duplicate prompts, by design. Prompts are drawn *with replacement* from each domain pool, so the subset contains repeated `(generator, prompt id, decoding)` keys—32 recur among the first 129 records, one three times. This is part of the sampling scheme rather than an error, but we state it because **anyone deduplicating by prompt would shrink the set unexpectedly**.

Decontamination. All `ok` records were audited against the FrontierBench train, dev and test splits by 200-character prefix hash. Arm 1 has 83 collisions and Arm 2 has 63. Inspecting them matters: *every* matched document is AI-labelled—not one human document—77 of 83 and 60 of 63 are exact full-text matches with a median length of about 60 characters, and they are almost entirely `headlines` whose benchmark counterparts were produced by the same or a sibling model from the same prompt pool. These are therefore *convergent short-form generations*, not training-data leakage: a tightly constrained headline prompt drives different frontier models to the same string. Only 2 (Arm 1) and 1 (Arm 2) collide with a split any detector was fitted on. We therefore report two exclusion policies: **policy B (primary)** drops only those fitted-split overlaps, giving 976 and 925 records; **policy A (sensitivity)** drops every collision, giving 895 and 863 and costing GPT-5.5 roughly a third of its samples to remove text no detector was ever fitted on. *The qualification outcome in Table 22 is identical under both.*

A warning that applies to this release, not only to others. Two kinds of recorded metadata in our artifacts turned out not to describe what ran, and we found both ourselves. Users should therefore (i) **confirm score-array orientation against a known-label subset** before using any released score array, rather than trusting a recorded orientation flag, and (ii) **treat a `decoding` block as *intended* configuration** unless it has been verified against current provider behaviour (Appendix B). We state these as properties of the artifact rather than as general advice because each is evidenced by a specific failure we had to correct.

B GENERATION GRID

Each provider is prompted with the same 1,860-entry prompt set (604 distinct texts) across 13 domains (essays, news, scientific abstracts, stories, reviews, question-answering, code explanations, headlines, tweets, one-liners, definitions, long essays, long stories) under a decoding grid (temperature $\in \{0.5, 0.7, 1.0\}$, top-p=0.95), with a length-extreme strategy targeting very short ($<$ 25-word)

Table 8: **Release coverage of AI records, by provider** — what the release package *contains*, counted from the split files (`verify/release_coverage_audit.py`; ledger `ledgers/release_coverage_audit.json`). Categories are mutually exclusive and sum to the 112,051 AI records; "eval-split" columns use the dev+test denominator (5,402 AI records). In the package audited here **every AI record carries raw text**, so the prompt/metadata-only and restricted categories are empty; the corpus-level "partly redistributable" caveat applies to *human* source families, whose terms differ by source (Table 6). **Two limits on this table.** It audits *content*, not permission: the `license` field is uniformly `CC-BY-4.0` as shipped by the pipeline, and we do not infer any redistribution right from that string. And it describes the local release package; **the live hosted copy was not inspected**, so correspondence between this package and what is served at the anonymized link is *unverified* here.

	whole corpus				eval splits	
Provider	raw text	prompt-only	restricted	% raw	raw text	% raw
Anthropic	25,845	0	0	100.0%	695	100.0%
Google	27,316	0	0	100.0%	2,033	100.0%
OpenAI	23,196	0	0	100.0%	1,008	100.0%
xAI	35,694	0	0	100.0%	1,666	100.0%
Total	**112,051**	**0**	**0**	**100.0%**	**5,402**	**100.0%**

Table 9: The 16 frontier generators in FrontierBench-2026 (2026 snapshot). **15 of the 16 are evaluated**: Gemini 3-flash has no held-out test cell, so it appears in the collection grid here but in no per-generator results table (Section 3, Table 4).

Provider	Variants
Anthropic	Claude Opus 4.7, Sonnet 4.6, Haiku 4.5
OpenAI	GPT-5.5, GPT-5.4-mini, o3, o4-mini
Google	Gemini 3.1-pro, 3-flash, 2.5-pro/flash/lite
xAI	Grok-4.20 (reas. & non), Grok-4-fast, Grok-3

and very long ($>$ 3000-word) outputs. The aim is to reduce *trivial* separation by length; it does **not** demonstrate that length carries no signal, and the intermediate band is not well characterised here — only 15 of 2,685 AI test documents fall between 200 and 999 characters. Reasoning models are queried in both reasoning and non-reasoning configurations where available, yielding matched pairs that isolate the effect of reasoning-tuned decoding. Provenance (model, prompt, decoding, timestamp) is retained per record.

Per-model decoding, and a metadata caveat. The grid above applies to the standard-sampling models. The three reasoning-mode variants (o3, o4-mini, Grok-4.20-reasoning) carry no top-p in their records, because those endpoints reject it.

One caveat, identified after generation and disclosed here rather than silently corrected: **GPT-5.5's records carry a `temperature` and a `top_p` that the endpoint does not accept.** Re-queried in August 2026, both surfaces refuse both parameters for this model, and the reported snapshot (`gpt-5.5-2026-04-23`) is contemporaneous with generation, so there is no evidence the API changed in between. The harness classified GPT-5.5 as a standard-sampling model—its records include top-p where the three reasoning variants' correctly omit it—so the recorded values are best read as *intended* rather than applied configuration. No generation log survives to establish whether the calls errored and were retried without the parameters or the parameters were dropped silently, and we do not claim to know. **The decoding parameters actually applied to these records remain unverified**; the text itself is retained and scored as generated. We flag it because a benchmark's per-record metadata is a claim like any other, and this one we cannot substantiate.

C Training objective and hyperparameters

We minimize

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_{\text{aux}}\mathcal{L}_{\text{aux}} + \lambda_{\text{dom}}\mathcal{L}_{\text{dom}} + \lambda_{\text{kl}}\mathcal{L}_{\text{kl}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \lambda_{\text{gate}}\mathcal{L}_{\text{gate}}, \tag{2}$$

where $\mathcal{L}_{\text{CE}}$ is label-smoothed cross-entropy on the fused logits, $\mathcal{L}_{\text{aux}}$ regresses the 65 features from the pooled representation, $\mathcal{L}_{\text{dom}}$ is a domain-adversarial cross-entropy passed through a gradient-reversal layer (Ganin et al., 2016), $\mathcal{L}_{\text{kl}}$ is a symmetric KL term aligning the two branches' logits, $\mathcal{L}_{\text{align}} = 1 - \langle z_g, z_b \rangle$ is a cosine term pulling the two branches' *pooled document* representations together: each branch's pooled vector is mapped into a shared 128-dimensional space by its own learned projection ($\text{Linear}(2d \to d)\,\text{ReLU}\,\text{Linear}(d \to 128)$ for the graph side, $\text{Linear}(1024 \to 256)\,\text{ReLU}\,\text{Linear}(256 \to 128)$ for the backbone), then L_2-normalised, and the loss is reduced over *documents* in the batch (not over sentences); and $\mathcal{L}_{\text{gate}} = \text{ReLU}(0.3 - \beta)^2$ (averaged over documents) is an asymmetric floor that penalizes only $\beta < 0.3$, discouraging collapse to graph-only while leaving $\beta \to 1$ unpenalized; schedules for all coefficients are in Table 10. We fine-tune for 5 epochs (effective batch 128), unfreezing the last 6 backbone layers at learning rate 2×10^{-6} and the graph/gate/aux heads at $1\text{–}2 \times 10^{-4}$, with label smoothing 0.15 and a warmup→mid→late schedule that anneals from a fixed β=0.5 to the learned gate. The 65 features are the cross-dataset-stable subset of 186 numeric T1–T10 features; using all 186 overfits corpus-specific idiosyncrasies (Table 14).

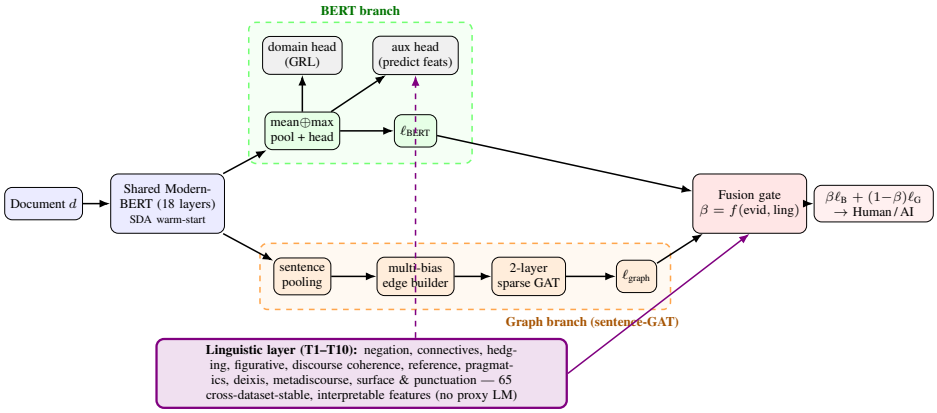

Figure 2: **Architecture.** One fine-tuned ModernBERT backbone is shared by a BERT branch (mean⊕max pool → head, with auxiliary feature-prediction and domain-adversarial regularizers) and a graph branch (sentence pooling → multi-bias edges → sparse GAT). A linguistic fusion gate computes β from graph evidence and the 65 stability-selected features and blends the two logits. *No proxy LM is queried for the linguistic signal.*

Operating-point convention. Every TPR-at-fixed-FPR figure we report uses one stated rule. Scores are oriented so that larger values indicate AI, and a document is called AI when its score is at or above the threshold. With n human dev documents sorted by score in descending order and $k = \lfloor \tau_{\text{FPR}} \cdot n \rfloor$, the threshold is the k-th largest human dev score, so that k human dev documents (plus any tied at the threshold) are called AI. This is an order-statistic rule, not a search for the maximum feasible TPR, and the target is a *nominal* dev FPR: with n=2,685, k=26 at the 1% target and 134 at 5%. Ties at the threshold can raise the achieved dev FPR above the target; among the eight *external* detectors of Table 19 (which also lists our two configurations, ten rows in all) this occurs only for RoBERTa-base-OpenAI (0.0108 and 0.0521), the rest landing at or below target. The k=0 branch sets the threshold strictly above the maximum human score and is not exercised here. Score orientation is verified per detector against dev labels before thresholding (the released MAGE checkpoint scores AI-low and is negated). Achieved test FPR at each dev-chosen threshold is reported alongside every TPR in Table 19; there is no interpolation between achievable operating points. Thresholds are chosen on dev and the test set is scored exactly once at the chosen threshold. We state this because TPR-at-fixed-FPR is sensitive to the choice—nearest-achievable FPR, most-conservative-threshold-under-target, and ROC interpolation can differ by more than a point at 1% FPR on a 5,402-item split—so an unstated convention is itself a reproducibility gap. The rule ships as an explicit, logged parameter of our evaluation harness rather than as a default buried in code.

Implementation details, corrected against the released code. A code audit for this revision found several places where the description above and in Table 10 did not match `fusion_clean.py`.

We state the verified behaviour here and have corrected the corresponding claims; line references are to the released script at the recorded commit.

Backbone. The encoder is **ModernBERT-large** (hidden size 1,024, 28 transformer layers), warm-started from a fine-tuned checkpoint and truncated by retaining layers 0–17, i.e. 18 **layers** (`:73`, `:76`, `:1161`). Earlier text did not state base-vs-large; the hidden size is 1,024, not 768.

Alignment loss. $\mathcal{L}_{\text{align}}$ operates on **document-level** representations, not sentence-level ones. The graph branch's pooled document vector and the backbone's pooled document vector are each mapped into a **shared** 128**-dimensional space** by two learned projections — $\text{Linear}(2d \to d)\,\text{ReLU}\,\text{Linear}(d \to 128)$ for the graph side and $\text{Linear}(1024 \to 256)\,\text{ReLU}\,\text{Linear}(256 \to 128)$ for the backbone side — then L_2-normalised, so the inner product is a cosine in $\mathbb{R}^{128}$ and the two dimensionalities never need to agree (`:181`, `:1192-1201`, `:1256-1257`). The reduction is a mean over the **batch**, i.e. over documents, not over sentences (`:1301-1302`).

Fusion gate. The gate consumes **six scalar evidence signals** plus the 65 linguistic features (71 inputs). Only three of the six come from the graph: edge *coverage* (fraction of valid sentence nodes receiving a learned top-k edge, $[0, 1]$), edge-score *margin* ($\tanh$ of the mean top-k margin over 5), and edge *density* ($\tanh$ of edges per valid node over 10). The other three are the backbone's decision *margin* (top-1 minus top-2 posterior), its *entropy* (base-2 normalised), and normalised document *length* ($\log n_{\text{sents}} / \log \text{MAX_SENTS}$) (`:1103-1137`). Calling all six "graph-evidence" signals, as earlier text did, is inaccurate. The gate is $\text{Linear}(71 \to 128)\,\text{ReLU}\,\text{Dropout}(0.1)\,\text{Linear}(128 \to 64)\,\text{ReLU}\,\text{Linear}(64 \to 1)$ with a learned scalar bias initialised to 1.5; β is a single scalar per document and the learned output is **clamped to** $[0.05, 0.95]$, not to $[0, 1]$ (`:1136`).

Short documents. Documents with fewer than 5 sentences are routed to the backbone by a **separate override** that sets β=1.0 after the clamp (`:1247-1251`); the override therefore reaches values the learned gate cannot. During the warmup and mid phases β is instead fixed at 0.5 for *all* documents (`:1242-1244`), so the override applies only once the learned gate is active. The graph branch is computed for every document regardless (`:1234`), and $\mathcal{L}_{\text{align}}$ and $\mathcal{L}_{\text{kl}}$ are **not masked** for 2–4-sentence documents: they contribute gradient from a graph representation the fused logits then bypass. $\mathcal{L}_{\text{gate}}$ is likewise unmasked but evaluates to exactly zero for them, since $\text{ReLU}(0.3-1.0)^2 = 0$. All three reduce with a mean over the full batch, so short documents sit in the denominator. This is a training/inference asymmetry we state rather than repair, since repairing it would invalidate the reported runs.

Domain-adversarial loss. $\mathcal{L}_{\text{dom}}$ classifies **legacy corpus provenance** — HC3, M4GT, MAGE, three classes — and not the 13 prompted generation domains (`:207-209`). Documents whose source is not one of those three receive `domain_id`$= -1$ (`:699`) and are excluded by the validity mask (`:1590-1597`). Every FrontierBench record's source is a generator name or `human`, so for all frontier-aware runs the mask selects nothing and $\mathcal{L}_{\text{dom}}$ is **identically zero**; training logs confirm it (`dom`=0.000 throughout, against 0.101–0.152 for legacy-corpus runs). The concern that AI-only domains make domain identity collinear with the AI label therefore cannot arise in these runs: the objective never sees a prompted domain. It is active only in the legacy regime.

T1–T10 feature specification (executed code). The ten families and the spaCy pipeline each was extracted with: **T1** negation & nuance, **T3** gradation & hedging, **T5** discourse structure & coherence, **T7** pragmatic & social cue density, **T9** meta-discursive & reflective language — all `en_core_web_trf`; **T2** logical bridging & connective diversity, **T4** figurative & idiomatic expression, **T6** referential & cultural breadth, **T8** grounding & deixis, **T10** rhythm, syntactic variability & advanced style markers — all `en_core_web_lg`. Extraction uses `spacy` (both models above), `numpy`, `pandas`, spaCy's `PhraseMatcher` (`attr="LOWER"`), NLTK's VADER `SentimentIntensityAnalyzer` (lexicon `vader_lexicon`) for sentiment-valenced features, and one lexical resource, `resources/idioms_en.txt` (one idiom per line), used by T4. **Missing-value handling**: if the idiom resource is absent the extractor degrades gracefully to an empty idiom list rather than failing, so a missing resource silently yields zero idiom hits — rerunners must ship the file, which is released alongside the code.

Units and standardisation. Raw feature units are heterogeneous (counts, per-token rates, ratios, entropies). *Standardisation exists and was used*: the 65 stability-selected features are z-scored with a `sklearn StandardScaler` **fit on the training split only** and then applied unchanged

to dev and test (`fusion_clean.py:507-511`); the fitted means and scales are written to `scaler_stats.json` in every run directory (`:518-522`), so the exact transform is recoverable per run. The extractor itself performs no scaling. The auxiliary head regresses these standardised targets from the pooled representation; the gate consumes the same standardised feature vector, identically at training and inference time. To score previously unseen raw text one therefore needs: both spaCy models, the VADER lexicon, the idiom resource, the released extractor, and the run's `scaler_stats.json`.

Table 10: Implementation record. "Source" is the line in the released `fusion_clean.py`. Values were read from the script that produced every reported number, not from a design document.

Item	Value	Source
Optimisation		
Optimiser	AdamW	`:1339-1345`
Weight decay	10^{-2}	`:154`
Gradient clipping	1.0 (global norm)	`:155`
Batch × accumulation	32×4 (effective 128)	`:151-152`
Epochs	5 (4 for leave-one-model-out)	`:153`
Learning rates	backbone 2×10^{-6}; graph 10^{-4}; fusion/aux/adv 2×10^{-4}	`:161-164, :212`
Schedule	warmup 5% → mid 70% → late	`:184-185`
Label smoothing	0.15	`:223`
Mixed precision	bfloat16	`:215`
Random seed	42 default, overridable via `SEED`	`:141-144`
Loss coefficients		
λ_{aux}	$0.5 \rightarrow 0.1$ (annealed)	`:202-203`
λ_{kl}	0.00/0.20/0.10 (warm/mid/late), T=2.0	`:192-195`
λ_{dom}	max 1.0, ramp over first 25%	`:210-211`
λ_{align}	0.05/0.05/0.01	`:188-190`
λ_{gate}	0.10/0.05 (mid/late)	`:197-198`
Backbone and text		
Backbone	ModernBERT, truncated after layer 17	`:73, :76`
Warm start	SDA checkpoint; last 6 layers unfrozen	`:74-75, :156`
Max tokens	256 (train) / 1024 (eval)	`:147-148`
Max sentences	128; min 2 for training	`:149-150`
Pooling	mean ⊕ max concatenation	`:1014-1020`
Dropout	0.15 (attention)	`:175`
Graph branch		
GAT	hidden 512, 4 heads, 2 layers	`:172-174`
Edges	window 12; learned top-k 4; local k 2	`:167-169`
Semantic edges	top-m 4, enabled for ≤ 50 sentences	`:170-171`
Similarity source	backbone sentence representations; *no external similarity model*	`:861-863`
Short-document augmentation (SDA)		
Truncation	p=0.5, to 1–5 sentences, 0.6 front-biased	`:218-219, :221`
Marker stripping	p=0.5, AI docs, SDA stage only	`sda_clean.py:327-328`
Features		
Gate features	65 of 186, by cross-corpus sign stability	`:487-492`
Selection statistic	sign agreement of feature/label correlation across 4 corpora	registry
Not recorded		
Temperature scaling	*not performed* — no calibration step exists in the pipeline	—
Hardware / runtime	one 48 GB GPU; wall-clock runtime not recorded	—

Pinned environment and feature list. The released `environment.yml` pins Python 3.10.19, `spacy` 3.8.7, `spacy-transformers` 1.3.9, `transformers` 4.49.0 and `torch` 2.9.0; the two pipelines are `en_core_web_trf` 3.8.0 and `en_core_web_lg` 3.8.0. **One honest discrepancy**: the interpreter used for the verification reruns reports `spacy` 3.8.11 rather than the pinned 3.8.7, so the environment file is the pinned intent and not a captured image of every historical run. **We did not record a frozen dependency snapshot at original training**

time, and we do not reconstruct one — NLTK and `scikit-learn` versions for those runs are simply not recoverable, and we state that rather than inventing them. The *selected* feature list is pinned, not described: the 65 stability-selected columns are enumerated as `stable65` in `harness/feature_schema_contract.json`, alongside the 186-numeric-feature full schema and `sha256` digests of the ordered column sets, so the exact feature identities and their order can be checked rather than inferred from prose.

Short-document augmentation. SDA is a **separate warm-start stage** (`sda_clean.py`), backbone-only, trained on the COLING-2025 English split as binary classification with label smoothing 0.15 — not masked language modelling — for 5 epochs under its **own optimiser, `AdamW`** (`sda_clean.py:445`; separate from the fusion-stage optimiser of Table 10), batch 16×8 accumulation (effective 128), LR 5×10^{-6}, weight decay 10^{-2}, clip 1.0, warmup fraction 0.05, seed 42, max 1,024 tokens, checkpoint selected on dev macro-F1. Human documents longer than 5 sentences are truncated to 1–5 sentences with probability 0.5, the window starting at sentence 0 with probability 0.6. **Marker stripping runs only in this stage and only on AI documents** (guard `label == 1 and random() < 0.5`, `sda_clean.py:327-328`). The patterns are not paraphrased here but pinned in the released script: the leading-preamble regex at `sda_clean.py:128-136`, the 15 inline phrase patterns at `:138-153`, and the seven markdown-normalisation patterns at `:157-163`. That file is the immutable specification; the line ranges are given so each pattern can be read verbatim. The identically named function in `fusion_clean.py` is defined but never called (`:318`); its constant is printed and logged but unused, because stripping text there would desynchronise the sentence map the graph branch depends on (`:663-665`). Fusion-stage truncation is applied, asymmetrically: probability 0.5 for human documents and 0.2 for AI ones (`:653`).

Complete implementation record. Table 10 lists every training and architecture constant, sourced per cell to the released training script (`fusion_clean.py`) so each value can be checked rather than taken on trust. Rows marked *not recorded* are stated as such and are *not* reconstructed: an unrecorded value that we later inferred would be indistinguishable, to a reader, from one we had logged all along.

D ADDITIONAL RESULTS

The temporal (human-era) confound of Finding 3. False positives are the number that matters for deployment, so we scored the frontier-aware configuration (EVAL_ONLY) on the matched post-2023 probe. Training against a pre-2015 human side buys a low false-positive rate on pre-2015 text, not on the text a deployed detector would actually see: across three seeds, the same configuration flags $0.02\% \pm 0.04\%$ of pre-2015 abstracts ($0.33\% \pm 0.04\%$ of all pre-2015 human test text) but $11.9\% \pm 3.8\%$ (seed s.d., not a CI) of the presumed-human post-2023 probe — which is dated 11.9% 2024, 13.2% 2025 and 74.9% 2026, i.e. weighted towards 2026. Every row of Table 11 is computed on one population per row. One limitation remains and is marked there: **Binoculars** retains only an *aggregate* post-2023 cell (119/1,105), so its Δ is indicative and it cannot enter any source-stratified or standardised analysis. Frontier-aware training closes the *generator* gap (Section 5.1) and leaves the *era* gap open. Predictions ship as `predictions/frontier_aware_post2023_probe.csv`; ledger `ledgers/post2023_frontier_aware.json`.

Notes to Table 11 — decision rules. By group. *Dev-selected macro-F1-maximising threshold*: RoBERTa-large-OpenAI, Fast-DetectGPT and DetectAnyLLM (the last at τ=30.1727, flag iff score $\geq \tau$, selected on FrontierBench dev, n=5,402, no ties) — and **also Binoculars**, at $\tau = -0.9364162087$ with AI-high orientation and flag iff score $\geq \tau$, selected on the FrontierBench dev split and on neither evaluation population. (An earlier caption called Binoculars a "published default"; that was wrong and is corrected here.) *Published model-card rule* $p \geq 0.92$: Advacheck. *Strict argmax* ($p_1 > p_0$): our legacy and frontier-aware detectors.

Notes to Table 11 — markers and provenance. ‡DetectAnyLLM is the author-released checkpoint, *not* a reproduction of the published DetectAnyLLM result (Table 19, Appendix D); its +0.4484 is the largest pooled difference in the table. †Binoculars' pre-2015 cell **is reproduced from the retained official score array** (189/1,421; PubMed 167/1,211, arXiv 22/210), but its post-2023 cell survives **only as an aggregate** (119/1,105): per-document scores were never retained, so that cell is *not* reproduced from per-document scores, its

Δ is indicative, and it is excluded from every source-stratified and standardised analysis (`ledgers/binoculars_threshold_verify.json`). The two populations are domain-matched but not source-matched (pre-2015 is 85.2% PubMed, post-2023 52.6% arXiv) and the post-2023 side is weighted towards 2026; Appendix D reports per-source rates.

Table 11: Temporal confound: human flagging rate on *matched* arXiv/PubMed abstracts, pre-2015 (n=1,421, provenance-verified) vs. post-2023 (n=1,105, presumed human). Both columns use the same population per row, and **each detector is scored under its own decision rule — these are not a common rule and the rows are not calibrated to a shared operating point**; the rules, the †/‡ markers and the provenance caveats are in the notes below. Δ is post-2023 minus pre-2015, so positive means more flagging on contemporary human text. The frontier-aware row is a three-seed mean $\pm$ s.d.; the originally reported single checkpoint is shown for traceability. Ledger: `ledgers/temporal_matched_abstracts.json`.

Detector (ref. era)	pre-2015 (n=1,421)	post-2023 (n=1,105)	Δ
RoBERTa-large-OpenAI (2019)	0.2006 (285)	0.4552 (503)	$+0.2546$
Advacheck (2025)	0.0929 (132)	0.1439 (159)	$+0.0510$
Ours, legacy-trained (2018–23)	0.2238 (318)	0.1683 (186)	-0.0555
Ours, frontier-aware (3 seeds)	**0.0002 ± 0.0004**	**0.1195 ± 0.0383**	$\mathbf{+0.1192}$
originally reported checkpoint	*0.0007 (1)*	*0.1747 (193)*	$+0.1740$
DetectAnyLLM (rel. ckpt)‡	0.0394 (56)	**0.4878** (539)	$\mathbf{+0.4484}$
Fast-DetectGPT (0-shot)	0.1161 (165)	0.0968 (107)	-0.0193
Binoculars (0-shot)†	0.1330 (189)	0.1077 (119)	-0.0253

Two detectors change direction between PubMed and arXiv. The legacy detector's pooled difference is negative (-0.0555), but that direction is not robust to source standardisation: the equal-source estimate is $+0.0050$ $[-0.0270, +0.0369]$ and the combined-population estimate $+0.0303$ $[-0.0035, +0.0652]$, both spanning zero — i.e. indistinguishable from no change once source mixture is held fixed, which is not the same as evidence of no effect. Fast-DetectGPT stays negative under every estimate (-0.0193 pooled, -0.0353 $[-0.0679, -0.0029]$ equal, -0.0040 $[-0.0334, +0.0256]$ combined): its sign does not reverse, only whether its interval covers zero. What both detectors do share is *opposite signs between sources* — positive within PubMed, negative within arXiv — so the mixture shifting from predominantly PubMed to predominantly arXiv *contributes* to their negative pooled values. That is not the whole explanation: Fast-DetectGPT stays negative under both fixed mixtures, so a within-source component persists.

The pooled and standardised figures are both valid descriptive estimates of different target populations: the pooled value describes the two released probe populations as they are, the standardised value a population whose source mixture is held fixed across eras. Neither is uniquely correct. Because the stratum-specific signs differ from each other rather than sharing a sign opposed to the pooled one, this is a composition-driven aggregation effect and *not* a classical Simpson's reversal. Aggregating across heterogeneous regions can destroy a local signal; Kempton et al. (2026) report that failure mode for token-level scores across a detector's hidden space, which we offer only as an analogy — their populations are token positions, ours are corpus sources, and we do not claim their mechanism operates here. RoBERTa-large-OpenAI, Advacheck, the frontier-aware detector and the released DetectAnyLLM checkpoint are positive pooled, within both sources, and under both weightings. Their *standardised* intervals in Table 12(b) exclude zero; that statement covers those supplied standardised estimates only and does not assert an established nonzero effect for every individual per-source contrast, which we do not interval-estimate separately.

The frontier-aware zero should be read with its interval: $0/267$ bounds the contemporary-news flagging rate at 1.42% at 95% confidence, which is not the same as zero risk, and the provenance limits above apply. Set beside the matched-abstract column of Table 11 — frontier-aware 0.1195 ± 0.0383, legacy 0.1683 — the two populations disagree in both level and direction: the frontier-aware detector flags roughly one contemporary abstract in eight but none of these contemporary news articles, while the legacy detector flags roughly one news article in three. Contemporary-human false alarms are therefore detector- and genre-dependent, and no single rate represents contemporary human text.

Table 12: Source-stratified temporal comparison and source-standardised sensitivity analysis. The pooled populations of Table 11 are matched at the scientific-abstract domain level but not in source composition: pre-2015 is 1,211 PubMed / 210 arXiv (85.2%/14.8%), post-2023 is 524 PubMed / 581 arXiv (47.4%/52.6%). Cells are flagging rates, pre / post. $\Delta^{\mathrm{std}} = \sum_s w_s(\hat{p}_{\mathrm{post},s} - \hat{p}_{\mathrm{pre},s})$ applies one weighting to both eras: *equal* $(0.5/0.5)$ targets a hypothetical evenly-sourced population, *combined* $(0.687/0.313)$ targets the observed pooled source distribution. Both weightings were fixed before the results were inspected and neither was chosen for its direction. Brackets are 95% stratified bootstrap percentile intervals over *documents*, resampled within each era×source cell (10,000 replicates, seed 20260918, detectors bootstrapped independently). For the frontier-aware detector the statistic is the mean over its three retained seeds and the document bootstrap does *not* absorb seed variation; that is reported separately as a seed s.d. and is not a confidence interval. Binoculars is absent: no per-document post-2023 official score array survives, so it cannot be stratified or standardised, and its within-source behaviour must not be inferred from its pooled value; its pre-2015 rates are PubMed 0.1379 (167/1,211) and arXiv 0.1048 (22/210). ‡DetectAnyLLM is the author-released checkpoint scored at the dev-selected threshold τ=30.1727, *not* a reproduction of the published result (Table 19, Appendix D); its intervals were computed with the same stratified document bootstrap as every other row (seed 20260918, 10,000 replicates) and it has no seed s.d. because it is a single released checkpoint. §The frontier-aware row is the *mean over seeds 42/43/44*, and its bracket is a genuine 95% document-bootstrap CI for that mean, **conditional on the three fitted models**: documents are resampled within each era×source stratum and the *same* resampled indices are applied to all three seeds, preserving document alignment, before averaging the per-seed standardised deltas (`verify/frontier_aware_meanseed_bootstrap.py`). **It covers document sampling only, not training-seed uncertainty** — that is the separate s.d. row (0.0362, 0.0285), and the two must not be conflated. Every bracket in this panel is a 95% document-bootstrap CI reproduced from its ledger at four decimal places, so table and prose endpoints are identical. Ledgers: `ledgers/temporal_source_standardised.json`, `ledgers/temporal_bootstrap_unified.json`.

(a) Within-source flagging rates.

	PubMed (1,211 → 524)		arXiv (210 → 581)	
Detector	pre / post	Δ	pre / post	Δ
RoBERTa-large-OpenAI	0.1990 / 0.3416	+0.1426	0.2095 / 0.5577	+0.3481
Advacheck (2025)	0.1057 / 0.2710	+0.1653	0.0190 / 0.0293	+0.0102
Ours, legacy-trained	0.2461 / 0.3187	+0.0726	0.0952 / 0.0327	−0.0625
Ours, frontier-aware	0.0000 / 0.0789	+0.0789	0.0016 / 0.1561	+0.1545
DetectAnyLLM (rel. ckpt)‡	0.0454 / 0.5153	+0.4699	0.0048 / 0.4630	+0.4582
Fast-DetectGPT (0-shot)	0.1082 / 0.1565	+0.0483	0.1619 / 0.0430	−0.1189

(b) Source-standardised difference Δ^{std} [doc-bootstrap 95%].

Detector	equal (0.5/0.5)	combined (0.687/0.313)
RoBERTa-large-OpenAI	+0.2454 [+0.2032, +0.2869]	+0.2070 [+0.1689, +0.2447]
Advacheck (2025)	+0.0878 [+0.0632, +0.1115]	+0.1167 [+0.0875, +0.1467]
Ours, legacy-trained	+0.0050 [−0.0270, +0.0369]	+0.0303 [−0.0035, +0.0652]
Ours, frontier-aware§	+0.1167 [+0.0996, +0.1339]	+0.1025 [+0.0856, +0.1201]
seed s.d. (not a CI)	*0.0362*	*0.0285*
DetectAnyLLM (rel. ckpt)‡	+0.4640 [+0.4330, +0.4945]	+0.4662 [+0.4338, +0.4995]
Fast-DetectGPT (0-shot)	−0.0353 [−0.0679, −0.0029]	−0.0040 [−0.0334, +0.0256]

The candidate pool behind Finding 1. Table 15 is the surviving record of the checkpoint/architecture search whose two endpoints Finding 1 reports. Both endpoints are scored on the *identical* 73,941-document COLING-2025 test split, so the +1.72 is a like-for-like difference and not an artefact of differing evaluation sets. We report what survives rather than the search as it was run: of an original pool of 37 ranked candidates (recoverable from the rank field retained in each run's results file), 14 still have per-sample predictions from which test macro-F1 can be recomputed, and only 3 retain the dev score that the leakage-free protocol selects on.

Table 13: Retained single-checkpoint contemporary Wikinews evaluation (n=267). A separate population from Table 11: contemporary news articles rather than scientific abstracts, reported here so the abstract-probe rate is not read as a genre-general figure. Decision rule is strict argmax (AI iff $p_1 > p_0$); one document is exactly tied at 0.5 and resolves to human. **These are retained evaluations supported by surviving predictions, not newly reproduced scores**: the 267 input documents, the loaded checkpoint identity and the seed are not recoverable from the run directories, so neither row is a multi-seed aggregate. The values are corroborated by `CLEAN_RUN/paper/8experiments_results.json`. Ledger: `ledgers/wikinews_probe.json`.

Detector	flagged / n	flagging rate	95% Wilson
Ours, frontier-aware (single retained checkpoint)	0/267	0.0000	[0.0000, 0.0142]
Ours, legacy-trained	89/267	0.3333	[0.2795, 0.3919]

Table 14: Ablations (COLING-2025 test, leakage-free). Feature *stability*, not quantity, drives generalization.

Configuration	macro-F1
Full: Graph-Fusion + SDA, 65 feats	**83.65**
− SDA (fusion only)	82.44
SDA backbone only (no graph/gate)	80.19
all 186 feats instead of 65-stable	82.74

Table 15: Candidate pool for Finding 1, as it survives. **The oracle maximum is over the 14 recoverable candidates, not the full 37, and only 3 candidates retain the dev score the protocol selects on — so "dev-argmax" here means the best of those three, not of the full pool, and the +1.72 is an endpoint difference rather than an identified selection effect, and is not a bound on the full-pool effect in either direction.** Dev macro-F1 is the leakage-free selection variable; where it is absent the run predates the current logging format and the row can support an oracle comparison but not a leakage-free one. Reporting the pool this way is the paper's own thesis applied to itself—a candidate ledger that quietly listed only the runs with complete records would reproduce the selection bias it exists to document.

Candidate	dev macro-F1	test macro-F1
SDA + graph fusion (256-token train)	—	**85.37**
Graph-Fusion + SDA (selected)	**86.14**	**83.65**
domain-adversarial + linguistic gate (a)	—	83.18
domain-adversarial + linguistic gate (b)	—	82.87
Graph-Fusion + SDA, all 186 features	86.00	82.74
Graph-Fusion, no SDA	84.67	82.44
layer-17 backbone, test-selected	—	82.40
linguistic-guided fusion gate (a)	—	81.17
linguistic-guided fusion gate (b)	—	79.33
shared backbone, asymmetric gate	—	79.19
graph multi-objective warm start	—	79.09
layer-17 backbone	—	78.79
full fine-tune	—	78.13
full fine-tune, multi-objective	—	77.33
Dev-argmax (leakage-free)	86.14	83.65
Test-argmax (oracle)	—	85.37
Endpoint difference (not a measured selection effect)		**+1.72**

Which baselines we ran, and which we did not. Table 16 states the disposition of every detector considered. **One detector could not be run at all** (Ghostbuster), and one was run as a pinned release without a successful reproduction (DetectAnyLLM); each is recorded with its specific cause rather than as a generic coverage limit.

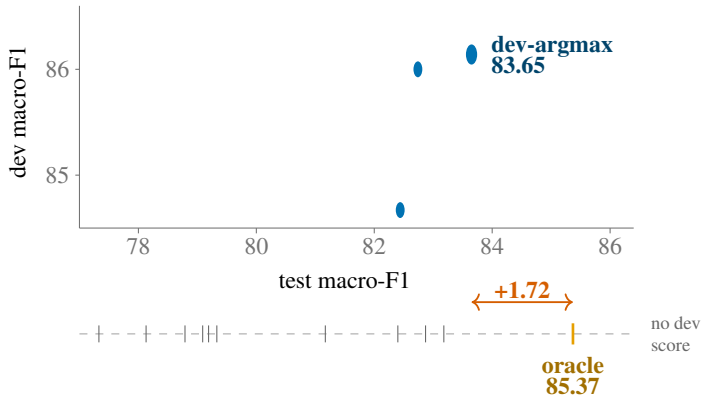

Figure 3: The candidate pool of Table 15, as it survives. **Only** 3 **of the** 14 **recoverable candidates retain a dev score**, so only those three can be placed in the plane; the remaining 11—including the oracle maximum—have a test score but no dev coordinate and are drawn as ticks on the test axis below. The figure therefore shows what the record supports and no more: 83.65 is the **argmax among the three candidates that retain a dev score**, *not* a reconstruction of the dev-argmax of the original 37-candidate pool, which the surviving artifacts cannot support; the test-oracle maximum is 85.37; and the $+1.72$ between them is a difference between retained endpoints — the controlled selection effect of Finding 1 is the MAGE replication, not this contrast. The missing dev scores are a property of the surviving artifacts, not of the selection: the original pool was 37. Table 15 remains authoritative for all values.

Why Ghostbuster could not be run. At the pinned commit, the stored per-token logprobs come from `davinci-002`/`babbage-002` and are tokenized with `cl100k_base` (100,277 types), while the n-gram vectors, the 2047-token truncation, and the trigram vocabulary are all `r50k_base` (50,257). The feature map combines these elementwise, so index i of one vector and index i of the other refer to different spans of text. Measured over 40 documents, the stored length matches the `cl100k` count in $40/40$ cases and the `r50k` count in $11/40$ (the 11 being documents where the tokenizers happen to agree). The mismatch runs in both directions and varies per document, so no offset or truncation repairs it. The released model, normalisation statistics, and feature list were fitted on `r50k`-indexed vectors, so "official implementation, applied unmodified" cannot produce a valid measurement at that commit—for anyone, not only for us. We report this as a property of the released artefact, not as a limitation of our effort, and we did not contact the authors before deciding, which is itself a limitation of the decision.

By provider. Aggregating detection recall by provider, OpenAI is the hardest family for our trained detector (0.65; Anthropic 0.70, xAI 0.74) and Google the easiest (0.93). The zero-shot detectors are weakest on Anthropic (Binoculars 0.36, Fast-DetectGPT 0.30)—reflecting Claude's fluent, low-perplexity style—and strongest on Google (0.78–0.84).

Which detector is hard for which generator. The per-generator columns of Table 4 do not order consistently across detectors, and o4-mini is the clearest case: RoBERTa-large-OpenAI recovers it at 0.977, whereas our legacy detector reaches 0.500 and Fast-DetectGPT and Advacheck 0.455 each — so it is hard for several detectors, not for ours alone, and the *spread across detectors* is the point. Difficulty is therefore detector-specific as well as generator-specific, which is why we report the full matrix rather than a single hardness ranking.

By domain. Difficulty varies sharply by domain (Table 17). News and headlines are easy (0.90–0.93 recall), but long-form creative writing evades detection: our detector recovers only 12% of frontier-generated long stories. Short headlines drive the highest human false-positive rate (0.39). Detection is thus weakest exactly where AI writing assistance is most consequential—creative and long-form composition.

Table 16: Baseline disposition. "Reproduced" means we recovered the method's originally reported number on its own benchmark before using it. Ghostbuster is omitted for the stated reason and discussed in text rather than silently dropped; Advacheck released no competition predictions, but its released checkpoint reproduces its reported score, so we generate aligned predictions from it and compare paired. DetectAnyLLM is the one entry evaluated *without* a successful reproduction; the row states what that does and does not license.

Detector	Type	Official code	Reproduced	Disposition
Binoculars	zero-shot	yes	yes	run on FrontierBench
Fast-DetectGPT	zero-shot	yes	yes	run on FrontierBench
RoBERTa-large-OpenAI	supervised	yes (weights)	yes	run on FrontierBench
RoBERTa-base-OpenAI	supervised	yes (weights)	yes	run on FrontierBench
RADAR	supervised	yes (weights)	yes	run on FrontierBench *and* the paraphrase slice
MAGE (released ckpt)	supervised	yes (weights)	yes	run on FrontierBench (67.66); the artifact scored in Table 19
MAGE-RoBERTa[‡] (ours)	supervised	retrained	n/a	*independently trained* replication used for the controlled selection experiment (Table 3); a **different model** from the released checkpoint above
Advacheck	supervised	yes (ckpt)	yes	**run on FrontierBench**: competition predictions unreleased, but the released checkpoint reproduces 83.07 (83.075), so the comparison is paired
DetectAnyLLM	trained discrepancy	yes (ckpt)	**no**	**run as a pinned release**, *not* as a reproduction: the official eval code on MIRAGE-SIG Rewrite gave AUROC 0.9794 vs. a published 0.9370 for the same backbone; unresolved. Integrity checks passed, so the checkpoint is eligible; correspondence to the published artifact is not established. **Not run on the paraphrase slice**
Ghostbuster	feature-based	yes (pinned)	**no**	**omitted**: the released pipeline combines two incompatible tokenizations (see below)

Table 17: Per-domain performance of our detector on FrontierBench-2026 (AI-detection recall and human false-positive rate).

Domain	AI recall	human FPR
news	0.93	0.06
headlines	0.90	**0.39**
abstracts	0.85	0.22
essays	0.83	0.09
long essays	0.71	0.01
long stories	**0.12**	0.04

The three residual failure modes. Separating the residuals is more useful than a single "still hard" caveat, because they have different causes and different remedies. *(1) Short-form text*: frontier-aware detection is perfect on documents $\geq 1{,}000$ characters and recovers little under 200 (Section 5.1); the transition between the two is uncharacterized, and the benchmark is too bimodal to locate it. *(2) Adversarial input*: our detector is near-chance on humanized/mixed text (the MixSet slice (Zhang et al., 2024), 51.7% accuracy), a DIPPER (Krishna et al., 2023) paraphrase attack drops its recall $0.82 \rightarrow 0.09$, and homoglyph substitution drives it to 0.000—though that last case, unlike the others, now has a measured mechanism and a measured defense (below). *(3) Off-the-shelf detection of*

reasoning and newest-flagship models, which remains open and is *not* the same problem as (1): the two fail in opposite length regimes. Axes (1) and (3) are properties of what detection currently reaches; (2) is an adversarial axis orthogonal to frontier accuracy, so paraphrase-robust *and* frontier-accurate detection remains unsolved (Zha et al., 2025; Rivera Soto et al., 2024).

Adversarial robustness, per attack type (Table 18). On the RAID (abstracts) subset, frontier-aware recall is significantly lower than *legacy* recall in nine of the eleven evaluated conditions after Holm correction: the unattacked condition and eight of the ten attack conditions. **Each test compares the two detectors within a single condition**; none compares attacked against unattacked performance within one detector. The correction family is those eleven comparisons. The legacy detector is essentially unaffected by the attacks themselves—the trade recorded in Table 21. Two attacks dominate. *Paraphrase* drops frontier-aware recall to 0.180 $[0.098, 0.308]$, corroborating the independent DIPPER result in Section 6 ($0.82 \rightarrow 0.09$) on a different corpus and attack implementation. *Homoglyph* substitution is worse still: recall falls to 0.000 $[0.000, 0.071]$ where the legacy detector retains 0.860 $[0.738, 0.930]$ on the same inputs.

Why homoglyphs break the detector. The mechanism is representational, not a learned orthographic cue: replacing 68% of non-ASCII characters across the whole frontier test set moves frontier-aware macro-F1 by -0.02 $[-0.06, +0.00]$ (Section 5), so no smooth cue is present over the range the model normally sees. What homoglyph substitution does instead is destroy tokenization: under our ModernBERT tokenizer the attacked documents expand to $3.87\times$ as many tokens (mean; range 1.94–5.14), with subword fragments averaging 1.73 characters against 4.80 for clean text and *no* unknown-token fallback—the byte-level vocabulary absorbs the substituted characters, so word identity is shattered rather than flagged. The effect is reversible: re-scoring the same subset after applying the UTS #39 confusables mapping restricted to non-ASCII source characters restores frontier-aware recall to 0.860 $[0.738, 0.930]$—43 of 50 predictions flip and none flips the other way (paired McNemar $p{=}1.5 \times 10^{-10}$)—while unattacked recall and the human false-positive rate are bit-identical before and after. We report this as a *defense evaluation* only; no normalization is shipped in the pipeline that produced any other number in this paper. Two things we do *not* claim: truncation is not the cause (attacked documents that exceed the evaluation token budget behave no differently from those that do not), and *why* fragmentation reads as "human" to the frontier-aware model specifically, when the legacy detector on identical input does not, is undetermined.

Table 18: Per-attack AI recall on the RAID (abstracts) subset, $n{=}50$ per attack type and $n{=}500$ for unattacked. Intervals are 95% Wilson; **bold** marks Holm–Bonferroni-significant legacy-vs-frontier differences across the 11 cells (paired McNemar), 9 of which are significant and all of which are frontier-aware degradations. A percentile bootstrap (2000 resamples) agrees with Wilson to within 0.045 at every endpoint *except* six cells where it degenerates to a zero-width interval—the five legacy cells at recall 1.000 and the frontier-aware homoglyph cell at 0.000—which is why Wilson is reported as primary.

Attack	n	Legacy	Frontier-aware	Δ	p_{Holm}
none (unattacked)	500	0.978 [.961,.988]	0.804 [.767,.836]	**−0.174**	<0.0001
homoglyph	50	0.860 [.738,.930]	**0.000 [.000,.071]**	**−0.860**	<0.0001
paraphrase	50	0.900 [.786,.957]	0.180 [.098,.308]	**−0.720**	<0.0001
whitespace	50	0.980 [.895,.997]	0.460 [.330,.596]	**−0.520**	<0.0001
article deletion	50	1.000 [.929,1.00]	0.700 [.562,.809]	**−0.300**	0.0021
alternative spelling	50	1.000 [.929,1.00]	0.720 [.583,.825]	**−0.280**	0.0031
synonym	50	1.000 [.929,1.00]	0.740 [.605,.841]	**−0.260**	0.0044
perplexity misspelling	50	1.000 [.929,1.00]	0.780 [.648,.873]	**−0.220**	0.0103
upper/lower case	50	0.940 [.838,.979]	0.760 [.626,.857]	**−0.180**	0.0230
insert paragraphs	50	0.980 [.895,.997]	0.920 [.812,.969]	−0.060	0.2672
number swap	50	1.000 [.929,1.00]	0.920 [.812,.969]	−0.080	0.2672

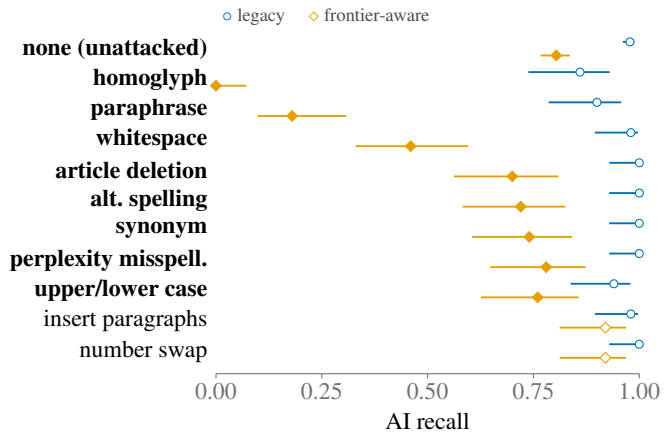

Figure 4: **Per-attack AI recall on the RAID (abstracts) subset, legacy vs. frontier-aware, with 95% Wilson intervals.** A visual reading of Table 18, which remains authoritative: every point and interval endpoint plotted here is a value from that table, and the axis ticks are scale marks. Circles are the legacy (COLING-trained) detector, diamonds the frontier-aware detector; bars are 95% Wilson intervals. *Filled* diamonds and **bold** attack names mark the nine losses that are significant under the same rule as Table 18—paired McNemar, Holm–Bonferroni across the family of 11, α=0.05—and every significant change is a loss for the frontier-aware detector; open diamonds mark the two that are not significant. Attacks are ordered as in the table.

Calibration. Our detector's expected calibration error on the COLING test set is 0.044 (10-bin confidence ECE on uncalibrated softmax outputs; mild overconfidence typical of fine-tuned encoders, versus the severe miscalibration of raw zero-shot scores). False negatives concentrate on "human-like" frontier outputs (Claude Sonnet 4.6, OpenAI o3); false positives concentrate on human text whose era is distant from the detector's reference.

Case study.

A representative Claude Sonnet 4.6 essay receives Binoculars and Fast-DetectGPT scores indistinguishable from human writing—their perplexity cues fail—yet our detector flags it correctly.

This is not a gate flip ($\beta \approx 0.95$): the auxiliary objective regularises the shared representation during training, and the BERT branch itself separates this essay from human writing where perplexity cues do not. Appendix I does not resolve whether the objective's benefit depends on the linguistic content of its targets, so we make no claim about which cues carry the distinction here.

Strong / purpose-built baselines (Table 19). Under the identical leakage-free protocol, RADAR and a MAGE-supervised RoBERTa-large both collapse on frontier text, no better than the perplexity detectors and far below frontier-aware training. At a nominal 1% dev FPR (thresholds set on dev; achieved test FPR in Table 19), *these three* detectors—including the supervised ones, using their own probability outputs—recover at most 12.4% of frontier AI text (RADAR 0.86%, MAGE-supervised 10.99%, RoBERTa-large-OpenAI 12.44%), versus 94.2% for frontier-aware training. The bound is specific to them and does *not* extend to off-the-shelf detection generally: at the same nominal operating point Advacheck recovers 47.19% and the released DetectAnyLLM checkpoint 31.99% (Table 19).

Advacheck degrades on frontier text but less severely, at a fixed operating point, than the older supervised baselines. At a nominal 1% dev FPR its dev-selected threshold is τ=0.999610, giving an *achieved* test FPR of 0.0129 and a test TPR of 0.4719; at a nominal 5% target, τ=0.995080, achieved test FPR 0.0582, test TPR 0.5993 (Table 19). Its weakest cells are the reasoning and newest-flagship models — o3 0.120 (3/25), GPT-5.5 0.273, Claude Opus 4.7 0.312 — the same cohort that is hardest for every other detector we evaluated.

Table 19: Strong/off-the-shelf detectors on FrontierBench-2026 (leakage-free): overall macro-F1 and TPR at a nominal dev FPR. Operating points use the stated convention—*the k-th largest human dev score, scores oriented so that larger values indicate AI*—with thresholds set on the dev split at a nominal target and the test set scored exactly once; "test FPR" is the achieved false-positive rate on test at that threshold, which is not guaranteed to equal the nominal target (Appendix C); achieved test FPRs differ across rows, so these are not comparisons at an identical test FPR. The "Ours (frontier-aware)" row is the originally reported single run, not a seed aggregate. ‡DetectAnyLLM is the author-released Qwen2-0.5B checkpoint (`JiachenFu/Qwen2-0.5B-detectanyllm-detector-en`, rev. `c1bcbefd`; official code `fjc2005/DetectAnyLLM ea82e853`); its correspondence with the checkpoint behind the published MIRAGE results is unresolved — our execution of the official evaluation code gave AUROC 0.9794 against a published same-backbone 0.9370 (Appendix D). Its FrontierBench macro-F1 and both operating points use the dev-selected threshold rule, not a published decision rule, since none is published.

			Nominal dev FPR 1%		Nominal dev FPR 5%	
Detector	COLING	Frontier	test TPR	test FPR	test TPR	test FPR
RADAR	54.8	66.29	0.0086	0.0107	0.0611	0.0460
MAGE (released ckpt)	80.4	67.66	0.1099	0.0103	0.2670	0.0582
Fast-DetectGPT	71.0	70.51	0.0019	0.0132	0.1471	0.0493
Binoculars	80.9	74.27	0.0298	0.0110	0.3970	0.0515
RoBERTa-base-OpenAI	53.7	58.03	0.0037	0.0081	0.0104	0.0515
RoBERTa-large-OpenAI	56.8	77.33	0.1244	0.0114	0.4600	0.0552
Advacheck (2025 winner)	83.08	76.67	0.4719	0.0129	0.5993	0.0582
DetectAnyLLM (rel. ckpt)‡	—	**84.60**	0.3199	0.0110	0.7330	0.0570
Ours (legacy-trained)	**83.7**	78.45	0.3281	0.0114	0.4983	0.0582
Ours (frontier-aware)	—	**92.76**	**0.9423**	0.0088	**0.9922**	0.0493

Table 20: Advacheck (COLING-2025 Task 1 winner, domain-aware multi-task DeBERTa-v3; Gritsai et al., 2025) on the FrontierBench-2026 test split, per generator, with 95% Wilson intervals. Checkpoint `OU-Advacheck/deberta-v3-base-daigenc-mgt1a` at revision `38b6f3f4`, scored under its model-card rule (softmax$[:, 1] \geq 0.92$, max_length 512, fp32); score orientation was verified against dev labels before thresholding. The same checkpoint and code path reproduce its published COLING-2025 macro-F1 (0.8308 against a pre-registered target of 0.8307 ± 0.0010), so the figures below are a degradation of a verified system rather than of a re-implementation. Ledger: `ledgers/advacheck_frontierbench.json`.

Generator	N	Recall	95% Wilson	TP/N
Gemini-2.5-flash-lite	408	0.865	[0.829, 0.895]	353/408
Gemini-2.5-flash	494	0.850	[0.816, 0.879]	420/494
Gemini-2.5-pro	50	0.700	[0.562, 0.809]	35/50
Grok-3	338	0.855	[0.814, 0.889]	289/338
Grok-4-fast (non-reas.)	231	0.524	[0.460, 0.587]	121/231
Claude Haiku 4.5	133	0.571	[0.486, 0.652]	76/133
Gemini-3.1-pro	62	0.419	[0.305, 0.543]	26/62
GPT-5.4-mini	366	0.598	[0.547, 0.647]	219/366
Claude Opus 4.7	77	0.312	[0.219, 0.422]	24/77
Grok-4.20 (non-reas.)	164	0.652	[0.577, 0.721]	107/164
Claude Sonnet 4.6	105	0.610	[0.514, 0.697]	64/105
GPT-5.5	44	0.273	[0.163, 0.418]	12/44
Grok-4.20 (reasoning)	144	0.604	[0.523, 0.680]	87/144
o4-mini	44	0.455	[0.317, 0.599]	20/44
o3 (reasoning)	25	0.120	[0.042, 0.300]	3/25
Overall macro-F1		*76.67*		
Aggregate AI recall	2685	*0.6912*	[0.6735, 0.7084]	1856/2685
Human FPR (pre-2015)	2717	*0.1557*	[0.1425, 0.1698]	423/2717

Cross-benchmark generality (Table 21). Our COLING-legacy-trained detector, evaluated leakage-free on balanced 2K-item subsets of two *external* benchmarks not used in its training, transfers well (MAGE 98.45, RAID 98.60)—its signal is not COLING-specific. Frontier-aware training *trades adversarial robustness for frontier accuracy*: it gives up part of this backward generality (MAGE recall $0.979 \rightarrow 0.564$), and on RAID *its own* recall falls from 0.804 unattacked to 0.618 under adversarial edits (z=6.63), where the legacy detector is unaffected ($0.978 \rightarrow 0.966$; n.s., z=1.15). Both figures in each pair are the same detector on the same split (Table 21). Per-attack cells, intervals and the mechanism behind the largest drop are given below (Table 18).

Expanded small-N generators (Table 22). The four generators with the fewest native test samples (n=25–50) carry wide intervals, so we regenerated 250 decontaminated samples each in August 2026, in two arms: one at the providers' default reasoning effort with output budgets raised so that visible text survives reasoning-token consumption, and one at low effort on the benchmark's native budgets, which doubles as a *decoding-configuration sensitivity* probe. Effective N after excluding calls that returned no visible text is 978 and 926 respectively (per-generator counts and the full status taxonomy in Appendix A).

Comparison is *cell-matched*: the arms have different surviving cell compositions, so every fresh-vs-native contrast runs on matched (generator $\times$ domain $\times$ length-cell) subsets rather than pooled. A fresh cell is compared to its native counterpart only where the decoding configuration we could send matches the configuration the benchmark records for that cell; on that basis 6 cells qualify and all 6 agree with their native cell. We label this outcome *no detectable disagreement* rather than replication: the native cells contain n=2–31 documents, so several native intervals span 40–66 points and a fresh estimate can fall inside one uninformatively. Cells whose decoding differs are reported side by side as decoding sensitivity, with no pass/fail.

The reason so few cells qualify is itself a result about benchmark durability. For three of the four generators the recorded native decoding *cannot* be reproduced against the current APIs: GPT-5.5 now rejects both the temperature and the `top_p` its own records carry, and the two reasoning models accept neither parameter (Appendix B). A benchmark's per-record decoding metadata is therefore perishable in a way its text is not.

Table 21: Cross-benchmark generality (leakage-free, N=2000 each): our detector on **MAGE** (test split; 14 base datasets, AI half weighted toward continuation domains with cnn/imdb/dialogsum/pubmed at n=9 each, human half uniform) and **RAID (abstracts subset)** (RAID's shards are domain-partitioned, so both halves are drawn entirely from `abstracts`: the AI/human comparison is domain-matched, but coverage is narrower than RAID as a whole). RAID AI half = 500 `attack=none` + 500 adversarial at exactly 50 per type across all 10 types; human half = 1000 `attack=none`. Legacy = COLING-trained; frontier = frontier-aware. *clean*/*adv* are AI recall on unattacked vs. adversarial RAID items; z is an unpooled (Wald) two-proportion test.

Detector	Bench	mF1	FPR	AI-rec.	clean	adv
Legacy	MAGE	98.45	0.010	0.979	—	—
Legacy	RAID	98.60	0.000	0.972	0.978	0.966
Frontier	MAGE	76.69	0.009	0.564	—	—
Frontier	RAID	85.24	0.000	0.711	0.804	**0.618**

Per-T-family importance (Table 23). Permuting each linguistic family (names exactly as defined in Section 4) in turn and re-evaluating the frontier-aware detector moves macro-F1 by at most 0.03 on average over 20 permutation seeds (200 evaluations in total), and every family's permutation interval spans zero. **Effects are small and variable across permutation draws; we do not establish a ranking among feature families.** That is a statement about what these draws can resolve, not a formal demonstration that the families are equivalent: the whole distribution spans 0–3 prediction flips on N=5402, so the design cannot separate small real differences from draw-to-draw noise. Appendix I shows the auxiliary objective's total contribution is itself comparable to the pooled seed range, so there is little per-family signal to resolve at this scale.

Table 22: **Fresh-sample cells that qualify for comparison against their native counterpart**, i.e. where the decoding configuration we could send in August 2026 matches what the benchmark records for that cell. All six are Gemini-2.5-pro, the only one of the four generators whose recorded decoding the current APIs still accept. Intervals are 95% Wilson. Arm 1 raised output budgets; Arm 2 used the benchmark's native budgets at low reasoning effort. The two long-stories rows are the *same* cell measured independently by each arm, not a duplicated row; their outcomes coincide. Agreement here means *no detectable disagreement*, not replication: native cells hold n=2–31 documents, so several native intervals are very wide. Cells whose decoding could not be matched are reported in Appendix A as decoding sensitivity, with no pass/fail.

		n		Legacy recall		Frontier-aware recall	
Arm	Cell (domain @ budget)	fresh	native	fresh	native	fresh	native
1	long stories @6000	10	4	0.000 [.00,.28]	0.000 [.00,.49]	1.000 [.72,1.00]	1.000 [.51,1.00]
2	news @1500	168	31	1.000 [.98,1.00]	1.000 [.89,1.00]	1.000 [.98,1.00]	1.000 [.89,1.00]
2	headlines @120	41	10	0.805 [.66,.90]	0.800 [.49,.94]	0.024 [.00,.13]	0.100 [.02,.40]
2	essays @1500	20	3	0.950 [.76,.99]	1.000 [.44,1.00]	1.000 [.84,1.00]	1.000 [.44,1.00]
2	long stories @6000	10	4	0.000 [.00,.28]	0.000 [.00,.49]	1.000 [.72,1.00]	1.000 [.51,1.00]
2	abstracts @1500	8	2	1.000 [.68,1.00]	1.000 [.34,1.00]	1.000 [.68,1.00]	1.000 [.34,1.00]

Table 23: Per-T-family permutation on the frontier-aware detector (FrontierBench test, baseline macro-F1 92.76, N=5402). Δ = baseline − permuted macro-F1, averaged over 20 permutation seeds; family names as in Section 4. Brackets are **permutation intervals**, not confidence intervals — see the notes below.

Family	Linguistic dimension	feats	mean Δ macro-F1 [95% perm. int.]
T1	negation	16	0.018 [0.000, 0.037]
T2	connectives	3	0.006 [0.000, 0.018]
T3	hedging	4	0.008 [0.000, 0.018]
T4	figurative	6	0.020 [0.000, 0.037]
T5	discourse coherence	3	0.019 [0.000, 0.037]
T6	reference	1	0.000 [0.000, 0.000]
T7	pragmatics	16	0.022 [0.000, 0.055]
T8	deixis	1	0.000 [0.000, 0.000]
T9	metadiscourse	5	0.001 [0.000, 0.018]
T10	surface & punctuation	10	0.030 [0.000, 0.037]

Table 24: **Canonical domain audit**, computed from the released split files rather than from prose (`verify/domain_audit.py`; ledger `ledgers/domain_audit.json`). Cells are human / AI counts. **"Train" status is split-scoped**: eight domains are AI-only in training (totalling exactly 54,638 AI documents), but *long essays* and *long stories* acquire human documents in dev and test, which is why they are evaluated and why Table 17 reports human FPRs for them. Taxonomy, the `essay` raw-label exception and the total reconciliation are in the notes below.

Domain	train h/ai	dev h/ai	test h/ai	train status	in eval	Table 17
abstracts	26,867/10,328	1356/1400	1421/1377	paired	yes	yes
code explanations	0/17,435	–/–	–/–	AI-only	–	–
definitions	0/1,568	–/–	–/–	AI-only	–	–
essays	11,271/12,797	51/58	65/58	paired	yes	yes
headlines	13,026/60	636/633	621/624	paired	yes	yes
long essays	0/1,360	114/101	85/98	AI-only	yes	yes
long stories	0/1,005	278/262	259/275	AI-only	yes	yes
news	7,460/14,385	245/257	258/246	paired	yes	yes
one liners	0/1,548	–/–	–/–	AI-only	–	–
qa	0/15,288	–/–	–/–	AI-only	–	–
reviews	0/14,861	–/–	–/–	AI-only	–	–
stories	40,273/14,441	5/6	8/7	paired	yes	–
tweets	0/1,573	–/–	–/–	AI-only	–	–
canonical subtotal	*98,897/106,649*	*2,685/2,717*	*2,717/2,685*			
raw-label adj. `essay`	*2/0*	*–/–*	*–/–*	*human*	–	–
total	**98,899/106,649**	**2,685/2,717**	**2,717/2,685**			

How to read Table 23. The bracketed range is a 95% **permutation interval**: the 2.5th–97.5th percentile of the 20 per-seed Δ values for that family. It describes *variability across permutation draws* and is neither a document-bootstrap interval nor a confidence interval for the mean, so it does not license a significance claim. "feats" is the number of consumed features in the family. T6 and T8 have one feature each, and their exact zeros mean permuting it moved no prediction — which cannot distinguish "carries no signal" from "not consumed by the gate", and is reported without interpretation. One prediction flip on N=5402 moves macro-F1 by 0.0184, so the whole distribution spans 0–3 flips. **Effects are small and variable across permutation draws; we do not establish a ranking among feature families.** Per-seed values: `verify/task_d005_multiseed_result.md`.

Paraphrase robustness (DIPPER). Running DIPPER (Krishna et al., 2023) over a frontier-test slice drops our frontier-aware detector's recall from 0.82 (329/400) to 0.09 (37/400). All values here are our measurements of these detectors, not numbers reported by Krishna et al. (2023); RADAR degrades least ($0.69 \rightarrow 0.59$) but is weak on native frontier text, and the RoBERTa-base-OpenAI detector breaks like ours ($0.68 \rightarrow 0.12$)—paraphrase-robust and frontier-accurate detection remain jointly unsolved.

Notes to Table 24. Seven domains are present in both dev and test; *stories* is evaluated but omitted from Table 17 as too small (8 human / 7 AI in test). **The prompted categories and the dataset labels are the same taxonomy, verified not assumed**: the 1,792 distinct AI-side prompt IDs in training reach exactly these 13 labels and no others, so no separate prompted-category mapping is required. (On the human side some reused source titles appear under both *headlines* and *news*; that is a human-side provenance artefact, not a second taxonomy.) Besides the 13 canonical labels the corpus carries one additional raw label, `essay` (two documents), **whose canonical mapping remains unresolved**: both are verified in `train.jsonl` as training *human* records, and their `prompt_id`s are Wikinews-style 2014 titles, which neither establishes nor excludes an *essays* mapping — so we report them as a separate raw-label adjustment rather than assigning them. **Totals reconcile exactly**: the canonical rows sum to 98,897 train-human; with the adjustment the corpus totals are 104,301 human / 112,051 AI = 216,352, and train = 205,548. This is a *reporting* aggregation only: the training pipeline keys on the raw `domain` string, so those two records carried the label `essay` during training and were not remapped at run time. No released data or prediction was altered.

E FRONTIER-AWARE TRAINING: FULL PER-GENERATOR RECOVERY

Length, not generator difficulty, orders the residual column. This length reading is confirmed on an independent sample. The fresh E2 arms were generated under a decoding configuration that does *not* match the native cells—156 of 180 qualification cells read "decoding configuration differs from native"—so they are a decoding-sensitivity probe and **cannot be substituted for the native per-generator recalls of Table 4**. What they do provide is an independent length-stratified sample at 4–13× the per-cell size of Table 22, and it reproduces the reading exactly: frontier-aware recall is 1.000 at $\geq$ 1,000 characters for *every* one of the four generators, including o3 (n=322), while the same detector recovers 0.140 of short-form text (n=516). The legacy detector is the mirror image—0.822 short-form, 0.533 long-form, and only 0.075 on long-form o3. The two detectors fail in opposite directions, and per-generator aggregates on these arms are ordered by short-form share rather than by generator difficulty. Instrument `harness/e2_length_stratified.py`; values `ledgers/e2_length_stratified.json`.

What the spread tracks is *document length*, not generator difficulty. Frontier-aware recall is 1.000 on every document of $\geq$ 1,000 characters—all fifteen generators, in the native split (n=2,053) and independently in both fresh arms (n=689, n=697)—while documents under 200 characters are the residual (0.387 native; 0.152, 0.123 fresh). Since generators differ in how much short text their sample contains, that composition alone orders the column: pooled recall correlates strongly with short-form share, and recall $\geq$ 200 characters is 1.000 for every one, as quantified below. The qualifying cells show it directly—at *headlines @120* frontier-aware recovers 0.024 fresh vs. 0.100 native where legacy holds 0.805 and 0.800 (Table 22). Read this way the column *understates* the Anthropic models: within long-form text alone Claude-Opus-4.7 moves from fifth to second hardest for off-the-shelf detection and Claude-Sonnet-4.6 from ninth to fifth. We give only these two coarse bands because the benchmark is bimodal—just 15 of 2,685 AI documents fall between 200 and

1,000 characters—so *the transition is not characterized by these data* and we claim nothing about where it lies.

Table 25: Length-stratified recall on the fresh E2 arms (decoding-sensitivity probe; not comparable to native cells). Frontier-aware detection is perfect on long-form text for all four generators and is the weaker detector on short-form; the legacy detector is the reverse.

Generator	Band	n	Legacy-tr.	Frontier-aware
o3	≥1000	322	0.075	**1.000**
gpt-5.5	≥1000	160	0.537	**1.000**
o4-mini	≥1000	492	0.486	**1.000**
*o4-mini has **zero** documents below 200 characters in these arms (verified), so it has no short-form row.*				
gemini-2.5-pro	≥1000	412	0.947	**1.000**
o3	<200	109	0.881	0.147
gpt-5.5	<200	322	0.820	0.143
gemini-2.5-pro	<200	85	0.753	0.118
pooled	≥1000	1386	*0.533*	***1.000***
pooled	<200	516	*0.822*	*0.140*

Table 26 reports per-generator AI-detection recall for the legacy-trained detector and the frontier-aware (in-distribution) detector of Section 5.1, both under the identical leakage-free protocol on the FrontierBench test set (N=5,402). Frontier-aware training raises overall macro-F1 from 78.45 to 92.76 and cuts the human false-positive rate from 0.219 to 0.004. Five of the fifteen per-generator changes are significant after correction, and all five are gains; *no* change in either direction is significant for the remaining ten, so the numerically lower cells for GPT-5.5 ($0.727 \rightarrow 0.477$, p_{Holm}=0.652), Gemini-3.1-pro ($0.790 \rightarrow 0.581$, 0.367) and Claude-Opus-4.7 ($0.675 \rightarrow 0.532$, 1.000) support no directional claim—they are small-N point estimates with wide intervals, and Section 5.1 shows their spread tracks document-length composition rather than generator difficulty. Quantitatively, pooled per-generator recall correlates with each generator's share of short-form (< 200-character) samples at $r = -0.99$ across the four generators we regenerated and $r = -0.78$ across all fifteen; *the tighter figure is specific to those four and should not be read as an all-generator value*. Recall on documents ≥ 200 characters is 1.000 for every generator in the table. **The two length claims nest, and the middle band is nearly empty.** "≥ 200" and "≥ 1,000" are not independent findings: on the FrontierBench AI test side only 15 of 2,685 documents fall in 200–999 characters, so the ≥ 200 statement is carried almost entirely by the ≥ 1,000 stratum. We therefore do not claim the 200–999 band is well characterised — it is too sparsely populated here to support a rate — and both statements describe the frontier-aware detector on this population, not the short-slice comparisons of Appendix D, which use different cohorts.

Aggregate paired tests. Over the same 5,402 documents scored by both detectors, McNemar on discordant pairs gives $p = 1.2 \times 10^{-92}$ for paired error disagreement between the two detectors (χ^2=416.9; b=331, c=1106) and $p = 4.9 \times 10^{-126}$ for the human false-positive rate (χ^2=570.3; b=6, c=590). The paired bootstrap CI on the macro-F1 difference (2000 replicates) is $[+13.06, +15.62]$. The gate's mean β falls from 0.958 (legacy reported run, `nufb_evalA`) to 0.634 (frontier-aware reported run, `frontier_trained_A`), with 28.5% of documents graph/linguistic-dominant ($\beta < 0.5$) versus 0% on legacy data, and auxiliary-feature correlation rises to 0.55. Both are document-weighted means over the same 5,402-document test split; the three-seed ablation counterparts are in Appendix I.

Table 26: Per-generator recall with 95% Wilson intervals: legacy-trained vs. frontier-aware (in-distribution) detector on FrontierBench-2026. **Bold** marks a *significant* change, not merely a larger number: paired McNemar on the same documents, Holm–Bonferroni-corrected across the family of 15 comparisons, α=0.05. Five cells qualify and all five are gains.

Generator	N	Legacy-tr.	Frontier-aware	p_{Holm}
o4-mini	44	0.500 [.36,.64]	**1.000** [.92,1.00]	<0.0001
Grok-4.20 (non-reas.)	164	0.591 [.52,.66]	**0.896** [.84,.93]	<0.0001
Grok-4.20 (reasoning)	144	0.590 [.51,.67]	**0.826** [.76,.88]	0.0022
GPT-5.4-mini	366	0.678 [.63,.72]	**0.888** [.85,.92]	<0.0001
Claude Haiku 4.5	133	0.692 [.61,.76]	**0.872** [.81,.92]	0.016
o3 (reasoning)	25	0.360 [.20,.55]	0.600 [.41,.77]	1.000
Claude Sonnet 4.6	105	0.724 [.63,.80]	0.848 [.77,.90]	0.652
Grok-4-fast (non-reas.)	231	0.714 [.65,.77]	0.758 [.70,.81]	1.000
Grok-3	338	0.885 [.85,.91]	0.891 [.85,.92]	1.000
Gemini-2.5-flash	494	0.931 [.91,.95]	0.931 [.91,.95]	1.000
Gemini-2.5-flash-lite	408	0.944 [.92,.96]	0.922 [.89,.94]	1.000
Gemini-2.5-pro	50	0.880 [.76,.94]	0.820 [.69,.90]	1.000
Claude Opus 4.7	77	0.675 [.56,.77]	0.532 [.42,.64]	1.000
Gemini-3.1-pro	62	0.790 [.67,.87]	0.581 [.46,.69]	0.367
GPT-5.5	44	0.727 [.58,.84]	0.477 [.34,.62]	0.652
Overall macro-F1		*78.45*	***92.76***	—
Human FPR		*0.219*	***0.004***	—

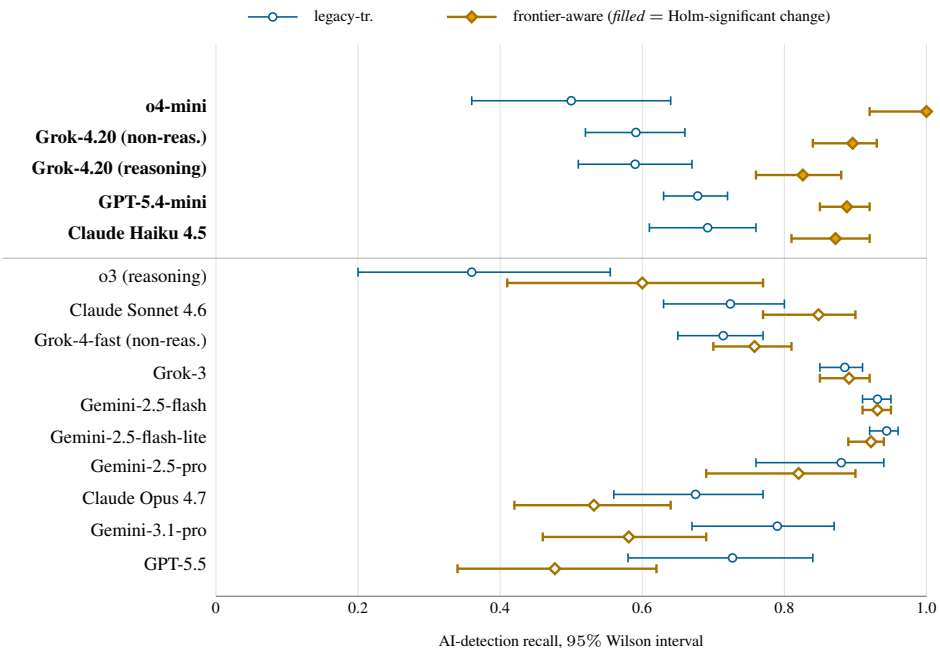

Figure 5: **Per-generator recall, legacy-trained vs. frontier-aware, with 95% Wilson intervals.** A visual reading of Table 26, which remains authoritative: every point and interval endpoint plotted here is a value from that table, and the axis ticks are scale marks. Circles are the legacy-trained detector, diamonds the frontier-aware detector; bars are 95% Wilson intervals. *Filled* diamonds and **bold** generator names mark the five changes that are significant under the same rule as Table 26—paired McNemar, Holm–Bonferroni across the family of 15, α=0.05—and all five are gains; open diamonds mark the ten that are not significant in either direction. Generators are ordered as in the table, significant group first.

F GENERALIZATION: DETAILED PER-PROVIDER, PER-MODEL, AND CAPABILITY-TIER TABLES

Table 27: **Frontier detection generalizes to unseen models, providers, and generations.** Held-out AI-detection recall at three increasing severities, each with the held-out data removed from training and dev. **Unless a row says otherwise, aggregates here are N-weighted means over the held-out documents**, so a 25-document cell does not count as much as a 1,014-document one; provider-macro rows are labelled as such and are *not* N-weighted. Because the four providers partition all 15 generators, the leave-one-model-out and leave-one-provider-out rows share the same legacy-trained and in-distribution references (both pool over the same 2,685 AI documents); only the held-out column differs. **The held-out column is a three-seed mean ± s.d. (seeds 42/43/44), each seed retrained from scratch**—40 additional runs. The legacy-trained and in-distribution columns are the single reference models. Every held-out recall spread is ≤ 0.006 across the three seeds tested. "Unseen" stays close to the in-distribution ceiling and far above the legacy-trained detector; the gap grows gracefully as the holdout becomes more severe. Human FPR ≤ 0.004 throughout. **The two blocks use different averaging rules and must not be read across:** in the capability-tier row the N-weighted contrast is 0.725 vs. 0.644 (+0.081) while the provider-macro contrast on the identical 621-document cohort is 0.644 vs. 0.670 (−0.025), i.e. the sign of the effect depends on the rule. In the leave-one-provider-out row both rules agree in sign (0.826 vs. 0.752 provider-macro; 0.845 vs. 0.788 N-weighted). Every cell in both blocks is recomputed from per-sample predictions on the identical per-provider cohorts. Detailed per-model and per-provider tables follow (Tables 28–30). Ledger: `ledgers/a4_legacy_provider_macro.json`.

Generalization test	Held out	Legacy-tr.	**Unseen**	In-dist
Leave-one-model-out (15 models)	one model (siblings kept)	0.788	**0.856** ± 0.001	0.859
Leave-one-provider-out (4 providers)	entire provider	0.788	**0.845** ± 0.001	0.859
Capability-tier (older → newer)	newer capability tier	0.644	**0.725** ± 0.006	0.754
unweighted provider macro-average — same cohort and predictions, different averaging rule				
Leave-one-provider-out	entire provider	**0.752**	0.826 ± 0.001	—
Capability-tier (older → newer)	newer capability tier	**0.670**	0.644 ± 0.004	—

The capability-tier aggregate is carried by one provider, and its sign depends on the averaging rule. On the identical 621-document cohort, both statistics for both detectors: document-weighted, tier 0.725 ± 0.006 vs. legacy 0.644 (+0.081); provider-macro, tier 0.644 ± 0.004 vs. legacy **0.670** (**−0.025**). **Under equal provider weighting the tier detector is behind, not ahead.** Comparing the tier provider-macro against the legacy *document-weighted* value — both 0.644 — reads a coincidence of two different statistics as equivalence. The divergence arises because the held-out tier is half Grok by document count. Splitting it: on the two Grok-4.20 cells (n=308) the capability-tier detector reaches 0.846 ± 0.004 against the legacy detector's 0.591, a gain of +0.255; on the remaining newer-tier documents (Gemini-3.1-pro, Claude Opus 4.7, Claude Sonnet 4.6, GPT-5.5 and o3, n=313) it reaches 0.606 ± 0.008 against the legacy detector's 0.697, i.e. 0.091 *below* the legacy baseline. The aggregate capability-tier gain is therefore carried by Grok-4.20 and does not describe uniform cross-provider transfer; per-cell differences on the non-Grok flagships are not individually significant at these sample sizes (n=25–105). Ledger: `ledgers/capability_tier_asymmetry.json`.

Tables 28–30 give the per-holdout detail behind the summary in Section 5.1 (Table 27). In every case the held-out generations are removed from both train and dev, and the model is retrained from scratch under the identical leakage-free protocol; "Legacy-tr." is the legacy-trained detector (never a zero-shot method in this paper's sense; see the Terminology note) and "In-dist" is the all-frontier-trained detector of Appendix E. Leave-one-model-out uses 4 epochs; all other runs use 5. The per-holdout columns below are the seed-42 runs; each holdout was additionally retrained at seeds 43 and 44 (40 runs), and the three-seed aggregate spreads are 0.001 (leave-one-model-out), 0.001 (leave-one-provider-out) and 0.006 (capability-tier). Per-holdout, per-seed values ship in `ledgers/generalization_multiseed.json`; the instrument is `harness/generalization_multiseed.py`. The largest single-cell movement across seeds is 0.013 (Claude-Opus-4.7, n=77); nine of the fifteen leave-one-model-out cells are identical across all three seeds. For the capability-tier holdout, the "newer" tier is defined by *capability tier* rather

than calendar release date: each provider's most capable 2026 variant is held out, which yields Gemini-3.1, Grok-4.20, GPT-5.5, o3, and Claude-Sonnet-4.6/Opus-4.7 as the held-out generation (see **Split criterion** below).

Table 28: Leave-one-provider-out. Each provider removed entirely; recall on its unseen generators.

Held-out provider	Legacy-tr.	**Generalize**	In-dist	Overall macro-F1	Human FPR
OpenAI (o3, o4-mini, GPT-5.x)	0.650	**0.808**	0.846	91.7	0.002
Anthropic (Claude $\times 3$)	0.698	**0.759**	0.781	92.4	0.003
Google (Gemini $\times 4$)	0.925	**0.891**	0.900	92.1	0.004
xAI (Grok $\times 4$)	0.737	**0.842**	0.846	92.6	0.004
N-weighted mean	*0.788*	***0.845***	*0.859*	*92.2*	*0.003*

Table 29: Leave-one-model-out: each model removed individually (siblings retained); recall on the unseen model. LOMO $\approx$ in-distribution throughout.

Held-out model	N	Legacy-tr.	**LOMO**
o4-mini	44	0.500	**1.000**
Grok-4.20 (non-reas.)	164	0.591	**0.884**
Grok-4.20 (reasoning)	144	0.590	**0.826**
GPT-5.4-mini	366	0.678	**0.885**
Claude Haiku 4.5	133	0.692	**0.865**
Claude Sonnet 4.6	105	0.724	**0.857**
Grok-4-fast (non-reas.)	231	0.714	**0.758**
Grok-3	338	0.885	0.888
Gemini-2.5-flash	494	0.931	0.931
Gemini-2.5-flash-lite	408	0.944	0.919
Gemini-2.5-pro	50	0.880	0.820
o3 (reasoning)	25	0.360	**0.600**
Gemini-3.1-pro	62	0.790	0.581
Claude Opus 4.7	77	0.675	0.519
GPT-5.5	44	0.727	0.455
Weighted mean		*0.788*	***0.856***

Split criterion. The "newer" generation holds out each provider's most capable flagship 2026 release—Grok-4.20 (both modes), Claude Opus 4.7 and Sonnet 4.6, GPT-5.5, Gemini-3.1-pro, and o3—while the "older" generation is the preceding capability tier (Grok-3 / Grok-4-fast, Claude Haiku 4.5, GPT-5.4-mini, the Gemini-2.5 family, and o4-mini). The split is by *capability tier*, not calendar release date: o3 (a flagship reasoning model) is held out while o4-mini (a smaller/faster sibling released the same day) is retained, so the test measures forward transfer to a stronger generation rather than to a later timestamp.

Table 30: Capability-tier holdout: trained only on the older-tier 2026 models; recall on the unseen newer/flagship generation.

Unseen newer model	N	Legacy-tr.	**Capability-tier**
Grok-4.20 (non-reas.)	164	0.591	**0.872**
Grok-4.20 (reasoning)	144	0.590	**0.812**
Claude Sonnet 4.6	105	0.724	**0.790**
o3 (reasoning)	25	0.360	**0.600**
Gemini-3.1-pro	62	0.790	0.581
Claude Opus 4.7	77	0.675	0.481
GPT-5.5	44	0.727	0.432
Weighted mean		*0.644*	***0.725***

G PROMPT AND PARENT-DOCUMENT OVERLAP BETWEEN SPLITS

FrontierBench's splits are decontaminated at the level of *text*: the audit reported in Section 3 (exact hashing, 100-word prefix hashing, 13-gram containment) finds 0 exact duplicates and 0 prefix collisions, and $>50\%$ 13-gram containment for 0.56% of AI and 11.2% of human documents. **Those two percentages are shares of the documents that are *eligible*** — long enough to yield at least one 13-gram — which is the denominator the audit script uses: 12 of 2,126 eligible AI documents and 263 of 2,357 eligible human documents. Of the 5,402 test documents, 919 (559 AI, 360 human) are too short to form a 13-gram and are excluded from those rates; as shares of all test documents the same hits are 0.45% of AI and 9.7% of human. They are **not** grouped by prompt or by parent document, and Table 31 reports that exposure directly rather than leaving a reader to discover it.

Table 31: Prompt and parent-document overlap between splits, as distinct units shared (percentage of the smaller split). Reproducible from the release with `verify/split_integrity.py`.

Audit	train–dev	train–test	dev–test
Exact `prompt_id` (AI)	517 (88.5%)	497 (88.0%)	491 (86.9%)
Normalised prompt text (AI)	238 (93.7%)	236 (93.7%)	232 (92.1%)
Parent document, by title (human)	294 (11.0%)	289 (10.7%)	11 (0.4%)
`template_id`	*no template field exists in the benchmark*		

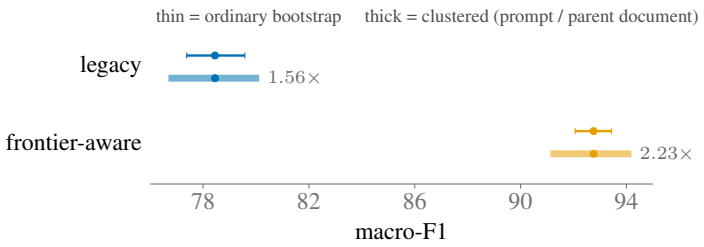

Figure 6: Interval width under ordinary versus clustered resampling. A visual reading of the clustered-bootstrap values reported in this section, which remain authoritative: every endpoint plotted here is one of those values and the axis ticks are scale marks. Thin bars are the ordinary bootstrap over documents; thick bars resample *clusters* — the prompt for AI documents, the parent document for human ones — which is the correct unit given the overlap in Table 31. Point estimates (dots) are identical under both: clustering changes the uncertainty, not the estimate. The two clustered intervals remain disjoint, which is the substantive result — the overlap widens uncertainty without threatening the conclusion.

At document level, 84.4% of test AI documents answer a prompt that also appears in training. On the human side the retained statistic is *title*-level, not document-level: 289 distinct parent-document titles are shared between train and test, which is 10.7% of the smaller split's distinct titles (Table 31). The 1,860-entry prompt set contains only 604 distinct texts, so prompt *text* rather than `prompt_id` is the content unit: 657 of 1,860 ids appear in *at least two* of the three partitions. That 657 is the size of the union of those ids, *not* the sum of the pairwise overlaps (517, 497, 491), which double-counts ids present in all three; by text the overlap is higher still. Concretely, prompt `abstract_0076` has 60 training documents across 11 generators and a test document from Grok-3 answering the identical prompt; on the human side a single Wikinews story contributes a headline record to test and a body record to train.

This is a design property, not a leak of text. A prompt pool shared across providers isolates provider and model signal from content novelty: the shipped split is a *shared-prompt* transfer setting, and the held-out-model, held-out-provider and capability-tier results of Appendix F all inherit that design. *Content-controlled* is reserved in this paper for the grouped split, where no prompt text and no parent document crosses a boundary. We price the consequence rather than assert it is harmless. Resampling *clusters* — the prompt for AI documents, the parent document for human ones — instead of individual documents widens the 95% interval by $2.23\times$ for the frontier-aware detector

(92.06–93.44 becomes 91.13–94.19) and by $1.56\times$ for the legacy detector (77.38–79.58 becomes 76.69–80.12). No point estimate moves, and the two clustered intervals remain disjoint. *The overlap widens uncertainty; it does not threaten the conclusion.* Both the audit and the clustered bootstrap ship as `verify/split_integrity.py` and `verify/clustered_bootstrap.py`.

The same overlap seen at token level. The train/test audit of Section 3 reports 0 exact duplicates and 0 100-word prefix collisions, yet 11.2% of *eligible* human test documents — 263 of the 2,357 long enough to yield a 13-gram — have $> 50\%$ of their 13-grams present in train. The two zero results are not in tension with the third: exact and prefix hashing are *boundary-sensitive*, so a passage cut at a different offset from the same source produces a different hash under both while its interior n-grams match almost entirely.

Three counts that must not be conflated. The audit records 275 hits in total across the whole test split; those are 12 AI and **263** human documents. Of the 263 human hits, **259** are long-form public-domain chapters (*long stories*), 3 are headlines and 1 is a short story — so the long-form share is $259/263 = 98.5\%$ of the human hits, **not** $259/275$. The remaining 12 hits are AI documents and belong to no part of that statement. The concentration is in fact total: all 259 long-story human test documents are hits, at a median 97.6% of their n-grams present in train. This is the parent-document overlap of Table 31 viewed at token level rather than by title. The two measurements point the same way but are *not* the same quantity and are not directly comparable: Table 31 counts 289 shared distinct parent-document *titles* (10.7% of the smaller split's distinct titles), whereas 11.2% counts *documents* whose n-grams are mostly present in train. Per-subset eligible counts, hits and the domain breakdown are recomputed in `verify/overlap_denominator_audit.py` (`ledgers/overlap_denominator_audit.json`); the same $275/12/263$ split and the 2,126/2,357 eligible denominators appear independently in the K-sweep of `verify/decon_variants.json`.

The human figure is structural, the AI figure is not. Across all 27 configurations we swept—$K \in \{10, 13, 20, 25\}$ word-level and $K{=}50$ character-level, thresholds 0.5 and 0.8, both containment directions—the human rate stays within 10.2–11.9%, so it is a property of the corpus rather than of the measurement. The AI rate is small and K-sensitive: 0.56% at the reported $K{=}13$, falling to $\approx 0.05\%$ by $K{=}20$ as borderline documents drop out. The direction reverses consistently too—1.23% of *train* documents are $> 50\%$ contained in test. Sweep and script: `verify/decon_variants.json`, `verify/decontamination_audit.py`.

Grouped-split challenge evaluation. Table 5 reports content-and-provider transfer: the frontier-aware model and all four leave-one-provider holdouts retrained on a re-split in which **no prompt text and no parent document crosses a boundary**. Zero overlap on every computable audit of Table 31 was verified *before* training by `verify/build_grouped_split.py`, and each holdout carve asserted its exclusion count against a measured expectation before its run began. Per-run values are in `verify/grouped_campaign.json`.

The campaign was repeated at seeds 43 and 44 for all five configurations, with the partition, the training configuration and the dev checkpoint-selection rule held fixed and the seed-42 holdout carves reused verbatim, so the seed was the only deliberately varied factor; the exclusion counts were re-asserted before launch. Across seeds 42, 43 and 44 ($\pm$ is sample s.d. across those three seeds), in-distribution macro-F1 is 89.78 ± 0.16, AI recall is 0.7984 ± 0.0030, the N-weighted held-out-provider recall is 0.7642 ± 0.0002 and the transfer gap is 0.0342 ± 0.0030 (per provider: OpenAI 0.7281 ± 0.0065, Anthropic 0.7220 ± 0.0031, Google 0.7570 ± 0.0035, xAI 0.8157 ± 0.0017). Neither the absolute level nor the doubling of the gap is therefore a single-seed artefact. Per-seed values are in `ledgers/grouped_multiseed.json`, produced by `harness/grouped_multiseed.py`.

Grouped capability-tier evaluation was not run. The content-controlled grouped split was evaluated for leave-one-provider-out transfer, but not for capability-tier transfer. The capability-tier results therefore use the shipped split and inherit its prompt overlap (84.4% of AI test documents answer a prompt also seen in training), which includes the provider-macro reversal reported in Appendix F. Evaluating this setting under content control would require a separate training campaign, which we did not run.

Roughly three points of in-distribution macro-F1 and eight points of held-out-provider recall separate the shipped and grouped splits, a joint sensitivity to content overlap and partitioning. We

state that plainly: the shipped split's absolute levels are optimistic, and the grouped column is the content-controlled estimate.

What survives is the claim this section makes. Under the N-weighted convention used throughout, the gap between in-distribution and held-out performance is 0.015 on the shipped split and 0.033 grouped: it roughly doubles under content control rather than staying constant, and we report it that way. Both remain small—an N-weighted mean recall gap of 1.5 points shipped and 3.4 ± 0.3 points grouped, an average over providers and seeds rather than a per-provider bound—and the human false-positive rate is lower under grouping. Frontier signatures transfer across providers; the shipped split scores higher on *every level* without changing *whether the transfer holds*. The evaluation was specified before the grouped runs began and is reported as it landed.

One bound on the reading: the grouped test split is a different partition of the same pool ($5{,}407$ documents against $5{,}402$), so this is a comparison across partitions rather than a paired before-and-after on identical documents. Under grouping, test documents answer prompts absent from training by construction, which is the mechanism under study.

H Can the era confound be trained away? A dose-response test, and a provenance audit that withdraws its conclusion

Intended design. Section 5 reports that the frontier-aware detector flags 0.0002 ± 0.0004 of matched pre-2015 abstracts but 0.119 ± 0.038 of the matched post-2023 probe. The obvious remedy is to put contemporary human text into training, and this experiment was designed to test that: add Wikinews/Wikipedia 2024+ text at increasing doses, retrain from scratch at seeds $42/43/44$, and measure the flagging rate on the matched post-2023 abstract probe. Doses are seeded subsets rather than nested prefixes; a fourth, exploratory 903-document dose was run once. Added text was verified disjoint from the probe on a 300-character prefix hash, so no training document can appear in the evaluation.

A post hoc provenance audit changes what this experiment can support. The added corpus is labelled `Wikinews-2024+` and `Wikipedia-2024+`, but only part of it carries a source timestamp. Wikinews articles open with a dateline, which lets the rest be dated independently: parsing only datelines *anchored at the start of the text* (so that a historical date inside an article cannot be mistaken for its publication date) resolves $2{,}190$ of the $2{,}664$ Wikinews documents, an 82.2% parse rate, and metadata and dateline never disagree on any record where both exist. Of the $2{,}211$ Wikinews records with no source timestamp, $2{,}158$ carry an anchored dateline earlier than 2024, 32 carry one from 2024 or later, and 21 remain unresolved ($2{,}158 + 32 + 21 = 2{,}211 = 2{,}664 - 453$). A null timestamp is never by itself treated as evidence of age. The 453 originally collected Wikinews documents carry real API dates and are genuinely 2024+, as is the entire $5{,}405$-document Wikipedia component.

Table 32: Realised temporal composition of each dose. "Verified 2024+" means a source timestamp or an anchored dateline of 2024 or later; "pre-2024" means an anchored dateline earlier than 2024; "unresolved" means neither. The 903 dose draws from the originally collected 903-document set and is clean; the $2{,}700$ and $8{,}069$ doses draw from the combined $8{,}069$-document pool and inherit its mixed-era Wikinews component. Ledger: `ledgers/dose_provenance_audit.json`.

Dose	Source	Total N	Verified 2024+	Pre-2024	Unresolved	% pre-2024
903	Wikinews	453	453	0	0	0.00%
	Wikipedia	450	450	0	0	0.00%
	whole dose	**903**	**903**	**0**	**0**	**0.00%**
2,700	Wikinews	892	170	713	9	79.93%
	Wikipedia	1,808	1,808	0	0	0.00%
	whole dose	**2,700**	**1,978**	**713**	**9**	**26.41%**
8,069	Wikinews	2,664	485	2,158	21	81.01%
	Wikipedia	5,405	5,405	0	0	0.00%
	whole dose	**8,069**	**5,890**	**2,158**	**21**	**26.74%**

Four denominators are in play for the expanded Wikinews component and they must not be interchanged: $2{,}158/2{,}190 = 98.5\%$ of expansion records with an anchored, successfully classified dateline; $2{,}158/2{,}211 = 97.6\%$ of the full expansion including the 21 unresolved records; $2{,}158/2{,}664 = 81.0\%$ of the complete Wikinews component; and $2{,}158/8{,}069 = 26.7\%$ of the entire largest dose. The last is the figure that characterises the intervention as administered.

Table 33: Era-confound dose-response, as realised. Flagging rate on the matched post-2023 human probe (n=1,105), three seeds per dose except the exploratory 903 row. The rightmost column is the realised pre-2024 share from Table 32: the two larger doses did not deliver the contemporary text they nominally specify.

Added contemporary human text	share	per-seed rate	mean	s.d.	realised pre-2024
none (frontier-aware)	0%	0.151 / 0.077 / 0.130	0.119	0.038	—
903 documents	0.45%	0.158 *(1 seed)*	0.158	—	0.00%
2,700 documents	1.33%	0.103 / 0.137 / 0.145	0.128	0.022	26.41%
8,069 documents	3.86%	0.122 / 0.107 / 0.094	0.108	0.014	26.74%

Statistical test as specified. The one-way ANOVA compares the three multi-run conditions — doses 0, 2,700 and 8,069, one observation per training run, three runs each — and gives F=0.443668, p=0.661150. The 903 condition is excluded because it has a single run and contributes no within-group variance. Including it as a fourth group gives p=0.469; that analysis mixes a one-run group with three-run groups and is not a comparable confirmatory test, so it is reported only to show the exclusion is not concealing a different answer. Within-dose s.d. (0.025) exceeds between-dose s.d. (0.010). Frontier detection is unharmed at every dose (macro-F1 92.4–92.8 against the 92.76 baseline).

What this experiment does and does not establish. A post hoc provenance audit found that 26.41% and 26.74% of the two larger nominal contemporary-human doses are pre-2024 Wikinews articles; within the successfully dated expanded Wikinews component, 98.5% are pre-2024. The reported p=0.661150 therefore describes a mixed-era, cross-domain intervention and does not establish that provenance-verified contemporary-human augmentation is ineffective. Four limitations compound: (i) roughly one document in four in the two larger complete doses is demonstrably pre-2024, much of it from 2005–2010, older than part of the pre-2015 reference itself; (ii) nearly all successfully dated records in the expanded Wikinews component are pre-2024; (iii) the outcome probe is scientific abstracts while the augmentation is news and encyclopaedic prose, so the test is cross-domain as well as mixed-era; (iv) the within-domain check that would have separated these — rescoring the dose checkpoints on the retained 267-article contemporary Wikinews probe — cannot be reproduced, because that probe's input documents, and the checkpoint identity and seed of the runs that scored it, were not preserved (Table 13). The 453 genuinely dated Wikinews records that remain sit inside the augmentation pool itself, so they cannot substitute as a held-out probe without training contamination.

We report the experiment rather than deleting it, because the audit is the finding. A valid successor requires verified contemporary dates at collection time, a domain-matched evaluation probe, evaluation samples disjoint from the augmentation pool, and preserved sample identifiers and inputs. We make no claim here about whether contemporary-human augmentation reduces the era effect.

Why the seed campaign mattered. Our own first pass ran one seed per dose and found what looked like a 41% reduction ($0.175 \rightarrow 0.103$). Three seeds dissolved it: 0.103 was the most favourable of three draws at that dose, and 0.175 came from a single earlier checkpoint that sits above every run in the three-seed dose-zero campaign ($0.151/0.077/0.130$). A dose-zero model at another seed reaches 0.077—better than any dosed model we trained.

The variation is not only across seeds. The 0.175 checkpoint and the campaign's seed-42 run are the *same configuration at the same seed*, their saved configs differing in no training field, yet they score 0.175 and 0.151—a 0.024 gap from run-to-run nondeterminism alone. Reporting a false-positive rate from one run is therefore unreliable even when the seed is fixed and stated. This is Finding 1b arriving in an experiment we designed and in the direction of our own hypothesis, and it is the reason the seed spread belongs in the main text.

I COMPONENT ABLATIONS UNDER MULTI-SEED EVALUATION

Scope. Unless explicitly labelled as legacy, results in this appendix use the *frontier-aware* configuration on the **FrontierBench-2026 test split** (N=5,402). Table 14 is a different experiment—the legacy configuration on the COLING-2025 test set—and nothing here revises it. Five variants were run at three seeds each; the variant set and the seed count were fixed before any result was seen (raw values: `verify/ablation_campaign.json`).

Table 34: Component ablations, frontier-aware configuration, FrontierBench test. Mean ± s.d. over three seeds (42/43/44); Δ is the paired difference against *full* at the same seed. With only three seeds we emphasize descriptive paired differences and sign consistency rather than significance claims.

Variant	macro-F1	s.d.	Δ vs full	sign-consistent
− domain-adversarial	92.821	0.096	+0.006	no
Full	92.815	0.082	—	—
Backbone only (no graph/gate)	92.729	0.122	−0.086	yes
− auxiliary objective	92.590	0.124	−0.226	yes
Shuffled auxiliary targets	92.565	0.171	−0.250	yes
Seed main effect, pooled over variants				
seed 42 / 43 / 44	92.594	92.691	92.827	(spread 0.233, monotone)

The pooled seed range exceeds the graph-branch effect. Pooled over variants, the three seeds span a **range** of 0.233 macro-F1 (max − min over three seeds — *not* a standard deviation, and not a significance threshold). Descriptively, that range is larger than the graph/linguistic branch's whole contribution (0.086) and comparable to the auxiliary objective's (0.226) and to the shuffled-target difference (0.250, which slightly exceeds it). A single-seed ablation on this configuration would therefore report the seed at least as much as the component under test, which is why the campaign was specified at three seeds rather than one.

The auxiliary objective helps, but three seeds do not isolate why. Removing the objective costs 0.226; *permuting its targets across the batch* (a real feature distribution, uninformative about the document) costs 0.250. Both are sign-consistent but comparable to the 0.233 pooled seed range, and the shuffled-target arm sits at the no-objective arm (92.565 against 92.590). We therefore claim no validated mechanism at this seed scale.

Activation is not contribution. Under frontier-aware training the gate demonstrably activates. **Across the three ablation seeds** mean β is 0.663 ± 0.024 (per seed 0.635, 0.679, 0.674) and $18.4\% \pm 9.1$ percentage points of documents are graph/linguistic-dominant ($\beta < 0.5$; per seed 28.7%, 11.4%, 15.1%). **The 28.5% quoted in Section 5.1 is the single reported run, not this three-seed mean**: the $\beta < 0.5$ *share* is far more seed-sensitive than the mean itself. Removing the branch entirely nevertheless costs 0.086 macro-F1, and removing the auxiliary objective pushes the gate back toward the backbone ($0.663 \pm 0.024 \rightarrow 0.755 \pm 0.015$, both three-seed means from this campaign), coherent with the objective being what brings the branch into play. All β statistics here are document-weighted means over the same 5,402-document test split, from the retained per-document `beta` column after the $[0.05, 0.95]$ clamp and the short-document override (`verify/gate_beta_audit.py`, `ledgers/gate_beta_audit.json`). The domain-adversarial term contributes nothing measurable (+0.006, sign inconsistent).

The legacy regime, and why the two regimes differ. Repeating the shuffled-target comparison on the legacy configuration (COLING test, N=73,941, three seeds per arm), deleting the auxiliary objective costs 0.54 macro-F1 (82.62 against full 83.16, sign-consistent), so there the objective does help — but permuting its targets costs nothing detectable (83.16 ± 0.24 full against 83.25 ± 0.32 shuffled, paired difference −0.09 with *mixed sign*). So in the legacy regime shuffled targets recover the objective's full benefit, consistent with a regularisation reading, whereas in the frontier-aware regime they do not (92.565 against 92.590), consistent with a content-specific effect at a magnitude comparable to the seed range. We report both as they landed and draw no mechanistic conclusion from three seeds. Finally, Table 14's single-run 83.65 reproduces exactly but sits about two standard deviations above the three-seed mean (83.16 ± 0.24) — the seed effect documented here.