Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Does Graph Structure Improve AI-Generated Text Detection? A Controlled Audit with Tuned Baselines

ABadila Alaktif

abadila.alaktif.doc23@ensem.ac.ma

University of Hassan II Casablanca

Meriyem Chergui

University of Hassan II Casablanca

Gmira Faiq

University of Hassan II Casablanca

Abdelkarim Ammoumou

University of Hassan II Casablanca

Research Article

Keywords: AI-generated text detection, graph neural networks, evaluation methodology, baselines, large language models, ablation studies

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11039287/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Does Graph Structure Improve AI-Generated Text Detection? A Controlled Audit with Tuned Baselines

Abadila Alaktif[1*], Meriyem Chergui[1], Gmira Faiq[2], Abdelkarim Ammoumou[3]

[1*]C3S Laboratory, École Nationale Supérieure d'Électricité et de Mécanique (ENSEM), Hassan II University of Casablanca, Casablanca, Morocco.

[2]University Sidi Mohamed Ben Abdellah, Fès, Morocco.

[3]C3S Laboratory, École Supérieure de Technologie (EST), Hassan II University of Casablanca, Casablanca, Morocco.

*Corresponding author(s). E-mail(s): abadila.alaktif.doc23@ensem.ac.ma;
Contributing authors: meriyem.chergui@ensem.ac.ma; faiq.gmira@etu.univh2c.ma; abdelkarim.ammoumou@univh2c.ma;

Abstract

Graph neural network (GNN) detectors of AI-generated text report state-of-the-art accuracy, and the architectural claim that graph topology captures signals invisible to sequential classifiers is now common in the literature. We report the complete lifecycle of one such detector, from construction to falsification. AEGIS, a graph attention network over k-nearest-neighbour text graphs with multi-signal edge weights, reaches 98.50% accuracy on scientific abstracts with only 73K trainable parameters, outperforming five established baselines. We then subject our own system to the evaluation its headline number deserves: generation-time length control, format normalization, leakage-free splits, a tuned frozen-embedding probe as the calibrated null, and a $\mathbf{10 \times 10}$ cross-generator transfer study over three seeds. Under this protocol the graph's apparent contribution collapses from $\mathbf{+0.240}$ AUROC over a surface baseline to $\mathbf{-0.0004}$ AUROC against the tuned probe, inside the measured replication noise floor, and the two detectors' transfer geographies are near-identical ($\boldsymbol{\rho = 0.986}$). A single untuned regularization default manufactured a $\mathbf{+0.027}$ phantom advantage. The structural differences the graph encodes are real, but they are a re-encoding of

information already present in the embeddings. We distil the audit into a practical checklist for evaluating architectural claims in text detection, and argue that accuracy numbers and mechanism claims must be validated separately.

Keywords: AI-generated text detection, graph neural networks, evaluation methodology, baselines, large language models, ablation studies

1 Introduction

Detectors of machine-generated text are multiplying faster than the methodology needed to evaluate them. Each new architecture arrives with a strong accuracy number and, usually, a mechanism story: the architecture is said to capture a kind of signal that simpler models cannot. Graph neural detectors are a prominent example. A growing line of work represents documents as graphs and reports that relational structure separates human from machine text [1–3], and a recent framework argues for the explanatory value of graph motifs in this task [4]. The accuracy numbers in this literature are often genuine. The mechanism claims are almost never tested against a properly calibrated alternative.

This paper reports what happened when we tested ours.

We first present AEGIS, a graph attention network [5] over k-nearest-neighbour text graphs with multi-signal edge weights, developed through a transparent three-stage process from a GloVe/GCN proof of concept (74.17% accuracy) to a multi-modal architecture reaching 98.50% accuracy on scientific abstracts with 73K trainable parameters—1,708× fewer than a fine-tuned RoBERTa baseline [6], and with perfect recall on human-written text. On the conventional evaluation diet (a single generator, no length control, test-set monitoring, untuned baselines), AEGIS beats GLTR [7], DetectGPT [8], Fast-DetectGPT [9], ZeroGPT [10], and fine-tuned RoBERTa by wide margins. When we originally wrote that result up, we interpreted it the way the literature primed us to: the graph was doing the work.

We then did what the field rarely does: we rebuilt the evaluation until the comparison was fair, and asked the question again. The fair protocol has five components: (i) length control at generation time, because every model we tested shortens text and length alone is a strong classifier; (ii) symmetric format normalization, because LaTeX-to-Unicode conversion is a formatting tell; (iii) source-identifier splits and validation-based model selection, closing leakage and test-set monitoring; (iv) a *tuned* frozen-embedding probe as the null model, because the default regularization setting of the standard baseline understates what embeddings alone achieve; and (v) a 10-generator corpus with a paired transfer design and a measured test–retest noise floor, so that margins can be judged against replication variance rather than against zero.

The result is unambiguous. Against the tuned probe, the graph's contribution is -0.0004 AUROC averaged over the full $10 \times 10 \times 3$-seed transfer study—inside the 0.001–0.003 noise floor—and the transfer geography of the GNN is statistically indistinguishable from that of the probe ($\rho = 0.986$). The 98.50% accuracy survives; the

mechanism story does not. AEGIS is accurate for the wrong reasons: it is an expensive reader of the same embedding signal a logistic regression reads for free.

This paper makes four contributions:

1. **A transparent system report.** We document AEGIS's architecture and its three-stage development honestly, including what did not work, providing a reproducible reference point for graph-based detection (Section 3).
2. **A confound-controlled evaluation protocol.** We characterize four confounds present in standard detection corpora (length, format, near-copies, topic leakage), introduce generation-time length forcing as a remedy, and quantify a replication noise floor for detection experiments (Section 4).
3. **A controlled isolation of the graph contribution.** With single-variable ablations against a tuned null, we show the architectural margin collapses to zero, identify the exact mechanism (an untuned baseline default) that manufactured the apparent advantage, and replicate the equivalence across ten generators and three seeds (Section 5).
4. **An audit checklist.** We distil the lifecycle into concrete practices for separating accuracy claims from mechanism claims in detector research (Section 6).

We emphasize what this paper does not claim. It does not claim that graph representations of text are useless, nor that the accuracy numbers of published graph detectors are wrong. It claims something narrower and, we believe, more consequential: on a fair evaluation, the *incremental* contribution of graph structure over a calibrated embedding baseline was zero for our detector, and nothing in the published literature currently rules out the same outcome for others, because the calibrated comparison is almost never run.

2 Related Work

Graph-based detection of machine-generated text.

Several systems precede or parallel ours in applying graph representations to this task. IDEATE [1] applies hierarchical graph convolution over internal and external factual structures, using entity graphs linked to a knowledge base. Gupta and Li [2] build similarity-guided graphs over masked reconstructions and classify with a GCN, reporting robustness to paraphrase. Valdez-Valenzuela et al. [3] report ablation-validated gains for text-graph neural networks in a shared-task setting. LM^{2}otifs [4] pursues explainable detection through graph motifs. None of these works compares against a regularization-tuned frozen-embedding probe; ablations, where present, alter several components at once. Our results suggest this is not a cosmetic omission: it is exactly the comparison that eliminated our own architectural margin. Outside detection, Galke and Scherp [11] found that a wide MLP over bag-of-words matches text-graph networks across general text classification benchmarks, a precedent for structure-free nulls matching structural models; Yao et al. [12] is the canonical text-graph architecture this line builds on.

Supervised and zero-shot detection.

Supervised detectors span fine-tuned transformers [6], classifiers over frozen embeddings [13], and feature-based approaches [7]. Zero-shot detectors exploit likelihood statistics: DetectGPT's perturbation curvature [8], its faster successor [9], and Binoculars' cross-perplexity ratio [14]. Theoretical work bounds detectability in terms of distributional divergence [15, 16]; watermarking sidesteps post-hoc detection altogether [17].

Evaluation practice.

The RAID benchmark [18] documents how sensitive detector rankings are to decoding and adversarial settings, and Liang et al. [19] show how uncontrolled domain factors can masquerade as detector bias. Our protocol extends this line to generation-time confound control: rather than filtering unbalanced data after the fact, we constrain generation so that the corpus is balanced by construction. To our knowledge, the test–retest replication ruler we report is the first quantified noise floor for detection experiments; without one, small architectural margins cannot be distinguished from seed variance.

Positioning.

We claim no architectural novelty. Every component of AEGIS—contextual embeddings, k-NN graph construction, graph attention, hierarchical pooling, multi-task heads—is established. The contribution of this paper is the audited lifecycle: a state-of-the-art-looking system, the fair re-evaluation its headline number deserves, and the methodological lessons of the gap between the two.

3 The AEGIS Detector

3.1 Architecture

AEGIS classifies a document in four stages (Figure 1).

Node construction. Each token is a node carrying a 790-dimensional feature vector: a 768-dimensional contextual embedding from frozen RoBERTa-base [6], a 19-dimensional part-of-speech one-hot encoding, a positional scalar, and two linguistic indicator flags.

Graph construction. Edges connect each node to its $k = 8$ nearest neighbours in feature space, weighted $w_{ij} = 1/(1+d_{ij})$ with d_{ij} the Euclidean distance; additional edges come from spaCy dependency parses, so the graph carries both semantic-similarity and syntactic relations.

Message passing. Three stacked graph-attention layers [5] with eight heads each, residual connections, layer normalization, and ELU activations propagate information over the graph.

Readout. Hierarchical pooling aggregates node states into a document representation, and a classification head predicts human versus machine; auxiliary heads for text complexity and readability act as regularizers during training.

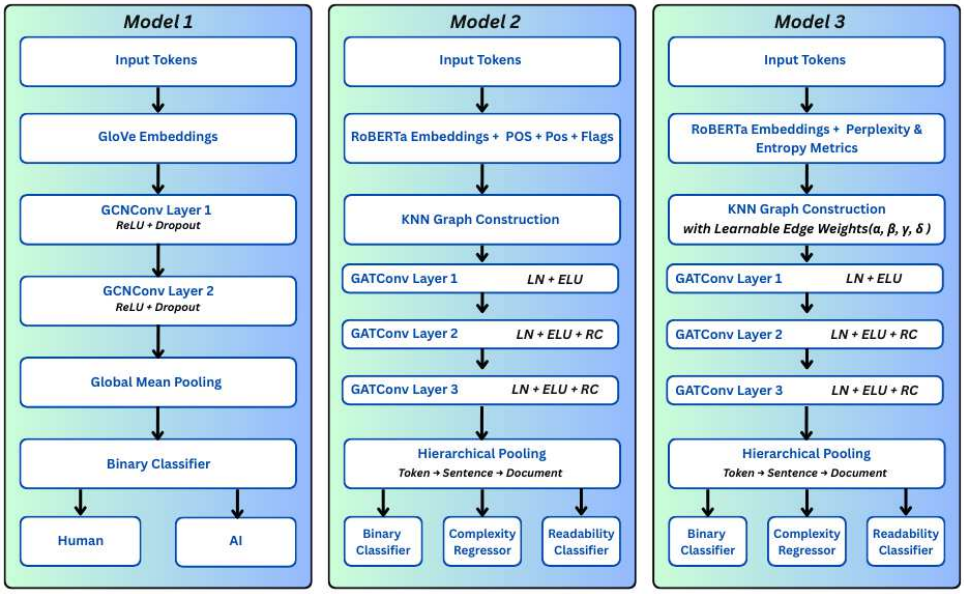


Fig. 1 The three-stage development path. Model 1 uses GloVe embeddings with GCN layers; Model 2 (AEGIS) integrates RoBERTa embeddings, part-of-speech features, and graph-attention layers with hierarchical pooling; Model 3 adds token-level uncertainty features with learnable edge weights.

The trainable footprint is 73K parameters (0.28 MB), because the transformer is frozen and serves purely as a feature extractor. We flag a fairness point directly: comparisons of "trainable parameters" favour frozen-backbone systems, and the full inference stack includes the 125M-parameter encoder. We report both figures throughout.

3.2 A transparent development path

AEGIS emerged from three deliberate stages, which we report with their failures intact.

Model 1 (GloVe + GCN). A minimal proof of concept: static embeddings, cosine-threshold graph construction, a two-layer GCN [20]. It reached 74.17% accuracy—clearly above chance, clearly below deployed utility. Its role was to establish that graph classification of text is feasible, not that it is preferable.

Model 2 (AEGIS). The architecture above: contextual embeddings, multi-modal node features, k-NN plus dependency edges, graph attention, hierarchical pooling, multi-task regularization. Accuracy jumped to 98.50% (+24.33 points), with perfect recall on human text (training behaviour in Figure 2). We note—because it matters for Section 5—that this step changed the embedding family, the graph construction, the message-passing operator, and the training objective *simultaneously*. Development-stage comparisons of this kind cannot attribute the gain to any single component, a limitation reviewers of the original system report correctly identified and that the audit in this paper repairs.

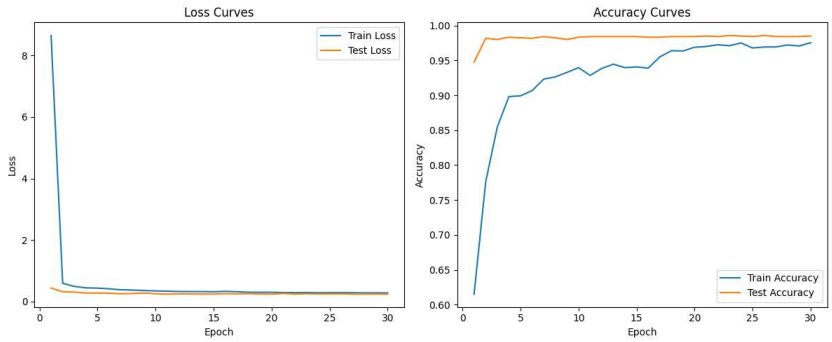


Fig. 2 Training convergence of Model 2 (AEGIS). Validation-based early stopping; the test set is never consulted during training or model selection.

Model 3 (uncertainty-aware variant). Adding token-level perplexity and entropy features with learnable edge weights did not improve on Model 2; the added complexity demanded tuning effort that did not pay for itself. We retain Model 2 as AEGIS throughout.

3.3 Legacy evaluation

Under the evaluation regime standard in the systems literature—a single generator (LLaMA-3.1), paired human/machine scientific abstracts, accuracy-based comparison—AEGIS dominates five baselines: GLTR at 59.42%, DetectGPT at 50.00%, Fast-DetectGPT at 45.92%, ZeroGPT near chance, and fine-tuned RoBERTa at 82.53% with 124.6M trainable parameters. These numbers are reproducible and, as far as they go, real. The remainder of this paper is about how far they go.

4 A Confound-Controlled Evaluation Protocol

Before interpreting any detector's accuracy, the data it is measured on must be characterized. This section describes the corpus, the four confounds our audit surfaced in it, and the control for each.

4.1 Corpus design

The corpus is built from 2,000 human-written arXiv abstracts (Figure 3). Scientific abstracts are a deliberately chosen domain: they are the text type most exposed to automated authorship screening in practice, they carry rich formatting (mathematical markup) that stresses format confounds, and their length distribution is narrow enough to make length effects measurable rather than overwhelming.

Each abstract is rewritten by *ten generator configurations*: nine unique models spanning four vendor families and both open-weight and API-only deployments—Llama-3.1-8B, Llama-3.3-70B (two configurations), Qwen-72B, GPT-4o, GPT-4o-mini, Gemini-Flash, Mistral-Large, Claude Sonnet, and DeepSeek-R1—plus one

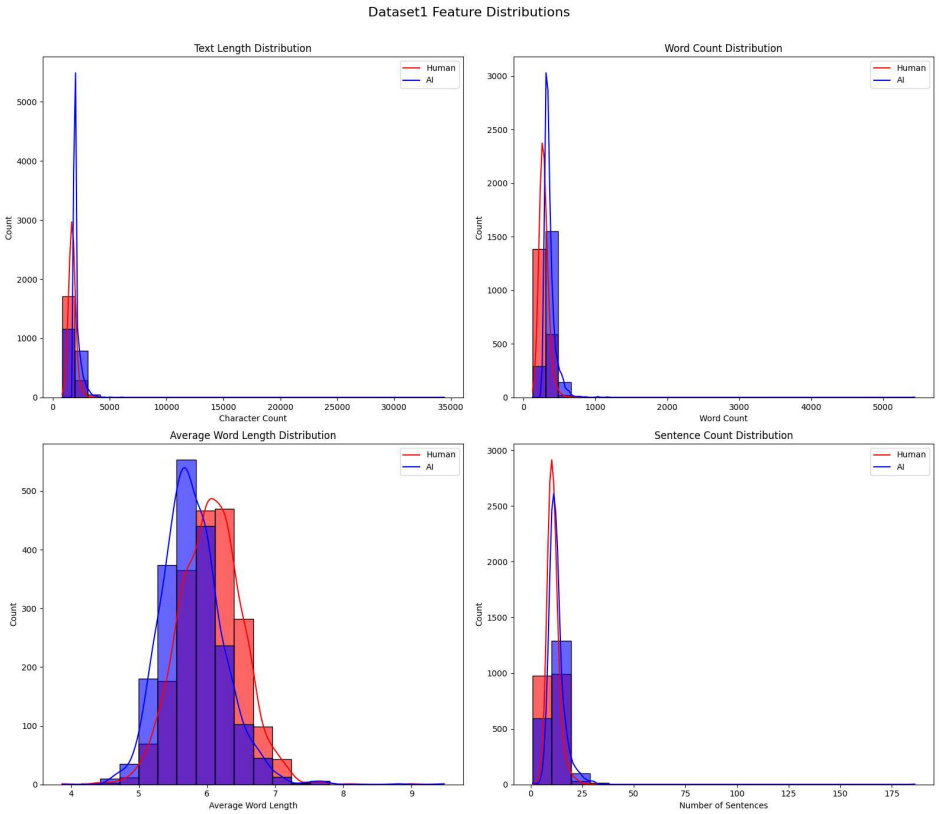


Fig. 3 Corpus statistics for the 2,000-abstract human source set and its machine rewrites: length distributions, vocabulary profiles, and class balance.

deliberate duplicate of Llama-3.3-70B under an independent sampling run, which provides the replication ruler of Section 4.6. The rewrite (rather than free-generation) design pairs every machine text with a human text on the same topic, so detectors cannot succeed by learning topical differences.

The paired design also fixes the evaluation unit: for every generator, human and machine texts descend from the same 2,000 sources, and splits are made at the source level (Section 4.4). Table 1 lists the resulting per-generator corpus statistics.

4.2 Length

Every generator we tested shortens text relative to its human source. Under natural prompting, compliance ranges from 6.8% (DeepSeek-R1) to 96.6% (GPT-4o-mini) (Table 1): a detector can score well by learning "short means machine." Our control is *generation-time length forcing*: the target length is embedded in the prompt, raising compliance substantially for every generator (e.g., Claude 70.1% to 99.2%; Llama-8B 49.8% to 82.7%; Figure 4). Filtering after generation—the occasional practice in

Table 1 Per-generator length compliance: fraction of rewrites within 20% of the source length, under natural prompting and under generation-time length forcing. Llama-70B-r denotes the replicate configuration of Llama-3.3-70B, generated only under length forcing.

Generator	Natural prompting	Length forcing
GPT-4o-mini	96.6%	97.0%
Claude	70.1%	99.2%
GPT-4o	50.9%	88.1%
Llama-8B	49.8%	82.7%
Mistral-Large	47.2%	75.1%
Gemini-Flash	38.4%	74.4%
Llama-70B	38.0%	78.1%
Qwen-72B	26.6%	66.7%
DeepSeek-R1	6.8%	13.2%
Llama-70B-r	—	78.7%

prior work—discards data and induces selection effects; constraining generation does neither.

4.3 Format

Human arXiv abstracts contain LaTeX markup that generators convert to Unicode; per-model retention of source math markup ranges from 42% to 86%, and a classifier reading only these markers reaches 0.551–0.635 AUROC. Symmetric normalization of both classes reduces the marker signal to at most 0.528 AUROC, approximately chance.

4.4 Near-copies and leakage

One generator (Llama-3.1-8B) returns near-verbatim copies of the source for 3.5% of inputs; these are excluded, since a rewrite that is not a rewrite belongs to neither class. Splits are made by source identifier, so no abstract contributes to both training and evaluation in any version, and all model selection and early stopping use a validation split, never the test set.

4.5 The baseline ladder and the tuned null

We evaluate every detector against a ladder of nulls: (i) a *surface* baseline of five shallow features; (ii) a frozen RoBERTa mean-pooled embedding followed by logistic regression with the library-default regularization ($C = 1.0$), the baseline as it typically appears in the literature; and (iii) the same probe with C selected on the validation set (chosen value 0.01), which we call the *tuned probe*. The distinction between (ii) and (iii) is a single hyperparameter, and Section 5 shows it is worth +0.027 AUROC—larger than most published architectural margins in this space.

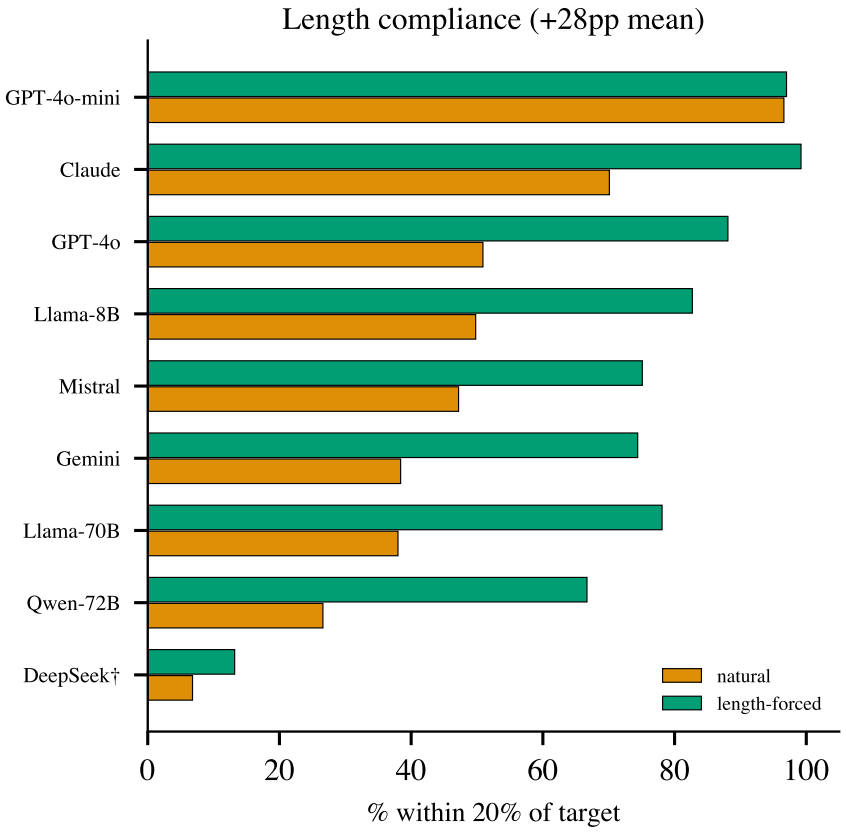


Fig. 4 Length compliance (fraction of rewrites within 20% of source length) under natural prompting versus generation-time length forcing, per generator.

4.6 The replication ruler

The corpus contains the same generator twice under independent sampling runs. The AUROC difference between these twins measures pure replication noise: 0.001–0.003. Any architectural margin smaller than this floor is indistinguishable from seed variance. We are not aware of a previously published noise floor for this task.

4.7 Zero-shot baselines, properly configured

Fast-DetectGPT and Binoculars are run with their reference configurations, and a scorer ablation (GPT-2 to GPT-2-medium on three generators) checks that conclusions are not artifacts of one scoring model.

Table 2 Per-generator AUROC under the confound-controlled protocol, sorted by probe score. δ_{graph} = AEGIS − probe. DeepSeek-R1 has $n = 263$ (directional only, †). The replication noise floor is 0.001–0.003.

Generator	Surface	Tuned probe	AEGIS	δ_{graph}
Llama-8B	0.681	0.922	0.921	−0.001
Qwen-72B	0.820	0.971	0.968	−0.002
GPT-4o	0.834	0.972	0.976	+0.003
Llama-70B-r	0.751	0.979	0.986	+0.007
Llama-70B	0.756	0.981	0.985	+0.004
GPT-4o-mini	0.821	0.984	0.988	+0.003
Mistral-Large	0.931	0.991	0.995	+0.005
Gemini-Flash	0.919	0.993	0.994	+0.001
DeepSeek†	0.980	0.993	0.995	+0.003
Claude	0.987	0.999	0.999	+0.000

5 Results

5.1 Fair-data accuracy: the ladder

Table 2 reports per-generator AUROC for the surface baseline, the tuned probe, and AEGIS under the full protocol. Two observations. First, AEGIS remains an excellent detector: 0.921–0.999 AUROC across nine unique generators. Second, the tuned probe is equally excellent, within −0.002 to +0.007 of AEGIS everywhere. The embedding stratum accounts for 92–101% of the improvement over surface features; the graph stratum accounts for the remainder, which the noise floor swallows. Figure 5 shows the same comparison graphically: in panel (a) the probe and AEGIS bars are visually inseparable on every generator, and in panel (b) the per-generator graph contribution δ_{graph} scatters around zero, with half the generators inside the replication noise band and the largest single value at +0.007.

5.2 Isolating the graph: the margin collapse

Figure 6 presents the central result as three successively fairer comparisons on the same data.

Against the surface null, AEGIS leads by +0.240 AUROC. This is the flattering comparison implicit in most system papers, and it establishes only that deep features beat shallow ones.

Against the untuned probe ($C = 1.0$), the margin is +0.027: the number a conventional "embedding baseline" row would print, and it looks like a real architectural gain.

Against the tuned probe ($C = 0.01$), the margin is −0.001 on the primary generator and −0.0004 averaged across the full transfer study—negative, and inside the 0.001–0.003 replication floor.

The phantom advantage has an exact mechanism. On the primary generator the untuned probe scores 0.894, the tuned probe 0.922, and AEGIS 0.921. A single library default, never questioned in our own original write-up and rarely questioned in the

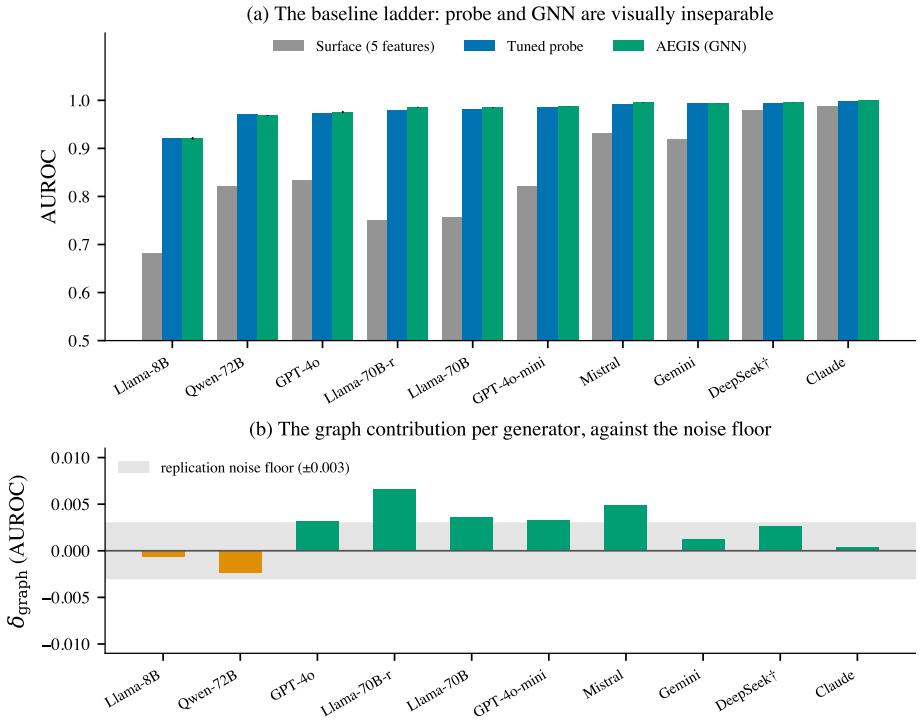


Fig. 5 (a) The baseline ladder per generator: surface features, tuned probe, and AEGIS. (b) The graph contribution $\delta_{\text{graph}} = \text{AEGIS} - \text{probe}$ per generator against the ± 0.003 replication noise band. Error bars in (a) are standard deviations over three seeds.

literature, suppressed the null by 0.028 and thereby manufactured the entire apparent contribution of a 73K-parameter graph architecture. Nothing about this mechanism is specific to AEGIS.

5.3 Transfer geography: the same detector twice

If the graph read a signal the embeddings do not carry, the two detectors should fail differently somewhere in a 10×10 cross-generator transfer matrix: ten specialists, each trained on one generator and tested on all ten, three seeds, run identically for AEGIS and the tuned probe.

They do not fail differently anywhere (Figure 7). Across the 90 off-diagonal cells the two transfer surfaces correlate at $\rho = 0.986$ (range 0.984–0.989 across seeds); AEGIS wins 47 of 90 cells—a coin flip; and the mean margin is -0.0004 AUROC. The transfer structure itself is scientifically meaningful (some generator families transfer to each other far better than others), but it is the *same* structure for both detectors: whatever map of generator space exists in this corpus, both models read it from the embeddings.

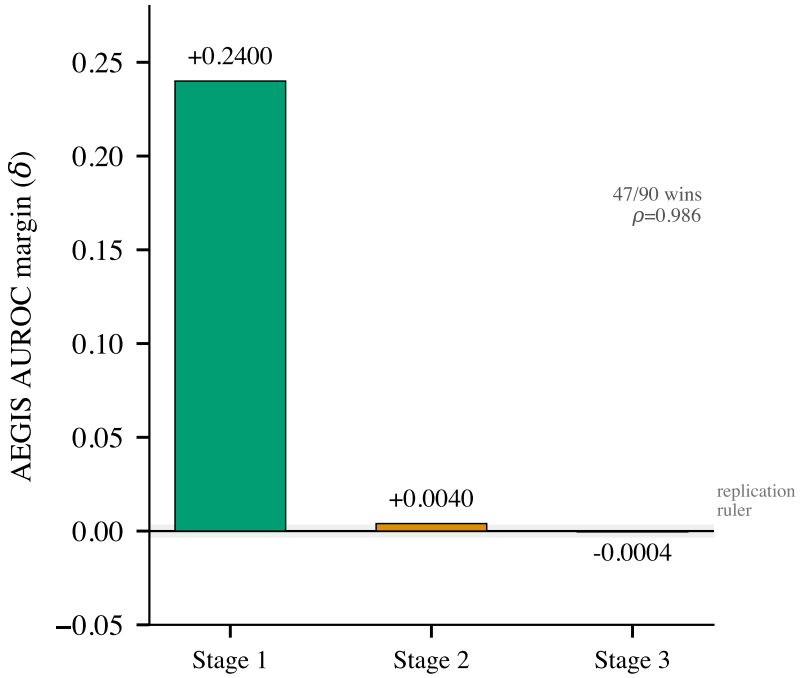


Fig. 6 The margin collapse. The graph detector's advantage under three successively fairer nulls: +0.240 over surface features, +0.027 over the untuned embedding probe, and −0.0004 over the validation-tuned probe (mean over the 10 × 10 × 3-seed transfer study). The shaded band is the replication noise floor.

5.4 Zero-shot detectors under length control

The protocol also re-scores the zero-shot baselines that the legacy evaluation beat. Under length control they do not merely lag—they collapse (Figure 8). Fast-DetectGPT averages 0.500 AUROC (range 0.382–0.544 over ten generators): chance. Binoculars averages 0.474 and *inverts* on the strongest generators, reaching 0.213 on Claude—it rates Claude's rewrites as more human than the human originals. The scorer ablation reproduces the pattern with a larger scoring model, so this is not an artifact of one scorer choice. Table 3 gives the per-generator breakdown, which carries a second observation: both statistics degrade most on the strongest commercial generators (Claude, GPT-4o, DeepSeek-R1) and retain most signal on the small open model (Llama-8B). Likelihood-based statistics thus fail exactly where detection matters most. Benign length control, introduced as an evaluation hygiene measure, is more damaging to these detectors than published adversarial attacks; statistics that ride on length and verbosity do not survive its removal.

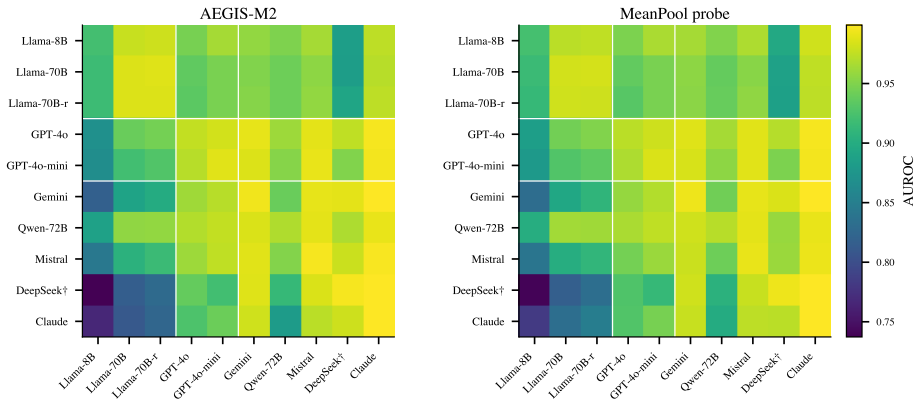


Fig. 7 Cross-generator transfer for AEGIS and the tuned probe: ten specialists × ten test sets × three seeds. The two geographies are near-identical ($\rho = 0.986$ over the 90 off-diagonal cells; AEGIS wins 47/90).

Table 3 Zero-shot detector AUROC per generator under generation-time length control. Values below 0.5 are systematically inverted. The tuned probe column from Table 2 is repeated for reference.

Generator	Fast-DetectGPT	Binoculars	Tuned probe
Claude	0.382	0.213	0.999
GPT-4o	0.474	0.429	0.972
DeepSeek-R1	0.476	0.303	0.993
GPT-4o-mini	0.501	0.457	0.984
Gemini-Flash	0.502	0.416	0.993
Llama-70B-r	0.517	0.627	0.979
Mistral-Large	0.532	0.472	0.991
Qwen-72B	0.533	0.580	0.971
Llama-70B	0.542	0.636	0.981
Llama-8B	0.544	0.607	0.922
Mean	0.500	0.474	0.979

5.5 The structure is real; the information is not new

Finally, the structural claims of the original system report deserve their own verdict. The measurements were correct: human text graphs show higher clustering coefficients (0.4681 vs. 0.4380, $p < 0.001$) and higher community modularity [21] (0.6159 vs. 0.5826, $p < 0.001$) than machine text graphs (Figure 9). What the audit overturns is the *interpretation*. Because the graphs are built from embedding-space neighbourhoods, their topology is a function of the embedding geometry; the structural statistics are a visible, human-readable re-encoding of class differences the embeddings already carry. They are useful as description—they show *what* the separation looks like in relational terms—but they are not an additional information source, which is precisely

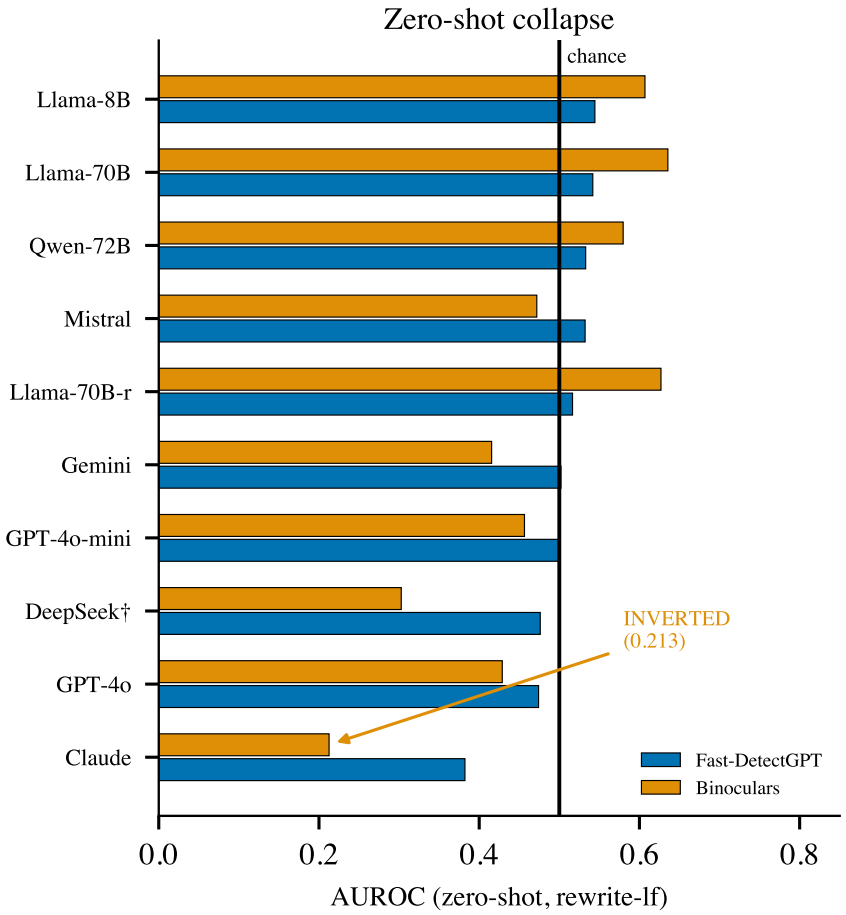


Fig. 8 Zero-shot detectors under generation-time length control. Fast-DetectGPT is at chance (mean 0.500); Binoculars inverts on strong generators (0.213 on Claude, i.e., systematically wrong).

why a detector reading them cannot beat a detector reading the embeddings directly. The point is visible without any graph at all: a two-dimensional t-SNE projection of the raw document embeddings (Figure 10) already separates the classes cleanly. The separation the graph statistics describe pre-exists them in the embedding space they are computed from.

6 Discussion

6.1 How phantom advantages arise

Our lifecycle exhibits, in one system, the three mechanisms by which an architecture acquires an undeserved mechanism story. First, *untuned nulls*: the default-regularization embedding probe is the standard baseline, and its default is not the right

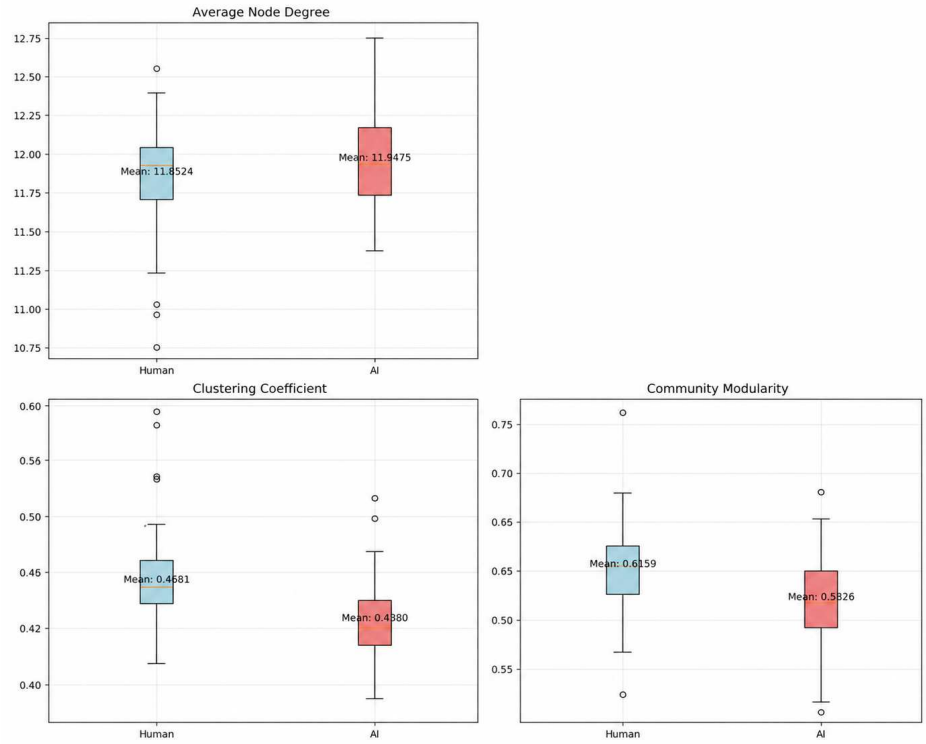


Fig. 9 Structural differences between human and machine text graphs (clustering, modularity, degree structure). The differences are statistically real but derive from embedding geometry: they re-encode, rather than extend, the embedding signal.

value for this task; the gap between default and tuned (+0.027 AUROC) exceeds the architectural margins typically reported. Second, *simultaneous-change development*: our own Model 1-to-Model 2 step changed four components at once, so the +24-point jump could be—and was—misread as evidence for the graph. Third, *missing noise floors*: without a replication ruler, a +0.004 margin reads as a small win rather than as seed noise. None of these mechanisms requires bad faith; all three are default practice.

6.2 An audit checklist for detector claims

The protocol compresses into five practices, none expensive: (1) control length at generation time, and normalize format symmetrically, before any training; (2) tune the null model with exactly the care given to the proposed model, on validation data; (3) change one component at a time when attributing gains; (4) measure a test–retest noise floor and refuse to interpret margins inside it; (5) evaluate transfer across generators, paired between the proposed model and the null, because equivalent aggregate accuracy can hide divergent failure modes—and, in our case, did not.

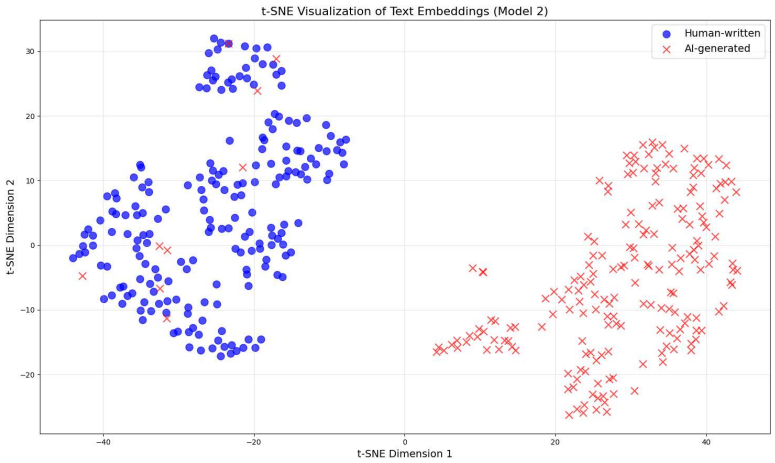


Fig. 10 t-SNE projection of the frozen RoBERTa document embeddings. The human/machine separation that the graph statistics describe is already present in the embedding space from which the graphs are constructed.

6.3 Implications for graph-based detection

Our findings do not overturn the accuracy of published graph detectors, and our audit concerns one system on one domain. But the burden of proof now sits differently. IDEATE, SGG-ATD, and text-graph systems [1–3] each report gains over baselines that are either untuned, non-embedding, or altered along several axes at once; the explainability agenda of Zheng et al. [4] presumes the motifs carry task information beyond the features they are built from. Every one of these claims is testable with the checklist above at the cost of one additional baseline column. Until that column exists, the parsimonious reading of our results—and of Galke and Scherp [11] in general text classification—is that graphs built from learned representations describe those representations; they do not exceed them.

6.4 Limitations

Our corpus is English scientific abstracts; other domains, languages and text lengths may distribute the signal differently. The audit evaluates one graph architecture family (GAT over k-NN-plus-dependency graphs); graphs built from information genuinely external to the embeddings—citation networks, factual knowledge bases as in IDEATE—could in principle escape the redundancy argument, and deserve the same test rather than the same verdict. Our probe null is built on one embedding family; the direction of our conclusion is robust to this choice in our experiments, but magnitudes are instrument-relative. Finally, all evaluations are in-distribution with respect to the human corpus; adversarial and out-of-domain robustness are separate questions.

7 Conclusion

We built a graph neural detector good enough to believe in, and then built the evaluation that showed why we should not have. The detector's accuracy is real. Its mechanism story—graph topology capturing what sequential models miss—did not survive a tuned baseline, a paired transfer study, and a noise floor. A single untuned default in the standard baseline manufactured the entire apparent architectural advantage.

The general lesson costs one sentence: *accuracy claims and mechanism claims require different evidence*. Accuracy needs fair data; mechanism needs a calibrated alternative. The field of machine-text detection currently produces the first kind of evidence in abundance and the second kind almost never. We offer our lifecycle—system, protocol, collapse, and checklist—as a template for closing that gap, and we suggest that any detector paper claiming an architectural mechanism include the one baseline column that would have saved ours: the same features, tuned, without the architecture.

Declarations

Funding. This research received no external funding.

Conflict of interest. The authors declare no conflict of interest.

Data availability. The corpus manifests, generation prompts, evaluation protocol, and per-experiment result files underlying every number in this paper are available from the corresponding author on request.

Author contribution. A. Alaktif: conceptualization, methodology, software, investigation, writing—original draft. M. Chergui: supervision, methodology, writing—review and editing. G. Faiq: validation, writing—review and editing. A. Ammoumou: supervision, writing—review and editing. All authors read and approved the final manuscript.

References

[1] Wang, Q., Zhang, L., Guo, Z., Mao, Z.: IDEATE: Detecting AI-generated text using internal and external factual structures. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp. 8556–8568 (2024)

[2] Gupta, N., Li, Q.: Seeing through the mask: AI-generated text detection with similarity-guided graph reasoning. In: Findings of the Association for Computational Linguistics: IJCNLP-AACL 2025, pp. 1350–1360 (2025)

[3] Valdez-Valenzuela, A., Gómez-Adorno, H., Montes-y-Gómez, M.: Text graph neural networks for detecting AI-generated content. In: Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect), pp. 134–139 (2025)

[4] Zheng, X., Chen, Z., Schafir, E., Chen, S., Salehi, H.A., Chen, H., Shirani, F., Cheng, W., Luo, D.: LM^{2}otifs: An explainable framework for machine-generated texts detection. arXiv preprint arXiv:2505.12507 (2025)

[5] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y.: Graph attention networks. In: Proceedings of the 6th International Conference on Learning Representations (2018)

[6] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)

[7] Gehrmann, S., Strobelt, H., Rush, A.M.: 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. Association for Computational Linguistics, ??? (2019)

[8] Mitchell, E., Lee, Y., Khazatsky, A., Manning, C.D., Finn, C.: DetectGPT: Zero-shot machine-generated text detection using probability curvature. In: Proceedings of the 40th International Conference on Machine Learning, vol. 202, pp. 24950–24962. PMLR, ??? (2023)

[9] Bao, G., Zhao, Y., Teng, Z., Yang, L., Zhang, Y.: Fast-DetectGPT: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In: Proceedings of the 12th International Conference on Learning Representations (2024)

[10] ZeroGPT: AI Text Detector. https://www.zerogpt.com (2023)

[11] Galke, L., Scherp, A.: Bag-of-words vs. graph vs. sequence in text classification: Questioning the necessity of text-graphs and the surprising strength of a wide MLP. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL) (2022)

[12] Yao, L., Mao, C., Luo, Y.: Graph convolutional networks for text classification. In: Proceedings of the 33rd AAAI Conference on Artificial Intelligence, pp. 7370–7377 (2019)

[13] Kuznetsov, K., Tulchinskii, E., Kushnareva, L., Magai, G., Barannikov, S., Nikolenko, S., Piontkovskaya, I.: Robust AI-generated text detection by restricted embeddings. In: Findings of EMNLP 2024 (2024)

[14] Hans, A., Schwarzschild, A., Cherepanova, V., Kazemi, H., Saha, A., Goldblum, M., Geiping, J., Goldstein, T.: Spotting LLMs with binoculars: Zero-shot detection of machine-generated text. In: Proceedings of the 41st International Conference on Machine Learning, pp. 17519–17537 (2024)

[15] Sadasivan, V.S., Kumar, A., Balasubramanian, S., Wang, W., Feizi, S.: Can AI-generated text be reliably detected? Transactions on Machine Learning Research (2025)

[16] Chakraborty, S., Bedi, A.S., Zhu, S., An, B., Manocha, D., Huang, F.: Position: On the possibilities of AI-generated text detection. In: Proceedings of the 41st International Conference on Machine Learning, pp. 6093–6115 (2024)

[17] Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., Goldstein, T.: A watermark for large language models. In: Proceedings of the 40th International Conference on Machine Learning, vol. 202, pp. 17061–17084. PMLR, ??? (2023)

[18] Dugan, L., Hwang, A., Trhlík, F., Zhu, A., Ludan, J.M., Xu, H., Ippolito, D., Callison-Burch, C.: 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 (ACL), pp. 12463–12492 (2024)

[19] Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., Zou, J.: GPT detectors are biased against non-native English writers. Patterns **4**(7) (2023)

[20] Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: Proceedings of the 5th International Conference on Learning Representations (2017)

[21] Newman, M.E.J.: Modularity and community structure in networks. Proceedings of the National Academy of Sciences **103**(23), 8577–8582 (2006) https://doi.org/10.1073/pnas.0601602103