
Alignment is the Watermark: How Making LLMs Helpful Makes Them Detectable

Anonymous Author(s)

Abstract

AI-generated text detection is often framed as an arms race between generators and detectors. We argue for a different view: alignment training—the process that makes large language models (LLMs) helpful, harmless, and honest—simultaneously creates a robust, implicit watermark that makes aligned outputs inherently more detectable than base model outputs. We test this hypothesis across three model families (Mistral, MPT, Cohere) using three complementary detection paradigms: distributional feature analysis, zero-shot statistical detection, and LLM-as-detector. Across 12 independent comparisons, aligned models are more detectable than their base counterparts in 11 cases (sign test $p = 0.006$). The strongest effect appears in zero-shot statistical detection, where alignment increases AUROC by 10–35 percentage points; base models like MISTRAL-7B and MPT-30B are near-random (AUROC ≈ 0.50), while their aligned variants reach AUROC 0.74–0.88. An LLM detector (GPT-4.1) achieves 97.5–100% true positive rates on aligned text versus 85–96.3% on base text. We trace the mechanism to alignment’s regularizing effect: aligned models produce text with lower sentence-length variability (Cohen’s $d = -0.45$ to -0.66) and higher token-level predictability. These findings reframe the detection problem: alignment *is* the watermark, and making AI useful guarantees its detectability.

1 Introduction

Can we reliably detect AI-generated text? This question has driven a growing body of research on watermarking [Kirchenbauer et al., 2023, Dathathri et al., 2024], statistical detection [Mitchell et al., 2023, Bao et al., 2024], and trained classifiers [He et al., 2024]. Yet a more fundamental question has received less attention: *why* is AI text detectable in the first place?

We propose a simple answer. A language model trained only to match the data distribution—a base model—approximates human text and should be difficult to detect. The moment we align that model to be helpful, harmless, and honest, we push its output distribution away from the human distribution. This distributional shift is not an accidental byproduct; it is the *purpose* of alignment. And it creates an inherent, implicit watermark: aligned models produce text that is systematically different from human text in ways that serve human preferences but simultaneously enable detection.

Why does this matter? If alignment is the watermark, then detection of aligned AI text is not a temporary arms race that better generators will eventually win. It is a fundamental consequence of making AI useful. This insight has immediate implications for AI policy (explicit watermarking may be redundant for aligned models), for the detection community (understanding *why* detection works is more valuable than building incrementally better detectors), and for AI safety (the detectability of aligned models provides a natural accountability mechanism).

What is the gap in prior work? Individual studies have shown that RLHF increases detectability [Xu and Zubiaga, 2025], that alignment collapses output diversity [Kirk et al., 2024], and that reward models can serve as detectors [Lee et al., 2024]. Mitchell et al. [2023] noted that LLMs “watermark themselves implicitly,” and Chakraborty et al. [2023] proved that detection is theoretically

possible whenever the machine and human distributions differ. However, no prior work has systematically tested the thesis that alignment itself constitutes an implicit watermark across multiple model families and detection paradigms.

Our approach. We conduct a multi-method empirical study using the RAID benchmark [Dugan et al., 2024], which provides matched base/aligned model pairs from three families: Mistral, MPT, and Cohere. We apply three complementary detection paradigms: (1) distributional feature analysis to measure *how* alignment changes text properties; (2) zero-shot statistical detection to test *whether* these changes enable detection; and (3) an LLM-as-detector approach to test whether aligned models are detectable “by other AIs.”

What do we find? Aligned models are more detectable than base models in 11 of 12 comparisons across three families and four detection methods (sign test $p = 0.006$). Zero-shot statistical detection shows the largest effect: alignment increases AUROC by 10–35 percentage points. Base MISTRAL-7B and MPT-30B are statistically indistinguishable from human text (AUROC ≈ 0.50), while their aligned variants reach AUROC 0.74–0.88. GPT-4.1 as a detector achieves 97.5–100% true positive rates on aligned text versus 85–96.3% on base text. The mechanism is consistent: alignment reduces sentence-length variability (Cohen’s $d = -0.45$ to -0.66) and increases token-level predictability across all three families.

In summary, our main contributions are:

- We formalize and empirically test the hypothesis that alignment training creates an implicit watermark, providing the first systematic multi-family, multi-method study of this phenomenon.
- We demonstrate that alignment increases AUROC by 10–35 percentage points for zero-shot statistical detection and pushes LLM-based detection to near-perfect rates (97.5–100%), across three model families.
- We identify the mechanism behind the alignment watermark—reduced structural variability and increased token predictability—and show it generalizes across model architectures (11/12 comparisons, $p = 0.006$).

2 Related Work

AI-generated text detection. Detection methods fall into three categories: trained classifiers, zero-shot statistical methods, and explicit watermarking. Trained classifiers fine-tune language models on labeled human/machine text [He et al., 2024, Li et al., 2024]. Zero-shot methods exploit statistical signatures of machine text without training data. DETECTGPT [Mitchell et al., 2023] uses probability curvature: machine text occupies regions of negative curvature in the model’s log-probability surface, so random perturbations consistently decrease its log-probability. FAST-DETECTGPT [Bao et al., 2024] replaces perturbation-based curvature estimation with conditional probability curvature, achieving similar accuracy 340 \times faster. Simpler baselines—thresholding on mean log-probability, log-rank, or entropy—remain competitive in many settings [Gehrmann et al., 2019]. Explicit watermarking modifies the generation process to embed a detectable signal [Kirchenbauer et al., 2023, Dathathri et al., 2024]. Our work is orthogonal to all three categories: we study a signal that arises naturally from alignment, without any modification to the generation process.

Effects of alignment on LLM outputs. Alignment training—spanning supervised fine-tuning (SFT), reinforcement learning from human feedback (RLHF), and direct preference optimization (DPO)—changes model outputs in well-documented ways. Kirk et al. [2024] provide the first rigorous demonstration that RLHF causes *mode collapse*: per-input lexical diversity drops by 75–90%, and across-input outputs converge to a consistent stylistic mode. The standard KL penalty fails to recover this diversity. Ouyang et al. [2022] show that instruction tuning with RLHF dramatically changes output characteristics including format compliance and safety properties. These distributional shifts are precisely what our hypothesis predicts will create a detectable watermark.

Alignment and detectability. The closest work to ours is Xu and Zubiaga [2025], who measure detectability at three alignment stages (base, SFT, PPO) for Llama-7B. They find that FAST-DETECTGPT AUROC increases monotonically from 0.68 (base) to 0.91 (PPO) for instruction-following text, with a concurrent 57% reduction in distinct n-gram scores. Lee et al. [2024] show that reward models—trained during RLHF—achieve 95–99% AUROC when used as detectors for aligned model outputs, but only 60–75% for base models, suggesting the alignment signal is directly

Base Model	Aligned Variant	Texts
MISTRAL-7B	MISTRAL-7B-CHAT	500 each
MPT-30B	MPT-30B-CHAT	500 each
COHERE	COHERE-CHAT	500 each
Human	—	500

Table 1: Dataset composition. All texts are scientific abstracts from the RAID benchmark. Each model variant contributes 500 clean (unattacked) texts.

encoded in the reward model’s learned preferences. Our work extends these findings in three ways: we test across three model families (not one), we use three complementary detection paradigms (not one), and we frame the result as a general principle rather than an empirical observation about a specific model.

Theoretical foundations. Chakraborty et al. [2023] prove that AI text detection is possible if and only if the total variation distance between the machine and human distributions is non-zero. For any positive TV distance δ , multi-sample detection AUROC approaches 1.0 exponentially with sample count n . Our hypothesis connects directly to this result: alignment training *guarantees* a non-zero TV distance by design, because its purpose is to shift the model’s output distribution toward human preferences and away from the data distribution. A perfect base model ($TV = 0$) would be undetectable but also unaligned; alignment ensures $TV > 0$ and thus ensures detectability.

Benchmarks. The RAID benchmark [Dugan et al., 2024] provides 6.2M+ generations across 11 LLMs, 8 domains, and 11 adversarial attacks. Critically for our study, it includes matched base/chat pairs for three model families: Mistral, MPT, and Cohere. Other benchmarks include MGTBench [He et al., 2024] with 13 detection methods and MAGE [Li et al., 2024] with 27 generators. We use RAID because its base/chat pair design directly enables our base-vs.-aligned comparison.

3 Methodology

We test the alignment watermark hypothesis using three complementary detection paradigms applied to matched base/aligned model pairs. Each paradigm addresses a different aspect of the hypothesis: distributional feature analysis measures *how* alignment changes text, zero-shot statistical detection tests *whether* these changes enable detection, and LLM-as-detector tests whether aligned models are “detectable by other AIs.”

3.1 Data

We use the RAID benchmark [Dugan et al., 2024], the largest publicly available AI text detection dataset. From RAID we extract matched base/aligned pairs from three model families, along with a human baseline (table 1). We restrict to domain `abstracts` (scientific abstracts) and attack `none` (clean, unmodified generations) to avoid confounding adversarial modifications with alignment signals. Each model variant contributes exactly 500 texts (250 greedy decoding + 250 sampling).

3.2 Experiment 1: Distributional Feature Analysis

We compute eight text-level features for each sample to quantify how alignment changes the distributional properties of generated text:

- **Type-token ratio (TTR):** vocabulary richness, defined as $|types|/|tokens|$.
- **Distinct n -gram ratios** ($n = 1, 2, 3$): the fraction of unique n -grams, measuring lexical diversity.
- **Mean and std. sentence length:** structural regularity of the text.
- **Mean word length:** a proxy for vocabulary complexity.
- **Hapax legomena ratio:** proportion of words appearing exactly once.
- **Token count:** overall text length.

For each feature, we compare base vs. aligned distributions within each model family using Mann-Whitney U tests with Bonferroni correction across features ($\alpha = 0.05/8$). We report Cohen’s d effect sizes to quantify the practical magnitude of each shift.

3.3 Experiment 2: Zero-Shot Statistical Detection

We test whether the distributional shifts identified in Experiment 1 translate to actual detectability differences using established zero-shot detection signals.

Reference model. We use GPT-2-LARGE (774M parameters) to compute per-token log-probabilities for each text. GPT-2 Large is the standard reference model in the detection literature [Mitchell et al., 2023, Bao et al., 2024] and provides a model-agnostic detection signal.

Detection signals. From the per-token log-probabilities, we extract three signals:

- **Mean log-probability:** higher values indicate text more expected by the reference model.
- **Mean log-rank:** lower values indicate tokens are more highly ranked by the reference model.
- **Mean entropy:** lower values indicate the reference model is more certain about the next token.

Evaluation. For each AI source (base or aligned) and each signal, we compute AUROC for the binary task of distinguishing human text from that AI source. We estimate 95% bootstrap confidence intervals with 1,000 iterations. The key comparison is AUROC(human vs. aligned) versus AUROC(human vs. base) for each family.

3.4 Experiment 3: LLM-as-Detector

We test the specific claim that aligned models are “detectable by other AIs” using GPT-4.1 as a zero-shot classifier.

Setup. We sample 80 texts from each of seven categories (human, base-Mistral, aligned-Mistral, base-MPT, aligned-MPT, base-Cohere, aligned-Cohere), yielding 560 texts total. Each text is classified independently using a zero-shot prompt that instructs GPT-4.1 to analyze writing style, vocabulary diversity, naturalness, and formulaic patterns, and return a binary label (human or AI) with a confidence score in $[0, 1]$. Temperature is set to 0 for deterministic classification.

Evaluation. We compute true positive rate (TPR, the rate at which AI text is correctly classified as AI), true negative rate (TNR, the rate at which human text is correctly classified as human), and AUROC using the confidence scores. We compare base vs. aligned TPR within each family using two-proportion z -tests.

3.5 Experiment 4: Cross-Family Consistency

We aggregate results across families and detection methods. For each of the 12 family \times metric comparisons (3 families \times 3 statistical signals + 3 families \times 1 LLM detector), we record whether aligned AUROC/TPR exceeds base AUROC/TPR. We use a binomial sign test to assess whether the fraction of positive comparisons significantly exceeds the null expectation of 50%.

4 Results

4.1 Alignment Shifts Text Distributions (Experiment 1)

Table 2 presents Cohen’s d effect sizes for distributional feature differences between base and aligned models within each family. Two patterns are consistent across all three families:

- **Alignment reduces sentence-length variability.** The standard deviation of sentence length decreases significantly with alignment in all three families ($d = -0.66$ for Mistral, $d = -0.64$ for MPT, $d = -0.45$ for Cohere; all $p < 0.001$ after Bonferroni correction).
- **Alignment produces shorter text.** Token count decreases across all families ($d = -1.57$ for Mistral, $d = -2.25$ for MPT, $d = -0.22$ for Cohere).

Lexical diversity measures show mixed patterns. Alignment *increases* distinct n -gram ratios for Mistral ($d = +0.92$ for distinct 2-grams) and MPT ($d = +1.13$), but *decreases* them for Cohere

Feature	MISTRAL	MPT	COHERE	Direction
Type-token ratio	+0.57	+0.64	−0.29	Mixed
Distinct 2-gram	+0.92	+1.13	−0.28	Mixed
Distinct 3-gram	+1.02	+1.21	−0.26	Mixed
Mean sent. length	−0.11	−0.52	+0.08	Mixed
Std sent. length	−0.66	−0.64	−0.45	Consistent ↓
Mean word length	+0.56	+0.01	+0.19	Mostly ↑
Hapax ratio	+0.68	+0.75	−0.27	Mixed
Num tokens	−1.57	−2.25	−0.22	Consistent ↓

Table 2: Cohen’s d effect sizes for base vs. aligned distributional features within each model family. Positive values indicate aligned > base. Bold entries mark features that shift consistently across all families. Alignment consistently reduces sentence-length variability and text length.

($d = -0.28$). This suggests that the universal alignment watermark lies in *structural regularity*, not necessarily vocabulary narrowing.

Comparing aligned text to human text reveals the same signature: aligned text has significantly lower sentence-length variation than human text ($d = -0.65$, $p < 10^{-144}$), is shorter ($d = -1.29$, $p < 10^{-137}$), and has higher type-token ratio ($d = +0.66$, $p < 10^{-33}$). Base model text is closer to the human distribution on most metrics.

4.2 Alignment Dramatically Increases Statistical Detectability (Experiment 2)

Table 3 presents AUROC values for distinguishing human text from each AI source using three zero-shot detection signals. The results strongly support the alignment watermark hypothesis.

Base models approach undetectability. Base MISTRAL-7B and MPT-30B have AUROC values near 0.50 for mean log-probability (0.514 and 0.504, respectively) and mean log-rank (0.535 and 0.506). These models are statistically indistinguishable from human text by these metrics, consistent with the prediction that a model trained to match the data distribution should approximate the human distribution.

Alignment dramatically increases detectability. Aligned variants reach AUROC 0.71–0.88 for the same metrics, representing increases of 10–35 percentage points (table 3). The largest effect is for MISTRAL-7B-CHAT on mean log-rank: AUROC jumps from 0.535 (base) to **0.881** (aligned), a Δ of +0.346. Even COHERE, which has higher baseline detectability (AUROC 0.84–0.86), shows a further 9–11 point increase with alignment.

Entropy is the weakest signal. Mean entropy shows smaller and less consistent alignment effects (Δ ranges from -0.014 to $+0.110$), suggesting that the alignment watermark is primarily about token-level predictability (log-probability, log-rank) rather than overall uncertainty.

4.3 LLMs Detect Aligned Text Near-Perfectly (Experiment 3)

Table 4 presents GPT-4.1 detection rates for each text category. Aligned model text is detected with near-perfect accuracy: GPT-4.1 achieves 100% TPR on aligned Mistral and MPT text, and 97.5% on aligned Cohere text. Base model text is also highly detectable (85–96.3% TPR), but aligned text consistently reaches higher rates.

The improvement is statistically significant for Mistral ($\Delta = +6.5$ pp, $p = 0.021$) and MPT ($\Delta = +15.0$ pp, $p = 0.0003$). Cohere shows a smaller, non-significant improvement ($\Delta = +1.3$ pp, $p = 0.650$), likely because its base model is already near-ceiling.

Human text is often misclassified. GPT-4.1 achieves only 56.2% TNR on human text, frequently labeling human scientific abstracts as AI-generated. This reflects the inherent formulaic nature of

Family	Signal	Base AUROC	Aligned AUROC	Δ
MISTRAL	Mean log-prob	0.514	0.831	+0.317
	Mean log-rank	0.535	0.881	+0.346
	Mean entropy	0.635	0.724	+0.088
MPT	Mean log-prob	0.504	0.711	+0.206
	Mean log-rank	0.506	0.740	+0.234
	Mean entropy	0.587	0.573	−0.014
COHERE	Mean log-prob	0.861	0.954	+0.092
	Mean log-rank	0.840	0.949	+0.109
	Mean entropy	0.620	0.731	+0.110

Table 3: AUROC for distinguishing human text from AI text using zero-shot statistical detection signals. Aligned models are substantially more detectable than base models across all families and primary signals. Best results per row in **bold**.

	MISTRAL		MPT		COHERE	
	Base	Aligned	Base	Aligned	Base	Aligned
TPR	0.935	1.000	0.850	1.000	0.963	0.975
Mean AI score	0.896	0.936	0.833	0.943	0.899	0.919
Δ TPR	+0.065 ($p = 0.021$)		+0.150 ($p = 0.0003$)		+0.013 ($p = 0.650$)	

Table 4: GPT-4.1 detection performance by model family. TPR is the rate at which AI text is correctly classified as AI. Human text TNR is 56.2%. Aligned models are detected at higher rates than base models, significantly so for Mistral and MPT. Best per-family results in **bold**.

scientific writing and suggests that the AUROC of ~ 0.77 – 0.80 is driven more by confidence calibration than perfect binary classification.

4.4 The Alignment Watermark Generalizes Across Families (Experiment 4)

Across the 12 family \times metric comparisons, aligned models are more detectable than base models in 11 cases. The single exception is MPT’s mean entropy ($\Delta = -0.014$), which is negligible. A binomial sign test confirms that this consistency significantly exceeds the null expectation of 50% ($p = 0.006$).

Table 5 summarizes the cross-family results. The alignment watermark is not an artifact of one model architecture or one detection method—it is a cross-family, cross-paradigm phenomenon.

5 Discussion

5.1 The Mechanism: Structural Regularity

Our results trace the alignment watermark to a specific mechanism: alignment produces text with lower structural variability. Sentence-length standard deviation decreases consistently across all three families (Cohen’s $d = -0.45$ to -0.66), and token-level predictability increases (log-probability AUROC jumps by 9–32 percentage points). This “smoothing” effect—well-structured, predictable prose—is precisely what makes aligned models helpful to users and simultaneously what makes them detectable.

The lexical diversity results provide a useful nuance. Contrary to Kirk et al. [2024], who found universal diversity collapse under RLHF, we observe that alignment *increases* distinct n -gram ratios

Signal	MISTRAL	MPT	COHERE	Consistent?
Mean log-prob	✓	✓	✓	Yes
Mean log-rank	✓	✓	✓	Yes
Mean entropy	✓	×	✓	No
GPT-4.1 TPR	✓	✓	✓	Yes

Table 5: Cross-family consistency of the alignment watermark. ✓ indicates aligned AUROC/TPR > base. 11 of 12 comparisons show aligned > base (sign test $p = 0.006$).

Finding	Our result	Prior work
AUROC base→aligned	+0.09 to +0.35	+0.23 [Xu and Zubiaga, 2025]
Structural regularity	$d = -0.45$ to -0.66	75–90% diversity reduction [Kirk et al., 2024]
Reward model detection	N/A	95–99% AUROC [Lee et al., 2024]
Cross-family generalization	3 families consistent	First systematic demonstration

Table 6: Comparison to prior work. Our results extend single-family findings to a cross-family setting.

for Mistral and MPT but *decreases* them for Cohere. This suggests that vocabulary narrowing is not the universal signature of alignment; structural regularity is. The alignment watermark manifests primarily in *how* text is organized (sentence structure, predictability), not necessarily in *which words* are used.

5.2 Comparison to Prior Work

Our findings are consistent with and extend the prior literature (table 6). The AUROC improvements we observe (+0.09 to +0.35) span the range reported by Xu and Zubiaga [2025] (+0.23 for FAST-DETECTGPT on Llama). Our sentence-length variability reduction ($d = -0.45$ to -0.66) aligns with the diversity collapse documented by Kirk et al. [2024]. The key extension is demonstrating that these effects hold across three model families, not just one.

5.3 Why Cohere Differs

COHERE base model is substantially more detectable than Mistral or MPT base models (AUROC 0.84–0.86 vs. ~ 0.50). This may reflect that Cohere’s “base” model already incorporates some instruction tuning, or that its pretraining data distribution diverges from human scientific abstracts. The alignment watermark Δ for Cohere is correspondingly smaller (+0.09 to +0.11 AUROC) but still positive. This observation suggests that the watermark strength depends on how far the base model already is from the “aligned distribution”: models that start closer to alignment show smaller marginal detectability gains.

5.4 The GPT-4.1 False Positive Problem

GPT-4.1 misclassifies 43.8% of human scientific abstracts as AI-generated. This is consistent with observations that modern scientific writing is increasingly formulaic [Guo et al., 2023], overlapping with AI text patterns. The high false positive rate means that GPT-4.1’s strong TPR on aligned text does not translate to reliable deployment as a standalone detector. Rather, its differential performance between base and aligned text confirms the existence of the alignment watermark signal.

5.5 Limitations

Single domain. All texts are scientific abstracts. Results may differ for creative writing, conversation, or code. Xu and Zubiaga [2025] found reversed effects on code-mixed text (StackExchange), suggesting domain-specific interactions.

Limited model families. Three families (Mistral, MPT, Cohere) provide convergent evidence, but testing on additional families (Llama, GPT, Gemma) would strengthen generalizability claims.

Alignment method not controlled. We compare base vs. aligned outputs but cannot isolate contributions of SFT, RLHF, or DPO, as the RAID dataset does not provide intermediate alignment stages. Xu and Zubiaga [2025] show that RLHF contributes the largest detectability increase, but SFT also shifts the distribution measurably.

Reference model. Our zero-shot detection uses GPT-2-LARGE as the reference model. A more modern reference model might produce different detection patterns, though GPT-2 remains the standard in the literature [Mitchell et al., 2023, Bao et al., 2024].

Decoding strategy confound. The RAID dataset includes both greedy and sampling decoding. We do not separate these, though both base and aligned variants use the same decoding mix, so this is balanced across conditions.

6 Conclusion

We have presented empirical evidence that alignment training creates an implicit watermark in LLM outputs. Across three model families, three detection paradigms, and 12 independent comparisons, aligned models are consistently and significantly more detectable than their base counterparts (sign test $p = 0.006$). The watermark manifests as increased structural regularity—reduced sentence-length variability and higher token-level predictability—and is detectable by both statistical methods (AUROC +0.09 to +0.35) and by other LLMs (TPR 97.5–100% on aligned text vs. 85–96.3% on base text).

These findings have practical implications. If alignment is the watermark, then explicit watermarking may be unnecessary for aligned models, which already carry a durable detection signal. Detection research should focus on exploiting the alignment watermark rather than imposing external marks. Policymakers should recognize that the very process of making AI helpful provides a natural accountability mechanism.

The fundamental insight is this: there is an inherent tension between making AI useful and making AI undetectable. A perfect base model approximates the human distribution and approaches undetectability. The moment we align that model to serve human preferences, we create a distributional shift that guarantees detectability. Alignment is not merely a target for detection—it *is* the detection signal.

Future work. Three directions are most promising. First, testing across additional domains (creative writing, conversation, code) and model families would establish the generality of the alignment watermark. Second, isolating the contributions of different alignment stages (SFT vs. RLHF vs. DPO) on the same base model would clarify which stage creates the strongest signal. Third, adversarial robustness testing—whether paraphrasing, back-translation, or prompt engineering can erase the alignment watermark while preserving helpfulness—would determine whether the watermark represents a fundamental trade-off or an erasable artifact.

References

- 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 *Proceedings of ICLR*, 2024.
- Souradip Chakraborty, Amrit Singh Bedi, Sicheng Zhu, Bang An, Dinesh Manocha, and Furong Huang. On the possibilities of AI-generated text detection. *arXiv preprint arXiv:2304.04736*, 2023.
- Sumanth Dathathri, Abigail See, Sumedh Ghaisas, Po-Sen Huang, Rob McAdam, Johannes Welbl, Vandana Bachani, Alex Krawczuk, Daniel Mayber, Nishant Puri, et al. Scalable watermarking for identifying large language model outputs. *Nature*, 2024.
- Liam Dugan, Alyssa Hwang, Filip Trhlík, Josh Magnus Ludan, Daphne Ippolito, and Chris Callison-Burch. RAID: A shared benchmark for robust evaluation of machine-generated text detectors. In *Proceedings of ACL*, 2024.
- Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. GLTR: Statistical detection and visualization of generated text. In *Proceedings of ACL*, 2019.

- 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.
- Xinlei He, Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. MGTBench: Benchmarking machine-generated text detection. In *Proceedings of ACM CCS*, 2024.
- John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *Proceedings of ICML*, 2023.
- Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of RLHF on LLM generalisation and diversity. In *Proceedings of ICLR*, 2024.
- Jaeyoung Lee, Junhyuk Tack, and Jinwoo Shin. ReMoDetect: Reward models recognize aligned LLM’s generations. *arXiv preprint*, 2024.
- Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Kong, Shuming Quan, Liang Shi, Yaqian Zhang, et al. MAGE: Machine-generated text detection in the wild. In *Proceedings of ACL*, 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 ICML*, 2023.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Proceedings of NeurIPS*, 2022.
- Boxi Xu and Arkaitz Zubiaga. Understanding the effects of RLHF on the quality and detectability of LLM-generated texts. *arXiv preprint arXiv:2503.17965*, 2025.

A Reproducibility

Data. We use the RAID benchmark [Dugan et al., 2024], publicly available on HuggingFace (liamdugan/raid). We extract 6,000 clean samples (500 per model variant \times 6 AI models + 500 human texts) from the scientific abstracts domain with no adversarial attacks applied.

Computational resources. Experiments were run on a machine with $4\times$ NVIDIA RTX A6000 GPUs (48GB each). Experiment 2 (zero-shot statistical detection) used a single GPU for GPT-2-LARGE inference. Experiment 3 (LLM-as-detector) used the OpenAI API for GPT-4.1 classification (560 API calls, \sim \$2 total cost). Total execution time was approximately 15 minutes.

Software. Python 3.12.8, PyTorch 2.10.0, Transformers 5.1.0, scikit-learn 1.8.0, SciPy 1.17.0, OpenAI API 2.20.0. Random seed was fixed at 42 for all experiments.

Hyperparameters. Table 7 lists all hyperparameters. No hyperparameter tuning was performed; all values were set before running experiments.

Parameter	Value	Rationale
Random seed	42	Reproducibility
Max tokens (Exp. 2)	512	Balance coverage vs. speed
Reference model (Exp. 2)	GPT-2-LARGE	Standard detection reference
LLM detector (Exp. 3)	GPT-4.1	State-of-the-art reasoning model
Samples per category (Exp. 3)	80	Balance cost vs. statistical power
Temperature (Exp. 3)	0.0	Deterministic classification
Bootstrap iterations	1,000	Confidence interval estimation
Significance level	$\alpha = 0.05$	Bonferroni-corrected where applicable

Table 7: Hyperparameters used across all experiments.