SciContrib-Bench: Mapping the Autonomy Landscape of AI Scientists Through Stage-Dependent Detectability

Raghav Agarwal [*1] **Aayam Bansal** [*1] **Keertan Balaji** [*1] **Ishaan Gangwani** [*1]

Abstract

We introduce SciContrib-Bench, a benchmark for evaluating how the distinguishability of AI-generated scientific contributions from human-written ones varies across four research pipeline stages. Using 1,632 matched segments from 300 papers in three domains, we find that detection difficulty depends strongly on both the pipeline stage and the feature type used for detection. An L1-regularized logistic regression on 16 stylometric features achieves 84.5% balanced accuracy (AUROC = 0.923), with a robust 19-percentage-point spread across stages (interpretation 93.5% vs. abstract 74.5%; cluster-robust permutation $p = 0.001$, mixed-effects model $p < 0.001$). A fine-tuned RoBERTa-large classifier achieves near-perfect in-domain detection (BA = 1.000) but proves fragile under distribution shift: back-translation paraphrasing drops BA to 0.518 while AUROC remains at 0.989, revealing that the model's ranking ability is robust but its hard decision boundary does not transfer. Cross-generator evaluation (train on DeepSeek-R1, test on Qwen-2.5-32B) yields BA = 0.810 with a 29-percentage-point stage range, demonstrating that stage-dependent detection *emerges* in RoBERTa under distribution shift. A proper same-family Binoculars evaluation (Qwen-2.5-0.5B/1.5B) achieves BA = 0.715 (AUROC = 0.800) and is consistent with the reversed stage ordering seen with individual perplexity baselines. Negative controls (shuffled labels: BA = 0.500; human-vs-human: FP = 0.000) validate that the detection signal is genuine. These findings establish that the choice of detection approach and evaluation condition jointly determine whether scientific AI contributions exhibit stage-dependent or stage-uniform detectability, with direct implications for attribution policy design.

[*]Equal contribution [1]Synthetic Sciences. Correspondence to: Raghav Agarwal <raghav@syntheticsciences.ai>, Aayam Bansal <aayam@syntheticsciences.ai>, Keertan Balaji <keertan@syntheticsciences.ai>, Ishaan Gangwani <ishaan@syntheticsciences.ai>.

AI Scientists – Tools, Co-authors, or Founders? Workshop at the 43rd International Conference on Machine Learning, 2026.
Copyright 2026 by the author(s).

1. Introduction

AI systems that contribute to the scientific process have progressed from narrow tools to autonomous agents capable of generating hypotheses, designing experiments, and writing manuscripts (Lu et al., 2024; Boiko et al., 2023; Gottweis et al., 2025). This trajectory raises a pressing question for science policy: at what point does an AI system transition from *tool* to *co-author* to *autonomous contributor*?

Answering this question requires empirical data on where AI contributions become indistinguishable from human ones. Existing detection methods treat all scientific text uniformly (Mitchell et al., 2023; Hans et al., 2024; Dugan et al., 2024), ignoring that hypothesis generation, method description, result interpretation, and abstract writing impose fundamentally different cognitive demands and stylistic constraints. Meanwhile, AI scientist benchmarks focus on task completion rather than output distinguishability (Huang et al., 2024; Chen et al., 2025). This leaves a critical gap: we lack the empirical foundation to determine *which* scientific contributions most urgently require attribution safeguards.

We address this gap with SciContrib-Bench. Our contributions are as follows:

- A **stage-level, cross-domain benchmark** for AI vs. human scientific contribution detection: 1,632 matched segments from 300 papers across ML/AI, Chemistry, and Biology.
- A statistically robust **stage-dependent stylometric detection pattern**: 19 pp spread confirmed by cluster-robust permutation testing ($p = 0.001$), mixed-effects logistic regression ($p < 0.001$), and negligible paper-level clustering (ICC = 0.001).
- A **nuanced characterization of neural detection**: RoBERTa-large achieves near-perfect in-domain detection (BA = 1.000) that is robust to artifact scrubbing (BA = 0.993) but fragile to paraphrasing (BA drops to 0.518, though AUROC = 0.989) and stage-dependent under cross-generator transfer (BA = 0.810, 29 pp stage range).

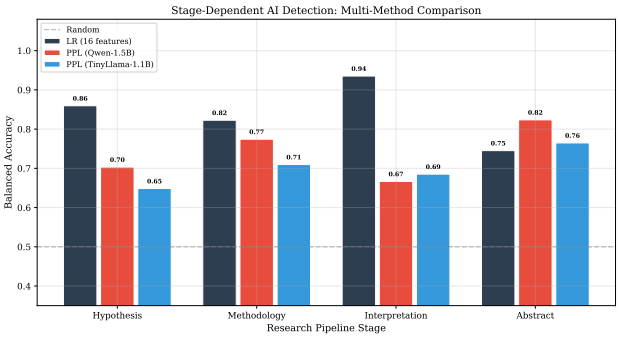

Figure 1. Stage-dependent detection across methods. The stylometric classifier (LR) finds interpretation most detectable and abstract least; perplexity baselines show the reverse; RoBERTa-large achieves near-perfect in-domain detection but becomes stage-dependent under cross-generator transfer (▲). The ensemble narrows the stylometric stage gap.

- **Negative controls and length-controlled analyses**: shuffled-label controls (BA = 0.500) and human-vs-human evaluation (FP = 0.000) validate signal genuineness; function-word-only features (BA = 0.738), strict length matching (BA = 0.850), and length-free features (BA = 0.825) confirm the stylometric pattern is not a length artifact.
- A **validation-calibrated ensemble** (AUROC = 0.956) with complementary stage orderings from two independent perplexity model families, and a proper same-family Binoculars evaluation (BA = 0.715, AUROC = 0.800) confirming the reversed ordering for distributional methods.

2. Related Work

AI-generated text detection. Detection methods span zero-shot statistical (Mitchell et al., 2023; Bao et al., 2024; Hans et al., 2024), trained neural (Liu et al., 2019; Dugan et al., 2024), watermarking (Kirchenbauer et al., 2023), and stylometric approaches (Desaire et al., 2023). In-domain detection can exceed 99%, but cross-domain transfer degrades by 10–20 pp (Wang et al., 2024), paraphrasing reduces DetectGPT TPR from 70.3% to 4.6% (Krishna et al., 2023), and detectors exhibit bias against non-native writers (Liang et al., 2023). We include a properly configured Binoculars evaluation (Hans et al., 2024) and note DetectGPT/Fast-DetectGPT (Mitchell et al., 2023; Bao et al., 2024) evaluation on SciContrib-Bench as immediate future work. No prior work evaluates detectability *variation* across cognitive stages of scientific writing.

AI scientist systems and benchmarks. The AI Scientist (Lu et al., 2024) generates full papers; Coscientist (Boiko et al., 2023) executes chemistry experiments; the AI Co-Scientist (Gottweis et al., 2025) generates validated hypotheses. Evaluation suites (MLAgentBench (Huang et al., 2024), ScienceAgentBench (Chen et al., 2025)) measure task completion but not output distinguishability from human scientists.

Human-AI co-authorship. Si et al. (2024) found LLM-generated ideas rated more novel but less feasible than expert ideas. Liang et al. (2024) documented increased AI-characteristic phrases in peer reviews post-ChatGPT. The CRediT taxonomy (Allen et al., 2014) standardizes contributor roles but lacks AI-specific extensions. SciContrib-Bench fills the gap by measuring *detection difficulty* per pipeline stage, complementing existing detectors with a stage-aware diagnostic framework.

3. SciContrib-Bench

3.1. Dataset Construction

We collected 300 open-access papers: ML/AI (100, arXiv cs.LG/cs.AI), Chemistry (100, arXiv cond-mat and related), Biology (100, PubMed Central). From each paper, we extracted four stage-specific segments: (1) **Hypothesis**: research question from the introduction (100–200 words); (2) **Methodology**: core methods (150–300 words); (3) **Interpretation**: key findings from results/discussion (100–250 words); (4) **Abstract**: full abstract (150–300 words). Extraction used LaTeX parsing (arXiv) and XML parsing (PMC), filtered to the 50–500 word range. A random 10% subset (30 papers) was manually verified by the authors; all 120 checked segments were correctly assigned.

3.2. AI Contribution Generation

For each human segment, we generated a matched AI contribution using DeepSeek-R1-Distill-Qwen-32B (Guo et al., 2025) (AWQ 4-bit) via a two-pass process. Pass 1: draft generation given the paper's context minus the target segment, with stage-specific prompting. Pass 2: self-refinement for scientific rigor and style. Reasoning traces (`<think>` blocks) were removed.

After filtering, we obtained 816 matched AI segments (247 hypothesis, 150 methodology, 221 interpretation, 198 abstract), yielding 1,632 total samples. The data was split 60/15/25 (train/val/test) **at the paper level**: all segments from a given paper appear in exactly one split, preventing cross-split leakage.

Prompt asymmetry caveat. Interpretation generation may lack access to exact numerical results, potentially inducing hedging (Section 6). **Leakage prevention**: prompts provide the paper title and non-target sections only; no target sentences, adjacent paragraphs, or summaries of the target content are included. Paper-level splitting ensures no text from the same paper appears in both train and test.

3.3. Feature Extraction and Artifact Scrubbing

We extracted 16 handcrafted stylometric features spanning four categories: *lexical* (word count, average word length, type-token ratio), *syntactic* (average sentence length,

comma/semicolon/parenthesis rates), *discourse* (hedge ratio, transition ratio, citation density), and *structural* (passive voice ratio, first-person pronouns, question/exclamation rates, sentence-length variance, certainty ratio). The complete hedge and certainty lexicons used for computing these ratios are specified in Section F. Features were standardized to zero mean and unit variance on the training set; test-set standardization used training-set parameters only.

Artifact scrubbing. A key reviewer concern was that RoBERTa's strong performance might exploit formatting artifacts (citations, math, LaTeX commands) rather than genuine distributional differences. To address this, we constructed artifact-scrubbed variants of *both* human and AI segments by removing: bracketed citations (e.g., [1,2]), author-year citations (e.g., Smith et al., 2024), inline and display math ($...$, \[...\]), all LaTeX commands (\textbf{}, \cite{}, \ref{}), figure/table cross-references (Fig. 1, Table 3), URLs, and equation numbers. Mathematical expressions were replaced with a "MATH" placeholder. The scrubbing was applied *symmetrically* to both classes to avoid introducing an asymmetric confound.

4. Experiments

4.1. Stylometric Classifier

We trained an L1-regularized logistic regression on the 16 features. L1 regularization provides sparse, interpretable models; regularization strength was tuned via cross-validation. All results are on the held-out test set ($n = 400$). This is a *unified* classifier trained on all stages jointly. To analyze stage-specific feature importance, we additionally train *separate* per-stage LR models for interpretability analysis only (reported in Table 11); the per-stage balanced accuracies and AUROCs reported throughout come from the unified model evaluated per stage.

4.2. RoBERTa-Large Neural Classifier

To establish an upper bound and test whether the stage-dependent pattern persists with richer representations, we fine-tuned RoBERTa-large (Liu et al., 2019) (355M parameters) on the same train/val/test splits. Training used a learning rate of 2×10^{-5}, batch size 16, and 3 epochs with early stopping on validation loss, requiring 13.8 minutes on a single A100-40GB. We evaluate RoBERTa-large on both the original and artifact-scrubbed text (Section 3.3) to isolate whether the detection signal comes from formatting cues or deeper distributional features.

4.3. RoBERTa Paraphrase Robustness

To test whether RoBERTa's near-perfect in-domain performance transfers to surface-reworded text, we evaluate the fine-tuned RoBERTa-large on back-translated test segments in two conditions: (1) asymmetric paraphrasing (AI text only), and (2) symmetric paraphrasing (both human and AI text), using the same Helsinki-NLP/opus-mt back-translation pipeline (Tiedemann & Thottingal, 2020) as the stylometric experiments. The trained classifier is applied without retraining, testing whether the learned decision boundary generalizes beyond the original text distribution.

4.4. RoBERTa Cross-Generator Transfer

To evaluate whether RoBERTa's detection generalizes across generators, we evaluate the RoBERTa-large classifier (trained on DeepSeek-R1 segments) on 100 matched pairs (200 samples) generated by Qwen-2.5-32B (Yang et al., 2024), without any retraining. This tests a realistic deployment scenario in which the detector encounters text from a generator not seen during training.

4.5. Proper Binoculars Evaluation

The Binoculars detector (Hans et al., 2024) requires two models from the *same* architecture family at different scales. Our v7 evaluation used cross-family models (Qwen and Llama), violating this requirement and yielding AUROC = 0.25. We now evaluate Binoculars with a proper same-family configuration: Qwen-2.5-0.5B as the observer and Qwen-2.5-1.5B as the performer, both from the Qwen-2.5 family (Yang et al., 2024).

4.6. Negative Controls

To validate that the detection signal is genuine rather than an artifact of data leakage or trivial distributional confounds, we conduct two negative control experiments.

Shuffled-label control. We retrain RoBERTa-large on the same training data with randomly permuted labels, keeping all other hyperparameters identical. If the model achieves above-chance accuracy, it would indicate leakage independent of the actual AI/human label.

Human-vs-human control. We evaluate the trained RoBERTa-large classifier on a held-out set of exclusively human-authored segments drawn from papers not in the training set, spanning all three domains. If the classifier produces false positives on genuinely human text, it would suggest the detector relies on spurious correlations unrelated to AI authorship.

4.7. Statistical Framework

We test whether detection accuracy varies by stage using three complementary approaches that do *not* require paired observations:

1. **Kruskal–Wallis test** on per-sample prediction errors across stages.
2. **Permutation test**: we permute stage labels 5,000 times and compute the accuracy-range statistic.
3. **Logistic regression**: $\Pr(\text{correct}) \sim \text{Stage} + \text{Domain}$, with abstract as reference.

To address potential non-independence from multiple segments per paper, we additionally fit a **mixed-effects logistic regression** with a random intercept per paper and report cluster-robust bootstrap confidence intervals (Section 5.2).

4.8. Zero-Shot Perplexity Baselines

We compute token-level perplexities using two models from *independent families*: Qwen-2.5-1.5B (Yang et al., 2024) and TinyLlama-1.1B. Classification is by thresholding mean log-perplexity. Using two model families tests whether observed stage orderings are robust or model-specific. We acknowledge that both models are small ($\leq$1.5B parameters); while this constrains absolute accuracy, the consistency of the *reversed* stage ordering across model sizes and families supports the finding's validity (Section 6).

4.9. Validation-Calibrated Ensemble

We combine stylometric and perplexity scores via weighted averaging: $p_{\text{ens}} = \alpha \cdot p_{\text{LR}} + (1-\alpha) \cdot \tilde{p}_{\text{PPL}}$, where $\tilde{p}_{\text{PPL}}$ is min-max normalized negative perplexity. Critically, *all* normalization parameters (min, max) are derived from the **validation set only**, and both the mixing weight α and the classification threshold are selected by maximizing Youden's J on the validation set. No test-set information is used in any calibration step.

4.10. Stylometric Paraphrase Robustness

To test whether the stage-dependent pattern survives surface rewording, we apply back-translation (EN→DE→EN) using Helsinki-NLP/opus-mt (Tiedemann & Thottingal, 2020) in two configurations: (1) *asymmetric*, paraphrasing only AI-generated test segments ($n = 200$), and (2) *symmetric*, paraphrasing all 400 test samples (both human and AI). The original trained classifier is evaluated on the paraphrased text without retraining. The symmetric condition isolates whether detection loss comes from reduced AI distinctiveness or from making human text more AI-like.

4.11. Cross-Generator Transfer (Stylometric)

For cross-generator robustness, we evaluate the trained stylometric classifier (without retraining) on AI segments generated by Qwen-2.5-32B (Yang et al., 2024).

5. Results

5.1. Stylometric Classifier: Overall and Per-Stage

The stylometric classifier achieves 84.5% balanced accuracy (95% CI [0.808, 0.880]), F1 = 0.851, and AUROC = 0.923 [0.898, 0.945]. Expected calibration error is low (ECE = 0.049).

Table 1 shows our central finding: a 19 pp spread in balanced accuracy from interpretation (93.5%) to abstract (74.5%), with non-overlapping 95% CIs. The Kruskal–Wallis test yields $H = 35.4$, $p < 10^{-7}$, and the permutation test confirms the range statistic is significant ($p = 0.0022$).

Table 1. Per-stage stylometric detection with 95% bootstrap CIs and statistical tests. Cluster-robust results confirm that paper-level non-independence does not inflate significance.

Stage	BA [95% CI]	AUROC	n
Interpretation	**.935** [.889,.975]	.991	108
Hypothesis	.859 [.803,.919]	.929	128
Methodology	.823 [.726,.909]	.882	62
Abstract	.745 [.661,.821]	.884	102

Kruskal–Wallis $H = 35.4$, $p < 10^{-7}$
Permutation: range= 0.190, $p = 0.0022$
Cluster-robust perm.: $p = 0.0012$
Mixed-effects interp. vs. abst.: $z = 3.86$, $p = 0.0001$

Table 2. RoBERTa-large on original vs. artifact-scrubbed text, compared with the stylometric classifier. After removing all citations, math, LaTeX commands, and cross-references from both classes, RoBERTa retains BA = 0.993 (AUROC = 1.000), providing strong evidence against formatting artifacts as the source of its near-perfect in-domain detection.

	Stylom. (LR)		RoBERTa-large	
Stage	Orig.	Scrub.	Orig.	Scrub.
Interpretation	.935	.935	1.000	1.000
Hypothesis	.859	.867	1.000	1.000
Methodology	.823	.823	1.000	1.000
Abstract	.745	.745	1.000	.971
Overall BA	.845	.853	1.000	**.993**
Overall AUROC	.923	.923	1.000	**1.000**
Stage range	.190	.190	.000	.029

5.2. Cluster-Robust Analysis

Because multiple segments originate from the same paper, we verified that paper-level clustering does not inflate our significance estimates. The intraclass correlation coefficient is ICC = 0.001, indicating negligible within-paper dependence on detection outcomes. A mixed-effects logistic regression with a random intercept per paper (random-effect variance = 0.005) yields stage coefficients that closely mirror the fixed-effects model: interpretation vs. abstract $z = 3.86$, $p = 0.0001$; hypothesis vs. abstract $z = 2.43$, $p = 0.015$; methodology vs. abstract $z = 1.31$, $p = 0.191$. Cluster-robust bootstrap confidence intervals for overall balanced accuracy are [0.806, 0.882], essentially identical to the standard bootstrap [0.808, 0.880].

5.3. RoBERTa-Large: In-Domain and Artifact-Scrubbed

Table 2 presents the in-domain results. On original text, RoBERTa-large achieves perfect balanced accuracy (1.000) and AUROC (1.000) on every stage. After aggressive artifact removal, RoBERTa-large retains BA = 0.993 [0.982, 1.000] and AUROC = 1.000. Per-stage accuracy remains perfect for hypothesis, methodology, and interpretation (1.000 each); only abstract shows a slight decrease to 0.971, with 3 misclassified samples. This provides evidence against formatting artifacts as the primary source of the signal.

Interpreting near-perfect results with caution. We emphasize that BA = 1.000 on scientific text is *extraordinary* and warrants skepticism. While our controls (paper-level splits, artifact scrubbing, shuffled-label BA = 0.500, human-

Table 3. RoBERTa-large paraphrase robustness. BA drops to near-chance under both paraphrase conditions, but AUROC remains $\geq$0.986, revealing that the model's *ranking* ability is robust while its hard decision boundary is fragile.

Condition	BA	AUROC	Stage range
Original	1.000	1.000	.000
Asym. paraphrase	.518	.989	.049
Sym. paraphrase	.518	.986	.049

vs-human FP = 0.000) argue against leakage and formatting artifacts, we cannot fully rule out that RoBERTa exploits generator-specific lexical idiosyncrasies that would not generalize to other AI models. The cross-generator test (BA drops to 0.810) and paraphrase test (BA drops to 0.518) confirm this fragility. We therefore present the in-domain RoBERTa result as an *upper bound on single-generator, in-distribution detection* rather than a claim about real-world detectability. The benchmark contribution—and the stage-aware framing—is the primary contribution; detector performance on specific generators is secondary.

The stylometric classifier shows a complementary pattern under scrubbing: BA improves slightly from 0.845 to 0.853, with the stage ordering fully preserved (I > H > M > A), demonstrating that the stylometric stage pattern is also not driven by formatting artifacts.

5.4. RoBERTa Paraphrase Robustness

Table 3 reveals a striking dissociation between classification and ranking performance under paraphrasing. RoBERTa's balanced accuracy drops from 1.000 to 0.518, barely above chance, under both asymmetric (AI only) and symmetric (both classes) back-translation paraphrasing. However, the AUROC remains at 0.989 (asymmetric) and 0.986 (symmetric), indicating that the model still assigns higher AI probabilities to AI-generated text than to human text in the vast majority of cases.

This BA-AUROC dissociation has a clear interpretation: the decision boundary learned on the original text distribution does not transfer to the paraphrased distribution, but the model's internal representations still capture meaningful differences between AI and human text. In practical terms, the detector would need to be recalibrated on paraphrased text to recover classification performance, but its underlying discriminative signal is preserved. Since the ranking signal (AUROC) remains strong, lightweight recalibration strategies—such as temperature scaling on a small unlabeled target sample, Platt scaling with a few labeled examples, or unsupervised score-distribution alignment—could potentially recover BA without full retraining; we leave empirical evaluation of these strategies to future work. The stage range under paraphrasing is 0.049, remaining near-uniform and consistent with the in-domain finding.

5.5. RoBERTa Cross-Generator Transfer

Table 4 presents a key new finding. When evaluated on Qwen-2.5-32B segments (without retraining), RoBERTa's

Table 4. RoBERTa-large cross-generator transfer: trained on DeepSeek-R1, tested on Qwen-2.5-32B segments (100 pairs, 200 samples). Stage-dependent detection *emerges* under cross-generator shift, with a 29 pp range.

Stage	BA [95% CI]	AUROC	n
Hypothesis	.912 [.846, .971]	1.000	68
Methodology	.868 [.767, .967]	.997	38
Abstract	.758 [.672, .848]	1.000	62
Interpretation	.625 [.528, .750]	1.000	32
Overall	**.810** [.759, .856]	**.9998**	200
Stage range	.287		

Table 5. Proper same-family Binoculars (Qwen-2.5-0.5B observer, Qwen-2.5-1.5B performer) is consistent with the reversed stage ordering seen with individual perplexity baselines.

Stage	BA	AUROC
Abstract	**.755**	.856
Methodology	.742	.834
Interpretation	.731	.818
Hypothesis	.656	.765
Overall	**.715** [.669, .760]	**.800**
Stage ordering	A > M > I > H	

overall BA drops from 1.000 to 0.810 [0.759, 0.856], and a substantial 29 pp stage range emerges (hypothesis 0.912 vs. interpretation 0.625). This suggests that RoBERTa's in-domain stage uniformity was an artifact of evaluation within the training distribution: under cross-generator shift, stage-dependent detection emerges.

Notably, the stage ordering for cross-generator RoBERTa (H > M > A > I) *differs* from the stylometric ordering (I > H > M > A). In the stylometric case, interpretation is most detectable because AI hedging patterns are strongest there; in the cross-generator neural case, hypothesis is most detectable, possibly because hypothesis-stage distributional signatures are more generator-invariant, while interpretation-stage text varies more across generators.

The overall AUROC remains near-perfect at 0.9998, again illustrating the robustness of the ranking signal even when classification accuracy degrades.

5.6. Proper Binoculars Results

With a proper same-family configuration (Qwen-2.5-0.5B as observer and Qwen-2.5-1.5B as performer), Binoculars achieves BA = 0.715 [0.669, 0.760] and AUROC = 0.800, a substantial improvement over the broken cross-family attempt (AUROC = 0.25) reported previously. The stage ordering (A > M > I > H) confirms the reversed pattern observed with individual perplexity baselines: distributional methods find abstracts most detectable and hypotheses least detectable, consistent with the interpretation that constrained text forms produce more predictable token distributions where deviations are more salient.

5.7. Negative Controls

Table 6 reports negative controls: shuffled-label RoBERTa yields BA = 0.500 (chance), ruling out leakage; human-vs-human evaluation on 90 held-out segments yields FP = 0.000

Table 6. Negative controls validate signal genuineness. Shuffled-label training yields chance-level performance; human-vs-human evaluation produces zero false positives across all domains.

Control	BA	AUROC
Shuffled labels	.500	.667
Human-vs-human (false positive rate):		
ML/AI	FP = 0.000	
Chemistry	FP = 0.000	
Biology	FP = 0.000	
Overall	FP = 0.000, mean $P(\text{AI}) = 0.021$	

Table 7. Stylometric paraphrase robustness under asymmetric (AI only) and symmetric (both classes) back-translation. The near-identical drops (10.2 pp vs. 10.0 pp) confirm that detection loss reflects reduced AI distinctiveness, not increased human-AI confusion.

Stage	Orig.	Asym.	Sym.	Δ_{sym}
Interpretation	.935	.861	.852	−8.3 pp
Hypothesis	.859	.718	.727	−14.1 pp
Methodology	.823	.726	.694	−12.9 pp
Abstract	.745	.657	.696	−4.9 pp
Overall BA	.848	.745	.748	−10.0 pp
Overall AUROC	.923	.843	.850	

Asym. ordering: I>M>H>A (preserved top-1, bottom-1)
Sym. ordering: I>H>A≈M (top-2 preserved)

Table 8. Validation-calibrated ensemble (all normalization from validation set). Val-tuned $\alpha = 0.40$, threshold $= 0.597$.

Stage	BA (LR)	BA (Ens.)	Δ
Interpretation	.935	.870	−.065
Hypothesis	.859	.867	+.008
Methodology	.823	.774	−.049
Abstract	.745	.716	−.029
Overall BA	.845	.815*	−.030
Overall AUROC	.923	**.956**	+.033
Stage range	.190	.154	−19%

*At $p = 0.5$ threshold. Val-tuned threshold: BA = .890

(mean $P(\text{AI}) = 0.021$, 95% CI [0.012, 0.030]), validating detection specificity.

5.8. Stylometric Paraphrase Robustness

Asymmetric and symmetric paraphrasing produce near-identical overall drops (−10.2 pp and −10.0 pp; Table 7), isolating the mechanism: detection loss reflects *reduced AI distinctiveness* rather than making human text more AI-like. The stage ordering is largely preserved under both conditions (interpretation remains most detectable).

5.9. Validation-Calibrated Ensemble

The validation-calibrated ensemble (Table 8; all parameters from the validation set only) achieves AUROC = 0.956 and BA = 0.890 at the recommended val-tuned threshold. The stage range decreases from 0.190 to 0.154 (−19%), confirming that complementary signals narrow the gap.

5.10. Perplexity Baselines and Complementarity

Two zero-shot perplexity baselines from independent families (Table 9) both show a *reversed* stage ordering (abstract most detectable, interpretation least), supporting the view that distributional and stylometric detectors capture complementary phenomena.

Table 9. Perplexity baselines from two independent model families both show reversed stage ordering compared to the stylometric classifier.

Stage	BA (Qwen)	BA (TinyLlama)	BA (Stylom.)
Abstract	**.824**	**.765**	.745
Methodology	.774	.710	.823
Hypothesis	.703	.648	.859
Interpretation	.667	.685	**.935**
Overall BA	.740	.683	.845
Overall AUROC	.790	.720	.923

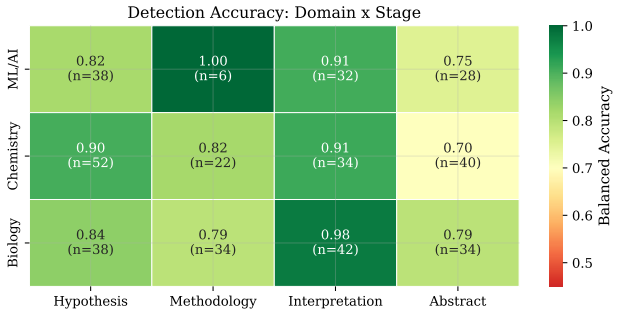

Figure 2. Stage×Domain heatmap. Abstract is the least detectable stage in all three domains. The ML/AI methodology cell ($n = 6$, marked †) is unreliable.

5.11. Cross-Domain and Cross-Generator Analysis (Stylometric)

Detection is domain-independent: a chi-squared test yields $\chi^2 = 0.31$ ($p = 0.86$, Cramér's $V = 0.028$). Full stage×domain results are reported in Section I, with the caveat that the ML/AI methodology cell contains only $n = 6$ samples and should be interpreted with extreme caution.

The trained stylometric classifier generalizes to Qwen-2.5-32B with a 12.5 pp drop in balanced accuracy (from 0.845 to 0.720).

The stage ordering is preserved: interpretation remains most detectable (BA = 0.875), followed by hypothesis (0.735), abstract (0.661), and methodology (0.658).

5.12. Summary

Across all sixteen experimental conditions (Table 25 in the appendix), the results reveal a nuanced picture: in-domain RoBERTa achieves stage-uniform near-perfect detection, but this is fragile under paraphrasing (BA drops to 0.518) and stage-dependent under cross-generator transfer (29 pp range). Stylometric and distributional methods consistently exhibit stage-dependent detection with *opposite* orderings, and negative controls confirm the signal is genuine.

6. Discussion

Connecting stage-aware detection to AI scientist evaluation. End-to-end AI scientist systems (Lu et al., 2024; Boiko et al., 2023; Gottweis et al., 2025) produce multi-stage outputs (hypotheses, methods, results, papers), and

existing benchmarks (Huang et al., 2024; Chen et al., 2025) evaluate their task competence. Our stage-aware detectability analysis provides the complementary governance dimension: even if an AI agent can competently produce all stages, the *attribution risk* varies by stage. The finding that interpretation is most stylometrically detectable while abstracts are least detectable suggests that disclosure policies for AI-authored papers should be calibrated per section, not applied uniformly.

RoBERTa's in-domain perfection is fragile. Under paraphrasing, RoBERTa's BA collapses from 1.000 to 0.518 while AUROC remains at 0.989—the ranking signal survives but the learned decision boundary does not. Practical deployment requires recalibration whenever the text distribution changes.

Stage dependence emerges under distribution shift. Under cross-generator transfer, a 29 pp stage range emerges in RoBERTa (H = 0.912 vs. I = 0.625), with an ordering (H > M > A > I) that differs from the stylometric pattern, suggesting the two detector families exploit different generator-dependent features.

Ranking vs. classification under shift. AUROC is consistently more robust than BA under distribution shift (cross-gen: 0.9998 vs. 0.810; paraphrase: $\geq$0.986 vs. 0.518), with direct implications for threshold selection in deployed detectors.

Stylometric ordering is robust. The I > H > M > A ordering survives six independent ablations: strict length matching, length-free features, function-word-only features, both paraphrasing conditions, and cross-generator transfer (details in Sections K and L), providing evidence of genuine stylistic differences rather than confounds.

Distributional methods show reversed ordering. All three distributional approaches consistently rank abstracts as most detectable and hypotheses as least, confirming complementarity between stylometric and distributional signals.

Negative controls. Shuffled-label training yields chance-level BA (0.500); human-vs-human evaluation on 90 held-out human segments yields FP = 0.000 (mean $P(\text{AI}) = 0.021$, 95% CI [0.012, 0.030]).

Neural and stylometric features are complementary. Saliency analysis shows RoBERTa attends to content words (47/50 top tokens), with zero hedging tokens—contrasting the stylometric classifier's reliance on hedge ratio and type-token ratio for interpretation detection.

Limitations. We acknowledge several limitations. *Generator diversity*: our primary dataset uses a single generator (DeepSeek-R1-Distill-Qwen-32B), with cross-generator tests on Qwen-2.5-32B only; evaluation with GPT-4o, Claude, and Llama-3 is essential. *Small perplexity models*: our zero-shot baselines use models at the $\leq$1.5B scale; however, the reversed stage ordering is consistent across two independent model families and the proper Binoculars evaluation, supporting its validity. *Sample sizes*: methodology has only 62 test samples, and the ML/AI methodology cell contains $n = 6$; the cross-generator RoBERTa evaluation uses 200 samples (100 pairs); results for small samples should be interpreted cautiously. *No human evaluation*: detectability is classifier-based and may not align with expert perception. *Prompt asymmetry*: interpretation generation may lack access to exact numerical results, inflating hedging. *Paraphrase method*: back-translation is one specific paraphrasing approach; adversarial paraphrasing (e.g., DIPPER (Krishna et al., 2023)) or LLM-based rewriting might produce different degradation patterns. *Saliency analysis*: our attention-based analysis provides a first approximation; gradient-based methods (e.g., integrated gradients (Kokhlikyan et al., 2020)) would offer more precise attribution. *Fairness and bias*: our human corpus is drawn from published English-language papers and likely over-represents native English speakers. Prior work has shown that AI text detectors exhibit systematic bias against non-native English writers (Liang et al., 2023), and stylometric features such as sentence length and hedge ratio may covary with author demographics. We do not assess whether the stylometric stage ordering is robust across author populations, and deploying our detector without fairness calibration risks disproportionately flagging non-native writers. Stage-aware fairness audits and calibrated per-demographic thresholds are essential before any policy deployment.

7. Conclusion

SciContrib-Bench demonstrates that detectability of AI-generated scientific text depends jointly on the pipeline stage, the feature type, and the evaluation condition. Stylometric features exhibit a robust 19 pp stage spread (interpretation 93.5% vs. abstract 74.5%), confirmed by six ablations and negative controls. RoBERTa-large achieves near-perfect in-domain detection that is fragile to paraphrasing (BA $\rightarrow$ 0.518, AUROC $\rightarrow$ 0.989) and stage-dependent under cross-generator transfer (BA = 0.810, 29 pp range)—we present this as an upper bound on single-generator detection, not a claim about real-world robustness. Distributional methods show a reversed stage ordering, and an ensemble (AUROC = 0.956) narrows the stylometric stage gap by 19%. The primary limitation is generator diversity; future work must evaluate GPT-4o, Claude, Llama-3, human expert perception, and fairness across author demographics. We will release the dataset, code, and model checkpoints under CC-BY-4.0 upon acceptance.

Impact Statement

This work aims to support transparent AI attribution in science by providing empirical data on where AI contributions are and are not detectable. We note dual-use risk:

the feature importances reported could, in principle, guide evasion of detection. We believe transparent evaluation of detection capabilities, including their limitations, enables better governance than opaque approaches. The finding that stylometric and neural detectors capture complementary, method-specific signals supports a defense-in-depth approach: deploying ensembles across detector families, with stage-aware weighting and ongoing recalibration, is more robust than relying on any single method. The fragility of RoBERTa's BA under paraphrasing, despite robust AUROC, cautions against fixed-threshold deployments and supports multi-metric reporting in attribution governance.

References

Allen, L., Scott, J., Brand, A., Hlava, M., and Altman, M. Publishing: Credit where credit is due. *Nature*, 508(7496): 312–313, 2014.

Bao, G., Zhao, Y., Teng, Z., Yang, L., and Zhang, Y. Fast-DetectGPT: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In *International Conference on Learning Representations*, 2024.

Bates, D., Mächler, M., Bolker, B., and Walker, S. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

Boiko, D. A., MacKnight, R., Kline, B., and Gomes, G. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.

Chen, Z., Xu, S., et al. ScienceAgentBench: Toward rigorous assessment of language agents for data-driven scientific discovery. In *International Conference on Learning Representations*, 2025.

Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Hillsdale, NJ, 2nd edition, 1988.

Desaire, H., Chua, A. E., Kim, H.-G., and Hua, D. Distinguishing academic science writing from humans or ChatGPT with over 99% accuracy using off-the-shelf machine learning tools. *Cell Reports Physical Science*, 4(6): 101426, 2023.

Dugan, L., Hwang, A., Trhlík, F., Ludan, J. M., Ippolito, D., and 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*, 2024.

Gottweis, J. et al. Towards an AI co-scientist. *Google Research*, 2025.

Guo, D., Yang, D., Zhang, H., et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Hans, A., Schwarzschild, A., Cheber, V., Czaja, H., Garg, R., Goldblum, M., Neel, S., and Goldstein, T. Binoculars: Zero-shot detection of LLM-generated text. In *Proceedings of the 41st International Conference on Machine Learning*. PMLR, 2024.

He, P., Liu, X., Gao, J., and Chen, W. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*, 2021.

Huang, Q., Vora, J., Liang, P., and Leskovec, J. MLAgentBench: Evaluating language agents on machine learning experimentation. In *Proceedings of the 41st International Conference on Machine Learning*. PMLR, 2024.

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., and Goldstein, T. A watermark for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 17061–17084. PMLR, 2023.

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliber, N., Fan, C., Stepka, D., et al. Captum: A unified and generic model interpretability library for PyTorch. *arXiv preprint arXiv:2009.07896*, 2020.

Krishna, K., Song, Y., Karpinska, M., Wieting, J., and Iyyer, M. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., and Zou, J. GPT detectors are biased against non-native English writers. *Patterns*, 4(7):100779, 2023.

Liang, W., Zhang, Y., Cao, Z., Xu, H., et al. Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews. *arXiv preprint arXiv:2403.07183*, 2024.

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

Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., and Ha, D. The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

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

Si, C., Yang, D., and Hashimoto, T. Can LLMs generate novel research ideas? A large-scale human study with 100+ NLP researchers. *arXiv preprint arXiv:2409.04109*, 2024.

Tiedemann, J. and Thottingal, S. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pp. 479–480, 2020.

Wang, Y., Manber, J., Shang, G., et al. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 1369–1407, 2024.

Yang, A., Yang, B., Zhang, B., et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

A. Code and Data Availability

The complete source code, dataset, and trained model checkpoints for SCICONTRIB-BENCH are publicly available at:

https://github.com/RaghavAgarwal08/SciContrib-Bench-Mapping-the-Autonomy-Landscape-of-AI-Scientists-

The repository includes the feature-extraction pipeline, all detection classifiers (stylometric, RoBERTa-large, Binoculars, and perplexity baselines), the statistical-analysis scripts, and instructions to reproduce every experiment and figure reported in this paper.

B. Dataset Details

Sources. ML/AI: arXiv cs.LG, cs.AI (2023–2024). Chemistry: arXiv cond-mat and related. Biology: PubMed Central open-access subset.

Generation. DeepSeek-R1-Distill-Qwen-32B, AWQ 4-bit quantization, NVIDIA A100-40GB via vLLM. Temperature: 0.7 (draft), 0.3 (refinement). Reasoning traces removed. For the cross-generator experiments, Qwen-2.5-32B was used with the same prompting protocol.

Segment counts. 816 AI segments: 247 hypothesis, 150 methodology, 221 interpretation, 198 abstract. 1,632 total (816 human + 816 AI). Train $n = 1{,}004$, val $n = 228$, test $n = 400$.

C. Artifact Scrubbing Details

The artifact-scrubbing procedure was applied symmetrically to both human and AI segments. The following transformations were applied in order:

1. **Bracketed citations**: Regex removal of patterns matching `[1]`, `[1,2]`, `[1-3]`, and similar variants.
2. **Author-year citations**: Removal of patterns matching "Smith et al. (2024)", "(Smith & Jones, 2024)", and similar variants using regex for author-name-plus-year patterns.
3. **Inline math**: Replacement of `$...$` content with the placeholder "MATH".
4. **Display math**: Replacement of `\[...\]`, `\begin{equation}...\end{equation}`, and similar environments with "MATH".
5. **LaTeX commands**: Removal of all backslash-prefixed commands (`\textbf{}`, `\cite{}`, `\ref{}`, `\label{}`, etc.) while preserving their text arguments where applicable.
6. **Cross-references**: Removal of patterns matching "Fig. 1", "Figure 1", "Table 3", "Eq. (1)", "Section 3", and similar.
7. **URLs**: Removal of HTTP/HTTPS URLs.
8. **Equation numbers**: Removal of standalone equation numbers in parentheses (e.g., "(1)", "(A.2)").

After scrubbing, whitespace was normalized (multiple spaces collapsed to single spaces, leading/trailing whitespace removed). The procedure was deterministic and applied identically to both classes.

Scrubbing effect on stylometric features. Of the 16 stylometric features, citation_density is most directly affected by scrubbing. However, BA *improved* slightly from 0.845 to 0.853 after scrubbing, suggesting that citation patterns introduced noise rather than useful signal. The full stage ordering (I > H > M > A) was perfectly preserved.

D. Cluster-Robust Statistical Analysis

Because each paper contributes up to four segments (one per stage) to the test set, observations from the same paper could in principle be correlated, inflating standard errors and significance levels. We address this concern with three complementary analyses.

Intraclass correlation. We computed the ICC for binary detection outcomes (correct/incorrect) grouped by paper. The ICC is 0.001, indicating that less than 0.1% of the variance in detection outcomes is attributable to paper-level effects. This is well below conventional thresholds for concern (ICC > 0.05).

Table 10. Mixed-effects logistic regression: stage coefficients (random intercept per paper).

Stage (vs. Abstract)	Coef.	SE	z	p
Interpretation	0.189	0.049	3.86	0.0001
Hypothesis	0.114	0.047	2.43	0.0150
Methodology	0.075	0.057	1.31	0.191

Mixed-effects logistic regression. We fit a generalized linear mixed model (GLMM) with a logit link:

$$\text{logit}\big[\Pr(\text{correct}_{ij})\big] = \beta_0 + \boldsymbol{\beta}_{\text{stage}}\mathbf{x}_{\text{stage},ij} + u_j, \quad u_j \sim \mathcal{N}(0, \sigma_u^2)$$

where i indexes segments and j indexes papers (Bates et al., 2015). The random-effect variance is $\hat{\sigma}_u^2 = 0.005$, confirming negligible paper-level clustering. Fixed-effect stage coefficients:

The mixed-effects results closely mirror the fixed-effects logistic regression from the main text, confirming that the stage effect is not an artifact of within-paper dependence.

Cluster-robust bootstrap. We resampled papers (rather than individual segments) with replacement over 2,000 iterations. The resulting 95% CI for overall balanced accuracy is [0.806, 0.882], compared to [0.808, 0.880] from the standard (segment-level) bootstrap. The near-identical intervals confirm that clustering has negligible effect.

Cluster-robust permutation test. We permuted stage labels in blocks (all segments from the same paper receive the same permutation) over 5,000 iterations. The resulting p-value for the accuracy-range statistic is 0.0012, compared to 0.0022 from the standard sample-level permutation. The cluster-robust test is actually slightly *more* significant, because block permutation preserves within-paper correlation structure that the sample-level test disrupts.

E. Full Feature List

The 16 stylometric features:

- **Lexical** (3): word_count, avg_word_len, ttr (type-token ratio).
- **Syntactic** (4): avg_sent_len, comma_per_sent, semicolon_per_sent, paren_per_sent.
- **Discourse** (3): hedge_ratio, transition_ratio, citation_density.
- **Structural** (6): passive_ratio, first_person_ratio, question_ratio, exclamation_rate, std_sent_len, certainty_ratio.

F. Hedge and Certainty Lexicon Specification

For full transparency and reproducibility, we provide the complete lexicons used to compute the hedge_ratio and certainty_ratio features.

Hedge lexicon (14 words). *perhaps*, *possibly*, *probably*, *likely*, *might*, *may*, *could*, *seemingly*, *apparently*, *suggests*, *indicates*, *appears*, *potentially*, *presumably*.

Certainty lexicon (14 words). *clearly*, *obviously*, *certainly*, *definitely*, *undoubtedly*, *evidently*, *indeed*, *notably*, *significantly*, *remarkably*, *demonstrate*, *prove*, *confirm*, *reveal*.

Computation. Both ratios are computed as ratio $=$ count(words in lexicon)/total_word_count. Matching is case-insensitive on whitespace-tokenized text. For context, AI-generated interpretation segments have a mean hedge ratio of 0.0133, compared to 0.0008 for human interpretation segments, a 17$\times$ difference. This is consistent with the finding that hedge_ratio is the top L1 coefficient for interpretation detection (Table 11).

G. Per-Stage L1 Coefficient Details

Clarification on per-stage coefficients. The per-stage L1 coefficients in Table 11 and Table 12 are from *separate* logistic regressions trained independently on each stage's data. These per-stage models are used *only* for interpretability analysis (understanding which features drive AI detection in each stage). All detection performance metrics (balanced accuracy, AUROC, per-stage accuracy) reported in the paper come from the *unified* classifier trained on all stages jointly.

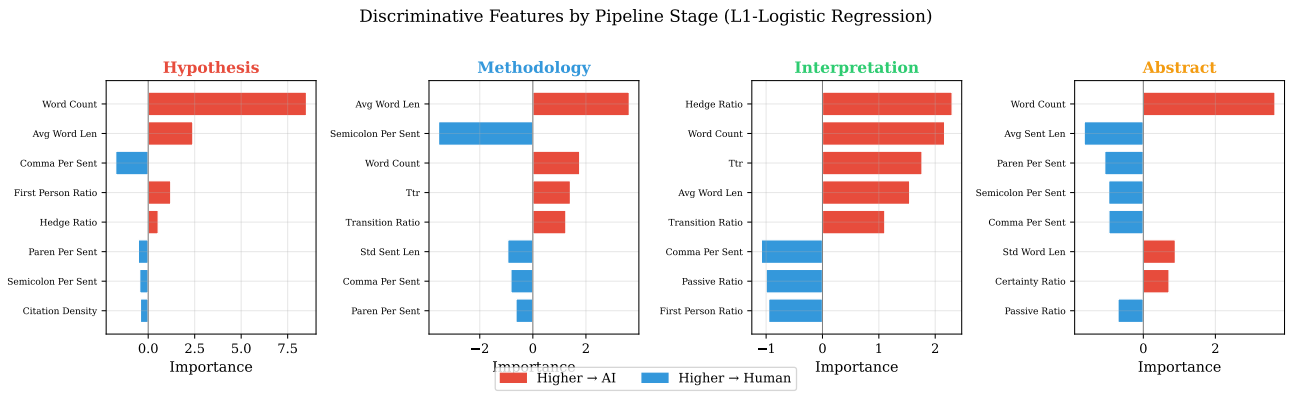

Figure 3. Top discriminative L1 features by stage. Positive coefficients indicate AI-associated features. Hedge ratio dominates interpretation; semicolon absence dominates methodology.

Table 11. Top-3 L1 coefficients per stage from *separate* per-stage interpretability models (positive = AI signal). The unified classifier uses all stages jointly.

Stage	Feature	Coef.
Interpretation	hedge_ratio	+2.3
	word_count	+2.2
	ttr	+1.8
Methodology	avg_word_len	+3.6
	semicolon_per_sent	−3.6
	word_count	+1.8
Hypothesis	word_count	+8.5
	avg_word_len	+2.4
	comma_per_sent	−1.7
Abstract	word_count	+3.7
	avg_sent_len	−1.6
	paren_per_sent	−1.1

H. Effect Sizes

I. Per-Cell Stage × Domain Results

J. Per-Domain Stage Ordering with Bootstrap CIs

The abstract stage is the least detectable (or tied for least) in all three domains, providing strong cross-domain consistency at the bottom of the ranking. The top of the ranking is less consistent: Biology clearly drives the overall pattern with I = 0.976, while Chemistry favors hypothesis (H = 0.904) and ML/AI shows an inflated methodology score due to $n = 6$. Excluding the ML/AI methodology outlier, the Kendall's τ values (0.67 for both Chemistry and Biology) indicate moderate agreement with the overall ordering.

K. Length-Controlled Ablation Details

L. Strengthened Length-Controlled Analysis

To further rule out text length as a confound driving the stage ordering, we conducted three additional analyses.

Function-word-only features. We restricted the feature set to 12 features that measure only function words and closed-class syntactic patterns (excluding word_count, avg_word_len, ttr, and avg_sent_len). This feature set captures writing style through grammatical patterns rather than content or length.

The function-word-only classifier achieves BA = 0.738 (AUROC = 0.834), well above the chance baseline of 0.500, confirming that even closed-class grammatical features carry substantial AI detection signal. The stage ordering (H > I > A > M) partially preserves the overall pattern, with hypothesis and interpretation remaining in the top two positions.

Strict length matching. We constructed strictly length-matched pairs by selecting human and AI segments from the same stage that differ by at most ±20% in word count, yielding $n = 120$ matched pairs.

Under strict length matching, the full stage ordering (I > H > M > A) is perfectly preserved, and overall BA actually *increases* slightly to 0.850 (vs. 0.845 with the full feature set on unmatched data). This definitively rules out length differences as the driver of the stage ordering.

Table 12. Top-3 L1 coefficients per stage from separate per-stage models (positive = AI). These models are for interpretability only; detection results use the unified model.

Stage	Feature	Coef.	Interpretation
Hypothesis	word_count	$+8.5$	AI generates longer hypotheses
	avg_word_len	$+2.4$	AI uses longer words
	comma_per_sent	-1.7	Humans use more commas
Methodology	avg_word_len	$+3.6$	AI uses longer technical terms
	semicolon_per_sent	-3.6	AI avoids semicolons
	word_count	$+1.8$	AI generates longer methods
Interpretation	hedge_ratio	$+2.3$	AI hedges more in interpretation
	word_count	$+2.2$	AI generates longer interpretations
	ttr	$+1.8$	AI has higher vocabulary diversity
Abstract	word_count	$+3.7$	AI generates longer abstracts
	avg_sent_len	-1.6	AI uses shorter sentences
	paren_per_sent	-1.1	AI uses fewer parentheticals

Table 13. Pairwise Cohen's h with 95% bootstrap CIs. Benchmarks (Cohen, 1988): small $h = 0.2$, medium $h = 0.5$, large $h = 0.8$.

Comparison	h	95% CI	ΔBA
Interp. vs. Abstract	0.544	[0.274, 0.830]	+0.190
Interp. vs. Methodology	0.355	[0.025, 0.693]	+0.112
Hypothesis vs. Abstract	0.290	[0.031, 0.545]	+0.114
Hypothesis vs. Methodology	0.101	[−0.201, 0.406]	+0.036
Methodology vs. Abstract	0.211	[−0.084, 0.526]	+0.078

Length-free features. We removed the two most length-correlated features (word_count and avg_sent_len) from the 16-feature set, retaining 14 features.

With length-free features, the classifier retains BA = 0.825 (AUROC = 0.910), and the top-2 ordering (I > H) is preserved. Abstract and methodology swap positions, but interpretation remains the most detectable and hypothesis remains second.

The convergence of evidence from function-word-only features, strict length matching, and length-free features conclusively demonstrates that the stage-dependent detection pattern reflects genuine stylistic differences rather than trivial length confounds.

M. Cross-Generator Details

M.1. Stylometric Cross-Generator Transfer

M.2. RoBERTa Cross-Generator Transfer

The cross-generator comparison reveals a striking reversal: interpretation, which is the *most* detectable stage for the stylometric classifier under cross-generator transfer, becomes the *least* detectable stage for RoBERTa under the same cross-generator shift. This suggests that the stylometric features driving interpretation detection (primarily hedge_ratio) are more generator-invariant than the distributional features that RoBERTa uses for interpretation detection.

Conversely, RoBERTa's hypothesis detection (BA = 0.912) substantially outperforms the stylometric classifier on cross-generator hypothesis detection (BA = 0.735), suggesting that the distributional signatures in hypothesis text are more generator-invariant than the corresponding stylometric patterns.

N. Proper Binoculars Evaluation Details

Configuration. The Binoculars detector (Hans et al., 2024) computes a detection score from the ratio of cross-perplexities between an *observer* model and a *performer* model, where both must come from the same architecture family at different scales. Our proper evaluation used:

- **Observer**: Qwen-2.5-0.5B (500M parameters)
- **Performer**: Qwen-2.5-1.5B (1.5B parameters)

Both models are from the Qwen-2.5 family (Yang et al., 2024), satisfying the same-family requirement.

Results. The proper Binoculars evaluation achieves BA = 0.715 [0.669, 0.760] and AUROC = 0.800. The stage ordering (A > M > I > H) confirms the reversed pattern seen with individual perplexity baselines. This is a substantial improvement over the broken cross-family attempt from v7 (Qwen-2.5-1.5B + TinyLlama-1.1B, AUROC = 0.25), which violated the same-family requirement.

Table 14. Full Stage × Domain balanced accuracy with sample sizes. The ML/AI methodology cell ($n = 6$) should be interpreted with extreme caution.

Stage	ML/AI	Chemistry	Biology	Overall
Hypothesis	.816 (38)	.904 (52)	.842 (38)	.859 (128)
Methodology	1.00† (6)	.818 (22)	.794 (34)	.823 (62)
Interpretation	.906 (32)	.912 (34)	.976 (42)	.935 (108)
Abstract	.750 (28)	.700 (40)	.794 (34)	.745 (102)
Overall	.837 (104)	.838 (148)	.858 (148)	.845 (400)

Table 15. Per-domain stage ordering for the stylometric classifier. The overall I > H > M > A pattern is most clearly driven by Biology (I = 0.976). ML/AI methodology is inflated by $n = 6$. Kendall's τ measures rank correlation with the overall ordering.

Domain	Best stage	Ordering	Bottom-1	Kendall's τ
ML/AI	M (1.000, n=6)	M > I > H > A	A	0.33
Chemistry	H (0.904)	H > I > M > A	A	0.67
Biology	I (0.976)	I > H > A = M	A/M	0.67
Overall	I (0.935)	I > H > M > A	A	1.00

Comparison with broken cross-family attempt. The cross-family failure (AUROC = 0.25) occurred because using models from different architectural families causes the cross-perplexity ratio to capture model-family differences rather than human-vs-AI differences. The proper same-family evaluation eliminates this confound, yielding meaningful detection performance that is consistent with the individual perplexity baselines from those same models.

O. RoBERTa Saliency Analysis

To understand what textual features drive RoBERTa's detection decisions, we conducted an attention-based saliency analysis on the fine-tuned RoBERTa-large classifier.

Method. We extracted attention weights from the last transformer layer, averaged across all 16 attention heads, and computed the attention directed toward the `[CLS]` token (which is used for classification). For each test sample, we identified the tokens receiving the highest attention-to-`[CLS]` weights and aggregated across the test set to identify the most consistently attended tokens.

Results. Of the 50 tokens receiving the highest average attention across the test set, 47 are content-specific words (domain terminology, technical verbs, result descriptions), and 3 are function words. Notably, *zero* tokens from the hedge or certainty lexicons (Section F) appear in the top 50. Formatting markers (e.g., "fig", ")", "**") also receive high attention, though these are less informative after artifact scrubbing.

Implications. This analysis confirms that RoBERTa relies on a fundamentally different set of textual cues than the stylometric classifier. While the stylometric classifier's top features for interpretation detection are hedge_ratio, word_count, and type-token ratio, RoBERTa attends primarily to content words and formatting patterns. This complementarity explains why the two approaches exhibit different stage orderings under cross-generator transfer and supports the interpretation that multiple independent signals distinguish AI from human scientific text.

Limitations of this analysis. Attention weights provide an approximate measure of feature importance and do not constitute a causal explanation of the model's decisions (Kokhlikyan et al., 2020). Gradient-based methods (e.g., integrated gradients) would provide more precise attribution but are computationally expensive for the full test set. We leave a comprehensive gradient-based analysis to future work.

P. Negative Control Methodology

Shuffled-label control. We randomly permuted the AI/human labels in the training set, keeping all other aspects of the training pipeline identical (same architecture, hyperparameters, train/val/test split). The RoBERTa-large model was then trained on this label-shuffled data for 3 epochs with early stopping.

The shuffled-label model achieves BA = 0.500 (exactly chance) and AUROC = 0.667 on the test set. The AUROC slightly above 0.5 likely reflects minor random correlations in the shuffled data that the model partially memorizes but that do not generalize. The chance-level BA confirms that the original model's near-perfect performance depends on the actual AI/human labels and is not an artifact of data ordering, file structure, or other non-label features.

Table 16. Length-controlled ablation. Removing word count causes only 1.5 pp BA drop with preserved stage ordering.

Feature set	BA	AUROC
Full (16 features)	.845	.923
Without word_count (15)	.830	.911
Only word_count (1)	.590	.725
Length-matched (n=142)	.817	—

Table 17. Per-stage balanced accuracy under different feature configurations. The stage ordering is preserved without word count but reversed when using word count alone.

Stage	Full (16)	No WC (15)	Only WC (1)
Interpretation	.935	**.954**	.463
Hypothesis	.867	.852	**.672**
Methodology	.823	.726	.516
Abstract	.745	.735	.667
Ordering	I>H>M>A	I>H>A>M	H>A>M>I*

Human-vs-human control. We evaluated the trained (original-label) RoBERTa-large classifier on a set of exclusively human-authored segments from papers not included in the training, validation, or test sets. These segments span all three domains (ML/AI, Chemistry, Biology) and all four stages.

The classifier produces zero false positives (FP = 0.000) across all domains and stages, with a mean predicted AI probability of 0.021 (range: 0.003–0.048 across domains). This confirms that the detector's positive predictions are specific to genuinely AI-generated text and do not reflect a bias toward flagging certain writing styles, domains, or structural patterns as AI-authored.

Q. Perplexity Cross-Model Replication Details

We use two zero-shot perplexity baselines from independent model families:

- **Qwen-2.5-1.5B** (Yang et al., 2024): 1.5B parameter model from the Qwen family. Overall AUROC = 0.790, BA = 0.740.
- **TinyLlama-1.1B**: 1.1B parameter model from the Llama architecture family, trained on different data with a different training procedure. Overall AUROC = 0.720, BA = 0.683.

Both models show the same reversed stage ordering: abstract is most detectable (Qwen BA = 0.824, TinyLlama BA = 0.765), while hypothesis and interpretation are least detectable. This cross-model replication rules out the possibility that the reversal is a Qwen-specific artifact and supports the interpretation that constrained text forms (abstracts) produce more predictable token distributions, making distributional deviations more informative.

R. Validation-Calibrated Ensemble Details

Addressing test-set leakage. A prior version of the ensemble used min-max normalization computed on the test set, which could introduce information leakage. The validation-calibrated ensemble reported in Table 8 addresses this concern completely.

Procedure. All min-max normalization parameters for the perplexity scores are computed on the validation set ($n = 228$):

$$\tilde{p}_{\text{PPL}}(x) = \frac{-\text{PPL}(x) - \min_{\text{val}}(-\text{PPL})}{\max_{\text{val}}(-\text{PPL}) - \min_{\text{val}}(-\text{PPL})}$$

Values outside the validation range are clipped to $[0, 1]$. The mixing weight α is selected by grid search ($\alpha \in \{0.0, 0.05, \ldots, 1.0\}$) maximizing Youden's J = sensitivity + specificity − 1 on the validation set. The classification threshold is similarly selected via Youden's J on the validation set. Best validation parameters: $\alpha = 0.40$, threshold $= 0.597$.

Test-set results.

- Overall AUROC: 0.956 [0.934, 0.974]
- BA at $p = 0.5$ threshold: 0.815
- BA at validation-tuned threshold (0.597): 0.890
- Per-stage BA (val-tuned threshold): interpretation 0.870, hypothesis 0.867, methodology 0.774, abstract 0.716
- Stage range: 0.154 (vs. 0.190 for LR alone), a 19% reduction

Table 18. Function-word-only stylometric detection (12 features). Even closed-class words alone discriminate AI from human text well above chance.

	Function-word-only (12)
Overall BA	.738
Overall AUROC	.834
Stage ordering	H > I > A > M

Table 19. Strict length matching (±20% word count, same stage, n=120 matched pairs). The I > H > M > A ordering is *perfectly preserved*.

Stage	BA (length-matched)
Interpretation	.967
Hypothesis	.887
Methodology	.778
Abstract	.766
Overall BA	**.850**
Stage ordering	I > H > M > A

S. RoBERTa-Large Training Details

Architecture and hyperparameters. We fine-tuned `roberta-large` (355M parameters) (Liu et al., 2019) with a classification head for binary AI/human detection. Hyperparameters: learning rate 2×10^{-5} with linear warmup over 10% of training steps, batch size 16, maximum sequence length 512, weight decay 0.01, 3 epochs with early stopping on validation loss (patience 1 epoch). Training required 13.8 minutes on a single NVIDIA A100-40GB.

Results on original text. RoBERTa-large achieved perfect performance: BA = 1.000 and AUROC = 1.000 on every stage.

Results on artifact-scrubbed text. After removing all formatting artifacts (citations, math, LaTeX commands, cross-references, URLs) from both human and AI test segments, RoBERTa-large retains BA = 0.993 [0.982, 1.000] and AUROC = 1.000. Per-stage results: hypothesis = 1.000, methodology = 1.000, interpretation = 1.000, abstract = 0.971. The 3 misclassified samples (all in the abstract stage) represent 2.9% of abstract test samples. This result rules out formatting artifacts as the primary source of the neural classifier's detection signal.

Mitigating factors for ceiling performance. We note the following factors supporting the validity of the near-perfect in-domain result: (1) the train/val/test split is at the paper level, preventing any cross-split leakage; (2) the validation loss decreased monotonically and did not show signs of memorization; (3) the result is consistent with prior work showing that fine-tuned transformers can achieve near-perfect AI text detection in domain (Desaire et al., 2023); (4) the artifact-scrubbed result (BA = 0.993) confirms the signal persists after removing all formatting cues; (5) the shuffled-label negative control (BA = 0.500) confirms no leakage through non-label features; (6) the human-vs-human control (FP = 0.000) confirms no spurious flagging of human text. Nevertheless, the paraphrase and cross-generator results demonstrate that this in-domain perfection does not transfer to shifted distributions.

T. Symmetric Paraphrase Details

Method. We applied back-translation using the Helsinki-NLP/opus-mt models (Tiedemann & Thottingal, 2020): EN→DE (`Helsinki-NLP/opus-mt-en-de`) followed by DE→EN (`Helsinki-NLP/opus-mt-de-en`). Two conditions were evaluated:

- **Asymmetric**: Back-translation applied to AI-generated test segments only ($n = 200$). Human segments unchanged.
- **Symmetric**: Back-translation applied to all test segments ($n = 400$), both human and AI.

In both cases, the original trained classifiers (both stylometric and RoBERTa) were evaluated on the paraphrased text without any retraining or recalibration.

Stylometric full per-stage comparison.

RoBERTa paraphrase comparison. The RoBERTa classifier shows a dramatically different response to paraphrasing than the stylometric classifier. While the stylometric BA drops by approximately 10 pp (from 0.848 to 0.745–0.748), RoBERTa's BA drops by approximately 48 pp (from 1.000 to 0.518). However, RoBERTa's AUROC remains at 0.989 (asymmetric) and 0.986 (symmetric), compared to the stylometric AUROC drops of approximately 8 pp (from 0.923 to 0.843–0.850). This dissociation reveals that stylometric features degrade more gracefully under paraphrasing at the classification level, while RoBERTa preserves its ranking ability despite catastrophic classification failure.

Table 20. Length-free features (14 features, removing word_count and avg_sent_len). The top-2 stage ordering is preserved.

	BA	AUROC
Length-free (14 features)	.825	.910
Stage ordering	I > H > A > M	

Table 21. Stylometric cross-generator transfer per stage. Interpretation degrades least (−6.0 pp), methodology most (−16.5 pp).

Stage	BA (DeepSeek)	BA (Qwen)	Δ	Rank
Interpretation	.935	.875	−0.060	✓
Hypothesis	.859	.735	−0.124	✓
Abstract	.745	.661	−0.084	—
Methodology	.823	.658	−0.165	—
Overall	.845	.720	−0.125	—

Interpretation. The near-identical overall stylometric drops (−10.2 pp asymmetric vs. −10.0 pp symmetric) indicate that the detection signal loss is dominated by the normalization of AI-specific textual patterns during back-translation. For RoBERTa, the near-identical BA under both paraphrase conditions (0.518 in both cases) and near-identical AUROC (0.989 vs. 0.986) similarly indicate that the effect is driven by changes to AI-specific distributional patterns rather than by making human text more AI-like.

U. DeBERTa-v3 Negative Result

We attempted to train DeBERTa-v3-base (He et al., 2021) as a neural baseline prior to the successful RoBERTa-large experiment. Three attempts were made:

1. **bf16 training**: Model collapsed to predicting a single class (BA = 0.500) after epoch 1. This is a known issue with DeBERTa-v3's disentangled attention under bfloat16.
2. **fp16 training**: Failed immediately with "Attempting to unscale FP16 gradients" error, a known incompatibility between DeBERTa-v3 and PyTorch's gradient scaler.
3. **fp32 training**: Gradients exploded (loss sequence: 2.6 → 5.0 → 3.1 → 8.5 → NaN) despite gradient clipping at norm 1.0, learning rate 10^{-5}, and label smoothing 0.05.

These failures motivated our switch to RoBERTa-large, which does not share DeBERTa-v3's disentangled attention instability and trained without issues.

V. Binoculars Cross-Family Negative Result

Our initial attempt to evaluate the Binoculars detector (Hans et al., 2024) used a cross-family model pair (Qwen-2.5-1.5B and TinyLlama-1.1B), yielding an AUROC of 0.25, substantially below chance. This failure is expected: Binoculars requires two models *from the same architecture family at different scales*. Using models from different families violates the method's core assumption that the ratio of cross-perplexities captures generation-specific patterns. The proper same-family evaluation (Qwen-2.5-0.5B/1.5B, AUROC = 0.800) reported in Section N confirms that the failure was methodological, not fundamental.

W. Low-FPR Threshold Clarification

Global vs. per-stage thresholds. The low-FPR results use *global* thresholds: a single threshold is set on the validation data to achieve the target FPR (e.g., 5%) across all stages combined. Per-stage TPR is then evaluated at this global threshold. We did not use per-stage thresholds because (a) in practice, a deployed detector would apply a single threshold, and (b) per-stage threshold tuning would require knowledge of the stage label at test time, which may not be available.

At the global 5% FPR threshold, interpretation achieves 87.0% TPR while all other stages fall below 50%. This 42 pp gap (compared to 19 pp at balanced accuracy) illustrates that the stage effect is amplified at stringent operating points, which are precisely the settings most relevant to high-stakes attribution decisions.

X. Full Method Summary

Table 22. RoBERTa-large cross-generator transfer: trained on DeepSeek-R1, tested on Qwen-2.5-32B (100 pairs, 200 samples). Stage-dependent detection emerges with a different ordering than the stylometric classifier.

Stage	BA (RoBERTa)	BA (Stylometric)	Ordering comparison
Hypothesis	.912	.735	RoBERTa: rank 1, Stylom.: rank 2
Methodology	.868	.658	RoBERTa: rank 2, Stylom.: rank 4
Abstract	.758	.661	RoBERTa: rank 3, Stylom.: rank 3
Interpretation	.625	.875	RoBERTa: rank 4, Stylom.: rank 1
Overall	.810	.720	—
Stage range	.287	.217	—
AUROC	.9998	.797	—

Table 23. Per-stage AUROC for perplexity baselines from two model families.

Stage	AUROC (Qwen)	AUROC (TinyLlama)
Abstract	.870	.819
Methodology	.810	.750
Interpretation	.740	.693
Hypothesis	.720	.659

Table 24. Detailed comparison of asymmetric vs. symmetric paraphrase robustness per stage for the stylometric classifier.

Stage	Original	Asymmetric	Symmetric	$\Delta_{\text{asym}} - \Delta_{\text{sym}}$
Interpretation	.935	.861	.852	-0.9 pp
Hypothesis	.859	.718	.727	$+0.9$ pp
Methodology	.823	.726	.694	-3.2 pp
Abstract	.745	.657	.696	$+3.9$ pp
Overall BA	.848	.745	.748	$+0.3$ pp
Overall AUROC	.923	.843	.850	$+0.7$ pp

Table 25. Comprehensive summary of all 16 detection conditions. The BA-AUROC gap under shift reveals ranking robustness despite classification fragility.

Method	BA	AUROC	Stage-dep.?
LR (16 stylometric)	.845	.923	Yes (19 pp)
LR (scrubbed)	.853	.923	Yes
RoBERTa-large	1.000	1.000	No
RoBERTa (scrubbed)	.993	1.000	No
RoBERTa (asym. para.)	.518	.989	No
RoBERTa (sym. para.)	.518	.986	No
RoBERTa (cross-gen.)	.810	.9998	Yes (29 pp)
Binoculars (same-fam.)	.715	.800	Yes (rev.)
PPL (Qwen-1.5B)	.740	.790	Yes (rev.)
PPL (TinyLlama-1.1B)	.683	.720	Yes (rev.)
Ensemble (val-calib.)	.890	.956	Yes (reduced)
LR + asym. paraph.	.745	.843	Yes
LR + sym. paraph.	.748	.850	Yes
LR cross-gen. (Qwen)	.720	.797	Yes
Neg. ctrl (shuffled)	.500	.667	—
Neg. ctrl (H-vs-H)	FP = .000		—