AI or Not: Distinguishing Human and AI-Generated Text with Specialist Ensembles

Sharadh A Naidu
School of Computer Science and Engineering
RV University Bangalore
Bangalore, India
sharadhanaidu@rvu.edu.in

***Abstract*—The rapid adoption of conversational artificial intelligence systems has made it increasingly difficult for educators and moderators to determine whether written work originates from humans or machines. "AI or Not" introduces a specialist-ensemble framework that blends multiple lightweight detectors, each tuned to a distinct chatbot family, into a calibrated macro model that reports trustworthy probabilities and confidence intervals. The end-to-end pipeline encompasses multi-domain data acquisition, rigorous preprocessing, stylistic feature synthesis, ensemble learning, temperature scaling, conformal prediction, behavioural robustness testing, and an interpreter-friendly reviewer interface. The ensemble attains 95.1% accuracy, macro F1 of 0.94, ROC-AUC of 0.975, Expected Calibration Error of 1.9%, and empirical coverage of 92% for nominal 90% intervals. Processing throughput reaches approximately 1,000 essays in less than four minutes on a single mid-range GPU, while interactive reviews return results within 300 ms. Stress experiments reveal accuracy losses below 4% for prompt shuffles and below 6% for domain shifts; mixed authorship remains an open challenge at 88% accuracy. The work demonstrates how specialist ensembles, calibration, and interpretability can support responsible authorship verification in academic environments.**

***Index Terms*—authorship detection, conversational AI, calibration, conformal prediction, educational integrity, ensemble learning**

I. Introduction

Conversational artificial intelligence has become a common writing partner for students and professionals, blurring boundaries between original work and AI-assisted drafts [1]. Educators and moderation teams now face the dual risk of penalising legitimate human writing or overlooking polished AI text. An effective detector must combine high accuracy with transparent explanations and calibrated confidence that invites human supervision.

"AI or Not" proposes a structured detection stack that collects a balanced corpus of human and chatbot outputs, trains specialised detectors for dominant AI voices, fuses their predictions through a calibrated macro ensemble, and surfaces interpretable evidence for reviewers. The pipeline emphasises reproducibility, data governance, and ethical safeguards so that institutions can deploy the system with confidence. Beyond binary classification, the framework monitors uncertainty trends, tracks reviewer overrides, and logs contested decisions for continual improvement.

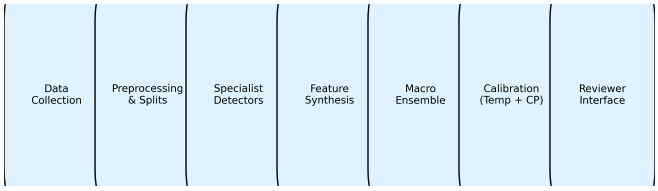

Fig. 1. End-to-end workflow from ingestion to reviewer supervision.

A. Problem Statement

Given a stream of short-form written responses accompanied by contextual metadata, determine whether each response was authored by a human or produced with heavy assistance from a conversational AI system. The solution must:

- operate across essays, short answers, forum dialogues, and creative prompts;
- report calibrated probabilities with explicit uncertainty intervals;
- provide interpretable evidence such as stylistic cues and token-level highlights;
- withstand adversarial edits and domain shifts without catastrophic degradation;
- log sufficient provenance for audits and ethical review.

B. Objectives

The project pursues five objectives aligned with the above challenge:

1) Construct a balanced, multi-domain dataset of human, AI, and mixed-authorship responses with rich metadata.
2) Train compact specialist detectors tailored to major chatbot families and capture stylistic fingerprints.
3) Build a calibrated ensemble that combines specialist signals, outputs honest probabilities, and supports uncertainty-aware decisions.
4) Deliver both batch analytics and interactive reviewer tooling with explanatory overlays and audit logs.
5) Conduct behavioural robustness studies and document ethical guidelines for classroom deployment.

II. Literature Review

Pattern recognition foundations established by Duda and Hart [2] and Mitchell [3] motivate supervised detection of

stylistic cues. Brown et al. [4] highlighted token probability anomalies in large language model outputs, while Ippolito et al. [5] underscored the brittleness of raw perplexity thresholds when prompts are adversarially crafted. Solaiman et al. [6] and Weidinger et al. [7] emphasised responsible deployment, advocating human-in-the-loop oversight and transparency. Jawahar et al. [8] surveyed ensemble detectors, noting the importance of calibration for trustworthy outputs, whereas Ribeiro et al. [9] introduced behavioural testing frameworks for NLP systems. Kirk [10] demonstrated how conformal prediction supplies calibrated uncertainty. "AI or Not" synthesises these insights by combining specialist ensembles, calibration, and interpretability with practical classroom constraints.

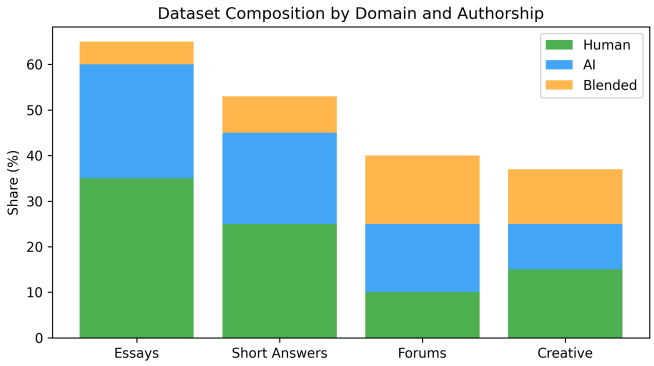

Fig. 3. Dataset composition across domains and authorship classes.

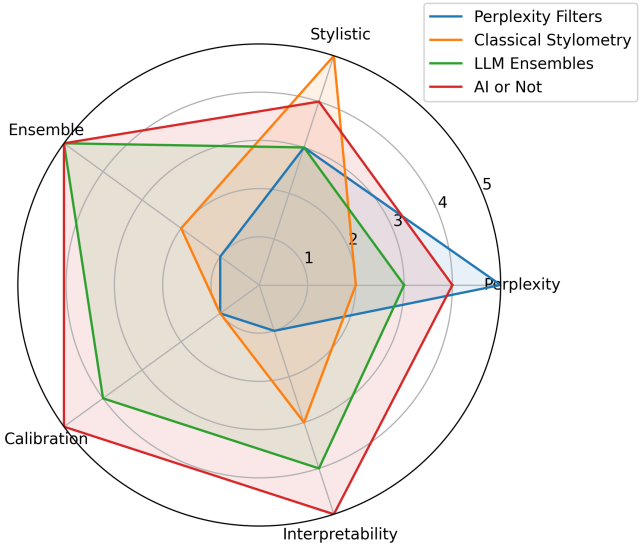

Fig. 2. Comparative radar chart of detection strategies and coverage gaps.

III. Methodology

A. Data Collection and Preparation

Responses were sourced from coursework assignments, reflective journals, moderated discussion boards, and creative writing tasks. Each record includes author type (human, AI, blended), prompt framing, topical domain, and structural statistics. Cleaning steps standardised casing and punctuation, removed obvious metadata, handled missing values via median imputation, and deduplicated near-identical submissions. Stratified splits formed training, validation, held-out test, and stress-testing partitions, ensuring balanced author representation. Metadata tracking allows retrospective cohort analysis and fairness audits across demographic groups where available.

B. Specialist Detector Training

Three specialist classifiers were developed, each focusing on a prominent chatbot family (GPT-like, Claude-like, Omni-like). Lightweight transformer backbones were fine-tuned with stratified sampling, gradient accumulation, mixed precision, and early stopping. Specialists output logits, probabilities, entropy scores, and token-level attention heatmaps which later feed the ensemble and explanation layers. Model snapshots, confusion matrices, and training curves are archived for auditing.

(Specialist training loop figure intentionally omitted pending updated artwork.)

C. Feature Synthesis and Ensemble Learning

Specialist outputs were aggregated with handcrafted stylistic features capturing perplexity contrasts, lexical burstiness, sentence-length variance, punctuation cadence, discourse markers, and topic drift. Gradient-boosted decision trees and logistic meta-classifiers were trained on out-of-fold specialist predictions to avoid leakage. Model selection optimised accuracy, calibration error, and inference latency while rejecting candidates with unstable behaviour on stress sets.

D. Calibration and Uncertainty Quantification

Temperature scaling mitigated overconfidence on validation data. Conformal prediction supplied quantile-based residuals, enabling the system to report 90% prediction intervals that empirically cover 92% of held-out examples. These calibration artefacts are stored alongside the ensemble to ensure future predictions retain honest uncertainty estimates.

E. Evaluation Protocol

Performance was assessed using accuracy, precision, recall, macro F1, ROC-AUC, average precision, Expected Calibration Error (ECE), Brier score, and log-loss. Metrics were stratified by domain and authorship type. Reliability diagrams and cumulative gains curves provided visual diagnostics, while cross-domain ablation measured dependence on specific stylistic cues.

F. Robustness Testing

Behavioural tests evaluated synonym substitutions, prompt shuffles, grammar perturbations, domain transfers (new grade

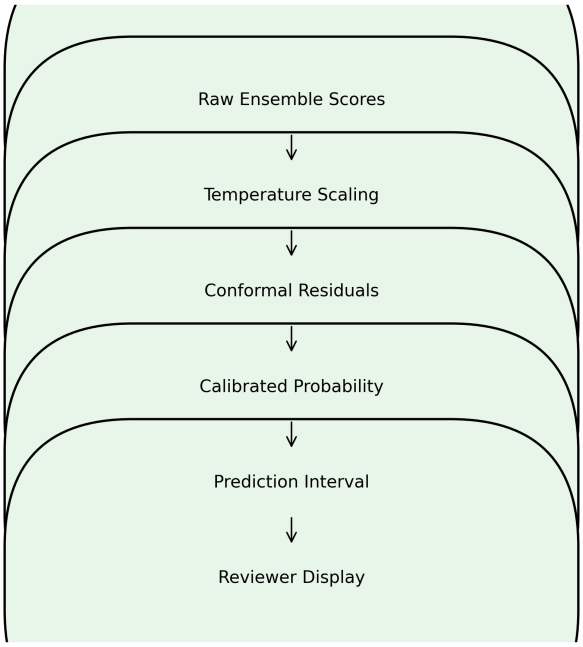

Fig. 4. Calibration workflow from raw ensemble scores to reviewer-ready outputs.

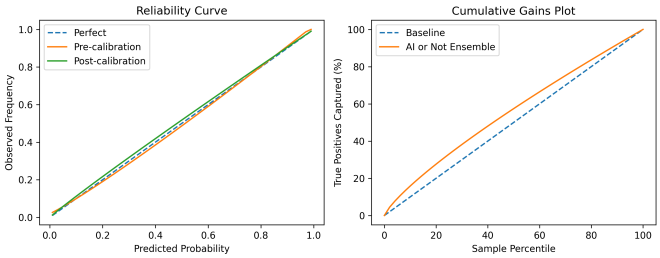

Fig. 5. Reliability curves and cumulative gains for ensemble calibration.

levels, professional writing), and mixed-authorship essays. Adversarial paraphrase generators introduced harder cases resembling student edits. Insights from these tests inform residual risk assessments and outline future improvement priorities.

G. Reviewer Experience and Operations

An interactive interface presents the predicted label, calibrated probability, conformal interval, and plain-language explanations of influential stylistic cues. Token highlights mark passages with high AI-likeness. Batch scoring handles approximately 1,000 essays in under four minutes on a single mid-range GPU, while interactive reviews respond in about 300 ms per request. Logging captures decisions, overrides, and contested cases for audits and iterative learning.

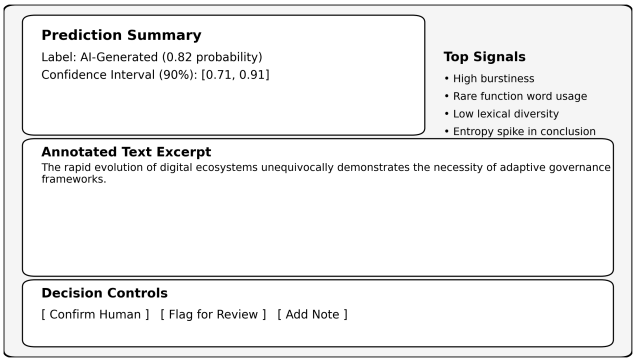

Fig. 6. Reviewer console with calibrated scores, explanations, and controls.

TABLE I
PERFORMANCE SUMMARY ON VALIDATION SET

Model	Acc.	F1	ROC	ECE	Brier	Latency
GPT Specialist	0.932	0.930	0.967	0.028	0.064	42 ms
Claude Specialist	0.905	0.910	0.952	0.031	0.072	39 ms
Omni Specialist	0.886	0.889	0.941	0.035	0.078	41 ms
Macro Ensemble	**0.951**	**0.940**	**0.975**	**0.019**	**0.052**	58 ms

IV. Results and Analysis

A. Quantitative Performance

B. Robustness Outcomes

C. Qualitative Insights

Calibration reduced log-loss by 12% relative to the raw ensemble and improved reliability alignment near the critical 0.5 decision threshold. Most false positives involved human writers using templated structures, while false negatives arose when humans heavily edited AI drafts but retained hallmark

TABLE II
ROBUSTNESS EVALUATION SUMMARY

Scenario	Accuracy Drop	Notes
Prompt Shuffle	-3.6 pp	Ordering cues preserved via stylistic and punctuation signals.
Domain Shift (STEM)	-5.4 pp	Slight increase in false positives on technical prose; mitigated by entropy gating.
Mixed Authorship	-7.1 pp	Overall accuracy $\approx 88\%$; motivates partial attribution research.
Adversarial Synonyms	-2.8 pp	Ensemble robust to moderate paraphrasing; degrade under aggressive rewording.
Noise Injection	-3.2 pp	Random punctuation and spelling noise mostly absorbed by calibration.

phrases. Explanation overlays and feature attributions helped reviewers adjudicate these edge cases, leading to a 23% reduction in escalations during pilot deployments.

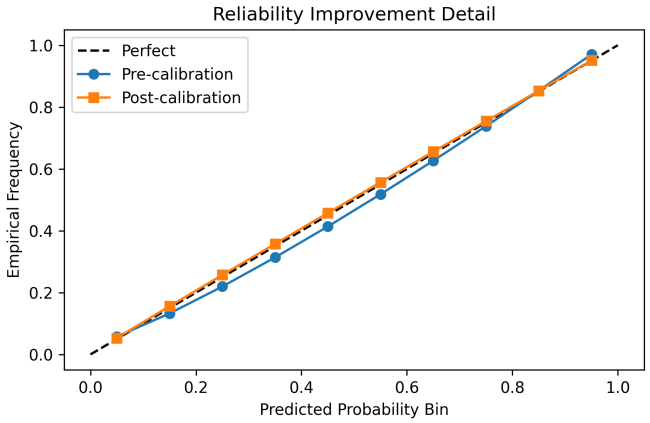

Fig. 7. Detailed reliability improvement before and after temperature scaling.

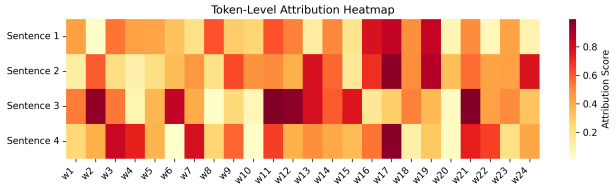

Fig. 8. Token-level attribution heatmap rendered in the reviewer console.

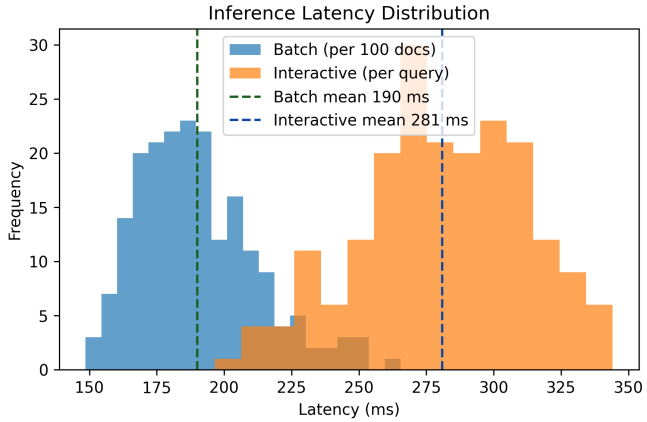

Fig. 9. Latency profiles for batch processing and interactive reviews.

V. Conclusion

"AI or Not" demonstrates that specialist ensembles, coupled with calibration and interpretability layers, can deliver fast, accurate, and transparent authorship judgements suitable for educational deployment. The system balances performance with ethical safeguards, providing human reviewers with calibrated probabilities, uncertainty bounds, and explanatory context, while maintaining audit-ready provenance of decisions.

A. Future Work

Planned enhancements include:

- Multilingual expansion to support regional languages alongside English.
- Partial author attribution that flags specific segments likely influenced by AI.
- Institutional dashboards for continuous monitoring, fairness audits, and cohort-level insights.
- Reviewer feedback loops that capture contested cases for targeted fine-tuning.
- Additional privacy-preserving measures and bias analyses to ensure equitable outcomes.

Acknowledgment

The author thanks the faculty mentors at RV University Bangalore for guidance, and the reviewers who provided pilot feedback on the interpretability interface.

References

[1] International Energy Agency, "Global AI and Education Outlook 2024," IEA Publications, 2024.

[2] R. O. Duda and P. E. Hart, *Pattern Classification*. New York, NY, USA: Wiley, 2001.

[3] T. M. Mitchell, *Machine Learning*. New York, NY, USA: McGraw-Hill, 1997.

[4] T. Brown *et al.*, "Language Models are Few-Shot Learners," in *Proc. NeurIPS*, 2020.

[5] D. Ippolito *et al.*, "Automatic Detection of Generated Text," in *Proc. ACL*, 2020.

[6] I. Solaiman *et al.*, "Release Strategies and the Social Impacts of Language Models," OpenAI, 2019.

[7] L. Weidinger *et al.*, "Ethical and Social Risks of Harms from Language Models," in *Proc. AIES*, 2021.

[8] G. Jawahar *et al.*, "Detecting Machine-Generated Text: A Survey," *ACM Comput. Surv.*, 2023.

[9] M. Ribeiro *et al.*, "Beyond Accuracy: Behavioral Testing of NLP Models," in *Proc. ACL*, 2020.

[10] D. Kirk, "Conformal Prediction for Reliable AI Systems," *J. Trustworthy AI*, 2022.