



VERIFYING HUMAN AUTHORSHIP IN AI-MEDIATED COMMUNICATION: A MULTI-LAYER FRAMEWORK COMBINING WRITING BEHAVIOR, LINGUISTIC SIGNALS, AND AI DETECTION

Indri Anjar Kartika Sari¹, Catur Suratnoaji², Agus Widiyarta³, Rifky Kurniawan⁴

^{1,2,3}Universitas Pembangunan Nasional “Veteran” Jawa Timur

⁴Cyber Security, IDS Digital College

¹25067020011@student.upnjatim.ac.id, ²catur_suratnoaji.ikom@upnjatim.ac.id

³agus_widiyarta.adneg@upnjatim.ac.id, ⁴rifky.kurniawan.MTI@ids.ac.id

Abstract

The rapid adoption of generative artificial intelligence (GenAI) has transformed digital communication and created new challenges concerning authorship, authenticity, and trust. Conventional AI-text detection approaches primarily examine the linguistic properties of completed texts, while providing limited evidence about how a text was produced. This study proposes a multi-layer human-authorship verification framework that integrates writing-behavior signals, linguistic characteristics, and external AI-detection signals. The proposed framework, implemented through the Creative Alibi prototype, records process-oriented indicators including typing rhythm, pauses, writing bursts, revisions, deletions, and copy-paste activity, while simultaneously extracting linguistic features such as moving-average type-token ratio, sentence burstiness, hapax legomena, sentence-starter diversity, and connective-word usage. External AI detectors are incorporated as a complementary rather than definitive source of evidence. The study is designed as an empirical comparative evaluation involving human-authored, AI-generated, and AI-assisted texts. Seven experimental configurations are proposed to examine the contribution of each analytical layer and their combinations. Performance will be evaluated using accuracy, precision, recall, F1-score, ROC-AUC, confusion matrices, and participant-level validation. The framework addresses a central limitation of product-oriented AI authorship detection by treating authorship verification as a process-oriented communication problem. The study contributes to emerging research on AI-mediated communication by proposing an evidence-based approach to authenticity that considers both the final linguistic product and the process through which communication is produced.

Keywords: AI-mediated communication; human authorship; generative AI; authenticity; writing behavior; keystroke dynamics; AI detection; digital communication

Article History

Received: September 2026

Reviewed: September 2026

Published: September 2026

Plagiarism Checker No

Copyright : Author

Publish by : Kohesi



This work is licensed under a [Creative Commons Attribution-NonCommercial 4.0 International License](https://creativecommons.org/licenses/by-nc/4.0/)



1. INTRODUCTION

Generative artificial intelligence has rapidly become integrated into digital communication, academic writing, professional communication, journalism, marketing, and creative production. Large language models can generate fluent and coherent texts that increasingly resemble human-authored communication. Consequently, determining whether a message was produced by a human, generated by artificial intelligence, or created through human-AI collaboration has become an increasingly important issue.

The problem is not limited to technological detection. Authorship is also a communication issue because perceptions of who produced a message can affect judgments concerning authenticity, trustworthiness, credibility, and the meaning attributed to the communication. Recent research on AI-mediated communication demonstrates that perceptions of communication can vary according to whether the author is human, AI-assisted, or fully AI-generated. A 2026 experimental study involving 547 participants found differences in perceived trustworthiness, authenticity, and willingness to use information depending on authorship and communication context. (ScienceDirect)

Similarly, recent scholarship has identified authorship and authenticity as important ethical challenges associated with AI-mediated communication. Rather than treating communication as a purely textual product, researchers have emphasized the importance of considering how meaning is produced and where the contribution to a message originates. (ScienceDirect)

These developments expose a limitation in conventional AI-text detection. Most detectors examine the final textual artifact and attempt to infer its origin from linguistic or statistical characteristics. However, the same text can potentially be produced through substantially different processes. A human may write a paragraph from scratch, an AI system may generate the paragraph automatically, or an AI-generated draft may be substantially revised by a human. Product-based analysis alone cannot fully represent these distinctions.

The reliability problem is also documented in recent empirical research. A 2024 evaluation of AI-text detectors found that detector performance can deteriorate substantially under different domains, models, and prompting strategies. (arXiv) Research published in 2025 likewise reported that human-written and AI-generated texts can be difficult to distinguish reliably, particularly when AI-generated writing reaches a high level of quality. (ScienceDirect)

Recent authorship-verification research has therefore begun exploring combinations of stylistic and semantic features rather than relying on a single signal. One 2025 study demonstrated that combining style and semantic features can improve authorship-verification performance under challenging conditions. (ScienceDirect)

At the same time, writing-process research provides another potentially valuable source of evidence. Keystroke-based research has demonstrated measurable differences between natural writing and copy-typing based on temporal and behavioral characteristics such as deletion rates, sentence-boundary behavior, and word-initiation patterns. (ScienceDirect)

Based on this background, this study proposes a process-oriented approach to human-authorship verification. Instead of asking only whether a completed text resembles AI-generated language, the proposed framework asks whether multiple independent sources of evidence—including writing behavior, linguistic characteristics, and external AI-detection signals—can jointly provide stronger evidence concerning authorship.



The study therefore addresses the following research questions:

- RQ1. To what extent can writing-behavior signals distinguish human-authored, AI-generated, and AI-assisted writing?
- RQ2. To what extent can linguistic signals distinguish the three authorship conditions?
- RQ3. What contribution do external AI-detection signals make to human-authorship verification?
- RQ4. Does combining behavioral, linguistic, and external AI-detection signals provide a more comprehensive authorship-verification framework than relying on a single analytical layer?

2. LITERATURE REVIEW

2.1 AI-Mediated Communication and Authorship

AI-mediated communication refers to communication in which artificial intelligence participates in the production, transformation, or delivery of messages. Generative AI expands this phenomenon by allowing users to delegate part of the linguistic production process to an AI system.

This development challenges traditional assumptions concerning authorship. A communication artifact may no longer have a binary origin of either “human” or “machine.” Instead, authorship can exist along a continuum ranging from entirely human production to AI-generated production, with multiple forms of human-AI collaboration between these extremes.

Recent research provides empirical evidence that perceived authorship influences communication outcomes. AI involvement has been associated with perceptions of authenticity and trustworthiness, although the direction and magnitude of these effects depend on communication context. (ScienceDirect)

Therefore, authorship verification should not be treated solely as a classification problem. It is also related to the social interpretation of communication.

2.2 Limitations of Product-Oriented AI Detection

Conventional AI detectors typically analyze the final text. Features may include token probabilities, sentence-level regularities, lexical patterns, perplexity, stylistic characteristics, or transformer representations.

The problem is that authorship is not directly observable from the final text. Detection therefore involves inference under uncertainty. Research has shown that AI detectors may experience false positives and reduced robustness across models, domains, writing styles, and prompting strategies. This limitation is especially important in academic and professional contexts where false accusations of AI use can have significant consequences.

2.3 Writing Behavior as Process Evidence

Writing is a temporal process rather than merely a static linguistic product. The process includes planning, typing, pausing, revising, deleting, and restructuring. Keystroke logging provides a means of observing selected aspects of this process without necessarily requiring access to the writer’s private cognitive state. Recent research demonstrates that keystroke timing and behavioral features can differentiate natural composition from copy-typing. (ScienceDirect). The present study extends this perspective to human-AI authorship verification.



2.4 Linguistic Authorship Signals

Linguistic features can provide complementary evidence concerning writing style. In the proposed framework, linguistic analysis includes lexical diversity, sentence-level variability, vocabulary uniqueness, sentence-starter diversity, and connective-word usage.

These measures are not treated as proof of authorship. Instead, they constitute one evidence layer within a broader verification framework. Recent authorship-verification research supports the use of combined stylistic and semantic information rather than relying exclusively on individual textual characteristics. (ScienceDirect)

3. CONCEPTUAL FRAMEWORK

The proposed framework consists of four layers.

Layer 1 – Behavioral Evidence

The behavioral layer captures process-oriented features:

- * typing duration;
- * inter-keystroke intervals;
- * pause frequency;
- * pause duration;
- * writing bursts;
- * insertion events;
- * deletion events;
- * revision frequency;
- * copy-paste activity.

These variables provide evidence concerning how the text was produced.

Layer 2 – Linguistic Evidence

The linguistic layer extracts:

- * MATTR;
- * sentence burstiness;
- * hapax-legomena ratio;
- * sentence-starter diversity;
- * connective-word frequency;
- * sentence-length variation;
- * lexical diversity.

Layer 3 – External Detection Evidence

External AI detectors are used as supplementary signals.

Importantly, an external detector's output is not treated as ground truth. It is treated as one probabilistic source of evidence.

Layer 4 – Authenticity Certificate

The final layer generates a digital certificate containing the analytical results and a SHA-256 hash associated with the analyzed artifact. The purpose is provenance rather than claiming mathematical proof that a human authored the text.



4. Methodology

4.1 Research Design

This study employs a quantitative experimental design. The unit of analysis is a writing session/document rather than an isolated sentence. Three authorship conditions are defined:

Human : the participant writes the text independently without generative AI.

AI-generated : the text is generated by a documented generative AI system without substantive human authorship.

AI-assisted/hybrid: the participant uses an AI-generated draft and subsequently modifies the text through human editing and rewriting. The three-class design is important because binary human-versus-AI classification does not adequately represent contemporary AI-mediated communication.

4.2 Dataset Construction

The study should construct a controlled corpus containing comparable writing tasks across the three authorship conditions. Each participant should receive matched prompts and writing requirements. The target length should be standardized, for example 250-400 words per session. Multiple topics should be used to reduce topic-specific bias. Where possible, more than one generative AI system should be used so that the model used to generate the AI condition does not become a hidden classification variable. The exact model, version, prompt, date, and generation conditions should be recorded for every AI-generated sample.

4.3 Behavioral Data Collection

The Creative Alibi prototype records process-level information during writing.

The behavioral dataset should include:

Variable	Description
duration_sec	total writing duration
keystroke_count	total keyboard events
mean_interkey_ms	mean interval between keystrokes
median_interkey_ms	median interval
pause_count	number of predefined pauses
mean_pause_ms	average pause duration
burst_count	number of writing bursts
revision_count	revision operations
deletion_count	deletion operations
copy_paste_count	copy-paste events
burst_ratio	proportion of burst activity

Raw keystrokes should not be retained unless necessary. The research dataset should prioritize derived behavioral features and participant anonymization.

4.4 Linguistic Feature Extraction

Each completed document is processed using the linguistic engine. MATTR measures lexical diversity while controlling for document length. Sentence burstiness measures variability in sentence length. Hapax-legomena ratio estimates the proportion of words occurring once. Sentence-starter diversity captures variation in sentence openings. Connective-word frequency measures the use of selected discourse connectives. These features are treated as statistical indicators rather than deterministic markers of human or AI authorship.



4.5 Experimental Configurations

To determine the contribution of each layer, the study conducts an ablation experiment.

Model	Behavioral	Linguistic	External AI
M1	✓	-	-
M2	-	✓	-
M3	-	-	✓
M4	✓	✓	-
M5	-	✓	✓
M6	✓	-	✓
M7	✓	✓	✓

M7 represents the complete proposed framework. This design allows the contribution of each analytical layer to be empirically evaluated rather than assumed.

4.6 Model Development

The current Creative Alibi prototype uses predetermined heuristic weights to combine multiple evidence sources. For the research version, these weights should not be treated as scientifically established values. The training data should be used to estimate model parameters, while a separate validation set should be used for model selection and a held-out test set should be used for final evaluation. Potential classifiers include:

- * logistic regression;
- * random forest;
- * gradient boosting;
- * XGBoost;
- * calibrated ensemble models.

The original heuristic scoring system can be retained as a baseline.

4.7 Data Splitting

A critical methodological requirement is participant-level separation. The same participant must not appear in both training and test sets. Otherwise, the model may learn individual writing style rather than authorship condition. A suitable initial configuration is:

- * 70% training;
- * 15% validation;
- * 15% test.

Alternatively, grouped cross-validation can be used when the number of participants is sufficient.

4.8 Evaluation Metrics

The study reports:

- * Accuracy;
- * Precision;
- * Recall;
- * F1-score;
- * macro-F1;



- * ROC-AUC;
- * confusion matrix.

Because the study contains three authorship conditions, macro-F1 should be emphasized rather than relying only on overall accuracy. Confidence intervals should also be reported where feasible.

4.9 Statistical Analysis

Performance differences between models should be evaluated using paired statistical procedures appropriate to the experimental design. For classification outcomes, McNemar's test can be considered for paired binary comparisons, while bootstrap confidence intervals can be used to quantify uncertainty around performance estimates. The study should avoid claiming that M7 is superior until the empirical results demonstrate such a difference.

5. PROPOSED RESULTS STRUCTURE

The results section should contain four major analyses.

5.1 Dataset Characteristics

Report:

- * number of participants;
- * number of writing sessions;
- * number of documents per condition;
- * average word count;
- * writing duration;
- * distribution of behavioral variables.

5.2 Behavioral Signal Analysis

Compare the behavioral distributions between:

- * Human;
- * AI-generated;
- * AI-assisted.

Visualizations should include:

- * pause distributions;
- * typing duration;
- * deletion frequency;
- * revision frequency;
- * burst ratio.

5.3 Linguistic Signal Analysis

Compare:

- * MATTR;
- * sentence burstiness;
- * hapax ratio;
- * sentence-starter diversity;
- * connective usage.

Statistical tests should determine whether observed differences are systematic rather than anecdotal.



5.4 Ablation Results

The principal result table should follow this structure:

Model	Accuracy	Precision	Recall	Macro-F1	ROC-AUC
Behavioral	TBD	TBD	TBD	TBD	TBD
Linguistic	TBD	TBD	TBD	TBD	TBD
External AI	TBD	TBD	TBD	TBD	TBD
Behavioral + Linguistic	TBD	TBD	TBD	TBD	TBD
Linguistic + External	TBD	TBD	TBD	TBD	TBD
Behavioral + External	TBD	TBD	TBD	TBD	TBD
Full Multi-Layer	TBD	TBD	TBD	TBD	TBD

TBD must remain until the experiment is actually run.

6. DISCUSSION

The proposed framework shifts the conceptual focus of authorship verification from product-only detection toward process-aware verification. This distinction is important because a human-authored text, an AI-generated text, and an AI-assisted text can potentially produce similar final linguistic artifacts while having substantially different production histories. Recent research on AI-mediated communication demonstrates that authorship affects perceptions of authenticity and trust, while other research indicates that AI-assisted communication can produce different outcomes depending on context and disclosure. (ScienceDirect) The proposed framework therefore does not assume that linguistic evidence alone can establish authorship. Instead, it treats authorship as a multi-source inference problem.

The behavioral layer contributes process evidence.

The linguistic layer contributes product evidence.

The external detector contributes an independent computational signal.

The authenticity certificate contributes provenance information.

This combination provides a more nuanced representation of AI-mediated writing than a binary AI detector.

Importantly, the framework does not establish that every human writer will demonstrate a particular typing pattern or that every AI-generated text will display a particular linguistic pattern. Individual differences, writing expertise, language background, editing practices, and assistive technologies may affect all measured signals. Recent research has specifically highlighted the possibility that writing style can confound AI-detection systems, including in non-native academic writing. Consequently, a process-oriented framework should be interpreted probabilistically rather than as an infallible authorship judge.

7. THEORETICAL CONTRIBUTION

The study contributes to research on AI-mediated communication in three ways. First, it conceptualizes authorship as a communication-production process rather than merely a property of the final message. Second, it connects communication authenticity with observable behavioral and linguistic evidence. Third, it introduces a multi-layer framework for examining human-AI co-production.



This perspective is consistent with emerging research arguing that AI-mediated communication complicates conventional concepts of authorship, authenticity, and human agency. (ScienceDirect)

8. PRACTICAL CONTRIBUTION

The proposed system has potential applications in:

- * academic writing;
- * journalism;
- * publishing;
- * freelance writing;
- * digital creative industries;
- * professional communication;
- * content provenance.

Rather than automatically accusing a writer of using AI, the system can provide a structured evidence record concerning the writing process. This distinction is particularly important in contexts where false accusations may have academic, professional, or reputational consequences.

9. Limitations

Several limitations must be acknowledged. First, behavioral data require controlled collection and cannot be reconstructed reliably from ordinary finished documents. Second, copy-paste behavior does not necessarily establish AI authorship because text may be copied from human sources. Third, AI-assisted writing exists on a continuum, making categorical labels inherently imperfect. Fourth, linguistic characteristics vary according to language, genre, education, editing, and individual writing style. Fifth, external AI detectors are themselves imperfect and should not be treated as ground truth. Existing research demonstrates substantial variation in detector performance under different conditions. Finally, the current Creative Alibi prototype contains heuristic scoring parameters that require empirical calibration before they can be presented as validated scientific thresholds.

10. CONCLUSION

This study proposes a multi-layer framework for human-authorship verification in AI-mediated communication. The framework combines writing-process evidence, linguistic characteristics, and external AI-detection signals rather than relying exclusively on the completed textual artifact.

The central contribution is a shift from AI detection toward authorship verification. The distinction is important because contemporary communication increasingly involves collaboration between humans and generative AI. A binary classification of “human” versus “AI” may therefore be insufficient to describe real communication practices.

The proposed experimental design enables the independent contribution of behavioral, linguistic, and external detection signals to be evaluated through controlled ablation experiments. The final scientific claim should be based on the empirical results obtained from the controlled dataset. Until those experiments are completed, the Creative Alibi prototype should be considered a research framework and technological implementation rather than a validated authorship-verification instrument.



REFERENCES

- Dik, S., Erdem, O., & Dik, M. (2025). Assessing GPTZero's accuracy in identifying AI vs. human-written essays.
- Kirk, C. P., & Givi, J. (2025). The AI-authorship effect: Understanding authenticity, moral disgust, and consumer responses to AI-generated marketing communications. *Journal of Business Research*, 186, 114984. (ScienceDirect)
- Yao, G., & Liu, Z. (2025). Can AI simulate or emulate human stance? Using metadiscourse to compare GPT-generated and human-authored academic book reviews. *Journal of Pragmatics*, 247, 103-115. (ScienceDirect)
- "AI, communication aids and the challenge of authentic authorship - Whose line is it anyway?" (2025). *Journal of Enabling Technologies*, 19(2), 102-112. (ScienceDirect)
- "The AI penalty and disclosure paradox: Trust, authenticity and knowledge uptake in AI-mediated communication." (2026). *Computers in Human Behavior: Artificial Humans*, 8, 100304. (ScienceDirect)
- "Writing with AI boosts trust-building efficiency." (2025). *iScience*. (ScienceDirect)
- "Combining style and semantics for robust authorship verification." (2025). *Machine Learning with Applications*, 22, 100732. (ScienceDirect)
- "Disentangling copy typing and natural writing behaviors using keystroke logs and deep learning models." (2026). *Assessing Writing*, 69, 101070. (ScienceDirect)