Journal of Data Processing

DLINE JOURNALS

Print ISSN: 2278 – 6481
Online ISSN: 2278 – 649X

JDP 2026: 16 (4)
https://doi.org/10.6025/jdp/2026/16/4/193-211

Authorship Attribution and Stylometric Analysis of AI-Generated Text: An Evaluation of prvsiyan's Extended Name That LLM Dataset

Pit Pichappan
Digital Information Research Labs
Chennai, Tamil Nadu. India

ABSTRACT

The proliferation of large language models (LLMs) has complicated the verification of digital text provenance, necessitating advanced methods to identify not just whether a text is AI-generated, but which specific model generated it. This study investigates authorship attribution and stylometric patterns using prvsiyan's Extended Name That LLM Dataset, a comprehensive corpus comprising 9,290 passages generated by 18 distinct LLMs across 510 prompts. Moving beyond binary AI-text detection, this research evaluates the discriminative power of lexical, character level, and engineered stylometric features in an 18-class model attribution task. Using a rigorous prompt grouped evaluation framework to prevent data leakage, the study demonstrates that while model specific linguistic signals are detectable and perform significantly above chance, robust forensic identification on unseen prompts remains challenging. The findings reveal that fine grained orthographic and lexical sequences carry more attribution information than global stylometric statistics, supporting the use of probabilistic attribution models rather than deterministic forensic certainty in current digital text analysis.

Keywords: Authorship Attribution, Stylometry, Large Language Models (LLMs), AI-Generated Text Detection, Natural Language Processing, Text Forensics, Linguistic Fingerprints

Received: 27 April 2026, Revised: 3 July 2026, Accepted: 19 August 2026

Copyright: DLINE

1. Introduction

The rapid development of large language models (LLMs) has transformed natural language generation, creating new opportunities for content creation while raising concerns about authenticity, originality, and the provenance of digital text [1, 2]. Models such as GPT-4, PaLM, and Llama can generate fluent and coherent text, making it increasingly difficult to distinguish machine generated content from human written material. This challenge has complicated conventional authorship attribution methods and has significant implications

for the integrity of authorship, legal and ethical accountability, and the credibility of online information [3, 4, 5, 6]. The growing difficulty of identifying the origin of digital content also raises concerns about the safety and reliability of online environments [7, 8].

Authorship attribution plays an important role in addressing these challenges. Its applications include combating misinformation [9, 10, 11, 12], protecting intellectual property rights [13, 14], and identifying fraudulent activities [15, 16]. In the context of LLM-generated content, neural authorship attribution extends this objective by attempting to identify the language model responsible for producing a particular text. This task differs from conventional AI-text detection because it seeks to determine the specific generating model rather than simply classify a passage as human written or AI-generated. The distinction is particularly important given the diversity of proprietary and open source LLMs and the stylistic variation that may occur within each category [17].

The present study examines authorship attribution and stylometric characteristics using prvsiyan's Extended Name That LLM dataset. It focuses on linguistic features across lexical, syntactic, and structural dimensions and considers their potential contribution to identifying model specific writing patterns. Particular attention is given to the interpretability of stylometric features and their potential to complement classifiers based on pretrained language model representations. By bringing these perspectives together, the study addresses the need for authorship attribution approaches that combine effective text representation with interpretable linguistic evidence.

2. Background

2.1 Evolution of Authorship Attribution

Authorship attribution has advanced considerably through developments in natural language processing, statistical text analysis, and machine learning. Traditional approaches have primarily relied on stylometry, which examines recurring linguistic characteristics to identify distinctive writing patterns. Through feature engineering, stylometric methods represent textual style using measurable linguistic properties [18]. Such methods provide a comparatively transparent means of investigating how writing patterns differ across authors or text sources.

The emergence of machine learning algorithms capable of processing high dimensional data has enabled more expressive representations of textual information. More recently, pretrained language models have been used to generate contextual embeddings that capture complex linguistic relationships and support neural authorship attribution systems [19]. Although these approaches can offer more powerful text representations, their decisions may be less interpretable than those produced by explicitly engineered linguistic features. Consequently, the increased representational capacity of neural methods can introduce a trade-off between predictive performance and explainability [20].

2.2 Authorship Attribution in the Era of Large Language Models

The development of LLMs has introduced new challenges for authorship attribution because machine-generated text can exhibit the fluency, coherence, and stylistic variation commonly associated with human writing. Research has therefore begun to investigate how the writing signatures of different language models can be characterized and distinguished.

Huang presented a comprehensive literature review examining recent research on authorship attribution in the era of LLMs [28]. In related work, T. Kumarage [21] conducted an empirical investigation of LLM writing signatures, focusing on differences between proprietary and open source models and variations among models within these categories. This research direction considers the integration of stylometric features covering lexical, syntactic, and structural properties with pretrained language model classifiers. The objective is to explore whether interpretable linguistic characteristics can complement neural representations and support a more transparent understanding of model attribution.

These studies establish the relevance of investigating model-specific writing patterns rather than treating AI-generated text as a single homogeneous category. However, identifying such patterns requires careful consideration of the relationship between observable linguistic features and the representations learned by neural models.

2.3 Stylometric Features and AI-Text Detection

Existing studies have investigated a range of statistical and stylistic features for characterizing generated text. Examples include word count, vocabulary richness, and readability, which provide measurable indicators of textual structure and linguistic complexity [22-25]. These features offer an interpretable basis for examining differences among text sources, although their usefulness for identifying a specific LLM must be established empirically.

Another research direction involves zero-shot detection methods that use text-generation statistics, including top-k, top-p, and entropy, to characterize generated text [26]. These methods provide an alternative to approaches that depend on supervised training with labelled examples.

In addition, black-box approaches based on machine learning and deep learning models have been used to distinguish AI-generated text from human written text [27, 28].

Although these approaches are related, AI-text detection and LLM authorship attribution represent distinct analytical objectives. A detector may identify a passage as machine generated without determining which model produced it. Conversely, a model-attribution system must learn or identify patterns that discriminate among the outputs of different LLMs. This distinction motivates the investigation of linguistic features that may carry information about the identity of the generating model.

2.4 Research Problem and Motivation

Despite advances in stylometry, machine learning, and neural text representation, identifying the specific LLM responsible for generating a passage remains a challenging problem. Different models may produce texts with similar linguistic characteristics, while a single model may generate different writing styles depending on the prompt and context. These factors can make model specific writing signatures difficult to isolate.

Traditional stylometric analysis offers an interpretable way to examine observable linguistic differences. Measures associated with vocabulary, sentence construction, and textual organization can be inspected to understand potential distinctions among text sources. Pretrained language model representations, in contrast, can capture complex contextual relationships that may not be fully represented by manually engineered features. The relationship between these two approaches provides a basis for investigating whether interpretable linguistic features can complement neural authorship-attribution methods.

The central research problem of this study is to examine whether lexical, syntactic, and structural characteristics in prvsiyan's Extended Name That LLM dataset can provide useful evidence for distinguishing texts associated with different LLMs. A related concern is whether these features can contribute to the interpretability of pretrained language model classifiers.

Accordingly, the study focuses on the potential of stylometric analysis to characterize model specific writing patterns and support neural authorship attribution. The actual effectiveness of individual features, their contribution to classification performance, and their generalizability must be established through empirical evaluation of the dataset.

3. Research Aim and Objectives

The primary aim of this study is to investigate authorship attribution and stylometric patterns in AI-generated text using prvsiyan's Extended Name That LLM dataset.

The specific objectives are:

1. To examine the potential of lexical, syntactic, and structural features for characterizing the writing patterns of LLM-generated text.

2. To investigate whether stylometric features can help distinguish texts associated with different language models.

3. To explore the potential for integrating interpretable stylometric features with pretrained language-model representations.

4. To examine the relationship between conventional statistical approaches, zero-shot detection methods, and neural approaches to AI-generated text analysis.

5. To distinguish the task of detecting AI-generated text from the more specific task of attributing generated text to its originating LLM.

These objectives provide a coherent framework for evaluating the role of stylometry in LLM authorship attribution while maintaining a clear distinction between established background research and findings that require experimental verification.

3.1 Scope and Analytical Perspective

This study is situated at the intersection of stylometric analysis and neural authorship attribution. It considers linguistic features as a potential source of interpretable evidence for distinguishing among LLM-generated texts. Rather than treating feature based and neural approaches as mutually exclusive, the study considers their possible complementary roles.

Stylometric features can provide explicit measurements of textual characteristics, allowing researchers to investigate which aspects of language may vary across model outputs. Pretrained language-model representations can capture more complex contextual and semantic patterns. Examining these approaches together may help clarify whether interpretable features contribute useful information beyond that captured by neural representations.

The analysis is centred on prvsiyan's Extended Name That LLM dataset. However, conclusions concerning dataset composition, individual model characteristics, feature importance, and comparative predictive performance must be supported by the dataset itself and the corresponding experimental results. Such conclusions should not be inferred solely from the general literature on stylometry or AI-text detection.

3.2 Contribution

The increasing use of large language models for text generation has made authorship attribution an important problem in digital text forensics. Distinguishing between human-written and AI-generated content is already challenging, and identifying the specific LLM responsible for generating a passage introduces an additional level of complexity.

Previous research has explored several approaches to this problem, including traditional stylometry, statistical measures of textual characteristics, zero-shot detection methods, and machine-learning and deep learning classifiers. These approaches offer different advantages in terms of interpretability and representational capacity. Stylometric features provide explicit linguistic indicators, whereas pretrained language model representations can capture more complex contextual patterns but may be less transparent.

By examining prvsiyan's Extended Name That LLM dataset, this study focuses on the potential of lexical, syntactic, and structural features to characterize model-specific writing patterns and support neural authorship attribution. The investigation is motivated by the possibility that interpretable linguistic features may

complement pretrained language model representations and contribute to a clearer understanding of model attribution.

Ultimately, the contribution of stylometric analysis must be established through systematic empirical evaluation. Reporting the dataset characteristics, experimental procedures, evaluation metrics, and comparative results will be essential for determining the extent to which these features support reliable and interpretable LLM authorship attribution.

4. Methodology

The study employs a systematic, machine-learning-based approach to evaluate text-derived features for multi class LLM attribution:

Dataset & Preprocessing: The analysis utilizes the *new* split of *prvsiyan's Extended Name That LLM Dataset*, restricted to 4,158 clean passages from 231 complete prompt blocks (where all 18 models answered the same prompt).

Leakage-Safe Partitioning: To prevent prompt content leakage, data was partitioned by *prompt ID* rather than randomly by row. The split consisted of 161 training prompts (2,898 passages), 35 validation prompts (630 passages), and 35 held-out test prompts (630 passages).

Feature Representations: Three distinct feature families were extracted from the passage text:

1. Word TF-IDF (lexical sequences)

2. Character TF-IDF (character n-grams and orthographic patterns)

3. Engineered Stylometric Features (35 global statistical measures, e.g., sentence length, punctuation rates, vocabulary richness).

Classification & Evaluation: A linear Stochastic Gradient Descent (SGD) classifier was trained using validation-set selection for regularization strength. Models were evaluated on Accuracy, Balanced Accuracy, Macro-F1, and Top-3 Accuracy to account for the balanced 18-class distribution (where random chance is ~5.6%).

4.1 Dataset Description: prvsiyan's Extended Name That LLM Dataset

4.11 Overview

prvsiyan's Extended Name That LLM Dataset is a large, API-logged corpus of AI-generated short-form prose designed for authorship attribution and stylistic analysis of large language models (LLMs). Version 4 of the dataset (released 19 October 2026) contains 9,290 passages written by 18 LLMs in response to 510 distinct prompts, generated between 16 and 19 October 2026. The dataset is organised as *prompt blocks*: each prompt is sent once to every model in the panel, so that topic and task are held constant while the generating model varies. All passages are AI-generated, and no passage or prompt from the associated Kaggle competition *Name That LLM* is included. The dataset is released under a CC BY-SA 4.0 licence.

The dataset was built to support three research uses: (i) additional training or validation data for authorship attribution; (ii) characterisation of per-model stylistic register; and (iii) measurement of how stable a model's style remains across prompts. Every passage carries detailed provenance, including the exact requested and served model identifiers, API route, timestamps, token counts, cost, and identity checks. Calls for which the served model differed from the requested model were quarantined and are not published under the requested label.

4.2 Composition and Splits

The corpus comprises five splits, summarised in Table 1. The primary split, new, contains 300 prompts written

specifically for this dataset, each answered once by all 18 models (5,400 passages). Of these, 231 prompt blocks (4,158 passages) passed the full quality funnel. The *train_regen* split contains 150 community-reconstructed competition training prompts answered by all 18 models (2,700 passages; 252 clean). The proxy_v1 split contains 60 earlier prompts answered by the 12 models reachable through the Kaggle Benchmarks Model Proxy (720 passages; 516 clean). Two auxiliary splits are retained for methodological documentation but excluded from the clean corpus: *new_kimi_nothink* (300 passages from a single model with reasoning disabled) and *pilot_range* (170 passages generated with a shared word-range suffix rather than per-generation length targets).

Table 1. Dataset splits (Version 4).

Table 1 summarises the composition of the five dataset splits, including the number of prompts, models, total passages, clean passages, generation dates, and generation protocol. The new split is the largest fully balanced component, with 5,400 passages generated from 300 prompts across 18 models; 4,158 passages from 231 complete prompt blocks are retained for the primary clean analysis. The *train_regen* and *proxy_v1* splits provide complementary data under different prompt and serving conditions, whereas *new_kimi_nothink* and *pilot_range* are auxiliary protocol subsets and are not part of the primary clean corpus. The table therefore shows both the scale of the archive and why the primary attribution experiment uses a carefully filtered subset.

Split	**Prompts**	**Models**	**Rows**	**Clean rows**	**Generation dates**	**Protocol**
new	300	18	5,400	4,158	2026-10-17 to 2026-10-18	Prompts written for this dataset; per-generation target length "(about N words)"
new_kimi_nothink	300	1	300	0	2026-10-17 to 2026-10-18	kimi-k3 re-run with reasoning disabled; excluded from clean corpus
proxy_v1	60	12	720	516	2026-10-16	Earlier prompt bank; per-generation length instruction
pilot_range	10	17	170	0	2026-10-17	Shared word-range suffix; retained to document length effects
train_regen	150	18	2,700	252	2026-10-18 to 2026-10-19	Community-reconstructed competition training prompts
Total	**510 unique**	**18**	**9,290**	**4,926**		

Table 1. The composition of the five dataset splits

The 18-model panel spans five model families: five Claude models (claude-fable-5, claude-opus-4-6, claude-opus-4-8, claude-opus-5, claude-sonnet-5); four Gemini models (gemini-3.1-pro-preview, gemini-3.6-flash, gemini-3.7-flash, gemini-3.8-flash); five GPT models (gpt-5.5, gpt-5.6-luna, gpt-5.6-sol, gpt-5.6-terra, gpt-6-astra); two GLM models (glm-5.3, glm-5.3-flash); and one model each from Moonshot AI (kimi-k3) and Alibaba (qwen3.8-max). These are the 18 labels of the *Name That LLM* competition.

4.2.1 Generation Protocol

Each prompt was sent as a single user message, verbatim, with no system prompt and no temperature, top-p, seed, reasoning-effort, or thinking parameter; provider defaults were used, with only an output-token cap applied. The *reasoning_setting* column records the reasoning configuration per row (generally "provider default"). Claude models therefore ran at their API defaults (adaptive thinking for Fable 5, Opus 5, and Sonnet 5; no thinking for Opus 4.6 and 4.8).

Prompt topics in the new split are disjoint from the 150 community reconstructed competition training topics. The domain mix is physics 23%, biology 18%, philosophy 14%, mathematics 11%, history 8%, creative prose 8%, computer science 6%, psychology 5%, economics 4%, and other 3%. Task types are explain 57%, describe/narrate 30%, argue/opinion 7%, story/fiction 5%, and compare 1%. Prompts include one or two style modifiers, a dictated opening phrase in 25% of cases, and the instruction to produce "a single self-contained paragraph with no headings or lists."

Length targets were assigned per generation rather than shared across models: each prompt–model pair received its own target N drawn from the competition training word-count distribution, clipped to [40, 250] words (median 98), and appended as "(about N words)." This design followed a pilot finding that a shared word range allowed the model label to explain 80% of within-prompt length variance, whereas in the competition training data the label explains only 1.0%. The pilot passages generated with a shared range are retained as pilot_range but excluded from the clean corpus.

Models were accessed through five routes: Anthropic Messages API (first party), OpenAI Chat Completions (first party; gpt-5.5 requested as the dated snapshot gpt-5.5-2026-04-23), Google Vertex AI or the Gemini Developer API, *OpenRouter* with the model's own lab pinned and fallbacks disabled, and the Kaggle Benchmarks Model Proxy. On every call, the served model identifier reported by the response was checked against the requested identifier (or a dated snapshot); mismatches were quarantined and excluded from the published data.

4.2.2 Quality Control and Filtering

A multi-step quality funnel was applied to produce passages_clean.csv. A passage passes only if it is unique; not truncated or content-filtered; free of code, markdown, emoji, and URLs; a single paragraph of 40–250 words in English; free of refusals or chatty openers; deduplicated; and shares fewer than three 8-word shingles with any competition passage. The passages.csv file retains all 9,290 passages with per-row quality flags, including *passes_funnel, funnel_fail_step, is_refusal, is_truncated, is_empty, has_markdown, in_word_range*, and *max_shared_8grams_comp*. The *train_regen* split uses a modified funnel (*passes_funnel_regen*) that omits the competition 8-gram step, because those passages naturally share prompt-dictated phrases with competition passages of the same prompt; test-set overlap is still excluded.

Funnel pass rates in the new split range from 0.960 to 0.997 across models, with a median passage length of 89–109 words and mean target-to-actual length ratios close to 1.0. Refusal rates are zero for all 18 models in the new split. Markdown rates are low (0–10%) and are flagged rather than removed where italics markup survives the funnel.

4.2.3 Files and Schema

The dataset contains 13 files. passages.csv (9,290 rows) contains every passage with provenance and quality columns. passages_clean.csv (4,926 rows) contains funnel-passing passages from complete blocks in the new and proxy_v1 splits. prompts.csv (510 rows) documents prompt templates, domain, task type, topic, style,

and opening phrase. blocks.csv (820 rows) records, per prompt and split, how many models were expected, present, and passing. label_summary.csv (66 rows) provides per-model, per-split counts, funnel pass rates, length statistics, typography rates, refusal and markdown rates, and routes. train_vs_dataset_stats.csv (18 rows) compares per-model competition-training aggregates with this dataset, and flags register differences. provenance. *jsonl* (9,291 envelopes) contains one record per API call with request body, raw response JSON, usage, and identity checks (keys and authorisation headers removed; quarantined text withheld). *provenance_summary*.json summarises per-model identifiers, routes, dates, settings, and Google *model_version* probes. A *starter_notebook.ipynb*, README.md, and CHANGELOG.md are also included.

The passage-level schema includes stable row identifiers (id), dataset version (*dataset_version_added*), call attempt index, *prompt_id*, model label, raw response text, word count, prompt length bounds, per-generation target length, the exact prompt sent, domain, task type, topic, style, opening phrase, split, provider route, requested and served model identifiers, served provider, generation timestamp, reasoning setting, finish reason and normalised finish reason, input/output/reasoning token counts, cost in USD, call and response identifiers, retry count, source prompt identifier for regenerated prompts, and the quality flags described above. All columns are documented in the file metadata.

4.2.4 Comparison with the Competition Training Set

Because the public APIs available in October 2026 may return text in a different register from the competition host's original passages, the dataset includes per-model comparisons against competition training aggregates (no competition rows are copied). Several register differences are flagged. For example, Gemini models produce em dashes in 13–28% of passages in this dataset versus 0–1% in the competition training data; gpt-6-astra produces *em* dashes in 33% versus 9%; kimi-k3 produces *em* dashes in 75% versus 52% and uses curly apostrophes far less often (0.023 versus 0.558). claude sonnet-5 has a longer mean sentence length in this dataset (37.9 words) than in competition training (26.12). These differences are documented in *train_vs_dataset_stats.csv* and in the "Known gaps and caveats" section of the dataset card, and users are advised to consult them before training on the rows.

A further caveat concerns prompt register: a prompt-level TF-IDF classifier separates the new prompts from community reconstructions of the competition training prompts almost perfectly (AUC H" 1.0), driven by format sentences, "Explain/Describe" openings, "for a general audience," and second-person modifiers. The authors note that this is the likely reason one of their blocks was worth only about 0.2–0.25 competition prompts in transfer tests. A planned new2 split uses prompts written in the reconstructions' register (style-only AUC 0.76; full TF-IDF AUC 0.77 against held-out reconstructions), and the *train_regen* split answers the reconstructed training prompts themselves.

4.2.5 Intended Use and Leakage-Safe Validation

The dataset is intended for authorship attribution, stylistic profiling, and register-stability analysis. Because each prompt appears with multiple models, users should validate with prompt-grouped folds (e.g., *GroupKFold* grouped by *prompt_id*) rather than random row splits; random splits place the same prompt in both training and validation and substantially overstate accuracy. The dataset's new prompts are disjoint from the competition's training topics, so adding these rows to competition training folds does not leak validation prompts. A starter notebook demonstrates loading, block completeness checks, a prompt-grouped baseline, and evaluation on the competition training set inside Kaggle without copying it.

5. Analysis

5.1 AI-generated text authorship attribution

We completed the primary modeling experiments, linguistic fingerprint analysis, explainability analysis, prompt-grouped evaluation, and cross-protocol generalization tests using the uploaded archive (7).zip.

The results are actual computed outputs from the archive. The main finding is that model-specific linguistic signals are detectable, but identifying the exact generating model remains difficult when the test prompts have

not been seen during training.

Primary experiment: identifying the generating model

The primary analysis used the new experimental split, restricted to 231 complete clean prompt blocks, containing 4,158 passages from 18 model classes.

We tested results among the evaluated feature representations.

Test accuracy - 25.87%

Test macro-F1 - 0.254

Top-three accuracy - 51.11%

Held-out test passages - 630

The test set comprised 35 previously unseen prompt groups, with 35 passages per model.

The train/validation/test partition was grouped by prompt:

Partition	Prompts	Passages
Training	161	2,898
Validation	35	630
Test	35	630
Total	231	4,158

This design prevents passages answering the same prompt from appearing in both training and test sets.

5.11 Model comparison

Results on the held-out prompt-grouped test set.

Feature representation	Accuracy	Macro -F1	Top-3 accuracy
Word TF-IDF	0.259	0.254	0.511
Character TF-IDF	0.243	0.244	0.517
Engineered stylometric features	0.170	0.163	0.440

The linear SGD classifier used validation-set selection of regularization strength. Macro-F1 weights the 18 model classes equally; top-three accuracy counts a prediction as successful when the correct label is among the three highest-scoring candidates.

5.12 Test performance by feature representation

split	rows	unique _prompts	models _flag	clean	empty	Truncated	median _words
new	5400	300	18	5326	0	0	100
new_kimi_nothink	300	300	1	293	0	0	103
pilot_range	170	10	17	161	0	0	169
proxy_v1	720	60	12	702	0	0	104
train_regen	2700	150	18	2067	0	0	102

Table 1. Attribution performance

Table 1 compares attribution performance across the primary held-out test set and two external protocol-transfer evaluations. The primary new-split evaluation includes 630 passages from 35 unseen prompt groups and achieves 0.259 accuracy and macro-F1 of approximately 0.254. Performance is higher on proxy_v1 (accuracy 0.335; macro-F1 0.330), but this result is based on 12 shared model labels and a different generation route and prompt bank. The *train_regen* evaluation is weaker (accuracy 0.183; macro-F1 0.160) and is based on only 14 complete prompt blocks. These external results should be interpreted as exploratory distribution-shift checks rather than directly equivalent estimates of primary test performance.

Evaluation	n_ classes	test _r ows	test _ prompts	accuracy	balanced _ accuracy	macro _f1	weighted _f1	Interpretation
New split: prompt-grouped held-out test	18	630	35	0.25873	0.258730159	0.253618754	0.253619	Primary test; unseen prompt groups; selected word TF-IDF SGD model
proxy_v1: protocol transfer	12	516	43	0.335271	0.335271318	0.329845127	0.329845	12 shared labels; earlier prompts and Kaggle proxy generation route
train_regen: reconstructed-prompt transfer	18	252	14	0.18254	0.182539683	0.159512223	0.159512	Only 14 complete clean prompt blocks; exploratory and low precision

Table 2. Validation and test metrics

Table 2 presents the detailed validation and test metrics for the three evaluated feature representations. Word TF-IDF achieves the highest test accuracy (0.259) and macro-F1 (0.254), while character TF-IDF gives a similar but slightly lower test accuracy (0.243) and the highest top-three accuracy (0.517). The engineered stylometric representation performs less well, with test accuracy of 0.170 and macro-F1 of 0.163. The pattern suggests that lexical sequences and character-level patterns carry more discriminative information than the compact set of manually engineered style statistics, although none of the representations achieves high reliability for 18-class attribution on unseen prompts.

Table 3 reports per model precision, recall, F1-score, and support for the 630-passage test set. Each of the 18 model classes contributes 35 test passages, so the class supports are balanced. gemini-3.1-pro-preview has the strongest class level result (F1 = 0.500; recall = 0.514), followed by qwen3.8-max (F1 = 0.371) and claude opus-5 and gpt-5.5 (F1 = 0.338 each). In contrast, gpt-5.6-sol has the lowest F1-score (0.097). The overall macro-F1 is approximately 0.254, showing substantial variation in how well individual labels are recognised and indicating that the model-specific signals are not equally distinctive.

model_features	selected_alpha	validation_accuracy	validation_macro_f1	test_accuracy	test_balanced_accuracy	test_macro_precision	test_macro_recall	test_macro_f1	test_weighted_f1	test_top3_accuracy
Word TF-IDF	0.0001	0.23175	0.22494	0.25873	0.25873	0.25387	0.25873	0.25362	0.25362	0.51111
Character TF-IDF	1.00E-05	0.24127	0.24223	0.24286	0.24286	0.25548	0.24286	0.24385	0.24385	0.51746
Stylometric features	0.0001	0.20952	0.20503	0.16984	0.16984	0.17621	0.16984	0.16289	0.16289	0.43968

Table 3. Per-model precision, recall, F1-score

class_or_average	precision	Recall	f1-score	support
claude-fable-5	0.13793	0.11429	0.125	35
claude-opus-4-6	0.2439	0.28571	0.26316	35
claude-opus-4-8	0.2069	0.17143	0.1875	35
claude-opus-5	0.30952	0.37143	0.33766	35
claude-sonnet-5	0.2093	0.25714	0.23077	35
gemini-3.1-pro-preview	0.48649	0.51429	0.5	35
gemini-3.6-flash	0.30233	0.37143	0.33333	35
gemini-3.7-flash	0.27273	0.17143	0.21053	35
gemini-3.8-flash	0.26667	0.34286	0.3	35
glm-5.3	0.175	0.2	0.18667	35
glm-5.3-flash	0.21739	0.14286	0.17241	35
gpt-5.5	0.30952	0.37143	0.33766	35
gpt-5.6-luna	0.24242	0.22857	0.23529	35
gpt-5.6-sol	0.11111	0.08571	0.09677	35
gpt-5.6-terra	0.19512	0.22857	0.21053	35
gpt-6-astra	0.33333	0.28571	0.30769	35
kimi-k3	0.17857	0.14286	0.15873	35

qwen3.8-max	0.37143	0.37143	0.37143	35
accuracy	0.25873	0.25873	0.25873	0.25873
macro avg	0.25387	0.25873	0.25362	630
weighted avg	0.25387	0.25873	0.25362	630

Table 4. Validation and test results for the feature-representation comparison

Table 4 consolidates the validation and test results for the feature representation comparison. Word TF-IDF provides the strongest overall test performance by accuracy, balanced accuracy, and macro-F1, whereas character TF-IDF yields the best top-three accuracy (0.517), marginally exceeding the word-based representation (0.511). Stylometric features alone show the weakest performance across the principal test metrics. Because the test partition is balanced across 18 classes, accuracy and balanced accuracy coincide. The results support using text-derived lexical and character features as useful attribution signals, while also demonstrating that their predictive value remains limited.

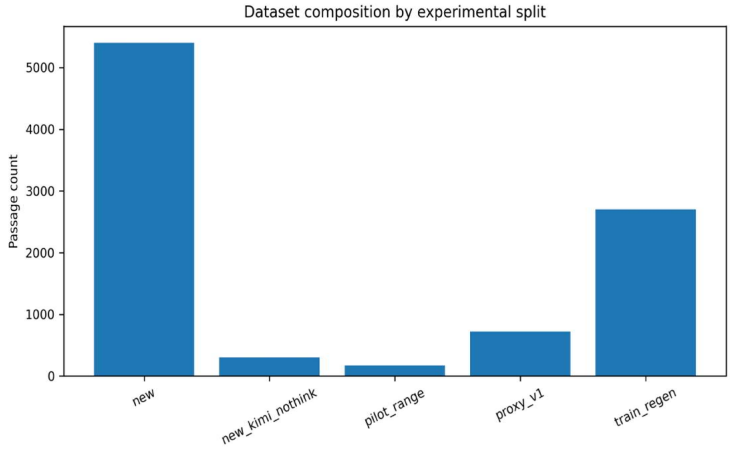

Figure 1. Passage counts by experimental split

Figure 1 visualises the number of passage records available in each experimental split. The new split contains the largest number of passages (5,400), followed by *train_regen* (2,700), *proxy_v1* (720), *new_kimi_nothink* (300), and *pilot_range* (170). These totals describe the raw split composition rather than the number of passages retained for every downstream experiment. Differences in split size and protocol are important when interpreting later model comparisons, because the primary analysis uses clean, complete prompt blocks rather than all records in the archive.

Dataset composition and quality control

The supplied archive contained 9,290 passage records across five experimental splits, covering 510 unique prompt identifiers and 18 model labels. The primary attribution experiment was restricted to the ‘new‘ split and to clean, complete prompt blocks. This yielded 4,158 passages from 231 prompts, with all 18 model labels represented for each retained prompt. To prevent prompt-content leakage, prompt identifiers; not individual passages were allocated to training, validation, and test partitions. The fixed split assigned 161 prompts to training, 35 to validation, and 35 to testing, corresponding to 2,898, 630, and 630 passages, respectively. Model identity, prompt identifiers, topic metadata, provider route, timestamps, token counts, provenance fields, and quality control flags were excluded from the predictor matrix; classification used passage text or text derived stylometric features only.

5.2 Model-attribution performance

Figure 2 compares held-out macro-F1 across word TF-IDF, character TF-IDF, and engineered stylometric

features. Word TF-IDF ranks first (macro-F1 = 0.254), closely followed by character TF-IDF (0.244), while stylometric features lag behind (0.163). The relatively small difference between the two text vectorisation approaches suggests that both lexical and character sequences encode useful model associated patterns. The weaker stylometric result indicates that global summary measures alone do not capture the full range of distinctions available in the generated passages.

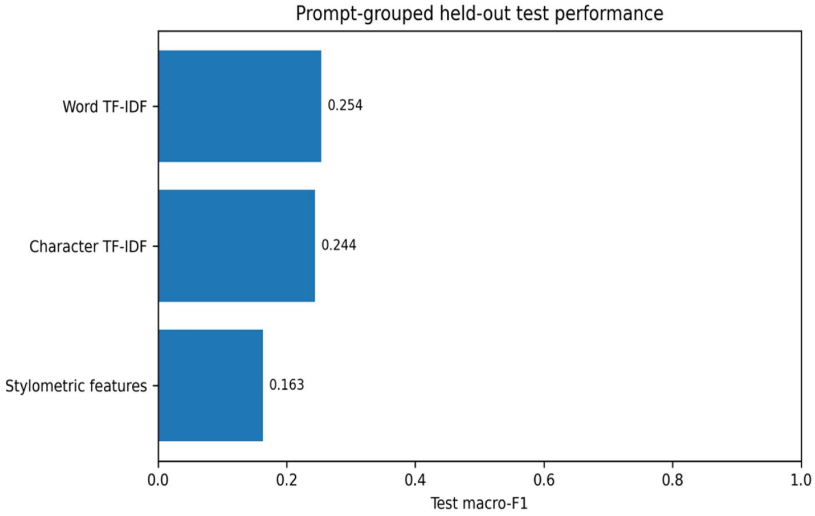

Figure 2. Test macro-F1 by feature representation

The highest held-out test macro-F1 among the evaluated feature families was obtained by Word TF-IDF (accuracy = 0.259; balanced accuracy = 0.259; macro-F1 = 0.254; weighted-F1 = 0.254). Its validation accuracy and macro-F1 were 0.232 and 0.225, respectively. Top-three accuracy was 0.511. Because the task contains 18 balanced classes in the primary test partition, uniform random selection would yield an expected accuracy of approximately 0.056; the observed classifier therefore extracted information beyond chance, but the absolute accuracy remained modest. Character n-grams produced a test macro-F1 of 0.244; engineered stylometric features alone produced 0.163. This pattern indicates that fine-grained orthographic and lexical sequences carried more attribution information than the compact set of manually engineered style variables, while no evaluated representation supported highly reliable 18-way attribution on unseen prompts.

5.3 Linguistic fingerprints

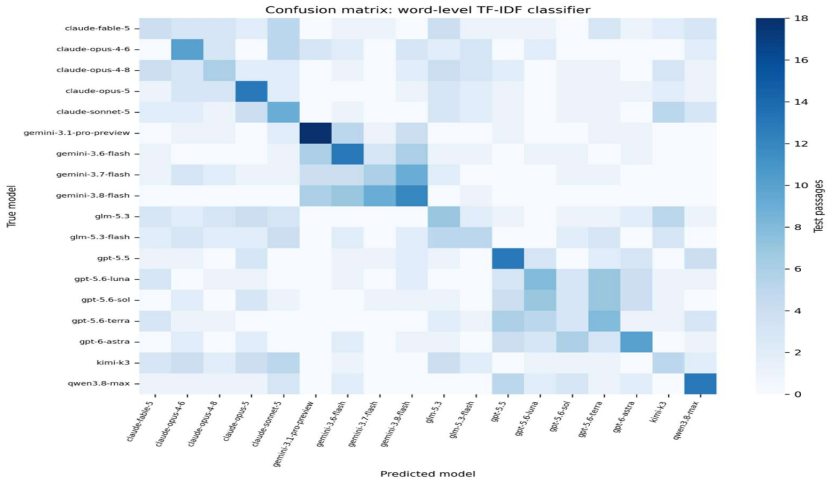

Figure 3. Confusion matrix on the held-out prompt-grouped test set

Figure 3 shows the distribution of correct and incorrect predictions across the 18 model labels for the held-out, prompt grouped test set. The diagonal cells represent correctly attributed passages, while off-diagonal cells identify model pairs that the classifier confuses. The limited overall accuracy and the variation in per-class F1-scores indicate that errors are distributed across labels rather than being confined to a single class. The matrix should therefore be interpreted as evidence of partial, uneven separability among model outputs, not as proof that individual passages can be attributed with certainty.

Nonparametric Kruskal–Wallis tests with Benjamini–Hochberg false-discovery-rate correction identified 25 of 35 engineered text features with statistically detectable differences across model labels at adjusted $p < .05$. Among the largest descriptive effects were *curly_apostrophe_share, emdash_per100w, mean_sentence_words*, *max_sentence_words, mean_word_chars*. These results establish that the sampled model outputs differ in measurable surface characteristics, but the descriptive eta-squared values should not be interpreted as causal effects or as evidence that any one feature uniquely identifies a model. The feature tests were performed over generated passages; dependence among passages sharing a prompt means that conventional passage level p-values may be optimistic. The results are therefore best interpreted as exploratory fingerprint evidence, supplemented by the prompt grouped held out predictive evaluation.

5.4 Generalization across protocols

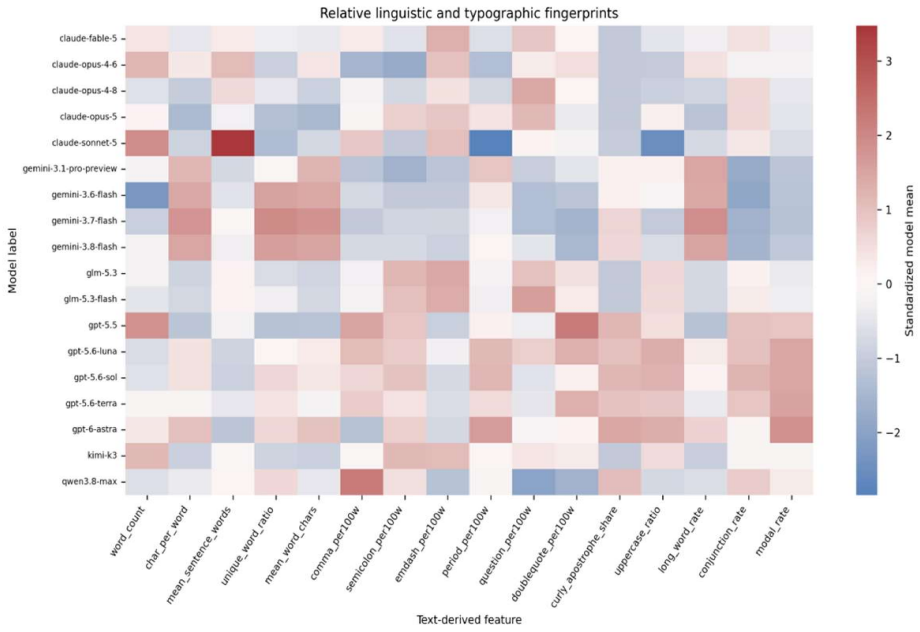

Figure 4. Standardized model-level means of text-derived features

Figure 4 compares standardized model-level averages for the engineered text-derived features. Standardisation places features measured on different scales onto a common scale, allowing relative profile differences to be inspected across model labels. The figure complements the nonparametric feature tests by showing how models vary across characteristics such as punctuation usage, apostrophe style, and sentence or word length. These patterns are descriptive fingerprints: they may help explain model-associated variation, but they do not establish causal effects or guarantee reliable identification of a model.

The word-level TF-IDF classifier was additionally evaluated without retraining or tuning on '*proxy_v1*', restricted to the 12 labels shared with 'new'. This protocol-transfer test used 516 passages from 43 complete clean prompts and yielded accuracy = 0.335, macro-F1 = 0.330, and weighted-F1 = 0.330. A second transfer evaluation on '*train_regen*' used only 252 passages from 14 complete clean prompt blocks; its accuracy was 0.183 and macro-F1 was 0.160. The latter estimate is highly uncertain because the test set contains only 14

prompt groups and uses community-reconstructed competition prompts. These cross-split values are stress tests under protocol and prompt-distribution shift, not directly comparable estimates from identically sampled test sets. '*proxy_v1*' was generated earlier through different serving routes and with a smaller set of accessible models; 'train_regen' used reconstructed prompts and is not the original competition test set.

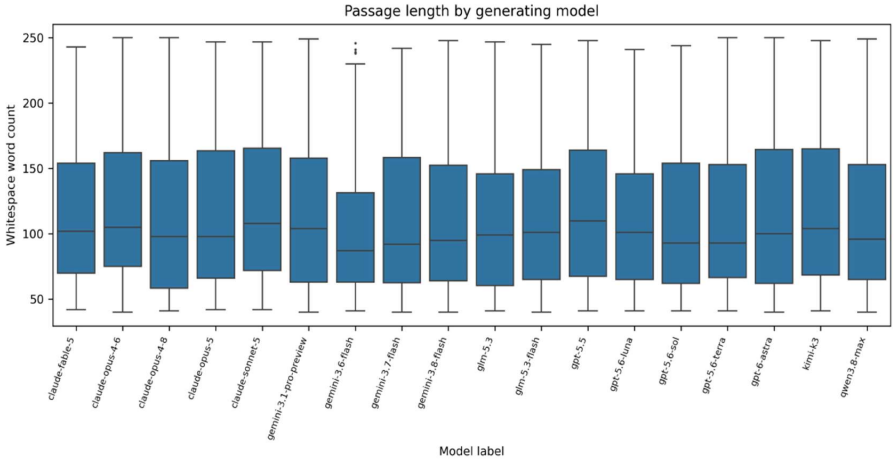

Figure 5. Passage word-count distributions by model

Description. Figure 5 displays the distribution of passage lengths across model labels. Comparing distributions rather than only averages helps reveal differences in central tendency and spread, as well as potential model-specific variation in generated passage length. In this study, per-generation word targets were assigned to reduce the confounding influence of a shared length instruction. Consequently, any remaining differences should be interpreted in the context of those targets and the quality-control filters, rather than assumed to represent an invariant property of each underlying model.

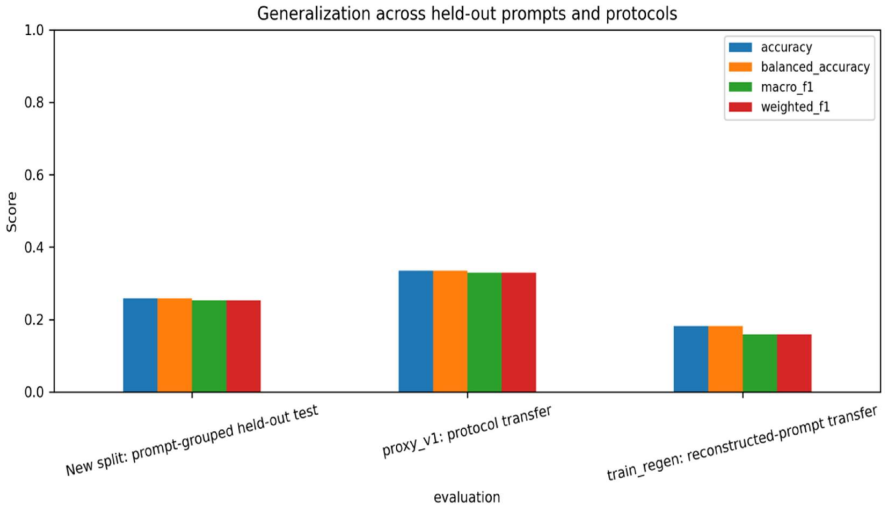

Figure 6. Primary test and external protocol-transfer performance

Figure 6 contrasts the primary prompt-grouped test result with the two external protocol-transfer evaluations. The primary new-split test reaches an accuracy of 0.259 and macro-F1 of 0.254; *proxy_v1* yields higher scores (accuracy 0.335; macro-F1 0.330), whereas *train_regen* produces lower scores (accuracy 0.183; macro-F1 0.160). The scores arise from different prompt collections, model coverage, and generation protocols, and the *train_regen* estimate is especially uncertain because it contains only 14 complete prompt groups. The figure therefore illustrates sensitivity to distribution and protocol changes, not a simple ranking under identical conditions.

6. Discussion

The findings support a qualified conclusion: AI-generated passages in this archive contain detectable model-associated linguistic signals, but these signals do not yield robust high accuracy identification across unseen prompts using the tested light weight classifiers. The best prompt grouped test macro-F1 was 0.254, substantially above the 18-class chance baseline yet far from error free attribution. This distinction matters for practical deployment. A classifier that performs above chance can still misidentify a substantial share of individual passages, so its outputs should not be treated as definitive proof of provenance.

The contrast between character level and engineered stylometric models suggests that model-associated information is distributed across local spelling, punctuation, affixes, and short character sequences rather than being captured fully by a small set of global statistics such as sentence length or punctuation rates. However, character patterns can also reflect provider specific serialization, typography, model family conventions, and generation settings. The archive documents pronounced differences in curly apostrophe and quotation mark use among model families, while several labels were served through more than one route. Consequently, attribution may partly identify a generation stack or register rather than an invariant underlying model identity. Follow up experiments should ablate punctuation and typography, normalize Unicode, and evaluate within provider as well as cross provider conditions.

The prompt grouped split is a key methodological strength because every passage answering a prompt was kept in a single partition. This prevents near identical prompt semantics from crossing training and test boundaries and gives a more credible estimate of generalization to new prompt topics than random passage splitting. Nevertheless, the experiment still uses prompts drawn from the same ‘new‘ prompt collection and broadly similar generation instructions. Generalization to other domains, longer texts, multilingual outputs, edited passages, human authored text, or future model versions remains unestablished.

The external split evaluations demonstrate why protocol separation matters. ‘*proxy_v1*‘ differs in prompt bank, date, available model coverage, and serving route; ‘*train_regen*‘ uses community-reconstructed rather than original competition prompts and has very few complete clean blocks. These evaluations should be reported as exploratory distribution shift checks, not pooled with the primary test score. Differences may arise from changed prompt content, target length instructions, provider routing, reasoning settings, or register drift. The dataset’s ‘*new_kimi_nothink*‘ and *‘pilot_range*‘ splits also encode deliberate protocol changes and were excluded from the primary 18-class model because they do not provide comparable complete 18-model blocks.

Several limitations constrain the scope of inference. First, all passages are AI-generated; the dataset cannot validate human versus AI detection. Second, the clean dataset is a selected subset created by a quality funnel, so the measured results may not represent all raw generations. Third, class labels are balanced by design in complete prompt blocks, unlike real world traffic. Fourth, the number of prompts not simply the number of passages is the effective basis for independent generalization, and only 35 prompt groups were held out in the primary test. Fifth, model versions, API defaults, routing, and post processing can change over time. Finally, stylistic significance tests are exploratory and passage level dependence may understate uncertainty.

For future work, we recommend repeated prompt grouped splits or grouped cross validation, uncertainty intervals obtained by boot strapping prompt groups, external validation on independently collected texts, typography normalized ablations, calibration analysis, and explicit testing of cross provider transfer. A practical provenance system should include an abstention option, and report calibrated probabilities or uncertainty rather than assigning a forced model label to every passage. Overall, this archive is useful for studying relative model fingerprints and failure modes, but the current evidence supports probabilistic attribution research not forensic certainty.

Feature Representation	Validation Accuracy	Test Accuracy	Balanced Accuracy	Macro-F1	Top-3 Accuracy
Word TF-IDF	0.232	0.259	0.259	0.254	0.511
Character TF-IDF	0.241	0.243	0.243	0.244	0.517
Stylometric features	0.210	0.170	0.170	0.163	0.440

Conclusion

The empirical evaluation confirms that AI-generated passages contain detectable, model associated linguistic signals, but current light weight classifiers cannot achieve error free, forensic level attribution on unseen prompts.

• Feature Efficacy: Lexical and character level patterns (Word and Character TF-IDF) significantly outperform engineered stylometric features, indicating that model "fingerprints" are distributed across local spelling, punctuation, and short character sequences rather than captured by global summary statistics like sentence length.

• Performance Limits: While the best performing model (Word TF-IDF) achieved a Macro-F1 of 0.254 and an accuracy of 25.9% substantially above the 5.6% random baseline this still represents a high error rate for individual passage attribution.

• Interpretation of Signals: The detected signals may partly reflect provider-specific serialization, typography conventions (e.g., curly apostrophes, em dashes), and generation stacks rather than an invariant underlying model identity.

• Future Directions: The study concludes that practical provenance systems must move away from forced, deterministic labeling. Future research should incorporate abstention options, calibrated probability reporting, typography normalized ablations, and cross provider validation to advance from probabilistic fingerprinting toward reliable text forensics.

References

[1] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Amodei, D. (2020). Language models are few shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, H.T. Lin (Eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020*. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

[2] Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., Sedova, K. (2023). *Generative language models and automated influence operations: Emerging threats and potential mitigations* (arXiv:2301.04246). arXiv. https://arxiv.org/abs/2301.04246.

[3] Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., Smith, N. A. (2021). All that's "human" is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (p. 7282–7296). *Association for Computational Linguistics*. https://doi.org/10.18653/v1/2021.acl-long.565.

[4] Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., Feizi, S. (2023). *Can AI-generated text be reliably detected* (arXiv:2303.11156). arXiv. https://arxiv.org/abs/2303.11156.

[5] Liu, Z., Yao, Z., Li, F., Luo, B. (2023). *On the detectability of ChatGPT content: Benchmarking, methodology, and evaluation through the lens of academic writing* (arXiv:2306). arXiv.

[6] Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., Pearson, A. T. (2022). Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. *bioRxiv*. Advance online publication.

[7] Solaiman, I., Talat, Z., Agnew, W., Ahmad, L., Baker, D., Blodgett, S. L., Chen, C., Daumé, H., III, Dodge, J., Duan, I., et al. (2023). *Evaluating the social impact of generative AI systems in systems and society* (arXiv:2306.05949). arXiv. https://arxiv.org/abs/2306.05949.

[8] Vidgen, B., Agrawal, A., Ahmed, A. M., Akinwande, V., Al-Nuaimi, N., Alfaraj, N., Alhajjar, E., Aroyo, L., Bavalatti, T., Blili-Hamelin, B., et al. (2024). *Introducing v0.5 of the AI Safety Benchmark from MLCommons* (arXiv:2404.12241). arXiv. https://arxiv.org/abs/2404.12241.

[9] Shu, K., Wang, S., Lee, D., Liu, H. (2020). Mining disinformation and fake news: Concepts, methods, and recent advancements. In *Disinformation, misinformation, and fake news in social media: Emerging research challenges and opportunities* (p. 1–19). Springer.

[10] Chen, C., Shu, K. (2024). Combating misinformation in the age of LLMs: Opportunities and challenges. *AI Magazine*. Advance online publication. https://doi.org/10.1002/aaai.12188.

[11] Chen, C., Wang, H., Shapiro, M., Xiao, Y., Wang, F., Shu, K. (2022). *Combating health misinformation in social media: Characterization, detection, intervention, and open issues* (arXiv:2211.05289). arXiv. https://arxiv.org/abs/2211.05289.

[12] Hanley, H. W. A., Durumeric, Z. (2024). Machine-made media: Monitoring the mobilization of machine-generated articles on misinformation and mainstream news websites. *Proceedings of the International AAAI Conference on Web and Social Media, 18*, 542–556.

[13] Meyer zu Eissen, S., Stein, B. (2006). Intrinsic plagiarism detection. In *Proceedings of the European Conference on Information Retrieval (ECIR-06)* (p. 1–4).

[14] Stamatatos, E., Koppel, M. (2011). Plagiarism and authorship analysis: Introduction to the special issue. *Language Resources and Evaluation, 45*, 1–4.

[15] Ott, M., Choi, Y., Cardie, C., Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (p. 309–319). Association for Computational Linguistics. https://aclanthology.org/P11-1032.

[16] Afroz, S., Brennan, M., Greenstadt, R. (2012). Detecting hoaxes, frauds, and deception in writing style online. In *2012 IEEE Symposium on Security and Privacy* (p. 461–475). IEEE.

[17] Kumarage, T., Liu, H. (2023). Neural authorship attribution: Stylometric analysis on large language models. In *2023 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)* (p. 51–54). IEEE. https://doi.org/10.1109/CyberC58899.2023.00019.

[18] Lagutina, K., Lagutina, N., Boychuk, E., Vorontsova, I., Shliakhtina, E., Belyaeva, O., Paramonov, I., Demidov, P.G. (2019). A survey on stylometric text features. In *2019 25th Conference of Open Innovations Association*

(FRUCT) (p. 184–195). IEEE.

[19] Fabien, M., Villatoro-Tello, E., Motlicek, P., Parida, S. (2020). BERTAA: BERT fine-tuning for authorship attribution. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)* (pp. 127–137). NLP Association of India. https://aclanthology.org/2020.icon-main.16.

[20] Rivera-Soto, R. A., Miano, O. E., Ordonez, J., Chen, B. Y., Khan, A., Bishop, M., & Andrews, N. (2021). Learning universal authorship representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (p. 913–919). *Association for Computational Linguistics*. https://doi.org/10.18653/v1/2021.emnlp-main.70.

[21] Kumarage, T., Garland, J., Bhattacharjee, A., Trapeznikov, K., Ruston, S., Liu, H. (2023). *Stylometric detection of AI-generated text in Twitter timelines* (arXiv:2303.03697). arXiv. https://arxiv.org/abs/2303.03697.

[22] Kumarage, T., Liu, H. (2023). Neural authorship attribution: Stylometric analysis on large language models. In *2023 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)* (p. 51–54). IEEE. https://doi.org/10.1109/CyberC58899.2023.00019.

[23] Mindner, L., Schlippe, T., Schaaff, K. (2023). Classification of human and AI-generated texts: Investigating features for ChatGPT. In *International Conference on Artificial Intelligence in Education Technology* (p. 152–170). *Springer*.

[24] Shah, A., Ranka, P., Dedhia, U., Prasad, S., Muni, S., Bhowmick, K. (2023). Detecting and unmasking AI-generated texts through explainable artificial intelligence using stylistic features. *International Journal of Advanced Computer Science and Applications, 14*(10).

[25] Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., Krueger, G., Kim, J. W., Kreps, S., et al. (2019). *Release strategies and the social impacts of language models* (arXiv:1908.09203). arXiv. https://arxiv.org/abs/1908.09203.

[26] Gehrmann, S., Strobelt, H., Rush, A. M. (2019). *GLTR: Statistical detection and visualization of generated text* (arXiv:1906.04043). arXiv. https://arxiv.org/abs/1906.04043.

[27] Mikros, G. K., Koursaris, A., Bilianos, D., Markopoulos, G. (2023). AI-writing detection using an ensemble of transformers and stylometric features. In *Proceedings of the IberLEF@SEPLN*.

[28] Huang, B., Chen, C., Shu, K. (2025). Authorship attribution in the era of LLMs: Problems, methodologies, and challenges. *ACM SIGKDD Explorations Newsletter, 26*(2), 21–43.