Stylometric Fingerprints of AI Rewriting: Punctuation, Syntax, and Model Attribution Across 60,786 Paired Texts

TextPulse Research. Working Paper.

Abstract

When a language model rewrites a human text, it changes more than the words. This study measures what occurs to the rest of the linguistic style. We compared 60,786 human-written academic texts with AI rewrites of the same texts produced by eight AI models across six model families, and computed 49 stylometric features for every text, namely, punctuation rates, passive voice, nominalization, first person, contractions, sentence openers, lexical diversity, word length, and a set of hedge and connective phrases. The paired design controls topic and content, so every change is caused by the model itself. The shifts are consistent. AI rewrites use longer words (Cohen's d = 1.57 for words of seven or more letters), more nominalizations (d = 1.12), higher lexical diversity (d = 1.31), more passive voice, and more commas, while first person, contractions, questions, and repeated sentence openers are mostly nonexistent. Each model has its own version of this profile. A multinomial logistic regression reading only the 49 features identifies which of six models produced a rewrite with 64.3 percent accuracy against a 16.7 percent chance rate, and a gradient boosting model raises this only to 66.4 percent, which shows the fingerprint is carried almost entirely by simple, interpretable features. DeepSeek is the most identifiable model and OpenAI is the least. Qwen changes the style of the human source text least and DeepSeek changes it most. Per-text feature data and all analysis code are made publicly available.

1. Introduction

Large language models are widely used to rewrite human text, and the question of what that rewriting does to writing style has mostly been answered with single observations. Individual marker words have been counted, and single properties such as sentence-length variation have been measured. An investigation on what rewriting does across the full range of stylometric properties, measured on the same texts, is lacking.

This paper aims to fill this gap. It is the third study in a series built on one paired corpus: 60,786 human-written academic texts, each with an AI generated rewrite of the same text produced by one of eight AI model configurations across six model families. The first study in the series measured vocabulary [7] and the second measured sentence-length burstiness [8]. This study measures the remaining stylometric properties: punctuation, passive voice, nominalization, first person, contractions, questions, sentence openers, adverbs, word length, lexical diversity, and hedge and

connective phrases. It then asks an important question. If every model changes style in its own way, can the AI model be identified from style alone?

Comparisons of unrelated human and AI corpora confuse the model with the topic, the register, and the era of the writing. Here every rewrite is compared with the human source of the same content, so any stylometric difference is caused by the model itself.

Three results are worth considering. First, the direction of AI rewriting is uniform and text becomes lexically heavier and grammatically more formal, while the direct cues of a human writer, first person, contractions, questions, and repeated sentence openers, are nonexistent. Second, several widely repeated claims about AI style are not supported by measurement in this corpus. The word “delve” is barely injected during rewriting, the word “crucial” is not injected at all, and human writers repeat their sentence openers 2.7 times more often than the models that supposedly write repetitively. Third, the per-model stylometric differences are large enough to act as a fingerprint. Style features alone identify the AI model at 3.9 times chance accuracy, and the most and least identifiable models differ by a factor of two in F1 score.

2. Related work

Word-frequency studies have established that AI-assisted writing has changed the vocabulary of published academic work. Liang et al. estimated that 6.5 to 16.9 percent of AI-conference peer review text was heavily modified by language models, based on the frequency shifts of characteristic tokens [1]. Kobak et al. tracked 15 million PubMed abstracts and found an abrupt surge in style words such as “delve” and “underscore” after 2022, and showed that at least 13.5 percent of 2024 abstracts were processed with a model [2]. Juzek and Ward identified 21 focal words that models overrepresent and reported evidence that reinforcement learning from human feedback (RLHF) contributes to the overuse [3]. These studies count words in large uncontrolled corpora. The present study measures style in a paired corpus, where each AI text is a rewrite of a known human-written source.

Apart from vocabulary, Muñoz-Ortiz et al. compared human news text with output from six models and found that human texts show more scattered sentence-length distributions, different dependency structures, and stronger negative emotions [4]. Their comparison uses model-generated text on similar topics rather than AI rewrites of identical sources.

Attributing a text to its author from style features is has been extensively studied. Mosteller and Wallace attributed the disputed Federalist papers from function-word rates in 1964 [5], and Stamatatos surveyed the modern methods used [6]. Uchendu et al. first investigated the attribution question for neural text generators, and asked which of several models produced a text [9], and Huang et al. survey the LLM-era version of the same problem [10]. The present study contributes a controlled attribution testbed whereby six AI models rewrite identical human sources, so the attribution signal cannot come from the topic or content.

In our prior work, a vocabulary study established per-family word injection and deletion lexicons [7], and a per-doc burstiness study established that all eight flagship AI models flatten sentence-length variation, with the flattening varying eight times in strength across these models [8]. Both features are considered in the set used in this study.

3. The stylometric profile of human writing

Stylometry describes a text through measurable surface properties rather than through its content. The main families of these properties are lexical features such as word length and vocabulary richness, syntactic features such as sentence length and its variation, passive voice, and nominalization, punctuation habits, voice markers such as first person, contractions, and questions, and recurring phrases. Figure 1 organizes the properties measured in this paper. Human writers differ from one another on all of these properties, and those stable individual differences are what made classical authorship attribution possible long before language models existed [5, 6].

Human academic writing has a recognizable stylometric character. In this corpus, the human sources mix short and long sentences freely, with a mean sentence-length coefficient of variation of 0.449 [8]. They open consecutive sentences with the same word 7.2 times per 100 sentence transitions, a natural repetitiveness that comes from topic continuity. They use first person 4.3 times per 1,000 words, ask an occasional question, and even in formal register keep a small number of contractions. Their vocabulary is moderately heavy, with 357 long words and 58 nominalizations per 1,000 words, and their hedging templates such as “in conclusion” appear at low but steady rates. Around these averages the variation between individual authors is wide. Human writing is heterogeneous, and comparable spread has been reported for human news text as well [4].

AI writing occupies a narrower band. As the following sections measure in detail, the rewrites in this corpus converge on a uniform register that is lexically heavier, more nominal, more passive, and stripped of the voice markers listed above. The distinction matters for context. The properties in Figure 1 are not arbitrary detector inputs. They are the same properties that have always distinguished one human writer from another, and AI rewriting shifts nearly all of them in one direction at once.

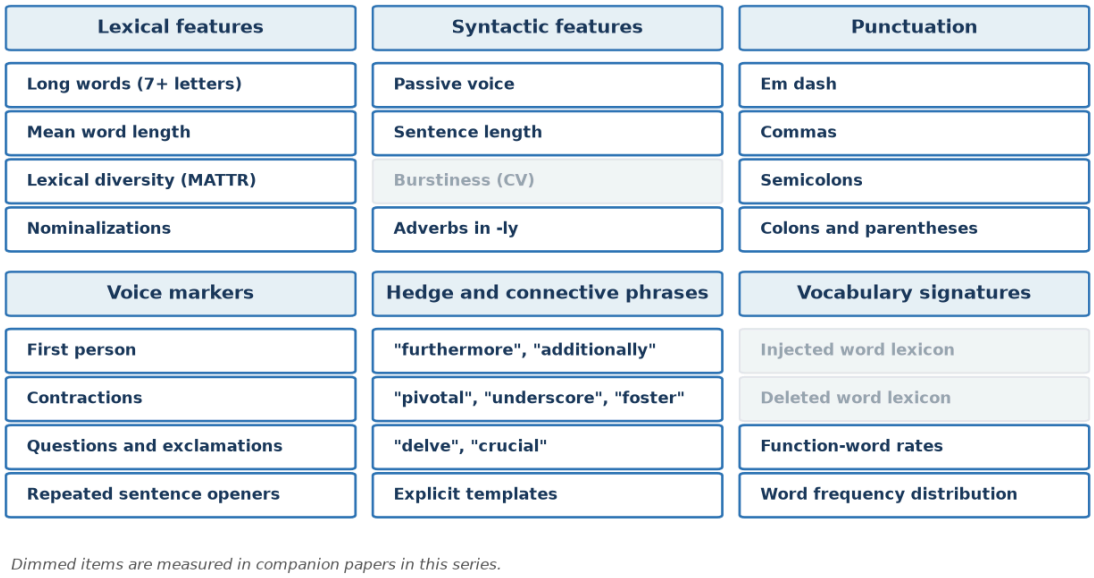

Figure 1. Stylometric properties measured in this paper, grouped by family.

4. Data and methods

4.1 Corpus

The corpus is the same collection of human-AI paraphrase pairs used in the previous two studies, built from open-access academic text. Corpus 1 holds 38,323 pairs rewritten by two production configurations (a DeepSeek chat configuration and a Gemini Flash Lite configuration). Corpus 2 holds 22,463 pairs in which identical human sources were rewritten by six models, namely, DeepSeek, Gemini, Grok, Mistral, OpenAI, and Qwen. They provide 60,786 pairs and 121,572 texts in total. Human sources are on average 343 words, and their AI rewrites 321 words. Cross-model comparisons and the attribution experiment use corpus 2 only, because its six models received identical inputs.

4.2 Features

We computed 49 features per each text. Twenty-four are structural: word and sentence counts, mean word length, the rate of words with seven or more letters, type-token ratio, moving-average type-token ratio (window 50), sentence-length mean, standard deviation, and coefficient of variation, six punctuation rates per 1,000 words (em dash, en dash, semicolon, colon, parenthesis, comma), passive voice rate, nominalization rate (words ending in -tion, -sion, -ment, -ness, -ity), first-person rate, the share of sentences that open with the same word as the previous sentence, the share of sentences that open with a connective, contraction rate, question mark and exclamation mark rates, and the rate of adverbs ending in -ly. The remaining 25 features are the per-1,000-word rates of hedge and connective phrases such as “it is important to note”, “in conclusion”, “moreover”, “furthermore”, “pivotal”, and “foster”. Passive voice, nominalization, and the -ly

adverb rate are pattern-based approximations, which are computed identically on both sides of each pair, so paired comparisons are unaffected by any approximation.

The human sources in this corpus contain no em or en dashes at all, because dash characters were normalized to plain hyphens when the corpus was constructed. The em dashes measured in the rewrites are therefore genuine insertions by the models, but the human baseline of zero is a property of the corpus construction, not of the human authors. We accordingly exclude the two dash features from all human vs. AI comparisons and use them only as differences between models on the AI side.

4.3 Statistics

All human vs. AI results are paired. For each feature we report the human mean, the AI mean, the mean paired delta with a bootstrap 95 percent confidence interval (2,000 resamples), Cohen's d computed on the paired deltas, and the percentage of pairs in which the feature increased. With 60,786 pairs, every delta reported in Section 5 has a confidence interval that excludes zero.

4.4 Attribution experiment

The attribution question is as follows: Given only the 49 features of a rewrite, which of the six AI models produced it? We use multinomial logistic regression on standardized features (scikit-learn 1.9.0, L2 regularization, C = 1.0), a simple model whose coefficients can be read directly to allow for an even playing field with all other variables constant. Accuracy is estimated with five-fold cross-validation grouped by pair, so no source text appears in both the training and test data. A held-out 20 percent split (4,493 texts) provides the confusion table. To assess how much signal a simple linear model misses, we also train a gradient boosting classifier on the same splits as a non-linear ceiling. Chance accuracy is 16.7 percent, and always predicting the largest class yields 24.8 percent.

5. What rewriting does to style

5.1 Text becomes lexically heavier

The largest changes in the corpus are all in the direction of heavier vocabulary. The rate of long words (seven+ letters) increases from 357 to 444 per 1,000 words (d = 1.57), and it increases in 96.4 percent of all pairs. Mean word length increases from 5.51 to 6.12 characters (d = 1.54). Nominalizations increase from 58 to 75 per 1,000 words (d = 1.12). Lexical diversity rises as well. The moving-average type-token ratio increases from 0.785 to 0.827 (d = 1.31), in 92.3 percent of pairs. Adverbs in -ly increase from 12.3 to 17.9 per 1,000 words (d = 0.60). The models also compress in general. Rewrites are on average 22 words shorter than their sources while containing slightly more sentences, which shortens the mean sentence from 26.8 to 24.8 words.

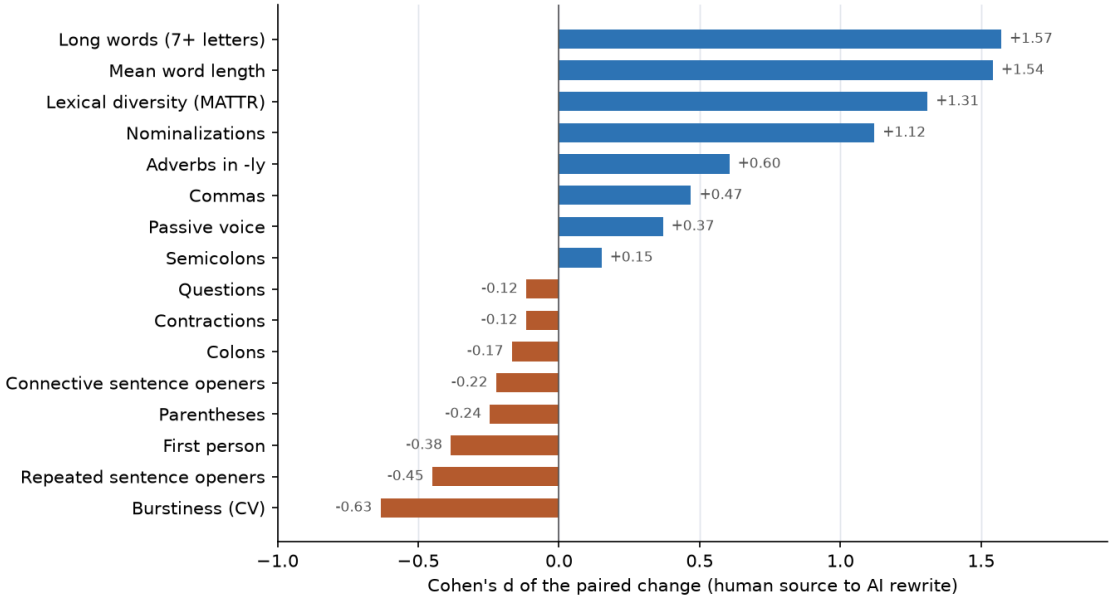

Figure 2. Effect size (Cohen's d) of the paired change for each structural feature, ordered from largest increase to largest decrease.

Figure 2 summarizes the direction and size of every structural change. The pattern is clearly one-sided. Everything that makes text denser and more formal increases. Everything that shows a person speaking decreases.

5.2 The writer's voice is removed

Four features show the direct presence of a writer: first person, contractions, questions, and exclamations. All four are heavily removed in rewrites. First person falls from 4.34 to 2.19 occurrences per 1,000 words, a 50 percent reduction, and increases in only 9.0 percent of pairs. Contractions fall from 0.32 to 0.15 per 1,000 words and increase in 0.9 percent of pairs, the most one-sided change in the corpus. Questions fall from 0.34 to 0.22 per 1,000 words and exclamation marks from 0.033 to 0.020.

A fifth marker behaves against expectation. Human writers open consecutive sentences with the same word 7.2 times per 100 sentence transitions. The rewrites do so 2.7 times per 100, a 63 percent reduction (d = -0.45). Repetitive sentence openings are commonly described as an AI trait. In this corpus the opposite is true. Repetition of openers is a human trait, and models remove it. The share of sentences opened with a connective also decreases, from 7.5 to 5.6 percent.

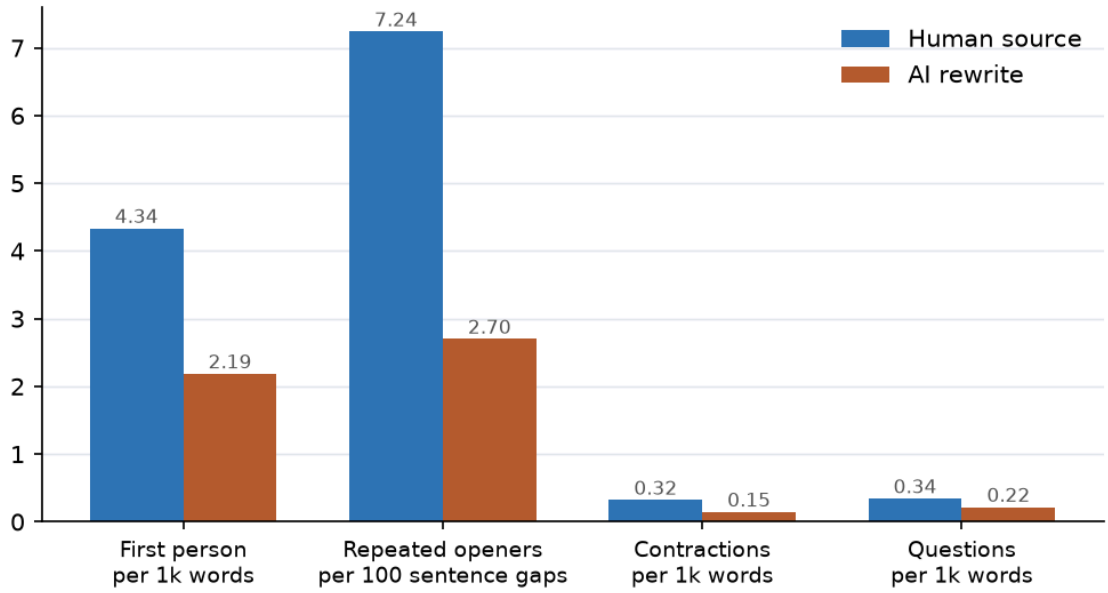

Figure 3. Voice markers before and after rewriting: first person, repeated sentence openers, contractions, and questions.

5.3 Punctuation

Commas increase from 62.7 to 71.2 per 1,000 words (d = 0.47), consistent with the denser, more subordinated sentences described above. Semicolons increase from 3.87 to 4.47. Colons fall from 1.98 to 1.50 and parentheses decrease from 15.6 to 14.0 per 1,000 words.

The em dash requires a separate account. Since the corpus construction removed all dashes from the human sources (Section 4.2), every em dash in a rewrite is an insertion by the model. The insertions are significant and extremely uneven. Across the corpus the rewrites contain 4.06 em dashes per 1,000 words. Across the six corpus-2 models on identical inputs, OpenAI inserts 3.02 per 1,000 words, Mistral 2.57, Grok 0.93, Qwen 0.40, Gemini 0.11, and DeepSeek 0.015, a 200-fold spread. The two corpus-1 production models insert more, namely, 6.98 (DeepSeek chat) and 4.65 (Gemini Flash Lite) per 1,000 words. The comparison between the two DeepSeek entries is instructive. The same model family inserts em dashes at 6.98 per 1,000 words in one configuration and 0.015 in another. The em dash is a strong AI signal, but it is a property of a model configuration, and not of a model family, and its absence proves nothing.

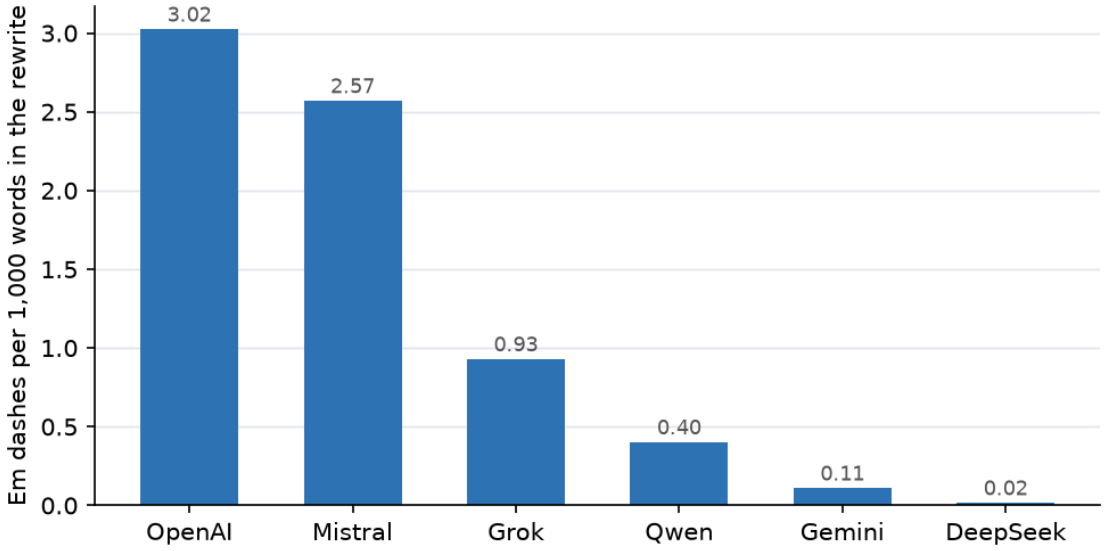

Figure 4. Em dashes inserted per 1,000 words by each corpus-2 model on identical inputs.

5.4 Hedge and connective phrases

The phrase results are split into three groups. The first group is strongly injected: “furthermore” triples from 0.28 to 0.83 per 1,000 words, “foster” rises six-fold from 0.06 to 0.37, “pivotal” eleven-fold from 0.015 to 0.17, “underscore” and “underscores” together roughly fifteen-fold from 0.017 to 0.25, “notably” increases more than three times, “additionally” doubles, and “multifaceted” increases over four times. The second group is flat, where “crucial” increases from 0.152 to 0.158 per 1,000 words, “moreover” declines slightly, and “delve” in all its forms stays below 0.01 per 1,000 words on both sides. The third group declines. The explicit templates “it is important to note”, “it is worth noting”, “in conclusion”, “in summary”, and sentence-initial “overall,” are all used less by the models than by the human authors.

The second and third groups matter for practice. “Delve” and “crucial” are the two most cited AI marker words, and neither is meaningfully injected when models rewrite academic text. Word lists compiled from chatbot answers or from text generated from scratch do not transfer directly to the rewriting setting. The words that do get injected in rewriting, “pivotal”, “underscore”, “foster”, “furthermore”, are related but not identical, consistent with the task-dependence of model vocabulary documented in the earlier studies [2, 3, 7].

6. Model fingerprints

6.1 Which model changes your voice least

For each corpus-2 model we computed a style distance. The mean absolute paired change across the 22 structural features (dashes excluded) was used, each standardized by the corpus-wide

spread of that change. The ranking is as follows: Qwen 0.48, OpenAI 0.63, Mistral 0.75, Gemini 0.81, Grok 0.83, DeepSeek 1.02. Qwen changes the style of the source text roughly half as much as DeepSeek does. The ranking is consistent with our prior study on burstiness, where DeepSeek also flattened sentence-length variation most and left 98.5 percent of its rewrites flatter than the source [8].

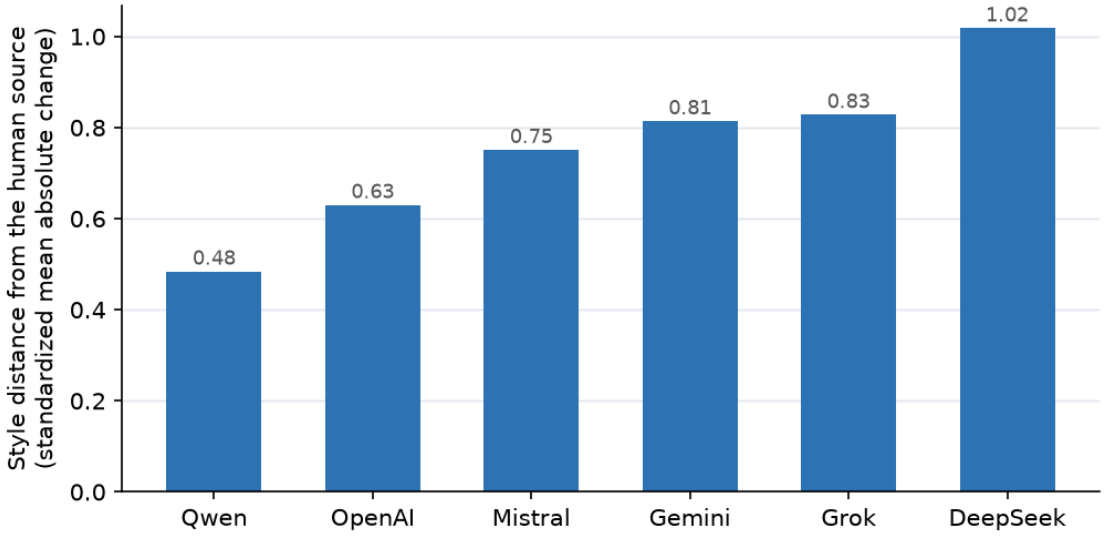

Figure 5. Style distance from the human source by model: standardized mean absolute change across 22 structural features.

6.2 Style alone identifies the model

The multinomial logistic regression identifies the rewriting model with 64.3 percent accuracy (five-fold grouped cross-validation, standard deviation 0.4 percentage points), against 16.7 percent chance and a 24.8 percent majority baseline. The gradient boosting ceiling on the same splits is 66.4 percent. The gap of two percentage points between the interpretable linear model and the non-linear ceiling is the second result, where the model fingerprint is carried almost entirely by simple, individually meaningful features, not by interactions a linear model cannot see.

Per-model performance on the held-out test set (4,493 texts):

Model	Precision	Recall	F1	Test texts
DeepSeek	0.803	0.856	0.829	1,100
Grok	0.651	0.704	0.676	1,115
Mistral	0.553	0.564	0.559	968
Gemini	0.513	0.443	0.475	531
Qwen	0.481	0.462	0.471	409

Model	Precision	Recall	F1	Test texts
OpenAI	0.467	0.349	0.399	370

Confusion on the held-out test set (rows are the true model, columns the predicted model):

True model	DeepSeek	Gemini	Grok	Mistral	OpenAI	Qwen
DeepSeek	942	53	55	27	10	13
Gemini	82	235	95	62	24	33
Grok	62	60	785	147	28	33
Mistral	61	52	179	546	56	74
OpenAI	9	32	38	111	129	51
Qwen	17	26	54	94	29	189

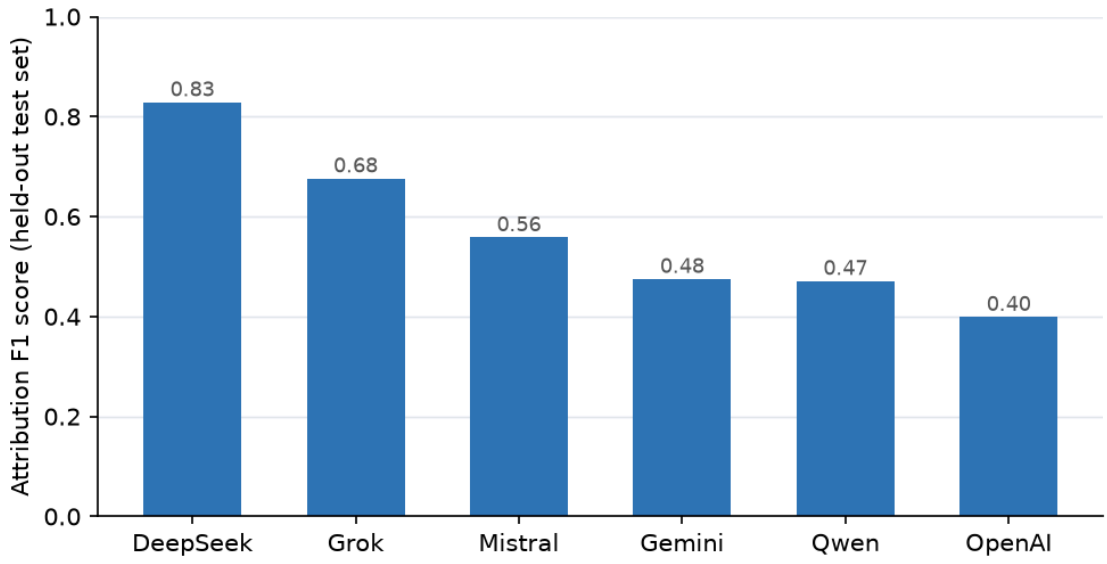

Figure 6. Attribution F1 score per model on the held-out test set.

DeepSeek is the most identifiable model by a wide margin, and the identifiability ranking tracks the style-distance ranking. The more a model changes the source, the easier it is to detect. OpenAI, the second-least intervening model, is the hardest to attribute, and its errors flow mostly to Mistral. This is the practical property of the AI fingerprint. A model reveals itself exactly to the degree that it rewrites text.

6.3 What identifies each model

The regression coefficients identify each model's markers. DeepSeek is identified by aggressive flattening and strongly negative coefficients on sentence-length variation and the em dash. Gemini

is identified by inflating sentence counts while lowering diversity. Grok is identified by preserved sentence-length variation combined with a high rate of “moreover”. Mistral’s strongest marker is em-dash and en dash insertion. OpenAI is identified by producing fewer, longer texts with low sentence-length spread but frequent em dashes. Qwen is identified by restraint and fewer sentences, shorter words, and fewer -ly adverbs compared to other models. None of these markers requires machine learning to check, since each is a rate a reader can count.

7. Discussion

Across 60,786 pairs and every model tested, rewriting changes academic text toward a heavier, more formal register and longer words, more nominalizations, more passive voice, more commas, higher lexical diversity, while stripping out first person, contractions, questions, and the natural repetitiveness of human sentence openings. The result reads as more polished and less personal. The models do not share one fixed style, but they share this general direction, and each applies it with a characteristic pattern strong enough to identify the model at 3.9 times chance from surface features alone.

Two of the findings are worth noting. The most widely cited marker words are task-dependent: “delve” and “crucial”, the most repeated examples of AI vocabulary, are not meaningfully injected when models rewrite existing academic text, while “pivotal”, “underscore”, and “foster” are injected at rates up to fifteen times the human baseline. The claim that AI writing is repetitive is, at the level of sentence openers, backwards. Human authors repeat their openers 2.7 times more often than the AI rewrites do. Style checkers and detection heuristics built on the unmeasured versions of these claims will misdetect in both directions.

The em-dash result has a practical implication. The em dash is the strongest single punctuation signal in this corpus, but its rate varies 200-fold between configurations of the same model families, and 465-fold across all eight configurations. Any detection rule of the form “em dashes imply AI” is a rule about specific configurations at a specific time, not about AI text in general.

For attribution, the result supports two conclusions. Sixty-four percent six-way accuracy from 49 surface rates, with a non-linear maximum only two points higher, shows that model fingerprints are simple and stable when scaled across thousands of texts. It also shows the limit. A third of AI rewrites are misattributed even in this controlled setting, models change with every release, and the least intervening models are already close to the attribution floor. Style-based attribution can support forensic hypotheses, but it cannot carry individual verdicts, the same conclusion the burstiness study reached for detection [8].

8. Limitations and conclusion

The corpus is human-written academic content rewritten by AI models, so conclusions may not transfer to text generated from scratch, to other registers, or to other languages. Passive voice,

nominalization, and -ly adverb rates are pattern-based approximations rather than parses. The dash normalization in the corpus construction removes the human baseline for two features, which we handle by restricting those features to model-side comparisons only. The eight configurations were captured at fixed points in 2026, and model updates may change the rates reported here. The attribution experiment covers six models on identical inputs. Accuracy against a wider or unknown set of models would be lower, and the experiment does not address texts a model did not write.

AI rewriting moves academic text in one consistent direction across every model tested. It makes the text lexically heavier and grammatically more formal while removing the surface traces of a human author. Each model applies this transformation with its own measurable signature, strong enough that six models can be told apart at almost four times chance accuracy from surface rates alone, and the signature is carried by features anyone can count, such as word length, passive voice, em dashes, and sentence openers.

Data availability

Per-text stylometric features for all 121,572 texts (opaque pair identifiers; corpus, model, discipline, side, and all 49 features), the attribution and analysis code, and the figure code are openly available on Zenodo at https://doi.org/10.5281/zenodo.22040683.

References

[1] Liang, W., Izzo, Z., Zhang, Y., Lepp, H., Cao, H., Zhao, X., Chen, L., Ye, H., Liu, S., Huang, Z., McFarland, D. A., and Zou, J. Y. (2024). Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews. Proceedings of the 41st International Conference on Machine Learning (ICML 2024). https://arxiv.org/abs/2403.07183

[2] Kobak, D., González-Márquez, R., Horvát, E.-Á., and Lause, J. (2025). Delving into LLM-assisted writing in biomedical publications through excess vocabulary. Science Advances, 11, eadt3813. https://doi.org/10.1126/sciadv.adt3813

[3] Juzek, T. S., and Ward, Z. B. (2025). Why does ChatGPT “delve” so much? Exploring the sources of lexical overrepresentation in large language models. Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025), 6397-6411. https://aclanthology.org/2025.coling-main.426/

[4] Muñoz-Ortiz, A., Gómez-Rodríguez, C., and Vilares, D. (2024). Contrasting linguistic patterns in human and LLM-generated news text. Artificial Intelligence Review, 57(9). https://arxiv.org/abs/2308.09067

[5] Mosteller, F., and Wallace, D. L. (1964). Inference and Disputed Authorship: The Federalist. Addison-Wesley.

[6] Stamatatos, E. (2009). A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology, 60(3), 538-556. https://doi.org/10.1002/asi.21001

[7] TextPulse Research (2026). The vocabulary fingerprint of AI rewriting: common words AI language models prioritize. Working paper. https://textpulse.ai/research/ai-vocabulary-fingerprint. https://doi.org/10.5281/zenodo.22028377

[8] TextPulse Research (2026). Sentence-length burstiness as a cross-disciplinary and cross-model signal of AI rewriting. Working paper. https://textpulse.ai/research/ai-burstiness-sentence-length. https://doi.org/10.5281/zenodo.22033062

[9] Uchendu, A., Le, T., Shu, K., and Lee, D. (2020). Authorship attribution for neural text generation. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020). https://aclanthology.org/2020.emnlp-main.673/

[10] Huang, B., Chen, C., and Shu, K. (2025). Authorship attribution in the era of LLMs: problems, methodologies, and challenges. ACM SIGKDD Explorations Newsletter, 26(2), 21-43. https://arxiv.org/abs/2408.08946