Modelometry: Classification of Flagship AI Model Families Based on Inter-Model Stylometry

TextPulse Research. Working Paper.

Abstract

Studies that measure the output of more than one AI language model consistently find that the models write differently, in the same manner that human authors have different writing styles. We coin the term **modelometry** for the measurement and attribution of the inter-model writing style of AI systems, and **modelolect** for the style itself, formed on the pattern of idiolect and sociolect. The study asks whether the modelolect of a flagship model family is strong enough for a classifier that reads only surface stylometric features to classify the AI family that originally wrote a text. We test seven families (GPT, Claude, Gemini, DeepSeek, Grok, Llama and Qwen) against human text, in two registers (formal academic text, and informal chat text) and with three tiers of features. Every AI text used is from corpora generated for our prior studies or from open datasets. On academic text, a gradient boosting classifier over 47 interpretable stylometric features attributes 7 classes (human and six families) at 73.7% accuracy against a 14.3% chance rate, and 6 classes on a second corpus at 79.2%. On chat responses from the LMArena preference dataset, the same 47 features attribute all seven families at 63.9%, and a character n-gram model with model names masked reaches 87.0%, so the small interpretable feature set accounts for about three quarters of the attributable signal. The confusion structure is also informative, since DeepSeek is rarely confused with GPT (3% in the corpus where DeepSeek is most identifiable), so the writing-style evidence does not support the claim that 'DeepSeek behaves as a distillation of GPT'. The largest confusion in the chat register is between Qwen and GPT, at about 16%. Claude is the most identifiable family in both registers. A classifier trained on academic AI generated rewrites achieves only 18.6% on chat text from the same families, so a modelolect is specific to *register*, and to *model version*, and attribution requires training data from the register it will judge. As a byproduct we release inter-family vocabulary lexicons built with the log-odds method of our study on AI vocabulary. As a final experiment, after one TextPulse humanization, *p(human)* under the family classifier increases for 87 to 98% of the texts, a majority of the humanized texts classify as human, and the source family is recovered for at most 3% of them. All features, statistics, scores and code are made publicly available for future work.

1. Introduction

Our earlier work established that AI-generated academic text differs from human text on measurable surface features (TextPulse Research, 2026e, 2026f) and that individual model families occupy different positions on the resulting human-to-AI spectrum (TextPulse Research, 2026h).

Differences between families were measured as well. Em dash rates differ up to 200-fold between models, and vocabulary that one model injects is suppressed by another (2026d, 2026e). The next question is whether the AI model family that wrote a given text can be identified from the text alone, without any additional metadata.

Attribution of machine text to its AI generator is an established task in the detection literature, but the public benchmarks were built in the GPT-3.5 era. RAID (Dugan et al., 2024) covers eleven generators, M4GT-Bench (Wang et al., 2024) nine, and MAGE (Li et al., 2024) twenty-seven, and none of the three contains Claude, Gemini, DeepSeek, Grok or Qwen in a current version. A survey of the field (Huang, Chen and Shu, 2024) lists model attribution as an open problem because the model set changes faster than the benchmarks. The families that dominate current use were not considered in the earlier benchmarks.

Modelometry is the measurement and attribution of per-model writing style. A *modelolect* is the style a given model or family writes in, as an *idiolect* is the style of one human writer. The measurements already exist in our prior work on AI detection, alignment and stylometry. The paper conducts the attribution experiment on seven current families in two registers, academic text (formal) and open chat (informal), under a strict zero-generation constraint. The academic corpora were produced for prior studies and have AI family labels. The chat register is from the open LMArena human preference dataset (Chiang et al., 2024), and the academic Llama and GPT-4o continuations are from the open HAP-E corpus (Reinhart et al., 2025). We also measure whether a rewriting (i.e., humanization) engine removes the modelolect and whether the rewritten text reads as human to an AI family classifier.

2. Background and related work

2.1 Human and machine text differ on measurable features

The stylometric distance between human and machine text is well documented in the literature. Munoz-Ortiz et al. (2023) compared human news text with LLaMA output and found the human text had a more scattered sentence-length distribution, shorter constituents and a wider emotional range, while the model used more numbers, auxiliaries and pronouns. Herbold et al. (2023) compared 1,000 student essays with ChatGPT essays and found the model text more uniform in structure. Reviriego et al. (2024) measured lower lexical diversity in ChatGPT-3.5 than in humans on the same tasks, with GPT-4 narrowing the vocabulary gap without matching the underlying distribution. Park et al. (2025) tested the latent community structure of paraphrased text and rejected the hypothesis that model text and human text share a distribution. Our prior studies constructed a 47-feature stylometric fingerprint (2026e), a classifier and spectrum from it (2026f), a 1,057-word lexicon of vocabulary that models inject and suppress (2026d), and a burstiness study showing that sentence-length variation is the largest single separating feature (2026b). Kobak et al. (2025) tracked the model vocabulary into the scientific literature across 15 million PubMed

abstracts, and Liang et al. (2024) found the same vocabulary in 6.5 to 16.9% of AI-conference peer reviews. These measurements treat AI text as one category.

2.2 Inter-model differences in writing style

A smaller set of studies compares inter-model linguistic differences rather than AI model-human comparison. Reinhart et al. (2025) compared six LLMs with matched human continuations across registers and found the models distinguishable from humans and from each other in grammatical and rhetorical style, with instruction-tuned variants further from human text than their original base versions. Gude et al. (2026) compared two generations of models and found the more aligned generation less diverse in grammar and lexicon, which implies that alignment recipes leave measurable and provider-specific traces. Our prompting study (2026h) found a stable ranking of four families on a stylometric spectrum, and the ranking was preserved under three different prompts, which indicates a persistent per-family style and shows that the prompt has little effect on AI model writing style. Our fingerprint study (2026e) measured per-family punctuation and syntax signatures directly, including the 200-fold spread in em dash rates. None of these studies trains a classifier over currently used AI flagship families (e.g. ChatGPT, Claude, Gemini, etc.).

2.3 Attribution of machine text as a task

Multi-class attribution predates the current model generation. TuringBench (Uchendu et al., 2021) posed attribution across nineteen generators of the GPT-2 era and reported that it was harder than binary detection. RAID (Dugan et al., 2024), M4GT-Bench (Wang et al., 2024) and MAGE (Li et al., 2024) carry per-generator labels and support attribution experiments, and all three predate the modern AI families, as previously mentioned. Huang, Chen and Shu (2024) report LLM attribution as one of four problems in modern authorship analysis and note that closed models and fast version turnover make benchmarks ‘stale’ within a year. Wang et al. (2025) studied the reverse task, whether a model can imitate the implicit style of a specific human author, and found that current models fail at this task, which suggests that a model’s own style is not fully under prompt control. In the code domain, a 2025 study attributes C programs among five current models at 95% accuracy (arXiv 2506.17323), which establishes that a per-model signal exists in that modality. For natural language text and current flagship families, to the best of our knowledge, classifying AI model families based on inter-model stylometry (i.e., *modelometry*) is lacking in the literature.

2.4 Why family style should exist

The underlying tuning mechanism makes machine text detectable in the first place. Preference tuning ‘narrows’ the distribution a model samples from. Kirk et al. (2024) measured each stage of the Reinforcement Learning from Human Feedback (RLHF) pipeline and found that it reduced output diversity on every measure they considered. O’Mahony et al. (2024) attributed this mode collapse to specific fine-tuning components and to bias in the preference data. Padmakumar and He (2024) found that only preference-tuned models reduced the diversity of what human co-

writers produced. Singhal et al. (2023) found that most of the reward gained in RLHF was explained by response length. Yun et al. (2025) showed that the chat template itself reduces output diversity even at high sampling temperature, and Mahapatra (2026) found that instruction-tuned systems lose entropy along discourse and structural dimensions regardless of scale. Janus (2022) documented the earliest public case in text-davinci-002. Xu et al. (2026) found that base models, before preference tuning, read as human to commercial AI detectors, which places the origin of the machine register in the tuning stage instead of the pretraining stage.

Each provider narrows, or 'limits', its models toward its own preference data, raters and reward models. Two providers running the same recipe on different preference data should therefore produce different styles, and models trained on shared data should produce similar ones. Attribution accuracy and its confusion structure measure this. The confusion matrix is an empirical map of which training pipelines produce similar text. If a family were trained substantially on another family's outputs, its modelolect should be close to the teacher's.

3. Data and methods

3.1 Design under a zero-generation constraint

Four sources were used for the texts, which are summarized in Table 1.

Experiment	Register	Source	Classes	Texts	Model versions
A	Academic	Rewrite Corpus 1 (2026e)	human, GPT, Gemini, DeepSeek, Grok, Qwen	39,778	the endpoints of the 2026e study
A2	Academic	Rewrite Corpus 2	human, GPT, Claude, Gemini, DeepSeek, Llama, Qwen	61,114	gpt-5.4-mini, claude-haiku-4-5, gemini-2.5-flash-lite, deepseek-chat, Llama-3.3-70B-Instruct, Qwen3.5-122B
B	Chat	LMArena preference dataset	GPT, Claude, Gemini, DeepSeek, Grok, Llama, Qwen	36,179	37 versions, April to July 2025 (Claude 3.5 to Opus 4, Gemini 2.0 to 2.5 Pro, o3 and GPT-4.1, DeepSeek R1

Experiment	Register	Source	Classes	Texts	Model versions
					and V3, Grok 3 to 4, Llama 3.3 to 4, Qwen 2.5 to 3)
C	Academic	HAP-E continuations	human, GPT-4o, Llama-3-70B-Instruct, and Claude with n=42	3,723	gpt-4o-2024-08-06, Meta-Llama-3-70B-Instruct
D	Academic	Prior-study AI texts with their TextPulse humanizations	probe only	346	four flagship families of August 2026

Table 1. The five experiments. Texts counts the rows entering each classifier, including the human class where present.

The two rewrite corpora are collections of paired rewrites built for our earlier studies. In each pair, a model rewrote a human academic text from a human-written corpus, so content is preserved constant within a pair and family labels cannot act as a proxy for topic. Rewrite Corpus 1 is the corpus used for our prior work on AI fingerprint and classification (2026e, 2026f). It comprises five families. Rewrite Corpus 2 is a later internal collection, generated in a single pipeline, and it comprises Claude (claude-haiku-4-5, 13,844 rewrites) and Llama (Llama-3.3-70B-Instruct, 1,953 rewrites) along with GPT, Gemini, DeepSeek and Qwen. Families are capped at 6,000 texts by seeded reservoir sampling, and the human class is the deduplicated set of source texts (22,001 in Experiment A and 30,557 in Experiment A2).

The chat register is from the open LMArena arena-human-preference-140k dataset, which contains 136 thousand crowd-sourced battles, each carrying two model-labeled responses. We kept English, non-code battles, took the first assistant turn of each side, required 200 to 1,500 words, mapped the 37 model versions into the seven families, excluded all other providers, and capped each family at 6,000 by seeded reservoir sampling. Experiment C uses the academic split of the open HAP-E corpus (Reinhart et al., 2025). Each of its 1,227 seed chunks of published academic text has the human author's actual continuation and continuations by GPT-4o and Llama-3-70B-Instruct, which allows for a content-controlled three-class problem with a Llama class. The 42 academic Claude texts from our earlier papers join as a small fourth class. Experiment D employs the 120 fresh generations of our study on prompting (2026h, four families under three prompts) and the 48 source texts of our humanizer comparison (2026g), along with the TextPulse humanization of each, as held-out probe texts.

3.2 Three tiers of features

Tier one is the 47-feature set of our study on AI fingerprint (2026e, 2026f). It covers the sentence-length distribution, lexical diversity, punctuation rates, passive and nominalization rates, sentence openers, contraction and adverb rates, and the phrase counts, with the two dash features excluded as before. Tier two extends the log-odds lexicon method of our vocabulary study (2026d, after Monroe et al., 2008) from AI-versus-human to one-versus-rest per family, and produces a distinctive vocabulary list per family and register (Section 7). Tier three is a character 3-to-5-gram TF-IDF model, the standard authorship attribution baseline, employed to estimate the maximum attainable accuracy. In the chat register all model and vendor names are masked before vectorization, since arena responses occasionally state their own model name inline.

3.3 Classifiers and leakage control

The protocol is inherited from our study on classification (2026f). Multinomial logistic regression on standardized features is the primary model, histogram gradient boosting is the stronger reference. Both of them use balanced class weights, five-fold grouped cross-validation and pooled out-of-fold predictions. In the AI rewrite corpora the same human text was rewritten by up to six AI models, so all AI rewrites of one source, and the source itself, share a group and never cross a fold boundary. In the chat register, both of the responses answer the same prompt, so the group is the prompt hash. In HAP-E the group is the seed chunk.

The two rewrite corpora were generated years apart with different pipelines, and a probe classifier separates DeepSeek rewrites of the two collections at 99.5% accuracy, so corpora cannot be pooled without the collection signature as a family signature. Experiments A and A2 therefore run within one corpus each, and the seven-family academic claim is based on Experiment A2 alone.

3.4 The humanized probe

Experiment D trains the family classifier on Rewrite Corpus 1 and scores four held-out sets. These are the 120 fresh generations of our study on prompting, the same 120 after one TextPulse humanization, the 48 humanizer-comparison texts, and the same 48 after one TextPulse humanization, with the ten human texts of the prompt study as a control set. The experiment measures whether the source family is still detectable even after humanization, and whether the humanized text is classified as human by a classifier that was trained to categorize AI families from humans. These probe texts are fresh generations while the training texts are AI rewrites, so the probe also measures how the classifier behaves when away from its training distribution, and the pre-humanization rows should be read with that in mind.

4. Inter-model attribution in the academic register

Figure F1 and Table 2 present the main results. On Rewrite Corpus 2 the gradient boosting model attributes seven classes at 73.7% accuracy (logistic regression 69.2%) against a 14.3% chance rate. Balanced to 1,953 texts per class, the logistic model reaches 64.5% with a macro-F1 of 0.638. On Rewrite Corpus 1, six classes reach 79.2% (logistic 75.3%) against a 16.7% chance rate. The signal is not produced by class imbalance, since every family exceeds chance by a wide margin in the balanced runs.

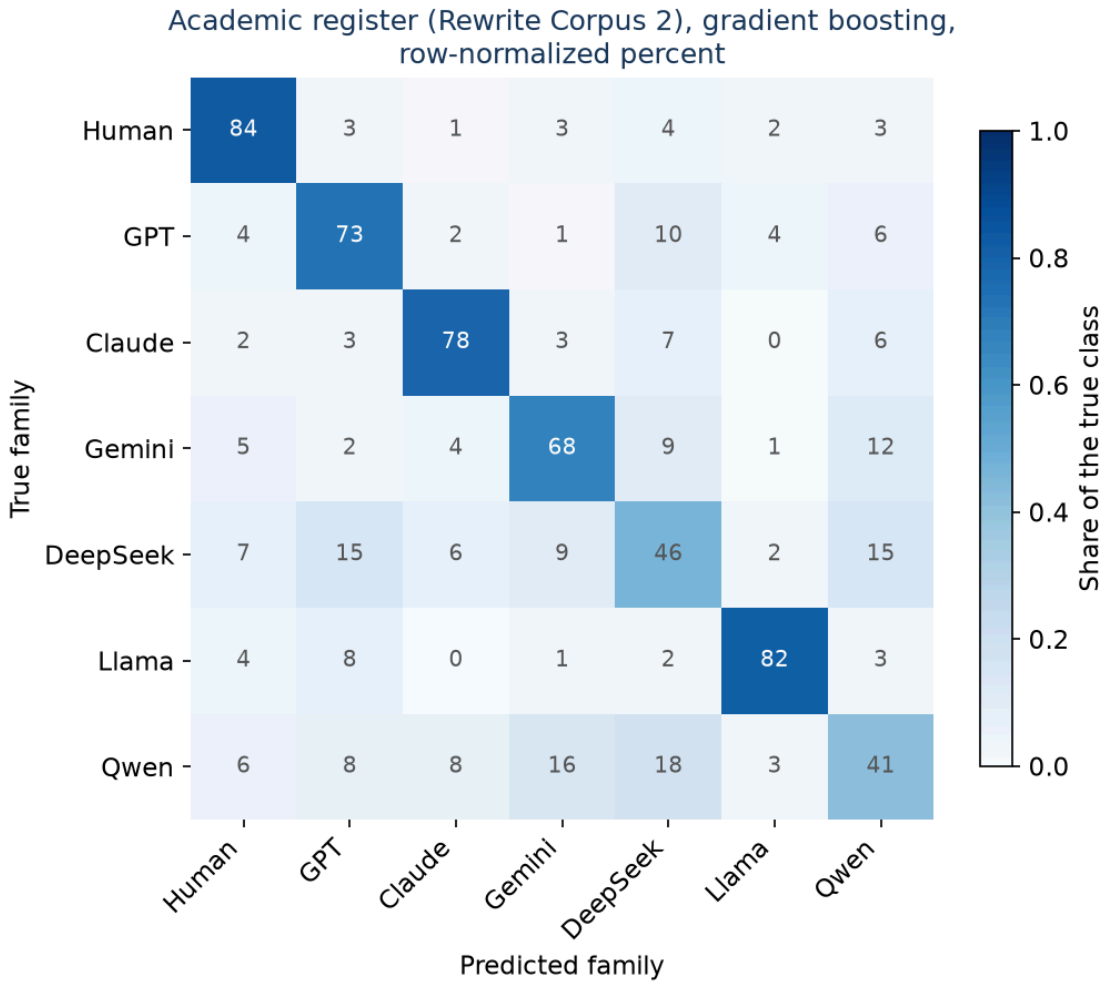

Figure F1. The academic confusion matrix (Experiment A2, gradient boosting, row-normalized). The diagonal is the per-family recall.

Class	n	Precision	Recall	F1
Human	30,557	0.947	0.839	0.890
Llama	1,953	0.601	0.819	0.693
Claude	6,000	0.768	0.783	0.775
GPT	6,000	0.629	0.733	0.677
Gemini	6,000	0.612	0.678	0.643

Class	n	Precision	Recall	F1
DeepSeek	provided, additionally, along, argued, originally, applying, caused, drives			
Qwen	4,604	0.361	0.408	0.383

Table 2. Experiment A2 per-class results, gradient boosting, pooled out-of-fold. Chance accuracy is 0.143.

The confusion matrix shows three notable patterns. First, DeepSeek is rarely attributed to GPT. In Rewrite Corpus 1, where DeepSeek recall is 0.84 and its one-versus-rest AUC is 0.977, only 3% of DeepSeek texts are attributed to GPT. In Experiment A2 the confusion increases to 15% and remains far from a collapse. Regardless of the provenance of DeepSeek's training data, its written style is distinct, and in Rewrite Corpus 1 it is the most identifiable family of the five.

Second, the weak classes are Qwen and, in Experiment A2, DeepSeek, and their errors are assigned mostly to each other, to Gemini and to GPT. Qwen is the least distinctive family in both academic corpora. Third, the human class keeps precision above 0.94 in both experiments. The classifier rarely labels an AI text as human, and when families are confused they are confused with each other.

The logistic regression coefficients identify the measurable content of each modelolect. Claude's academic signature has its largest coefficients on word length, comma rate, passive rate and parentheses. Gemini's are on the sentence-length distribution and lexical diversity. Llama's and GPT's are on sentence-length variation, and Qwen's are on type-token ratio and repeated sentence openers. Sentence-length geometry and lexical diversity, the same features that separate AI text from human text in our earlier studies, also do most of the work separating the AI families themselves from each other.

5. Inter-model and intra-model attribution in the chat register

The arena data tests whether the academic results depend on the rewrite task. These texts are free-form answers to real user prompts, written by 37 current model versions. Figure F3 and Table 3 present the results obtained. Gradient boosting over the 47 features reaches 63.9% (logistic regression 54.1%) against a 14.3% chance rate.

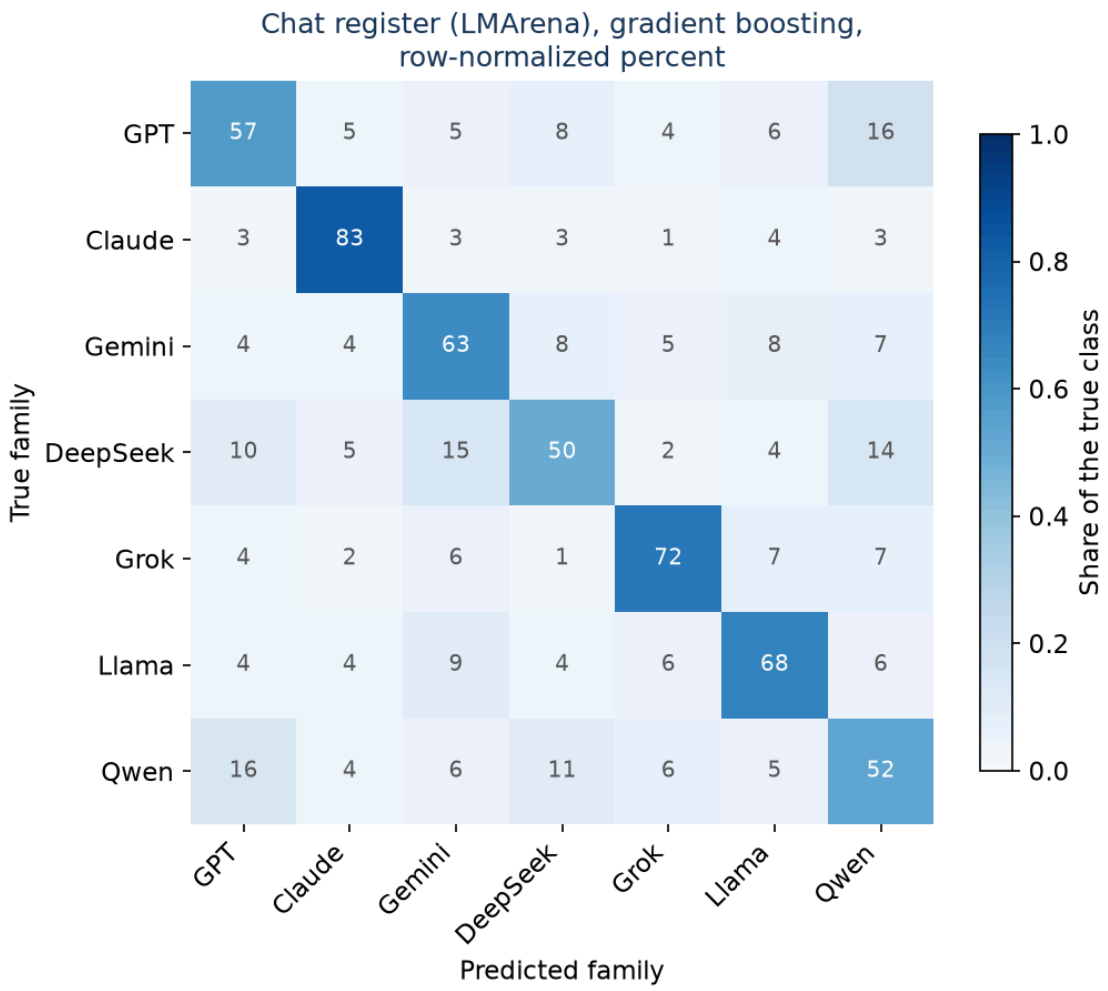

Figure F3. The chat confusion matrix (Experiment B, gradient boosting, row-normalized).

Family	n	Precision	Recall	F1
Claude	6,000	0.808	0.826	0.817
Grok	3,958	0.675	0.718	0.696
Llama	4,946	0.649	0.677	0.663
Gemini	6,000	0.656	0.631	0.643
GPT	6,000	0.620	0.573	0.595
Qwen	6,000	0.541	0.520	0.530
DeepSeek	provided, additionally, along, argued, originally, applying, caused, drives			

Table 3. Experiment B per-family results, gradient boosting, pooled out-of-fold. Chance accuracy is 0.143.

Claude is the most identifiable family in open chat by a wide margin, as it is in the academic register once present at scale. The largest off-diagonal cell in the matrix is Qwen predicted as GPT (16%) along with GPT predicted as Qwen (16%), the only 'symmetric' confusion of that size under study. Under the interpretation of Section 2.4, the preference pipelines of these two families produce the most similar text of any pair among the seven. DeepSeek is again weak in this register, and its errors are assigned mostly to Gemini and Qwen.

Figure F5 divides the family recalls into versions, and the intra-family versions differ by a wide margin. OpenAI's o3 is attributed at 0.84 while chatgpt-4o-latest reaches 0.30. Claude 3.5 Haiku reaches 0.89 while Claude 3.7 Sonnet reaches 0.60. Grok 4 (0.79) is more identifiable than the Grok 3 preview (0.59). A family label is an aggregate over versions with visibly different styles, and reasoning-focused versions (o3, the Grok mini variants, QwQ) achieve higher recall than their conversational counterparts.

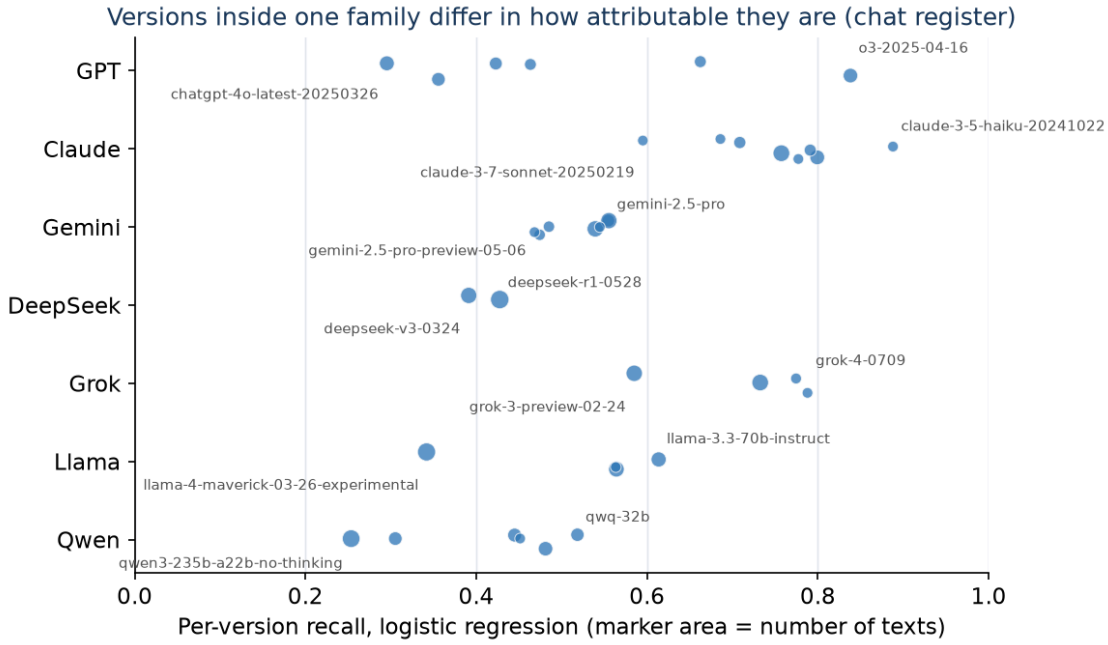

Figure F5. Per-version recall inside each family in the chat register. Marker area is the number of texts, and the extremes of each family are labeled.

The classifier trained on Rewrite Corpus 1, applied to the same five families' arena texts, scores 18.6%, below the 20% chance rate of a five-way task, so transfer between formal-informal registers fails. The modelolect measured here is specific to *register* and to *model version*. A family's chat style and its academic rewriting style are different objects, and training data for attribution must match the register of the target text. This repeats, at family level, the domain-shift fragility that the detection literature reports for detectors (Dugan et al., 2024).

6. The maximum attainable accuracy

Two experiments were carried out to estimate the maximum degree of attribution. On the HAP-E academic split, where the human author's own continuation and two models' continuations complete the same 500-word seed, gradient boosting distinguishes human, GPT-4o and Llama-3-70B at 98.6% (Figure F2). Under matched content, register and length, attribution among these three classes is almost perfect. Adding the 42 academic Claude texts from our earlier papers as a fourth class keeps the accuracy at 98.5%, with a Claude recall of 1.00, and the different provenance of those 42 texts means their number is an upper estimate only.

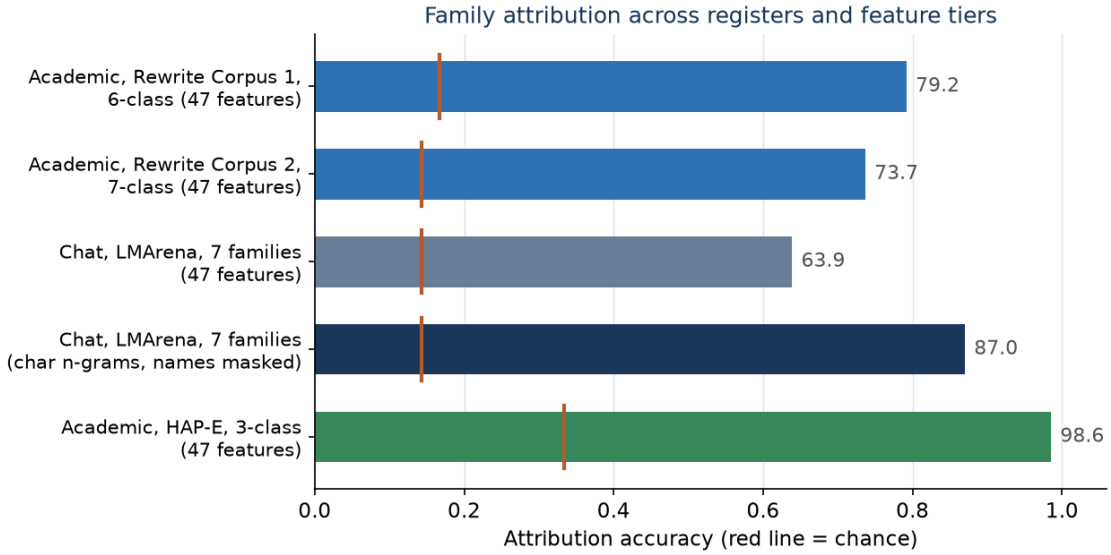

Figure F2. Attribution accuracy across registers and feature tiers. The red line in each bar is the chance rate of that task.

The character n-gram model quantifies the share of the signal that the interpretable features capture. On the same arena texts, folds and groups, TF-IDF character 3-to-5-grams with model names masked reach 87.0% accuracy (macro-F1 0.862) against 63.9% for the 47 features. The 47 interpretable features account for about three quarters of the attributable signal, and the remainder is lexical and sub-lexical detail that only a high-dimensional representation captures. For practical attribution the n-gram model is the stronger tool. For understanding what a modelolect consists of, the interpretable tier is informative, since each of its features names a verifiable property.

7. Family vocabulary lexicons

Tier two produces, per register, a ranked list of words each family over-uses relative to the other families pooled, computed with the log-odds method of our vocabulary study. Table 4 shows the top of the academic lists. These are the words that distinguish each family's academic text. Claude over-uses institutional vocabulary ("substantially", "mechanisms", "frameworks", "fundamentally"). Gemini over-uses connectives and formal verbs ("furthermore", "utilizing",

“posited”). Qwen over-uses formal connectives (“regarding”, “alongside”, “wherein”). The full lists, forty words per family and register with rates and z-scores, are released in the data.

Family	Most distinctive academic vocabulary (top 8 by z-score)
Claude	through, substantially, across, mechanisms, institutional, frameworks, documented, populations
DeepSeek	provided, additionally, along, argued, originally, applying, caused, drives
Gemini	furthermore, individuals, specifically, significant, utilizing, posited, ascertain, concerning
GPT	be, so, used, may, together, accordingly, moreover, because
Llama	being, context, result, highlighting, which, because, due, virtue
Qwen	regarding, while, specific, author, alongside, executed, year, utilized

Table 4. Per-family academic vocabulary, one-versus-rest log-odds with an informative Dirichlet prior, Rewrite Corpus 2. Every listed word appears at a minimum of twice the rate of the other families pooled, and each lexicon reflects the single model version in the corpus. The public release includes rates per 10,000 tokens.

Prior work on AI-vocabulary has so far treated model vocabulary as one list (our 2026d lexicon, and Kobak et al., 2025). The per-family lists show that the aggregate list is a union of distinct family vocabularies. “Furthermore” is used by Gemini, “moreover” by GPT, and “substantially” by Claude. A population-scale word-tracking study such as that by Kobak et al. could attribute literature contamination to families, and we note that application for future work.

8. Attribution after humanization

Experiment D connects family attribution to the humanization task. The family classifier trained on Rewrite Corpus 1 scores the 168 fresh AI generations of our earlier papers and the same texts after one TextPulse humanization (Figure F4 and Table 5). As a control, eight of the ten human texts are classified as human.

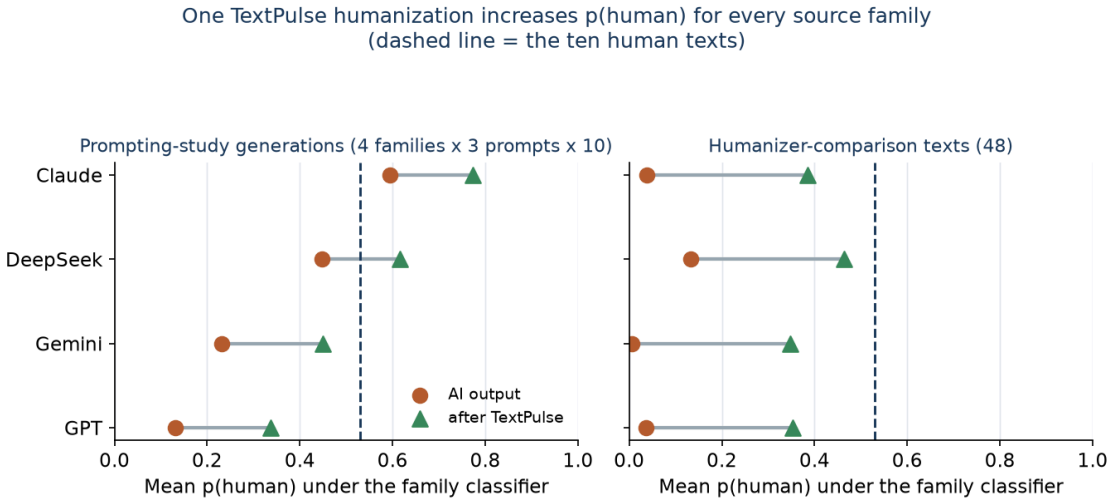

Figure F4. Mean p(human) per source family before and after one TextPulse humanization, for both probe sets. The dashed line is the ten human texts.

Probe set	n	Source family recovered	Classified human	Mean p(human)
Humanizer-comparison texts (AI)	48	14% within trained families	2%	0.05
Humanizer-comparison texts after TextPulse	48	0%	58%	0.39
Prompting-study generations (AI)	120	9% within trained families	43%	0.35
Prompting-study generations after TextPulse	120	3%	68%	0.54

Table 5. The humanized probe. Source family recovered counts texts attributed to their true family. Claude texts cannot be recovered by this classifier, which has no Claude class, and are excluded from the recovery rates. The prompting-study generations include instruction-loaded prompts that already imitate human style, which is why 43% read as human before any humanization, while the humanizer-comparison texts, plain generations, read as human in 2% of cases.

The recovery rates are informative even before humanization. On the original generations, the classifier recovers the source family in only 14% (humanizer-comparison texts) and 9% (prompting-study generations) of the eligible texts. These probe texts are fresh generations by August 2026 flagship versions while the training texts are rewrites by earlier versions, so this repeats the transfer failure of Section 5 and reinforces the practical conclusion. An attribution model is only as good as the match between its training data and the target text's *register* and

version era. Family removal by the humanizer therefore cannot be separated from the domain shift in this experiment, and we report the recovery rates.

The probe result is the change toward the human class, which does not rely on recovering the source family. After one TextPulse humanization, p(human) increases for 87.5% of the prompting-study pairs and 97.9% of the humanizer-comparison pairs, the mean p(human) increases by 0.19 and 0.33, and the share classified human goes from 2% to 58% on the humanizer-comparison set and from 43% to 68% on the prompting-study set. The classifier here is the stylometric classifier we developed in prior work, and the result complements the commercial-detector measurements of the humanizer comparison (2026g) with an attribution-based reading. Regardless of the family signal the classifier can still find in a fresh generation, the humanization removes most of it and replaces it with statistics the classifier reads as human.

9. Conclusion

The attribution experiments support modelometry as a measurement task. A classifier over 47 interpretable surface features identifies the family that wrote a text at four to five times the chance rate in both registers tested, a character n-gram model reaches 87% on seven current families, and on matched-content academic continuations attribution is near perfect. The modelolects differ in strength, with Claude the most identifiable family and Qwen the least identifiable. The confusion structure maps the similarity of training pipelines. Modelolects are bound to register and version, so attribution requires training data from the register it judges. The interpretable tier accounts for about three quarters of the attributable signal, which means most of a modelolect consists of countable properties, namely, sentence-length geometry, lexical diversity, punctuation and connective habits, and a family-favored vocabulary. A purpose-built humanizer outperforms the classifier, and after one TextPulse humanization the probe texts change toward the human class in 87 to 98% of pairs and a majority read as human to the same classifier.

10. Limitations

Figure F5 shows the reason version-level intra-family claims need version-level data, since recall varies by a factor of two among versions of one family. The academic corpora are rewrites of human texts, so Experiments A and A2 measure the modelolect of AI rewriting. Experiments B and C cover free generation and continuation, and the family rankings agree across all three experiments, but the tasks are distinct and the accuracy numbers are not comparable across experiments. The corpora span model eras, which is inherent to a zero-generation design, and it causes cross-register transfer to fail. Version change and register change are confounded in that number. Arena responses vary in mean length by family (351 to 785 words), and although the grouped design prevents prompt leakage, length-linked features contribute to the chat result. The HAP-E experiment, where length is controlled, shows that attribution survives without that contribution. The Claude class in Experiment C has a different provenance from its co-classes and it obtained

perfect recall there. The probe classifier lacks Claude and Llama classes, so its recovery rates exclude Claude texts. These limitations are deferred for future work.

Data availability

Per-text features, per-text scores, classifier results, confusion matrices, the per-family lexicons and every script are released with this paper. The archive is deposited on Zenodo (DOI 10.5281/zenodo.22237709). Raw texts from the LMArena and HAP-E datasets are not redistributed. Both are openly downloadable, and the release includes the deterministic sampling code (fixed seed) that reconstructs the exact text sample from the public files. The in-house corpora used are described in the data statements in our earlier papers.

References

Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J. E., & Stoica, I. (2024). Chatbot Arena: An open platform for evaluating LLMs by human preference. Proceedings of the 41st International Conference on Machine Learning. https://arxiv.org/abs/2403.04132

Dugan, L., Hwang, A., Trhlik, F., Ludan, J. M., Zhu, A., Xu, H., Ippolito, D., & Callison-Burch, C. (2024). RAID: A shared benchmark for robust evaluation of machine-generated text detectors. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. https://arxiv.org/abs/2405.07940

Gude, A., Santos-Rios, R., Bond, F., Flickinger, D., Gomez-Rodriguez, C., & Zamaraeva, O. (2026). More aligned, less diverse? Analyzing the grammar and lexicon of two generations of LLMs. arXiv:2605.06030. https://arxiv.org/abs/2605.06030

Herbold, S., Hautli-Janisz, A., Heuer, U., Kikteva, Z., & Trautsch, A. (2023). A large-scale comparison of human-written versus ChatGPT-generated essays. Scientific Reports, 13, 18617. https://doi.org/10.1038/s41598-023-45644-9

Huang, B., Chen, C., & Shu, K. (2024). Authorship attribution in the era of LLMs: Problems, methodologies, and challenges. arXiv:2408.08946. https://arxiv.org/abs/2408.08946

Janus. (2022). Mysteries of mode collapse. LessWrong. https://www.lesswrong.com/posts/t9svvNPNmFf5Qa3TA

Kirk, R., Mediratta, I., Nalmpantis, C., Luketina, J., Hambro, E., Grefenstette, E., & Raileanu, R. (2024). Understanding the effects of RLHF on LLM generalisation and diversity. International Conference on Learning Representations. https://arxiv.org/abs/2310.06452

Kobak, D., Gonzalez-Marquez, R., Horvat, E.-A., & Lause, J. (2025). Delving into LLM-assisted writing in biomedical publications through excess vocabulary. Science Advances, 11(27), eadt3813. https://doi.org/10.1126/sciadv.adt3813

Li, Y., Li, Q., Cui, L., Bi, W., Wang, Z., Wang, L., Yang, L., Shi, S., & Zhang, Y. (2024). MAGE: Machine-generated text detection in the wild. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. https://arxiv.org/abs/2305.13242

Liang, W., Izzo, Z., Zhang, Y., Lepp, H., Cao, H., Zhao, X., … & 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. https://arxiv.org/abs/2403.07183

Mahapatra, R. (2026). From context shift to stylistic collapse: Why training objectives matter more than scale. arXiv:2605.28826. https://arxiv.org/abs/2605.28826

Monroe, B. L., Colaresi, M. P., & Quinn, K. M. (2008). Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. Political Analysis, 16(4), 372-403. https://doi.org/10.1093/pan/mpn018

Munoz-Ortiz, A., Gomez-Rodriguez, C., & Vilares, D. (2023). Contrasting linguistic patterns in human and LLM-generated text. arXiv:2308.09067. https://arxiv.org/abs/2308.09067

O'Mahony, L., Grinsztajn, L., Schoelkopf, H., & Biderman, S. (2024). Attributing mode collapse in the fine-tuning of large language models. ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models. https://openreview.net/pdf?id=3pDMYjpOxk

Padmakumar, V., & He, H. (2024). Does writing with language models reduce content diversity? International Conference on Learning Representations. https://arxiv.org/abs/2309.05196

Park, M., Choi, Y., & Jeon, J.-J. (2025). Does a large language model really speak in human-like language? arXiv:2501.01273. https://arxiv.org/abs/2501.01273

Reinhart, A., Brown, D. W., Markey, B., Laudenbach, M., Pantusen, K., Yurko, R., & Weinberg, G. (2025). Do LLMs write like humans? Variation in grammatical and rhetorical styles. Proceedings of the National Academy of Sciences, 122(8). https://doi.org/10.1073/pnas.2422455122

Reviriego, P., Conde, J., Merino-Gomez, E., Martinez, G., & Hernandez, J. A. (2024). Playing with words: Comparing the vocabulary and lexical diversity of ChatGPT and humans. Machine Learning with Applications, 18, 100602. https://doi.org/10.1016/j.mlwa.2024.100602

Singhal, P., Goyal, T., Xu, J., & Durrett, G. (2023). A long way to go: Investigating length correlations in RLHF. arXiv:2310.03716. https://arxiv.org/abs/2310.03716

TextPulse Research. (2026b). Sentence-length burstiness as a cross-disciplinary and cross-model signal of AI rewriting. TextPulse Working Paper. https://doi.org/10.5281/zenodo.22033063

TextPulse Research. (2026d). The vocabulary fingerprint of AI rewriting: A 1,057-word lexicon from 60,786 paired academic texts. TextPulse Working Paper. https://doi.org/10.5281/zenodo.22028377

TextPulse Research. (2026e). Stylometric fingerprints of AI rewriting: Punctuation, syntax, and model attribution. TextPulse Working Paper. https://doi.org/10.5281/zenodo.22040684

TextPulse Research. (2026f). Human versus AI text classification from stylometric features across 121,092 academic texts. TextPulse Working Paper. https://doi.org/10.5281/zenodo.22048477

TextPulse Research. (2026g). A controlled comparison of AI text humanizers on academic writing. TextPulse Working Paper. https://doi.org/10.5281/zenodo.22152404

TextPulse Research. (2026h). Do AI models speak human? Prompting cannot undo preference tuning. TextPulse Working Paper. https://doi.org/10.5281/zenodo.22159494

Uchendu, A., Ma, Z., Le, T., Zhang, R., & Lee, D. (2021). TURINGBENCH: A benchmark environment for Turing test in the age of neural text generation. Findings of the Association for Computational Linguistics: EMNLP 2021. https://arxiv.org/abs/2109.13296

Wang, Y., Mansurov, J., Ivanov, P., Su, J., Shelmanov, A., Tsvigun, A., … & Nakov, P. (2024). M4GT-Bench: Evaluation benchmark for black-box machine-generated text detection. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. https://arxiv.org/abs/2402.11175

Wang, Z., Tripto, N. I., Park, S., Li, Z., & Zhou, J. (2025). Catch me if you can? Not yet: LLMs still struggle to imitate the implicit writing styles of everyday authors. Findings of the Association for Computational Linguistics: EMNLP 2025. https://arxiv.org/abs/2509.14543

Xu, Y. E., Zhong, Z., Raghunathan, A., Fang, F., & Kolter, J. Z. (2026). Base models look human to AI detectors. arXiv. https://arxiv.org/abs/2605.19516

Yun, L., An, C., Wang, Z., Peng, L., & Shang, J. (2025). The price of format: Diversity collapse in LLMs. arXiv:2505.18949. https://arxiv.org/abs/2505.18949