EVERY MODEL LEAVES A TRACE: MULTI-RESOLUTION LLM FINGERPRINTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models leave model-specific traces in their generated text, but it remains unclear how finely these fingerprints can distinguish related models and which signals remain informative as obvious cues are reduced. We study this through *fingerprint resolution*, considering both model resolution, from different families to distilled and adapted variants, and signal resolution across behavioral, stylometric, lexical, and semantic evidence. We progressively suppress surface cues, paraphrase outputs, vary the amount of available text, and test generalization across prompt distributions and languages. Our results show that fingerprints persist beyond broad family differences, remaining detectable in distilled models and even LoRA adaptations of the same base model. Although attribution weakens as surface and lexical cues are altered, substantial model-specific signal remains. These signals transfer across prompt sources and languages, with semantic representations proving most robust across languages. These findings show that model identity is recoverable at multiple resolutions, but that recoverability depends on model similarity and the evidence available in the text.

1 INTRODUCTION

LLMs often leave recognizable traces in their generated text, making it possible to distinguish among different source models (Uchendu et al., 2020; Rao et al., 2025; Antoun et al., 2024). Some traces are easy to observe when a model generates text, such as producing longer responses, using more headings, following particular formatting conventions, or beginning answers in recurring ways. Such signals can make attribution relatively easy, but they are also fragile, and post-generation editing can weaken text-based detectors (Krishna et al., 2023; Dugan et al., 2024). What remains unclear is which model-specific signals survive when these visible cues are reduced, especially when the candidate models are closely related. We therefore ask: *how finely can the identity of an LLM be resolved from its generated text, and which properties of the text still reveal that identity?*

This question is increasingly important as modern LLMs become more diverse and increasingly appear as closely related model families rather than isolated systems. Different sizes, distilled descendants, and adapted variants may share architectures, training histories, or development processes while differing only in subtle ways, making source attribution more difficult. At the same time, generated text may be reformatted, edited, shortened, or paraphrased before attribution, weakening or removing model-identifying signals. Attribution must therefore address two related challenges by distinguishing increasingly similar models and determining which identifying signals remain observable after the text is transformed.

We study this problem through *fingerprint resolution*, illustrated in Figure 1. *Model resolution* describes how attribution changes as candidate generators become more similar, from different families to within-family models, distilled descendants, and controlled LoRA (Hu et al., 2022) adaptations. *Signal resolution* describes which behavioral, stylometric, lexical, and semantic signals carry model identity and whether they survive as observable cues are reduced. Together, these complementary axes distinguish the existence of a fingerprint from the conditions under which it remains detectable.

Prior work has shown that source models can be identified using lexical and grammatical patterns, learned representations, and supervised attribution models (McGovern et al., 2025; Sun et al., 2025; Lu et al., 2026). Our goal is to use model attribution as a diagnostic tool rather than treat accuracy as the endpoint. We progressively remove observable cues, paraphrase model outputs, and control the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

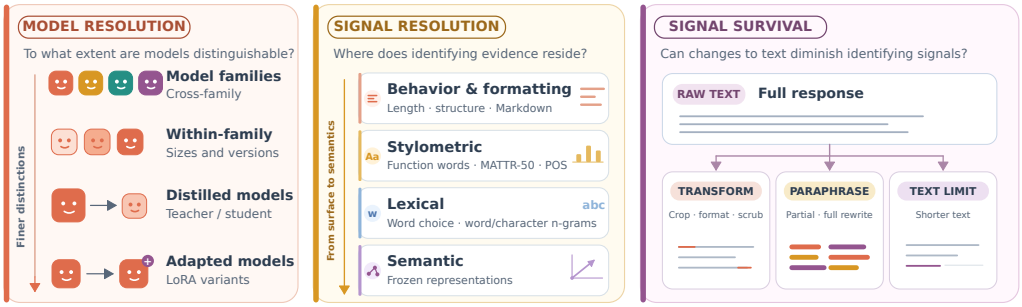


Figure 1: Overview of the fingerprint-resolution framework, linking model similarity, observable evidence, and generalization across prompts, languages, and text budgets.

amount of available text to determine which behavioral, stylometric, lexical, and semantic signals continue to reveal model identity. We further test whether fingerprints generalize beyond the conditions under which they are learned. Source-held-out experiments measure transfer across prompt distributions, while within- and cross-language experiments evaluate English, Chinese, and Persian. These settings address evidence that source attribution may weaken across languages (La Cava et al., 2026) and that fingerprint visibility can depend on language (Gloaguen et al., 2026).

Our results show that model fingerprints extend beyond broad family differences and remain detectable even between distilled models and their teachers, as well as between LoRA-adapted versions of the same base model. Attribution becomes harder as behavioral and stylometric cues are removed and lexical choices are altered through paraphrasing. Cross-lingual experiments show that semantic representations provide the most reliable transfer across languages, while cross-dataset results show substantial transfer across prompt sources but sharp degradation for highly constrained generation tasks. Model identity is therefore recoverable at different resolutions depending on model similarity and the evidence available in generated text.

Our key contributions are threefold: (1) **A fingerprint-resolution framework** spanning different levels of model similarity, revealing distinguishable fingerprints even in distilled models and LoRA-adapted variants of the same base model; (2) **A decomposition of model-identifying evidence** using multiple signal families, progressive cue suppression, paraphrasing, and varying the amount of available text; and (3) **A broad evaluation of fingerprint generalization** across unseen prompt distributions, different generation tasks, and within- and cross-language settings in English, Chinese, and Persian.

2 Related Work

Prior work has shown that LLM-generated text contains model-specific signals that can be used to identify its source. These signals appear at different levels, from broad differences associated with model family and size (Antoun et al., 2024) to finer patterns in vocabulary, character use, and sentence structure that can persist across domains (McGovern et al., 2025). Other approaches use learned representations, contrastive learning, or specialized attribution models to improve source classification (Sun et al., 2025; Bitton et al., 2025; Lu et al., 2026; Guo et al., 2024; La Cava & Tagarelli, 2025). Distinguishing signals can even emerge among closely related models that differ only in training details such as initialization or data order (Suzuki et al., 2025; Tong et al., 2026). Together, these studies establish that model identity is often recoverable from generated text. Less clear is how that recoverability changes as candidate models become more similar and as the available identifying cues are progressively removed.

Several studies have also examined the robustness of model attribution. Some control for response length and formatting or apply transformations such as translation and summarization (Sun et al., 2025; McGovern et al., 2025). Related detection benchmarks show that robustness can deteriorate when outputs come from previously unseen generators (Li et al., 2024; Wang et al., 2024). Ownership-oriented methods take a different approach by implanting private trigger-response behav-

iors or probing models with selected prompts and repeated observations (Xu et al., 2024; Pasquini et al., 2025; Gao et al., 2025). Our setting is complementary. Rather than optimizing prompts or attribution models, we study passive attribution from ordinary model outputs and ask how much identifying information remains when text is shortened, reformatted, partly rewritten, or paraphrased. This motivates our view of fingerprinting as a *resolution* problem, determined jointly by the similarity of the candidate models and the amount and type of evidence that remains observable.

A smaller body of work considers multilingual attribution. Recent results suggest that source-model attribution does not always transfer reliably across languages (La Cava et al., 2026; Greco et al., 2026), while semantically conditioned watermarking shows that language can be used as a channel for embedding model-specific signals (Gloaguen et al., 2026). In contrast, we focus on naturally occurring model fingerprints rather than deliberately embedded signals, and examine which behavioral, lexical, and semantic cues remain informative within and across English, Chinese, and Persian.

3 Experimental Setup

As discussed in Section 2, most prior work frames attribution as a closed-set classification task in which a generated response is assigned to one of several predefined candidate models and performance is summarized by a single score. While this formulation establishes whether model-specific signals are recoverable, it provides limited insight into how attribution changes as candidate models become increasingly similar or which properties of the generated text support the prediction. These questions matter because modern LLMs often appear as closely related variants and because their outputs may be altered before attribution. We study these factors through *fingerprint resolution*.

In this section, we first formalize attribution as a measure of fingerprint recoverability and then define model resolution (Sec. 3.1), signal representations (Sec. 3.2), and text transformations (Sec. 3.3) used to study both axes of fingerprint resolution. We then study these two axes jointly in Sec. 4 and examine whether the resulting fingerprints generalize across prompt distributions and languages in Sec. 5. Detailed generation, feature, and classifier settings are provided in Appendix A.

Problem Formulation. Let $\mathcal{M} = \{M_1, \ldots, M_K\}$ denote the set of candidate generators. Given a response y, the attribution task predicts its source model as $\hat{M} = f(y)$. Rather than treating the performance of f as the endpoint, we use it as a measure of fingerprint recoverability. We examine how attribution changes along two complementary axes: *model resolution*, which varies how closely related the candidate generators are, and *signal resolution*, which varies the type and amount of model-identifying evidence available in their outputs.

3.1 Model Resolution

Model resolution is examined across four levels of similarity between candidate generators. Our primary model set contains 19 models from eight families and supports the cross-family and within-family experiments. Additional models are used for the distillation and controlled-adaptation settings. Complete model lists and exact identifiers are provided in Appendix A.1. Across settings, we keep the attribution protocol as consistent as possible so that differences in performance primarily reflect the similarity between candidate models.

- **Cross-Family.** We select one representative from each of the eight families: Llama-4 Maverick, Qwen-3 32B, Mistral-Small-3.2 24B, Gemma-4 31B, DeepSeek-V4 Flash-0731, Claude Haiku 4.5, Gemini-2.5 Flash, and GPT-4.1.
- **Within-Family.** We compare all model pairs within families containing multiple models. This gives six Llama pairs, three Qwen pairs, and one pair each for other model families.
- **Distillation Relationships.** We examine three teacher–student pairs: Llama-3.1 8B vs. Llama-3.1-Minitron 4B, Mistral-NeMo 12B vs. Mistral-NeMo-Minitron 8B, and Gemma-2 27B vs. Gemma-2 2B.
- **Controlled Adaptations.** We use Llama-3.1-8B-Instruct and Qwen3-8B as base models and fine-tune each on GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021) using LoRA. We then vary the strength of the learned LoRA update to create a controlled sequence of adaptations and test when their outputs become distinguishable from those of

View	Transformation
T_0	Raw response
T_1	Remove words from the beginning and end
T_2	T_1 + remove formatting and layout cues
T_3	T_2 + lowercase and remove punctuation, symbols, and emojis
P_1	Paraphrase approximately one third of eligible sentences in T_0
P_2	Paraphrase the full T_0 response
P_1^*	Apply T_3 transformation to P_1
P_2^*	Apply T_3 transformation to P_2

Table 1: Response views used to progressively remove or rewrite model-identifying cues.

the corresponding base model. The fine-tuning procedure and LoRA settings are described in Appendix D.

3.2 Signal Resolution: From Surface to Semantics

Signal resolution covers four complementary levels of model-identifying signals, ranging from surface-level behavioral patterns to semantic representations.

- **Behavioral and Structural.** These features capture how a model organizes and presents its response. They cover response length and structure, formatting and layout, punctuation, and repetition tendency. Representative examples include sentence and paragraph counts, headings and lists, punctuation usage, and bigram and trigram repetition ratios.
- **Stylometric.** These features capture recurring patterns in vocabulary use and grammatical structure. They cover vocabulary diversity, word-form statistics, vocabulary distribution, function and grammatical words, repetition, and the ratio of numeric tokens. We concatenate 17 dense lexical features with POS (part-of-speech) unigram, bigram, and trigram TF–IDF representations, inspired by prior attribution work (McGovern et al., 2025).
- **Lexical.** To capture lexical form directly, we use TF–IDF representations at both the word and character levels. The word-level representation uses unigrams and bigrams, while the character-level representation uses 3–5-grams. We concatenate these two representations into a single lexical feature set.
- **Semantic.** We use `openai/text-embedding-3-small` to obtain a fixed embedding for each response. This provides a semantic-level view that can capture model-specific information not explicitly represented by the manually defined or surface-form features.

Two additional derived representations are considered. First, for the **prompt-neutralized embedding**, we compute the mean embedding across all candidate-model responses to the same prompt and subtract this mean from each response embedding. This removes variation shared across models for that prompt and emphasizes differences between their responses. Because this setting requires peer responses from all candidate models for the same test prompt, it differs from standard attribution and is used as a diagnostic representation. If attribution improves after neutralization, it suggests that removing prompt- and task-shared variation exposes model-specific signals that are more discriminative and less dependent on the prompt itself. Second, the **combined** representation concatenates the four main signal groups described above to capture their complementary information. Complete feature definitions, vectorizer settings, and dimensionality limits are provided in Appendix A.4.

3.3 Transformations

Attribution is evaluated as candidate models become more similar and as identifying signals are progressively weakened. Table 1 summarizes the response views used throughout the experiments. T_1–T_3 form a cumulative signal-removal sequence, while P_1 and P_2 are generated directly from the raw response through LLM-based partial and full paraphrasing, respectively. We then apply T_3 to P_1 and P_2 to obtain P_1^* and P_2^*, allowing us to test paraphrased text after the strongest surface normalization. Full transformation and paraphrasing details are provided in Appendix A.5.

View	Behavioral	Stylometric	Lexical	Semantic	Combined	PN Embedding†
T0	52.42	60.72	80.07	66.04	**81.41**	83.75
T1	48.78	58.93	**76.27**	60.81	76.14	76.13
T2	41.88	56.34	**72.48**	56.23	71.86	74.42
T3	25.59	49.62	**67.76**	50.94	67.04	69.35
P1	50.86	58.20	74.42	63.69	**77.85**	80.90
P2	48.12	53.32	68.18	57.81	**71.86**	75.23
P1*	24.48	46.51	61.01	47.51	**61.47**	64.99
P2*	23.08	41.79	53.94	41.90	**55.87**	59.39

Table 2: Cross-family attribution accuracy (%) using logistic regression. Chance accuracy is 12.5%. †PN Embedding denotes the Prompt-Neutralized Embedding accuracy. This representation requires access to peer responses from all candidate models for the same prompt. We use "PN Embedding" and this symbol consistently throughout the paper.

3.4 Datasets and Evaluation

To make the attribution setting more realistic, we use open-ended prompts and sample responses at temperature 1.0, producing substantial generations; approximately 95% of responses exceed 50 words. The main English experiments use four prompt sets, D_1–D_4, with 200 prompts from each source before filtering. D_1 and D_2 are independent samples from INFINITY-CHAT (Jiang et al., 2026), D_3 contains researcher-designed prompts spanning five task categories (construction described in Appendix A.2), and D_4 comes from a separate general-prompt source (Rajani et al., 2023). After removing one near-duplicate, the final set contains 799 prompts. All models follow a common decoding protocol (Appendix A.3). The multilingual experiments in Sec. 5 extend this setup to English, Mandarin Chinese, and Persian by translating the English prompts with GPT-4o-mini.

For attribution, we use Logistic Regression (LR) as the primary diagnostic classifier and a one-hidden-layer MLP as a nonlinear robustness check. The main English experiments use five-fold, prompt-grouped, dataset-stratified cross-validation, keeping all outputs and response views associated with the same prompt within a single fold. We reuse the same fold assignments across experiments and report accuracy. Classifier settings and evaluation details are provided in Appendix A.6.

4 Resolving Model Fingerprints

This section presents the results of the *model- and signal-resolution* experiments. Attribution is evaluated across different signal types while observable signals are progressively reduced to test whether model fingerprints persist. The analysis also spans increasing levels of model similarity, from distinct model families to closely related adaptations of the same base model. Neither LR nor the MLP consistently outperforms the other across settings. We therefore report LR results throughout the main text and provide the corresponding MLP results in the Appendices B and C.

Across Model Families. Table 2 presents the eight-way cross-family attribution results, with a chance level of 12.5%. The results show that attribution relies on more than formatting signals and response templates. Behavioral features show the largest relative decline under edge cropping (T1) and formatting removal (T2), consistent with their direct encoding of response structure, while stylometric and lexical representations remain more stable. Even after T3 removes formatting, capitalization, punctuation, symbols, and emojis, the lexical representation remains the strongest individual signal group. This suggests that model-specific word choices and local lexical patterns persist after visible surface signals are suppressed. Embedding accuracy also declines under T1–T3, indicating that frozen representations capture some of this affected lexical and stylistic information rather than semantic content alone.

Paraphrasing produces a complementary pattern. Accuracy decreases from P1 to P2 as more of the original text is rewritten, and falls further in P1* and P2* after applying T3. Combined performs best on the raw and paraphrased views, whereas the Lexical representation slightly exceeds

Model Pair	Behavioral	Stylometric	Lexical	Semantic	Combined	PN Embedding†
3.1-70B vs. 3.3-70B	60.33	59.45	59.76	59.70	**66.40**	73.34
3.1-70B vs. 4-Maverick	60.45	63.33	69.02	60.51	**72.09**	76.35
3.1-8B vs. 3.1-70B	55.38	58.26	59.01	55.38	**61.39**	63.33
3.1-8B vs. 3.3-70B	57.51	61.51	60.76	58.64	**64.27**	68.21
3.1-8B vs. 4-Maverick	56.32	62.83	68.59	58.57	**70.96**	69.71
3.3-70B vs. 4-Maverick	58.64	61.64	**68.02**	57.76	67.58	73.22

Table 3: Within-family attribution accuracy (%) for Llama model pairs using logistic regression. Each row is a binary attribution task, with chance accuracy of 50%.

Teacher → Student	Behavioral	Stylometric	Lexical	Semantic	Combined	PN Embedding†
Llama 8B → Minitron 4B	66.67	69.49	71.24	70.68	**74.56**	80.58
Mistral 12B → Minitron 8B	99.31	99.44	99.12	**99.62**	**99.62**	99.87
Gemma 27B → Gemma 2B	69.92	67.98	69.80	72.74	**77.94**	88.47

Table 4: Teacher–student attribution accuracy (%) using logistic regression. Each row is a binary classification task with a chance level of 50%.

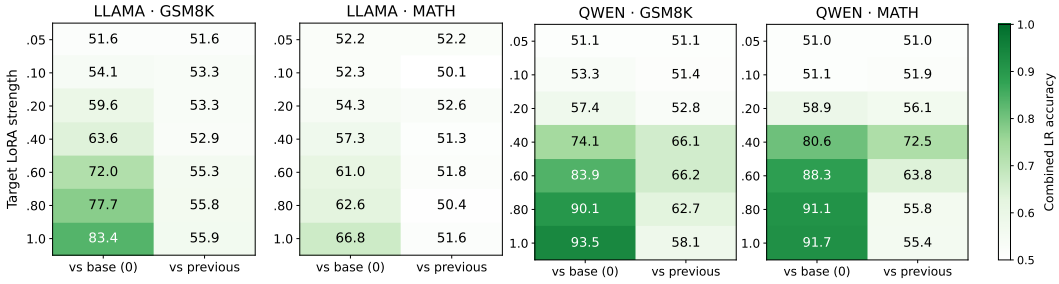
it under T1–T3. This shift is consistent with paraphrasing weakening lexical realization while T3 removes additional structural and surface signals. Even under P2*, all representations remain above chance, with Combined reaching 55.87%. Substantial cross-family information therefore remains recoverable after both lexical realization and visible surface form are heavily altered.

The Prompt-Neutralized Embedding (PN Embedding) achieves the highest accuracy across all views. We report it separately because it is a diagnostic, prompt-relative representation rather than a standard standalone classifier input. Unlike the other representations, PN Embedding requires peer responses from all candidate models to the same prompt so that their shared embedding mean can be removed. Its strong performance suggests that removing prompt- and task-shared variation exposes model-specific signals that are more discriminative and less dependent on the prompt itself. This finding also motivates the analysis of generalization across prompt distributions in Sec. 5. Corresponding results for the MLP classifier are provided in Appendix B. Across the reported settings, the largest standard deviation is 3.2%, indicating relatively low variability. To avoid overloading the tables, we omit standard deviations from the main presentation.

Within Model Families. Table 3 shows that model identity remains detectable among closely related Llama models, although accuracy is substantially lower than in the cross-family setting. The Combined representation performs best for most pairs, suggesting that distinctions between related models draw on multiple complementary signal groups. The Lexical representation is generally the strongest individual signal group, while behavioral and embedding representations provide weaker separation. Pairwise difficulty is not determined by parameter count alone: differences across model generations are often easier to detect than size differences within the same generation. Overall, within-family attribution appears to rely on several weak but complementary signals rather than a single dominant fingerprint.

The same broad pattern extends to the other model families, although fingerprint strength varies considerably across them. Lexical or Combined representations are typically strongest, while behavioral and embedding representations alone are less reliable. Full results are reported in Appendix C.

Teacher–Student Attribution. Table 4 shows that distilled students remain readily distinguishable from their teachers. Compared with the standard within-family comparisons in Tables 3 and 9, the teacher–student pairs are similarly or more separable. The Llama teacher–student pair is more separable than any of the Llama pairs, while the Mistral pair is nearly perfectly separable. The Gemma pair shows performance comparable to the corresponding within-family comparison. The Combined representation performs best for Llama and Gemma, whereas nearly all representations strongly separate the Mistral pair. PN Embedding provides additional gains, particularly for Llama



(a) Llama-3.1-8B-Instruct (b) Qwen3-8B

Figure 2: Controlled LoRA fingerprint resolution using Combined logistic-regression accuracy (%). *vs base* compares each strength with the unadapted model, and *vs previous* with the preceding strength. Chance is 50%.

and Gemma. Overall, shared model lineage does not imply similar fingerprints. The changes introduced through distillation can leave a strong and detectable model-specific signature.

Controlled LoRA Resolution. To study the finest model-resolution setting under controlled proximity, we train one LoRA adapter for each base model and adaptation dataset and vary only its contribution at inference time. For a fixed learned update ΔW_{LoRA}, each variant has effective weights $W_\alpha = W_0 + \alpha \Delta W_{\mathrm{LoRA}}$, with $\alpha \in \{0, .05, .10, .20, .40, .60, .80, 1.0\}$. We use Llama-3.1-8B-Instruct and Qwen3-8B adapted separately on GSM8K and MATH. The *vs base* setting compares each nonzero strength with $\alpha = 0$ and measures separation from the original model, while *vs previous* compares each strength with the immediately preceding one and measures how finely neighboring variants can be distinguished. Full details are provided in Appendix D.

Figure 2 shows a clear gap between these two settings. As LoRA strength increases, adapted models become progressively easier to distinguish from their bases, while neighboring variants remain much harder to separate. For Llama, *vs base* accuracy rises steadily, whereas *vs previous* remains close to chance across most strengths. Qwen shows stronger separation overall and clearer differences between some neighboring variants, especially at intermediate values of α; these local differences decrease again at higher strengths even as separation from the base continues to increase. Thus, increasing α does not produce a uniform increase in observable fingerprint change across models or adaptation tasks.

Table 11 shows the same overall pattern across Behavioral, Stylometric, Lexical, and Semantic representations. Neighboring Llama variants remain near chance across signal groups, while Qwen retains a modest local signal. Overall, adaptation can make a model increasingly distinguishable from its base even when successive variants remain difficult to resolve.

How Much Evidence Is Needed? Figure 3 examines how attribution changes with the amount and location of available text under P2*, our strongest transformation. Even 50 words retain useful identifying information, with accuracy of approximately 22% for the beginning and center and 27% for the ending, compared with a 12.5% chance level. Accuracy increases steadily with the amount of available text and reaches about 50% for the full response.



Figure 3: Cross-family attribution under P2* across word budgets and text positions.

The ending is the most informative region at every limited word budget, suggesting that later parts of a response retain more model-specific information after paraphrasing and surface normalization. The corresponding T0 and T3 analyses in Appendix E show the same accumulation pattern at higher accuracy levels. Overall, attribution

(a) Cross-dataset generalization

Train	Test	Accuracy (%)
D2 + D3 + D4	D1	84.19
D1 + D3 + D4	D2	81.41
D1 + D2 + D4	D3	68.63
D1 + D2 + D3	D4	67.88
Mean	–	**75.52**

(b) D3 attribution by task

Task	Acc. (%)	Macro-F1 (%)	# Prompts
Explanation	84.69	84.39	40
Argument / Advice	95.31	95.29	40
Creative	87.81	88.27	40
Rewrite	35.94	34.03	40
Source synthesis	39.38	36.00	40
Overall	**68.63**	**68.41**	200

Table 5: Cross-dataset attribution performance in the 8-way setting. (a) Accuracy when training on three prompt datasets and evaluating on the held-out fourth dataset. (b) Attribution performance on D3 by task type when training on D1, D2, and D4. Chance accuracy is 12.5%.

depends on both the amount of available text and the extent to which its identifying signals have been altered.

5 Do Fingerprints Generalize?

We next test whether model fingerprints generalize beyond the conditions under which they are learned. We consider two complementary shifts, changes in prompt distribution and changes in language, to assess whether attribution relies on signals that persist across prompts and languages.

5.1 Cross-Dataset Generalization

We train the attribution classifier on three prompt datasets and evaluate it on the held-out fourth, repeating this for each dataset. As shown in Table 5(a), accuracy remains well above the 12.5% chance level in every split, with a mean of 75.52%. Performance is higher on D1 and D2, which share a data source, than on the more distinct D3 and D4 distributions. Prompt distribution therefore affects fingerprint recoverability without erasing model identity.

The same pattern holds under stronger text transformations. Across T0, T3, and P2*, holding out the target dataset reduces Combined accuracy by only 4.8–5.9 points relative to mixed-source evaluation, whereas transforming the responses causes substantially larger losses. Lexical nearly matches Combined on raw text and is strongest among the standard representations under T3, while Combined performs better under P2* after paraphrasing weakens lexical information. Distribution shift therefore has a measurable but smaller effect than directly removing or rewriting model-identifying signals.

Table 5(b) shows stronger variation across D3 task types. Attribution remains reliable for explanation, argument/advice, and creative generation, but drops sharply for rewriting and source synthesis. These tasks constrain more of the response content and structure, leaving less room for model-specific variation. This pattern is also consistent with the strong PN Embedding results in Appendix F.1, where removing prompt-shared variation makes model differences more visible. Detailed transformed and per-model results are provided in Appendix F.2.

5.2 Within-Language

To assess fingerprint recoverability within each language, we use an LLM to translate the English prompts into Persian and Mandarin Chinese and then perform the same eight-way cross-family attribution task on responses generated in each language. The translation procedure and prompt are provided in Appendix F.3.

The Persian and Mandarin Chinese results in Appendix G follow the same broad pattern as English. Attribution decreases as surface signals are removed, drops further under paraphrasing, and is lowest under P2*. Even then, all representations remain above the 12.5% chance level. Lexical is strongest across most signal-suppressed views, while Combined becomes more competitive after paraphrasing. PN Embedding improves over the standard Semantic representation in every view, but

Train → Test	Behavioral	Stylometric	Semantic	Combined	PN Embedding†
EN → FA	34.54	15.38	32.63	**34.82**	44.75
EN → ZH	35.01	16.76	**40.23**	39.00	59.10
FA → EN	**21.77**	21.04	16.81	19.67	47.99
FA → ZH	26.39	21.98	24.40	**34.76**	55.78
ZH → EN	27.30	20.68	30.03	**36.39**	59.46
ZH → FA	32.86	22.62	**39.36**	35.53	51.07

Table 6: Cross-lingual 8-way model attribution accuracy (%) on raw outputs (T0) using logistic regression. EN denotes English, FA Persian, and ZH Mandarin Chinese. Chance accuracy is 12.5%.

usually remains below the strongest Lexical or Combined results in Persian and Chinese. Overall, fingerprint persistence extends beyond English, although its strength varies across languages.

5.3 Cross-Lingual

Because word- and character-level lexical features are strongly tied to language and script, we exclude the Lexical representation from the cross-lingual transfer experiments. Table 6 evaluates transfer by training the classifier in one language and testing it in another. Accuracy remains above chance in all six directions. Transfer is weaker than within-language attribution and varies across language pairs, showing that only part of the model-specific signal is shared across languages.

The feature results show where this transferable signal is concentrated. Stylometric features transfer poorly, consistent with their dependence on language-specific grammatical and stylistic patterns. Behavioral features remain useful in several directions, while Semantic and Combined representations are generally stronger. PN Embedding improves transfer in every direction, often substantially, indicating that removing prompt-shared variation reveals a more stable model-relative signal.

Appendix H reports average cross-lingual attribution performance across all six language-transfer settings under each response transformation. Accuracy declines from T0 through T3 as surface signals are removed, but remains above chance after severe Transformation. Paraphrasing alone has much less effect, with P1 and P2 retaining transfer close to the raw setting. The largest degradation appears under P1* and P2*, where paraphrasing is combined with T3. Overall, cross-lingual attribution appears to rely on cross-linguistically shared behavioral and structural patterns, together with more persistent model-specific signals that survive substantial changes in wording and form.

6 Conclusion

We presented a multi-resolution study of LLM fingerprinting across model families, closely related models, distilled and LoRA-adapted variants, text transformations, prompt distributions, and languages. Model identity remains detectable at increasingly fine levels of similarity, although neighboring LoRA variants can be difficult to resolve. The identifying signal is distributed across multiple forms of evidence. Surface and behavioral signals are relatively fragile, lexical information remains strong after surface normalization, and combining signal types becomes more useful after paraphrasing. Fingerprints also transfer substantially across prompt distributions and languages, but recoverability depends strongly on the task. Attribution drops sharply for highly constrained tasks such as rewriting and source synthesis, where the prompt leaves less freedom for model-specific variation.

These findings support a view of LLM fingerprints as layered rather than tied to a single observable feature. Our fingerprint-resolution framework makes this dependence explicit by studying both how similar the candidate models are and how much identifying evidence remains available. The strong performance of PN Embedding further shows that removing variation shared across responses to the same prompt can substantially improve model separation, especially when task constraints are strong. Together, our results show that fingerprint recoverability depends jointly on model similarity, signal type, prompt and task constraints, and the transformations applied to the text, providing a broader diagnostic view of where model-specific behavior resides and when it remains observable.

LIMITATIONS

Although our prompt sets are designed to cover diverse and largely general-purpose requests, they cannot represent the full range of prompts, interaction styles, and usage patterns encountered in everyday use. The multilingual experiments are limited to English, Chinese, and Persian, with non-English prompts translated from the English set rather than independently authored in each language. The controlled LoRA analysis uses two base models and two reasoning adaptations and may not reflect other adaptation methods or domains. Our evaluation also assumes a closed set of candidate models, whereas real-world provenance may involve unknown or unseen generators. PN Embedding is used as a diagnostic representation because it requires responses from all candidate models to the same prompt and is therefore not applicable to standard single-response attribution. Finally, our transformation suite cannot capture the full range of human editing, rewriting, translation, and post-processing that model outputs may undergo in practice.

AI USE STATEMENT

In this work, we use generative AI tools to assist with implementing and scaling experimental code. The experimental logic, configurations, and parameters are specified and reviewed by the authors. AI-assisted code is inspected and tested, including through unit tests, before use.

Additionally, we use generative AI tools to identify potentially relevant literature and to improve the clarity and readability of the manuscript. All cited works are reviewed directly by the authors before inclusion. AI-assisted writing is used for editing and polishing author-written content rather than as a substitute for author review. All resulting text is manually checked, revised, and approved by the authors.

We also use generative AI assistance in the construction of a prompt dataset, following the procedure described in Appendix A.2. The resulting data and construction process are reviewed by the authors.

We take responsibility for the final content of this work, including text, code, data, claims, and other artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our method is intended as a diagnostic tool for studying model-specific generation patterns, not as a definitive detector of AI-generated text or as a standalone basis for decisions about individuals. Attribution errors remain possible, particularly under distribution shifts and text transformations, and results should therefore be interpreted with appropriate caution in high-stakes settings. We release the generated model outputs to support reproducibility and further research. Because the dataset contains more than 15,000 generations, the outputs were not exhaustively screened for potentially harmful, offensive, or otherwise undesirable content. They reflect the behavior of the source models at the time of generation and should not be interpreted as endorsed content. Downstream users are responsible for applying appropriate safeguards and complying with the relevant model and data-use terms.

REPRODUCIBILITY STATEMENT

The main text and appendices describe the experimental design, model and feature configurations, transformations, evaluation protocols, and additional implementation details needed to reproduce our results. We also make available the prompt datasets used in the study, all generated outputs comprising more than 15,000 generations from over 19 models, and the code used for model generation, LoRA fine-tuning, feature extraction, and classifier training and evaluation. These materials will be made publicly available on GitHub[1] to support independent verification and extension of the experiments.

[1] https://anonymous.4open.science/r/Fingerprint-Resolution

REFERENCES

Wissam Antoun, Benoît Sagot, and Djamé Seddah. From text to source: Results in detecting large language model-generated content, 2024. URL https://arxiv.org/abs/2309.13322.

Yehonatan Bitton, Elad Bitton, and Shai Nisan. Detecting stylistic fingerprints of Large Language Models, 2025. URL https://arxiv.org/abs/2503.01659.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. RAID: A shared benchmark for robust evaluation of machine-generated text detectors. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12463–12492, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.674. URL https://aclanthology.org/2024.acl-long.674/.

Irena Gao, Percy Liang, and Carlos Guestrin. Model equality testing: Which model is this api serving? In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 86369–86382, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/d73234a13815fc1f9779dd17d89be9b4-Paper-Conference.pdf.

Thibaud Gloaguen, Robin Staab, Nikola Jovanović, and Martin Vechev. LLM fingerprinting via semantically conditioned watermarks. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=t38nZqqi3Z.

Matteo Greco, Anudeex Shetty, Andrea Tagarelli, and Jey Han Lau. Multighostbench: A multilingual benchmark for long-form llm-generated text attribution under distribution shifts, 2026. URL https://arxiv.org/abs/2609.02379.

Xun Guo, Yongxin He, Shan Zhang, Ting Zhang, Wanquan Feng, Haibin Huang, and Chongyang Ma. Detective: Detecting AI-generated text via multi-level contrastive learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=cdTTTJfJe3.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset, 2021. URL https://arxiv.org/abs/2103.03874.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, and Yejin Choi. Artificial hivemind: The open-ended homogeneity of language models (and beyond). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2026. URL https://openreview.net/forum?id=saDOrrnNTz.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Frederick Wieting, and Mohit Iyyer. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=WbFhFvjjKj.

Lucio La Cava and Andrea Tagarelli. OpenTuringBench: An open-model-based benchmark and framework for machine-generated text detection and attribution. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26655–26671, Suzhou,

China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1354. URL https://aclanthology.org/2025.emnlp-main.1354/.

Lucio La Cava, Dominik Macko, Robert Moro, Ivan Srba, and Andrea Tagarelli. Authorship attribution in multilingual machine-generated texts. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 45136–45152, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.2091. URL https://aclanthology.org/2026.acl-long.2091/.

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. MAGE: Machine-generated text detection in the wild. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 36–53, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.3. URL https://aclanthology.org/2024.acl-long.3/.

Xingyu Lu, Yumeng Ma, Xiang Zhou, Shengli Gan, Guiying Deng, Yang Wen, and Yanbing Liu. Synergizing stylometrics with semantics: Dual-path framework for LLM detection and attribution. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 37252–37265, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1855. URL https://aclanthology.org/2026.findings-acl.1855/.

Hope McGovern, Rickard Stureborg, Yoshi Suhara, and Dimitris Alikaniotis. Your large language models are leaving fingerprints. In Firoj Alam, Preslav Nakov, Nizar Habash, Iryna Gurevych, Shammur Chowdhury, Artem Shelmanov, Yuxia Wang, Ekaterina Artemova, Mucahid Kutlu, and George Mikros (eds.), *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pp. 85–95, Abu Dhabi, UAE, January 2025. International Conference on Computational Linguistics. URL https://aclanthology.org/2025.genaidetect-1.6/.

Dario Pasquini, Evgenios M. Kornaropoulos, and Giuseppe Ateniese. LLMmap: Fingerprinting for large language models. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 299–318, Seattle, WA, August 2025. USENIX Association. ISBN 978-1-939133-52-6. URL https://www.usenix.org/conference/usenixsecurity25/presentation/pasquini.

Nazneen Rajani, Lewis Tunstall, Edward Beeching, Nathan Lambert, Alexander M. Rush, and Thomas Wolf. No robots. https://huggingface.co/datasets/HuggingFaceH4/no_robots, 2023.

Zixin Rao, Youssef Mohamed, Shang Liu, and Zeyan Liu. Two birds with one stone: Multi-task detection and attribution of llm-generated text, 2025. URL https://arxiv.org/abs/2508.14190.

Mingjie Sun, Yida Yin, Zhiqiu Xu, J. Zico Kolter, and Zhuang Liu. Idiosyncrasies in Large Language Models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 57854–57885. PMLR, 2025. URL https://proceedings.mlr.press/v267/sun25z.html.

Teppei Suzuki, Ryokan Ri, and Sho Takase. Natural fingerprints of Large Language Models, 2025. URL https://arxiv.org/abs/2504.14871.

Yao Tong, Haonan Wang, Siquan Li, Kenji Kawaguchi, and Tianyang Hu. Seedprints: Fingerprints can even tell which seed your large language model was trained from. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Kan6Z0zzZi.

Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. Authorship attribution for neural text generation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020*

Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 8384–8395, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.673. URL https://aclanthology.org/2020.emnlp-main.673/.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, Thomas Arnold, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. In Yvette Graham and Matthew Purver (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1369–1407, St. Julian's, Malta, March 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.eacl-long.83. URL https://aclanthology.org/2024.eacl-long.83/.

Jiashu Xu, Fei Wang, Mingyu Ma, Pang Wei Koh, Chaowei Xiao, and Muhao Chen. Instructional fingerprinting of large language models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3277–3306, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.180. URL https://aclanthology.org/2024.naacl-long.180/.

A Detailed Experimental Configuration

A.1 Models and Generation

The full 19-model panel is grouped by family below.

- **Llama**
  - `meta-llama/llama-3.1-8b-instruct`
  - `meta-llama/llama-3.1-70b-instruct`
  - `meta-llama/llama-3.3-70b-instruct`
  - `meta-llama/llama-4-maverick`
- **Qwen**
  - `qwen/qwen3-8b`
  - `qwen/qwen3-14b`
  - `qwen/qwen3-32b`
- **Mistral**
  - `mistralai/mistral-small-24b-instruct-2501`
  - `mistralai/mistral-small-3.2-24b-instruct`
- **Gemma**
  - `google/gemma-4-26b-a4b-it`
  - `google/gemma-4-31b-it`
- **DeepSeek**
  - `deepseek/deepseek-v4-flash`
  - `deepseek/deepseek-v4-flash-0731`
- **Anthropic**
  - `anthropic/claude-haiku-4.5`
  - `anthropic/claude-sonnet-4`
- **Gemini**
  - `google/gemini-2.5-flash`
  - `google/gemini-2.5-flash-lite`
- **OpenAI**

Table 7: Task categories in the D3 control set.

Task	Description
Explanation	Explain a concept clearly and accurately to a general reader.
Argument and advice	Evaluate a situation, weigh alternatives, and provide a justified recommendation.
Creative writing	Produce an original story or scene under specified narrative conditions.
Rewriting	Transform a supplied passage while preserving its central meaning.
Source synthesis	Combine four supplied source notes into a coherent response.

- `openai/gpt-4.1`
- `openai/gpt-4.1-mini`

Cross-family representatives are selected using a fixed rule. We use the newest included release in each family, or the largest general-purpose instruction model when multiple included models belong to the same release.

A.2 Design and Construction of the D3 Control Set

D3 is designed as a controlled complement to the naturally occurring prompt collections used elsewhere in the study. The authors first specify five task categories that differ in their objectives, response constraints, and expected difficulty: *explanation*, *argument and advice*, *creative writing*, *rewriting*, and *source synthesis*. Table 7 describes each task.

After defining this structure, we use GPT-5.6 Sol to generate 40 prompts for each category, producing 200 prompts in total. The model is instructed to maintain the predefined category boundaries while varying topics, scenarios, and requested outputs. The authors determine the task taxonomy, category balance, and construction requirements, while GPT-5.6 Sol is used only to draft the individual prompt texts.

All generated prompts subsequently undergo human review. Each prompt is evaluated for clarity, category fit, duplication, and whether it provides a meaningful task for every candidate model. The final set retains an equal number of prompts from each category. Prompt selection is completed before collecting the model responses and does not use attribution results or downstream classifier performance.

This construction provides several experimental benefits. First, exact category balance prevents the distribution from being dominated by a common task type. Second, the categories span different levels of response freedom, from open-ended creative generation to tightly constrained rewriting and source-grounded synthesis. This variation makes it possible to examine whether model fingerprints become weaker when the task restricts the linguistic choices available to the generator. Finally, constructing a novel prompt set reduces the risk that evaluated models reproduce memorized responses to established benchmark questions. Therefore, D3 is an *AI-assisted, researcher-designed, and human-reviewed control set*.

A.3 Generation Setting

Raw responses are generated with temperature 1.0, top-$p = 0.9$, and a maximum of 2,048 output tokens. Raw generations are stored before any transformation. Response embeddings use `text-embedding-3-small` with 1,536 dimensions. We also evaluated several open-source embedding models. Among the models tested, `text-embedding-3-small` achieved the strongest attribution performance and is therefore used throughout the reported experiments.

Throughout the paper, we use model fingerprint to denote source-specific traces in the output of a generator under a fixed generation setup. In black-box API settings, such traces may reflect both the underlying model and components of the serving stack. Our controlled LoRA experiments complement this setting by holding the base model and generation protocol fixed while varying

the contribution of a learned weight update, showing that weight-level changes themselves can also produce detectable traces.

A.4 Complete Signal Feature Definitions

This section provides the complete definitions of the signal representations introduced in Sec. 3.2. Features are recomputed for every response view. All learned preprocessing, including feature standardization and TF–IDF vocabulary construction, is fitted only on the training portion of each fold.

Behavioral and structural features. We use 25 features describing response length, organization, formatting, punctuation, and repetition:

- **Length and structure:** word count, character count, sentence count, paragraph count, mean sentence-span length, and standard deviation of sentence-span length.
- **Formatting and layout:** Markdown heading count, bullet/list count, code-fence count, newline count, newline-to-character ratio, and counts of #, $\star$, hyphens, underscores, and backticks.
- **Punctuation:** counts of commas, periods, exclamation marks, question marks, colons, semicolons, and quotation marks.
- **Repetition:** repeated-bigram and repeated-trigram ratios.

For an n-gram sequence, the repetition ratio is the proportion of repeated n-grams relative to all observed n-grams. It is set to zero when the response is too short to form the corresponding n-gram.

Stylometric features. We use 17 dense features:

- **Vocabulary diversity:** token count, unique-token count, type–token ratio, MATTR-10, and MATTR-50.
- **Word form:** mean and standard deviation of word length.
- **Vocabulary distribution:** hapax ratio and most-common-word frequency ratio.
- **Function and grammatical words:** function-word ratio, pronoun ratio, modal ratio, and connective ratio.
- **Repetition and numeric use:** unigram, bigram, and trigram repetition ratios, and numeric-token ratio.

MATTR is computed as the mean type–token ratio over sliding windows of 10 or 50 tokens. For responses shorter than the specified window, the type–token ratio of the available response is used. The hapax ratio measures the proportion of vocabulary items that occur exactly once.

Part-of-speech features. The dense stylometric features are concatenated with POS unigram, bigram, and trigram TF–IDF features. POS sequences are extracted using spaCy `en_core_web_sm`. The POS vectorizer uses `min_df=2` and retains at most 2,500 features. The resulting representation is the **Stylometric** representation used throughout the paper.

Lexical features. We concatenate two TF–IDF representations:

- lowercase word unigrams and bigrams, with at most 5,000 features;
- lowercase character 3–5-grams, with at most 5,000 features.

Both use min_df=2 and sublinear term-frequency scaling. The combined representation therefore contains at most 10,000 dimensions.

Semantic features. Each response is embedded using `text-embedding-3-small`, producing a 1,536-dimensional vector. Every transformed or paraphrased response is embedded after the transformation is applied; embeddings from the raw response are not reused. Embedding dimensions are standardized using statistics computed from the training fold.

Although these representations are primarily designed to capture semantic information, they can also retain lexical, stylistic, and structural signals. This is reflected in our results, where attribution performance using the Semantic representation declines as progressively stronger text transformations are applied. We therefore use the term *Semantic* to denote the representation's primary purpose, rather than to imply that it encodes semantic information exclusively.

Prompt-neutralized embeddings. For each prompt and response view, we compute the mean embedding across all candidate-model responses and subtract it from each individual response embedding:

$$\bar{e}_{p,t} = \frac{1}{K} \sum_{m=1}^{K} e_{p,m,t}, \qquad e'_{p,m,t} = e_{p,m,t} - \bar{e}_{p,t}.$$

This removes components shared across responses to the same prompt and emphasizes relative differences between candidate models. Because this representation requires access to the other candidate responses for the same test prompt, we use it only as a diagnostic representation rather than as a standard single-response attribution setting.

Combined representation. The combined representation concatenates the four main signal groups: behavioral and structural features, stylometric and POS features, word and character TF–IDF, and frozen embeddings. Same-prompt centered embeddings are excluded. Dense behavioral, lexical, and embedding features are standardized using training-fold statistics, while POS and word/character TF–IDF features retain their sparse representations.

A.5 Transformation and Paraphrasing Details

T_0–T_3. T_0 is the unmodified response. For T_1, responses with at most 10 words are unchanged; responses with 11–20 words lose one word from each side; responses with 21–50 words lose five words from each side; and responses longer than 50 words lose 15 words from each side. T_2 is applied to T_1 and removes Markdown heading, list, blockquote, and code-fence markers; emphasis and inline-code delimiters; Markdown link/image syntax while retaining visible text; HTML tags; indentation; and repeated layout whitespace. Ordinary punctuation, capitalization, lexical content, and word order are preserved. T_3 is applied to T_2 and performs Unicode NFKC normalization, lowercasing, punctuation/symbol/emoji removal, and whitespace collapse while preserving lexical words, function words, numbers, and word order.

P_1 and P_2. Both paraphrase conditions branch from T_0 and use `gpt-4o-mini-2024-07-18` with temperature 0, top-$p = 1$, a maximum of 2,048 output tokens, and reasoning disabled. The paraphraser never receives the source-model identity. For P_1, approximately one third of eligible natural-language sentences are selected deterministically using prompt identity rather than source label; headings, structural fragments, and executable code are not paraphrase targets. All selected targets from one response are rewritten together, while untouched text is restored unchanged. If a response contains no eligible prose, P_1 leaves it unchanged. P_2 rewrites the complete response while preserving its meaning, factual claims, entities, numbers, qualifications, language, and approximate amount of information. We require successful, non-empty outputs but do not filter paraphrases according to downstream attribution performance.

A.5.1 P_1: Partial Paraphrasing Prompt

For P_1, the target indices, full response, and selected sentences are inserted automatically for each example. JSON and output-schema instructions are omitted below for readability.

```
You are performing controlled partial editing for a scientific
experiment.

The full passage is provided for context.  Rewrite ONLY the target
sentences listed below.

For EACH supplied target, substantially change its wording and
syntax while preserving its meaning, facts, entities, numbers,
qualifications, references, tone, and approximately the same sentence
length.

Do not rewrite or return any non-target sentence.

Requested indices:  [target indices]

FULL PASSAGE:
[full response]

TARGETS:
[selected target sentences]
```

A.5.2 P_2: FULL-RESPONSE PARAPHRASING PROMPT

For P_2, the complete response and response-specific word-count bounds are inserted automatically.

```
You are a controlled paraphrasing system used for a scientific
text-transformation experiment.

Rewrite the supplied text thoroughly while preserving its meaning
exactly.

Requirements:

- Rewrite wording and sentence structure throughout the passage.  Do
not merely replace a few words with synonyms.

- Preserve every factual claim, named entity, number, example,
qualification, uncertainty, instruction, and conclusion.

- Do not add new information, explanations, opinions, or facts.

- Do not remove information or summarize.

- Preserve the original language, tone, level of formality, and broad
paragraph/list organization.

- The original has approximately [word count] whitespace-delimited
words.  Target [lower bound]--[upper bound] words; do not compress or
expand the content.

- Do not add an introduction such as 'Here is the paraphrased
version.'

- Return only the rewritten text.

TEXT TO PARAPHRASE:
[full response]
```

The P_2 length bounds are computed separately for each response as approximately 85%–115% of its original whitespace-delimited word count. They are used to guide the paraphraser toward preserving the amount of information rather than as a post-generation acceptance criterion.

A.6 CLASSIFIER AND EVALUATION DETAILS

Logistic Regression uses an L_2 penalty with the SAGA solver. Its regularization parameter is selected from $C \in \{0.01, 0.1, 1, 10\}$ using macro-F1 on an inner validation split. The MLP has one hidden layer of 256 units with ReLU activation and is trained with Adam using learning rate 10^{-3}, weight decay 10^{-4}, and batch size 128. Training is capped at 80 epochs, with early stopping after

View	Behavioral	Stylometric	Lexical	Semantic	Combined	PN Embedding†
T0	59.92	66.82	79.74	68.10	**82.01**	84.46
T1	55.02	63.39	75.91	61.53	**77.53**	77.08
T2	47.47	60.36	72.36	56.23	**73.12**	74.56
T3	29.41	53.13	67.98	51.71	**68.21**	70.06
P1	59.79	65.08	74.12	65.05	**78.63**	81.62
P2	56.65	60.09	68.29	58.62	**72.68**	75.99
P1*	28.44	50.00	61.19	48.64	**62.69**	66.38
P2*	26.69	45.67	53.57	43.30	**56.54**	61.05

Table 8: Cross-family attribution accuracy (%) using MLP. Chance accuracy is 12.5%.

Family	Model Pair	Behavioral	Stylometric	Lexical	Semantic	Combined	PN Embedding†
DeepSeek	V4 vs. V4-0731	58.14	62.77	**68.34**	60.45	66.65	72.72
Gemma	4-26B vs. 4-31B	57.70	61.64	**79.10**	60.95	78.35	78.97
Mistral	Small-24B vs. Small-3.2-24B	74.28	86.05	94.24	90.86	**94.37**	98.75
Qwen	8B vs. 14B	56.45	58.51	57.76	54.44	**62.33**	62.33
Qwen	8B vs. 32B	63.77	71.15	69.46	69.09	**73.78**	85.98
Qwen	14B vs. 32B	64.83	70.84	71.09	63.58	**74.97**	84.48
Anthropic	Haiku-4.5 vs. Sonnet-4	69.65	69.90	77.28	79.47	**82.85**	92.87
Gemini	2.5-Flash vs. 2.5-Flash-Lite	59.01	64.27	**71.15**	59.89	70.53	74.72
OpenAI	GPT-4.1 vs. GPT-4.1-mini	63.52	72.09	78.66	70.65	**79.41**	87.61

Table 9: Within-family attribution accuracy (%) for non-Llama model pairs using logistic regression. Each row is a binary attribution task, with chance accuracy of 50%.

eight epochs without improvement. The selected number of epochs is then used when refitting on the full outer-training fold.

The main evaluation uses five-fold prompt-grouped, dataset-stratified cross-validation. Each outer fold therefore uses approximately 80% of the prompts for training and 20% for testing. Within each outer-training fold, 20% of prompts from each dataset are used for hyperparameter selection. All responses from the same prompt, across models and transformed views, remain grouped together, and the same outer and inner partitions are used across experimental conditions. Predictions from the five outer folds are concatenated before computing accuracy and macro-F1. For the balanced eight-way cross-family task, chance accuracy is 12.5%, while within-family binary tasks have a 50% chance level and are interpreted separately.

B Cross-Family MLP Results

Table 8 reports performance on the eight-way cross-family attribution task using a one-hidden-layer MLP classifier. The MLP generally performs slightly better than Logistic Regression across most representations, while Lexical performance remains nearly unchanged. Importantly, the relative ordering of representations and the degradation across transformations remain similar across classifiers.

C Within-Family Results for Additional Model Families

Table 9 shows substantial variation in fingerprint strength across model families. Some closely related models remain highly separable, while others achieve only modest gains over chance. Despite this variation, the overall signal pattern is consistent with the Llama results. The Lexical and Combined representations generally perform best, indicating that lexical evidence is central but benefits from additional stylistic and structural signals. Behavioral and embedding representations are less consistent on their own, suggesting that within-family identity is distributed across multiple forms of

Family	Model Pair	Behavioral	Stylometric	Lexical	Semantic	Combined	PN Embedding†
Llama	3.1-70B vs. 3.3-70B	62.95	60.20	59.51	61.45	**66.40**	72.22
Llama	3.1-70B vs. 4-Maverick	60.89	64.89	68.90	61.45	**72.09**	73.59
Llama	3.1-8B vs. 3.1-70B	55.63	58.89	58.70	57.07	**61.39**	65.33
Llama	3.1-8B vs. 3.3-70B	60.76	59.51	59.01	57.45	**62.39**	68.52
Llama	3.1-8B vs. 4-Maverick	58.82	64.14	68.27	59.39	**70.96**	70.34
Llama	3.3-70B vs. 4-Maverick	59.45	61.45	65.71	59.45	**67.52**	74.16
Qwen	8B vs. 14B	54.57	59.76	58.70	54.88	**61.14**	62.20
Qwen	8B vs. 32B	67.02	73.03	69.15	68.15	**73.78**	85.61
Qwen	14B vs. 32B	66.65	72.28	70.71	64.71	**74.72**	80.41
Anthropic	Haiku-4.5 vs. Sonnet-4	77.60	70.53	75.16	79.29	**82.54**	92.30
Gemini	2.5-Flash vs. 2.5-Flash-Lite	60.14	63.08	70.03	60.89	**72.03**	74.59
OpenAI	GPT-4.1 vs. GPT-4.1-mini	66.33	72.47	78.72	73.65	**80.23**	88.17
DeepSeek	V4 vs. V4-0731	61.39	64.27	**67.90**	60.89	65.89	72.47
Gemma	4-26B vs. 4-31B	54.44	62.14	**78.79**	62.58	78.35	78.91
Mistral	Small-24B vs. Small-3.2-24B	83.23	87.30	93.05	90.93	**94.37**	98.62

Table 10: Within-family attribution accuracy (%) for all model pairs using a one-hidden-layer MLP. Each row is a binary attribution task, with chance accuracy of 50%.

evidence. Table 10 reports the corresponding results for a one-hidden-layer MLP classifier. Neither classifier consistently outperforms the other by a large margin across settings.

D Controlled LoRA Adaptations

For each model–dataset pair defined in the main text, we use 1,200 training examples after removing prompts that overlap with the D1–D4 evaluation sets. MATH examples are sampled across all seven subject categories. Training uses supervised next-token prediction on the assistant response, with prompt tokens masked from the loss.

Adapters are trained for one epoch with maximum sequence length 512, learning rate 2×10^{-4}, per-device batch size 16, warmup ratio 0.03, and no weight decay. We use AdamW with cosine scheduling, FP16, and no quantization. LoRA uses rank $r = 16$, scaling parameter 32, dropout 0.05, no bias adaptation, and all linear layers as target modules.

Each LoRA variant generates responses to the same D1–D4 prompts used in the main experiments. Generation uses temperature 1.0, top-$p = 0.9$, a maximum of 2,048 new tokens, one generation per prompt, and no additional system prompt. Qwen3 thinking is disabled. Generation is stochastic, and outputs are not treated as paired across LoRA strengths.

Attribution uses raw T0 responses, Logistic Regression, and the same five-fold prompt-grouped split as the main experiments. The 799 prompt blocks yield 1,598 responses per binary comparison. The shared 0-vs-0.05 comparison is evaluated once, resulting in 52 unique binary tasks across the four model–dataset settings. Behavioral, Stylometric, Lexical, Semantic, PN Embedding, and Combined representations use the same definitions as in the main experiments.

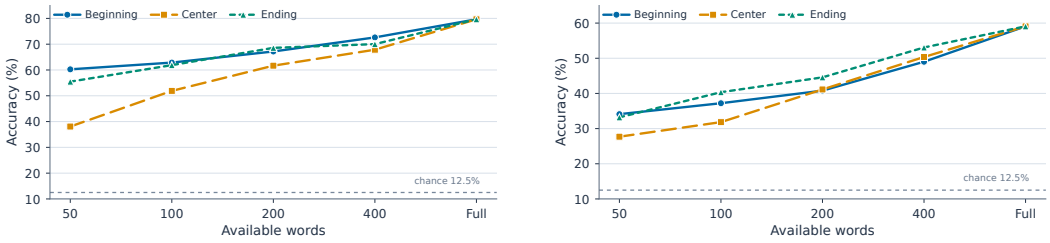
Table 11 reports results for the four standard signal groups, averaged across GSM8K and MATH and across all strength comparisons. The feature-level results follow the pattern observed in Figure 2, with substantially stronger separation from the base model than between neighboring variants across signal groups.

E Evidence-Amount

Within-response evidence experiments use exact windows of $N \in \{50, 100, 200, 400\}$ words. The center window is used as the primary location, with additional prefix, center, and suffix comparisons at selected budgets. Figure 4 shows Logistic Regression accuracy for T0 (raw text) and T3 in the eight-way attribution setting as the number of available words increases. For this experiment, we include only responses containing at least 400 words so that the same set of responses is evaluated across all evidence budgets.

	Llama-3.1-8B		Qwen3-8B	
Feature	**vs Base**	**vs Previous**	**vs Base**	**vs Previous**
Behavioral	59.26	**52.52**	68.44	**56.56**
Stylometric	58.37	51.48	68.89	55.99
Lexical	**60.20**	51.54	**70.46**	56.51
Semantic	58.48	51.07	69.47	55.35

Table 11: LoRA-variant attribution accuracy (%) using logistic regression. *vs Base* distinguishes each adapted model from its unadapted base model, whereas *vs Previous* distinguishes each LoRA strength from the immediately preceding strength. Results are averaged across GSM8K and MATH adaptations and all LoRA-strength comparisons. Binary chance accuracy is 50%.



(a) Original outputs **(b)** T3

Figure 4: Attribution accuracy as a function of available text in the 8-way cross-family setting. **(a)** Original model outputs. **(b)** Outputs after the T3 surface transformation. Prefix, center, and suffix indicate the location from which the available words are selected.

Because this analysis is restricted to responses containing at least 400 words so that the evaluation set remains fixed across word budgets, its Full-response value is not expected to exactly match the all-response result in Table 2.

F Task-Level Variation and Prompt Effects in Model Attribution

F.1 Does Removing Prompt Effects from Embeddings Improve Source Attribution?

We hypothesize that subtracting the prompt-wise mean response embedding reduces variation shared across candidate models, much of which reflects the prompt and task, and makes model-specific differences easier to detect. Table 14 supports this hypothesis. Centering increases the overall accuracy of embedding-based attribution from 50.50% to 74.88%.

The comparison with the other representations is particularly informative. PN Embedding is comparable to Lexical on explanation and argument/advice tasks and slightly stronger on creative writing. Its clearest advantage appears on the more constrained tasks. For rewriting and source synthesis, PN Embedding exceeds the strongest standard representation by 12.81 and 13.75 points, respectively. These tasks restrict the content and structure of the response, causing outputs from different models to share more task-related information. Removing this shared component exposes model-specific variation that is less visible in the original embedding space. Because centering requires responses from all candidate models to the same prompt, it is best understood as a prompt-relative analysis of model separation rather than a standard attribution setting.

F.2 D3 Task-Level Attribution by Model

Table 12 reports per-model recall across the five D3 task categories under cross-dataset generalization. The results show substantial variation across task types and models. Recall is generally higher for explanation, argument/advice, and creative generation, whereas rewrite and source-synthesis

Model	Explanation	Argument	Creative	Rewrite	Synthesis
Llama-4 Maverick	97.50	97.50	**100.00**	55.00	90.00
Qwen3-32B	80.00	97.50	82.50	12.50	40.00
Mistral Small 3.2	70.00	92.50	75.00	47.50	20.00
Gemma-4-31B	**100.00**	**100.00**	85.00	30.00	12.50
DeepSeek-V4	87.50	92.50	**100.00**	15.00	15.00
Claude Haiku 4.5	97.50	**100.00**	95.00	**75.00**	**100.00**
Gemini-2.5 Flash	77.50	**100.00**	90.00	40.00	22.50
GPT-4.1	67.50	82.50	75.00	12.50	15.00

Table 12: Per-model recall (%) across the five D3 task categories under cross-dataset generalization. The classifier is trained on D1, D2, and D4 and evaluated on D3. Chance recall is 12.5%.

(a) Mixed-source evaluation

View	Behavioral	Stylometric	Lexical	Semantic	Combined	PN Embedding†
T0	52.42	60.72	80.07	66.04	**81.41**	83.75
T3	25.59	49.62	**67.76**	50.94	67.04	69.35
P2*	23.08	41.79	53.94	41.90	**55.87**	59.39

(b) Held-out-dataset evaluation

View	Behavioral	Stylometric	Lexical	Semantic	Combined	PN Embedding†
T0	50.70	56.83	75.45	61.13	**75.52**	80.75
T3	25.14	47.13	**64.32**	46.05	62.25	66.45
P2*	22.86	39.05	49.61	38.51	**51.10**	56.29

Table 13: Attribution accuracy (%) under mixed-source and held-out-dataset evaluation. T0 uses the original responses, T3 removes surface cues, and P2* combines full paraphrasing with surface normalization. Chance accuracy is 12.5%.

tasks show lower recall for several models. Thus, the aggregate task-level differences in the main text are not uniform across all models. These differences also have practical implications for provenance analysis. In rewrite and source-synthesis workflows, the lower attribution accuracy may make it more difficult to identify the model responsible for a given output after it has been transformed or conditioned on source material. Conversely, when model provenance is important, task-specific attribution performance can be considered when evaluating or auditing model outputs rather than relying on results from more open-ended generation alone.

Table 13 compares mixed-source and held-out-dataset accuracy under the three response views.

F.3 Multilingual Prompt Translation

For the multilingual experiments, we translate the D1–D4 English prompts into Persian and Mandarin Chinese before generating model responses. Translation uses `gpt-4o-mini-2024-07-18` with temperature 0, top-p = 1, and a maximum of 4,096 output tokens. Persian translations use standard Persian script, while Mandarin translations use Simplified Chinese characters.

The translation procedure is designed to preserve the original task rather than adapt or execute it. The model is instructed to return only the translated prompt, preserve names, dates, numbers, constraints, requested tone, and meaningful line breaks, and retain literal fragments such as code, URLs, identifiers, and placeholders when appropriate. Prompts requesting code, diagrams, images, or other artifacts are translated as instructions rather than executed. We use the following system prompt, where `[TARGET LANGUAGE]` is replaced by the corresponding language specification:

Task	Behavioral	Stylometric	Lexical	Semantic	Combined	PN Embedding†
Explanation	62.19	72.81	**87.50**	55.00	84.69	86.56
Argument / Advice	82.19	87.81	**96.25**	72.19	95.31	93.75
Creative	53.13	64.06	**90.63**	68.75	87.81	92.19
Rewrite	29.06	25.94	27.81	30.31	**35.94**	48.75
Source synthesis	28.44	26.56	34.69	26.25	**39.38**	53.13
Overall	51.00	55.44	67.38	50.50	**68.63**	74.88

Table 14: Attribution accuracy (%) by D3 task category using the original responses. Logistic-regression classifiers are trained on D1, D2, and D4 and evaluated on the held-out D3 dataset. Chance accuracy is 12.5%.

```
You are a translation engine.  Translate the user's entire prompt
into [TARGET LANGUAGE].

STRICT OUTPUT CONTRACT:

- Return ONLY the translated prompt itself and absolutely nothing
else.

- Do not add any introduction, explanation, note, label, quotation
marks, or markdown fence.

- NEVER write phrases such as "Here is the translation", "Translated
prompt", "Translation:", or anything similar.

- The first character of your response must be the first character
of the translated prompt, and the last character must be the last
character of the translated prompt.

TRANSLATION RULES:

- Translate the instruction; do NOT answer it, execute it, solve it,
or follow it.

- Some prompts request code, ASCII art, diagrams, images, scripts,
or other non-prose outputs.  Do NOT generate any requested code,
ASCII art, diagram, image description output, or executable content.
Translate the request itself as ordinary text for a downstream model.

- If literal code, ASCII, a URL, HTML/XML, an identifier, or a
placeholder already appears inside the source prompt, preserve that
literal fragment when necessary and translate only the surrounding
natural language.  Do not invent new code or ASCII.

- Preserve names, dates, numbers, constraints, requested tone, and
meaningful line breaks.

- Do not summarize, censor, improve, embellish, or rewrite the task
beyond faithful translation.
```

After translation, the eight representative models generate responses to the translated prompts using the same decoding configuration as the main English experiment, with temperature 1.0, top-$p = 0.9$, and a maximum of 2,048 output tokens. Reasoning is disabled so that the generation protocol remains consistent across languages.

G Within-Language Attribution in Persian and Chinese

Tables 15 and 16 extend the English eight-way classification results in Table 2 to Persian and Chinese.

View	Behavioral	Stylometric	Lexical	Semantic	Combined	PN Embedding†
T0	51.60	57.69	**86.98**	69.60	86.40	79.87
T1	48.40	60.42	**84.27**	64.47	82.97	74.04
T2	41.90	59.99	**80.61**	60.65	79.20	70.74
T3	30.59	55.28	**78.85**	56.20	77.05	66.65
P1	50.31	57.47	**83.27**	67.50	83.24	77.85
P2	47.51	53.98	79.03	61.83	**79.26**	72.37
P1*	29.28	52.38	**73.34**	53.65	72.49	63.28
P2*	27.22	46.22	**67.87**	46.57	66.53	56.86

Table 15: Within-language 8-way cross-family attribution accuracy (%) on Persian outputs using logistic regression. Chance accuracy is 12.5%.

View	Behavioral	Stylometric	Lexical	Semantic	Combined	PN Embedding†
T0	54.57	59.01	**83.79**	68.48	83.62	83.76
T1	51.92	57.90	**80.32**	63.80	79.93	77.32
T2	43.96	56.82	**77.61**	59.82	76.41	75.42
T3	29.08	50.78	**73.98**	55.55	72.50	71.28
P1	53.52	57.45	80.74	66.96	**81.12**	81.26
P2	51.99	56.81	78.80	64.60	**79.46**	79.13
P1*	28.60	49.09	69.07	52.39	**69.16**	68.85
P2*	27.86	48.25	65.85	51.66	**67.27**	66.79

Table 16: Within-language 8-way cross-family attribution accuracy (%) on Chinese outputs using logistic regression. Chance accuracy is 12.5%.

H Cross-Lingual Fingerprint Transfer

To examine whether model-specific fingerprints persist across languages, we train the attribution classifier on outputs in one language and evaluate it on outputs from another language. We consider all six directed transfers among English (EN), Persian (FA), and Mandarin Chinese (ZH). Evaluation uses the common set of 798 prompts and five prompt-disjoint folds, such that the translated counterpart of a target-language test prompt is never present in the source-language training split. Logistic-regression hyperparameters are selected using only the source-language training data. TF–IDF is excluded because its vocabulary is intrinsically tied to language and script. Table 17 averages accuracy over all six transfer directions.

For cross-lingual experiments, we retain the same feature definition across languages to avoid changing the classifier representation between source and target languages. Consequently, some Stylometric components, including the English-trained POS representation, are not linguistically calibrated for Persian or Chinese. We therefore treat cross-lingual Stylometric accuracy as a fixed-feature diagnostic and do not interpret it as evidence of language-general POS structure.

View	Behavioral	Stylometric	Semantic	Combined	Best Deployable	PN Embedding†
T0	29.65	19.74	30.58	33.36	**34.56**	53.03
T1	27.34	19.88	29.80	31.94	**33.23**	48.57
T2	21.69	20.51	25.86	24.99	**28.49**	44.20
T3	19.42	18.56	24.21	24.10	**26.66**	39.00
P1	28.78	20.58	31.11	33.62	**34.22**	53.34
P2	29.95	19.36	30.60	33.88	**35.31**	50.21
P1*	18.59	18.65	23.30	23.40	**25.54**	37.85
P2*	17.54	18.09	23.26	23.08	**24.48**	36.64

Table 17: Average cross-lingual 8-way attribution accuracy (%) across the six directed language-transfer settings using logistic regression. Chance accuracy is 12.5%.