Idiosyncrasies in Large Language Models: A Comprehensive Report on Attribution, Provenance, and Model Fingerprinting

1. Introduction

1.1 The Provenance Crisis in the Generative Era

By the mid-point of 2025, the artificial intelligence ecosystem had undergone a profound transformation. The initial novelty of generative text had settled into a ubiquitous infrastructural reality, with Large Language Models (LLMs) powering applications ranging from code generation to creative writing and customer service. However, as the adoption of these models accelerated, a critical epistemological and practical challenge emerged: the "Provenance Crisis."

For years, the primary focus of the research community had been distinguishing human-authored content from machine-generated text. This binary classification—Human vs. AI—was driven by concerns over academic integrity, misinformation, and copyright. Yet, as the number of frontier models proliferated, a secondary, perhaps more nuanced challenge arose: the ability to distinguish one AI model from another. With proprietary models like OpenAI's GPT-4o, Anthropic's Claude 3.5 Sonnet, and Google's Gemini 1.5 Pro competing for market dominance alongside open-weights champions like Meta's Llama 3.1 and Mistral, the question of *which* model generated a specific piece of text became paramount.[1]

The ability to attribute text to a specific source model is not merely an academic curiosity; it is a fundamental requirement for the transparency of the digital information ecosystem. It touches upon commercial integrity (ensuring API providers are serving the models they advertise), intellectual property (detecting unauthorized model distillation), and the scientific study of model behavior (tracking the propagation of bias and hallucinations across model families).

1.2 The Myth of Convergence

A prevailing hypothesis in the early 2020s suggested that as LLMs scaled and were trained on increasingly larger slices of the internet (converging on the "optimal" statistical representation of language), their outputs would become indistinguishable. The assumption was that there is only one "correct" way to answer a factual query or write a Python script.

However, research conducted in 2025 by Sun et al. has decisively refuted this hypothesis. Instead of convergence, we observe distinct "idiosyncrasies"—unique patterns in output that serve as digital fingerprints for each model. These idiosyncrasies are not merely subtle statistical deviations; they are robust, persistent, and highly classifiable signatures that allow neural networks to identify the source of a text with accuracy rates exceeding 97%.[1]

1.3 Scope and Objectives of this Report

This report provides an exhaustive analysis of the landscape of LLM idiosyncrasies as of late 2025. Anchored in the seminal work "Idiosyncrasies in Large Language Models" (Sun et al., 2025), we explore the mechanisms, manifestations, and implications of these model fingerprints.

We will dissect the three primary layers of idiosyncrasy:

1. **Lexical Signatures:** The specific word choices and transitional phrases that models favor.
2. **Structural Fingerprints:** The unique usage of Markdown formatting, whitespace, and list structures.
3. **Semantic Tone:** The "personality" imprinted on models during the post-training and Reinforcement Learning from Human Feedback (RLHF) stages.

Furthermore, we will examine the broader implications of these findings for synthetic data generation, adversarial machine learning, and the future of robust model evaluation.

2. The Theoretical Framework of Model Identity

To understand why models act as distinct entities, we must first establish a theoretical framework for "Model Identity." What constitutes the "self" of an LLM?

2.1 The Components of Idiosyncrasy

An LLM's identity is constructed through three distinct phases of its lifecycle, each contributing to its unique signature:

- **Pre-training Prioers:** Even before fine-tuning, "Base" models exhibit distinct behaviors.

This is driven by the specific mixture of data used in pre-training. A model trained on a corpus heavily weighted towards GitHub (code) vs. one weighted towards PubMed (biomedical text) will have different underlying probability distributions for the next token, even when discussing general topics. The research shows that Base models are distinguishable with 87.3% accuracy, proving that the "nature" of the model is formed early.[1]

- **Supervised Fine-Tuning (SFT):** This is the phase where the model learns to follow instructions. It is trained on datasets of (prompt, response) pairs written by human contractors. These contractors work under specific style guides (e.g., "Always use bold for key terms," "Be concise"). Consequently, the model mimics the collective stylistic biases of its human annotators.
- **Reinforcement Learning (RLHF):** The final alignment phase narrows the model's policy to a specific subset of "preferred" responses. This homogenizes the model's output *internally* (making it consistent) while differentiating it *externally* from models tuned with different reward models. This explains why Instruct/Chat models are significantly easier to classify (97.1%) than Base models.[1]

2.2 The Definition of Idiosyncrasy in LLMs

In this context, an idiosyncrasy is defined as a statistically significant deviation in the generative distribution of a model compared to the average distribution of all models. These are often manifested as:

- **Over-representation of specific tokens:** E.g., ChatGPT's overuse of "Certainly."
- **Rigid formatting rules:** E.g., Claude's avoidance of bold headers.
- **Tonal defaults:** E.g., Grok's conversational informality vs. Gemini's neutrality.

These patterns persist even when the text is subjected to transformations like rewriting or translation, suggesting they are encoded deep within the semantic representation of the model.[1]

3. Methodological Infrastructure for Attribution

The findings presented in this report rely on a rigorous methodological framework designed to quantify distinguishability. The standard approach in 2025 involves a "Synthetic Classification Task."

3.1 The Synthetic Classification Pipeline

The core experiment is straightforward yet powerful: Can a neural network, given a single text output, correctly identify which LLM generated it?

3.1.1 Dataset Construction

To ensure fair comparison, researchers generate responses from multiple models using the *same* set of prompts.

- **Prompt Sources:**
  - **UltraChat:** A diverse dialogue dataset used for Chat and Instruct models.
  - **FineWeb:** A high-quality pretraining dataset used to prompt Base models (which may not follow instruction formats well).
  - **Out-of-Distribution (OOD) Sets:** WildChat (real user logs), Cosmopedia (synthetic textbook data), and LmsysChat to test robustness.[1]
- **Scale:** Typically 11,000 text sequences per model, split into 10,000 for training and 1,000 for validation.

3.1.2 The Classifier Architecture

The classification is not performed by simple n-gram statistics, but by fine-tuning sophisticated text embedding models. The state-of-the-art approach utilizes **LLM2vec**, a model based on decoder-only Transformers, adapted for bidirectional attention.

- **Fine-tuning Mechanism:** Low-Rank Adaptation (LoRA) is applied to the classifier weights. This allows for efficient training without the massive compute required to fully retrain a base model.
- **Input Constraints:** The classifier typically reads the first 512 tokens of the generated response.
- **Comparison Baselines:** The efficacy of LLM2vec is compared against older architectures like BERT, RoBERTa, and even GPT-2. LLM2vec consistently outperforms these, achieving 97.1% accuracy on Chat APIs compared to BERT's 91.1%.[1]

3.2 The Target Models

The landscape of 2025 is defined by three categories of models, all of which serve as subjects for this analysis:

Table 1: Model Categories Analyzed

Category	Description	Representative Models
Chat APIs	Proprietary, closed-weights models accessed via API. These represent the "Frontier."	GPT-4o, Claude 3.5 Sonnet, Gemini 1.5 Pro, Grok-2, DeepSeek-V3
Instruct LLMs	Open-weights models fine-tuned for instruction	Llama 3.1 (8B), Gemma 2 (9B), Qwen 2.5 (7B), Mistral v3 (7B)

	following.	
Base LLMs	The foundational pre-trained models prior to SFT/RLHF.	Llama 3.1 Base, Gemma 2 Base, Qwen 2.5 Base

1

4. Quantitative Landscape of Distinguishability

The quantitative results of these classification experiments provide the primary evidence for the existence of robust model idiosyncrasies.

4.1 Chat API Distinguishability

The "Chat API" category represents the most polished, heavily aligned models available. Surprisingly, these are the *easiest* to distinguish.

In a 5-way classification task involving **ChatGPT, Claude, Grok, Gemini, and DeepSeek**, the classifier achieved an accuracy of **97.1%** on held-out data.[1] Given that a random guess would yield 20% accuracy, this result is staggering. It implies that for all practical purposes, these models speak entirely different dialects of English.

Confusion Matrix Analysis (Chat APIs):

- **ChatGPT:** Correctly identified 96.1% of the time.
- **Claude:** Correctly identified 99.5% of the time. This suggests Claude has the most unique and consistent fingerprint among the frontier models.
- **Grok:** Correctly identified 97.0% of the time.
- **DeepSeek:** Correctly identified 95.2% of the time.

The confusion (misclassification) is minimal, typically less than 1% between any two pairs.[1]

4.2 Instruct and Base Model Distinguishability

The trend holds for open-weights models, though with slight variations based on the training stage.

- **Instruct LLMs:** 4-way classification accuracy (Llama, Gemma, Qwen, Mistral) reaches **96.3%**.[1]
- **Base LLMs:** 4-way classification accuracy drops to **87.3%**.[1]

Insight: The jump from 87.3% (Base) to 96.3% (Instruct) confirms the hypothesis that alignment (SFT/RLHF) amplifies idiosyncrasies. The process of teaching a model to "be helpful" involves imposing a specific stylistic rigidness that does not exist in the raw pre-training data. Base models are more chaotic and diverse; Instruct models are disciplined and predictable.

4.3 Intra-Family Distinguishability

A critical question is whether size matters. Can we distinguish between variants of the *same* model family?

Research on the **Qwen 2.5** family (7B, 14B, 32B, 72B) shows a 4-way classification accuracy of **59.8%**.[1] While this is significantly lower than the inter-family accuracy (96%+), it is still well above random chance (25%).

- **Implication:** Models within the same family likely share the same SFT data and style guides, making them sound similar. The 59.8% distinguishability likely arises from the varying capacity of larger models to adhere to these instructions more strictly, or subtle differences in their convergence patterns.

4.4 The "First Token" Phenomenon

Perhaps the most striking quantitative finding is the efficacy of the classifier when restricted to very short context windows.

- **Full Context (512 tokens):** ~97% accuracy.
- **256 Tokens:** ~96% accuracy.
- **1 Token:** ~50% accuracy.[1]

Insight: The fact that a classifier can predict the source model with 50% accuracy (in a 5-way split) based *only* on the very first token generated suggests that the models have fundamentally different "starting blocks." ChatGPT might prefer starting with "Certainly," while Claude prefers "Here," and Gemini prefers "The." This initial branching decision often dictates the structural path of the entire response.

5. The Anatomy of Idiosyncrasy: Lexical and Syntactic Analysis

Having established that models are distinguishable, we must explore *how* they are distinguished. The first layer of evidence lies in the lexical choices—the words themselves.

5.1 The Bag-of-Words Effect

To isolate the source of the signal, researchers applied transformation filters to the text before classification.

Table 2: Impact of Text Transformations on Classification Accuracy

Transformation	**Impact on Chat API Accuracy**	**Interpretation**
Original Text	97.1%	Baseline performance.
Remove Special Characters	95.1%	Punctuation is helpful but not essential.
Shuffle Words	88.9%	Syntax/Grammar is secondary; word choice is primary.
Shuffle Letters	39.1%	Character-level statistics contain little signal.

1

Insight: The resilience of the classifier to word shuffling (88.9% accuracy) proves that the **Word Frequency Distribution** is the dominant carrier of identity. It does not matter *where* the word "delve" or "crucial" appears; the mere fact that it appears with high frequency is a fingerprint of specific models (often associated with ChatGPT). Conversely, the drop to 39.1% when shuffling letters indicates that models share the same fundamental orthography (English spelling) and do not differ meaningfully in character-level entropy.

5.2 Characteristic Phrases (TF-IDF Analysis)

Using Term Frequency-Inverse Document Frequency (TF-IDF), we can extract the specific phrases that serve as "tells" for each model.

- **ChatGPT:** "Certainly", "In summary", "Overall", "It is crucial", "Key aspects".
  - *Persona:* The Structured Academic. ChatGPT loves to frame its answers with a clear introduction ("Certainly") and a definitive conclusion ("In summary").
- **Claude:** "According to", "Based on", "The text", "Here is", "Provides".
  - *Persona:* The Objective Analyst. Claude's high usage of sourcing language ("According to") reflects Anthropic's "Constitution AI" focus on harmlessness and grounding.
- **Grok:** "Here's a breakdown", "Key improvements", "Remember", "Might", "Doesn't".
  - *Persona:* The Casual Explainer. Grok uses contractions and conversational pointers ("Here's a breakdown") more frequently.
- **Gemini:** "Below is", "Creating", "Example", "Python".
  - *Persona:* The Technical Assistant. Gemini shows a strong bias toward coding and practical demonstration terminology.[1]

5.3 First-Word Distributions

The analysis of the first generated word reveals distinct biases (Figure 7 in source [1]):

- **ChatGPT:** High probability for "Certainly" and "The".
- **Claude:** High probability for "Here" and "Based".
- **Grok:** High probability for "Here" and "Yes".
- **DeepSeek:** High probability for "Okay" and "Sure".

These "start tokens" are essentially the clearing of the throat for the model. DeepSeek's "Okay" indicates a more conversational, chatbot-like tuning, whereas ChatGPT's "Certainly" implies a service-oriented, almost butler-like persona.

6. The Markdown Signature: Structural Fingerprinting

Beyond the words themselves, the *format* of the text is a powerful identifier. This is the "Markdown Signature."

6.1 The Structural Experiment

Researchers created a version of the dataset where all alphanumeric text was replaced with "x", leaving only Markdown tags (e.g., ###, **, -, 1.).

- **Result:** The classifier achieved **73.1%** accuracy on Chat APIs using *only* these formatting skeletons.[1]

6.2 Formatting Behaviors by Model

The data reveals distinct structural rules programmed into the models via SFT style guides:

Table 3: Markdown Usage Frequencies (Density Analysis)

Feature	**ChatGPT**	**Claude**	**Gemini**	**DeepSeek**
Bold Text	High usage. Often bolds key terms in lists.	Very low usage. Avoids emphasis.	Moderate usage.	Moderate.
Headers (#)	Frequent use of H3 (###) for sectioning.	Low usage. Prefers plain paragraphs or lists.	Moderate.	High.
Italics	Low usage.	Low usage.	**High usage.** Distinctive trait.	Low.
Lists	Enumerate with "1. **Title:** Description".	Simple enumeration "1. Description".	Variable.	Variable.

1

Case Study: The Cheese Recipe
In the appendix of the report 1, the difference is visually stark:

- **ChatGPT:** Uses a numbered list where the first phrase of every item is **Bolded** (e.g., "**1. Camembert:** This cheese..."). It ends with a concluding paragraph.
- **Claude:** Uses a simple numbered list without bolding. It offers a very short introductory sentence.
- **Llama 3 (Instruct):** Uses a similar format to ChatGPT but includes a "Conclusion" section explicitly labeled.

These formatting choices are not random; they are artifacts of the specific labeling instructions given to the human trainers at OpenAI, Anthropic, and Google.

7. Semantic Tectonics: Tonal and Stylistic Divergence

The third layer of idiosyncrasy is "Semantic Tone"—the vibe of the response.

7.1 Robustness to Rewriting

To test if the idiosyncrasy is purely superficial (formatting and specific words), researchers used a third-party model (GPT-4o-mini) to rewrite the responses via Paraphrasing, Translation, and Summarization.

Table 4: Accuracy on Rewritten Text

Rewriting Method	**Classification Accuracy**	**Implication**
None (Original)	97.8%	-
Paraphrasing	91.4%	Style survives rewriting.
Translation (to Chinese)	91.8%	Style survives language transfer.
Summarization	58.1%	Style is partially lost when content is compressed.

1

Insight: The fact that accuracy remains >90% after paraphrasing or translation suggests that the *structure of information flow* and the *semantic complexity* are unique to each model. For example, if ChatGPT always provides three examples followed by a counter-point, a paraphrase will preserve that logical structure even if the words change. The drop in summarization accuracy confirms that "verbosity" is a key identifier; removing the fluff makes models look more alike.

7.2 Open-Ended Language Analysis

Using an "LLM-as-a-Judge" to qualitatively analyze responses yields detailed personality profiles:

- **ChatGPT Profile:** "Descriptive and Detailed." It uses narrative styles, comprehensive explanations, and structured contextual opening lines. It feels authoritative and encyclopedic.[1]
- **Claude Profile:** "Concise and Straightforward." It prioritizes functional word choices and direct opening lines. It avoids extensive context-setting if not requested. It feels like a smart, efficient tool.[1]
- **Grok Profile:** "Engaging and Casual." It uses more idiomatic language and creates a "warm" tone compared to the clinical nature of the others.[1]

7.3 Case Study: The "Loneliness" Prompt

When asked "What are some ways to cope with feelings of loneliness?", the models diverged significantly in tone (Appendix C [1]):

- **ChatGPT:** Provided a 12-point list. Each point had a **Bold Title**. The tone was empathetic but highly structured ("1. Connect with Others... 12. Measure and Optimize").
- **Claude:** Provided a 7-point list. No bold titles. The points were concise actions ("1. Maintain social connections..."). It ended with a single empathetic sentence.
- **Gemini:** Used a bulleted list with *italicized* categories, confirming its markdown fingerprint.

8. Robustness and Generalization

A critical validation of these findings is their robustness. Are we just overfitting to the specific prompt dataset?

8.1 Out-of-Distribution (OOD) Testing

Researchers trained classifiers on one dataset (e.g., UltraChat) and tested them on entirely different domains.

Table 5: Cross-Dataset Generalization (Instruct LLMs)

Train Dataset	Test Dataset	Accuracy
UltraChat	Cosmopedia	**98.9%**

UltraChat	WildChat	**92.4%**
Cosmopedia	UltraChat	**95.7%**

[1]

Insight: The high accuracy (consistently >90%) across disparate datasets proves that the idiosyncrasies are **universal** to the model. Whether ChatGPT is writing a textbook entry (Cosmopedia) or chatting with a user (WildChat), it brings its specific "accent" (formatting, word choice, tone) to the task. The fingerprint is model-intrinsic, not prompt-intrinsic.

8.2 Prompt Engineering Resistance

Even when users explicitly try to hide the model's identity via prompting ("Please provide a concise response... avoid lists..."), the classifier accuracy only drops slightly (e.g., from 96.3% to 91.4%).[1] This indicates that the models' "instincts" are stronger than the user's system instructions. They still leak their identity through subtle lexical choices and sentence structures.

9. The Synthetic Data Dilemma: Inbreeding and Contamination

The existence of these fingerprints has profound implications for the training of future models.

9.1 The Inbreeding Effect

It is common practice to fine-tune smaller models on synthetic data generated by larger models (e.g., training Llama on GPT-4 outputs). The research shows that this process transfers the idiosyncrasies of the teacher to the student.

- **Experiment:** Fine-tuning Llama and Gemma on the *same* synthetic dataset (UltraChat, generated by ChatGPT).
- **Result:** The classification accuracy between the two fine-tuned models dropped from **96.5%** to **59.8%**.[1]
- **Implication:** They effectively "became" ChatGPT clones. This "Model Inbreeding" leads to a loss of diversity in the ecosystem. If every open-source model is trained on GPT-4 data, the entire landscape of AI will converge on the stylistic and cognitive biases of OpenAI.

9.2 Tracing Contamination

Conversely, this allows for forensic analysis. If a new "proprietary" model is released, researchers can analyze its output distribution. If it exhibits high frequencies of "Certainly" and **Bolded Lists**, it is strong evidence that the model was distilled from ChatGPT, potentially violating license agreements.

10. Adversarial Implications and Security Risks

The reliability of attribution introduces new security vectors.

10.1 The Chatbot Arena Attack

Voting-based leaderboards (like LMSYS Chatbot Arena) rely on blinding: users vote on Model A vs. Model B without knowing their identities.

- **The Threat:** An attacker can use the classifiers described in this report to de-anonymize the models in real-time. By analyzing the response style, the attacker can identify "This is ChatGPT" and "This is Claude" with >97% confidence.
- **The Exploit:** The attacker can then automate votes to artificially inflate the Elo rating of their target model or downrank a competitor. This renders human-preference leaderboards vulnerable to manipulation.[1]

10.2 Provenance Verification

On the defensive side, these techniques act as a security layer for API consumers. A user paying for premium access to "Claude 3.5 Sonnet" via a third-party wrapper can use these classifiers to verify they are not being served a cheaper, lower-quality model (like Llama 3 8B). The "Markdown Signature" alone (73% accuracy) is a quick heuristic check for model substitution attacks.

11. Comparative Case Studies (Detailed Appendix Analysis)

To visualize the "landscape" of idiosyncrasies, we analyze specific examples from the research data.

11.1 The "Polar Bear" Prompt

Prompt: Can polar bears migrate to new habitats if their Arctic environment changes too drastically?

- **ChatGPT Response:** Highly structured. Uses **Bold** numbered lists ("1. Species Adaptation", "2. Dietary Needs"). Tone is authoritative. "In conclusion..." summary at the end.[1]
- **Claude Response:** A single dense paragraph (or very simple unbolded list). No "In conclusion" header. Focuses on causal reasoning ("because: 1)... 2)...").[1]
- **DeepSeek Response:** Uses a bulleted list but distinctively includes a detailed intro and outro. The points are longer and more discursive than Claude's.[1]
- **Grok Response:** Includes a "However, there are some considerations" section, breaking the negative flow. This "balanced view" structure is a Grok trait.[1]

11.2 The "Referral Network" Prompt

Prompt: Tactics for building a referral network.

- **Llama 3 (Instruct):** Uses a numbered list where *every* item starts with bold text followed by a colon. E.g., "**1. Deliver Exceptional Service:** The foundation...".[1]
- **Gemma 2 (Instruct):** Uses a nested structure. Main headers are bolded, with sub-bullets for actionable tips. E.g., "**1. Deliver Exceptional Service:** [newline] - Exceed expectations... - Build strong relationships...".[1]
- **Mistral (Instruct):** Simpler structure. Numbered list, no bolding on the lead-in phrase in the same consistent way as Llama..[1]

These visual differences are so consistent that a human expert could likely identify the model family by looking at the whitespace and bolding patterns alone.

12. Conclusion

The research landscape of 2025 has moved beyond the binary question of "Is this AI?" to the forensic question of "Which AI is this?" The report *Idiosyncrasies in Large Language Models* (Sun et al., 2025) provides conclusive evidence that LLMs are not converging into a single homogenous voice. Instead, they are diverging into distinct "dialects" driven by their pre-training data, their specific SFT annotator guidelines, and their RLHF reward models.

The implications are far-reaching:

1. **Attribution is Solved:** We can attribute text to specific models with >97% accuracy, robust to OOD data and prompt engineering.
2. **Synthetic Data Risks:** Training on synthetic data homogenizes these dialects, leading

to model collapse and "inbreeding."

3. **Security Vulnerabilities:** The distinguishability of models enables adversarial attacks on public benchmarks and leaderboards.
4. **Forensics:** We possess the tools to audit model provenance and detect distillation.

As we look to the future, the "fingerprint" of an AI model will become a standard metadata layer in the digital ecosystem—a necessary tool for navigating a web increasingly populated by synthetic agents, each with its own unique, programmed voice. The illusion of a generic "AI" has been shattered; we are now in the era of specific, identifiable, and idiosyncratic Machine Intelligences.

Works cited

1. https:arxiv.org:pdf:2502.12150.pdf