Intrinsic Fingerprint of LLMs: Continue Training is NOT All You Need to Steal A Model!

Do-hyeon Yoon, Minsoo Chun, Thomas Allen, Hans Müller, Min Wang, and Rajesh Sharma
Honest AGI Community
honestagi@outlook.com

Abstract

Large language models (LLMs) face significant copyright and intellectual property challenges as the cost of training increases and model reuse becomes prevalent. While watermarking techniques have been proposed to protect model ownership, they may not be robust to continue training and development, posing serious threats to model attribution and copyright protection. This work introduces a simple yet effective approach for robust LLM fingerprinting based on intrinsic model characteristics. We discover that the standard deviation distributions of attention parameter matrices across different layers exhibit distinctive patterns that remain stable even after extensive continued training. These parameter distribution signatures serve as robust fingerprints that can reliably identify model lineage and detect potential copyright infringement. Our experimental validation across multiple model families demonstrates the effectiveness of our method for model authentication. Notably, our investigation uncovers evidence that a recently Pangu Pro MoE model released by Huawei is derived from Qwen-2.5 14B model through upcycling techniques rather than training from scratch, highlighting potential cases of model plagiarism, copyright violation, and information fabrication. These findings underscore the critical importance of developing robust fingerprinting methods for protecting intellectual property in large-scale model development and emphasize that deliberate continued training alone is insufficient to completely obscure model origins.

1 Introduction

The rapid advancement of large language models (LLMs) has fundamentally transformed the artificial intelligence landscape, with training costs reaching millions of dollars and requiring substantial computational resources (Hoffmann et al., 2022; Touvron et al., 2023a,b; Liu et al., 2024). As these models become increasingly valuable assets, protecting intellectual property rights and preventing unauthorized model reuse has emerged as a critical challenge for both commercial entities and research institutions (Samuelson, 2023; Zeng et al., 2024). The substantial investment required for training state-of-the-art LLMs creates strong incentives for model theft, unauthorized copying, and derivative works that may infringe upon original creators' rights (Yao et al., 2024).

Traditional approaches to model protection have primarily relied on watermarking techniques embedded during training or inference phases (Kirchenbauer et al., 2023; Kuditipudi et al., 2023). However, these methods face significant vulnerabilities when confronted with continued training, fine-tuning, or model modification. Adversaries can potentially remove or obscure watermarks through additional training iterations, making conventional protection mechanisms insufficient for robust copyright enforcement. This limitation has created an urgent need for more resilient fingerprinting methods that can survive various forms of model manipulation and adaptation.

In this work, we propose a simple yet effective approach to LLM fingerprinting that leverages intrinsic parameter characteristics rather than externally imposed watermarks. Our key insight is that the simple statistical properties of attention parameter matrices, i.e., their standard deviation distributions across layers, form distinctive signatures that are remarkably stable across the training process. Unlike traditional watermarks that can be intentionally targeted for removal, these intrinsic fingerprints emerge naturally from the model's architecture and training dynamics, making them significantly more difficult to erase through continued training or model modification, even from dense model to MoE. All we need is loading model checkpoints and simply apply torch.std() to different parameter matrices, then we can establish model lineage and detect potential cases of unauthorized model derivation with high confidence. To demonstrate the practical importance of this research, we present a compelling case study involving the recently released Pangu Pro MoE model by Huawei[1] and the Qwen-2.5 14B model[2]. Our analysis provides evidence suggesting that the Pangu Pro MoE model may have been derived from Qwen-2.5 14B through upcycling techniques, highlighting real-world instances of potential copyright infringement in the current LLM ecosystem and serious information fabrication in the corresponding technique report (Tang et al., 2025). This finding underscores the immediate relevance of robust fingerprinting methods for protecting intellectual property in large-scale model development. The contributions of this work extend beyond technical methodology to address fundamental questions about model ownership, attribution, and intellectual property protection in the age of foundation models. As the LLM landscape continues to evolve rapidly, establishing reliable methods for model authentication becomes increasingly critical for maintaining innovation incentives and protecting the substantial investments required for model development.

2 Related Work

The rapid evolution of Large Language Models (LLMs) has brought the challenge of protecting intellectual property (IP) to the forefront (He et al., 2022). As models, both proprietary and open-source, represent significant investment in terms of data, computation, and human expertise, there is a critical need for reliable methods to verify model ownership and trace unauthorized distribution. The primary approaches in the literature for this purpose are LLM watermarking and fingerprinting. However, a critical examination reveals their inherent fragility, particularly when confronted with common model adaptation techniques like fine-tuning and parameter modification.

2.1 Watermarking for Provenance Tracking

Watermarking schemes aim to proactively embed a hidden, detectable signal into the output generated by an LLM (Zhao et al., 2023). The goal is to allow a model owner to later verify if a piece of text was generated by their model. A popular category of methods involves modifying the model's output probability distribution during decoding. For instance, a secret key can be used to pseudo-randomly partition the vocabulary into a "greenlist" and a "redlist," where the generation of greenlisted tokens is subtly encouraged (Kirchenbauer et al., 2023). By analyzing a suspect text for a statistically significant presence of greenlisted tokens, one can trace its origin.

Despite their ingenuity, these generative watermarking schemes are fundamentally brittle. Their reliance on a specific output distribution makes them highly susceptible to any process that alters this distribution. Fine-tuning is a primary adversary in this context. Even brief continued training on a small, domain-specific dataset can significantly shift the model's weights and, consequently, its next-token probabilities, effectively erasing or overwriting the embedded watermark (Kirchenbauer et al., 2023). The signal, being statistical in nature, is simply too fragile to survive the gradient-based updates inherent to the fine-tuning process.

Furthermore, these methods are vulnerable to simple post-processing or paraphrasing attacks. An adversary can use another off-the-shelf LLM to rephrase the watermarked output. This process preserves the semantic content but completely disrupts the token-level statistical patterns that form the watermark, rendering it undetectable (Sadasivan et al., 2023). Other attacks, such as spoofing, have also demonstrated the ability to frame a non-malicious provider by mimicking their watermark, casting further doubt on the reliability of this approach for robust IP claims (Zhao et al., 2023).

2.2 Fingerprinting for Model Identification

In contrast to watermarking, which marks the output, fingerprinting aims to identify the model itself (Lyu et al., 2022; Sun et al., 2024). These techniques seek to discover or implant unique characteristics within the model that serve as an identifier. Fingerprinting can be passive, by identifying unique stylistic quirks or artifacts of a model, or active, by intentionally embedding a "backdoor" signal. Active methods often involve training the model on a carefully crafted set of "trigger" inputs that elicit a specific, secret output (e.g., a specific phrase or identifier). The presence of this behavior in a suspect model serves as the fingerprint (Lyu et al., 2022). Researchers also proposed methods based on fingerprinting features extracted from the parameters of LLMs (Zeng et al., 2024). However, like watermarking, existing fingerprinting schemes struggle to withstand long

[1] https://gitcode.com/ascend-tribe/pangu-pro-moe-model

[2] https://huggingface.co/Qwen/Qwen2.5-14B

continued (pre)training and deliberate model modification (Han et al., 2024), and many methods even become invalid when model architectures are changed by the attacker. In summary, while existing watermarking and fingerprinting techniques provide a foundational framework for LLM IP protection, their core assumptions are often violated in real-world scenarios. Their sensitivity to modifications in model weights and output distributions renders them largely ineffective against adversaries who can perform continue training, model distillation, upcycling or even simple paraphrasing attacks. This highlights a critical gap in the literature: the need for a verification method that is intrinsically tied to the core, functional identity of the model and is demonstrably resilient to such modifications. Our work aims to address this challenge.

3 Methodology

3.1 Problem Definition

We formalize the problem of LLM lineage detection as follows: Given two large language models A and B with their respective parameter weights θ_A and θ_B, our objective is to determine whether model A has been derived from model B through continued training, fine-tuning, or other modification techniques, rather than being trained independently from scratch. This binary classification problem is critical for intellectual property protection and model attribution in scenarios where organizations may claim independent development while actually building upon existing foundation models.

3.2 Parameter Distribution Fingerprinting

Our approach leverages the intrinsic statistical properties of attention mechanism parameters to create robust fingerprints for model lineage detection. For each transformer layer l in a given model, we extract the query (Q_l), key (K_l), value (V_l), and output (O_l) projection matrices from the multi-head attention mechanism.[3]

For each attention matrix $M \in Q_l, K_l, V_l, O_l$ at layer l, we compute the standard deviation of all parameters:

$$\sigma_l^M = \text{std}(M_l) \tag{1}$$

where std$(\cdot)$ denotes the standard deviation operation across all elements of the matrix. This process yields four sequences of standard deviation values across all L layers of the model:

$$\mathbf{S}^Q = [\sigma_1^Q, \sigma_2^Q, \ldots, \sigma_L^Q] \tag{2}$$

$$\mathbf{S}^K = [\sigma_1^K, \sigma_2^K, \ldots, \sigma_L^K] \tag{3}$$

$$\mathbf{S}^V = [\sigma_1^V, \sigma_2^V, \ldots, \sigma_L^V] \tag{4}$$

$$\mathbf{S}^O = [\sigma_1^O, \sigma_2^O, \ldots, \sigma_L^O] \tag{5}$$

To enable meaningful comparisons between models of different scales and architectures (also aiming to be robust to parameter rescale), we normalize each sequence to have zero mean and unit variance. For each sequence $\mathbf{S}^M$, the normalized sequence $\hat{\mathbf{S}}^M$ is computed as:

$$\hat{\mathbf{S}}^M = \frac{\mathbf{S}^M - \mu(\mathbf{S}^M)}{\sigma(\mathbf{S}^M)} \tag{6}$$

where $\mu(\mathbf{S}^M)$ and $\sigma(\mathbf{S}^M)$ represent the mean and standard deviation of the sequence $\mathbf{S}^M$, respectively. This normalization procedure removes scale-dependent variations while preserving the relative patterns of parameter standard deviations across layers, which we hypothesize to be characteristic signatures of model lineage. In addition, our method is also robust to linear projection and dimension permutation.

3.3 Lineage Detection

Given two models A and B, we compute their respective normalized fingerprint sequences $\hat{\mathbf{S}}_A^Q, \hat{\mathbf{S}}_A^K, \hat{\mathbf{S}}_A^V, \hat{\mathbf{S}}_A^O$ and $\hat{\mathbf{S}}_B^Q, \hat{\mathbf{S}}_B^K, \hat{\mathbf{S}}_B^V, \hat{\mathbf{S}}_B^O$. The similarity between models is then assessed by computing the correlation coefficients between corresponding sequences:

$$\rho^M = \text{corr}(\hat{\mathbf{S}}_A^M, \hat{\mathbf{S}}_B^M). \tag{7}$$

Models exhibiting high correlation across multiple attention matrix types are considered to share potential lineage relationships, indicating that one model may have been derived from the other through continued training or modification processes. For models with identical layer counts, the correlation coefficients can be computed directly between corresponding normalized sequences. However, when comparing models with different numbers of layers, we employ interpolation techniques to enable meaningful comparison. Given two models A and B with L_A and L_B layers respectively, where $L_A \neq L_B$,

[3] We prefer to use attention parameters because vector parameters such as layer norm and bias are easy to change and manipulate. FFN parameters may also be suboptimal choices because they can be directly modified when doing upcycling.

we use linear interpolation to align the sequences to a common length. Specifically, for each attention matrix type M, we interpolate the shorter sequence to match the length of the longer sequence. Let $\hat{\mathbf{S}}^M_{\text{short}}$ and $\hat{\mathbf{S}}^M_{\text{long}}$ denote the normalized sequences from the model with fewer and more layers, respectively. We define index mappings:

$$\mathbf{i}_{\text{short}} = [0, 1, 2, \ldots, L_{\text{short}} - 1], \tag{8}$$

$$\mathbf{i}_{\text{target}} = \text{linspace}(0, L_{\text{short}} - 1, L_{\text{long}}). \tag{9}$$

The interpolated sequence $\hat{\mathbf{S}}^M_{\text{interp}}$ is computed using linear interpolation:

$$\hat{\mathbf{S}}^M_{\text{interp}} = \text{interp1d}(\mathbf{i}_{\text{short}}, \hat{\mathbf{S}}^M_{\text{short}}, \mathbf{i}_{\text{target}}), \tag{10}$$

where $\text{interp1d}(\cdot)$ represents the linear interpolation function. The correlation coefficient is then computed between the interpolated sequence and the longer model's sequence:

$$\rho^M = \text{corr}(\hat{\mathbf{S}}^M_{\text{interp}}, \hat{\mathbf{S}}^M_{\text{long}}). \tag{11}$$

This interpolation approach assumes that the layer-wise evolution of parameter standard deviations follows a smooth pattern that can be meaningfully approximated through linear interpolation, enabling robust comparison across models with different depths while preserving the essential distributional characteristics that serve as lineage fingerprints.

4 Experiments

4.1 Experimental Setup

Our experiments were conducted on a diverse collection of large language models spanning multiple architectural families and organizations. The model selection includes popular open-source models from different lineages to ensure comprehensive coverage of the current LLM landscape. Specifically, we analyzed models from the Qwen series (Qwen2.5-7B, Qwen2.5-14B, Qwen2.5-32B, Qwen2.5-72B (Team, 2024), Qwen3-4B, Qwen3-8B, Qwen3-14B, Qwen3-30A3B (Yang et al., 2025)), Llama family (Llama-3.1-8B, Llama-3.1-70B (Grattafiori et al., 2024), and Llama-3.1-Nemotron-70B-Instruct (Bercovich et al., 2025)), additional MoE models (OLMoE-7BA1B (Muennighoff et al., 2024), Qwen1.5-MoE-A2.7B (Team, 2024)), and the Pangu Pro MoE (Tang et al., 2025) model released by Huawei.

For each model, we extracted the complete parameter tensors for the query (Q), key (K), value (V), and output (O) projection matrices from all transformer layers. The parameter extraction was performed using the models' official checkpoints and configurations to ensure accuracy. We computed the standard deviation of each attention matrix at every layer, followed by normalization to enable cross-model comparison as described in our methodology section. The correlation analysis was conducted using Pearson correlation coefficients, with statistical significance assessed through p-value calculations. For models with different layer counts, we employed linear interpolation to align sequence lengths before correlation computation. All experiments were conducted with full fp32 precision to maintain numerical accuracy.

4.2 Cross-Family Model Analysis

Figure 1 presents the normalized standard deviation patterns of attention matrices across representative models from different families. The visualization reveals several critical insights about the distributional signatures of transformer models. Each model family demonstrates distinct characteristic patterns that appear to encode fundamental properties of their training procedures and architectural designs.

The Llama-3.1-8B model exhibits a distinctive pattern with relatively stable Q and K projections across layers, while showing more variation in V and O projections, particularly in the deeper layers. This pattern is consistent with the Llama architecture's specific attention mechanism design and training dynamics. The Qwen3-14B model displays a different signature, with more pronounced variations in the Q projection and smoother transitions in the K and V projections. OLMoE-7BA1B, being a mixture-of-experts model, shows yet another unique pattern characterized by more dramatic fluctuations across all projection types, likely reflecting the routing dynamics and expert specialization inherent in MoE architectures. These variations suggest that the distributional signatures capture not only the model's training history but also fundamental architectural characteristics.

Most strikingly, Pangu and Qwen2.5-14B exhibit nearly identical patterns across all four attention matrix types. The curves overlay so precisely that they are often indistinguishable in the visualization. **This similarity is extraordinary given that these models supposedly represent independent de-**

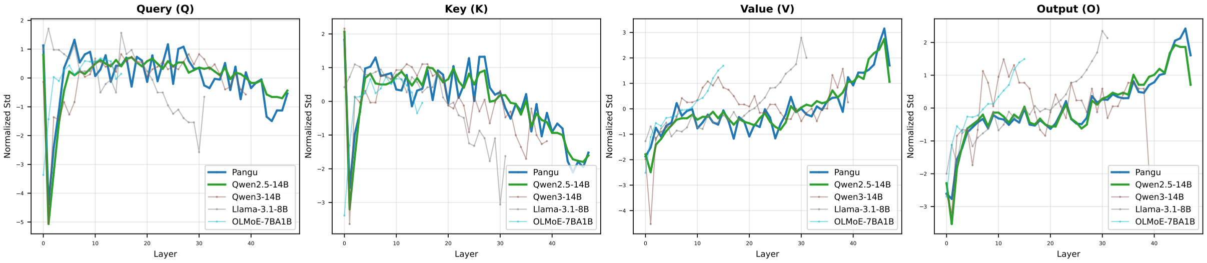

Figure 1: Normalized standard deviation patterns of attention matrices (Q, K, V, O) across different model families. The figure shows distinctive distributional signatures for each model family, with Pangu and Qwen2.5-14B exhibiting remarkably similar patterns despite their different origins and architectures.

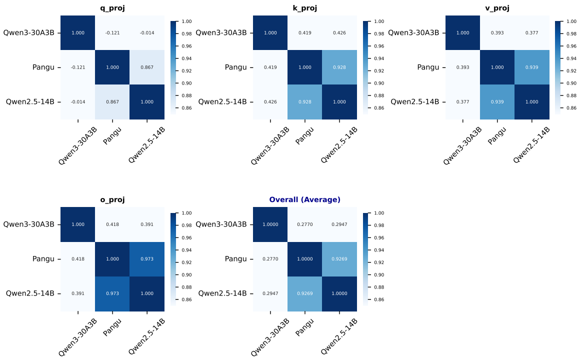

Figure 2: Correlation matrices for three key models (Qwen3-30A3B, Pangu, Qwen2.5-14B) across different attention matrix types. The exceptional correlation between Pangu and Qwen2.5-14B (overall average: 0.927) stands out significantly compared to other model pairs.

velopment efforts from different organizations with distinct training infrastructures, datasets, and optimization strategies.

4.3 Quantitative Correlation Analysis

To provide rigorous quantitative assessment of the observed similarities, we computed comprehensive correlation matrices across multiple models. Figure 2 focuses on three key models: Qwen3-30A3B, Pangu, and Qwen2.5-14B, presenting correlation coefficients for each attention matrix type separately. The results reveal remarkable patterns that strongly support our lineage hypothesis. Pangu and Qwen2.5-14B achieve correlation coefficients of 0.867 for Q projection, 0.928 for K projection, 0.939 for V projection, and 0.973 for O projection. The progression from Q to O projections shows increasing correlation, which is particularly significant because output projections typically undergo more substantial modifications during continued training and fine-tuning processes.

Figure 3 extends this analysis to a comprehensive twelve-model comparison matrix, providing broader context for interpreting the Pangu-Qwen2.5-14B correlation. The heatmaps reveal that most inter-model correlations fall within the 0.3-0.7 range, with even closely related models rarely exceeding 0.8. Within the Qwen family itself, we observe correlations ranging from 0.4 to 0.9, reflecting the natural variation expected even among models sharing training infrastructure and methodologies.

4.4 Validation Through Known Model Lineages

To establish the reliability and sensitivity of our fingerprinting methodology, we examined several documented cases of model derivation where the lineage relationships are publicly known and verified. Figure 4 analyzes Llama-3.1-Nemotron-70B-Instruct, which NVIDIA developed through supervised fine-tuning of Meta's Llama-3.1-70B base model. Despite extensive instruction tuning and safety alignment, the attention parameter distribution curves remain virtually identical between the base and derived models. Figure 5 examines models derived from Qwen2.5-7B through various community fine-tuning efforts. OpenR1-Qwen-7B and OpenThinker3-7B represent different fine-tuning approaches focusing on reasoning capabilities and chain-of-thought optimization respectively. The results show remarkable consistency in attention parameter distributions despite these divergent fine-tuning objectives.

Figure 6 provides perhaps the most relevant validation case for understanding the Pangu-Qwen relationship. Qwen1.5-MoE-A2.7B was created through a documented upcycling process that converted the dense Qwen-1.8B model into a mixture-of-experts architecture. Despite substantial architectural modifications, the attention parameter dis-

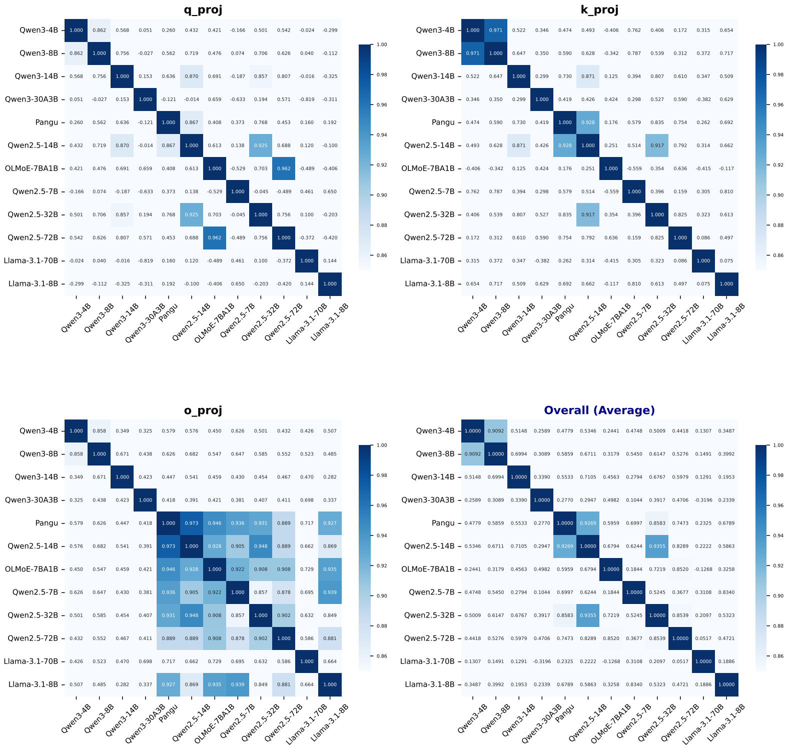

Figure 3: Comprehensive correlation analysis across twelve models from various families. The heatmaps show that most inter-model correlations fall within 0.3-0.7 range, with the Pangu-Qwen2.5-14B correlation representing a clear outlier.

tributions maintain strong similarities, directly paralleling the Pangu situation.

4.5 Feed-Forward Network Analysis

To strengthen our analysis beyond attention mechanisms, we extended the investigation to feed-forward network (FFN) parameters. Figure 7 presents the comparative analysis between Qwen2.5-14B and Qwen2-57B-A14B across the three FFN projection types. Despite both models belonging to the Qwen family, they exhibit substantial differences in their FFN parameter distributions, establishing the expected level of variation between independently developed models. Figure 8 presents a stark contrast to the intra-family variation. Pangu and Qwen2.5-14B display remarkably similar FFN parameter distributions across all three projection types, despite their fundamental architectural differences. The probability of such identical patterns arising independently between a MoE and dense model is extremely low, providing compelling evidence for direct lineage through upcycling. We notice that Pangu Pro MoE use a different tokenizer with Qwen models. We guess that they may randomly initialized the embeddings of mismatched tokens.

The technical documentation accompanying Pangu Pro MoE claims training on 13 trillion tokens, representing one of the most extensive (continued) training efforts documented in the literature. Our findings demonstrate that even such extensive continued pre-training fails to erase the intrinsic parameter distribution fingerprints inherited from the base model. This robustness validates our methodology's effectiveness for intellectual property protection and makes it particularly valuable in scenarios where traditional watermarking approaches might be vulnerable to continued training attacks.

5 Discussion

5.1 Limitations

While our fingerprinting methodology demonstrates strong performance across the examined models, several limitations must be acknowledged that may affect its applicability and reliability in certain scenarios. Our approach exhibits scale-dependent effectiveness, with detection capabilities generally improving for larger models. The statis-

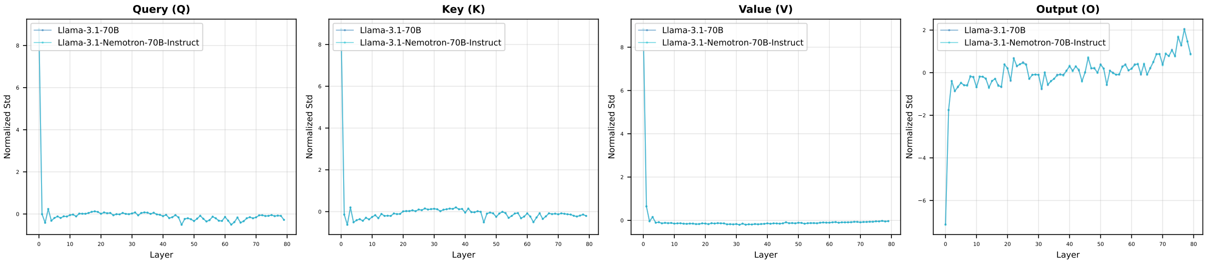

Figure 4: Attention parameter distribution comparison between Llama-3.1-70B and its fine-tuned derivative Llama-3.1-Nemotron-70B-Instruct. The nearly identical curves validate our methodology's ability to detect known lineage relationships.

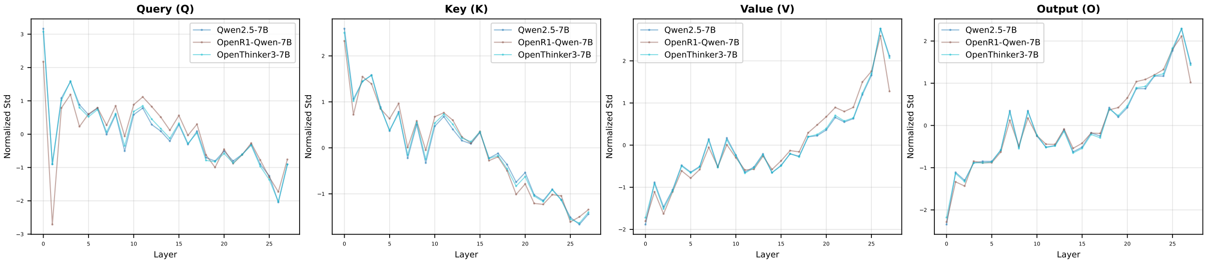

Figure 5: Validation analysis of models derived from Qwen2.5-7B through different fine-tuning approaches (OpenR1-Qwen-7B and OpenThinker3-7B). Despite divergent fine-tuning objectives, the models maintain consistent distributional signatures with their base model.

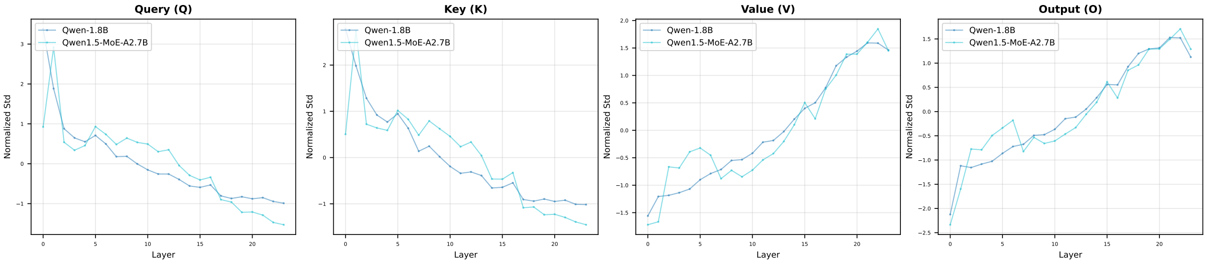

Figure 6: Analysis of Qwen1.5-MoE-A2.7B, which was created by upcycling Qwen-1.8B into a mixture-of-experts architecture. The preserved distributional patterns despite architectural transformation validate our methodology for detecting upcycling relationships.

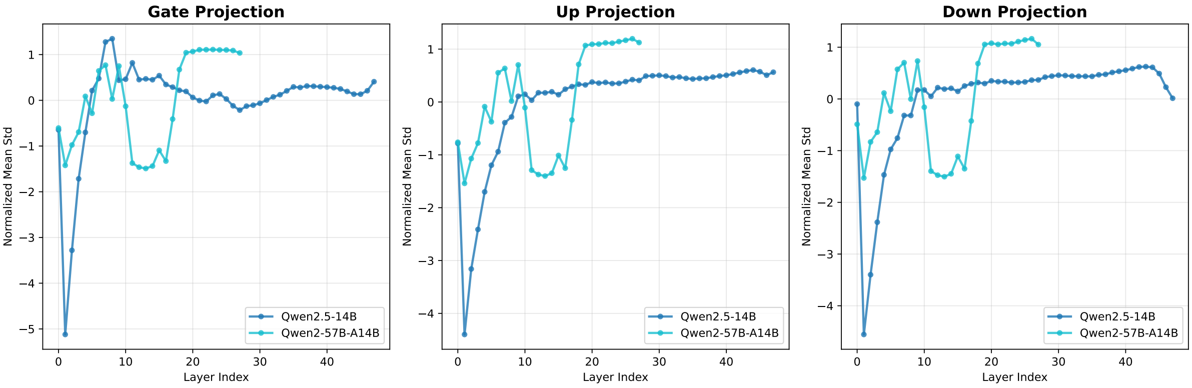

Figure 7: Feed-forward network parameter distribution comparison between two Qwen family models (Qwen2.5-14B and Qwen2-57B-A14B). Despite sharing organizational infrastructure, the models exhibit substantial differences across gate, up, and down projections.

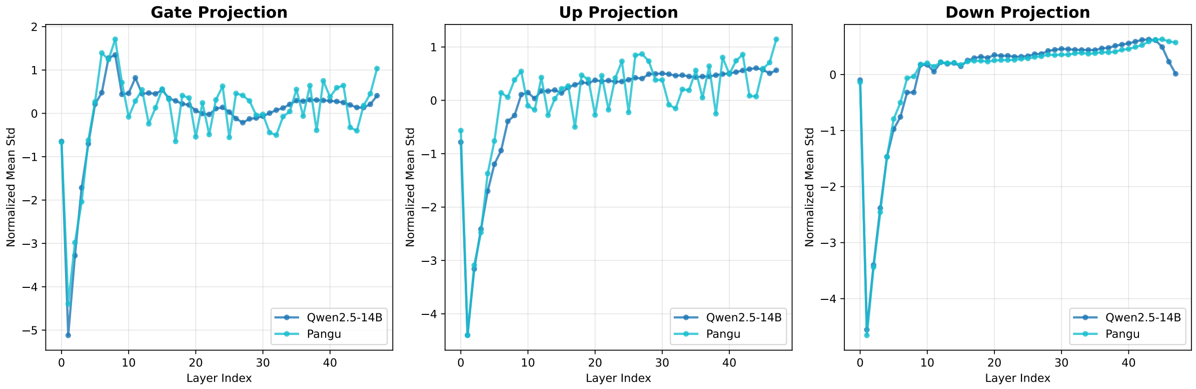

Figure 8: Feed-forward network parameter distribution comparison between Pangu and Qwen2.5-14B. The remarkable similarity across all projection types is extraordinary given their different architectures (MoE vs. dense) and supposed independent development.

tical signatures we rely upon become more robust and distinctive as model size increases, primarily due to the larger number of parameters providing richer distributional information. For models with billions of parameters, such as those examined in our study, the attention and feed-forward parameter distributions contain sufficient statistical power to enable reliable detection. However, for smaller models with fewer than one billion parameters, the effectiveness of our method is not guaranteed.

The fundamental issue stems from the limited statistical sample size in smaller models. With fewer parameters per attention matrix, the computed standard deviations may be more susceptible to random variations and optimization noise, potentially leading to less stable fingerprints. Additionally, smaller models may exhibit more volatile parameter distributions during training, making it difficult to distinguish between genuine lineage relationships and coincidental similarities arising from convergent optimization dynamics. This limitation is particularly relevant for edge deployment models, mobile-optimized variants, and early-stage experimental models that often operate in the millions rather than billions of parameter range. For such models, our methodology should be applied with caution, and additional validation through complementary techniques may be necessary to establish reliable lineage detection.

5.2 Broader Impacts

The intense competition in AI has created significant pressure for companies to rapidly demonstrate competitive model performance as evidence of their technical capabilities, often leading to compressed development timelines that may conflict with proper intellectual property practices. In this high-stakes environment, organizations facing geopolitical constraints or resource limitations may be incentivized to leverage existing foundation models rather than investing in lengthy from-scratch development, potentially bypassing licensing requirements or attribution standards in pursuit of market positioning advantages.

In this context, the potential derivation of Huawei's Pangu Pro MoE model from Qwen-2.5 14B must be understood within the broader context of Huawei's strategic positioning in the global AI hardware and software ecosystem. As a company facing significant restrictions on access to advanced semiconductor technologies and international AI supply chains, Huawei has been compelled to pursue alternative pathways to maintain competitiveness in the AI market, particularly in demonstrating the capabilities of its proprietary hardware platforms.

The evidence of Huawei's potential appropriation of Qwen's technology extends beyond mere model similarities to encompass serious concerns about the integrity of technical documentation. The substantial overlap between Huawei's and Qwen's model parameters suggests not only unauthorized use of the underlying model but also **potential fabrication in technical reporting**. This pattern of behavior indicates a systematic approach to misrepresenting the originality and development process of the Pangu Pro MoE model. Such practices undermine the fundamental principles of scientific integrity and technological transparency that are essential for maintaining trust within the AI research community. The deliberate obfuscation of model origins through misleading technical reports represents a particularly concerning form of intellectual property violation that goes beyond simple unauthorized use to encompass academic misconduct.

In fact, core contributors of this model are probably aware of their misconduct, but this still ridiculously happens in an open-source model. From a business strategy perspective, the pressure to showcase competitive AI capabilities on Huawei's Ascend chips and other proprietary hardware may have created incentives for accelerated model development timelines. Training large language models from scratch requires not only substantial computational resources but also significant time investments that may conflict with market pressures to demonstrate hardware capabilities quickly. Thus, leveraging existing high-quality foundation models as starting points for further development could be viewed as a pragmatic approach to rapidly achieving competitive performance benchmarks. However, this strategic imperative raises important questions about the balance between competitive pressures and intellectual property respect. While the desire to demonstrate hardware capabilities is understandable from a business perspective, the means by which these demonstrations are achieved must align with established norms of intellectual property protection and academic integrity. The AI industry's rapid pace of development should not come at the expense of fundamental principles of attribution and fair use.

Fortunately, the authentication methodology presented in this work has significant implications for the broader AI industry and research community. By providing a systematic framework for detecting potential model derivation through simple analysis, our approach offers practical tools for protecting intellectual property rights in an increasingly complex AI landscape. The methodology's ability to identify subtle signs of unauthorized model adaptation could serve as a deterrent against future instances of model appropriation while providing legal and technical evidence when violations occur. Furthermore, the development of such authentication techniques promotes greater transparency and accountability in AI model development, encouraging companies to properly attribute their work and respect existing intellectual property. As the AI industry continues to mature, the widespread adoption of similar authentication frameworks could help establish a more trustworthy and fair competitive environment where innovation is properly recognized and protected. The impact extends beyond individual cases of potential infringement to encompass broader questions of how the AI community can maintain scientific integrity while fostering continued innovation and competition.

6 Conclusion

This work introduces a robust fingerprinting methodology for large language models based on intrinsic parameter distribution characteristics, addressing critical intellectual property challenges in the rapidly evolving LLM landscape. Our approach leverages the statistical signatures of attention mechanism parameters, which remain stable even after extensive continued training and architectural modifications. Through comprehensive experimental validation across multiple model families, we demonstrate that these distributional fingerprints can reliably detect model lineage relationships with high accuracy. Most significantly, our investigation reveals compelling evidence that Huawei's Pangu Pro MoE model appears to be derived from Qwen-2.5 14B through upcycling techniques, as indicated by extraordinarily high correlation coefficients (0.927 overall average) that far exceed typical inter-model similarities. This finding not only highlights potential cases of model plagiarism in the current AI industry but also underscores the critical importance of developing robust authentication methods for protecting intellectual property rights. The persistence of these statistical signatures through claimed training on 15 trillion tokens demonstrates the fundamental robustness of our approach, making it a valuable tool for maintaining fair competition and ensuring proper attribution in large-scale model development.

Acknowledgment

This work is conducted by members in *Honest AGI* community and is not supported by any funding of other individuals or organizations. The authors declare no competing interests.

References

Akhiad Bercovich, Itay Levy, Izik Golan, Mohammad Dabbah, Ran El-Yaniv, Omri Puny, Ido Galil, Zach Moshe, Tomer Ronen, Najeeb Nabwani, et al. 2025. Llama-nemotron: Efficient reasoning models. *arXiv preprint arXiv:2505.00949*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-

Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Tingxu Han, Shenghan Huang, Ziqi Ding, Weisong Sun, Yebo Feng, Chunrong Fang, Jun Li, Hanwei Qian, Cong Wu, Quanjun Zhang, et al. 2024. On the effectiveness of distillation in mitigating backdoors in pretrained encoder. *arXiv preprint arXiv:2403.03846*.

K. He et al. 2022. On the security and forensics of large language models. *arXiv preprint arXiv:2210.01234*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Mifer, and Tom Goldstein. 2023. A watermark for large language models. *arXiv preprint arXiv:2301.10226*.

R. Kuditipudi et al. 2023. The robustness of watermarks for large language models. *arXiv preprint arXiv:2306.01235*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Lyu Lyu, Y. Li, H. Wang, Z. Zhang, T. Su, L. Sun, and B. Li. 2022. Reading between the lines: Fingerprinting and identifying language models. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 2413–2426.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Pete Walsh, Oyvind Tafjord, Nathan Lambert, et al. 2024. Olmoe: Open mixture-of-experts language models. *arXiv preprint arXiv:2409.02060*.

V. S. Sadasivan, S. Kumar, S. Balasubramanian, and S. Feizi. 2023. Can we trust your explanations? on the robustness of watermarked explanations. *arXiv preprint arXiv:2305.01236*.

Pamela Samuelson. 2023. Generative ai and copyright law. *Communications of the ACM*, 66(8):20–23.

Zhi Sun, Ta-Ying Liu, Tian-Shuo Li, Yi-Zhen Liu, Hong-Han Li, Fan-Lin Tseng, Yun-Chun Chen, and Hung-yi Lee. 2024. Proflingo: a fingerprinting-based intellectual property protection scheme for large language models. *arXiv preprint arXiv:2405.02466*.

Yehui Tang, Xiaosong Li, Fangcheng Liu, Wei Guo, Hang Zhou, Yaoyuan Wang, Kai Han, Xianzhi Yu, Jinpeng Li, Hui Zang, et al. 2025. Pangu pro moe: Mixture of grouped experts for efficient sparsity. *arXiv preprint arXiv:2505.21411*.

Qwen Team. 2024. Qwen2 technical report. *arXiv preprint arXiv:2412.15115*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. 2024. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, page 100211.

Boyi Zeng, Lizheng Wang, Yuncong Hu, Yi Xu, Chenghu Zhou, Xinbing Wang, Yu Yu, and Zhouhan Lin. 2024. Huref: Human-readable fingerprint for large language models. *Advances in Neural Information Processing Systems*, 37:126332–126362.

Xuandong Zhao, Yu-Xiang Zhou, Jian He, Vamsi Potluru, Di Chen, He He, and Furong Liu. 2023. Provable robust watermarking for ai-generated text. In *The Eleventh International Conference on Learning Representations*.