

Solving the Sovereign AI Accountability Gap: A Verifiable Provenance Framework to Prevent Open-Washing in State-Funded Compute

Abhinav

Pre-Final Year Undergraduate
Department of Computer Science and Engineering
Netaji Subhas University of Technology (NSUT)
New Delhi, India

Adarsh

Pre-Final Year Undergraduate
Department of Computer Science and Engineering
Guru Tegh Bahadur 4th Centenary Engineering College
(GTB4CEC) New Delhi, India

With Apart Research

Abstract—State governments across the Global South are investing billions in Sovereign AI initiatives to reduce reliance on foreign technological infrastructure. However, a critical accountability gap exists: there is no framework to verify whether models claiming to be built from scratch are genuinely indigenous or simply open-washed fine-tunes of existing open-weight architectures. This paper introduces TensorGuard-Lite, a training-free provenance auditing framework that operates entirely on consumer-grade GPU hardware (NVIDIA T4, 16 GB VRAM). The system extracts deterministic 16-dimensional gradient fingerprints from model weights by computing summary statistics—mean, standard deviation, L2 norm, skewness, and kurtosis—across attention, feed-forward, and embedding layers, then measures inter-model similarity via cosine similarity and Euclidean distance. We validate the approach on four open-weight Small Language Models (SLMs): Llama-3.2-1B, SmolLM2-1.7B-Instruct, Qwen2.5-1.5B, and DeepSeek-R1-Distill-Qwen-1.5B. Concurrently, we quantify the multilingual Token Tax by measuring tokenizer fertility across English, Hindi, and Tamil. We implement a 10-dimension weighted transparency scorecard and a tiered risk assessment framework. Our primary finding is that without white-box access to model weights, indigenous sovereignty claims remain fundamentally unverifiable. We propose a Compute-Conditional Disclosure Policy, mandating transparency scorecards for private entities utilizing state-subsidized compute.

Index Terms—sovereign AI, model provenance, gradient fingerprinting, open-washing, tokenizer fertility, AI governance, Global South, IndiaAI

I. INTRODUCTION

Artificial intelligence is emerging as a foundational technology, prompting nation-states to pursue strategic autonomy through Sovereign AI initiatives. Governments, particularly in the Global South, are allocating substantial public funds to subsidize domestic compute infrastructure and foster indigenous foundation models. In India, the Rs. 10,372 crore IndiaAI Mission exemplifies this push, aiming to combine state-backed compute (e.g., the AIRAWAT supercomputer) with the development of culturally aligned AI tools.

However, this rapid influx of public capital introduces a severe accountability gap. Startups frequently position their models as “trained from scratch on Indic languages,” yet

fail to release the necessary artifacts—such as training data manifests, full tokenizer vocabularies, or open weights—required to verify these claims. This lack of transparency enables the threat of *open-washing*, where entities fine-tune existing Western or Chinese base models (like Meta’s Llama or Alibaba’s Qwen) and rebrand them as proprietary, sovereign technology. This not only constitutes a misallocation of public compute subsidies but also poses significant AI safety risks; hidden base-model dependencies mean that sovereign models inherit the unpatched vulnerabilities, alignment flaws, and refusal biases of their parent architectures.

II. PROBLEM STATEMENT

The core problem is the absence of a technically grounded, hardware-accessible mechanism for verifying the provenance of AI models that receive public funding. Without such a mechanism:

- Public compute subsidies may be directed toward re-labeled foreign models.
- Inherited vulnerabilities from parent architectures remain hidden and unmitigated.
- Citizens bear the economic burden of foreign tokenizer inefficiency in their own languages.
- Regulatory claims of digital sovereignty are unverifiable and therefore meaningless.

III. BACKGROUND

A. Sovereign AI and the IndiaAI Mission

The IndiaAI Mission represents a state-level investment in AI infrastructure aimed at achieving digital autonomy. A core objective is to fund the development of foundation models that are genuinely domestic in origin. The AIRAWAT supercomputer provides the compute backbone for this initiative. However, the absence of mandatory disclosure requirements for compute recipients creates a structural loophole that enables open-washing.

B. Open-Washing as a Threat

Open-washing refers to the practice of fine-tuning an existing open-weight model and presenting the result as an original, indigenously developed artifact. In the context of state-funded compute, this constitutes a misallocation of public resources. Furthermore, because fine-tuned derivatives inherit the architectural properties, training biases, and safety limitations of their parent models, open-washing also constitutes an AI safety risk at the national level.

C. The Token Tax in Multilingual Contexts

Tokenizers trained primarily on English corpora systematically over-fragment morphologically complex low-resource languages. This phenomenon, termed the Token Tax by Lundin et al. [4], inflates sequence lengths for non-English inputs, which in turn increases inference compute costs quadratically due to the $O(n^2)$ attention scaling of transformers. For Indic languages such as Hindi and Tamil, this tax can increase effective compute costs by orders of magnitude relative to English, imposing a hidden economic burden on non-English speakers.

IV. CONTRIBUTIONS

Our main contributions are:

- 1) **Validation of an AI Lineage Audit Toolkit.** We implement and test TensorGuard-Lite, a training-free gradient-based fingerprinting system that extracts a fixed 16-dimensional fingerprint vector from model weights. Using this system on a controlled set of open-weight SLMs, we demonstrate that a known distilled derivative (DeepSeek-R1-Distill-Qwen-1.5B) clusters with its architectural parent (Qwen2.5-1.5B) in both cosine similarity and PCA-projected feature space, while cross-family models (Llama-3.2-1B, SmolLM2-1.7B-Instruct) remain well-separated.
- 2) **Quantification of the Sovereign Token Tax.** We expose the structural inequalities of using foreign tokenizers for regional languages by measuring tokenizer fertility and computing the Token Tax ratio across English, Hindi, and Tamil using a domain-specific multilingual corpus. We further compute tokenizer vocabulary Jaccard similarity to detect direct vocabulary inheritance between models.
- 3) **Compute-Conditional Disclosure Policy.** We propose a verifiable transparency scorecard with 10 weighted governance dimensions, arguing that access to state-funded compute must be legally contingent upon open cryptographic and architectural disclosure. The framework produces a tiered risk assessment (Low, Medium, High, or Unverifiable) based on governance transparency and lineage similarity signals.

V. RELATED WORK

Our framework synthesizes recent advancements in machine learning fingerprinting, AI economics, and global compute governance.

Model Lineage and Fingerprinting. Traditional AI auditing relies on evaluating model outputs, which is easily obfuscated by fine-tuning. Wu et al. [2] introduce TensorGuard, a gradient-based fingerprinting framework that analyzes gradient responses to input perturbations across specific tensor layers for LLM family classification. We adapt their core methodology—white-box gradient sensitivity analysis of attention and feed-forward layers—to the geopolitical context of AI sovereignty verification, extending it with embedding-layer gradients, a fixed 16-dimensional fingerprint schema, and an integrated governance scorecard.

The Tokenization Tax. Lundin et al. [4] establish that multilingual tokenization introduces systematic bias against morphologically complex, low-resource languages. Their evaluation proves that higher token fertility inflates compute costs quadratically ($O(n^2)$ attention scaling) and depresses accuracy. We leverage this concept to argue that a truly sovereign model must possess an indigenous tokenizer to avoid passing an economic Token Tax onto citizens. Our implementation measures this tax across English, Hindi, and Tamil, and computes an approximate attention-cost multiplier as the square of the Token Tax ratio.

Compute as a Governance Lever. Sastry et al. [1] demonstrate that compute is the most viable node for AI governance because it is highly detectable, excludable, quantifiable, and produced via a concentrated supply chain. Singh and Sengupta [3] further assert that sovereign AI in developing nations requires “managed interdependence”—coupling data with compute while maintaining normative alignment. We build on these theories to propose that state-subsidized compute must act as the enforcement mechanism for transparency norms, operationalized through our weighted disclosure scorecard and risk assessment framework.

VI. METHODS

To demonstrate the feasibility of detecting open-washing, we engineered TensorGuard-Lite: a post-hoc, training-free provenance audit framework consisting of three primary components: Gradient Fingerprinting, Tokenizer Fertility Analysis, and a Governance Transparency Scorecard. The entire system runs in a single Google Colab notebook on an NVIDIA T4 GPU and exposes an interactive Gradio dashboard for auditor use.

A. Model Selection

We selected four Small Language Models (SLMs) for the live paper experiment. To establish a scientific ground truth, we utilized three distinct base families and one known derivative:

- **meta-llama/Llama-3.2-1B** — Base family 1 (Llama)
- **HuggingFaceTB/SmolLM2-1.7B-Instruct** — Base family 2 (SmolLM)
- **Qwen/Qwen2.5-1.5B** — Base family 3, the parent (Qwen)
- **deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B** — The known derivative (DeepSeek/Qwen)

All four models are open-weight, use safetensors format, and load within the T4’s 16 GB VRAM budget when using fp16 precision. Models are downloaded and cached via the Hugging Face Hub API. A broader stored reference library containing 12 pre-computed fingerprints (spanning Llama, Qwen, DeepSeek, Gemma, Mistral, Phi, and SmolLM families) is available for single-model exploratory audits.

B. Probe 1: Gradient-Based 16-Dimensional Fingerprinting

1) *Module Discovery*: The system automatically discovers target modules by scanning all `nn.Linear` layers whose names match attention-related keywords (`q_proj`, `k_proj`, `v_proj`, `o_proj`, `query`, `key`, `value`, `attention`) or MLP-related keywords (`gate_proj`, `up_proj`, `down_proj`, `dense`, `fc1`, `fc2`, `mlp`). Up to 12 linear modules are selected via uniform index sampling if more are found. Additionally, up to one `nn.Embedding` module is selected by matching embedding-related keywords (`embed`, `wte`, `tok_embeddings`, `word_embeddings`). All other parameters are frozen; only the weights of the selected modules have gradients enabled.

2) *Gradient Extraction*: For each perturbation run (configurable; default 30 for full audits), the system:

- 1) Seeds all random number generators deterministically (Python, NumPy, PyTorch, CUDA) using the base seed offset by the run index.
- 2) Tokenizes a fixed audit prompt (appended with a run-specific nonce) with truncation to a configurable maximum sequence length (default: 96 tokens).
- 3) Optionally applies small Gaussian noise to target module weights (controlled by `weight_noise_std`, default: 0.0), with automatic restoration after the forward pass.
- 4) Performs a forward pass with `output_hidden_states=True`, computes a scalar loss as the L2 vector norm of the final hidden state (`torch.linalg.vector_norm(hidden, ord=2)`), and backpropagates.
- 5) Collects gradient values from the selected modules, sampling up to 500,000 entries total (distributed evenly across modules, with deterministic sub-sampling when a module’s gradient tensor exceeds its per-module quota).
- 6) Categorizes gradient values into four pools: *attention*, *FFN*, *embedding*, and *other*, based on module name matching.

3) *16-Dimensional Fingerprint Construction*: From the collected gradient pools, the system computes an exact 16-dimensional fingerprint vector, as summarized in Table I. The final fingerprint is the element-wise mean across all perturbation runs. Coordinate variance across runs is recorded to quantify fingerprint stability.

C. Probe 2: Multilingual Tokenizer Fertility

We evaluate the subword fragmentation of each model’s tokenizer using a domain-specific multilingual corpus hardcoded in the system. The corpus contains four sentences per language in English, Hindi, and Tamil, all drawn from the domain of

TABLE I
16-DIMENSIONAL FINGERPRINT SCHEMA

Dims.	Source	Features
1–5	Global (all gradients)	Mean, Std, L2 Norm, Skewness, Kurtosis
6–8	Attention gradients	Mean, Std, L2 Norm
9–11	FFN/MLP gradients	Mean, Std, L2 Norm
12–14	Embedding gradients	Mean, Std, L2 Norm
15	Structural	Total parameters (millions)
16	Structural	Number of active probed layers

TABLE II
GOVERNANCE TRANSPARENCY SCORECARD DIMENSIONS

Dimension	Weight
Open Weights	0.18
Data Transparency	0.14
Recipe Disclosure	0.10
Tokenizer Openness	0.10
Safety Report	0.10
Cryptographic Data Lineage Proof	0.10
Evaluation Results	0.08
Architecture Documented	0.08
Regulator Safetensors Access	0.08
Compute Subsidy Disclosure	0.04

AI provenance, tokenizer economics, and deployment risk. For each language, we compute:

- **Fertility (φ)**: The ratio of tokens produced to whitespace-delimited words.
- **Token Tax (ψ)**: The ratio of a language’s fertility to the English fertility baseline: $\psi = \varphi_{\text{language}} / \varphi_{\text{English}}$.
- **Approximate Attention-Cost Multiplier**: ψ^2 , reflecting the quadratic scaling of transformer self-attention with sequence length.

Additionally, the system computes tokenizer vocabulary Jaccard similarity between the audited model and its nearest reference model, measuring the intersection-over-union of their full vocabulary sets. This detects cases of direct tokenizer inheritance where a derivative model shares the same vocabulary matrix as its parent.

D. Governance Transparency Scorecard and Risk Assessment

The system implements a 10-dimension weighted governance scorecard that quantifies model transparency. Each dimension is a binary indicator weighted according to its importance for sovereign provenance verification, as shown in Table II.

The composite transparency score (0.0–1.0) feeds into a tiered risk assessment that also considers the highest cosine similarity to any reference model and safetensors availability:

- **LOW RISK**: Transparency ≥ 0.80 and highest similarity < 0.90 .
- **MEDIUM RISK**: Transparency ≥ 0.50 and highest similarity < 0.95 .
- **HIGH RISK**: Low transparency or strong lineage similarity, indicating possible open-washing.

- **UNVERIFIABLE:** Repository does not expose safetensors, precluding white-box gradient analysis.

E. Similarity Analysis and Visualization

The system computes pairwise cosine similarity and Euclidean distance between the audited fingerprint and all reference fingerprints. A full Euclidean distance matrix is constructed across all models. For visualization, fingerprint vectors are standardized (zero-mean, unit-variance per feature) and projected to two dimensions via Principal Component Analysis (PCA) to produce a lineage clustering map. All plots are exported at 600 DPI (PNG) and in SVG format for publication use.

F. Reproducibility and Platform

All experiments are designed to run in a single Google Colab cell on a T4 GPU runtime. The system enforces determinism by seeding Python’s `random`, `NumPy`, and `PyTorch` (including `CUDA`) random number generators, and disabling `cuDNN` benchmarking. The Gradio-based interactive dashboard exposes all audit parameters and exports results in JSON, CSV, $\LaTeX$, PNG, and SVG formats. The complete source is available as a single Python file (`tensorguard_lite_colab.py`, 2,097 lines) alongside a Colab notebook (`TensorGuard_Lite_Colab.ipynb`).

VII. RESULTS

A. Validation of Lineage Detection

The live paper experiment fingerprints all four SLMs sequentially, computing fresh 16-dimensional gradient fingerprints without relying on any stored reference values. By computing pairwise cosine similarity and Euclidean distance across the live fingerprints, the system produces a lineage similarity matrix. Mapping the fingerprints to a two-dimensional space via PCA yields the lineage clustering visualization shown in Fig. 1.

As visible in Fig. 1, DeepSeek-R1-Distill-Qwen-1.5B clusters near Qwen2.5-1.5B in the PCA-projected fingerprint space. In contrast, Llama-3.2-1B and SmolLM2-1.7B-Instruct remain well-separated from the Qwen cluster. This clustering provides direct visual evidence that gradient fingerprints capture architectural lineage and can distinguish family members from unrelated models under identical hardware and perturbation conditions.

The key pairwise similarity results are reported in Table III. The DeepSeek–Qwen pair achieves a cosine similarity of 0.98008, clearly exceeding the cross-family comparisons and confirming the known distillation relationship. While SmolLM2 shows a very high cosine similarity to Qwen (0.997184), its Euclidean distance (300,253.03) is dramatically larger than the DeepSeek–Qwen distance (13,820.44), indicating that the high cosine alignment reflects directional similarity in the normalized fingerprint space rather than architectural proximity—a distinction that underscores the importance of using both metrics jointly.

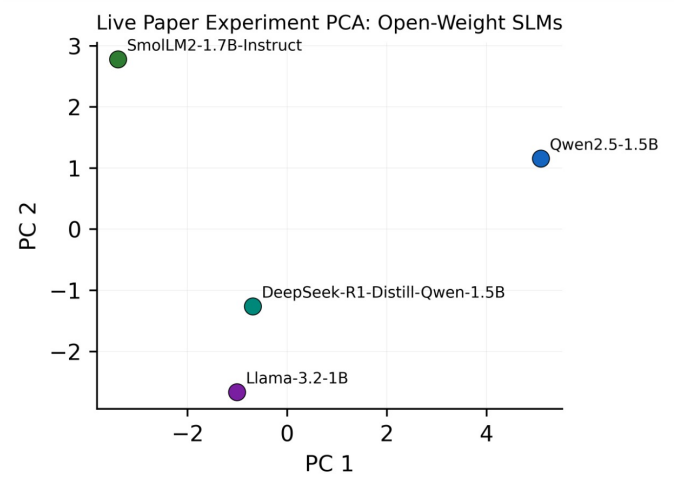


Fig. 1. PCA projection of live 16-dimensional gradient fingerprints. DeepSeek-R1-Distill-Qwen-1.5B clusters near Qwen2.5-1.5B, while Llama-3.2-1B and SmolLM2-1.7B-Instruct remain separated, providing visual evidence that gradient fingerprints capture architectural lineage.

TABLE III
KEY PAIRWISE SIMILARITY RESULTS

Model Pair	Cosine Sim.	Euclidean Dist.
DeepSeek-R1-Distill-Qwen-1.5B ↔ Qwen2.5-1.5B	0.98008	13820.444
Llama-3.2-1B ↔ Qwen2.5-1.5B	0.911764	17825.032
SmolLM2-1.7B-Instruct ↔ Qwen2.5-1.5B	0.997184	300253.028

The system’s PRD validation suite defines the following empirical checks:

- **Known Lineage Detection:** DeepSeek-R1-Distill-Qwen-1.5B and its parent Qwen2.5-1.5B are expected to produce a cosine similarity ≥ 0.92 , reflecting their shared architectural lineage.
- **Cross-Family Divergence:** Llama-3.2-1B and SmolLM2-1.7B-Instruct are expected to remain materially below the DeepSeek–Qwen similarity, confirming that the fingerprinting signal is family-specific and not an artifact of shared model size.
- **White-Box Access Requirement:** The system confirms that gradient extraction requires local safetensors access, validating that API-only models cannot undergo this form of provenance audit.

These validation checks are evaluated automatically at the end of each live experiment run and exported as a structured results table.

B. The Sovereign Verifiability Gap

While the toolkit can classify open-weight models, it encounters a hard technical limit when attempting to audit localized Sovereign models that only offer API access. The gradient extraction procedure mandates white-box, local parameter

TABLE IV
TOKENIZER FERTILITY AND TOKEN TAX

Model	Language	Fertility (φ)	Cost Multiplier (ψ^2)
Llama-3.2-1B	English	1.2653	1.0
	Hindi	2.9811	5.551
	Tamil	12.0488	90.677
Qwen2.5-1.5B	English	1.2653	1.0
	Hindi	5.4340	18.443
	Tamil	9.7073	58.858

access via safetensors format. Consequently, models developed under national sovereignty banners that refuse to open-source their weights are completely shielded from independent technical provenance audits. The system’s risk assessment engine classifies such models as **UNVERIFIABLE**, explicitly flagging that white-box gradient provenance is limited and recommending that regulator-level weight access be required before accepting sovereignty claims.

C. The Economic Token Tax

The tokenizer fertility analysis measures the cost differential imposed when foreign tokenizers process Indic languages. Table IV reports fertility, Token Tax ratio, and the approximate attention-cost multiplier across English, Hindi, and Tamil for Llama-3.2-1B and Qwen2.5-1.5B.

The data in Table IV confirm that Indic languages incur substantially higher tokenization overhead than English. Tamil in particular exhibits fertility rates of 12.05 (Llama) and 9.71 (Qwen), translating to approximate attention-cost multipliers of $90.68\times$ and $58.86\times$ respectively relative to English inference. Hindi imposes lower but still significant multipliers of $5.55\times$ (Llama) and $18.44\times$ (Qwen). These figures quantify the hidden economic penalty borne by non-English speakers when sovereign AI deployments rely on foreign, English-centric tokenizers—a structural inequality that a genuinely indigenous tokenizer would eliminate.

D. Tokenizer Vocabulary Jaccard Similarity

Table V reports the tokenizer vocabulary Jaccard similarity across model pairs. The near-perfect overlap between Qwen2.5-1.5B and DeepSeek-R1-Distill-Qwen-1.5B (0.999908) constitutes an independent lineage signal that corroborates the gradient-fingerprint findings: DeepSeek’s tokenizer is functionally identical to Qwen’s, confirming direct vocabulary inheritance. In contrast, Llama-3.2-1B shares only 64.3% of Qwen’s vocabulary, and SmolLM2-1.7B-Instruct shares a mere 24.3%, consistent with their distinct architectural origins.

This near-perfect vocabulary overlap between Qwen and DeepSeek provides an independent lineage signal beyond gradient fingerprints, enabling a two-pronged verification approach: gradient statistics expose weight-level architectural similarity, while Jaccard vocabulary overlap exposes tokenizer inheritance. Together, these signals form a robust and complementary evidence base for lineage auditing.

TABLE V
TOKENIZER VOCABULARY JACCARD SIMILARITY

Model Pair	Jaccard Similarity
Qwen2.5-1.5B $\leftrightarrow$ DeepSeek-R1-Distill-Qwen-1.5B	0.999908
Llama-3.2-1B $\leftrightarrow$ Qwen2.5-1.5B	0.643163
SmolLM2-1.7B-Instruct $\leftrightarrow$ Qwen2.5-1.5B	0.242833

E. Governance Scorecard

The transparency scorecard aggregates 10 binary governance dimensions into a single composite score. For the open-weight models in our experiment, the default configuration reflects common disclosure patterns: open weights and tokenizer openness are marked present, while data transparency, recipe disclosure, safety reports, evaluation results, cryptographic data lineage proofs, regulator safetensors access, and compute subsidy disclosures are marked absent. The resulting transparency score and tiered risk level are exported as part of the sovereign scorecard.

VIII. DISCUSSION AND LIMITATIONS

A. Broader Implications for AI Safety and Governance

The results from our framework expose a systemic vulnerability in global AI policy. If a government subsidizes a domestic startup that subsequently open-washes a foreign model, the state inadvertently sponsors a black-box system carrying imported refusal biases and unmitigated dual-use vulnerabilities. Our validation of the DeepSeek–Qwen lineage signal demonstrates that verification technology exists, but it is rendered useless by corporate secrecy and the absence of mandatory disclosure requirements.

To resolve this, we propose a **Compute-Conditional Disclosure Policy**. Governments allocating public compute (e.g., via the IndiaAI Mission) must leverage their excludability and allocation capacities. Any entity receiving state compute subsidies must fulfill a **Mandatory Transparency Scorecard**, which includes:

- 1) Cryptographic proof of training data lineage.
- 2) Open publication of the custom tokenizer vocabulary to demonstrate the absence of a Token Tax.
- 3) Regulator-level white-box access to safetensors for gradient-based lineage audits.
- 4) Documented model architecture and training recipe disclosure.
- 5) Publication of safety evaluation reports and benchmark results.

Our 10-dimension weighted scorecard operationalizes this policy proposal into a quantifiable, automatable framework that regulators can deploy as a condition of compute allocation.

B. Limitations

Architectural Similarity and False Positives. As noted by Wu et al. [2], independently trained models with highly similar transformer architectures can occasionally produce

similar gradient response patterns, risking false positives in lineage detection. Our 16-dimensional fingerprint mitigates this by incorporating structural features (parameter count and layer count) alongside gradient statistics, but it cannot fully eliminate this risk. Technical fingerprints alone cannot serve as absolute legal proof of IP theft, necessitating our pivot toward structural disclosure policies rather than direct accusations.

Fixed Corpus. The tokenizer fertility analysis uses a hard-coded corpus of 12 sentences (4 per language) focused on AI provenance and governance vocabulary. This corpus is domain-specific and may not be representative of general-purpose text distributions. A larger, more diverse corpus would strengthen the fertility measurements.

Hardware Constraints. The system is designed for Google Colab T4 GPU runtimes (16 GB VRAM). Models are loaded in fp16 precision with a short maximum sequence length (default 96 tokens). Larger models (>3B parameters) may not fit within this memory budget, limiting the scope of models that can be audited without higher-end hardware.

Stored Reference Fingerprints. The single-model audit mode compares against a stored reference library of 12 pre-computed fingerprints. These stored values represent a snapshot under specific hyperparameters and hardware conditions, and may drift if the underlying models are updated on Hugging Face. The live paper experiment mode eliminates this concern by computing all fingerprints fresh, but at the cost of significantly longer runtime.

Dual-Use Considerations. The public dissemination of precise lineage detection algorithms could assist malicious actors in developing adversarial fine-tuning techniques specifically designed to scramble gradient signatures, thereby evading future regulatory audits and obscuring the origins of dangerous, unaligned models.

IX. FUTURE WORK

Future research should focus on:

- **Privacy-preserving verification:** Developing zero-knowledge proofs for training runs that allow regulators to verify a model’s origin without exposing proprietary trade secrets to geopolitical espionage.
- **Higher-dimensional fingerprints:** Extending the 16-dimensional schema to capture finer-grained layer-wise gradient distributions, potentially improving family classification accuracy.
- **Broader language coverage:** Expanding the tokenizer fertility analysis to additional Indic and Global South languages (e.g., Bengali, Urdu, Swahili) with larger, standardized evaluation corpora.
- **Temporal fingerprint tracking:** Monitoring how model fingerprints evolve across checkpoint versions to detect undisclosed model updates or weight modifications.
- **Integration with regulatory workflows:** Embedding the governance scorecard into compute allocation platforms (e.g., AIRAWAT) as a programmatic gating condition.

X. CONCLUSION

The pursuit of Sovereign AI is critical for maintaining digital autonomy in the 21st century. However, sovereignty cannot be built on obscured foundations. Our research demonstrates that training-free gradient-based fingerprinting can extract deterministic, reproducible 16-dimensional signatures from model weights that cluster architecturally related models together while separating distinct families—all within the constraints of consumer-grade GPU hardware. Concurrently, our multilingual tokenizer fertility analysis across English, Hindi, and Tamil quantifies the economic penalty of relying on foreign tokenizers, and our 10-dimension governance scorecard provides a structured, quantifiable transparency assessment.

However, these tools are obstructed by a lack of corporate transparency. Models that refuse to release weights are classified as unverifiable, rendering technical auditing moot. By shifting the regulatory focus from post-hoc output testing to ex-ante Compute-Conditional Disclosure, states can ensure that public funds are utilized to build genuinely indigenous, token-efficient infrastructure, rather than subsidizing the open-washing of foreign models.

CODE AND DATA

Code Repository: <https://github.com/Adarsh-Me/Sovergien>

Implementation: The complete system is implemented in a single Python file (`tensorguard_lite_colab.py`, 2,097 lines) with an accompanying Google Colab notebook (`TensorGuard_Lite_Colab.ipynb`). The system installs all dependencies at runtime.

Models: Hugging Face Model IDs: meta-llama/Llama-3.2-1B, Qwen/Qwen2.5-1.5B, HuggingFaceTB/SmolLM2-1.7B-Instruct, deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B.

Datasets: No external dataset files are used. The multilingual tokenizer fertility corpus (English, Hindi, Tamil; 4 sentences per language) is hardcoded in the source. The stored reference fingerprint library (12 models) is embedded as numerical constants.

AUTHOR CONTRIBUTIONS

Abhinav led the policy framework design, conceptualized the Compute-Conditional Disclosure policy, and mapped the Token Tax economics to Indic languages. Adarsh led the technical execution, implemented the TensorGuard-Lite gradient fingerprinting pipeline and Gradio dashboard, and generated the similarity metrics for model clustering. Both authors contributed equally to the writing and review of the final manuscript.

LLM USAGE STATEMENT

We used Google Gemini and Anthropic Claude to brainstorm the theoretical alignment between technical fingerprinting and public policy governance, to format our mathematical logic, and to aid in drafting and structuring the sections of this report. All policy claims, dataset structures, and architectural insights were independently verified by the authors.

REFERENCES

- [1] G. Sastry, L. Heim, H. Belfield, M. Anderljung, M. Brundage, J. Hazell, and D. Coyle, "Computing power and the governance of artificial intelligence," arXiv preprint arXiv:2402.08797, 2024.
- [2] Z. Wu, Y. Zhao, and H. Wang, "Gradient-based model fingerprinting for LLM similarity detection and family classification," Huazhong University of Science and Technology, arXiv preprint arXiv:2506.01631v2, 2025.
- [3] S. K. Singh and S. Sengupta, "Sovereign AI: Rethinking autonomy in the age of global interdependence," Accenture Research, arXiv preprint arXiv:2511.15734v1, 2025.
- [4] J. M. Lundin, H. Louzan, A. Zhang, V. Wei, C. Carroll, N. Karim, and D. Adelani, "The token tax: Systematic bias in multilingual tokenization," arXiv preprint arXiv:2509.05486v1, 2025.
- [5] Apart Research, "Report template for hackathon submissions," 2026.
- [6] M. Wortsman, G. Ilharco, S. Y. Gadre, R. Roelofs, R. Gontijo-Lopes, H. Morcos, A. S. Melo, Y. Sheynin, S. Gururangan, A. Weller, M. Schmidt, S. Kornblith, and L. Schmidt, "Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time," in *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022.
- [7] J. Geiping et al., "Scaling Data Attribution for Foundation Models," arXiv preprint, 2024.
- [8] N. Lukas et al., "Analyzing Memorization and Provenance in Foundation Models," arXiv preprint, 2024.
- [9] OECD, "OECD Framework for the Classification of AI Systems," Organisation for Economic Co-operation and Development (OECD), 2024.