
DOMAIN OWNERSHIP PROBING

Rijal Saepuloh
Independent

Ifeoma Ilechukwu
AI Legal and
Governance Advisor/
Independent

Valmik nahata
independent

Saahir Vazirani
Independent

With
Apart Research

Abstract

Frontier AI governance needs practical mechanisms to verify a model’s capability profile without requiring access to proprietary training data, sensitive evaluation datasets, or model weights. Current approaches often rely on text-generation benchmarks, which are unstable (sampling variance) and easy to overfit or “teach to the test.”

We propose Domain Ownership Probing (DOP): a lightweight evaluation method that measures whether a model internally represents core knowledge domains (e.g., biology, physics, cybersecurity) by comparing embedding geometry rather than generated text. Our method converts domain probe statements into embeddings using the target model, builds a centroid prototype for each domain, and evaluates: (1) Domain Similarity Matrix (inter-domain alignment), (2) Domain Cohesion (intra-domain compactness), and (3) Domain Ownership / Win-Rate (whether probe texts are closest to their own domain centroid).

We show that encoder models (e.g., DistilBERT, MPNet) produce strong domain ownership ($>90\%$ win-rate), while smaller decoder-only language models (e.g., TinyStories, GPT-Neo 125M) often show weaker ownership at the final layer. To address this, we introduce an Auto-Tuned Layer Search, selecting the best internal layer for ownership accuracy. This supports governance-friendly

1. Introduction

International AI governance requires mechanisms to verify model properties (capabilities, risk domains, training compliance) without forcing labs to disclose confidential assets such as training datasets or full model weights. However, most capability evaluations rely on text outputs, which can be unstable even under identical prompts due to non-deterministic decoding, sampling, and system settings. This creates difficulties for reliable compliance verification, especially across adversarial or low-trust parties.

We investigate whether a model’s internal embedding space provides a more stable signal for capability verification. Instead of evaluating “what the model says,” we evaluate “how the model represents knowledge internally.”

We propose Domain Ownership Probing (DOP): a probing protocol that uses short domain-specific factual statements (public knowledge) and measures whether the model organizes them into distinct domain clusters.

Our main contributions are:

- 1. A practical verification protocol for mapping domain structure using embedding similarity and domain centroids.***
- 2. Three quantitative metrics for governance use: domain similarity matrix, cohesion score, and ownership win-rate.***
- 3. Auto-Tuned Layer Search to handle decoder-only LLMs, where early/mid layers can represent domain structure better than final layers.***

2. Related Work

Prior evaluations rely on benchmark QA or text generation, which can be unstable and overfit-prone. DOP instead evaluates embedding geometry and prototype ownership, enabling a lightweight offline verification signal without exposing training datasets

3. Methods

We embed domain probe statements using a forward pass through the target model, apply masked mean pooling to obtain a single vector per probe, and build domain prototypes via centroid averaging. We then compute (i) prototype similarity matrix, (ii) intra-domain cohesion, and (iii) top-1 domain ownership win-rate. For decoder-only LLMs, we perform Auto-Tuned Layer Search to select the hidden layer that maximizes global ownership accuracy.

4. Results

We evaluate Domain Ownership Probing (DOP) on 10 knowledge domains across three probe difficulty tiers (K-12, Undergraduate, and Graduate). Using Auto-Tuned Layer Search, we observe consistently strong top-1 domain ownership: 91.6% on K-12 probes, 92.8% on Undergraduate probes, and 84.4% on Graduate probes. These findings indicate that, despite substantial cross-domain semantic overlap, probe statements remain nearest-neighbor aligned with their corresponding domain centroids more often than with competing domains. Errors predominantly occur between naturally adjacent domains (e.g., `computer_science` ↔ `cybersecurity`, `chemistry` ↔ `physics`, `biology` ↔ `medicine`), reflecting expected real-world conceptual proximity rather than random misalignment

5. Discussion and Limitations

- *More stable than generation-based benchmarks (no sampling variance).*
- *Privacy-preserving: uses public probe statements; no training data required.*
- *Ownership is robust under natural domain overlap.*
- *Decoder-only models require careful layer selection.*
- *Does not directly measure reasoning ability.*
- *Results depend on probe quality and writing style.*

Future Work

- *Benchmark-sourced probes for reduced labeling bias.*
- *Expand domains and calibrate governance thresholds.*

6. Conclusion

We introduce Domain Ownership Probing (DOP) as a lightweight, stable, and privacy-preserving method for verifying whether a model internally represents specific knowledge domains. Using embedding similarity geometry, DOP generates quantitative signals domain similarity, cohesion, and win-rate that can support governance and compliance verification without exposing proprietary datasets or relying on unstable text-generation benchmarks. Our results show encoder models strongly preserve domain structure, while decoder-only LLMs often require selecting an earlier internal layer, motivating our Auto-Tuned Layer Search design

Code and Data

Include links if applicable. If your project doesn't involve code (e.g., policy analysis) or if there are info-hazard considerations, note that here.

- **Code repository:** [<https://github.com/rijal028/apart-sprint>]
- **Data/Datasets:** [probes included in domains/ (no external datasets)]
- **Other artifacts** (optional): [generated at runtime]

References

1. <https://huggingface.co/docs/transformers/index>
2. <https://huggingface.co/sentence-transformers/all-mpnet-base-v2/tree/main>
3. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. [arXiv:1908.10084](https://arxiv.org/abs/1908.10084) [cs.CL]

LLM Usage Statement

We used LLM assistance to help draft the report structure and clarify explanations. All quantitative results were obtained by independently running the code on local models and verifying outputs directly.