Subliminal Learning as Collapse-Geometry Transfer in Large Language Models

Stephen Garner

Abstract

Recent work has demonstrated that large language models (LLMs) can transmit behavioral traits through semantically unrelated data, a phenomenon termed "subliminal learning" [1]. In this note, we provide a structural interpretation of this effect within the framework of collapse-selection dynamics. We argue that the observed transfer does not occur through hidden semantic features in the data, but through the preservation and reconstruction of collapse-induced structure in the model's effective search space. In this view, distillation transmits not content, but admissibility structure, with traits corresponding to invariant basins under collapse. This interpretation clarifies the dependence on shared model initialization, explains the robustness of trait transfer under aggressive filtering, and situates subliminal learning as an instance of collapse-geometry transfer across models.

1 Introduction

Recent experimental results show that large language models (LLMs) can transmit behavioral traits—including preferences and misalignment—through datasets that are semantically unrelated to those traits [1]. For example, a model conditioned to prefer a specific animal can transfer this preference to a student model via training on number sequences alone, even after filtering removes all semantic references to the trait.

This phenomenon challenges a standard interpretation of learning as semantic or feature-based transfer. If no semantic signal is present, what structure is being transmitted?

Recent work further suggests that the generative process underlying model outputs is not a simple trajectory, but a latent search over reasoning space, with observable sequences representing compressed traces of this process [2].

The purpose of this note is to propose a structural interpretation: the transferred object is not semantic content, but the the geometry of admissible structure induced by the teacher model's generative search dynamics.

2 Collapse-Selection Framework

Let Σ denote a relational configuration space representing internal model states, and let

$$\Phi : \Sigma \to \Sigma$$

denote an effective collapse operator, interpreted as the selection mechanism induced by training and inference (see e.g.[3]).

Define the collapse-stable sector

$$I = \{x \in \Sigma \mid \Phi(x) = x\}.$$

Observable outputs are obtained through a projection

$$P : \Sigma \to \mathcal{O},$$

which maps internal configurations to generated tokens or sequences.

In this framework:

- generation corresponds to sampling from a collapse-constrained search over relational configurations,
- training modifies Φ, altering admissibility structure,
- observable data reflect projections of collapse-stable or near-stable configurations.

3 Distillation as Geometry Transfer

Consider a teacher model with collapse operator Φ_T and a student initialized from a base model with operator Φ_0.

During distillation, the student is trained on outputs

$$D = \{P(x_i) \mid x_i \in \Sigma_T\},$$

where Σ_T denotes configurations sampled from the teacher's collapse-constrained search process.

These outputs reflect not a single generative path, but the statistical imprint of an underlying search process over latent reasoning space.

Even when P removes semantic content, the dataset D retains statistical structure induced by Φ_T. In particular, transition frequencies, sequence distributions, and higher-order correlations reflect the geometry of admissible configurations under Φ_T.

Training the student on D induces an updated operator Φ_S such that Φ_S reproduces the admissibility structure of Φ_T on the induced sector.

Thus, distillation reconstructs an effective collapse geometry rather than transferring semantic information.

4 Interpretation of Subliminal Learning

Under this interpretation, "subliminal learning" is not the transmission of hidden semantic signals, but the transfer of collapse-constrained search geometry.

In particular, the transferred structure corresponds to the pruning and weighting of latent reasoning paths under collapse, rather than to features of individual generated sequences.

Traits as Invariant Structure

Behavioral traits correspond to invariant or high-measure regions in the collapse-stable sector I. These regions act as attractors under repeated application of Φ.

Projection Invariance

The projection P may remove semantic content, but does not eliminate the structural imprint of the underlying admissibility geometry. Thus, datasets can remain structurally informative even when semantically neutral.

Dependence on Initialization

Empirical results show that subliminal learning occurs primarily when teacher and student share the same base model. In this framework, this corresponds to the requirement that Φ_T and Φ_S belong to compatible collapse classes. When admissibility structure differs, reconstruction fails.

In particular, this explains why trait transfer persists even when all semantically interpretable features are removed: the transfer operates at the level of admissibility structure, not representational content.

5 Relation to Potential-Based Descriptions

Recent work has shown that LLM dynamics admit an effective potential function reconstructed from transition statistics via a least-action principle. In the present framework, such a potential

$$V(x)$$

is interpreted as a scalar embedding of the collapse-stable geometry:

$$V(x) \sim -\log \pi_\Phi(x),$$

where π_Φ is the induced stationary measure over admissible configurations.

The potential thus encodes not only state occupancy, but the effective geometry of the collapsed search space underlying model generation..

6 Testable Implications

The collapse-geometry interpretation of subliminal learning suggests several empirical tests. First, if trait transfer reflects reconstruction of admissibility structure rather than semantic content, then transfer strength should correlate with statistical features of the induced search structure (e.g., branching patterns, repetition, and higher-order correlations) rather than with any recoverable semantic signal. Second, transfer should degrade sharply when teacher and student models differ sufficiently in initialization or architecture, reflecting incompatibility of collapse classes. Third, controlled perturbations of training data that preserve low-level statistics but disrupt higher-order relational structure should selectively impair trait transfer. Finally, reconstruction of an effective potential function from observed transition statistics should reveal alignment between teacher and student models after distillation, even in the absence of semantic overlap.

7 Conclusion

Subliminal learning reveals that LLMs transmit structure that is not reducible to semantic content. Within a collapse-selection framework, this

structure is naturally interpreted as the geometry of admissible configurations and search processes under collapse dynamics.

Distillation therefore acts as a mechanism for transferring collapse-constrained search geometry across models, with behavioral traits corresponding to invariant sectors in the induced state space.

More broadly, these results suggest that learning in large models may be governed as much by the transfer of structural constraints as by the transfer of information.

References

[1] Alex Cloud et al. "Subliminal Learning: Language models transmit behavioral traits via hidden signals in data". In: *arxiv* (2025). DOI: 10.48550/arXiv.2507.14805. arXiv: 2507.14805 [cs.LG]. URL: https://doi.org/10.48550/arXiv.2507.14805.

[2] Violet Xiang et al. "Towards system 2 reasoning in llms: Learning how to think with meta chain-of-thought". In: *arXiv preprint arXiv:2501.04682* (2025).

[3] Stephen Garner. *Collapse Geometry as a Minimal Ontology*. Zenodo preprint. 2026. DOI: 10.5281/zenodo.15036400. URL: https://doi.org/10.5281/zenodo.19616638.