Beyond Neighbours: A Geometry Aware Study of Representational Convergence

Do the Platonic and Aristotelian hypotheses disagree only because we measure with the wrong geometry?

Mentors: Giovanni Marraffini & Trinidad Borrell

DISCLAIMER: We are open to other directions/techniques that align better with our mentees' interests/knowledge. However, the base of the project should contain something along the lines of what is written here.

Summary

What: Recent interpretability and control work suggests that steering along curved concept manifolds is more faithful than steering along linear directions (Wurgaft et al., 2026; Engels et al., 2025). That gain, however, comes at a cost: the manifold is estimated inside one model's latent space, so the intervention is model-dependent by construction and must be refit for every new model. This project asks whether that cost is necessary: *is the geometry those methods exploit private to each model or shared across models?*
Current evidence on cross-model sharing is measured with the wrong instrument. The Platonic Representation Hypothesis (PRH) held that as models scale their representations converge to a shared model of reality (Huh et al., 2024). The Aristotelian revision (Gröger et al., 2026) showed this is an illusion for global metrics: once CKA is calibrated against a permutation null, its scaling trend vanishes for both linear and RBF kernels, and only local metrics (mkNN, CKNNA) survive. Yet both linear and RBF CKA compare representations in ambient Euclidean space: the very assumption manifold steering abandons. If representations live on curved manifolds, two models could share the same intrinsic geometry while looking globally misaligned to any Euclidean metric. **We test whether geometry-aware comparisons recover a shared global signal CKA cannot see, and characterise when alignment emerges during pretraining.**

Motivation: Linear steering already works, at least locally, and is the workhorse of current activation-based control. Groups such as Goodfire argue that many concepts instead lie on curved manifolds, and that steering along those manifolds is more faithful than linear interventions (Wurgaft et al., 2026; Engels et al., 2025). That is a claim about structure within a single model, and it is logically separate from what this project tests: whether that structure is shared across models. This is the bottleneck for the method.
We attack the question where it is tractable: concepts, rather than the behaviours (deception, sycophancy) steering ultimately targets. Concepts are the only place where a calibrated, open, cross-model pipeline already exists (Huh et al., 2024; Gröger et al., 2026), letting us borrow both the models and the statistical discipline. The Aristotelian result leaves two questions open: (1) existence: does any pair of models share intrinsic geometry once compared with a curvature-respecting metric? and, (2) only if existence holds, generality: does that sharing extend across families and scale, as local alignment already appears to? That is the distinction between manifold steering being a per-family trick and a model-agnostic tool.

Background, Motivation & Hypotheses

The raw Huh et al. (2024) scores are confounded by width and depth; Gröger et al.'s permutation-null calibration removes the apparent global convergence (CKA; Kornblith et al., 2019) while local convergence persists. Crucially, RBF CKA fails too, so the problem is not

kernel nonlinearity: what linear and RBF CKA share is a dependence on the ambient Euclidean metric.
Two manifolds can be isometric under their intrinsic (geodesic) metric yet differ arbitrarily in how they sit in ambient space, and a Euclidean metric will call them misaligned. Two tools sidestep this: (i) kernels built from intrinsic distances: geodesic-graph or diffusion distances (Coifman & Lafon, 2006); and (ii) Gromov-Wasserstein (GW) discrepancy (Mémoli, 2011), which compares two metric-measure spaces by their internal distance structure alone, needing no shared embedding. **If global convergence is genuinely absent these confirm it cleanly; if it was hiding behind the wrong geometry, they reveal it.**

We test two hypothesis:

- **H1 (Geometry aware PRH):** The failure of global CKA reflects its reliance on the ambient Euclidean metric, not a true absence of global structure.
- **H2 (Alignment emergence):** Local (topological) alignment emerges early in pretraining and rises monotonically with tokens/compute. Global alignment, if present, emerges later and more abruptly: a signature of the manifold organising itself after neighbourhoods are set.

Manifold-based steering is the frontier of activation control, and its open weakness is portability: the manifold is fit per model, so every claim that a steering vector, probe, or detector transfers rests on untested shared structure. Much interpretability-based oversight assumes such universality (Jha et al., 2025); the Aristotelian result says it fails globally, but only through a Euclidean instrument manifold methods have already been rejected.
This project tests the assumption in stages. If geometry-aware metrics recover a signal where CKA fails, curved concept structure is real, not a measurement artefact. If calibrated GW or geodesic signals recover alignment between specific pairs, those models share shape. If the signal holds across the grid and with scale, it licenses the strongest claim: a manifold for a safety-relevant concept (deception, sycophancy, hazardous knowledge) could be mapped once and reused across models. If instead sharing is pairwise or absent, that is itself an actionable safety finding: transferred vectors or probes can silently fail on an unfamiliar model.

Methods

We build directly on the open Aristotelian codebase (https://github.com/mlbio-epfl/Aristotelian), reusing its models, its 1024 image-text pairs, and its permutation-null calibration, so every new metric is reported as a calibrated effect size rather than a raw score.

1) Reproduce the baseline.

Re-run calibrated linear-/RBF-CKA and mKNN/CKNNA on the VLM pair grid to confirm the global-collapse / local-survival pattern in our pipeline before adding anything.

2) Pick and probe one latent space.

Rather than averaging over everything, we select a single well-behaved latent space (the model/layer with the highest calibrated alignment to the rest) and study its geometry in depth: local intrinsic dimension (TwoNN, Facco et al., 2017; MLE), and diffusion-map structure. This exploratory step tells us which geometry-aware kernel is even appropriate before we scale up.

3) Test H1.

Replace the Euclidean/RBF kernel with a kernel built from geodesic (kNN-graph shortest-path, Isomap-style) or diffusion distances, then apply the identical permutation-null calibration. Compute Gromov-Wasserstein between the two intrinsic distance matrices and calibrate it against the same null. Report the calibrated global signal from each geometry-aware metric against CKA (fails) and mKNN (survives) on identical pairs, as a function of scale.

4) Test H2:

Repeat the calibrated local and geometry-aware global metrics across public pretraining checkpoints (Olmo as a cross-check) to plot alignment vs. tokens and locate emergence points / phase transitions.

Backup Plans

- **We know GW is too expensive at scale.** In case this becomes an issue we will fall back to entropic / low-rank GW with concept subsampling, or lean on the cheaper geodesic-kernel CKA.
- **We risk having noisy manifold estimates at n ~ 1024.** In that case we will increase n (find another dataset of pairwise images-text), report every result across multiple intrinsic-dimension / curvature estimators, and restrict deep analysis to the single well-populated latent space from Step 2.
- **If H1 and H2 finish early, the highest-value extension is a first, small-scale test of the downstream claim:** fit a manifold-based intervention for one concept in model A, transfer it to model B, and check whether our calibrated geometry-aware alignment score predicts transfer success across pairs where linear CKA does not.

Theory of Change

- **Removes the main structural limitation of the best non-linear control method we have.** Manifold steering is more faithful than linear steering, but is currently model-bound: each manifold is an artefact of one latent space. We test whether the geometry it exploits is shared. A positive result gives control researchers the licence (and the first calibrated evidence) to attempt a shared, model-agnostic manifold for a safety-relevant concept, mapped once and reused; a negative result tells them, before that effort is spent, that manifold control is inherently per-model and must be designed with refitting and validation built in. Concretely: it changes whether Goodfire-style manifold interventions are engineered as transferable objects or as per-model artefacts, and it tells anyone shipping a transferred probe whether they owe the field a transfer validation.
- **Corrects the universality picture underlying oversight tools.** Much safety tooling assumes a detector or steering vector found in one model transfers to another because they share structure. We test whether that sharing is real at the level manifold control needs (intrinsic geometry) or only at the weaker level already confirmed (neighbour rank), preventing labs from trusting cross-model transfer claims that current evidence doesn't actually support and preventing the field from abandoning a real shared structure because it was measured with a Euclidean ruler.

Scope & Concrete Output

Minimum output expected: On the fixed Aristotelian VLM grid, implement one geometry-aware global metric with permutation-null calibration, compare it against linear-CKA and mKNN, and answer H1 for that setting.

Deliverables: A GitHub repo extending the Aristotelian framework with geometry-aware metrics and calibrated GW. Depending on the impact of the results/amount of work done: A LessWrong blog post plus a reproducible notebook or paper sent for a safety workshop on a computer-science venue.

Risks and downsides (externalities)

Cross-model geometric alignment is dual-use: the machinery that lets a safety-relevant manifold transfer is close to work on recovering embeddings across models (Jha et al., 2025), with implications for model extraction and embedding-store privacy. We mitigate by

working on public VLM/LLM checkpoints and generic image-text concepts, and by reporting calibrated effect sizes rather than releasing tooling to transfer representations between specific proprietary models.

Bibliography

[1] Huh, M., Cheung, B., Wang, T. & Isola, P. (2024). The Platonic Representation Hypothesis. ICML 2024. arXiv:2405.07987.

[2] Gröger, F., Wen, S. & Brbić, M. (2026). Revisiting the Platonic Representation Hypothesis: An Aristotelian View. arXiv:2602.14486.

[3] Park, K., Choe, Y. J. & Veitch, V. (2024). The Linear Representation Hypothesis and the Geometry of Large Language Models. ICML 2024. arXiv:2311.03658.

[4] Engels, J., Liao, I., Michaud, E. J., Gurnee, W. & Tegmark, M. (2025). Not All Language Model Features Are Linear. arXiv:2405.14860.

[5] Wurgaft, D., Rager, C., Kowal, M., Shyam, V., Feucht, S., Bhalla, U., Haklay, T., Bigelow, E., Sarfati, R., McGrath, T., Lewis, O., Merullo, J., Goodman, N. D., Fel, T., Geiger, A. & Lubana, E. S. (2026). Manifold Steering Reveals the Shared Geometry of Neural Network Representation and Behavior. arXiv:2605.05115.

[6] Kornblith, S., Norouzi, M., Lee, H. & Hinton, G. (2019). Similarity of Neural Network Representations Revisited. ICML 2019. arXiv:1905.00414.

[7] Mémoli, F. (2011). Gromov-Wasserstein Distances and the Metric Approach to Object Matching. Foundations of Computational Mathematics, 11(4), 417–487.

[8] Coifman, R. R. & Lafon, S. (2006). Diffusion Maps. Applied and Computational Harmonic Analysis, 21(1), 5–30.

[9] Facco, E., d'Errico, M., Rodriguez, A. & Laio, A. (2017). Estimating the Intrinsic Dimension of Datasets by a Minimal Neighborhood Information. Scientific Reports, 7, 12140.

[10] Jha, R., Zhang, C., Shmatikov, V. & Morris, J. X. (2025). Harnessing the Universal Geometry of Embeddings. arXiv:2505.12540.