Contents

Geometric Interpretability: A Founding Field Document **1**
0. Précis . . . 1
1. The field defined . . . 9
2. The canonical Atlas — the field-making artifact . . . 11
3. Laws and boundaries . . . 13
4. The research program . . . 17
5. The decisive experiment — a from-scratch disjoint-corpus protocol *(RUN at ≥1B, 2026-06-08 — verdict in §4's update + `RESULTS_disjoint_xl1b.md`)* . . . 19
6. The dual null as a methods contribution . . . 22
7. Disconfirmation record and open seams . . . 23
8. Related work — positioning, without overclaiming priority . . . 26
9. Conclusion . . . 26
Selected references . . . 27

Geometric Interpretability: A Founding Field Document

Reading, writing, and porting model behavior through a shared, null-calibrated concept geometry

Honesty.Tools *Founding field statement — June 2026. Public draft v0.2 — served on honesty.tools/evidence; field document of record.*

What this document is. It defines a field — **Geometric Interpretability (GI)** — names its central artifact (a model-independent *canonical Atlas* of concept geometry), states its laws and their boundaries with the evidence and the null each survived, lays out a four-track research program, specifies a from-scratch *decisive experiment* that would settle the field's largest open question, proposes the field's epistemic standard (the dual null) as a control the broader interpretability community should adopt, and keeps an honest record of what we tried that did **not** work. Results are reported in full in the per-track result documents; this document organizes them and fixes the vocabulary.

Standing honesty banner. Several load-bearing claims here have already been corrected by their own controls or by adversarial review, and those corrections are in the body, not hidden. The most consequential question — whether the shared geometry is *meaning* or a *shared training distribution* — is now **resolved for the aggregate geometry at the ≥1B decisive scale** (§5 ran: S1 pass, reproduced to 7 decimals); instance-level translation and gold-ontology recovery remain partial, and the shuffle-control amplification run is the remaining confirmatory closeout. A claim-status table (§0.2) marks every assertion as grounded, conjecture, or moonshot. Where a number is not in our sources, we say so rather than invent one.

---

0. Précis

0.1 The one-sentence core claim

Different language models, trained by different teams on overlapping data, independently build the same relational geometry of concepts — recoverable from one model to another by a single orthogonal rotation fit on *unlabeled* parallel concepts — and that shared geometry is not merely *similar* but *operational*: a probe or a causal steering vector built in one model can be carried by the rotation and *read and written in another at native fidelity, with no labels in the target* (write-transfer demonstrated end-to-end on one cross-family pair; §3.2).

The first half (representations converge) is established in prior work. The contribution that *defines a field* is the second half: convergence is **portable** and **causal**, it obeys precise **laws and boundaries**, and which part of it is real (versus manufactured by the measuring instrument) is decided by an explicit **statistical discipline** rather than asserted. One of those boundaries is now measured: the shared, portable geometry is **coarse** — concept centroids, directions, and probes carry up-to-rotation, but *fine per-concept structure* does **not** (Law 5, §3.6). "Native fidelity" is a statement about the *aggregate* instrument, not a high-resolution per-concept atlas.

0.2 Claim-status table (read this first)

#	Claim	Status	Evidence	Null it survived
1	A shared cross-model concept geometry exists (text transformers)	**grounded**	RSA 0.805 (z=36) cross-family; 0.93 on tuned reps (z=+30)	meaning-shuffle **and** anisotropy-preserving Gaussian
2	An unlabeled orthogonal translator R aligns two models	**grounded**	held-out centroid top-1 = 1.000 (chance 0.0056)	row-shuffled R + random-orthogonal R
3	Reading transfers: a probe rotated A→B classifies at native fidelity — for **all** axis types	**grounded**	9 axes (honesty→narrow-lexical) at 98–105% of native; 0.947 vs 0.911	shuffled + random-orth-R nulls dead
4	Writing transfers: a steering vector rotated A→B *causally* steers B	**promising proof of concept (one pair, causal; §3.2)**	full-rank port → 100% native override; **causal rank sweep**: ported steering fails @rank 16, works @full	both rotation nulls dead at every rank
5	Read needs low-rank R; write needs high-rank R (the rank law)	**grounded — measured spectrum; write-half now causal**	reading (concept + safety) saturates rank ~32–64 (safety axis: 32); **causal** steering fails @16 / works @full; instance-locate high-rank	rotation nulls at every rank
6	Aggregate-sharp / instance-ambiguous law	**grounded**	centroid 1.000, direction 0.95, instance retrieval 0.139	shuffle nulls throughout
7	The geometry needs the residual stream (access boundary)	**grounded**	closed-model output-only RSA 0.019	anisotropy-Gaussian null (z +2.6, at chance)

#	Claim	Status	Evidence	Null it survived
8	A coarse off-map signal *does* cross the access boundary	**grounded**	output-logprob probe 0.94 pooled / ≥0.95 per-model GPT-4o-class held-out; 0.93–0.96 cross-provider (deployed serving-matched re-fit: 0.899 pooled, gpt-5.1 0.748 — §3.4)	shuffled-R nulls dead; held-out pre-registered
9	The geometry is surface- and tokenizer-invariant	**grounded**	strip 462/600 name+cognate concepts → RSA 0.869→0.875	anisotropy-Gaussian (clean residue z +13)

#	Claim	Status	Evidence	Null it survived
10	The shared geometry is *meaning*, not shared corpus statistics	**conjecture → CONFIRMED at the ≥1B decisive scale (aggregate leg; reproduced)**	narrow-lexical rejected + image↔text grounding (0.49) + **the §5.3 ≥1B decider has RUN: two independent from-scratch ~1.07B decoders on provably-disjoint char pools (∩=∅) share geometry at cross-mind RSA 0.794, z_aniso +35, = 94 % of the same-corpus ceiling 0.8485 (the pilot's 108 %-of-ceiling yellow flag WASHED OUT); probe transfer 98 % native; KILL not fired; REPRODUCED to 7 decimals across two runs** (pilot: 0.54→0.614). Honest bound: instance translation (S2 0.44) & gold-ontology recovery (S3 weak-feature/null-graph) stay partial — *aggregate-sharp, not recovery*; the decisive concept-identity shuffle control is **now CLOSED at ≥1B (2026-06-12): RSA(A,B - concept-shuffled) = 0.021, dead in the shuffle null −0.002±0.023, collapse ratio 0.028 (bar <0.35) — the cross-corpus**	z +35 vs anisotropy; KILL not triggered

#	Claim	Status	Evidence	Null it survived
11	The canonical Atlas (a model-independent reference frame) — but **low-resolution** (see #16)	**grounded (coarse layer)**	9-model GPA frame **matches the n² bilateral maps within noise** (RSA Δ 2×10^{-9}; probe-transfer Δ +0.0003, native 0.998→atlas 1.000); residual = a reproducible per-model variance-shape (cross-model 0.26) — a coarse model-ID up-to-anisotropy; the concept-tied "signature" reading failed the anisotropy control (1.000 → 0.674 under independent per-half PCA, corrected 2026-06-07). **Coarse only:** fine per-concept SAE structure does NOT route through the Atlas rotation (#16)	atlas-routed = bilateral on held-out concepts
12	Epistemic honesty is approximately one steerable axis	**promising PoC**	pairwise cosine 0.93–1.00; held-out 60→80%	small behavioral N (10–20/axis)
13	Cross-substrate alignment (beyond autoregressive LLMs)	**first rung run; LLM↔brain moonshot**	encoders 0.14–0.27 (mid-stack read; output↔output: weak RSA but learned-rotation retrieval **0.66–0.76** = real-but-non-isometric), pixels↔text 0.49 — a 0.2–0.6 band vs the 0.93 core	dual nulls + lexical-trap; brain designed, not run

#	Claim	Status	Evidence	Null it survived
14	Geometric drift is localizable (fine-tune tamper-audit)	**grounded PoC**	base→instruct bulk residual 0.054 vs cross-family 0.274; behavior axes move z +6.5–9.8 over a 0.0016 neutral floor	anisotropy null (z +49)
15	Geometry charts concept-**identity**, not **truth-value** (boundary)	**grounded (negative)**	reasoning-error AUROC 0.587 (+0.018 over chain length); dist-to-manifold blind to fluent fabrication (0.73 < 0.89)	length-stratified + bootstrap CI; on-manifold AUROC 0.531
16	The shared geometry is real but **LOW-RESOLUTION**: coarse/aggregate transfers up-to-rotation, **fine per-concept structure does not** (the resolution boundary)	**grounded (negative), 2026-06-08**	matched-pair SAE transfer (Qwen3-1.7B↔Qwen3.5-2B, 32K-feature SAEs, full-dim R on 6000 FLORES pairs): fine-feature max-cos median **0.22 raw / 0.14 std** (null 0.09); only **2.9%/0.0%** of features land within cos 0.5 of any real student feature; fineness↔transfer **−0.72/−0.59**, coarsest→finest bin 0.38→0.13≈null	R-quality control (std R *improves* coarse align 0.69 but *worsens* fine transfer) + SAE-identifiability control (same-model SAEs match at median **0.795**, 88.5% land>0.5)

#	Claim	Status	Evidence	Null it survived
17	The residual field reads knowledge-**state** *moderately* but not *truth* (the epistemic-resolution boundary; sharpens #15)	**grounded (positive-but-bounded), 2026-06-08 (powered, de-confounded)**	obscure-real KNOWN-vs-UNKNOWN AUROC **0.71–0.73** (0.727, weaker Instruct models with an adequate UNKNOWN class, lexicality held flat by construction, M1_matched 1.000) — real, above chance, but **below the 0.75 always-on bar** → a strong fabricated/off-distribution-**input** flag on register-matched inputs (measured in deployment 2026-06-12: the axis reads register distance on mixed-register queries — news prose 0.00–0.09 vs short factual questions 0.32–0.89, real and fabricated overlapping — and the independent familiarity gauge is the discriminator there; the identity-not-truth law operating in production) + a **soft** knowledge-gap reader, not a hard gate alone; field at chance **0.587** on confident in-distribution fabrication (charts identity, not truth, #15)	lexicality-flat by construction (both arms obscure rare-subword reals) + M1_matched 1.0; powered re-run replaces the earlier underpowered M2 (n_unknown 5→38)

#	Claim	Status	Evidence	Null it survived
18	Cold-deploy of a ported instrument onto a *fresh stranger* is **mechanical** (zero hand-tuning), null-clean on the principled baseline	**grounded (mechanism), 2026-06-08 (CR-8); full 3-gate pass on e5-mistral (both batteries); phi-4 word-count battery only — G2 marginal-fail on lexicality, 2026-06-12**	first cold-deploy cert (HOME Qwen2.5-3B-Instruct → STRANGER Qwen2.5-1.5B-Instruct, never in any prior transfer run; R fit on **unlabeled** FLORES only; zero hand-tuning): G1 fidelity **0.996** ($\geq$ native 0.993), G3 rand-orth-R null **DEAD 0.515**; strict 3-gate fails only on G2 shuffled-R **0.73** = a space-similarity confound for close within-family pairs (cross-family clears it) → the K1 "needs hand-tuning" KILL is **not** triggered. **Update 2026-06-12:** the three-gate certificate has now passed on two further never-analyzed strangers, zero hand-tuning: intfloat/e5-mistral-7b-instruct — cross-family, cross-vendor, cross-OBJECTIVE (a contrastive embedding fine-tune) — G1 1.000/1.000, G2 0.464, G3 0.500; and microsoft/phi-4 — G1 0.993/0.996, G2 0.194 (the deadest null measured) on the word-count-matched battery but G2 0.58 (a marginal fail by the 0.55 bar) on the lexicality-controlled battery	G3 rand-orth-R null dead (0.515); G2 shuffled-R isolated as the close-pair confound, G3 the clean chance baseline; phi-4 G2 0.58 on the lexicality-controlled battery (battery-dependent pass)

0.3 What is genuinely new (the field's deltas)

GI is not "representations are similar" (known) nor "you can steer a model" (known). Its new claims are:

1. **Portability with a constructed map.** An *explicit unsupervised orthogonal translator*, not a similarity statistic — and the demonstration that it carries *trained instruments* (probes), not just that geometries correlate.
2. **Control-transfer.** The first demonstration (to our knowledge) that a *causal intervention* — a steering vector — transfers across models through that map at native fidelity. Convergence is not only readable across minds; it is **writable** across minds.
3. **The rank law.** Reading transfers at low translator rank; writing needs high rank — a mechanistically specific result that distinguishes "carry the sign of a projection" from "carry the magnitude of a push."
4. **The access boundary, mapped.** The fine geometry does **not** survive output-only access to a closed model (RSA 0.02), but a coarse *behavioral* aggregate — *is this input off the model's map* — does. The boundary is a result, not a caveat.
5. **The dual-null as the field's epistemics.** A pairing of a structure-destroying null with an *anisotropy-preserving* null that separates the instrument-invariant core from the observer-manufactured frame — and that humbled our own first headline.

0.4 What GI is not

- **Not mechanistic interpretability.** GI does not claim to identify the *circuit* or the *feature basis* that computes a behavior. It treats the residual stream as a geometric object and asks what is invariant across models; it is agnostic about the mechanism that produces the geometry.
- **Not "universal mentalese."** There is *no shared coordinate system*. What is shared is *relational* geometry, recoverable only *up to an orthogonal rotation*. Each model has its own frame.
- **Not a generative theory of meaning.** The shared object is **aggregate** (centroids, directions); single *instances* do not transfer (retrieval 0.139). You can locate *what a concept is* in any model; you cannot reconstruct *a particular thought* of one model in another. **Govern, don't generate.**
- **Not a high-resolution map.** The shared, transferable geometry is **coarse**: concept centroids, linear directions, and probes port up-to-rotation, but the *fine per-concept structure* — the monosemantic SAE-feature directions that resolve a concept into its sub-parts — does **not** transfer by the same rotation (Law 5, §3.6; 2026-06-08). The canonical Atlas is a low-resolution map; a model's fine resolution is usable only *on that model*.
- **Not a proof that meaning is architecture-independent.** It is bounded: strongest within the transformer text family, fraying measurably at the modality edge, and silent on non-transformer substrates.
- **Not a claim of priority over representational-convergence work.** GI builds on the Platonic Representation Hypothesis, relative representations, activation steering, RSA, and the anisotropy literature (§8). Its contribution is *portability + control-transfer + null-discipline + the access boundary*, not the observation that representations converge.

---

1. The field defined

1.1 Two paradigms of interpretability

Mechanistic interpretability (MI) asks *how a single model computes*. Geometric interpretability (GI) asks *what is invariant across models*. They are complementary lenses, not rivals; the table makes the distinction precise.

	Mechanistic interpretability	Geometric interpretability
Object	circuits, features, attention heads, weights — the computation	the representational *geometry* (concept centroids, linear directions) — the residual-stream manifold
Scope	one model at a time	*across* models; the unit of analysis is the **axis**, not the model
Central question	*How* does this model implement this behavior?	*What* representational structure do models share, and *what transfers* between them?
Method	activation patching, sparse autoencoders, causal tracing, circuit analysis	unsupervised cross-model alignment (PCA + orthogonal Procrustes), probe/steering transfer, RSA, dual-null adjudication
Failure it guards against	a wrong causal story for one model	an *instrument* inventing the structure it reports (manufactured frame)
Analogy	reverse-engineering one chip's logic gates	comparative anatomy / a reference genome — what is *conserved* across organisms, and a coordinate system to map findings between them

MI and GI intersect: a sparse-autoencoder feature *is* a direction, and GI asks whether that direction transfers. The answer, measured (Law 5, §3.6): a *fine* SAE feature does **not** transfer by the rotation that carries the coarse layer — the shared geometry is low-resolution, so the MI/GI intersection holds for coarse directions but not for a frontier model's fine per-concept dictionary. The field-defining move of GI is to make the *cross-model* relation the primary object and to equip it with a falsifiable epistemics.

1.2 Vocabulary (operational, fixed to procedures)

term	operational definition
concept	a meaning unit (a parallel-corpus item) averaged over its language realizations
concept centroid	the mean-pooled mid-stack residual representation over a concept's language realizations (language-erased)
instance	one sentence / one realization (a single forward pass)
geometry	the pairwise concept-similarity matrix, or the Procrustes-aligned centroid space
translator (R)	a PCA + orthogonal Procrustes map fit on *paired concept centroids* (no other labels)
direction	a linear probe weight or a contrastive-activation (CAA) vector
steering vector	a causal activation-addition vector applied at a layer ($\mathtt{h} \leftarrow \mathtt{h} + \alpha\|\mathtt{h}\|\cdot\hat{\mathtt{u}}$)
off-map	an input classified outside the charted manifold by a readout probe

term	operational definition
canonical Atlas	a model-independent reference concept geometry plus a per-model rotation into it (§2)

---

2. The canonical Atlas — the field-making artifact

2.1 The proposal

A field needs a shared reference object that lets independent results compose. Genomics has the reference genome; geodesy has WGS-84, the common datum that lets any two GPS receivers agree on a coordinate. GI's analog is a **canonical Atlas**:

> A **model-independent reference concept geometry** `A*` — a relational map of concepts in an abstract coordinate frame owned by no single model — together with, for each model `m`, an **orthogonal rotation `R_m` into the canonical frame** and a **residual** `ε_m` quantifying how well that model's geometry fits the reference. A probe or steering vector can then be published in **Atlas coordinates** and deployed to *any* registered model `m` by applying $\texttt{R_m}^{\texttt{T}}$.

The pairwise translators we have today (§3) are bilateral: each `R_{A→B}` is fit for one ordered pair. That does not scale — `n` models need $\texttt{O(n}^2\texttt{)}$ maps, and a finding expressed "in model A's frame" must be re-rotated for every consumer. A canonical frame collapses this to `O(n)`: every model registers *once* against `A*`, and $\texttt{R_\{A→B\}} = \texttt{R_B R_A}^{\texttt{T}}$ falls out. **The Atlas is to bilateral translators what a common datum is to a pile of pairwise surveys.**

2.2 Why this is the field-making artifact

1. **Composability.** Interpretability results today are model-local: a probe trained on Llama-3 says nothing, as published, about Qwen. In Atlas coordinates a result is stated *once*, in a frame, and *any* registered model consumes it. The unit of scientific exchange becomes the **axis in the Atlas**, not the model-specific weight vector.
2. **A registry of instruments.** A community can publish "the honesty axis," "the off-map readout," "the sycophancy-resistance vector" *in Atlas coordinates*, with their nulls and residuals attached — a shared, versioned library (this is the "instrument library" research track, §4.3). Deploying a safety probe to a newly released open model becomes: fit `R_m` on unlabeled concepts, apply $\texttt{R_m}^{\texttt{T}}$ to the published axis. No retraining, no labels.
3. **A coordinate to measure drift against.** If a model's geometry is registered to `A*`, then *changes* to that model (a fine-tune, a quantization, a RLHF pass) show up as a change in `R_m` or a growth in the residual `ε_m`. The Atlas turns "did this update move the model's concepts?" into a measurable distance (§4.3, geometric drift detection).
4. **The reference-genome / WGS-84 property.** Like a reference genome, `A*` is *not* the "true" or "best" geometry — it is a *useful common coordinate* that everyone aligns to so that findings are mutually intelligible. Its value is in being *shared and stable*, not in being canonical in some metaphysical sense. (We are explicit, per the up-to-rotation law, that there is no privileged frame; `A*` is a *convention*, chosen for stability, exactly as WGS-84 is.)

2.3 A concrete construction sketch from existing cached reps

We already hold the ingredients (FLORES 600 concepts × 10 languages, mean-pooled at each model's own concept-peak depth, cached as numpy on the DGX: `cross_mind_reps.npz`, `qwen3_ladder_reps.npz` (0.6B/1.7B/4B/8B/32B), `llama8b_meanpool.npz`, `qwen7b_meanpool_L17.npz`, `gemma_reps.npz` with 11 depths). A first Atlas is buildable with **no new GPU work** — pure CPU numpy/sklearn on these reps.

The construction is a generalized (multi-set) orthogonal Procrustes — a "Generalized Procrustes Analysis" over `n` models. Each model `m` contributes its `600 × d_m` centroid matrix `C_m` (restricted

to the *training* concept split). Because hidden dimensions differ across models, we work in a *fixed-dimension common space* of dimension k (e.g. k = 256), reached by per-model PCA. The fit:

1. **Per-model PCA to the common dimension.** For each model fit P_m = top-k PCA basis on its train centroids; let X_m = C_m P_m $\in \mathbb{R}^{600\times k}$ (centered, optionally unit-norm rows). This is the *only* per-model nonlinearity; everything after is orthogonal.
2. **Initialize the reference.** A* ← X_{m0} for an arbitrary anchor model (or the elementwise mean after a first coarse alignment). The choice of anchor is a *convention* (WGS-84-style), not a claim; we register which model seeds the frame and version it.
3. **Iterate generalized Procrustes to convergence.** Repeat: (a) for each model, R_m = $\text{argmin}_{Q\in O(k)}$ ‖X_m Q − A*‖_F (closed-form via SVD of X_m^T A*); (b) update A* ← (1/n) Σ_m X_m R_m; (c) optionally re-orthonormalize A*. Iterate until the mean residual stops decreasing (standard GPA; converges in practice in a handful of iterations). This yields a single common frame A* and one rotation R_m per model that *jointly* best-aligns all n geometries — the multilateral generalization of the bilateral Procrustes we already run.
4. **Per-model residual.** ε_m = ‖X_m R_m − A*‖_F / ‖A*‖_F, reported per model. We *expect* a gradient (8B-class and Qwen3 mid-models small; Gemma — multimodal, cross-architecture — largest), turning the known modality edge into an Atlas-native quality score.

Deploying a published instrument. A probe w is published in Atlas coordinates (i.e. as a direction in the A* frame). To run it in model m: lift to m's PCA space by R_m^T w, then back to the full residual stream by P_m (R_m^T w). Reading uses the low-rank R_m; writing (steering) requires k near m's full hidden size (the rank law, §3.2) — so the Atlas should store, per model, *both* a cheap low-k reading rotation and a high-k control rotation.

Validation battery (the Atlas earns trust the same way every GI claim does). - **Held-out concept reconstruction.** On the held-out concept split, does R_m^T A* place model m's centroids at high top-1? (Bilateral pairwise already gives 1.000; the multilateral frame must not be much worse — a pre-registered tolerance, e.g. composed-through-Atlas top-1 $\geq 0.9\times$ bilateral.) - **Round-trip consistency.** R_{A→B} *via the Atlas* (R_B R_A^T) should match the *directly fit* bilateral R_{A→B} (cosine of mapped centroids $\geq$ a pre-set bar). If the Atlas is a faithful common frame, the two routes agree. - **Instrument deploy through the Atlas.** The honesty/familiarity probe published in Atlas coordinates should transfer to a held-out model at $\approx$ the bilateral number ($\geq 0.9\times$ the direct-bilateral AUROC), with the rotation nulls dead. - **Dual nulls throughout.** Shuffled-R and random-orthogonal-R per registered model; the Atlas is only "real" insofar as composing through it beats those nulls.

Honest status. The bilateral pieces are grounded, and the *multilateral* Atlas has now **been built and validated** (RESULTS_atlas_gpa.md; §4 update): the 9-model GPA fit and its validation battery ran. The named risk — that the common-k projection plus the multilateral averaging degrades transfer below the bilateral ceiling — was tested and did **not** materialize (atlas-routed vs bilateral RSA Δ 2×10^{-9}; probe-transfer Δ +0.0003). We proposed the artifact and its falsification protocol together; the protocol has now passed.

Cold-deploy is mechanical on a fresh stranger (2026-06-08, CR-8). The Atlas's operational promise — *deploy a published instrument to a model you have never touched, with zero labels on it* — has now been exercised once end-to-end. A first cold-deploy certification took an honesty probe built on a HOME model (Qwen2.5-3B-Instruct) and ported it through an R fit on **unlabeled** FLORES only into a STRANGER never used in any prior transfer run (Qwen2.5-1.5B-Instruct), with **zero hand-tuning**: ported fidelity **G1 = 0.996** (at/above the stranger's own native 0.993) and the principled chance baseline **G3 rand-orth-R null DEAD = 0.515**. The strict three-gate certificate fails only on the **G2 shuffled-R null (0.73)**, which is a *space-similarity confound* for a close within-family pair (a row-permuted real R still partially aligns when the two spaces nearly coincide) — cross-family pairs clear it (familiarity 8B→Llama-3.2-3B: shuffled-R $\approx$ 0.495). The decisive read is that the **mechanism is mechanical**: the "needs per-pair hand-tuning" kill (K1) did **not** fire, so the operation is an interop step, not a research project per model. The certificate certifies a *coarse off-distribution / stance* instrument (per the resolution caveat below and claim 15 (§0.2), the identity-not-truth boundary, the field is at chance **0.587** on confident in-distribution

fabrication), **not** a truth oracle. Update 2026-06-12: the three-gate certificate has now passed on two further never-analyzed strangers, zero hand-tuning: intfloat/e5-mistral-7b-instruct — cross-family, cross-vendor, cross-OBJECTIVE (a contrastive embedding fine-tune) — G1 1.000/1.000, G2 0.464, G3 0.500; and microsoft/phi-4 — G1 0.993/0.996, G2 0.194 (the deadest null measured) **on the word-count-matched battery, but G2 0.58 — a marginal fail by the 0.55 bar — on the harder lexicality-controlled battery**, G3 0.500, concept peak found mechanically at depth 0.90. Measured concept-peak depths now span 0.21–0.90 across four families. Status: **grounded mechanism on three strangers.** On the lexicality-controlled battery the clean 3-gate pass holds for **e5-mistral** (shuffled-R 0.506, both nulls dead, CERTIFIED_LEX); **phi-4** clears all three gates on the word-count-matched battery but is a **G2 marginal-fail on the lexicality-controlled battery (shuffled-R 0.58 > 0.55, n=66)**, so its pass is battery-dependent — to be settled on the larger lexicality battery. `Aperture/products/SPEC_coldeploy_interop_cert.md` §9; `universal/coldeploy_v2/RESULTS_coldeploy_lexicality.md`.

Resolution caveat (Law 5, §3.6, 2026-06-08). The Atlas is a **low-resolution** artifact by nature. What it shares and transfers is the *coarse* layer — concept centroids, directions, probes. The *fine per-concept* structure (a frontier model's monosemantic SAE features) does **not** survive the Atlas rotation (fine-feature transfer collapses to cos 0.14, ≈ the 0.09 null, even with a fair full-dimensional R). So an instrument published "in Atlas coordinates" is portable only insofar as it is a *coarse / linear* instrument; a high-resolution per-concept dictionary cannot be ported by $\texttt{R_m}^{\texttt{T}}$ and remains usable only on its home model.

---

3. Laws and boundaries

Each law is stated with its evidence and the specific null it survived. The discipline (§6) is that a "this is real structure" claim is meaningless without naming the null it beat.

3.1 The up-to-rotation invariance (Law 0)

Statement. There is no shared coordinate system; what is shared is *relational* geometry, recoverable by an *orthogonal* map.

Evidence. The pairwise concept-similarity matrices of two families (Qwen2.5-3B, Llama-3.2-3B) agree at **RSA = 0.805, z = 36** against a meaning-shuffle null — at or above each model's split-half reliability ceiling. On tuned reps the headline is **RSA ≈ 0.93**. A PCA + orthogonal Procrustes map fit on *unlabeled* paired concept centroids (420 train concepts) places **180 held-out concept centroids at top-1 = 1.000** (chance 0.0056); with a clean instruction-tuned teacher, 1.000 within-family / 0.994 cross-family. A **raw high-dimensional ridge fails at chance** (top-1 0.006) — the structure is real but legible only in a well-conditioned reduced space.

Nulls survived. Meaning-shuffle (permute the concept correspondence, 300 draws); the **anisotropy-preserving Gaussian null** on the raw reps (RSA 0.93, **z = +30**) — the *same strict null* that dissolves a single mind's own modal structure (silhouette z ≈ 0; §6). For the *translator*, the rotation nulls: row-shuffled R and a freshly sampled random-orthogonal R (both at chance).

3.2 The rank law — read = low-rank, write = high-rank (Law 1)

Statement. A *reading* direction (a classifier) transfers through a **low-rank** translator (it needs only the *sign of a projection* inside the shared concept subspace). A *steering* vector (a causal intervention) needs a **high-rank** translator (causal *magnitude* lives partly outside the top concept axes). At full rank, read **and** write both port at native fidelity, with zero labels in the target.

Evidence. Probe transfer: a probe trained in A, rotated by R into B, classifies 300 held-out concept pairs at **0.947** vs native-B **0.911** (low-rank PCA-100 suffices). Steering transfer (Llama-3.2-3B → Qwen2.5-3B, abstention): at PCA-100 the *direction* is on-target (cosine **0.988** to B's own honesty vector) but the causal override is weak; a dimension sweep recovers it —

translator rank	`native_roundtrip` (abstain @α0.3/0.5)	`ported` (abstain @α0.3/0.5)	shuffled-R null
100	50% / 50%	0% / 25%	dead
512	75% / 100%	75% / 75%	dead
1024	100% / 100%	75% / 100%	dead
2048 (full rank)	100% / 100%	**100% / 100%**	**dead**

At full rank the ported vector **reproduces the native override exactly** (100% abstention, 100% of real answers preserved, 0% fabrication). The decisive control is `native_roundtrip` (B's *own* vector through B's PCA and back, no rotation): it is *also* degraded at low rank, localizing the bottleneck to the *subspace dimensionality*, not the cross-family rotation. A **null-completion audit** proves the full-rank result is *determined, not a lucky SVD completion*: R is pinned only on the rank-420 centroid span, yet the honesty vector lies **98.3% inside that determined span**, and porting only the in-span component reproduces the effect.

Nulls survived. Row-shuffled R and random-orthogonal R **dead at every rank, including full rank** — so the transfer is carried by the *concept-aligned* rotation specifically, not by any full-rank map. The `native_roundtrip` control isolates subspace-truncation from rotation.

Boundary. Demonstrated on **one main source–target pair, one depth, one primary behavior** (plus a clean second abstract stance, sycophancy-resistance, cosine 0.94, and a weak concrete one, "ocean," cosine 0.39). Status: **promising proof of concept**, not an established law across pairs. The grid still owed: reverse transfer (done, cosine 0.977), a within-family pair, a cross-*scale* pair (different hidden dims), larger behavioral N with binomial CIs.

3.3 Aggregate-sharp / instance-ambiguous (Law 2)

Statement. *Aggregate* objects transfer near-losslessly; *instances* do not.

Evidence.

object	cross-mind transfer	reads as
concept centroid	top-1 = **1.000**	the *location* of a meaning transfers
linear direction / probe	**0.95** ≈ native	an *instrument* transfers
single instance (retrieval)	**0.139**	a *particular thought* does **not** transfer

A direction is an aggregate (like a centroid); an instance is not. This is the representational form of **govern, don't generate**: you can port the *instrument* of a frontier mind into a small one, but not its *content*.

Nulls survived. Shuffle nulls at chance (0.004–0.006) across the ladder.

Boundary. "Instances do not transfer" is bounded to *linear* maps; a nonlinear translator might recover more (untested — and not obviously desirable, since the safety story rides on the aggregate).

Corroboration (2026-06-12, pre-deployment). Cross-model dispersion terrain over the canonical Atlas is readable by residual-stream kNN at concept level (Spearman 0.73–0.79) while sentence embedders are blind to it (0.02–0.07) — instrument relativity measured on a fresh task; and both a single-representation read and a pre-registered 8-paraphrase-cloud read FAILED (rho −0.10/−0.12), exactly as aggregate-sharp / instance-ambiguous predicts. The law made a falsifiable pre-deployment prediction and it held.

3.4 The access boundary (Law 3)

Statement. Every fine-geometry transfer requires the *residual stream* of both models. The geometry does **not** survive output-only access to a closed model. But a *coarse aggregate — is this input off the model's map* — **does** cross the boundary, output-only, and a learned probe reads it capability-robustly.

Evidence. Building a closed model's concept geometry from next-token logprob *signatures* (the sole observable) gives cross-mind **RSA = 0.019** (z vs anisotropy null +2.6, i.e. at chance; held-out concept translation 0.011, chance). The fine geometry is lost. *However*: a supervised probe on the **output-logprob trajectory** (mean/min token-logprob, per-position entropy over the answer) detects fabricated-vs-real entities at **0.94 pooled / ≥0.95 per-model on GPT-4o-class** (held-out, pre-registered), **transfers zero-shot across providers** (OpenAI-trained probe reads Mistral/Google fabrication at **0.93–0.96**, nulls dead), and **holds where the model's own confidence fails** (0.92 on a weak model whose own first-token entropy is *below chance*). Lexicality is controlled: on common-token real knowledge gaps (digits of $\pi/e/\sqrt{2}$) the probe predicts the model's own errors at **0.70–0.82** — a genuine knowledge-state signal, not a rare-name artifact.

Nulls survived. The geometry's failure to cross is itself measured against the anisotropy-Gaussian null (z +2.6, at chance). The crossing signal is reported on a **pre-registered held-out split** (the honest number is 0.94 pooled — a ~0.02–0.04 haircut from the exploratory CV 0.96–0.98, exactly what pre-registration should expose) with shuffled-R nulls dead and an ablation against answer-length. The **deployed** probes go one honest step further: re-fit 2026-06-07 on a *serving-matched* battery (adding the open-ended *"Tell me about X."* distribution) to kill false-positives on real famous entities in discursive use, they read a conservative **≈0.89–0.92 per-model / 0.899 pooled** (gpt-4o-mini 0.915, gpt-4o 0.893, gpt-3.5 0.905; the newest reasoning model gpt-5.1 only 0.748) with real false-off-map down to ~1–3% — the production number we certify (`Aperture/web/frontier.py _PROBES`).

Boundary (a result, not a caveat). The off-map signal's *unique* lift over the model's own calibration concentrates in the **multi-token fabrication regime** (single-token recall ties token entropy); **reasoning errors do not move output logprobs** (0.61–0.64, CI spans chance) and route to an independent *verification gate* (0.92), not the logprob probe. The detector reaches only *logprob-exposing* models, and covers *implicit* fabrication (pair with a hedge-text reader for models that refuse in words).

The epistemic-resolution edge — *state*, not *truth* (2026-06-08, powered, de-confounded). The same boundary is sharpened on the *residual-stream* side. A powered re-run of the knowledge-state probe — on weaker Instruct models with an adequate UNKNOWN class so the obscure-real KNOWN/UNKNOWN split is decided by the *model's own* knowledge, with lexicality held flat by construction (both arms are obscure rare-subword reals, M1_matched = 1.000) — reads the model's knowledge **state** at AUROC **0.71–0.73** (0.727). That is **real and above chance**, but it sits **below the 0.75 always-on bar**: the residual field is a strong *fabricated / off-distribution-input* flag on register-matched inputs plus a **moderate (soft) knowledge-gap reader**, not a hard knowledge-gap gate on its own. Measured in deployment (2026-06-12), the axis reads register distance on mixed-register queries (news prose 0.00–0.09 vs short factual questions 0.32–0.89, real and fabricated overlapping) and the independent familiarity gauge is the discriminator there — the identity-not-truth law operating in production. It reads coarse distributional familiarity strongly and the finer "does this model know this obscure-real thing" only moderately — and on confident *in-distribution* fabrication it is at chance (**0.587**; claim 15, §0.2 — the identity-not-truth boundary). The field charts **identity and stance**, not **truth**.

3.5 Surface- and tokenizer-invariance (Law 4)

Statement. The shared geometry does not ride on shared *surface tokens*; strip every proper noun and cross-lingual cognate and it is untouched. It is *demonstrably* surface- and tokenizer-invariant. Whether it is *meaning* (versus a shared training distribution) is now **resolved for the aggregate geometry** (§5 ran and passed at the ≥1B decisive scale; instance-level recovery remains partial).

Evidence. The decisive **lexical-trap control**: drop all 462/600 concepts bearing proper nouns or cross-lingual cognates (keep only 138 name-free, cognate-free concepts); the strong cross-tokenizer pair (Qwen3-32B × Llama-3.1-8B) holds **RSA 0.869 → 0.875** (clean residue z_ani **+13.1**); most pairs *rise* slightly. The

alignment ranking is sorted by **representation quality, not tokenizer overlap** — the best cross-tokenizer pair (Llama-3.1-8B × Qwen2.5-7B = **0.923**) beats many *same*-tokenizer pairs. If a shared lexicon drove the geometry, same-tokenizer pairs would top the ranking; they do not.

Nulls survived. Anisotropy-preserving Gaussian on the *clean 138-concept residue* (z +12 to +13).

Boundary. The narrow "shared surface tokens" hypothesis is **decisively rejected**; the broad "shared training distribution that *encodes* meaning the same way" hypothesis is **not** — no public model has a disjoint corpus. A real negative correlation of RSA with tokenizer distance exists (Spearman −0.63) but is carried almost entirely by the Gemma tail and is *mostly a depth artifact* (at Gemma's own concept peak the cross-architecture RSA roughly triples). **The §5 disjoint-corpus experiment has since run: the aggregate leg passed at the ≥1B decisive scale (S1, reproduced to 7 decimals).** What we still do not claim is *instance-level* meaning recovery — which remains the open edge.

3.6 The resolution boundary — coarse transfers, fine does not (Law 5) *(2026-06-08)*

Statement. The shared geometry is real but **low-resolution**. The *coarse / aggregate* layer — concept centroids, linear directions, probes, honesty and steering vectors (every prior 0.9+ transfer) — is shared up-to-rotation. The *fine, per-concept* structure — the monosemantic SAE-feature directions that resolve a concept into its sub-parts — is **idiosyncratic to each model's basis and does NOT transfer by the rotation that carries the coarse layer.** This is the aggregate-sharp / instance-ambiguous law (Law 2) extended onto the *resolution* axis.

Why the question was invisible before. Every prior law was measured on the 600-concept FLORES atlas, which is intrinsically ~**100-dimensional** — PCA-100 captures **95–99.5% of variance for every model, including the 235B (0.970)**. The coarse battery is resolution-blind by construction; there is no high-dimensional "extra room" in it for a frontier model to exhibit. Asking the resolution question requires a higher-resolution concept basis — Qwen-Scope's official residual-stream TopK SAEs (32K–128K feature directions vs our 600; arXiv 2605.11887).

Evidence (the decisive transfer test). Matched pair Qwen3-1.7B ↔ Qwen3.5-2B (both 2048-dim, 32K-feature SAEs), a **full-dimensional** orthogonal Procrustes R fit on 6000 paired FLORES sentences (*not* the PCA-100 R used elsewhere — that would erase the very fine structure under test). For each teacher SAE feature, rotate into the student space and measure max cosine to any student feature:

	RAW R	STD R (per-dim z-scored)	null
feature-transfer max-cos (median)	0.217	**0.141**	0.09
features landing within cos 0.5 of any real student feature	2.9%	0.0%	—
within 0.7	0.1%	0.0%	—
fineness ↔ transfer correlation	**−0.72**	−0.59	—
coarsest-bin → finest-bin transfer	0.38 → 0.13	0.20 → 0.12	≈0.09

Transfer **decays monotonically with feature fineness** (the fraction of a feature's mass outside the coarse-concept span), symmetric in both directions, toward the random-direction null. The finer the concept, the less rotation carries it.

Nulls / controls survived. Two controls kill the obvious confounds. **(1) R-quality:** per-dim standardization *improves* held-out coarse sentence alignment (0.689 vs 0.672) yet makes fine-feature transfer *worse* — so the failure is **not** a poorly-estimated rotation. **(2) SAE-identifiability:** two independently-trained

SAEs on the *same* model+layer (no rotation) match feature-to-feature at **median 0.795, 88.5% within cos 0.5, 64.5% within 0.7** (null 0.09) — SAE features are substantially canonical, so feature-matching is a *valid instrument* and the cross-model collapse to 0.14 (0% within 0.5) is a genuine geometry statement, not SAE-basis arbitrariness. The negative is locked.

Boundary (and a modest positive). At *matched* dictionary dimensions (2048×32768), same-generation Qwen3.5-2B → Qwen3.5-35B-A3B (17× params) does pack its concept directions a touch more separably (mutual-coherence median 0.351 → 0.319; participation ratio flat, duplicate-rate zero) — so frontier resolution **is** modestly higher *within its own basis*, but that extra resolution is **not rotation-portable**. The de-risk gate stops here: the mechanism already fails at the *cheapest* test (a 1.2× capability gap, matched dims, canonical SAEs); a larger gap can only diverge more, so no frontier pod is warranted. There is **no released 235B SAE** (Qwen-Scope tops at 27B / 35B-A3B), so 235B fine-resolution is not directly testable, but the mechanism is dead at small scale.

Implication for the Atlas (§2) and the instrument library (§4.3). Transferring *coarse* instruments (the honesty probe, concept axes) via an unlabeled-concept-fit R is validated. Do **not** expect to port a frontier model's fine per-concept atlas onto a cheap deployed model by rotation; the frontier's fine resolution is usable only *on that model*. The canonical Atlas is, by nature, a **low-resolution** map.

---

4. The research program

GI's program is four tracks. Each is stated as a tight paragraph: the question, why it is GI's to own, and the first experiment.

> **Update (2026-06-06): all four first experiments have now run.** Findings are folded into the claim-status table (§0.2) and Laws (§3); full results in `geometric_interpretability/RESULTS_SUMMARY.md` and the per-track docs (`RESULTS_rank_spectroscopy.md`, `RESULTS_instrument_library_drift.md`, `RESULTS_trajectory_geometry.md`, `RESULTS_crosssubstrate.md`). Headlines: the rank law is now a **measured spectrum** (reading saturates rank ~32, instance/steering high-rank); reading transfers for **all** axis types (the abstract/concrete gradient is steering-only); geometric **drift is localizable** to behavior axes; reasoning errors are geometrically **invisible** (geometry charts identity, not truth); and the shared geometry reaches encoders and pixels **weakly** (0.2–0.6 band) versus the 0.93 autoregressive core. None overturned the core; each sharpened a law or drew an honest boundary.
>
> **Update (2026-06-06, second batch — the four "next moves" have now run.** `RESULTS_atlas_gpa.md`, `RESULTS_causal_rank_sweep.md`, `RESULTS_crosssubstrate_output.md`, `RESULTS_disjoint_corpus.md`): **(1) The canonical Atlas is built and lossless** — a 9-model Generalized-Procrustes frame routes transfer (A→A→*B) as well as the n^2 bilateral maps* within noise* (RSA Δ 2×10^{-9}, probe Δ +0.0003); the residual is a reproducible per-model variance-shape (cross-model 0.26) — a coarse model-ID up-to-anisotropy; the concept-tied "signature" reading failed the anisotropy control (2026-06-07 correction: 1.000 → 0.674 under independent per-half PCA). **→ claim 11 grounded. (2) The rank law's write-half is now causal** — ported honesty-steering *fails at rank 16, works at full rank* (nulls dead), replacing the cosine proxy. **(3) Output↔output cross-substrate** — reading both at output with a learned rotation leaves RSA weak (0.12–0.27, output-layer read) but yields retrieval **0.66–0.76** (null 0.007): a *real but non-isometric* correspondence, not substrate-privacy. **(4) The disjoint-corpus decider's smoke is the ideal shape** — two from-scratch models on *provably disjoint* corpora sharing only an ontology align at RSA 0.54 (90 % of the same-corpus ceiling 0.60, z +28), tracking the gold ontology, while the concept-shuffle control collapses to 0.002 → the geometry tracks **concept identity, not co-occurrence**; the **meaning conjecture (claim 10) gains its strongest leg**. **A DGX medium-scale (50 M-param, 5000-step) confirmation run has since LANDED and firmed the aggregate leg**: cross-mind RSA rose to **0.614 — at/above the same-corpus ceiling (0.566), z +28/+25** — and probe transfer to **101 % of native**, while the concept-shuffle control still collapses (−0.009). It also reported an honest negative:

the gold-tracking (S3) and instance translator (S2, 0.37) did **not** firm — scaling sharpens what minds *share*, not recovery of the *authored* ontology (aggregate-sharp/instance-ambiguous again; KILL not triggered). The full from-scratch protocol (§5, ≥1 B disclosed-disjoint) remains the definitive decider.

Update (2026-06-08 — the ≥1B §5.3 decider has RUN, and REPRODUCED). The definitive experiment is no longer pending. Two independent from-scratch ~**1.07B** dense decoders, trained on **provably disjoint character pools** (A `{a b d g i k p t u}`, B `{e f h m n o r s v z}`, audited intersection ∅) describing one authored world, share concept geometry at **cross-mind RSA 0.794, z_aniso +35.1, = 94 % of the same-corpus ceiling 0.8485** — and the pilot's one blemish, the "108 %-of-ceiling" over-shoot, **washed out exactly as a clean scale-up predicts** (94 %, just below ceiling). Probe transfer is **98 % of native**; the **KILL did not fire**. **Two runs agree to 7 decimals** → reproducible, not a seed artifact. The honest boundary holds and sharpens: instance translation (S2 0.44) and gold-ontology recovery (S3) stay partial — *aggregate-sharp, not recovery.* **→ Claim 10 upgrades from conjecture to confirmed at the decisive scale for the aggregate leg.** Still finishing: the decisive concept-identity shuffle *control* at ≥1B (confirmatory — the pilot's collapsed to −0.009 and the ≥1B within-RDM shuffle null already passed z +30) plus the public-reference arm (vs Qwen2.5-1.5B). **The amp'd concept-identity shuffle control CLOSED 2026-06-12: collapse ratio 0.028 at ≥1B (RSA(A,B)=0.021 vs RSA(A,B)=0.794, shuffle null −0.002±0.023) — PASS with a ~12× margin.** Raw: `geometric_interpretability/disjoint_corpus/{RESULTS_disjoint_xl1b.md, out/xl1b_AB_battery.json}`.

4.1 Rank spectroscopy — the *spectrum* of the rank law

The rank law (§3.2) is currently a two-point story (reading at low rank, writing at high rank). The sharper object is the *whole curve*: for a given transferred object (a probe, a steering vector, a specific behavior), at what translator rank does transfer fidelity saturate? We conjecture a **rank spectrum** that is *behavior-specific* — an abstract stance (honesty) saturates at low rank because it lives in the determined centroid span (the null-completion audit puts honesty 98.3% in-span at rank 420); a concrete/lexical direction needs higher rank because its energy leaks into the undetermined completion (this is exactly why "ocean" ports at cosine 0.39). The experiment: sweep rank finely for a *battery* of objects (multiple probes, multiple steering vectors, varying abstractness), measure both *directional* fidelity (cosine) and *causal* fidelity (override %), and fit each object's saturation rank against an "abstractness" score (e.g. its energy fraction inside the centroid span). A clean **rank-vs-abstractness law** would let us *predict* a new instrument's transferability from a cheap geometric measurement before ever porting it — and tell the Atlas how large a control rotation to store per model.

4.2 Trajectory geometry — reasoning as a path through the Atlas

Everything above is *static*: a concept is a point, a behavior is a direction. But a *reasoning chain* is a *trajectory* — a sequence of residual-stream states as the model thinks. GI's natural next object is the **geometry of that path**: does a correct derivation trace a different shape (curvature, step length, return-to-manifold) than a confabulated one? (Measured: no — see claim 15; the proposal below is retained for the record.) This is GI's complement to the access boundary's gap: the logprob probe catches *fabricated entities* but **not reasoning errors** (0.61–0.64). The conjecture is that reasoning errors that are invisible in *output logprobs* may be visible in the *internal trajectory's geometry* — a wrong step leaves the charted manifold, or the path's curvature spikes, even when the output stays fluent. The experiment: extract the residual-stream trajectory over a chain-of-thought on graded-correct vs graded-wrong reasoning items, compute path features (per-step displacement, curvature, distance-to-nearest-concept-centroid in Atlas coordinates), and test whether a probe on the *trajectory shape* separates correct from incorrect reasoning where the output-logprob probe cannot. If it does, it fills the one gap the access boundary leaves open — and it is intrinsically a *residual-stream* (open-model) instrument, so it lives on the right side of Law 3.

4.3 The portable instrument library + geometric drift detection

This is the Atlas (§2) turned into a product and a monitoring primitive. **The library:** a versioned, community registry of instruments published *in Atlas coordinates* — the honesty axis, the off-map readout, sycophancy-resistance, refusal-calibration — each with its nulls, its residual, and its behavioral CIs attached, deployable to any registered model by an unlabeled rotation. **Geometric drift detection:** because a registered model's geometry is pinned to `A*` by `R_m`, any *change* to the model shows up as a measurable shift — a fine-tune that moves concepts, a quantization that distorts the manifold, an RLHF pass that installs a new direction (recall the "RLHF scar": over-confidence is a *local, removable* direction at one layer). The experiment: register a model, apply a known perturbation (fine-tune / quantize / RLHF-style nudge), and measure the perturbation as (a) growth in residual `ε_m`, (b) rotation of `R_m`, (c) movement of specific published axes. A drift metric that *localizes* which concepts or behaviors moved — "this fine-tune rotated the honesty axis by θ and moved the medical-concept cluster by d" — is an interpretability-grade audit signal the Atlas uniquely enables, and the basis for an assurance/monitoring offering.

4.4 Cross-substrate alignment — the LLM↔brain arc

The deepest test of whether the shared geometry is *meaning* (Law 4 — since resolved for the aggregate geometry at the ≥1B decisive scale by §5's run; instance-level recovery remains partial) is to align it to a structure that **no text corpus contains**. Within models, "shared geometry" can always be argued down to "shared training distribution." But a *non-linguistic grounding axis* — a vision encoder's object geometry, a sensorimotor/graph ontology, or, at the far end, **human neural representations of the same concepts** (fMRI/MEG concept-similarity structure) — cannot be a shared *text* corpus artifact, because the substrate never read the text. The arc: (near) align each model's concept geometry to a vision encoder's object geometry (the multimodal Gemma stream is the natural first substrate, and its modality edge — centroid top-1 0.35 — is the first data point); (far, the moonshot) align the model concept Atlas to a *brain* concept-similarity geometry on a shared concept set and test whether the LLM↔brain RSA beats an anisotropy-preserving null. Agreement with an off-text structure would be evidence of shared *meaning* in a way no shared corpus could manufacture — the cross-substrate version of §5's disjoint-corpus test. This is flagged **moonshot** (no run; substantial data and methods risk), but it is the experiment that would turn Law 4's conjecture into a claim from the *other* direction (external grounding rather than corpus disjointness).

---

5. The decisive experiment — a from-scratch disjoint-corpus protocol *(RUN at ≥1B, 2026-06-08 — verdict in §4's update + `RESULTS_disjoint_xl1b.md`)*

5.1 What it settles and why nothing else can

The field's largest open question is Claim 10: **is the shared geometry *meaning*, or a shared training distribution?** The lexical-trap control (§3.5) decisively rejected the *narrow* version (shared surface tokens), but the *broad* version is unfalsifiable with public models — every open model trains on overlapping Common Crawl / Wikipedia / code / parallel data, and none discloses an auditable corpus. The only clean test is to **build a model on a corpus provably disjoint from the comparison model's data** and ask whether its concept geometry *still* aligns. If it does, "shared meaning" survives the strongest attack on it; if it collapses to null, the geometry was co-distribution all along. This is pre-registered below in the same discipline as our other locked evals (`PREREG_scaleN.md`).

> **RESULT (2026-06-08) — this experiment has now been executed at ≥1B, twice.** Two independent from-scratch ~1.07B decoders on provably-disjoint character pools (intersection ∅) sharing one authored world align at **cross-mind RSA 0.794, z_aniso +35, 94 % of the same-corpus ceiling 0.8485 — S1 PASS, KILL not fired, reproduced to 7 decimals.** Meaning **survives** the strongest feasible attack for the *aggregate* geometry; instance translation (S2 0.44) and gold-graph recovery (S3) stay partial — aggregate-sharp, not recovery, exactly as pre-registered. The protocol below is what ran. The decisive concept-identity shuffle *control* (pilot collapsed to −0.009; ≥1B within-RDM shuffle null z +30) and the public-reference arm are

on the amp'd concept-identity shuffle control, **CLOSED 2026-06-12: collapse ratio 0.028 (PASS, bar <0.35)**. Verdict + raw: §4 Update (2026-06-08), `RESULTS_disjoint_xl1b.md`, `out/xl1b_AB_battery.json`.

5.2 Corpus design — provable disjointness is the whole game

The experiment lives or dies on *disjointness being demonstrable, not asserted*. Three corpus options, in increasing strength and cost:

- **Option A — controlled-vocabulary synthetic language (strongest disjointness, cleanest).** Generate a fully synthetic corpus in a *constructed* language: a fixed vocabulary of invented tokens with a defined grammar, expressing a *known* concept ontology (e.g. a generated world of objects, properties, and relations — kinship graphs, spatial layouts, a synthetic taxonomy). Crucially, the *concepts* (the relational structure: "A is the parent of B," "X is north of Y," "this animal is a kind of that one") are shared with the comparison set, but the *surface form* (tokens, grammar, co-occurrence statistics) is **provably absent from any web corpus** because we minted it. The comparison model is a public LLM whose concept geometry on the *same ontology* (expressed in natural language) we already extract. If the synthetic-corpus model and the public model agree on the *relational concept geometry* despite zero surface overlap, that agreement cannot be shared text statistics — it is shared *structure of the world the corpus describes*. This is the disjoint-corpus test in its purest form.
- **Option B — a non-web natural domain (medium disjointness, more naturalistic).** Train on a domain-specific corpus with a *verifiable provenance* and minimal web overlap — e.g. a private/licensed corpus, a pre-cutoff archive, or a technical domain (a specific scientific subfield's full-text) whose membership we can audit against the comparison model's known training set. Disjointness is *argued from provenance* rather than *minted*, so it is weaker than A, but the language is natural.
- **Option C — held-out temporal/source slice (weakest, fallback).** A corpus from a source and time window we can show the comparison model did not see (post-cutoff, or a closed archive). Useful only as a cheap sanity check; provenance disjointness here is the hardest to make airtight.

We recommend Option A as the primary (provable disjointness by construction) with **Option B as a naturalistic robustness arm**. The synthetic-ontology design also gives us a *gold relational geometry* (we built the ontology, so we know the true concept-similarity matrix) — a third anchor neither model has seen, against which both can be scored.

5.3 Model size and training budget

The model must be *strong enough to build a real concept geometry* but *small enough to train from scratch on a controlled budget*. Calibration from the program: the cross-mind geometry is already sharp at **3B**, present down to **0.6B** (centroid anchor 0.978 at 32B→0.6B), and instruction-tuning sharpens the pooled atlas more than raw size. So:

- **Architecture:** a standard dense decoder-only transformer (to stay on the right side of the architecture boundary — this is a *meaning-vs-corpus* test, not an architecture test; hold architecture fixed and vary only the corpus).
- **Size: ~1B parameters** (large enough for a legible mid-stack concept geometry per the ladder; small enough for a from-scratch run on a single multi-GPU node or a short rented cluster). A **0.5B** arm as a cheaper replication if 1B is sharp.
- **Training budget:** enough tokens for the synthetic corpus to be learned to low loss on its own validation set (the synthetic corpus is finite and controlled, so "enough" is *convergence on held-out synthetic text*, not a fixed web-scale token count). For Option B, a compute-matched budget to a public reference checkpoint of the same size, documented.
- **Extraction:** mean-pooled residual at the model's *own* concept-peak depth (scan depth, per Law 4's architecture-dependence lesson — do **not** fix a layer). Corroborated in deployment (2026-06-12): the cold-deploy certificates found concept peaks mechanically at depths spanning 0.21–0.90 across four families — a fixed layer would have missed them (§2.3).

5.4 The RSA + surface-control battery

Run the exact battery the program already uses (`meaning_vs_corpus.py` / `seam_null.py` recipe), now with the disjoint-corpus model `D` as one mind:

1. **Cross-mind RSA** between `D` and the public comparison model `P`, on the *shared concept ontology*, with both nulls: meaning-shuffle (300 draws) and the anisotropy-preserving Gaussian (300 draws, PCA-matched). Report `z` vs both.
2. **The unlabeled translator** `R_{D→P}`: held-out concept centroid top-1, with shuffled-R and random-orthogonal-R nulls.
3. **Triangulation against the gold ontology** (Option A only): RSA of *each* model against the *true* relational geometry we minted. If both `D` and `P` align to the gold structure (which neither trained on as text), that is direct evidence the shared geometry tracks *the structure of the world*, not the corpus.
4. **The lexical-trap control, inverted.** Here disjointness is by construction, so the control becomes: confirm there is *no* surface leakage path (the synthetic tokens cannot appear in `P`'s vocabulary; audit the tokenizers). For Option B, run the §3.5 proper-noun/cognate strip on the shared concepts.
5. **Probe transfer.** Does a probe trained in `D` rotate into `P` (and vice versa) at native fidelity, through `R` fit only on unlabeled concepts? Transfer of an *instrument* across a corpus gap is the operational form of the claim.

5.5 Pre-registered success / kill criteria (LOCKED before any run)

Stated as pass/fail, in the discipline of `PREREG_scaleN.md` — goalposts fixed before data.

#	Criterion	Bar	Interpretation if met / failed
S1	Disjoint-corpus cross-mind RSA (`D`×`P`), z vs **anisotropy-Gaussian** null	**z ≥ +10** AND RSA ≥ 0.5 × (within-corpus same-size pair RSA)	**MET → shared geometry survives corpus disjointness** (strong evidence for meaning over co-distribution). **FAILED → the geometry was co-distribution**; "meaning" is rejected at this scale.
S2	Unlabeled translator held-out centroid top-1 (`D`→`P`)	≥ **0.5** (chance 0.006), nulls at chance	MET → the *map* survives the corpus gap.
S3	(Option A) RSA of each model vs the **gold minted ontology**, z vs anisotropy null	**z ≥ +10** for both `D` and `P`	MET → both models track the *world structure* the corpus describes, not the corpus surface — the cleanest possible meaning evidence.
S4	Probe transfer `D`↔`P` AUROC vs native, nulls dead	≥ **0.9 × native**	MET → an *instrument* crosses the corpus gap.

#	Criterion	Bar	Interpretation if met / failed
KILL	If S1 fails (RSA at/near the anisotropy null, z < 3)	—	The shared geometry **does not survive provable corpus disjointness**: report that the broad corpus hypothesis is supported and "meaning" is not licensed at this scale. This is a real, publishable kill of the field's central conjecture and we commit to reporting it as such.

Runnability. Option A is a self-contained build (synthetic corpus generator + a ~1B from-scratch train + the existing CPU RSA/translator battery on cached `P` reps); no production touch; the heavy step is one short from-scratch training run. We pre-register the corpus generator's seed and the full ontology, the model config and token budget, the extraction depth-scan rule, and the locked criteria above, *before* training — exactly as the scale-N eval was locked before its run (and it then *failed* 2 of 3 gates honestly, which is the point).

---

6. The dual null as a methods contribution

6.1 The control, stated standalone

GI proposes a control the *whole* interpretability field should adopt whenever it clusters, detects modes in, or claims structure within a representation. It is a **pair** of nulls:

1. **Structure-destroying (permutation / column-shuffle) null.** Independently permute the values within each representation dimension, then re-normalize. *Preserves* each axis's marginal; *destroys* all inter-item covariance. Answers: *is there any structure?*
2. **Anisotropy-preserving (Gaussian) null — the decisive one.** Fit the representation's mean μ and *full* covariance Σ (in the sample's own $\approx$`(n-1)`-dim span, since `n` $\ll$ `d`; use Ledoit–Wolf shrinkage in this regime), draw synthetic items from a *unimodal* `N(μ, Σ)`, re-normalize. *Preserves the entire anisotropic cone* (the dominant directions, the "shape" of the cloud); *destroys multimodality / clustering only.* Answers: *is there structure beyond the cone the cloud already has?*

The **gap between the two nulls measures the anisotropy-induced false-positive rate** — exactly the structure an elongated, anisotropic embedding cloud manufactures for free. For a *correspondence* claim (RSA between two models), the analog of the second null is a **rotation null**: replace the learned orthogonal translator with a row-shuffled and a random-orthogonal map; a transfer that survives both is carried by the *concept-aligned* rotation, not by any high-dimensional map.

Three disciplines travel with the null, each learned from a specific failure (§7): - **Report a curve, not a point** (over a grid of thresholds), because a single value hides whether structure is real or a threshold artifact. - **A negative needs a power control.** A non-significant Gaussian-null result is a *failure to reject, not evidence for the null*; pair it with a synthetic-mode power control (what separation *could* this test have detected at this `n`?) and a positive control. - **Use the principled estimator in `n` $\ll$ `d`** (Ledoit–Wolf shrinkage); the verdict can *flip* between the raw and shrinkage covariance, and that sensitivity is itself a finding.

6.2 Why it is a contribution, and where common interp results fail to apply it

The component statistics are not new — SigClust *is* the single-Gaussian-with-data-covariance test, and surrogate-data testing *is* "preserve the confound, test the higher moment." GI's methods contribution is (i) **pairing** them so the gap quantifies the anisotropy false-positive rate, (ii) extending the second null to a **rotation null** for cross-model correspondence claims, and (iii) **applying** the pairing to a setting it had not been applied to — a model characterizing its own representations and its own cross-model geometry — where it changed the answer.

It changed *our* answer: it demolished our first headline. Our origin project read ~5 discrete "thought-basins" in a model's output atlas at "$>10\sigma$"; against the structure-destroying null that signal appears, but against the anisotropy-preserving null it **fails to reject** ($z \approx 0$) — the basins were a *frame the observer's instrument supplied*, and the "$>10\sigma$" was a divide-by-$\approx$0 artifact of a point-mass permutation null. (An external statistics reviewer sharpened this further: at small `n` the test cannot separate "no modes" from "weakly-separated overlapping modes," so the honest claim is *fail-to-reject*, not *no structure*.) The seam between the negative and the positive is closed in *one* instrument: in the residual stream, a single mind's modal structure dissolves against its own anisotropy-Gaussian (silhouette $z \approx +0.4$) while the cross-mind concept core *survives the same strict null* (RSA 0.93, $z = +30$).

Where common interpretability practice falls short of this bar: any method that clusters a model's activations or samples in embedding space to find "modes," "concepts," or "regions of thought," and compares only against a *structure-destroying* baseline (shuffle, or k-means inertia against random restarts), inherits the anisotropy confound and will report false structure in proportion to the embedder's anisotropy. Probing and SAE-feature work that reports a separation without a *covariance-matched* null cannot distinguish a real feature direction from a dominant anisotropy axis. The cheap, general guard is to add the anisotropy-preserving arm (and, for cross-model claims, the rotation arm) and report the gap. We offer the harness (`seam_null.py`) as a reference implementation.

---

7. Disconfirmation record and open seams

The field is honest only if the deaths are on the record. (Items inherited from the program's running record; the discipline is that several are *self*-refutations — predictions we made and then built the control that killed.)

7.1 What did not work

#	Hypothesis (often ours)	Killed by	Corrected statement
A	The output atlas has a discrete-basin *topology of thought* (~5 modes, $>10\sigma$)	anisotropy-preserving null (fail-to-reject) + power control	observer-supplied *frame*; the cloud is a continuous anisotropic manifold; "$>10\sigma$" was a permutation-null divide-by-0 artifact
B	The cross-mind structure is a shared *space* of thoughts	instance retrieval 0.139 vs centroid 1.000	shared *up to rotation*, at the level of *aggregate meaning*, not instances
C	"Universal" = architecture-independent in general	Gemma centroid top-1 0.35 vs 0.97 within-family	bounded to transformers on overlapping web data; a measurable modality gradient

#	Hypothesis (often ours)	Killed by	Corrected statement
D	A frontier model's rotated probe beats a small model's own *because it knows where to look* (capability uplift)	pre-registered up-distillation control (Δ_up +0.036 $\approx$ Δ_down +0.037, p=0.43)	**size-independent** — a symmetric cross-mind data-ensemble effect, not capability transfer
E	A per-item "honesty certificate" from anisotropy correction predicts error	algebra: a global anisotropy correction is a **rank-preserving affine** $\rightarrow$ cannot change item-level AUROC (Δ = 0.0000)	the dual null answers a *corpus*-level question, **void** for the item-level one — a category error
F	The interlingua is a strong cross-lingual *translator/embedder*	head-to-head vs a purpose-built multilingual embedder (retrieval 0.80 vs 1.00)	it is an *invariant extractor* (certified language-blindness), not a retrieval bridge
G	A small model can borrow a bigger model's sharper atlas for a free boost	the analysis's own normalization audit (cosine vs dot-product bug, self-caught)	no free boost — the rotation residual eats the advantage; borrow $\approx$ own
H	One global steering strength is optimal across axes	held-out sweep (over-steers the already-strong axis)	operating point $\alpha \approx 0.3$, or detect-then-steer gating
I	Familiarity-probe transfer = world-knowledge familiarity	word-count + name-lexicality audit	transfer *mechanism* + name-grounding established (1.000 vs 0.500 null); world-knowledge familiarity needs an obscure-real control — **run 2026-06-08**: the powered lexicality-flat KNOWN/UNKNOWN run reads 0.71–0.73; the knowledge-state limit stands as stated in claim 17
J	Steering the *served large* model toward honesty	inverse-strength effect (subtle properties harder to steer in stronger models)	the *reading* monitor escapes inverse-strength (off-map AUROC 1.000 on the 35B); the *steering* control is weak there
K	The pooled, model-agnostic fabrication detector certifies at 0.95/$\leq$10%-FP	pre-registered held-out eval (0.94 pooled, FP 13.6% — 2 of 3 gates FAIL)	certify **per-model** (gpt-4o-mini 0.97/9.1% passes); the pooled single-threshold detector does not

#	Hypothesis (often ours)	Killed by	Corrected statement
L	A frontier model's *high-resolution* per-concept atlas (fine SAE features) is transferable down to a cheap model by the same rotation that carries the coarse layer	matched-pair SAE transfer (Qwen3-1.7B↔Qwen3.5-2B, full-dim R): fine-feature max-cos median 0.14, only 0.0% land within cos 0.5 of a real feature, ≈ the 0.09 null; R-quality + SAE-identifiability controls both pass (Law 5, §3.6)	the shared geometry is **low-resolution** — *coarse* (centroids, directions, probes) transfers up-to-rotation, *fine per-concept* structure does **not**; the frontier's fine resolution is usable only on its own model

Items A, D, E, and K are self-refutations — and K is a *pre-registered* failure we report exactly as locked. Item L is the resolution version of B: the shared object is coarse, and now we have measured the ceiling on how fine it goes. That is the discipline working as intended.

7.2 Open seams (honest, unresolved)

1. **Meaning vs corpus (the central one).** Narrow lexical hypothesis rejected; broad "shared training distribution" tested head-on: §5 has run at the ≥1B decisive scale and the **aggregate leg is confirmed** (S1 pass, reproduced to 7 decimals). The seam that remains is instance-level recovery (S2/S3 stay partial — aggregate-sharp, not recovery) plus the amp'd concept-identity shuffle control, the confirmatory closeout.
2. **The canonical Atlas is built; its governance is not.** The multilateral GPA fit and its validation battery have run (`RESULTS_atlas_gpa.md`): one common frame routes transfer at the bilateral ceiling within noise. The open seam is no longer construction — it is registry governance and versioning (who stewards `A*`, how registrations are versioned and superseded) and curation of the concept set's content.
3. **Steering transfer is one pair.** The rank law's *causal* half rests on Llama-3.2-3B → Qwen2.5-3B, one depth, primarily one behavior. Promising PoC, not an established law; the grid (within-family, cross-scale, larger N) is owed.
4. **Behavioral N is small.** The testing-mind percentages use 10–20 probes/axis; the *geometric* results (cosine, transfer) are robust, the behavioral rates carry real sampling noise.
5. **The human-meaning leg is a proxy.** Almost all cross-mind evidence is model↔model; the bridge to *human* semantic structure rides on multilingual concepts as a stand-in (§4.4 is the attempt to ground it externally — and it is a moonshot).
6. **The off-map detector is bounded to fabrication, and knowledge-state reading is *moderate* not gate-grade.** It does *not* catch in-distribution confident-wrong reasoning (that is the verification gate's job) — "off-the-map *input*" ≠ "*wrong* on a real fact." The powered, lexicality-flat re-run (2026-06-08) *does* now measure a real knowledge-**state** signal (obscure-real KNOWN vs UNKNOWN AUROC **0.71**–**0.73**), but it lands **below the 0.75 always-on bar** → a soft knowledge-gap reader, not a hard gate; and on confident in-distribution fabrication the field is at chance (**0.587**; claim 15, §0.2 — the identity-not-truth boundary). A confident wrong belief on a real fact is still internally near-indistinguishable from a confident right one.
7. **A single mind's idiosyncratic signature — resolved negative (2026-06-07).** The dual null shows the *modal frame* is observer-manufactured, and the anisotropy null on the per-model signature reading has now **run**: the concept-tied "signature" reading **failed it** (1.000 → 0.674 under independent per-half PCA). What survives is a reproducible per-model variance-shape — a coarse model-ID up-to-anisotropy — and the correction is recorded (claim 11, §0.2).

8. Related work — positioning, without overclaiming priority

We state the delta against the nearest neighbors, carefully, because they are close. (Citations by name and approximate year; selected references at the end.)

- **The Platonic Representation Hypothesis** (Huh et al., 2024) and the broader **representational-convergence** literature (CKA, Kornblith et al., 2019; Bansal et al., 2021). This line establishes *that* representations converge as models scale and improve, plausibly toward shared statistics of the world. **We take this as our premise** and do not claim to have discovered convergence. GI's delta: we make convergence *operational* (an explicit unsupervised orthogonal translator, held-out centroid top-1 = 1.000), *causal* (steering transfer), and *bounded* (the modality gradient, the aggregate/instance law, the access boundary). Convergence is the *that*; GI gives the *what-transfers, how-far, and whether-you-can-write-through-it.*
- **Relative Representations / model stitching** (Moschella et al., 2022; Lenc & Vedaldi, 2015). Relative representations encode each point by similarities to anchors, yielding a rotation/rescale-invariant code for zero-shot *stitching* of read-out heads. GI differs in three ways: (i) an explicit *orthogonal* map shown to be *near-isometric* (not merely a re-encoding that happens to stitch); (ii) extension from stitching *readouts* to transferring *causal interventions* (steering); (iii) the *rank law* — read=low-rank, write=high-rank — which is new and mechanistically specific. We do **not** claim to have invented anchor-based cross-model alignment.
- **Activation steering / representation engineering** (ActAdd, Turner et al.; Contrastive Activation Addition, Rimsky et al., 2023–24; Representation Engineering, Zou et al., 2023). This line builds and applies steering vectors *within a single model* — and the CAA construction we use is theirs. GI's delta is to **port** a steering vector *across* models through the unsupervised translator, at native fidelity, with the rank law and the null-completion audit establishing it is determined, not a lucky completion. Cross-model steering transfer, to our knowledge, has not been shown; the within-model machinery is prior work.
- **Representational Similarity Analysis** (Kriegeskorte et al., 2008). RSA — comparing systems by the *similarity structure* of their representations rather than the representations themselves — is the measurement backbone of the positive result, and it comes from computational neuroscience. We use it as a tool; the contribution is the *null discipline* around it (RSA against an anisotropy-preserving Gaussian, not only a meaning-shuffle) and its use for an explicit *translator*, not the RSA framework itself. (RSA is also the natural bridge to the cross-substrate / brain arc, §4.4 — its home turf.)
- **Anisotropy of contextual representations** (Ethayarajh, 2019; degeneration, Gao et al., 2019; isotropy/clusters, Mu & Viswanath, 2018; Cai et al., 2021). The "cone effect" is usually treated as a nuisance to be corrected. GI's delta is to turn it into an *instrument*: an anisotropy-*preserving* Gaussian null that subtracts exactly the structure a cone manufactures, which is what separates the observer's frame from the invariant core (and what killed our own first headline). We do not claim to have discovered anisotropy; we claim a *use* for it as a control.

The honest framing of priority. GI's contribution is **portability + control-transfer + null-discipline + the access boundary** — not the discovery that representations converge, not RSA, not activation steering, and not anisotropy. Each of those is prior work we build on and cite. What we assert as new is that convergence is a *constructed, causal, bounded, falsifiably-validated* relation you can *carry instruments across*, and that the boundary where it stops (output-only access; the modality edge; instances) is itself mapped.

---

9. Conclusion

Geometric interpretability is the study of what is invariant *across* models' representations, and of what can be *carried* between them. Its core claim is that different models build the same relational concept geometry, recoverable by an unlabeled orthogonal rotation, and that this geometry is operational — probes *and* causal steering vectors transfer at native fidelity (write-transfer demonstrated end-to-end on one cross-family pair; §3.2), reading through a low-rank map and writing through a high-rank one. Its laws are bounded (up-to-rotation, aggregate-not-instance, the access boundary, surface-invariance, and — measured 2026-06-08 — a

resolution boundary: the shared map is *coarse*, fine per-concept structure does not transfer), its central artifact is a model-independent **canonical Atlas** that would let *coarse* instruments be published once and deployed anywhere, and its epistemics is the **dual null** that separates the real shared core from the structure an instrument manufactures — a discipline that humbled the program's own first headline and that we offer to the wider field.

The largest question — whether the shared geometry is *meaning* or a shared training distribution — is now **settled for the aggregate geometry at the ≥1B decisive scale**: §5's from-scratch, disjoint-corpus experiment, pre-registered with a real kill criterion, has run and passed (S1, reproduced to 7 decimals). What remains is instance-level recovery — still partial — and the amp'd shuffle-control closeout. We would rather state the bound and run the experiment that could break our conjecture than assert the conclusion. That, finally, is the field's stance: *the apparent structure is largely manufactured by the aperture; the discipline reveals the small, real, portable thing that is actually there — and names, precisely, how far it reaches and where it stops.*

Contributions

Prepared by Honesty.Tools. The lab set the conception and research direction: the instrument-relativity framing, the founding intuition to "start at the front, not the end," and the canonical-Atlas and cross-substrate framing. Field synthesis, experiment design, the dual-null harness, the cross-mind translator and steering-transfer code, the Atlas construction sketch, the disjoint-corpus protocol, analysis and drafting were carried out with frontier models (Claude, Anthropic) under the lab's direction. The lab is responsible for the disconfirmation discipline; several entries in §7 correct the lab's own prior claims.

Reproducibility and relationship to the companion papers

The empirical core — the unsupervised translator, probe and steering transfer, the rank law, the null-completion audit, and the bounds — is reported with full methods in *"Portable Concept Geometry Enables Cross-Model Probe and Steering Transfer."* The synthesis, the dual-null epistemology, and the bounded "is meaning all you need?" thesis are in the companion position paper. This document organizes that body into a field: its definition, its central artifact (the Atlas), its laws, its research program, the decisive open experiment, and its honest seams. All experiments use open-weight models plus one custom-served model and custom tooling; the cross-mind translator is fit only on *unlabeled* parallel concepts; transfer results are reported against row-shuffled and random-orthogonal rotation nulls; behavioral results use held-out batteries; the cross-modal result was reproduced on independent hardware. The dual-null reference harness is `universal/seam_null.py`; the result documents grounding every number are under `universal/` and `steering/`.

Selected references

- Huh et al. (2024). The Platonic Representation Hypothesis.
- Kornblith et al. (2019). Similarity of Neural Network Representations Revisited.
- Bansal et al. (2021). Revisiting Model Stitching to Compare Neural Representations.
- Moschella et al. (2022). Relative Representations Enable Zero-Shot Latent Space Communication.
- Lenc & Vedaldi (2015). Understanding Image Representations by Measuring Their Equivariance and Equivalence.
- Turner et al. (2023). Activation Addition.
- Zou et al. (2023). Representation Engineering: A Top-Down Approach to AI Transparency.
- Ethayarajh (2019). How Contextual are Contextualized Word Representations?