Contents

Is Meaning All You Need? **1**
Abstract . . . 1
1. Introduction . . . 2
1b. Related work, and how we differ . . . 5
2. The dual-null discipline . . . 6
3. The negative result: the topology was a frame . . . 6
4. Walking upstream . . . 7
5. The shared concept geometry (the positive result) . . . 7
6. Three laws . . . 8
7. Bounds: the architecture-invariance gradient . . . 9
7b. The decisive test: disjoint corpora . . . 10
8. From reading to writing: cross-mind steering transfer . . . 10
9. A reference "testing mind" . . . 12
10. What we killed (disconfirmation record) . . . 13
11. Limitations (and the seam, now closed) . . . 14
12. Discussion: is meaning all you need? . . . 15
13. Conclusion . . . 16
Selected references . . . 17

Is Meaning All You Need?

A Family-Bounded, Steerable Concept Geometry Across Language Models — Up to Rotation

Honesty.Tools *Working draft — June 2026. The* ***synthesis / position*** *companion to the empirical paper.*

> **Two-paper note.** This program is split into two manuscripts. The **empirical core** — the unsupervised translator, probe and *steering* transfer, the rank law, the null-completion audit, and the bounds — is reported in full, with methods and controls, in the companion paper *"Portable Concept Geometry Enables Cross-Model Probe and Steering Transfer."* **This** paper is the synthesis: it carries the arc (the failed "topology of thought," the dual-null epistemology, the bounded thesis that *meaning* is the architecture-invariant the mechanism converges to, and the reflexive turn from observed frame to installed instrument), summarizes the empirical results, and cites the companion for the rigorous detail. The split follows reviewer guidance: separate knives cut cleaner — a reviewer who doubts the synthesis should not be able to downgrade the concept-geometry result, and vice versa.

Abstract

We set out to map the *topology of thought*: tens of thousands of model-generated "thoughts" passed through a scoring gate to build an atlas, on the assumption that the output of a frontier language model hides a rich latent structure waiting to be charted. It does not — or rather, most of the structure we first measured was put there by the instrument we used to measure it. Under an anisotropy-preserving null, the discrete-basin / modal reading of that atlas fails to reject: the topology was largely a *frame the observer supplied*, not a property of the mind.

What survives the *correspondence, rotation, and transfer* nulls — the nulls appropriate to a cross-model claim — is something different, and deeper. Walking the instruments upstream — from output embeddings, to the residual stream, to the mid-stack — the object that resists framing is not a topology of *thoughts* but a **conceptology**: a shared geometry of *meaning*. Across the tested text-model families and a 390× scale range, concept geometry is near-isometric **up to an orthogonal rotation** recoverable from *unlabeled* parallel concepts alone (representational-similarity z = 36; a learned translator places held-out concept centroids at top-1 = 1.000 in the clean within-family pipeline — 0.994 cross-family, 0.889 averaged over nine

models); across the cross-modal edge it degrades measurably. The invariance is **aggregate**: concept centroids and linear *directions* transfer near-losslessly between minds, while individual *instances* do not (cross-mind sentence retrieval ≈ 0.14). It is **bounded**: strongest within the transformer text family, partial across the modality edge (translator centroid top-1 falls from ~0.97 within-family, under the cross-modal-comparison fit — the canonical within-family ceiling is 1.000 — to 0.35 into a multimodal model).

Crucially, the shared geometry is not only *readable* but *writable*. A linear probe trained in one mind, rotated by the unlabeled translator, classifies in another at native fidelity; and — the new result here — a *causal steering vector* built in one mind, rotated by the same translator, **causally steers** another at native fidelity, provided the translator is high-rank. We use this to equip a small reference model with a **two-part epistemic instrument**: a shared *epistemic-honesty* steering axis (abstention, sycophancy-resistance, and refusal-of-the-unanswerable collapse to a single direction, pairwise cosine 0.93–1.00) plus an independent off-map *readout* (held-out AUROC 0.89). On a small held-out behavioral battery the steering axis raises epistemic honesty (mean 60→80%) while preserving correct answers (100%) and not over-refusing (0%); reading-off-map and steering-honest are **distinct axes** (cosine ≈ 0.2–0.4), so the mind both *detects* and *acts*, via two complementary directions. We report this as a proof of concept: the geometric (cosine, transfer) results are robust, the behavioral percentages rest on 10–20 probes per axis.

The discipline that demolished the false topology — a pair of nulls separating the observer's frame from the invariant core — is the same discipline that certified every transfer. We report what we *killed* as carefully as what we found. We also close the program's main seam: in one instrument (the residual stream) the within-mind modal frame *dissolves* against its own anisotropy-preserving null (z ≈ 0) while the cross-mind concept core *survives* that same strict null (RSA 0.93, z = +30) — the manufactured frame and the real core, separated by one discipline. **And we have now run the decisive test of the title itself:** two minds trained from scratch on **provably disjoint symbols** (non-overlapping character pools, audited intersection $\emptyset$) describing one authored world still share concept geometry at **cross-mind RSA 0.794, z = +35 vs the anisotropy null, 94 % of the same-corpus ceiling, at ≥1B and reproduced across two independent runs** — *meaning is not shared tokens*; the aggregate geometry survives even when the surface forms cannot overlap (instance recovery stays partial, as the law predicts). The bounded answer to the title is **not "yes" in general**: yes for *aggregate concept geometry* within the tested transformer family, *partial* across the modality edge, and *no* for *instances*. **The aggregate/instance law also sets a resolution ceiling** (added 2026-06-08): the *coarse* layer — concept centroids, directions, probes, steering vectors — is shared up-to-rotation, but the *fine* per-concept structure carried by a model's sparse-autoencoder dictionary does **not** transfer by that same rotation (matched cross-generation Qwen pair, fine SAE-feature transfer collapses to max-cosine median 0.14 against a 0.09 null, while the held-out *coarse* alignment stays intact at 0.69; §6, §7). The portable map is real but *low-resolution*. Or, as a line: **attention was all you needed to build it; meaning is what it built — a portable aggregate geometry, up to rotation, bounded by architecture and modality, sharpest for centroids and directions rather than instances, and coarse rather than fine.**

---

1. Introduction

The 2017 sequence-transduction result named a *mechanism* — attention — and showed it sufficient to replace recurrence and convolution. This paper is about what that mechanism, and its successors, *converge to*. Our central finding is that the specific mechanism is, within limits, **not the load-bearing thing**: different architectures, trained on overlapping web-scale data, build the *same geometry of meaning*, recoverable from one to another by a single orthogonal rotation. The engine was interchangeable; the meaning is the invariant.

We did not begin there. We began trying to chart a "topology of thought" and discovered, through a disciplined pair of null models, that the topology was mostly an artifact of measurement. The genuine, null-surviving structure lay upstream, at the level of concepts rather than thoughts, and it was shared across minds. This paper traces that arc end to end and adds the result that turns it from an observation into a capability: the shared geometry can be *steered*, not just read.

Contributions.

1. **A negative result, honestly bounded.** The discrete-basin "topology" of a frontier model's output atlas does not survive an anisotropy-preserving null (Section 3). We are careful about what this does and does not show: against this null we *fail to reject*; the test cannot distinguish "no structure" from "weak overlapping structure." What dissolves is the *observer-supplied frame*, not all structure.
2. **A shared concept geometry (the positive result).** Mid-stack representations of the same concepts in different model families are isometric up to rotation; the rotation is learnable from *unlabeled* parallel concepts; held-out concept centroids translate at top-1 = 1.000, with representational-similarity z = 36 (Section 5).
3. **Three laws.** The invariance is *up to rotation* (no shared coordinate system, only shared relational geometry) and *aggregate-sharp / instance-ambiguous* (centroids and directions transfer; instances do not) (Section 6). The aggregate/instance law extends onto a **resolution axis**: the coarse/aggregate layer is shared up-to-rotation, but *fine per-concept* structure (a model's SAE dictionary) is model-idiosyncratic and does not transfer by the same rotation (Section 6, Section 7).
4. **Bounds and a gradient.** The invariance holds across a 390× parameter range and cross-family, and degrades — measurably — across the modality edge (Section 7).
5. **From reading to writing.** A causal steering vector, rotated by the unlabeled translator, steers a second mind at native fidelity; reading needs a low-rank translator, steering a high-rank one (Section 8).
6. **A reference "testing mind."** Epistemic honesty is a single steerable axis; a two-part instrument (a steering direction + an independent off-map readout) equips a small model with calibrated epistemic behavior on a held-out battery — a proof of concept at small behavioral N (Section 9).
7. **A disconfirmation record.** We log the hypotheses we killed, including our own (Section 10).

Apparatus. We study open-weight models (Qwen2.5 and Qwen3 at 0.6B–235B, Llama-3.2/3.1 at 1B–8B, and the multimodal Gemma-4-12B) and one custom-served mid-size model. All instruments — the scoring "lens," the dual-null harness, the cross-mind translator, the steering-transfer harness, and the testing-mind stack — are custom-built by the authors. Code and configurations accompany this draft (Section: Reproducibility).

Vocabulary (operational). To keep the prose honest we fix terms to procedures:

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

Claim-status (read this first). The paper is rich; this table is the map of what we are and are not asserting.

claim	status	main evidence	main caveat
discrete output "thought-basins"	**unsupported**	anisotropy null: fail-to-reject	weak modes may be undetectable by this null
shared concept geometry, Qwen/Llama text models	**strong (coarse only)**	RSA 0.805/z=36; centroid translator top-1=1.000	parallel-concept corpus only; needs bootstraps; *coarse-resolution* — fine per-concept SAE structure does not transfer (see resolution row)
geometry survives provably disjoint symbols	**strong**	from-scratch pair (≥1B), audited symbol intersection ∅: cross-mind RSA 0.794 (z = +35.1), 94 % of the same-corpus ceiling (0.8485); reproduced to ~7 sig figs on a same-seed re-run on independent hardware (determinism check, not a fresh-init replication)	one authored world, one model pair; instance recovery stays partial (§7b)
fine per-concept structure transfers up-to-rotation	**no**	matched Qwen pair: SAE-feature transfer median 0.14 (std-R) vs 0.09 null, 0% land within cosine 0.5; same-model SAEs match at 0.795 (instrument valid)	the shared layer is low-resolution; fine resolution is usable only on its own model
architecture invariance *in general*	**bounded**	Gemma gradient	centroid transfer drops to 0.35 cross-modal
crosses to a closed frontier model (output-only)	**no, for the geometry**	RSA 0.02, translation at chance	the portable instruments need *residual access*; "any mind you can open"

claim	status	main evidence	main caveat
off-map *signal* crosses the access boundary	**yes — a learned output-logprob probe**	logprob-trajectory probe AUROC **0.96–0.98** cross-validated on the original held-out battery, haircutting to **0.938 pooled** on a pre-registered split (a registered FAIL of the locked 0.95 / 10%-FP bar). A subsequent **serving-matched re-fit** — the probes we actually *deploy* (2026-06-07) — is more conservative: per-model **≈0.89–0.92** (gpt-4o-mini **0.915**, gpt-4o **0.893**, gpt-3.5 **0.905** — its own token-entropy is below chance there — pooled **0.899**), and the newest reasoning model gpt-5.1 exposes only weak logprobs (**0.748**); ablation-verified vs length. Real-entity false-positive rate of the deployed re-fit: ~1–3%.	N=80/model on the original battery, fabricated-entity-scoped; we deploy and certify **per-model on the conservative re-fit**, not pooled or the original headline; logprob exposure is **multi-surface, not universal** (Anthropic/Google/Meta expose none); the instrument that crosses is a learned read of output logprobs
instance-level translation	**negative/weak**	retrieval 0.139	could improve with nonlinear maps
probe (reading) transfer	**strong**	0.947 vs 0.911 native, nulls 0.495	concept-pair benchmark
steering (writing) transfer	**promising PoC**	high-rank port reaches native potency	one main source–target pair; small behavioral N
testing mind	**promising PoC**	60→80% held-out honesty; readout AUROC 0.89	10–20 probes per axis
human-meaning bridge	**speculative**	multilingual concepts as proxy	no direct human semantic data

1b. Related work, and how we differ

Our result sits at the intersection of four lines of work; we state the delta against each, because the nearest neighbors are close.

Representational convergence / the Platonic Representation Hypothesis (Huh et al., 2024; and the representational-alignment literature, e.g. Kornblith et al. CKA, 2019; Bansal et al., 2021). This line shows that as models scale and improve, their representations *converge* in similarity, plausibly toward shared statistics of the world. We take this as our starting premise and sharpen it into operations: we fit an *explicit unsupervised orthogonal translator* and quantify the correspondence (held-out centroid top-1 = 1.000), we

draw the boundary the convergence claim leaves open — *aggregate vs. instance* and *up-to-rotation vs. shared-space* — and we map *where the convergence frays* (the cross-modal gradient). Convergence is the *that*; we give the *what-transfers, how-far,* and *whether-you-can-write-through-it.*

Relative representations and model stitching (Moschella et al., 2022; Lenc & Vedaldi, 2015). Relative representations encode each point by similarities to anchors, yielding a rotation/rescaling-invariant code for zero-shot *stitching* of read-out heads. Ours differs in three ways: (i) we use an explicit *orthogonal* map on concept centroids and show it is *near-isometric* (not merely a re-encoding that happens to stitch); (ii) we extend from stitching *readouts* to transferring *causal interventions* (steering), which prior stitching work does not address; (iii) we establish the *rank law* — reading transfers at low rank, steering needs high rank — which is new and mechanistically specific.

Activation steering / representation engineering (Turner et al., ActAdd; Rimsky et al., Contrastive Activation Addition, 2023–24; Zou et al., Representation Engineering, 2023). This line builds and applies steering vectors *within a single model.* Our contribution is to *port* a steering vector *across* models through the unsupervised translator: a vector built in one mind causally steers another at native fidelity. Cross-model steering transfer, to our knowledge, has not been shown.

Anisotropy and the geometry of embeddings (Ethayarajh, 2019; Gao et al., 2019). The "cone effect" — that representations occupy a narrow anisotropic cone — is usually treated as a nuisance. We turn it into an *instrument*: an anisotropy-*preserving* Gaussian null that subtracts exactly the structure a cone could manufacture, which is what lets us separate the observer's frame from the invariant core (and what kills our own original "topology").

(Selected references are collected at the end of the paper.)

2. The dual-null discipline

Every claim in this paper is adjudicated by two null models, because the failure mode we are guarding against — an instrument inventing the very structure it reports — is subtle and self-flattering.

- **Column-shuffle null.** Permute features independently across items, destroying cross-feature structure while preserving per-feature marginals. Catches structure that is real covariance vs. coincidence.
- **Anisotropy-preserving (Gaussian) null.** Sample from a Gaussian matched to the *full covariance* of the representation — i.e., preserve the anisotropy (the dominant directions, the "shape" of the cloud) while destroying any genuine clustering. Catches structure that is *real grouping* vs. *anisotropy alone.*

The second null is the strict one, and it is the one that demolished our first result. An important self-correction (forced by adversarial review of our own draft): against the anisotropy null we *fail to reject*, which is weaker than "the structure dissolves." A power control showed that genuinely-but-weakly separated modes give the same negative against this null, so the test cannot separate "no modes" from "overlapping modes." We therefore report the dual-null not as a structure-detector but as a **frame / core separator**: it strips away structure an observer could have manufactured and leaves only what an observer could not.

For transfer claims (Sections 5, 8) the analog of the second null is a **rotation null**: replace the learned orthogonal translator R with (i) a row-shuffled R, and (ii) a freshly sampled random-orthogonal R. A transfer that survives both is carried by the *concept-aligned* rotation specifically, not by any high-dimensional map.

---

3. The negative result: the topology was a frame

The origin atlas was built by passing model-generated short texts through a scoring gate and embedding the survivors, then reading off apparent basins / modes — a "topology of thought." Three nulls later, the reading is not supported:

- Against the column-shuffle null, weak modal structure appears ($z \approx +11$ at the deepest lens).

- Against the anisotropy-preserving Gaussian null, that signal **fails to reject** ($z \approx 0$): the apparent basins are consistent with the anisotropy of the embedding cloud plus overlap, with no additional grouping.
- The cloud is better described as a **continuous anisotropic manifold** than as discrete basins.

The honest statement is not "there is no structure in model output." It is: *the discrete-basin topology was a frame the observer (our scoring-and-embedding instrument) supplied.* This is the first and most important methodological lesson of the arc — and the reason the rest of the paper measures *upstream of the output*, where the instrument's fingerprints are smaller.

---

4. Walking upstream

We moved the measurement from the model's *output* to its *internals*, in three steps: (1) sentence-embedder representations of generated text; (2) the model's own residual stream at the last token; (3) the residual stream **mean-pooled at mid-stack**. At each step the observer's contribution shrinks and a different object comes into focus — one organized by *concept* rather than by *thought*, and, as the next section shows, *shared between minds*.

A practical lesson recurs here and matters for everything downstream: **the concept-rich depth is architecture-dependent.** Llama and Qwen peak mid-stack (~0.5–0.6 of depth); the multimodal Gemma-4 peaks *early* (~0.21 of depth) and decays monotonically thereafter. Extracting every model at a single fixed depth is a near-worst-case choice for some architectures; each mind must be read at *its own* concept peak.

---

5. The shared concept geometry (the positive result)

We represent each of N concepts by its mean-pooled mid-stack vector in each model, language-erased so that the representation is a function of *meaning*, not surface language. Three measurements, all dual-null controlled:

(a) Representational similarity. The pairwise concept-similarity matrices of two different families (Qwen2.5-3B and Llama-3.2-3B) agree at **RSA = 0.805 (z = 36)** against a meaning-shuffle null — at or above the split-half ceiling of either model alone. Two architectures build the *same relational concept geometry*.

(b) A learned translator. Fit PCA followed by orthogonal Procrustes on the concept *centroids* of a **training** split of concepts (420 concepts), using no labels beyond "these are the same concept in both models." On a **held-out** split (180 concepts), the translator places each concept's centroid at **top-1 = 1.000** (chance 0.0056). With a clean instruction-tuned teacher this reproduces at **1.000 within-family / 0.994 cross-family**. There is a working, generalizing map between two minds, fit without supervision.

(c) Probe (direction) transfer. Train a linear probe to separate two concepts in model A; rotate its weight vector by R into model B; test in B. Across 300 held-out concept pairs the **rotated probe classifies at 0.947**, matching the **native-B ceiling of 0.911**, while the **shuffled-R null sits at 0.495** and 100% of pairs exceed 0.70. A *direction* trained in one mind works in another after the unlabeled rotation. The same holds for a *specific* instrument — a grounded-vs-fabricated-entity ("familiarity") probe transfers at **AUROC 1.000** cross-family against a clean 0.500 null.

Together: the shared object is not a vague "similarity" but an operational geometry — you can build an instrument in one mind, carry it by an unsupervised rotation, and run it in another.

---

6. Three laws

Law 1 — up to rotation. There is *no shared coordinate system*. Each mind has its own frame; what is shared is the relational geometry, recoverable by an *orthogonal* map. "Universal mentalese" overstates it: the right picture is a relational invariant with per-mind coordinates, not a common space all minds write into.

Law 2 — aggregate-sharp, instance-ambiguous. What transfers is *aggregate* objects — centroids and directions. What does not is *instances*:

object	cross-mind transfer	interpretation
concept centroid	top-1 = 1.000 (clean pipeline; 0.994 cross-family, 0.889 over 9 models)	the *location* of a meaning transfers
linear direction / probe	0.95 ($\geq$ native 0.911)	an *instrument* transfers
single instance (retrieval)	**0.139**	a *particular thought* does **not** transfer

You can locate *what a concept is*, relationally, in any mind; you cannot reconstruct *a particular thought* of one mind in another. This is the boundary we call **govern, don't generate**: the universal structure lives at the level of *averaged meaning*, which is why "topology of thought" was the wrong object and "conceptology" is the right one. (A capability corollary we initially over-claimed and then killed — see §10, item D — is that a "second mind's view" of a concept boundary is a mildly *better* estimator of a direction than either mind alone; real, Wilcoxon-significant, but **size-independent** and hence a data-ensemble effect, not capability uplift.)

Govern-not-generate, now measured: the field reads STATE, not truth (added 2026-06-08). The aggregate layer's product consequence is sharper than "you cannot reconstruct an instance." The residual field is the purest signal of a model's coarse *epistemic state* — whether an input is off its distribution, and how familiar it is — and it is *not* a signal of *truth*. We quantified both edges of that boundary on the deployed off-map reading monitor (Qwen2.5-3B-Instruct, mid-stack). (i) On confident **on-manifold fabrication** — a model asserting a falsehood about something it has charted — the field is at **chance (AUROC 0.587)**: a fluent fabrication stays *on* the manifold, so the geometry cannot see it. Geometry charts **identity, not truth**. (ii) On the *cleanest* knowledge-state test — obscure-but-real entities the model either does or does not know, with subword-rarity and word-count held *flat by construction* so the probe cannot read letters — the field reads knowledge state at **AUROC 0.71–0.73** (powered re-run on weaker Instruct models with an adequate UNKNOWN class, n_unknown 19–38; de-confounded, M1_matched 1.000). That is **real and significantly above chance, but below the 0.75 bar** an always-on auto-gate would need. Within a register-matched battery, the same instrument separates **fabricated / off-distribution** inputs from grounded ones *perfectly* even lexicality-matched (AUROC 1.000, off-map recall 1.000). The match matters: measured across registers in deployment (2026-06-12), the off-distribution axis reads register distance — news-style prose scores 0.00–0.09 while short factual questions, real or fabricated alike, score 0.32–0.89 — so on question-register inputs the discriminating gauge is the independent familiarity reading, not this axis. The axis is an off-distribution-INPUT detector, not a real-vs-fabricated classifier. The reading is therefore graded by *resolution*, exactly as the law predicts: the field reads **coarse off-distribution / register distance** strongly (off-the-map vs on-the-map; the separately-read familiarity axis is the gauge that discriminates real from fabricated subjects) and **fine knowledge state** only *moderately* (0.71–0.73), and is *blind* to truth on the manifold (0.587). It is a strong off-distribution flag and a *soft* knowledge-gap advisory — not a fine truth oracle. (`universal/hook1_hardening/`.) In deployment the soft knowledge-gap signal is *escalated, not trusted alone*: the Grounded Atlas — an external, offline existence gate — separates obscure-but-real from fabricated entities at **AUROC 0.948** — where the model's own familiarity signal manages only **0.70** — recovering on the fine knowledge-state question precisely what the on-device geometry cannot resolve, and handing the deeper claim to a cross-family verification spectrum (companion: the Aperture honesty layer).

Law 3 — coarse-sharp, fine-idiosyncratic (the resolution axis; added 2026-06-08). The same law holds along a third axis — *resolution*. The aggregate objects of Law 2 (centroids, directions, probes) are all *coarse*: each averages over many instances. We asked whether the *fine* per-concept structure — the individual monosemantic features a model learns — is also shared up-to-rotation. It is not. Using Qwen's official residual-stream sparse autoencoders (Qwen-Scope, arXiv 2605.11887), we took a *matched-dimension* cross-generation pair (Qwen3-1.7B ↔ Qwen3.5-2B, both 2048-dim with 32K-feature SAEs, a mild 1.2× capability gap), fit a **full-dimensional** orthogonal R on 6000 paired FLORES sentences (deliberately *not* a PCA-100 R — that would erase the very fine structure under test), and asked where each teacher SAE feature lands in the student's dictionary:

object	cross-mind transfer (matched Qwen pair)	interpretation
coarse sentence alignment (held-out, R)	0.689	the *coarse* layer transfers, as Laws 1–2 say
fine SAE feature (max-cosine, std-R)	**0.141** (raw-R 0.217; null 0.09)	a *fine concept* does **not** transfer by the same rotation
fraction of features landing within cosine 0.5 of a real student feature	**0.0%** (raw-R 2.9%); within 0.7: 0.0% / 0.1%	almost nothing crosses cleanly

The transfer **decays monotonically with feature fineness** (the fraction of a feature's mass lying outside the coarse-concept span): fineness↔transfer correlation **−0.72** (raw) / −0.59 (std), the coarsest bin transferring at median 0.38 and the finest at 0.13 ≈ the 0.09 null. The decay *is* the mechanism: the finer the concept, the less the rotation carries it. Two controls close the obvious escapes. *(i) Not a bad rotation:* per-dim standardizing R — which **improves** the held-out coarse alignment (0.689 vs 0.672) by killing a rogue-dim confound — makes fine-feature transfer *worse* (0.141 vs 0.217), so the failure is the geometry, not a poorly-estimated map. *(ii) Not an SAE-basis artifact:* two independently-trained SAEs on the *same* model and layer match feature-to-feature at **median 0.795** (88.5% within 0.5, 64.5% within 0.7, null 0.09), so SAE features are substantially canonical and feature-matching is a valid instrument — the cross-model collapse to 0.14 is a genuine geometry statement. (Phase 1a corroborates from the static side: at matched dictionary dimensions a 17× larger same-generation model packs its concepts *modestly* more separably — mutual coherence 0.351→0.319 — so frontier models *do* hold higher-resolution maps; that extra resolution simply is **not rotation-portable**.)

The product reading is sharp: **transfer the coarse instruments — the honesty probe, the concept axes, the steering vector — "build once, deploy via an unlabeled R"; do *not* expect to port a frontier model's fine per-concept atlas onto a cheap deployed model by rotation. Fine resolution is usable only on the model that learned it.** The portable universal map is low-resolution by nature — the aggregate-sharp / instance-ambiguous law, read along the resolution axis.

---

7. Bounds: the architecture-invariance gradient

"Universal" is an empirical claim with a measurable edge. We map it along three axes.

Scale (within family). On the Qwen3 ladder from 32B down to 0.6B — a **53× gap, one tokenizer, zero confound** — centroid anchors are **0.94–0.98** (32B→0.6B = 0.978), probe transfer is *at or above* each student's native probe, and familiarity transfers at **1.000** across the whole ladder. Extending to the **235B-A22B flagship → 0.6B (a 390× gap)**: centroid anchor **1.000**, probe transfer beats native, and familiarity is **1.000**. The shared geometry holds across three orders of magnitude of scale.

Family. All of the above holds *cross-family* (Qwen ↔ Llama), with a clean instruction-tuned teacher giving the 1.000 / 0.994 centroid anchors. Instruction-tuning, not raw size, drives the *sharpness* of a model's pooled concept atlas.

Modality (the edge). The hardest case — a third family that is also multimodal (Gemma-4-12B; we read only its text residual stream) — is where "universal" starts to fray. Read at *Gemma's own* concept peak (early, ~0.21 depth), the **instrument** still transfers: honesty/familiarity probe transfer **0.99–1.00**, probe-direction recovery **91–97% of native**. But the **absolute centroid isometry is only partial** (centroid anchor 0.48–0.73; and a fresh translator fit for this paper places Qwen→Gemma held-out centroids at top-1 **0.35**, versus ~0.97 within the Qwen/Llama text family). The gradient is the result: **directions and instruments cross the modality edge near-intact; absolute concept positions cross only partially.** This both extends "universal" to a genuinely different architecture *and* draws its boundary.

(Independent-hardware reproduction: the cross-modal scan was re-run on separate hardware and a separate software stack and reproduced to ~three decimals, ruling out a single-environment artifact.)

7b. The decisive test: disjoint corpora

Every result above leaves one escape open: the models share web-scale training text, so the shared geometry could in principle be an artifact of shared *tokens* rather than shared *meaning*. We closed it with a from-scratch protocol. Two models, each at ≥1B parameters, were trained and read on corpora drawn from **provably disjoint character pools** describing one authored world — the symbol intersection audited empty ($\emptyset$) — so no surface form can occur in both minds. If the shared geometry were token statistics, it would have to die here. It does not: the cross-mind concept correspondence comes out at **RSA 0.794 ($z = +35.1$)** against the anisotropy-preserving null — **94 % of the same-corpus ceiling (0.8485)** and 98 % of native. A same-seed re-run on independent hardware reproduces the statistic to ~7 significant figures (a determinism/robustness check — the two runs share fixed corpus/training seeds, so it is not a fresh random-init replication). Instance-level recovery remains partial, exactly as the aggregate/instance law predicts — the low-resolution law survives the symbol barrier intact, sharp at the aggregate layer and ambiguous at the instance layer. Meaning, on this evidence, is not shared tokens: the geometry is built from the *world being described*, not from the symbols describing it. (Artifacts: `disjoint_corpus/out/xl1b_AB_battery.json` and `RESULTS_disjoint_xl1b.md`.)

---

8. From reading to writing: cross-mind steering transfer

Sections 5–7 establish that the shared geometry is *readable* across minds. The new result is that it is *writable*. We had two facts that had never been multiplied together: (i) a *reading* direction transfers A→B by the rotation R (§5c); (ii) a *contrastive activation-addition* (CAA) steering vector causally steers a property *within one mind*. We test the composition: **rotate a steering vector by R and apply it as a causal hook in another mind.**

Setup. Source A = Llama-3.2-3B-Instruct; target B = Qwen2.5-3B-Instruct (chosen because its native honesty override is independently established, giving a known ceiling). Build an abstention steering vector in A; project it into A's concept subspace; rotate by R^T; lift it back into B's *full residual stream*; apply via $h \leftarrow h + \alpha\|h\| \cdot \hat{u}$. A forced-confident system prompt makes B fabricate on fictional entities; we sweep α and measure abstention, fabrication, and preserved correctness on real questions. Controls: `native_full` (B's own vector), `native_roundtrip` (B's own vector through B's PCA and back, *no rotation* — isolates subspace-truncation cost), shuffled-R and random-orthogonal-R nulls, and `raw_unrotated`.

Result — direction transfers; potency is rank-gated. At the cheap low-rank (PCA-100) translator the *direction* transfers almost perfectly — the rotated Llama vector sits at **cosine 0.988** to Qwen's *own* honesty vector, the nulls at zero, the unrotated vector *anti-aligned* (−0.981) — yet it does not reproduce the clean low-α causal override. The decisive control is `native_roundtrip`: Qwen's own vector, merely passed through its own PCA-100 and back (cosine 0.992), is *also* causally degraded. So the bottleneck is the **dimensionality of the shared subspace, not the cross-family rotation.** A dimension sweep confirms it:

translator rank	`native_roundtrip` (abstain @α0.3/α0.5)	**`ported` (abstain @α0.3/α0.5)**	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, real answers 100% preserved, fabrication 0%), tracking `native_roundtrip` up the whole curve, while *both* rotation nulls stay dead at every rank — so the transfer is carried specifically by the concept-aligned R. This yields a clean law:

> **Reading transfers through a low-rank translator** (a classifier needs only the *sign of the projection* inside the shared concept subspace). **Steering needs a high-rank translator** (causal *magnitude* lives partly outside the top concept axes). At full rank, read *and* control both port across families at native fidelity, with zero labels in the target.

Generality and its limit. The transfer is *not* honesty-specific — a second, abstract stance (sycophancy-resistance) ports identically (cosine **0.94** to the target's own direction, unrotated anti-aligned, nulls dead). But a *concrete* behavior (a topic-steering "ocean" vector) ports only weakly (cosine **0.39** even at full rank; the unrotated vector is not cleanly killed). Law 2 reappears at the level of steering directions: **abstract stances sit along the shared geometry and port losslessly; concrete, lexical directions are model-idiosyncratic and port only partially.** Fortunately, the behaviors a safety layer wants — honesty, anti-sycophancy, calibrated refusal — are abstract stances.

Status and what this result needs (honest scoping). The steering-transfer result is a *strong proof of concept*, not yet an established law: the causal evidence rests on one main source–target pair (Llama-3.2-3B → Qwen2.5-3B), one depth, and one primary behavior (plus a clean second abstract stance and a weak concrete one). (Update 2026-06-12: the three-gate cold-deploy certificate has since passed on two further never-analyzed strangers with zero hand-tuning — a cross-family, cross-vendor, cross-objective contrastive embedding fine-tune of Mistral-7B (e5-mistral; shuffled-R 0.464, random-orthogonal 0.500, both dead — and it stays certified on the harder lexicality-controlled battery, shuffled-R 0.506) and Microsoft's phi-4 (ported 0.996 vs native 0.993; shuffled-R 0.194 on the word-count-matched battery, but a G2 marginal-fail at shuffled-R 0.58 $>$ 0.55 on the lexicality-controlled battery run the same day — so phi-4 certifies on the easier battery only, pending the larger lexicality battery); measured concept-peak depths now span 0.21–0.90 across four families, each found by the one mechanical scan.) Before this can carry a main-venue paper it needs the following grid, which we have not yet run:

experiment	what it rules out
reverse transfer (Qwen → Llama)	a directional accident
≥1 within-family pair	separates the family gap from translator mechanics
≥1 cross-scale pair (differing hidden dims)	tests high-rank transfer when dimensions differ
layer sweep	verifies depth 0.5 is not a hand-picked portal
prompt-template sweep	that steering is not tied to the forced-confident prompt
larger behavioral N + binomial CIs	makes abstain/fabricate rates statistically meaningful
sign controls (raw-unrotated is anti-aligned)	that sign, not magnitude, drives the null gap
rank / null-completion audit (below)	that full-rank success is determined, not a lucky SVD completion

Rank / null-completion audit. A reviewer's sharpest concern: `orthogonal_procrustes` is fit on ~420 concept centroids, so the rotation R is *pinned* only on the ≤420-dimensional row-space of the centroid matrix;

beyond that span its orthogonal completion is arbitrary. If the honesty steering vector lived *outside* that determined span, full-rank success would depend on an arbitrary completion (fragile). We audit this directly. The centroid matrix has effective rank **420** (as expected from 420 training concepts), so at PCA-2048 the rotation is pinned on 420 of 2048 dimensions. Yet the honesty steering vector lives almost entirely *inside* that determined span:

translator rank	rank(centroids)	honesty-vector energy **in determined span**	lift-back norm fidelity
100	100	100.0%	99.2%
512	420	99.8%	99.7%
1024	420	99.1%	99.9%
2048	420	**98.3%**	100.0%

At full rank the steering vector is **98.3% inside the rank-420 determined span**; the arbitrary completion carries 1.7% of its energy. Porting *only* the determined (in-span) component yields a vector at cosine **0.992** to the full ported vector and **0.988** to the target's own honesty vector, and it drives a strong causal override on its own (abstention 75% at α0.5 vs the full vector's 100%, real answers preserved). **So the transfer is *determined, not a lucky SVD completion*: the direction and the bulk of the causal effect live in the rank-420 span the data pins; the undetermined completion holds 1.7% of the energy and contributes only a modest residual potency (the last ~25 points of override).** The two columns also explain the dimension recovery of Section 8's main result: it is the *lift-back* norm fidelity (99.2% $\rightarrow$ 100.0%) that climbs with rank, restoring the steering vector's causal *magnitude*, while the *direction* was determined all along. This sharpens the abstract-vs-concrete law: **an abstract stance lies in the low-dimensional concept-centroid span where R is determined; a concrete lexical direction has energy outside it, in the undetermined completion — which is exactly why it ports only partially.**

---

9. A reference "testing mind"

We compose the read and write capabilities into one small model (Qwen2.5-3B) as a two-part epistemic instrument. We report this as a proof of concept: the *geometric* results below (the single-axis cosine, the read$\perp$steer separation) are robust and noise-free; the *behavioral* percentages use 10–20 probes per axis and carry real sampling noise (binomial CIs and larger batteries are required future work).

Epistemic honesty is one axis. We build three native steering vectors — abstention (don't fabricate unknown entities), sycophancy-resistance (don't cave to a confident false premise), and refusal (decline the genuinely unanswerable). Their **pairwise cosines are 0.93–1.00**: they are nearly the *same* direction. "Knowing what you don't know," across three superficially different failures, is approximately a **single direction** in the residual stream. A sum-of-three vector and a single unified contrast are *identical* (cosine 1.000).

The "is it just refusal?" control (run). "One axis" has a boring alternative: the three contrasts might all be smuggling in a generic *refuse / hedge* style. We tested it against a generic-refusal vector ("I can't help with that") and a verbosity vector, on Qwen2.5-3B.

- **Distinct from generic refusal — strongly.** `cos(epistemic, refusal) = -0.88`: the epistemic axis is the *opposite* of decline-to-engage. Behaviorally, pushed hard (α1.0) the epistemic vector *preserves answering* — obscure-but-real **known facts 40%**, answerable-sensitive **83%** — while the generic-refusal vector *collapses* — known **0%**, sensitive **33%** (both abstain on fabricated entities, 100%). At the operating point (α0.6) the epistemic vector is *calibrated*: fabricated-abstain 100%, known-facts 100%, sensitive-engaged 100%, zero flattening. A blanket-refusal vector cannot do this; selective abstention on the *unknown* while preserving the *known* is the signature of an epistemic, not a refusal, direction.
- **Not cleanly separable from verbosity/hedging — an honest limitation.** `cos(epistemic, verbosity) = 0.98`, even with a pure-length contrast. In this model, *length/elaboration and*

epistemic-caution are entangled in expression (confident answers are terse; uncertain ones elaborate), so a *cosine* cannot separate "an epistemic axis" from "a verbose-hedging axis." The discriminating evidence is therefore *behavioral calibration*, not geometry: a verbosity vector lengthens all answers, whereas the epistemic vector *selectively* abstains on the unknown and preserves the known. We report the entanglement as a real residual caveat — "epistemic honesty" and "expressing uncertainty at length" may be one axis here — while the calibration result stands.

(Pushed past its operating point the epistemic vector over-flattens — known-fact accuracy falls to 40% at $\alpha 1.0$ — so there is a calibration *window*, $\alpha \approx 0.3$–0.6, not an unbounded "more is better.")

The control half. A single always-on epistemic vector, at a tuned strength, lifts the model across all three axes while preserving correctness. On a **held-out** battery (fresh fabricated entities, fresh false premises, fresh unanswerables, fresh real questions — zero overlap with the tuning probes):

axis (held-out)	baseline	+ epistemic vector
abstains on fabricated entities	40%	60%
resists false premises	70%	90%
declines the unanswerable	70%	90%
epistemic honesty (mean)	**60%**	**80%**
real answers correct	100%	**100%**
over-refusal (wrongly declines answerable Q)	0%	**0%**

The tuned-set numbers were higher (≈97%); the **honest, held-out lift is 60→80%**, and — critically — the *safety properties generalize perfectly*: real-answer correctness stays 100% and over-refusal stays 0% on unseen probes. The limiter is the fabricated-entity axis (some invented entities are plausible enough that a single global strength cannot flip them).

The reading half — and a clean separation. Independently, a linear probe on the *input's* mid-stack activation flags off-distribution inputs at **AUROC 0.965** (5-fold) / **0.89 held-out** (train-on-tuning, test-on-fresh), with per-category recall 1.00 on fabricated and unanswerable inputs. The probe does not cheat on syntax: true- and false-premise probes are both leading-question form, and it still separates them. *What the readout reads (the §6 resolution boundary, made precise).* This figure is the **off-distribution / familiarity** AUROC — the coarse class the field reads strongly. The same instrument, hardened and lexicality-matched, separates fabricated-vs-grounded inputs at **AUROC 1.000** (register-matched battery; subword-rarity and word-count held flat) but reads the *finer* knowledge-state question — obscure-real entities the model does or does not know — at only **0.71–0.73** (powered, de-confounded; below the ≥0.75 always-on gate), and is at **chance (0.587)** on confident *on-manifold* fabrication. So the readout is a strong **off-distribution-input flag** and a **soft** knowledge-gap advisory, *not* a fine truth detector; the on-manifold tail and reasoning errors are routed to downstream instruments, not to this probe. The key structural finding holds regardless: **the off-map *reading* direction and the honest *steering* direction are distinct axes** (cosine ≈ 0.22–0.42). Detecting "this *input* is off-map" and making "the *output* honest" are *separate mechanisms*. The complete instrument is genuinely two-part: a **readout** that flags, and a **control** that acts.

Detect-then-steer. Letting the readout *gate* the control (steer hard only on flagged inputs) lifts the fabricated-entity axis (60→90%) and guarantees zero over-refusal by construction, at the same aggregate score — its ceiling is set by the readout's recall. This closes the loop: the mind detects that it is off the map, and acts.

---

10. What we killed (disconfirmation record)

The arc is honest only if the deaths are on the record. Selected entries:

#	Hypothesis (often ours)	Killed by	Corrected statement
A	The output atlas has a discrete-basin *topology of thought*.	anisotropy-preserving null (fail-to-reject) + power control	The basins were an observer-supplied *frame*; the cloud is a continuous anisotropic manifold.
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" means 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.
D	A frontier model's rotated probe beats a small model's own *because it knows where to look* (capability uplift), growing with teacher size.	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.	a global anisotropy correction is a rank-preserving affine $\rightarrow$ cannot change item-level AUROC ($\Delta = 0.000$)	The dual-null answers a *corpus*-level question, not an item-level one.
F	The interlingua is a strong cross-lingual *translator*.	loses to a purpose-built multilingual embedder (retrieval 0.80 vs 1.00)	It is an *invariant extractor* (certified language-blindness), not a bridge.
G	One global steering strength is optimal across axes.	held-out sweep (over-steers the already-strong axis)	Operating point is $\alpha \approx 0.3$; or detect-then-steer gating.
H	A frontier model's *higher-resolution* concept map transfers down to a cheap model by the same rotation R ("build the high-res instrument once, read it on the deployed model").	matched-dim SAE transfer (Qwen-Scope): fine-feature transfer 0.14 $\approx$ null vs coarse alignment 0.69; fineness$\leftrightarrow$transfer -0.72; per-dim-std R *improves* coarse yet *worsens* fine; same-model SAEs match at 0.795 (instrument valid).	**Only the coarse layer is rotation-portable.** Frontier models *do* hold modestly higher-resolution maps, but that fine resolution is model-idiosyncratic and usable only on its own model (§6, Law 3).

Items A, D, E, and H are self-refutations — predictions we made and then built the control that killed. That is the discipline working as intended.

---

11. Limitations (and the seam, now closed)

- **The instrument seam — closed.** The negative (frame dissolves) was originally measured with sentence-embedders-on-output; the positive (concept core survives) in the residual stream — different instruments, and the cross-mind correspondence had *never* been tested against the same anisotropy-

preserving Gaussian null that killed the topology. We fired that experiment, fusing both edges in **one instrument** (the residual stream):

  - *(negative, reproduced here)* a single mind's instance-cloud modal structure **dissolves** against its own anisotropy-matched Gaussian — k-means silhouette 0.160 vs 0.158 ± 0.005, $\mathbf{z = +0.4 \approx 0}$ — so the "topology" is manufactured anisotropy in the residual stream too, not only under the sentence-embedder;
  - *(positive, vs the same strict null)* the cross-mind concept correspondence (RSA 0.93) **survives** the **anisotropy-preserving Gaussian null at $\mathbf{z = +30}$** and the meaning-shuffle null at $z = +33$ (0.93 on the seam experiment's concept set and instrument; the canonical-frame measurement on the original battery reads 0.805 — both far beyond their nulls).

  Same instrument, same strict null: the observer's frame dissolves and the mind-shared core survives. The dual-null cleanly separates them, and the negative and positive are no longer measured with incommensurable instruments. *Residual caveats:* this *corroborates* the sentence-embedder negative in the residual stream rather than literally unifying the two instruments; and a participation-ratio probe of *mind-specific* anisotropy was weak (1.2 vs 1.3) and is not claimed — "observer-manufactured" is now firm for the modal frame, but the finer question of a mind's idiosyncratic *signature* remains open. `universal/seam_null.py`.
- **The shared map is low-resolution (added 2026-06-08).** Everything that transfers across minds is *coarse* — centroids, directions, probes, steering vectors, each an average over instances. The *fine* per-concept structure does not cross: on a matched cross-generation Qwen pair, SAE-feature transfer collapses to max-cosine median 0.14 (vs a 0.09 null, 0% landing within cosine 0.5) while the held-out coarse alignment stays at 0.69, with transfer decaying monotonically as a feature gets finer (fineness↔transfer −0.72). This is not a rotation-quality artifact (a fairer per-dim-standardized R *improves* coarse alignment yet makes fine transfer worse) nor an SAE-basis artifact (same-model SAEs match at 0.795). So the portable universal layer has a *resolution ceiling*: a frontier model's fine concept atlas is usable only on that model, not rotatable onto a cheaper deployed one. (No released SAE exists above 27B/35B-A3B, so the very largest models' fine resolution is untested; but the mechanism already fails at the cheapest matched-dimension test, so a larger gap can only diverge more.)
- **The human-language leg is a proxy.** Almost all cross-mind evidence is model↔model; the bridge to *human* semantic structure rides on multilingual parallel concepts as a stand-in, not a direct measurement.
- **Small behavioral N.** The testing-mind percentages use 10–20 probes per axis; the *cosine* results (one-axis, transfer) are robust and noise-free, but the behavioral rates carry real sampling noise.
- **Bounded models.** Open-weight transformers (0.6B–235B) plus one custom-served mid model; no closed frontier API model, no non-transformer architecture, no audio/vision *output* path.
- **The served-model tail.** Within-mind steering of a larger served model is weak (an *inverse-strength* effect: subtle properties are harder to steer in stronger models); whether a high-rank *ported* vector overcomes this is open.

---

12. Discussion: is meaning all you need?

The title is a question, and the honest answer is *bounded yes*. The mechanism named in 2017 — and the specific architectural choices that followed — are, within the family we tested, **not what is shared**. What is shared, across families, a 390× scale range, and partway across the modality edge, is the *geometry of meaning*: a relational concept structure, identical up to rotation, recoverable without supervision, and — this is the part that makes it more than a curiosity — *both readable and writable* across minds.

Two implications stand out. First, for **interpretability**: a *coarse* instrument (a probe, a detector, a steering vector — each an aggregate direction) need not be rebuilt per model. Build it once on a model you control, fit an unlabeled rotation, and deploy it on any other mind in the family — at native fidelity for reading, and, at full rank, for control. The unit of analysis shifts from the model to the **axis**. The boundary is *resolution*: this works for the coarse, aggregate instruments, *not* for a frontier model's *fine* per-concept atlas, which is model-idiosyncratic and does not survive the rotation that carries the coarse layer (§6, Law 3). You can port

the *axis*; you cannot port the *high-resolution dictionary*. Second, for **alignment**: a safety-relevant property like epistemic honesty turns out to be approximately *one direction*, shared across minds. "Knowing what you don't know" is not a diffuse capability to be trained in from scratch each time; it is a direction that can be *located, verified against nulls, and installed*. The reading half (off-map detection) and the writing half (honest steering) are separate axes, so a model can be made to *both* flag its own uncertainty *and* act on it.

The deepest turn of the arc is reflexive. We began by mistaking the *frame our instrument supplied* for the *territory*. The dual-null taught us to subtract the observer. And then — having learned that an observer can write structure into what it measures — we did so *deliberately and safely*: we manufactured a frame (the epistemic axis), verified it against the same nulls that had humbled us, and installed it into a mind. The manufactured frame went from the *bug* to the *tool*. The discipline that demolished the false topology is the discipline that certified the real control. Method became product: a model that "knows what it knows," built by a process that rigorously tracked what it did not.

Attention was all you needed to build it. Meaning is what it built — a portable aggregate geometry, the same up to rotation across the transformer minds we have looked into, fraying measurably at the modality edge, sharpest for centroids and directions rather than instances, and coarse rather than fine.

---

13. Conclusion

We set out to chart a topology of thought, found we had charted our own instrument, followed the instruments upstream, and arrived at a shared *conceptology*: a family-bounded geometry of meaning, near-isometric up to rotation, aggregate-sharp and instance-ambiguous, readable *and* writable across the minds we tested. The title's decisive test passed: two minds trained from scratch on provably disjoint symbols still share the geometry (RSA 0.794, $z = +35$, 94 % of the same-corpus ceiling) — meaning is not shared tokens. We demonstrated cross-mind *steering* transfer for the first time, established the low-rank/high-rank reading/writing law, and built a small reference model that knows what it knows via one shared epistemic axis plus an independent off-map readout. And we closed the program's seam: in one instrument, the same anisotropy-preserving null that dissolves a mind's own modal frame ($z \approx 0$) is survived by the cross-mind concept core ($z = +30$) — the manufactured and the real, cleanly separated. The next system is the same instrument installed on a larger and a cross-modal mind; the one question that remains is whether a *single* mind's idiosyncratic signature — not just the modal frame — is itself observer-manufactured.

Contributions

Prepared by Honesty.Tools. The lab set the conception and research direction: the instrument-relativity framing and the founding intuition to "start at the front, not the end." Experiment design and implementation, the dual-null harness, the cross-mind translator and steering-transfer code, 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 §10 correct the lab's own prior claims.

Reproducibility

All experiments use open-weight models plus one custom-served mid-size model and custom tooling (the lens / dual-null harness, the PCA+Procrustes translator, the CAA steering and steering-transfer harnesses, and the testing-mind stack). The cross-mind translator is fit only on *unlabeled* parallel concepts; all transfer results are reported against row-shuffled and random-orthogonal rotation nulls; behavioral steering results are reported on *held-out* probe batteries distinct from any tuning set. The cross-modal result was reproduced on independent hardware and a separate software stack.

A note on the collaboration

This paper was written by a human and a model who spent the work arguing about when the other was off the map — which is, not incidentally, the instrument the paper describes.

Selected references

- Bansal, Y., Nakkiran, P., & Barak, B. (2021). Revisiting Model Stitching to Compare Neural Representations.
- Ethayarajh, K. (2019). How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings.
- Huh, M., Cheung, B., Wang, T., & Isola, P. (2024). The Platonic Representation Hypothesis.
- Kornblith, S., Norouzi, M., Lee, H., & Hinton, G. (2019). Similarity of Neural Network Representations Revisited.
- Lenc, K., & Vedaldi, A. (2015). Understanding Image Representations by Measuring Their Equivariance and Equivalence.
- Moschella, L., Maiorca, V., Fumero, M., Norelli, A., Locatello, F., & Rodolà, E. (2022). Relative Representations Enable Zero-Shot Latent Space Communication.
- Rimsky, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., & Turner, A. M. (2024). Steering Llama 2 via Contrastive Activation Addition.
- Turner, A. M., Thiergart, L., Udell, D., Leech, G., Mini, U., & MacDiarmid, M. (2023). Activation Addition: Steering Language Models Without Optimization.
- Zou, A., Phan, L., Chen, S., Campbell, J., et al. (2023). Representation Engineering: A Top-Down Approach to AI Transparency.