A GÄRDENFORSIAN PERSPECTIVE ON THE CONCEPTUAL GEOMETRY OF LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Gärdenfors' conceptual spaces offer a geometric language for describing conceptual knowledge. We show that this cognitive theory can be approximately realized in the representation geometry of LLMs. Specifically, we propose the Gärdenforsian Representation Hypothesis (GRH), which organizes LLM representations into a hierarchy of quality dimensions, domains, properties, concepts, and prototypes. GRH provides a common geometry in which existing representation hypotheses can be understood as level-specific instantiations, with linear representations arising at the level of individual quality dimensions and lattice representations at the level of discretized property structure. Empirically, we progressively reconstruct this conceptual space from LLM representations using THINGSplus and Binder across Llama-3.1-8B, Gemma-7B, and Mistral-7B, and achieve consistent alignment with human judgments at each level of conceptual organization, providing evidence that LLM representations naturally admit a conceptual space in the sense of Gärdenfors.

1 INTRODUCTION

Concepts are basic units of human cognition. They organize semantic knowledge in ways that support categorization, abstraction, analogy, and reasoning (Goldstone et al., 2018; Murphy, 2002). Large language models (LLMs) (Achiam et al., 2023) exhibit many of these abilities, suggesting that a rich conceptual structure emerges during pretraining (Wu et al., 2023; Lin & Ng, 2022). Yet, it remains fundamentally unclear how this structure is organized in the geometry of LLM representations. Understanding this conceptual geometry in LLMs is crucial for interpretability (Bereska & Gavves, 2024), model steering (Han et al., 2024), and AI safety (Arditi et al., 2024; Ratnakar & Vats, 2026).

A natural line of work studies conceptual knowledge through the geometry of LLM representations. The dominant view is the *Linear Representation Hypothesis*, which represents semantic features, attributes, or concepts as linear directions or low-dimensional subspaces (Park et al., 2025; Lee et al., 2025; Gurnee & Tegmark, 2024). Recent work has extended this geometric perspective to increasingly higher levels of abstraction, from multidimensional feature representations (Modell et al., 2025) to categorical concepts (Park et al., 2025), concept algebra (Wang et al., 2023), and formal concept lattices (Xiong, 2026). Despite these advances, existing approaches largely use the same geometric vocabulary across different conceptual primitives (Rosch & Lloyd, 2024). Objects, quality dimensions, properties, domains, and concepts are not explicitly distinguished by their conceptual roles. It therefore remains unclear how these primitives relate geometrically across levels of abstraction and how they combine into a coherent conceptual structure.

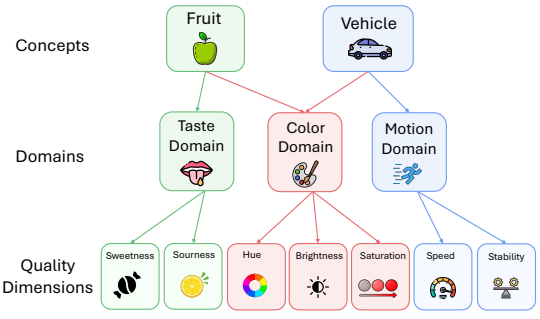


Figure 1: **Concepts, domains, and quality dimensions in Gärdenfors' conceptual spaces.** Concepts are organized by subsets of domains, each of which is composed of a collection of quality dimensions.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Notably, Gärdenfors' *Conceptual Spaces* (Gärdenfors, 2004; 2014) provide exactly such a geometric theory. Instead of representing all semantic information with a single geometric object, they organize quality dimensions, domains, properties, concepts, and prototypes into a coherent conceptual structure, where higher-level primitives are built upon lower-level ones. For example, as illustrated in Fig. 1, quality dimensions *hue*, *brightness*, and *saturation* constitute the *color* domain, while the high-level concept *fruit* is defined only over its relevant domains, namely *color* and *taste*. Despite its broad influence on cognitive science and knowledge representation, conceptual-space theory has rarely been considered in the context of LLM representation geometry. Recent work has shown that individual conceptual-space dimensions can be recovered from LLM representations (Kumar et al., 2024; 2025), suggesting that LLMs capture some of the fundamental building blocks of conceptual spaces. Conceptual-space theory, however, is fundamentally concerned not with individual quality dimensions, but with how they organize into higher-level conceptual structure. This raises a natural question: *Do LLM representations admit such a Gärdenfors-style conceptual space?*

We propose the *Gärdenforsian Representation Hypothesis* (GRH), which establishes a correspondence between the conceptual primitives of Gärdenfors' conceptual spaces and the geometric primitives of LLM representations. Specifically, GRH realizes (1) **quality dimensions as linear directions**, where an object's projection onto a direction defines its strength on the corresponding quality dimension; (2) **domains as linear subspaces** spanned by their constituent quality dimensions; (3) **properties as convex regions** within individual domains; (4) **concepts as convex regions** spanning one or more domains; and (5) **prototypes as the geometric centers** of regions that gives rise to graded typicality.

Our theoretical analysis shows that GRH provides a common geometric foundation for existing representation hypotheses: the Linear Representation Hypothesis arises as a consequence of prototype-based conceptual geometry, while the Lattice Representation Hypothesis (Xiong, 2026) emerges as a discrete abstraction of the same continuous conceptual space. Empirically, we progressively reconstruct this conceptual space from LLM representations using THINGSplus and Binder, and evaluate it across Llama-3.1-8B, Gemma-7B, and Mistral-7B. Across datasets and models, each level of conceptual organization, from quality dimensions and domains to properties, concepts, and prototypes, shows consistent alignment with human judgments. These results suggest that seemingly disparate geometric structures in LLMs are manifestations of a single Gärdenfors' conceptual space.

2 Preliminaries

2.1 Representation Hypotheses in LLMs

The *Linear Representation Hypothesis* (LRH) (Park et al., 2024) states that semantic attributes admit approximately linear representations in LLM embedding spaces. Each attribute is associated with a direction $d \in \mathbb{R}^d$, and the value of an object x along that attribute is given by the scalar projection $z_d(x) = d^\top v_x$. Moving along d selectively changes the corresponding semantic attribute while preserving causally unrelated attributes. Building on LRH, the *Lattice Representation Hypothesis* (Xiong, 2026) extends linear attribute representations to symbolic concept organization.

Definition 1 (Lattice Representation of Concepts (LRH-Lattice))**.** *Let $\mathcal{D} = \{d_1, \ldots, d_m\}$ be a collection of linearly represented attributes. Each attribute induces a binary incidence relation $x\, I\, d_i \iff d_i^\top v_x \geq \tau_i$, where τ_i is an attribute-specific threshold. This defines a formal context $(\mathcal{X}, \mathcal{D}, I)$. For any closed attribute set $B = B'' \subseteq \mathcal{D}$, the pair (B', B) forms a formal concept, where B' is the set of objects satisfying all attributes in B. The resulting formal concepts form a lattice whose partial order captures semantic subsumption.*

Thus, LRH and LRH-Lattice characterize two levels of LLM representation geometry: linear attribute representations and symbolic concept organization. They do not, however, distinguish the conceptual primitives that lie between these levels, nor explain how quality dimensions, domains, properties, concepts, and prototypes are geometrically related within a common conceptual organization.

2.2 Gärdenfors' Conceptual Space Theory

A *conceptual space* (Gärdenfors, 2000; Gärdenfors, 2004) is a geometric representation space defined over a collection of *quality dimensions* $\mathcal{Q} = \{q_1, \ldots, q_m\}$. Each quality dimension q_i represents an aspect along which entities can be compared, such as brightness, sweetness, or size, and is associated

with a value space $\mathcal{X}_i$. An entity x is represented by a point $z(x) = \big(z_1(x), \ldots, z_m(x)\big) \in \mathcal{X}$, where $\mathcal{X} = \prod_{i=1}^{m} \mathcal{X}_i$ is the conceptual space and $z_i(x)$ denotes the value of x on quality dimension q_i.

Quality dimensions may be *integral* or *separable*. Integral dimensions cannot be specified independently, whereas separable dimensions can vary independently. A *domain* $\delta \subseteq \mathcal{Q}$ groups mutually integral quality dimensions and is separable from dimensions outside the domain. Let $\Delta = \{\delta_1, \ldots, \delta_K\}$ denote the collection of domains. Writing $\mathcal{X}_\delta = \prod_{q_i \in \delta} \mathcal{X}_i$, the conceptual space decomposes as $\mathcal{X} = \prod_{\delta \in \Delta} \mathcal{X}_\delta$. Each domain is equipped with a metric $d_\delta : \mathcal{X}_\delta \times \mathcal{X}_\delta \to \mathbb{R}_{\geq 0}$, which determines similarity within that domain. Similarity in the full conceptual space is obtained by combining distances across the domains relevant to the comparison.

Gärdenfors characterizes properties and concepts geometrically through two representation criteria.

Definition 2 (Property Region)**.** *A property P is represented by a convex region $R_P \subseteq \mathcal{X}_\delta$ within a single domain δ. A region $R \subseteq \mathcal{X}_\delta$ is convex if $x, y \in R$ implies $\lambda x + (1-\lambda)y \in R$ for every $\lambda \in [0, 1]$.*

Definition 3 (Concept Region)**.** *Let $\Delta_C \subseteq \Delta$ denote the domains relevant to concept C. A concept is represented by $\mathcal{G}(C) = \left(\{R_{C,\delta}\}_{\delta \in \Delta_C}, \Gamma_C\right)$, where $R_{C,\delta} \subseteq \mathcal{X}_\delta$ is the region associated with C in domain δ, and Γ_C captures dependencies among its domain-specific regions.*

The two definitions correspond to Gärdenfors' Criteria P and C, respectively: properties form convex regions within individual domains, while concepts combine regions across relevant domains.

Finally, each concept is associated with a prototype p_C, representing a geometrically central point of the concept. Typicality decreases with distance from the prototype under the relevant conceptual metric, so objects closer to p_C are predicted to be more representative instances of the concept.

3 A Gärdenforsian Perspective on the Geometry of LLMs

3.1 Gärdenforsian Representation Hypothesis

Let $\mathcal{X}$ denote a collection of objects, and let $\phi : \mathcal{X} \to \mathbb{R}^d$ be the representation map induced by the final layer of an LLM. For each object $x \in \mathcal{X}$, we write $v_x = \phi(x)$. The image of ϕ defines the representation geometry $\mathcal{H} = (\phi(\mathcal{X}), \langle \cdot, \cdot \rangle)$. A conceptual realization specifies how the conceptual primitives introduced in Section 2.2 are realized in this representation geometry.

Definition 4 (Conceptual Realization)**.** *A conceptual realization of $\mathcal{H}$ consists of (i) a collection of quality dimensions with linear realizations $\mathcal{D} = \{d_1, \ldots, d_m\} \subseteq \mathbb{R}^d$; (ii) a partition of the quality dimensions into domains $\Delta = \{\delta_1, \ldots, \delta_K\}$; (iii) a domain-specific metric d_δ for each domain; (iv) a collection of property regions $\{R_P\}$; (v) a collection of concept representations $\{\mathcal{G}(C)\}$; and (vi) a prototype p_C for every concept.*

For notational convenience, let $D = [d_1, \ldots, d_m] \in \mathbb{R}^{d \times m}$, so that the conceptual coordinates of x are $z(x) = D^\top v_x \in \mathbb{R}^m$. For a domain $\delta \in \Delta$, let I_δ index its quality dimensions and define $D_\delta = [d_i]_{i \in I_\delta}$. The domain is realized by $\mathcal{S}_\delta = \operatorname{span}\{d_i : i \in I_\delta\} \subseteq \mathbb{R}^d$, with coordinate space $\mathcal{Z}_\delta = \mathbb{R}^{|\delta|}$ and coordinates $z_\delta(x) = D_\delta^\top v_x \in \mathcal{Z}_\delta$. Thus, $\mathcal{S}_\delta$ lies in the LLM representation space, whereas domain-specific metrics and regions are defined in $\mathcal{Z}_\delta$.

Figure 2 illustrates how these geometric objects form a conceptual realization of the LLM representation geometry. We formalize this correspondence in the Gärdenforsian Representation Hypothesis.

Definition 5 (Gärdenforsian Representation Hypothesis)**.** *The representation geometry $\mathcal{H}$ satisfies the* Gärdenforsian Representation Hypothesis *if there exists a conceptual realization satisfying:*

- ***A1. Quality Dimensions.*** *Each quality dimension admits a linear realization in the representation geometry. The scalar projection $z_i(x) = d_i^\top v_x$ defines the coordinate of object x along that quality dimension.*
- ***A2. Domains.*** *Quality dimensions are partitioned into domains $\Delta = \{\delta_1, \ldots, \delta_K\}$. Each domain δ is realized in the representation space by $\mathcal{S}_\delta = \operatorname{span}\{d_i : i \in I_\delta\} \subseteq \mathbb{R}^d$ and induces a conceptual coordinate space $\mathcal{Z}_\delta = \mathbb{R}^{|\delta|}$ equipped with a metric d_δ.*

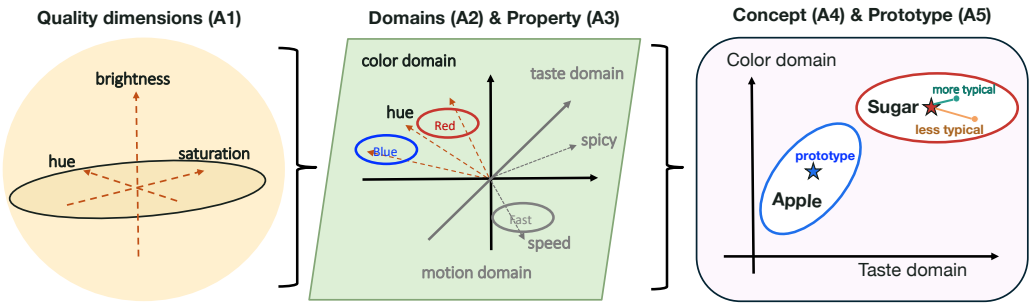


Figure 2: **Conceptual realization under GRH. A1:** quality dimensions are realized as linear coordinates; **A2:** related dimensions span a subspace for a particular domain; **A3:** properties correspond to convex regions within individual domains; **A4:** concepts are represented by regions across their relevant domains; and **A5:** prototypes capture the geometric centrality of concept regions.

A3. Properties. *Each property P is associated with a single domain δ and is represented by a convex region $R_P \subseteq \mathcal{Z}_\delta$.*

A4. Concepts. *Let $\Delta_C \subseteq \Delta$ denote the domains relevant to concept C. The concept is represented by $\mathcal{G}(C) = (\{R_{C,\delta}\}_{\delta \in \Delta_C}, \Gamma_C)$, where $R_{C,\delta} \subseteq \mathcal{Z}_\delta$ is the region associated with concept C in domain δ, and Γ_C captures correlations across domain-specific regions.*

A5. Prototypes. *Each concept C is associated with a prototype $p_C \in \mathcal{Z}_C$, where $\mathcal{Z}_C = \prod_{\delta \in \Delta_C} \mathcal{Z}_\delta$. The prototype represents a point of geometric centrality within the concept representation, and typicality decreases with distance from the prototype in $\mathcal{Z}_C$.*

A1–A5 jointly describe a layered conceptual organization in LLM representations. Quality dimensions provide graded semantic coordinates, domains organize related dimensions into structured subspaces, properties occupy convex regions within individual domains, and concepts combine such domain-specific structure across multiple domains. Prototypes then introduce graded category structure through geometric centrality.

3.2 Relations to Existing Representation Hypotheses

GRH extends existing representation hypotheses by placing their geometric objects within a common conceptual organization. In particular, linear representations arise at the level of individual quality dimensions, while lattice representations arise after discretizing continuous property geometry.

Linear representation as a special case. **A1** directly recovers the geometric structure assumed by the Linear Representation Hypothesis.

Proposition 1 (Linear Representation as a Special Case)**.** *Assume GRH and restrict attention to a single quality dimension d_i. Then the semantic value of an object x is represented by the linear coordinate $z_i(x) = d_i^\top v_x$. Hence, at the level of individual quality dimensions, GRH reduces to a linear representation of semantic attributes.*

Thus, LRH describes the first level of the conceptual geometry specified by GRH: individual graded semantic dimensions are linearly realized, while GRH further specifies how these dimensions organize into domains, properties, concepts, and prototypes.

Linear boundaries from prototype geometry. Under a shared conceptual metric, prototype geometry also induces linear concept boundaries.

Theorem 1 (Linear Boundaries from Prototype Geometry)**.** *Consider two concepts C_i and C_j represented in the same conceptual coordinate space $z = D^\top v$ and classified by nearest prototype under a shared quadratic metric $M \succeq 0$. Assume $DM(p_i - p_j) \neq 0$. Then their decision boundary in representation space is the hyperplane $n_{ij}^\top v = \tau_{ij}$, where $n_{ij} = 2DM(p_i - p_j)$ and $\tau_{ij} = p_i^\top M p_i - p_j^\top M p_j$.*

Thus, linear concept boundaries arise as a special case of prototype geometry when concepts share the same metric. More generally, GRH allows concept-specific domain weights and relevant domains.

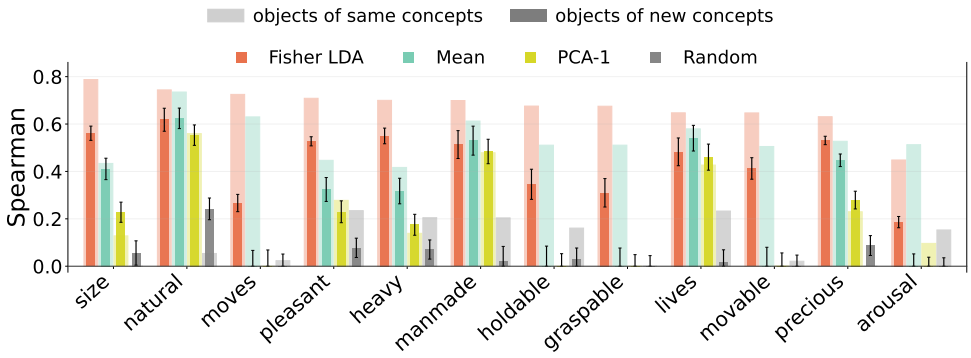


Figure 3: **A1: Recovered quality dimensions align with human judgments.** Linear coordinates recovered from LLM representations correlate with human ratings across the 12 THINGSplus quality dimensions and remain predictive on unseen concepts. Light bars show held-out objects from concepts represented during training. Dark bars show objects from unseen concepts.

When these induce different metrics, the corresponding prototype boundary is generally quadratic instead of linear.

Lattice representation as a discrete special case. The Lattice Representation Hypothesis is recovered when continuous property geometry is discretized into binary attributes.

Proposition 2 (Lattice Representation as a Special Case)**.** *Assume GRH and threshold each quality coordinate to define the binary incidence relation* $x\ I\ P_i \iff d_i^\top v_x \geq \tau_i$*. Then the formal context* $(\mathcal{X}, \mathcal{P}, I)$ *induces a formal concept lattice. For any closed set of properties* $B = B''$*, the pair* (B', B) *forms a formal concept, where* B' *consists of the objects satisfying all properties in* B*.*

Hence, LRH-Lattice can be viewed as a discrete, order-theoretic realization of the continuous conceptual geometry described by GRH.

4 Constructing Conceptual Spaces from LLM Representations

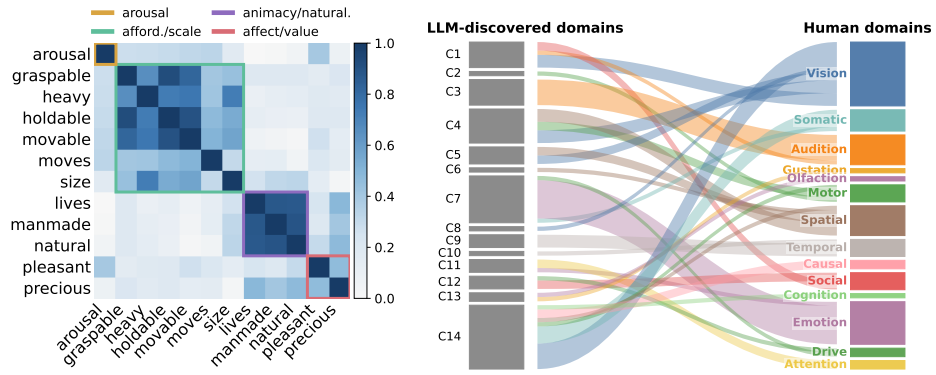
We evaluate to what extent the geometry of LLMs realizes a Gärdenforsian conceptual space. Following prior work on representation geometry (Park et al., 2025; Xiong, 2026), we focus on static object-level representations extracted from isolated lexical forms at the final hidden layer. Following GRH, we progressively reconstruct quality dimensions, domains, properties, concepts, and prototypes from the representation geometry of LLMs, where each stage builds on the previous one.

Datasets and Models. We use two public datasets with human judgments: 1) **THINGSplus** (Stoinski et al., 2024) provides human judgments of 12 quality dimensions together with concept membership and typicality annotations, supporting the evaluation of **A1** and **A3–A5**. Binder feature norms (Binder et al., 2016) provide human ratings for 65 semantic attributes organized into 14 human-annotated semantic domains, supporting the evaluation of **A1** and **A2**. Detailed dataset statistics are provided in Appendix C.1. We evaluate three decoder-only LLMs with similar size: Llama-3.1-8B (Dubey et al., 2024), Gemma-7B (Team et al., 2024), and Mistral-7B (Jiang et al., 2023). Unless otherwise specified, we report the results of Llama-3.1-8B using final-layer hidden representations, and report the robustness across all three backbone models in Section 4.6.

4.1 A1: Quality Dimensions

Construction. We first construct the representation geometry $\mathcal{H} = (\phi(\mathcal{X}), \langle \cdot, \cdot \rangle)$, which serves as the common space for recovering all conceptual primitives. Let $\mathcal{X}$ denote the evaluation objects. For each object $x \in \mathcal{X}$, we obtain its representation $v_x = \phi(x)$ by averaging the final-layer hidden representations over multiple lexical variants of the object name, following Xiong (2026).

To recover a quality dimension, let $y_q(x)$ denote the human rating of object x on quality dimension q. We estimate a linear direction from the extreme examples $G_q^+ = \{x : y_q(x) \geq Q_{75}(y_q)\}$, $G_q^- =$



(a) Affinity matrix exhibits block-like structure (b) Discovered domains vs human domains

Figure 4: **Recovery of domains from quality dimensions.** (a) Pairwise affinity matrix of the recovered quality directions on THINGSplus. The block structure reveals coherent domains. (b) Alignment between domains recovered from Binder and the 14 human-annotated domains.

$\{x : y_q(x) \leq Q_{25}(y_q)\}$, using a regularized Fisher discriminant, $d_q = (\tilde{\Sigma}_+ + \tilde{\Sigma}_- + \lambda I)^{-1}(\mu_+ - \mu_-)$, where $\mu_\pm$ and $\tilde{\Sigma}_\pm$ denote the class means and Ledoit–Wolf covariance estimates.

Results. Figure 3 shows that Fisher directions consistently recover human quality judgments and substantially outperform mean-difference, PCA-1, and random baselines. The recovered quality coordinates also generalize well across unseen categories, indicating that they capture generalizable qualities instead of category-specific distinctions. These results support **A1** by showing that human quality dimensions admit stable linear realizations in LLM representation geometry.

Takeaway 1. A1: Quality Dimensions

The results support **A1**: human quality dimensions admit stable linear realizations in the representation geometry of LLMs, and these representations generalize across concepts.

4.2 A2: Domain Organization and Metrics

Construction. A2 states that quality dimensions organize into semantic domains, each realized by a linear subspace equipped with its own similarity metric. We recover such domains from the quality directions identified in **A1**.

Let $D = [\hat{d}_1, \ldots, \hat{d}_K] \in \mathbb{R}^{d \times K}$ denote the matrix of recovered quality directions. We construct two affinity matrices. The first measures directional similarity, $A_{ij} = |\cos(\hat{d}_i, \hat{d}_j)|$, while the second measures the correlation of the induced quality coordinates, $B_{ij} = \mathrm{Corr}_{x \in \mathcal{X}}(\hat{d}_i^\top v_x, \hat{d}_j^\top v_x)$. Spectral clustering on either affinity yields a partition $c : \{1, \ldots, K\} \to \{1, \ldots, L\}$, where the recovered domain is represented by the subspace $S_\ell = \mathrm{span}\{\hat{d}_j : c(j) = \ell\}$. For each recovered domain, we define the corresponding domain metric by projecting object representations onto S_ℓ. Implementation details are provided in Appendix C.3.

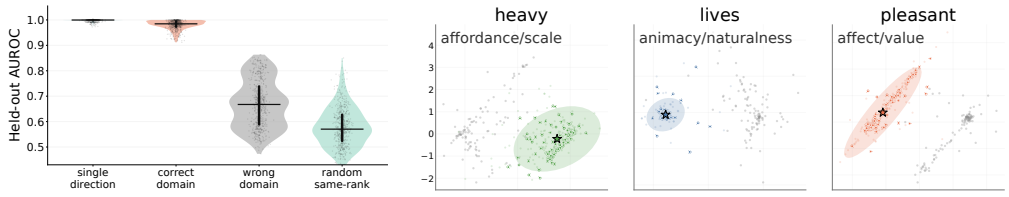
Table 1: Alignment between recovered domains and human-annotated domains in Binder.

Model	Aff.	ARI	NMI	Purity
Llama-3.1	A	0.324	0.660	0.631
	B	0.229	0.602	0.585
Mistral	A	0.284	0.635	0.600
	B	0.231	0.604	0.585
Gemma	A	**0.371**	**0.687**	**0.662**
	B	0.229	0.602	0.585

Results. Figure 4a shows that the recovered quality directions in THINGSplus organize into coherent semantic domains, including *animacy/naturalness*, *affordance/scale*, and *affect/value*. To evaluate whether these domains are consistent with human conceptual organization, we compare them against the 14 human-annotated Binder domains using adjusted Rand index (ARI), normalized mutual information (NMI), and purity. As shown in Fig. 4b(b) and Table 1, affinity derived from the geometry of recovered quality directions consistently produces better agreement with human domains than object-induced coupling. Finally, we evaluate the induced domain metrics on THINGS odd-one-out similarity judgments by comparing each recovered domain against 200 random subspaces of the

Table 2: **Recovered domains induce meaningful similarity metrics.** Domain-restricted metrics consistently predict THINGS odd-one-out similarity judgments better than random same-rank subspaces. Both the unscaled Euclidean metric and the shared diagonal metric (scaled) exhibit the same trend.

		Unscaled			Scaled		
Domain	Rank	Target	Random	Δ(%)	Target	Random	Δ(%)
animacy / naturalness	3	0.492	0.374	+31.6%	0.490	0.372	+31.7%
affordance / scale	6	0.429	0.388	+10.6%	0.431	0.385	+11.9%
affect / value	2	0.394	0.366	+7.7%	0.396	0.365	+8.5%
arousal	1	0.365	0.354	+3.1%	0.365	0.354	+3.1%



(a) (b)

Figure 5: **A3: Properties are realized as convex regions within individual domains.** (a) Quantitative comparison of property representations. (b) Visualization of three representative properties.

same dimension. Table 2 shows that recovered domains consistently outperform random subspaces, indicating that they capture behaviorally meaningful similarity structure.

Takeaway 2. A2: Domains

The results support **A2**: quality dimensions organize into domains that align with human judgments and induce meaningful similarity metrics.

4.3 A3: Properties as Convex Regions

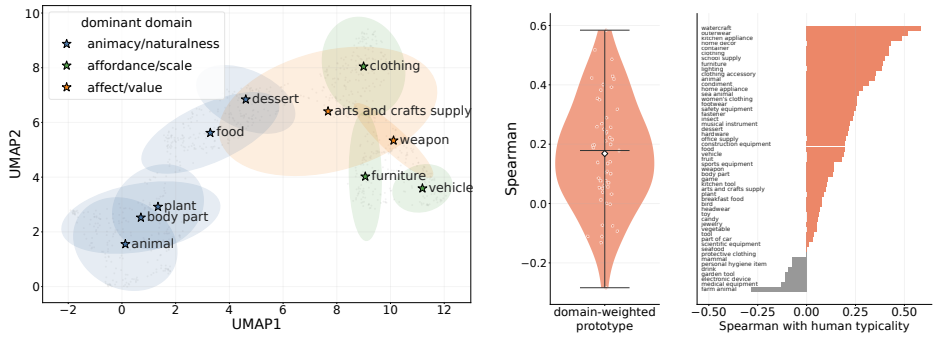
Construction. A3 states that properties are represented by convex regions within individual domains. We instantiate property regions using the semantic domains recovered in **A2**. For a quality direction $\hat{d}_j$, the induced quality coordinate is $q_j(x) = \hat{d}_j^\top v_x$. A scalar property such as *heavy* is naturally represented by a half-space $R_j(\tau) = \{x : q_j(x) \geq \tau\}$. More generally, GRH predicts that properties occupy convex regions within their corresponding domains. Our evaluation operationalizes properties using thresholded THINGSplus quality ratings, providing a controlled test of whether such graded properties also form convex regions within their recovered domains.

For each THINGSplus property, we estimate a convex region from the training positives in its recovered domain. Let $\mathcal{D}(j)$ denote the corresponding domain and $z_{\mathcal{D}(j)}(x)$ the domain coordinates of object x. The property center is estimated as $\mu_j = \frac{1}{|P^+_{j,\text{train}}|} \sum_{x \in P^+_{j,\text{train}}} z_{\mathcal{D}(j)}(x)$, and the property region is defined by

$$R_j = \left\{ x : (z_{\mathcal{D}(j)}(x) - \mu_j)^\top M_{\mathcal{D}(j)} (z_{\mathcal{D}(j)}(x) - \mu_j) \leq r_j^2 \right\}, \tag{1}$$

where $M_{\mathcal{D}(j)}$ is the shared domain metric from **A2**. We compare the recovered region against three alternatives: a single-direction half-space, an ellipsoid estimated in an incorrect domain, and an ellipsoid estimated in a random same-rank subspace. Implementation details are in Appendix C.4.

Results. Figure 5a shows that the recovered domain-specific regions achieve strong held-out property prediction and substantially outperform regions estimated in incorrect or random domains. The single-direction half-space performs slightly better, as expected, because the evaluated THINGSplus properties are defined by thresholding individual scalar ratings. The domain-level ellipsoid tests the complementary question of whether these properties also occupy compact convex regions within the broader recovered semantic domain. Figure 5b visualizes three representative properties. In each case, positive objects occupy compact convex regions within the recovered domain, and the estimated regions generalize well to held-out positives.



(a) (b)

Figure 6: **Recovering concept geometry and prototype structure.** (a) Representative concepts form compact prototype-centered regions across multiple semantic domains. The recovered regions generalize to held-out category members while remaining well separated from non-members. (b) Human typicality is consistently positively correlated to the recovered prototype.

Takeaway 3. A3: Properties

The results support **A3**: properties admit convex realizations within individual domains and generalize substantially better than regions estimated in incorrect or random domains.

4.4 A4: Concepts as Multi-Domain Regions

Construction. **A4** states that concepts are represented by collections of regions across multiple cognitive domains. In our realization, we use a separable approximation that omits explicit cross-domain covariance terms in Γ_C. We instantiate each concept by combining the recovered domains into a unified concept geometry and evaluate whether it generalizes to held-out category members.

For each concept C, let $\Delta_C \subseteq \Delta$ denote its relevant domains. Within each domain $\delta \in \Delta_C$, we estimate a prototype $\mu_{C,\delta}$ from the training members and define the corresponding domain distance $d_{C,\delta}(x)^2 = (z_\delta(x) - \mu_{C,\delta})^\top M_\delta (z_\delta(x) - \mu_{C,\delta})$, where M_δ is the shared metric recovered in **A2**. The concept distance is $d_C(x)^2 = \sum_{\delta \in \Delta_C} \alpha_{C,\delta} d_{C,\delta}(x)^2$, and the concept region is defined by $R_C = \{x : d_C(x)^2 \le r_C^2\}$. The domain weights $\alpha_{C,\delta}$ and radius r_C are estimated from the training split. We evaluate the recovered concept geometry using held-out category membership prediction on THINGS. To assess the contribution of domain organization, we compare the recovered concept regions against three baselines: uniform domain weights, permuted domain weights, and random same-rank domain decompositions. Additional implementation details are in Appendix C.

Results. Figure 6a visualizes a few representative concepts. Objects belonging to the same concept form compact prototype-centered regions spanning multiple semantic domains, and the recovered regions generalize well to held-out category members. Table 3 shows that the recovered multi-domain regions achieve strong held-out membership prediction. Uniform and permuted domain weights perform slightly worse, whereas replacing the recovered domains with random same-rank decompositions leads to a substantial degradation. These results indicate that concept membership is driven primarily by the recovered quality-domain geometry, while concept-specific domain weighting provides only a modest additional refinement.

Takeaway 4. A4: Concepts

The results support **A4**: concepts admit compact prototype-centered realizations spanning multiple semantic domains, and the recovered geometry generalizes well to unseen objects.

4.5 A5: Prototype Centrality and Typicality

Construction. **A5** states that prototype centrality determines graded concept typicality. We evaluate this prediction using the human typicality ratings in THINGSplus. The recovered concept geometry, including prototypes, domain weights, and metrics, is estimated entirely from category membership

Table 3: **Concept membership from recovered multi-domain regions.** Concept regions recovered from the semantic domains identified in **A2** consistently outperform random domain decompositions.

Domain Geometry	Weight	AUROC	AP	Coverage	Specificity
Recovered	Learned	**0.874**	**0.295**	0.839	**0.746**
	Uniform	0.868	0.291	**0.847**	0.722
	Permuted	0.863	0.284	**0.847**	0.712
Random	Learned	0.746	0.164	0.840	0.463

Table 4: **Robustness of the recovered conceptual geometry across LLM backbones.** Columns correspond to the empirical evaluations of GRH assumptions **A1–A5**. Held-out and category-disjoint Spearman correlations are reported for **A1**, normalized mutual information (NMI) for **A2**, AUROC for **A3** and **A4**, and category-wise Spearman correlation with human typicality for **A5**.

Model	A1 Held-out ρ	A1 Cat.-disjoint ρ	A2 Domain NMI	A3 Property AUROC	A4 Concept AUROC	A5 Typicality ρ
Llama-3.1-8B	0.674	0.442	0.660	0.981	0.874	0.168
Mistral-7B	0.665	0.413	0.635	0.981	0.874	0.185
Gemma-7B	0.675	0.446	0.687	0.983	0.879	0.214

as described in **A4**. For each category C, we define the predicted typicality as $\widehat{T}(x, C) = -d_C(x)$, where $d_C(x)$ is the recovered concept distance. We then compute the Spearman correlation $\rho_C = \rho(\widehat{T}(x, C), T_{\text{human}}(x, C))$ for every category.

Results. Figure 6b shows that prototype centrality is positively correlated with human typicality in 45 of the 53 THINGSplus categories, with both the sign test and the Wilcoxon signed-rank test indicating significance ($p < 10^{-4}$). Thus, objects closer to the recovered prototype are consistently judged as more typical, despite human typicality annotations never being used during prototype estimation. Full comparisons across backbones and alternative prototype metrics, including the raw hidden-space centroid, are reported in Appendix C.6.

Takeaway 5. A5: Prototypes

The results support **A5**: geometric prototype centrality predicts human typicality, demonstrating that graded category structure emerges naturally from the recovered concept geometry.

4.6 Robustness across LLM Backbones

Besides using Llama-3.1-8B, we also evaluate whether the recovered conceptual geometry is model-specific. We repeat the complete reconstruction pipeline on models with similar size, namely Gemma-7B and Mistral-7B, and independently recover the conceptual realization defined by **A1–A5** for each model. Table 4 shows that the recovered conceptual geometry is highly consistent across all three backbones. Quality dimensions (**A1**), semantic domains (**A2**), property regions (**A3**), concept regions (**A4**), and prototype centrality (**A5**) all exhibit comparable performance, with only minor quantitative variation. These results suggest that the conceptual geometry recovered by GRH is a stable property of modern decoder-only LLMs instead of an artifact of a particular architecture. Detailed comparisons across all three models are provided in Appendix C.

5 Conclusion

We introduced the Gärdenforsian Representation Hypothesis (GRH), which provides a unified view of geometric structures that have largely been studied individually and shows how higher-level conceptual organization can be progressively recovered from lower-level geometric primitives. This perspective suggests that LLM representations can be analyzed as structured conceptual systems with explicit relations across levels of abstraction. More generally, GRH offers a framework for connecting cognitive theories of concepts with the internal organization of modern foundation models. Future work can examine how this conceptual geometry emerges during training, how it changes with context and composition, and whether its structure plays a causal role in model behavior.

AI USE STATEMENT

Generative AI tools were used to assist with grammar checking, proofreading, submission-format compliance checking, and code for result visualization. They were not used to generate synthetic data, develop the conceptual framework, formulate hypotheses or mathematical claims, construct proofs, design the methodology or experiments, implement the core methods, process the experimental data, analyze or interpret the results, or formulate the conclusions. The authors reviewed all AI-assisted outputs and independently verified the resulting text, citations, visualization code, and figures. The authors take full responsibility for the accuracy, originality, and final content of the paper.

ETHICS STATEMENT

This work analyzes the internal representations of publicly available language models using public benchmark datasets. It involves no human-subject experiments, private data, or high-stakes deployment. The work is interpretive in scope, with the main ethical considerations arising from the broader use of methods for understanding and modifying language-model representations.

REPRODUCIBILITY STATEMENT

Section 3 states the assumptions and formal results underlying GRH, with proofs in Appendix D. The appendix documents dataset statistics, preprocessing, construction procedures, evaluation protocols, and implementation details. Cross-backbone results are summarized in Section 4.6 and reported in full in Appendix C. Code for reproducing the experiments is included in the supplemental material and will be released upon publication.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *NeurIPS*, 2024.

Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for ai safety–a review. *arXiv preprint arXiv:2404.14082*, 2024.

Jeffrey R Binder, Lisa L Conant, Colin J Humphries, Leonardo Fernandino, Stephen B Simons, Mario Aguilar, and Rutvik H Desai. Toward a brain-based componential semantic representation. *Cognitive neuropsychology*, 33(3-4):130–174, 2016.

Stefano Colamonaco, David Debot, Pietro Barbiero, and Giuseppe Marra. Prototype-grounded concept models for verifiable concept alignment. *CoRR*, abs/2604.16076, 2026.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The LLaMA 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *CoRR*, abs/2209.10652, 2022.

Peter Gärdenfors. *Conceptual spaces - the geometry of thought*. MIT Press, 2000.

Peter Gärdenfors. Conceptual spaces as a framework for knowledge representation. *Mind and matter*, 2(2):9–27, 2004.

Peter Gärdenfors. *The geometry of meaning: Semantics based on conceptual spaces*. The MIT Press, 2014.

Robert L Goldstone, Alan Kersten, and Paulo F Carvalho. Categorization and concepts. *Stevens' handbook of experimental psychology and cognitive neuroscience*, 3:275–317, 2018.

Wes Gurnee and Max Tegmark. Language models represent space and time. In *ICLR*. OpenReview.net, 2024.

Chi Han, Jialiang Xu, Manling Li, Yi Fung, Chenkai Sun, Nan Jiang, Tarek F. Abdelzaher, and Heng Ji. Word embeddings are steers for language models. In *ACL (1)*, pp. 16410–16430. Association for Computational Linguistics, 2024.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Yibo Jiang, Goutham Rajendran, Pradeep Ravikumar, Bryon Aragam, and Victor Veitch. On the origins of linear representations in large language models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Nitesh Kumar, Usashi Chatterjee, and Steven Schockaert. Ranking entities along conceptual space dimensions with LLMs: An analysis of fine-tuning strategies. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 7974–7989, 2024.

Nitesh Kumar, Usashi Chatterjee, and Steven Schockaert. Extracting conceptual spaces from LLMs using prototype embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 9275–9298, 2025.

Andrew Lee, Melanie Weber, Fernanda Viégas, and Martin Wattenberg. Shared global and local geometry of language model embeddings. *arXiv preprint arXiv:2503.21073*, 2025.

Ruixi Lin and Hwee Tou Ng. Does BERT know that the IS-A relation is transitive? In *ACL (2)*, pp. 94–99. Association for Computational Linguistics, 2022.

Alexander Modell, Patrick Rubin-Delanchy, and Nick Whiteley. The origins of representation manifolds in large language models. *CoRR*, abs/2505.18235, 2025.

Aaron Mueller, Andrew Lee, Shruti Joshi, Ekdeep Singh Lubana, Dhanya Sridhar, and Patrik Reizinger. From isolation to entanglement: When do interpretability methods identify and disentangle known concepts? In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17188–17210, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.782. URL https://aclanthology.org/2026.acl-long.782/.

Gregory L Murphy. *The big book of concepts*. The MIT press, 2002.

Burin Naowarat, Hao Tang, and Sharon Goldwater. A framework for analyzing concept representations in neural models. In Claire Bonial and Yevgeni Berzak (eds.), *Proceedings of the 30th Conference on Computational Natural Language Learning*, pp. 574–587, San Diego, California, USA, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-410-1. doi: 10.18653/v1/2026.conll-main.34. URL https://aclanthology.org/2026.conll-main.34/.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. *ICLR*, 2025.

Shivam Ratnakar and Kartikeya Vats. The geometry of refusal: Linear instability in safety-aligned llms. *CoRR*, abs/2606.22686, 2026.

Eleanor Rosch and Barbara B Lloyd. *Cognition and categorization*. Taylor & Francis, 2024.

Laura M Stoinski, Jonas Perkuhn, and Martin N Hebart. Thingsplus: New norms and metadata for the things database of 1854 object concepts and 26,107 natural object images. *Behavior Research Methods*, 56(3):1583–1603, 2024.

Lihao Sun, Lewen Yan, Xiaoya Lu, Andrew Lee, Jie Zhang, and Jing Shao. Valence-arousal subspace in llms: Circular emotion geometry and multi-behavioral control, 2026. URL https://arxiv.org/abs/2604.03147.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Deepika SN Vemuri, Sayanta Adhikari, Ankit Saha, Krishn Vishwas Kher, and Vineeth N Balasubramanian. Formal concept lattices are good semantic scaffolds for concept-based learning, 2026. URL https://arxiv.org/abs/2606.05471.

Johanna Vielhaben, Dilyara Bareeva, Jim Berend, Wojciech Samek, and Nils Strodthoff. Beyond scalars: Concept-based alignment analysis in vision transformers. In *NeurIPS*, 2025.

Zihao Wang, Lin Gui, Jeffrey Negrea, and Victor Veitch. Concept algebra for (score-based) text-controlled generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://openreview.net/forum?id=SGlrCuwdsB.

Weiqi Wu, Chengyue Jiang, Yong Jiang, Pengjun Xie, and Kewei Tu. Do plms know and understand ontological knowledge? In *ACL (1)*, pp. 3080–3101. Association for Computational Linguistics, 2023.

Bo Xiong. The lattice representation hypothesis of large language models. In *ICLR*, 2026.

Bo Xiong and Steffen Staab. From tokens to lattices: Emergent lattice structures in language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

A Limitations

Our analysis concerns how conceptual primitives are organized in LLM representation geometry and does not imply that LLMs acquire or use concepts through the same mechanisms as humans. Our experiments also focus on static object representations. This setting supports the recovery of quality dimensions, domains, properties, concepts, and prototypes within a common framework, but does not address contextualized concepts, semantic composition, or meaning that changes across contexts. Some components of Gärdenfors' conceptual-space theory are only partially instantiated. In particular, our empirical treatment of concepts uses a separable approximation and does not explicitly model the full cross-domain dependency structure encoded by Γ_C. Domain salience weights are also treated operationally. Richer cross-domain interactions and alternative conceptual metrics remain to be studied. Finally, our experiments cover decoder-only language models and concept-level semantic benchmarks. The consistency observed across the three tested backbones does not establish that the same organization holds for multimodal models or other representation-learning systems.

B Supplemental Related Work

Linear Geometry of LLM Representations. Semantic features in language models are often associated with approximately linear directions or low-dimensional subspaces (Elhage et al., 2022; Park et al., 2024). Other work shows that feature representations can also be distributed, entangled, or heterogeneous, limiting a strict one-concept–one-direction view (Mueller et al., 2026; Naowarat et al., 2026). At a higher level of organization, the *Lattice Representation Hypothesis* shows how thresholded attribute directions induce formal concepts and concept-lattice structure through intersections of half-spaces (Xiong, 2026).

Conceptual and Structured Geometry of LLMs. Gärdenfors' theory of conceptual spaces organizes quality dimensions into domains with domain-specific metrics, convex properties, multi-domain concepts, and prototypes (Gärdenfors, 2000; Gärdenfors, 2004). Kumar et al. (2025) show that

Table 5: Summary of the datasets used for evaluating GRH.

Dataset	# Objects	Annotations	GRH Assumptions
THINGSplus	1,854	12 quality dimensions, concept membership, typicality	A1, A3, A4, A5
Binder Feature Norms	535	65 quality dimensions, 14 human domains	A1, A2

perceptual quality dimensions can be recovered from LLM representations using prototype embeddings. Structured low-dimensional geometries have also been reported for relational and affective representations (Jiang et al., 2024; Sun et al., 2026). These findings motivate studying whether such structures participate in a common conceptual organization within the model representation space.

Regions, Prototypes, and Lattice Structures. Concepts have also been modeled using geometric objects beyond single directions, including manifolds and learned prototypes (Vielhaben et al., 2025; Colamonaco et al., 2026). Work on Formal Concept Analysis has connected lattice structure to language-model representations (Xiong & Staab, 2025; Xiong, 2026) and used externally specified concept lattices as semantic scaffolds (Vemuri et al., 2026). GRH connects these lines of work by placing property regions, concept regions, prototype structure, and lattice discretization within the same conceptual geometry.

C Experimental Details and Full Results

This appendix provides the implementation details and complete cross-backbone results for the empirical evaluation of A1–A5. For each backbone, including Llama-3.1-8B, Gemma-7B, and Mistral-7B, we independently extract object representations, estimate quality directions, recover semantic domains, and fit property and concept regions. No direction, cluster assignment, metric, or region parameter is transferred across models. Unless otherwise specified, all experiments use final-layer hidden representations.

C.1 Datasets and Evaluation Setup

Table 5 summarizes the datasets used throughout the paper. THINGSplus serves as the primary benchmark for evaluating quality dimensions, properties, concepts, and prototypes. Binder feature norms provide a substantially richer collection of semantic attributes together with human-defined domains, enabling evaluation of the domain organization predicted by GRH.

For every object, we average the final-layer representations obtained from its lexical variants to obtain a single object representation. The same representation extraction procedure is applied independently to each backbone. All subsequent directions, domains, metrics, and regions are estimated within the corresponding model's representation space.

C.2 A1: Quality Dimensions

Construction. For each quality dimension q, we construct positive and negative sets from objects in the upper and lower quartiles of the corresponding human ratings. Let μ_+ and μ_- denote the mean representations of the two sets, and let $\widetilde{\Sigma}_+$ and $\widetilde{\Sigma}_-$ denote their regularized covariance estimates. We recover the quality direction using a regularized Fisher discriminant, $d_q = (\widetilde{\Sigma}_+ + \widetilde{\Sigma}_- + \lambda I)^{-1}(\mu_+ - \mu_-)$. The direction is normalized to obtain $\hat{d}_q$, and the recovered quality coordinate of an object x is $z_q(x) = \hat{d}_q^\top v_x$.

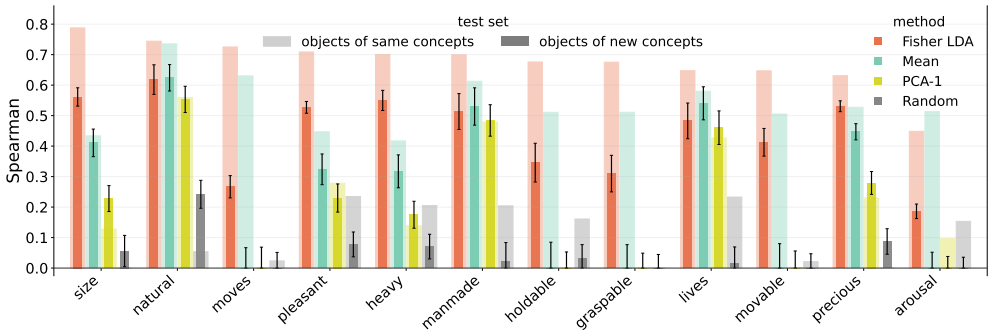
Evaluation. We evaluate each recovered coordinate by its Spearman correlation with held-out human quality ratings. We consider both ordinary held-out splits and a more stringent category-disjoint setting in which the semantic categories used for estimating the direction do not appear in the evaluation set. We compare Fisher directions with mean-difference, first-principal-component (PCA-1), and random directions.

Results. Table 6 reports the complete comparison across backbones. Fisher is the strongest method for every model and evaluation split. On held-out objects, mean correlations are tightly concentrated across models (0.665–0.675), despite larger differences among the weaker baselines. The category-

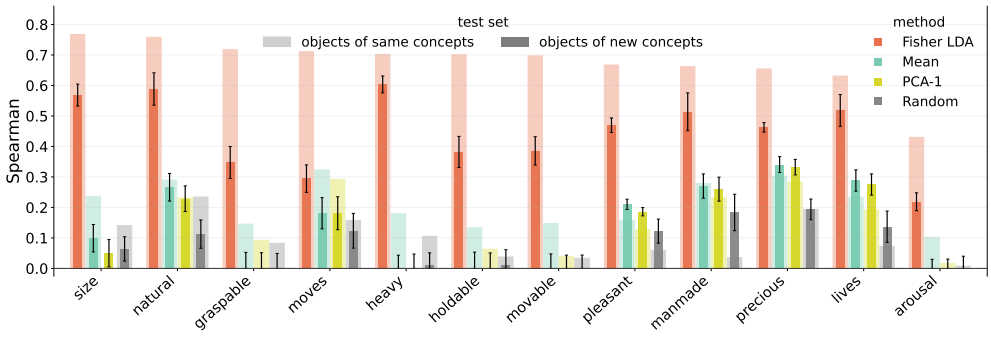
Table 6: **Quality-dimension recovery across backbones.** Entries are mean Spearman correlations across the 12 THINGSplus quality dimensions. Held-out evaluates new objects from the same concept inventory; category-disjoint evaluates objects from unseen concepts.

Model	Direction estimator	Held-out ρ	Category-disjoint ρ
Llama-3.1-8B	Fisher	**0.674**	**0.442**
	Mean difference	0.535	0.249
	PCA-1	0.177	0.130
	Random	0.106	0.041
Gemma-7B	Fisher	**0.675**	**0.446**
	Mean difference	0.211	0.085
	PCA-1	0.128	0.070
	Random	0.096	0.062
Mistral-7B	Fisher	**0.665**	**0.413**
	Mean difference	0.420	0.162
	PCA-1	0.178	0.055
	Random	0.127	0.038

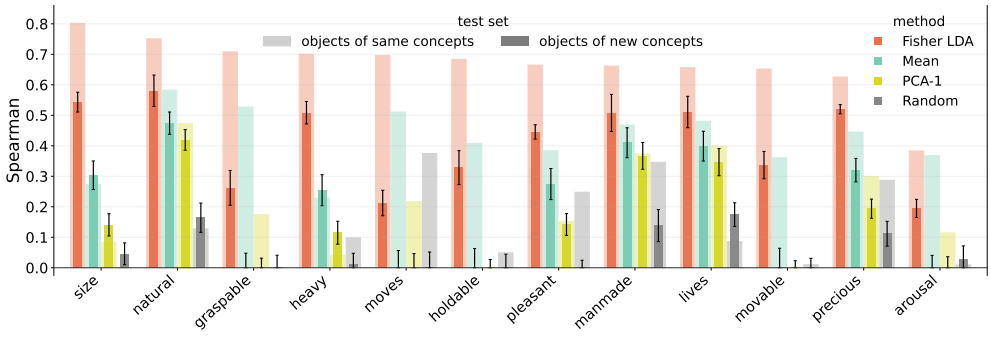
disjoint setting provides the more demanding test. Fisher directions retain correlations of 0.442, 0.446, and 0.413 for Llama-3.1-8B, Gemma-7B, and Mistral-7B, respectively. By contrast, the best non-Fisher baseline reaches only 0.249, 0.085, and 0.162. Thus, directions estimated from one set of semantic categories recover graded orderings among previously unseen concepts. Figures 7–10 show that this effect is distributed across qualities and remains visible when directions are estimated from few extreme examples.



(a) Llama-3.1-8B

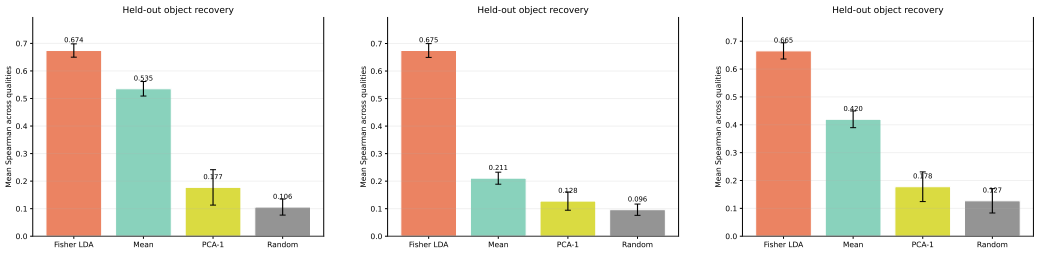


(b) Gemma-7B



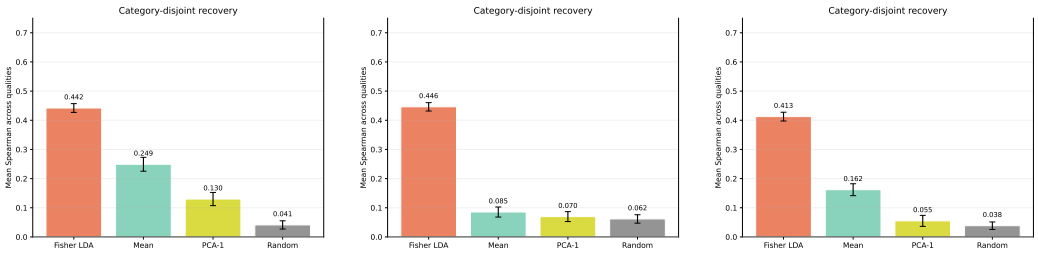
(c) Mistral-7B

Figure 7: **Quality directions recover graded human judgments across attributes.** Light bars report held-out objects from the same concept inventory. Dark bars report objects from unseen concepts.



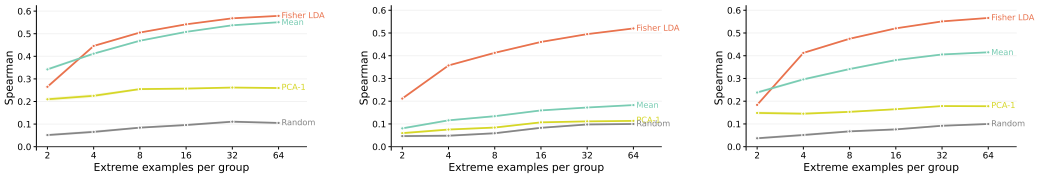
(a) Llama-3.1-8B (b) Gemma-7B (c) Mistral-7B

Figure 8: **Fisher directions outperform alternative direction estimators.** Bars report the mean held-out Spearman correlation across quality dimensions.



(a) Llama-3.1-8B (b) Gemma-7B (c) Mistral-7B

Figure 9: **Quality directions generalize to unseen concepts.** All methods are evaluated under the same repeated category-disjoint splits.



(a) Llama-3.1-8B (b) Gemma-7B (c) Mistral-7B

Figure 10: **Quality dimensions can be recovered from few extreme examples.** Curves show means over repeated samples; error bars show sampling variability.

Table 7: **Recovered Binder domains align with human domain annotations.** Each model clusters 65 independently estimated quality directions into 14 domains. A denotes direction affinity and B object-induced coupling.

Model	Affinity	ARI	NMI	Purity
Llama-3.1-8B	A	0.324	0.660	0.631
	B	0.229	0.602	0.585
Gemma-7B	A	**0.371**	**0.687**	**0.662**
	B	0.229	0.602	0.585
Mistral-7B	A	0.284	0.635	0.600
	B	0.231	0.604	0.585

C.3 A2: Domain Organization and Metrics

Affinity matrices. Given the recovered quality directions $\{\hat{d}_1, \ldots, \hat{d}_K\}$, we construct two affinity matrices. The first measures directional similarity, $A_{ij} = |\cos(\hat{d}_i, \hat{d}_j)|$, while the second measures the correlation of the induced quality coordinates, $B_{ij} = \mathrm{Corr}_{x \in \mathcal{X}}(\hat{d}_i^\top v_x, \hat{d}_j^\top v_x)$. Unless otherwise specified, the main paper reports results obtained from the direction affinity matrix A, while B provides an alternative construction.

Domain recovery. Spectral clustering on the affinity matrix yields a partition $c : \{1, \ldots, K\} \to \{1, \ldots, L\}$. Each recovered domain is represented by $S_\ell = \mathrm{span}\{\hat{d}_j : c(j) = \ell\}$. Let $D_\ell = [\hat{d}_j : c(j) = \ell]$. The corresponding orthogonal projector is $P_\ell = D_\ell (D_\ell^\top D_\ell)^+ D_\ell^\top$.

Domain metrics. The Euclidean distance induced by a recovered domain is $d_\ell(x, y) = \|P_\ell(v_x - v_y)\|_2$. Equivalently, if U_ℓ is an orthonormal basis of S_ℓ, then $d_\ell(x, y) = \|U_\ell^\top v_x - U_\ell^\top v_y\|_2$. We additionally evaluate a diagonal metric obtained by standardizing each domain coordinate according to its pooled variance, $M_\ell = \mathrm{diag}(s_{\ell,1}^{-2}, \ldots, s_{\ell,r}^{-2})$.

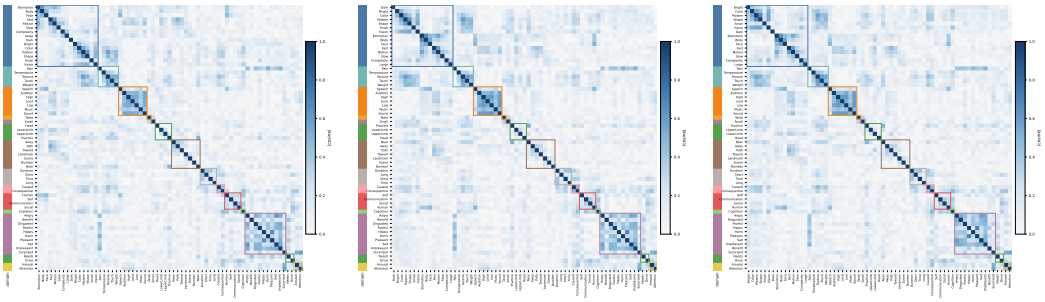
Evaluation. On Binder, recovered clusters are compared with the 14 human-annotated semantic domains using adjusted Rand index (ARI), normalized mutual information (NMI), and purity. On THINGSplus, we test whether recovered domains induce behaviorally meaningful similarity metrics using 88,448 THINGS odd-one-out judgments. Each recovered domain is compared with 200 random subspaces of the same rank.

Results. Table 7 shows that direction affinity A consistently provides stronger alignment with the Binder domains than object-induced coupling B. Its NMI ranges from 0.635 to 0.687, whereas B remains near 0.602–0.604. The same ordering holds for ARI and purity. Because the Binder directions are estimated independently for every backbone, these blocks are not inherited from a shared clustering solution.

Figures 11 and 12 show both successful recovery and cross-domain mixing. Precise boundaries vary across models, but sensory, motor, spatial, affective, and cognitive groupings recur.

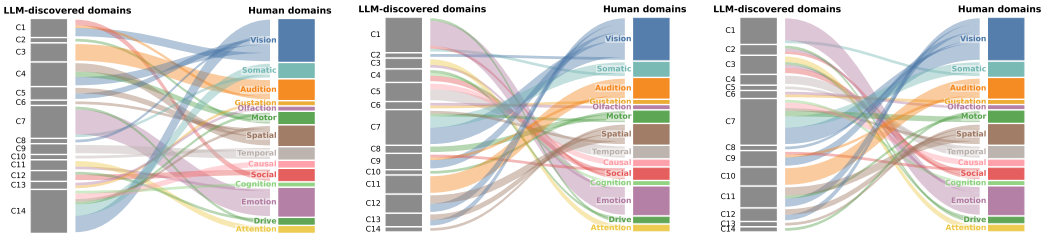
Domain-metric results. Table 8 compares the recovered THINGSplus domains with rank-matched random subspaces. The animacy/naturalness domain produces the largest and most stable relative gain, ranging from 31.6% to 35.9%. Affordance/scale and affect/value also outperform their controls for every backbone. The main exception is the one-dimensional arousal domain for Llama and Mistral, where the gain is small.

Unscaled and variance-normalized metrics yield the same qualitative conclusions, indicating that the recovered subspaces contribute more than the particular choice of coordinate normalization. Figures 13–17 further show that the global domain structure is stable while the domain most useful for human similarity varies across object categories.



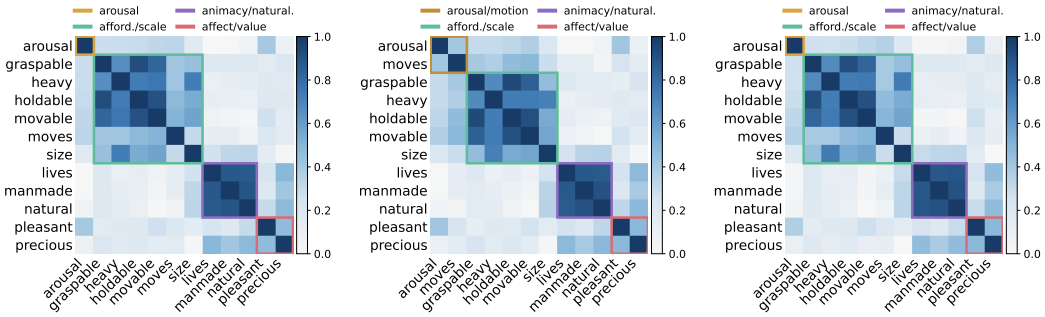
(a) Llama-3.1-8B (b) Gemma-7B (c) Mistral-7B

Figure 11: **Binder quality directions exhibit human-domain block structure.** Affinity matrices are ordered by human domain annotations; colored boundaries mark the 14 Binder domains.



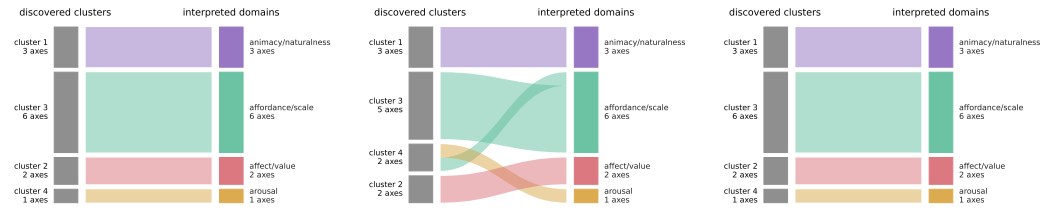
(a) Llama-3.1-8B (b) Gemma-7B (c) Mistral-7B

Figure 12: **LLM-discovered domains partially recover human domains.** Flow width equals the number of quality dimensions shared by a discovered cluster and a human-annotated domain.



(a) Llama-3.1-8B (b) Gemma-7B (c) Mistral-7B

Figure 13: **THINGSplus quality dimensions organize into coherent blocks.** The ordering and domain boundaries are recovered directly from the quality geometry.

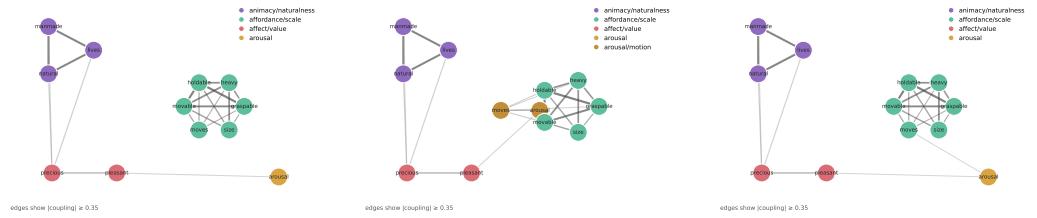


(a) Llama-3.1-8B (b) Gemma-7B (c) Mistral-7B

Figure 14: **The two domain-induction views recover related structure.** Alluvial mappings compare direction geometry (A) with object-induced coupling (B).

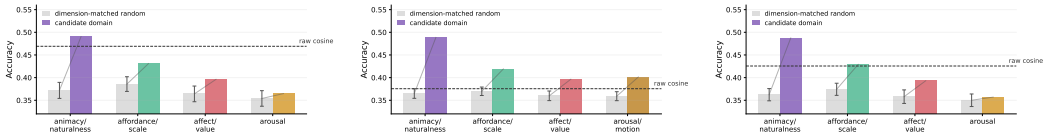
Table 8: **Recovered domains induce meaningful similarity metrics across backbones.** Target and random report THINGS odd-one-out accuracy. Random subspaces match the rank of the corresponding target domain; Δ is relative improvement over random accuracy.

			Unscaled			Scaled		
Model	Domain	Rank	Target	Random	$\Delta(\%)$	Target	Random	$\Delta(\%)$
Llama-3.1-8B	animacy / naturalness	3	0.492	0.374	+31.6%	0.490	0.372	+31.7%
	affordance / scale	6	0.429	0.388	+10.6%	0.431	0.385	+11.9%
	affect / value	2	0.394	0.366	+7.7%	0.396	0.365	+8.5%
	arousal	1	0.365	0.354	+3.1%	0.365	0.354	+3.1%
Gemma-7B	animacy / naturalness	3	0.489	0.364	+34.4%	0.489	0.365	+33.9%
	affordance / scale	5	0.425	0.367	+15.8%	0.419	0.370	+13.3%
	affect / value	2	0.399	0.360	+11.0%	0.397	0.360	+10.2%
	arousal / motion	2	0.406	0.359	+13.3%	0.401	0.359	+11.8%
Mistral-7B	animacy / naturalness	3	0.494	0.363	+35.9%	0.488	0.362	+34.6%
	affordance / scale	6	0.433	0.377	+15.1%	0.429	0.374	+14.7%
	affect / value	2	0.392	0.359	+9.2%	0.393	0.358	+9.8%
	arousal	1	0.357	0.350	+2.0%	0.357	0.350	+2.0%



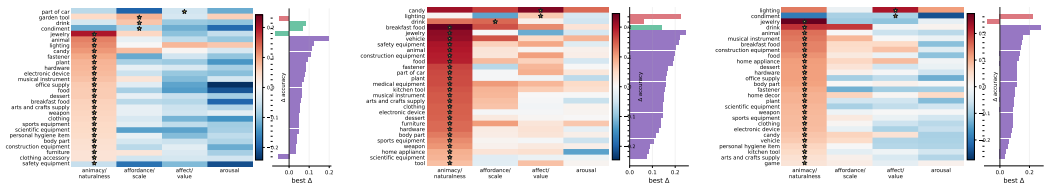
(a) Llama-3.1-8B (b) Gemma-7B (c) Mistral-7B

Figure 15: **Quality dimensions form structured domain networks.** Nodes denote quality directions; edges retain their strongest geometric relationships.



(a) Llama-3.1-8B (b) Gemma-7B (c) Mistral-7B

Figure 16: **Target domains outperform random subspaces of the same rank.** Panels report odd-one-out accuracy and the target-minus-random difference for each recovered domain.



(a) Llama-3.1-8B (b) Gemma-7B (c) Mistral-7B

Figure 17: **Different object categories rely on different domain metrics.** A domain is marked as dominant when its score is at least 0.9 times the best domain score for that category.

Table 9: **Held-out property-region performance across backbones.** All regions are estimated from training positives and evaluated on held-out positive and negative objects. Random subspaces match the rank of the correct domain.

Model	Region	AUROC	AP	Coverage	Specificity
Llama-3.1-8B	Correct-domain ellipsoid	0.981	0.985	0.896	0.977
	Single-direction half-space	**0.998**	**0.996**	0.898	**0.998**
	Wrong-domain ellipsoid	0.668	0.642	0.895	0.263
	Random same-rank ellipsoid	0.578	0.577	0.894	0.174
Gemma-7B	Correct-domain ellipsoid	0.983	0.988	0.895	0.989
	Single-direction half-space	**0.998**	**0.996**	0.895	**0.999**
	Wrong-domain ellipsoid	0.666	0.644	0.895	0.295
	Random same-rank ellipsoid	0.550	0.550	0.897	0.143
Mistral-7B	Correct-domain ellipsoid	0.981	0.987	0.896	0.975
	Single-direction half-space	**0.998**	**0.996**	0.896	**0.999**
	Wrong-domain ellipsoid	0.662	0.638	0.894	0.268
	Random same-rank ellipsoid	0.565	0.562	0.896	0.161

C.4 A3: Properties as Convex Regions

Construction. For each THINGSplus property j, we identify positive and negative examples using the upper and lower quartiles of the human ratings, $P_j^+ = \{x : y_j(x) \geq Q_{75}(y_j)\}$, $P_j^- = \{x : y_j(x) \leq Q_{25}(y_j)\}$. Positive and negative examples are randomly divided into training and held-out subsets. Only the training positives are used to estimate the property region.

Let $\mathcal{D}(j)$ denote the recovered semantic domain associated with property j, and let $z_{\mathcal{D}(j)}(x)$ denote the corresponding domain coordinates. The property center is $\mu_j = \frac{1}{|P_{j,\text{train}}^+|} \sum_{x \in P_{j,\text{train}}^+} z_{\mathcal{D}(j)}(x)$. The radius is chosen as the 90th percentile of training distances, $r_j = Q_{90}\left(\{d(x, \mu_j) : x \in P_{j,\text{train}}^+\}\right)$. The resulting property region is $R_j = \left\{x : (z_{\mathcal{D}(j)}(x) - \mu_j)^\top M_{\mathcal{D}(j)}(z_{\mathcal{D}(j)}(x) - \mu_j) \leq r_j^2\right\}$.

Evaluation. Property regions are evaluated on held-out positives and negatives using AUROC, average precision (AP), coverage, and specificity. We compare the recovered region with three baselines: (i) a single-direction half-space, (ii) an ellipsoid estimated in an incorrect semantic domain, and (iii) an ellipsoid estimated in a random same-rank subspace.

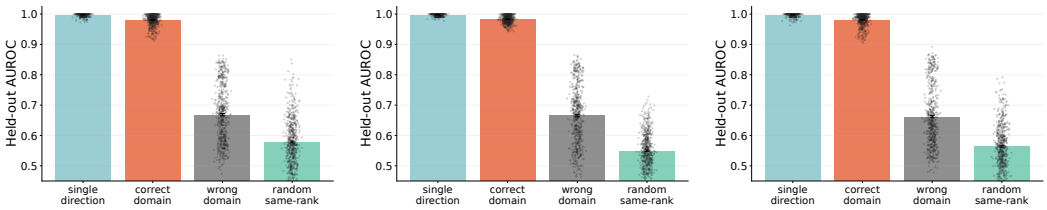
Results. Table 9 reports the complete comparison. Correct-domain ellipsoids reach AUROC above 0.98 for every backbone, with AP above 0.984 and specificity above 0.975 at approximately 0.895 coverage. Random same-rank ellipsoids are substantially weaker (AUROC 0.550–0.578), while wrong-domain regions occupy an intermediate range (0.662–0.668). Thus, low dimensionality alone does not explain the effect; the region must lie in the domain associated with the property.

Single-direction half-spaces obtain AUROC near 0.998. This is expected because THINGSplus positives are defined by a single graded human quality, making a threshold along the recovered quality coordinate the matched minimum-complexity solution. The domain ellipsoid instead tests whether the larger semantic domain supports a compact convex region that generalizes to held-out objects. Figures 18–23 show that the correct-domain advantage is distributed across qualities and repeated splits.



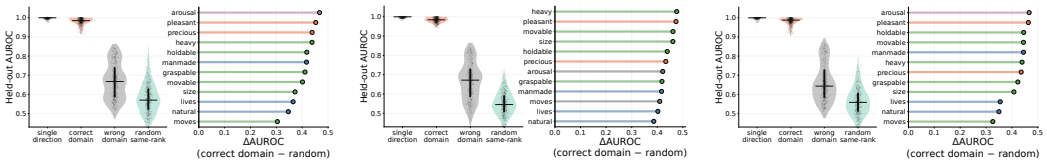
(a) Llama-3.1-8B (b) Gemma-7B (c) Mistral-7B

Figure 18: **Properties occupy compact regions in their corresponding domains.** Regions are fitted from training positives and displayed with held-out positive and negative objects.



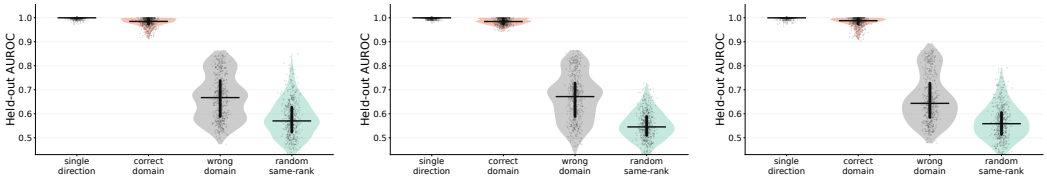
(a) Llama-3.1-8B (b) Gemma-7B (c) Mistral-7B

Figure 19: **Matched property regions outperform wrong-domain and random controls.** Results are aggregated across qualities and repeated train–test splits.



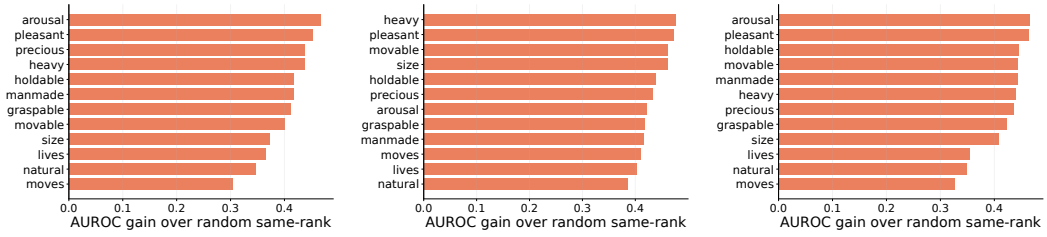
(a) Llama-3.1-8B (b) Gemma-7B (c) Mistral-7B

Figure 20: **The property-region effect is consistent across quality dimensions.** Panels report held-out performance separately for each THINGSplus property.



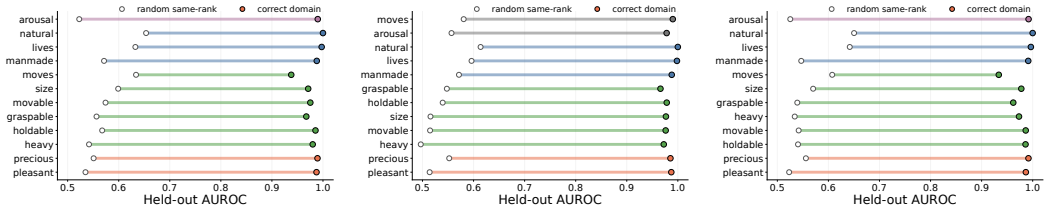
(a) Llama-3.1-8B (b) Gemma-7B (c) Mistral-7B

Figure 21: **Property-region performance is stable across repeated splits.** Raincloud plots show the full distribution, including variation around the mean.



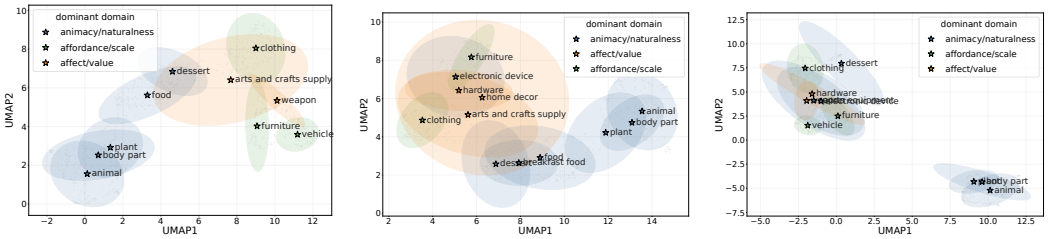
(a) Llama-3.1-8B (b) Gemma-7B (c) Mistral-7B

Figure 22: **Correct domains yield positive AUROC gains across properties.** Here $\Delta\text{AUROC} = \text{AUROC}_{\text{correct domain}} - \text{AUROC}_{\text{random same-rank}}$.



(a) Llama-3.1-8B (b) Gemma-7B (c) Mistral-7B

Figure 23: **Property prediction benefits from the recovered domain geometry.** Lollipop plots show the correct-domain advantage over the dimension-matched random-subspace control.



(a) Llama-3.1-8B (b) Gemma-7B (c) Mistral-7B

Figure 24: **Concepts form prototype-centered regions in a UMAP layout.** Stars mark the visual centroids of the projected regions.

C.5 A4: Concepts as Multi-Domain Regions

Construction. For each concept C and each recovered domain $\delta \in \Delta_C$, we estimate a domain-specific prototype from the training members, $\mu_{C,\delta} = \frac{1}{|C_{\text{train}}|} \sum_{x \in C_{\text{train}}} z_\delta(x)$. The corresponding domain-specific distance is $d_{C,\delta}(x)^2 = (z_\delta(x) - \mu_{C,\delta})^\top M_\delta (z_\delta(x) - \mu_{C,\delta})$.

Domain weighting. For every recovered domain, we compute the separation between training members and non-members using $d_{C,\delta}(x)$. The resulting separation scores are normalized to obtain non-negative weights $\alpha_{C,\delta}$ satisfying $\sum_{\delta \in \Delta_C} \alpha_{C,\delta} = 1$.

Concept region. The multi-domain concept distance is $d_C(x)^2 = \sum_{\delta \in \Delta_C} \alpha_{C,\delta} d_{C,\delta}(x)^2$. The concept radius is chosen as the 90th percentile of the training distances, $r_C = Q_{90}\left(\{d_C(x) : x \in C_{\text{train}}\}\right)$, yielding $R_C = \{x : d_C(x)^2 \le r_C^2\}$.

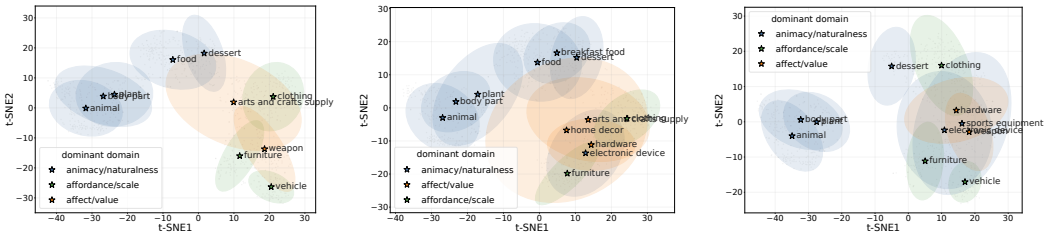
Evaluation. Concept regions are evaluated on held-out category members and non-members using AUROC, AP, coverage, specificity, and precision. We compare the recovered concept geometry with three controls: (i) *uniform weighting*, where $\alpha_{C,\delta} = 1/|\Delta_C|$; (ii) *permuted weighting*, where learned weights are randomly permuted across concepts while preserving the recovered domains; and (iii) *random same-rank decomposition*, where each recovered domain is replaced by a random subspace of identical dimension.

Results. Table 10 reports the complete concept-membership comparison. Recovered domain-weighted regions obtain AUROC 0.874–0.879 across backbones. Replacing the recovered domains with random same-rank subspaces lowers AUROC to 0.631–0.746 and substantially reduces AP and specificity at comparable coverage. The result is therefore not explained by the dimensionality or parameter count of the region.

Uniform and permuted weights remain close to learned weights but are consistently weaker. The large recovered-versus-random gap shows that the domain decomposition contributes most of the useful structure, while the smaller learned-versus-uniform/permuted gap indicates that concept-specific domain salience provides an additional refinement. The projections in Figures 24– 26 are illustrative only; all membership metrics are computed in the original domain coordinates.

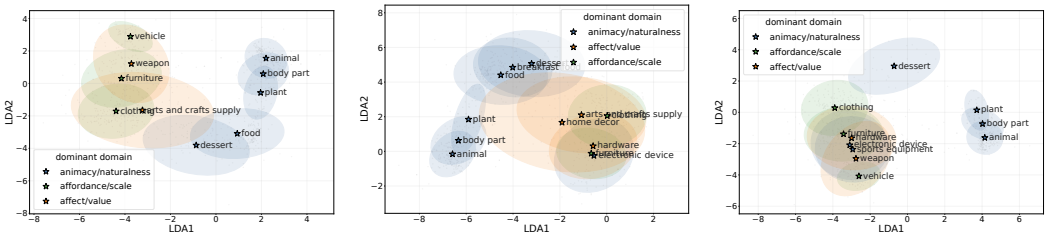
Table 10: **Concept membership from recovered multi-domain regions.** Centers, radii, and learned domain weights are estimated from training members only; metrics are evaluated on held-out members and non-members.

Model	Domain Geometry	Weight	AUROC	AP	Coverage	Specificity
Llama-3.1-8B	Recovered	Learned	**0.874**	**0.295**	0.839	**0.746**
	Recovered	Uniform	0.868	0.291	**0.847**	0.722
	Recovered	Permuted	0.863	0.284	0.847	0.712
	Random	Learned	0.746	0.164	0.840	0.463
Gemma-7B	Recovered	Learned	**0.879**	**0.298**	0.842	**0.755**
	Recovered	Uniform	0.873	0.291	0.852	0.726
	Recovered	Permuted	0.871	0.286	**0.852**	0.722
	Random	Learned	0.631	0.092	0.849	0.279
Mistral-7B	Recovered	Learned	**0.874**	**0.304**	0.848	**0.735**
	Recovered	Uniform	0.869	0.299	0.850	0.720
	Recovered	Permuted	0.866	0.292	**0.854**	0.707
	Random	Learned	0.691	0.118	0.837	0.376



(a) Llama-3.1-8B (b) Gemma-7B (c) Mistral-7B

Figure 25: **Concept regions remain visible under a local-neighborhood-preserving layout.** The t-SNE projection provides a complementary view of the same fitted multi-domain regions.



(a) Llama-3.1-8B (b) Gemma-7B (c) Mistral-7B

Figure 26: **Supervised projections reveal separation among concept regions.** LDA emphasizes between-concept separation while retaining the prototype-centered organization.

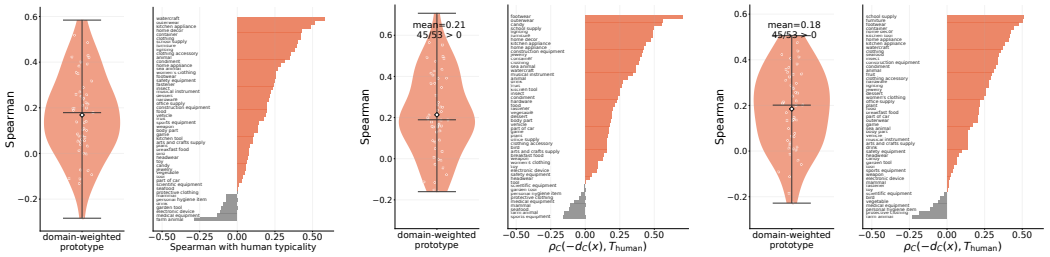
C.6 A5: Prototype Centrality and Typicality

Evaluation. Concept prototypes and domain weights are estimated solely from concept membership; human typicality ratings are never used in fitting. For each object x and concept C, the primary geometric typicality prediction is $\widehat{T}(x, C) = -d_C(x)$. We evaluate the prediction using category-wise Spearman correlation with human typicality judgments over the 53 THINGSplus categories.

We additionally compare domain-weighted prototype distance with Euclidean and axis-scaled distances in quality space, as well as distance to the raw hidden-space centroid. These alternatives test the sensitivity of the typicality effect to the choice of metric and the contribution of prototype centrality.

Table 11: **Prototype centrality and human typicality across backbones.** Entries are category-wise Spearman correlations over 53 THINGSplus categories. Positive categories and significance tests are reported for the domain-weighted prototype, the primary A5 test.

Model	Prototype metric	Mean ρ	Median ρ	Positive cats.	Wilcoxon p
Llama-3.1-8B	Domain-weighted quality space	0.168	0.179	45/53	2.1×10^{-7}
	Euclidean quality space	0.170	0.186	–	–
	Axis-scaled quality space	0.176	0.198	–	–
	Raw hidden centroid	**0.290**	**0.300**	–	–
Gemma-7B	Domain-weighted quality space	0.214	0.189	44/53	1.2×10^{-8}
	Euclidean quality space	0.205	0.176	–	–
	Axis-scaled quality space	**0.215**	**0.190**	–	–
	Raw hidden centroid	0.107	0.072	–	–
Mistral-7B	Domain-weighted quality space	0.185	0.201	46/53	3.6×10^{-8}
	Euclidean quality space	0.190	0.202	–	–
	Axis-scaled quality space	**0.194**	**0.217**	–	–
	Raw hidden centroid	0.185	0.195	–	–



(a) Llama-3.1-8B (b) Gemma-7B (c) Mistral-7B

Figure 27: **Geometric prototype centrality predicts human typicality.** Category-wise correlations are broadly positive for every backbone, although their magnitude varies across concepts and models.

Results. Domain-weighted prototype centrality is positively related to human typicality for all three models. Mean category-wise Spearman correlation is 0.168, 0.214, and 0.185 for Llama-3.1-8B, Gemma-7B, and Mistral-7B, with positive correlations in 45, 44, and 46 of the 53 categories. Both sign and one-sided Wilcoxon tests reject zero-centered effects for every backbone. Prototype centrality therefore predicts graded typicality even though typicality annotations are never used to estimate the prototypes.

Table 11 also clarifies the role of the metric. Gemma benefits strongly from moving out of raw hidden space, whereas Mistral's domain-weighted and raw-centroid means are nearly identical. For Llama-3.1-8B, the raw hidden centroid is stronger than the quality-space metrics. The robust A5 conclusion is therefore that centrality around a membership-estimated prototype predicts graded typicality; the stronger claim that one particular domain weighting must dominate every prototype metric is not supported uniformly across backbones.

Takeaway 6. Full Cross-Backbone Analysis

All five geometric predictions are recovered across Llama-3.1-8B, Gemma-7B, and Mistral-7B. Quality directions generalize across concepts; quality geometry yields stable domains and meaningful domain metrics; properties and concepts generalize as convex regions; and prototype centrality predicts human typicality. The precise domain boundaries and best-performing metric vary across models, but the central comparisons with matched random or wrong-domain controls remain stable.

D PROOFS FOR SECTION 3

D.1 PROOF OF PROPOSITION 1

Proof. Fix a quality dimension d_i and consider the restriction of the GRH conceptual realization to this dimension. By A1, $d_i \in \mathbb{R}^d$ defines the linear functional

$$\ell_i : \mathbb{R}^d \to \mathbb{R}, \qquad \ell_i(v) = d_i^\top v.$$

For every object $x \in \mathcal{X}$, its coordinate along the restricted quality dimension is therefore

$$z_i(x) = \ell_i(v_x) = d_i^\top v_x.$$

Hence the semantic value associated with this quality dimension depends linearly on the representation v_x. Moreover, for any two objects $x, y \in \mathcal{X}$,

$$z_i(x) - z_i(y) = d_i^\top (v_x - v_y),$$

so differences in the corresponding graded attribute are represented along the same direction d_i.

Thus, after restricting GRH to a single quality dimension, the resulting representation is a one-dimensional linear encoding of that semantic attribute. This recovers the geometric structure assumed by the Linear Representation Hypothesis at the level of individual graded attributes. □

D.2 PROOF OF THEOREM 1

Proof. Let

$$z = D^\top v$$

denote the conceptual coordinates of an object. Consider two concepts C_i and C_j with prototypes p_i and p_j, classified by nearest prototype under the same symmetric positive semidefinite metric $M \succeq 0$.

The decision boundary consists of points equidistant from the two prototypes:

$$(z - p_i)^\top M(z - p_i) = (z - p_j)^\top M(z - p_j).$$

Since M is symmetric, expanding both sides gives

$$z^\top Mz - 2p_i^\top Mz + p_i^\top Mp_i = z^\top Mz - 2p_j^\top Mz + p_j^\top Mp_j.$$

Because the two concepts share the same metric M, the quadratic terms $z^\top Mz$ cancel. Rearranging yields

$$2(p_i - p_j)^\top Mz = p_i^\top Mp_i - p_j^\top Mp_j.$$

Substituting $z = D^\top v$ gives

$$2(p_i - p_j)^\top MD^\top v = p_i^\top Mp_i - p_j^\top Mp_j.$$

Using symmetry of M,

$$2(p_i - p_j)^\top MD^\top v = \big(2DM(p_i - p_j)\big)^\top v.$$

Therefore the decision boundary can be written as

$$n_{ij}^\top v = \tau_{ij},$$

where

$$n_{ij} = 2DM(p_i - p_j), \qquad \tau_{ij} = p_i^\top Mp_i - p_j^\top Mp_j.$$

By the non-degeneracy assumption $DM(p_i - p_j) \neq 0$, we have $n_{ij} \neq 0$. Hence the solution set is a non-degenerate affine hyperplane in the representation space. □

Concept-specific metrics. The shared-metric assumption is essential to the linearity result. Suppose instead that C_i and C_j use symmetric metrics M_i and M_j, respectively. Their boundary satisfies

$$(z - p_i)^\top M_i (z - p_i) = (z - p_j)^\top M_j (z - p_j).$$

Expanding and collecting terms gives

$$z^\top (M_i - M_j) z - 2(M_i p_i - M_j p_j)^\top z + p_i^\top M_i p_i - p_j^\top M_j p_j = 0.$$

When $M_i \neq M_j$, the quadratic term $z^\top (M_i - M_j) z$ generally does not vanish, so the boundary is quadratic. Theorem 1 therefore characterizes the shared-metric special case of the more general concept geometry allowed by GRH.

D.3 Proof of Proposition 2

Proof. Let $\mathcal{P} = \{P_1, \ldots, P_m\}$ denote the set of thresholded properties. For each P_i, define the binary incidence relation

$$x \, I \, P_i \quad \Longleftrightarrow \quad d_i^\top v_x \geq \tau_i.$$

This determines a formal context

$$(\mathcal{X}, \mathcal{P}, I).$$

For $A \subseteq \mathcal{X}$ and $B \subseteq \mathcal{P}$, define the derivation operators

$$A' = \{P \in \mathcal{P} : x \, I \, P \text{ for every } x \in A\},$$

and

$$B' = \{x \in \mathcal{X} : x \, I \, P \text{ for every } P \in B\}.$$

The maps $A \mapsto A'$ and $B \mapsto B'$ form the standard Galois connection between subsets of objects and subsets of properties.

A pair (A, B) is a formal concept precisely when

$$A' = B \qquad \text{and} \qquad B' = A.$$

Equivalently, an attribute set B determines a formal concept when it is closed under the induced closure operator,

$$B = B'',$$

in which case the corresponding formal concept is (B', B). The collection of all such formal concepts, ordered by extent inclusion (or equivalently by reverse intent inclusion), forms the concept lattice associated with $(\mathcal{X}, \mathcal{P}, I)$.

The same construction has a continuous geometric interpretation. Each thresholded property defines the half-space

$$H_i = \{v \in \mathbb{R}^d : d_i^\top v \geq \tau_i\}.$$

For any property set $B \subseteq \mathcal{P}$,

$$B' = \left\{ x \in \mathcal{X} : v_x \in \bigcap_{P_i \in B} H_i \right\}.$$

Thus, $\bigcap_{P_i \in B} H_i \subseteq \mathbb{R}^d$ is a continuous polyhedral region, whereas $B' \subseteq \mathcal{X}$ is the discrete set of observed objects lying in that region. For closed $B = B''$, this discrete extent together with its intent forms a formal concept.

Hence, thresholding the continuous quality coordinates of GRH induces the binary incidence structure whose closed attribute sets form a formal concept lattice. The lattice representation is therefore obtained as a discrete, order-theoretic abstraction of the underlying continuous geometry. □