BEYOND THE LATTICE: A GRADED REPRESENTATION HYPOTHESIS FOR LLM CONCEPT HIERARCHIES

Anonymous authors
Paper under double-blind review

ABSTRACT

The *Lattice Representation Hypothesis* holds that large language models (LLMs) encode concept hierarchies as intersections of attribute half-spaces. Testing it on SNOMED CT, an ontology whose defining attributes are human-curated, we find that half-spaces are recovered reliably, but the lattice's inclusion score measures attribute overlap rather than subsumption: it orients *parent–child* pairs at only 59.5% and barely separates siblings from parents. Much of the directional information is lost in the inclusion aggregation, but a single linear direction in the embedding recovers it. In this paper, we propose the ***Graded Representation Hypothesis***: hierarchy is realized by a relatedness subspace (the lattice) together with a specificity direction whose projection assigns each concept a *grade*, a real-valued grading that approximately preserves the order of the *is-a* poset. Experiments on SNOMED CT and WordNet show that the learned grade correctly orients 94% of held-out *is-a* edges, generalizes to unseen subtrees, transfers across ontologies without retraining, and, combined with one relatedness score, accounts for most of the linearly decodable subsumption signal with 68 parameters. These results suggest that LLMs encode not only which concepts are related but how abstract each concept is, as a single linear coordinate. Code will be available.

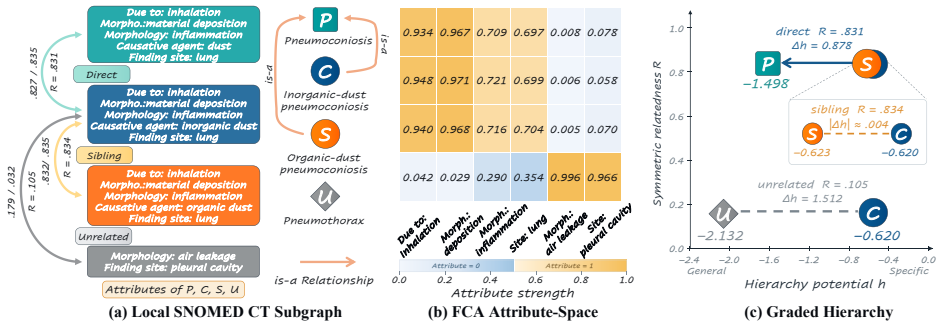
1 INTRODUCTION

Large language models (LLMs) (Achiam et al., 2023; Dubey et al., 2024; Mesnard et al., 2024) capture a remarkable amount of conceptual knowledge: what things are, how they are categorized, and how categories relate (Lin & Ng, 2022; Xiong & Staab, 2025). How this knowledge is laid out in the geometry of their representations remains only partially understood (Petroni et al., 2019; Wu et al., 2023). The *Linear Representation Hypothesis* (Mikolov et al., 2013; Park et al., 2024; Gurnee & Tegmark, 2024) is the most productive lens so far: semantic attributes correspond to directions, recoverable as embedding differences, probe weights, or steering vectors (Liu et al., 2025; 2024; Li et al., 2026; Han et al., 2024). Park et al. (2025) showed that hierarchy leaves a geometric signature, the vector from a child to its parent being orthogonal to the parent, and Xiong (2026) took the step from directions to structure: if each attribute defines a half-space, intersections of half-spaces define concepts, and the induced object–attribute relation yields a concept lattice in the sense of Formal Concept Analysis (Ganter et al., 2005) (FCA), with geometric meet and join. This *Lattice Representation Hypothesis* (LRH) is the most explicit account to date of how subsumption, intersection and generalization could be realized inside a linear representation.

Its evidence, however, rests on two choices that leave the central claim untested. First, WordNet (Miller, 1995), the testbed, records which concepts subsume which but not why; the attributes the lattice is built from were generated by another LLM (Achiam et al., 2023), so the structure being recovered is itself a model's output. Second, subsumption was evaluated against random negatives, which are almost always unrelated pairs; any similarity measure separates related from unrelated concepts, so the harder question is never asked: *given two related concepts, which is more specific?*

We test the LRH on SNOMED CT (Bos & Donnelly, 2006), a clinical ontology whose concepts carry human-curated defining attributes, so that the intents the LRH presupposes are available as ground truth. Figure 1 previews what we find. One half of the lattice picture is reliable: curated attributes are linearly recoverable as half-spaces, and the binary incidence in panel (b) is what discriminant directions reproduce on held-out concepts across three LLMs (Dubey et al., 2024; Mesnard et al.,

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.



(a) Local SNOMED CT Subgraph (b) FCA Attribute-Space (c) Graded Hierarchy

Figure 1: **Testbed on SNOMED CT.** (a) Local SNOMED CT subgraph and curated attributes: P, C, S, U represent a shared parent, anchored child, sibling, and unrelated concept, respectively. *Is-a* arrows point from child to parent. Forward/Reverse inclusion scores $\mathrm{Incl}(A, B)/\,\mathrm{Incl}(B, A)$ and symmetric relatedness R are reported with arrows. (b) Heat map of FCA attribute-space: rows correspond to P, C, S, U, with blue and yellow indicating binary attribute incidence. C and S are indistinguishable from each other and from P. (c) Graded hierarchy: a learned direction assigns each concept a scalar grade h, with more specific concepts at larger values. The grade gap Δh is large for (C, P), near zero for (C, S), and large but uninformative for (C, U), which has low R.

2024; Cho et al., 2025). The other half is not. The inclusion score $\mathrm{Incl}(A, B)$ the lattice uses to decide subsumption is nearly identical for the true child–parent pair (C, P) and the sibling pair (C, S), and nearly identical in both directions (panel (a)): it separates related from unrelated concepts but only weakly indicates which of two related concepts is more specific. We show this is structural. Incl decomposes into a global activation term plus a covariance term and is invariant to every attribute the child has that the parent lacks, so it measures *relatedness* far more than subsumption. Nor can attributes repair it: even with oracle intents, strict containment holds on only a quarter of *is-a* edges, because SNOMED's attribute values form their own hierarchy that a flat formal context cannot see. The inclusion aggregation obscures much of this information; a single linear readout from the embedding recovers it far more effectively. A learned direction $\mathbf{u}$ assigns each concept a scalar *grade* $h(\mathbf{x}) = \mathbf{u}^\top \mathbf{x}$ that increases along *is-a* edges, places siblings at the same height, and separates C from P by a margin two hundred times the sibling gap (panel (c)). The grade is stable across models, generalizes to subtrees withheld from training, and transfers between SNOMED CT and WordNet without retraining. Relatedness and grade are complementary: the unrelated pair (C, U) has a large grade difference but low relatedness, and together they recover subsumption where either alone fails.

We formalize these findings as the ***Graded Representation Hypothesis*** (GRH): in an LLM's representation space, a concept hierarchy is realized by an attribute subspace whose half-space intersections encode *relatedness* (the lattice), together with a single direction whose projection is a real-valued grading of the *is-a* poset, encoding *specificity* (the grading). A lattice tells which concepts are comparable; a grading tells which is more specific. Subsumption requires both, and two scalars carry most of it: a 68-parameter model on one relatedness score and one grade difference matches the best linear pair probe under strict subtree holdout and recovers 96% of its performance under standard cross-validation, at $1/22$ of its parameter count.

Our contributions are threefold: (i) the first test of the LRH on an ontology whose intents are human-curated ground truth, with a decomposition result showing that its inclusion score carries only weak directional information and an oracle analysis showing that flat attributes underdetermine it; (ii) the discovery and characterization of a single linear direction that grades specificity in LLM concept embeddings, stable across folds and models, approximately order-preserving and additive over the *is-a* poset, and transferable across ontologies; (iii) the GRH and a two-factor result showing that relatedness and grade together account for most of the linearly decodable subsumption signal.

Related work. Attribute directions underlie probing and steering (Park et al., 2024; Gurnee & Tegmark, 2024; Nanda et al., 2023). Work on hierarchy treats it as a relation between two concepts: an orthogonality constraint on each edge (Park et al., 2025), half-space intersections (Xiong, 2026), or a depth-specific linear map (Sakata et al., 2026); none asks whether specificity is a property

of a single concept. Encoding generality in one coordinate has been a training objective, in order embeddings (Vendrov et al., 2016), hyperbolic embeddings and entailment cones (Nickel & Kiela, 2017; Ganea et al., 2018), norm-based hypernymy models (Nguyen et al., 2017; Vulić & Mrkšić, 2018) and hierarchy-tuned LLMs (He et al., 2024); we find that a linear probe reads such a coordinate from frozen representations never trained for it. Levy et al. (2015) treated a per-word "prototypical hypernym" score as an artifact; we identify such a coordinate with a relation-supervised linear probe on frozen representations and find that it transfers across held-out branches and ontologies without retraining (Appendix K).

2 From Lattice to Grading

This section makes three claims: what the lattice's inclusion score does and does not measure (Proposition 1); what a formal context built from flat attributes can and cannot encode (Proposition 2); and the GRH together with the object it posits, a linear, real-valued grading of the *is-a* poset (Proposition 3). Proofs are in Appendix A.

2.1 Preliminaries

Concepts, attributes, and half-spaces. Let G be a finite set of concepts with embeddings $\mathbf{x}_g \in \mathbb{R}^d$ read from a frozen LLM, and M a finite set of binary attributes. The incidence $I \subseteq G \times M$ makes (G, M, I) a *formal context* (Ganter et al., 2005). The *intent* of a concept is its attribute set $Y_g = \{m : (g, m) \in I\}$; formal concepts are ordered by reverse intent inclusion, $A \sqsubseteq B$ iff $Y_B \subseteq Y_A$. The LRH posits that each attribute is a half-space $\{\mathbf{x} : \mathbf{d}_m^\top \mathbf{x} \geq \tau_m\}$; with the centered projection $q_g(m) = \mathbf{d}_m^\top \mathbf{x}_g - \tau_m$, the predicted intent is $\widehat{Y}_g = \{m : q_g(m) > 0\}$.

The inclusion score. Since predicted intents are noisy, the LRH tests $A \sqsubseteq B$ with a soft version of $Y_B \subseteq Y_A$, the proportion of B's attributes that A retains. With the parent gate $W_B(m) = \mathbb{I}[q_B(m) > 0]$ and child satisfaction $S_A(m) = \sigma(q_A(m))$,

$$\mathrm{Incl}(A, B) = \frac{\sum_m W_B(m)\, S_A(m)}{\sum_m W_B(m)} = \mathbb{E}_m[S_A(m) \mid W_B(m) = 1]\,, \tag{1}$$

where $\mathbb{E}_m$ averages uniformly over M. This is the form in the released implementation of Xiong (2026); the softplus-weighted form in their paper replaces the hard gate by nonnegative weights and performs worse on every metric (Appendix D.1). $R_{\mathrm{Incl}}(A, B) = \frac{1}{2}[\mathrm{Incl}(A, B) + \mathrm{Incl}(B, A)]$ denotes its symmetrization.

The is-a poset and gradings. The ontology supplies a second order, the transitive closure $(G, \preceq)$ of its *is-a* edges, with $c \prec p$ when c is a descendant of p and $c \lessdot p$ when (c, p) is a direct edge (a cover). A poset is *graded* if it admits a rank function with $\rho(c) = \rho(p)+1$ on every cover. SNOMED CT and WordNet permit multiple inheritance and provide no canonical unit-step rank, so we do not assume combinatorial gradedness and use a real-valued notion that drops the unit-step requirement.

Definition 1 (Grading). *$h : G \to \mathbb{R}$ is a* grading *of $(G, \preceq)$ if $h(c) > h(p)$ for every cover $c \lessdot p$; it is* linear *if $h(g) = \mathbf{u}^\top \mathbf{x}_g$ for some $\mathbf{u} \in \mathbb{R}^d$.*

Unlike a combinatorial rank function, the increment $h(c) - h(p)$ need not be constant across covers. A grading is monotone along every path, orders comparable concepts by specificity, says nothing about incomparable ones, and is unique only up to increasing transformations; an estimated h satisfies the inequality on most covers, and we report the fraction. The LRH builds its lattice from (G, M, I) and evaluates it against $(G, \preceq)$; the next two propositions show the two orders are related more loosely than this assumes.

2.2 What the Lattice Encodes

Equation equation 1 is a conditional mean and therefore decomposes.

Proposition 1 (Inclusion mainly measures relatedness). *For any A, B with $\mathbb{E}_m[W_B] > 0$,*

$$\mathrm{Incl}(A, B) = \mathbb{E}_m[S_A] + \frac{\mathrm{Cov}_m(W_B, S_A)}{\mathbb{E}_m[W_B]}. \tag{2}$$

Consequently: (i) $\mathrm{Incl}(A, B)$ *depends on* A *only through* $\{S_A(m) : W_B(m) = 1\}$*, so it is invariant to every attribute that* A *possesses and* B *lacks; (ii) if* C *and* S *agree on* B*'s gated attributes, then* $\mathrm{Incl}(C, B) = \mathrm{Incl}(S, B)$*; (iii) in the hard-threshold limit* $S_A \to W_A$*,* $\mathrm{Incl}(A, B) = |W_A \cap W_B|/|W_B|$*, so whenever* $W_A \cap W_B \neq \emptyset$*,* $\mathrm{Incl}(A, B) > \mathrm{Incl}(B, A)$ *iff* $|W_A| > |W_B|$.

The first term is A's global activation and does not involve B; the second is a covariance between the two profiles. The decomposition holds verbatim for any nonnegative parent weighting, including the softplus form, for which (i) holds approximately because attributes B lacks receive near-zero weight; the limitation is the aggregation, not the implementation. Neither term is directly sensitive to what subsumption requires. That P subsumes C means that C has every attribute of P and additional ones; by (i) the score is blind to the additional ones. A sibling that shares P's attributes, as siblings in a well-formed ontology do, receives the same score as C by (ii). This is the collision in Figure 1(a), where the forward inclusion scores of the direct and sibling pairs are 0.827 and 0.832. By (iii) the only directional information the score can carry is attribute count, which Section 4.1 shows is uninformative on most edges. Incl therefore measures whether A lies in the attribute region that B activates, a form of *relatedness*. Empirically, it separates direct pairs from unrelated pairs at AUROC 0.888 but from siblings at 0.624, and orients only 59.5% of true edges (Table 1).

2.3 What Attributes Can Encode

A natural fix is to use exact intents, for which $Y_P \subseteq Y_C$ should hold on every edge. On SNOMED CT it does not, for a structural reason. A SNOMED attribute is a typed pair $m = (r, v)$ of a relation type r (*finding site*, *causative agent*, ...) and a value v that is itself a SNOMED concept, so values carry their own *is-a* order $\preceq_V$. A child typically specializes its parent by replacing a value with a more specific one: *causative agent: dust* becomes *causative agent: inorganic dust* (Figure 1(a)). A one-valued formal context sees only whether (r, v) is present. FCA has a standard remedy, *conceptual scaling* of a many-valued context (Ganter et al., 2005), which adds a column for every ancestor of a value; the point of the proposition is that scaling requires the value order as input.

Proposition 2 (Flat contexts are blind to value order)**.** *Let* $M = \mathcal{R} \times V$ *with* V *partially ordered by* $\preceq_V$*, and let* $c \lessdot p$ *be an edge on which* c *refines a value:* $(r, v) \in Y_c$*,* $(r, v') \in Y_p$*,* $v \prec_V v'$*,* $(r, v') \notin Y_c$*,* $(r, v) \notin Y_p$*. Then* Y_c *and* Y_p *are incomparable in* (G, M, I)*, and any score that depends on the intents only through* $|Y_c \cap Y_p|$*,* $|Y_c \setminus Y_p|$ *and* $|Y_p \setminus Y_c|$*, including equation 1 on exact intents, assigns the edge the value of any incomparable pair with the same counts. In the value-closed context* $I^{\uparrow}$*, where* $(g, (r, a)) \in I^{\uparrow}$ *iff* $(g, (r, v)) \in I$ *for some* $v \preceq_V a$*,* $Y_p^{\uparrow} \subseteq Y_c^{\uparrow}$ *whenever every attribute of* p *is kept or refined by* c*.*

The proposition is elementary, but it changes how the weak edge orientation should be read. It is not solely a decoding failure: even with curated intents, strict containment holds on 26.7% of SNOMED edges and 65.3% are incomparable, almost all through clean value specializations (Section 4.1). Closing the context along $\preceq_V$, which is ordinal scaling, restores containment on every edge, but requires the value hierarchy as input, which is the hierarchy the lattice was meant to recover. Set inclusion over flat intents therefore cannot orient most edges, however well the intents are decoded. The profile retains some directional signal (a logistic regression over it orients 83% of edges, Appendix D.2), but Section 4.2 shows that one linear readout of the embedding recovers it better.

2.4 The Graded Representation Hypothesis

A linear grade from pair orientation. Consider the simplest supervised task that isolates direction: given a known edge, decide whether it is presented as (C, P) or (P, C). We fit a linear classifier without intercept on the embedding difference, $D(C, P) = \mathbf{w}^\top \mathbf{\Phi}(\mathbf{x}_C - \mathbf{x}_P)$, with $\mathbf{\Phi}$ a fixed linear map (a label-free PCA projection to k dimensions, Section 3). Although fitted on pairs, the classifier is equivalent to a scoring of single concepts.

Proposition 3 (Pair orientation induces a linear grading)**.** *Let* $D(C, P) = \mathbf{w}^\top \mathbf{\Phi}(\mathbf{x}_C - \mathbf{x}_P)$ *with* $\mathbf{\Phi}$ *linear and no intercept, and set* $\mathbf{u} = \mathbf{\Phi}^\top \mathbf{w}$*,* $h(g) = \mathbf{u}^\top \mathbf{x}_g$*. Then (i)* $D(C, P) = h(C) - h(P) = -D(P, C)$*; (ii) if* $D(c, p) > 0$ *on every cover,* h *is a linear grading of* $(G, \preceq)$*; (iii) along any path* $c = g_0 \lessdot \cdots \lessdot g_k = a$ *with* $k \geq 2$*,* $h(c) - h(a) = \sum_{i<k} [h(g_i) - h(g_{i+1})]$*, so the gap to a non-direct*

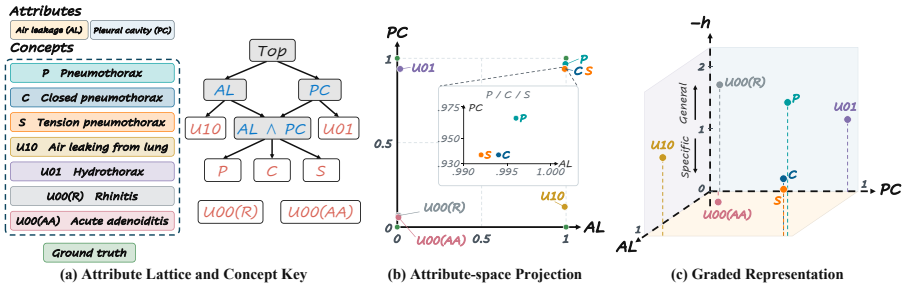


Figure 2: **From lattice to grade.** (a) Concept key and the attribute lattice generated by *air leakage* (AL) and *pleural cavity* (PC): *pneumothorax* (P) and its children C and S share the node $\mathrm{AL} \wedge \mathrm{PC}$; *rhinitis* and *acute adenoiditis* share the top. (b) Learned attribute scores recover every concept's lattice cell (green points are the binary ground truth) but leave P, C and S within 0.01 of one another. (c) Adding the grade h as a third axis separates the parent from its more specific children and orders the top-cell concepts by their own generality; the lattice coordinates are unchanged.

ancestor is the sum of the gaps along the path and exceeds the gap to the direct parent g_1 whenever the remaining gaps are positive.

We call $h(g)$ the *grade* of a concept and $\mathbf{u}$ the *specificity direction*; the appendix uses *hierarchy potential* for the same object. Property (iii) should be read as a prediction rather than a shortcoming: a grading measures how far along the hierarchy a concept sits, not whether two concepts are adjacent, so a grandparent should receive a larger gap than a parent.

The hypothesis. Figure 2 places the two structures side by side. In panel (b) the learned half-spaces put every concept in its correct lattice cell, and the three concepts sharing $\mathrm{AL} \wedge \mathrm{PC}$ land within 0.01 of one another. This is the intended behavior of the lattice, and it is also the reason the lattice cannot order them. Panel (c) adds the grade as a third axis: the parent sits above its children, the children sit at the same height, and the concepts the lattice cannot distinguish at the top are ordered by their own generality. The lattice determines which concepts are comparable, and the grade determines which is more specific.

Hypothesis 1 (Graded Representation Hypothesis)**.** *In the representation space of an LLM, a concept hierarchy $(G, \preceq)$ is realized by two complementary linear structures: (i) an attribute subspace whose half-space intersections form a concept lattice, in which inclusion measures* relatedness *$R(A, B)$; and (ii) a direction $\mathbf{u}$ whose projection $h(g) = \mathbf{u}^\top \mathbf{x}_g$ is a linear grading of $(G, \preceq)$, encoding* specificity*. Subsumption $A \preceq B$ is decodable from $R(A, B)$ and $\Delta h(A, B) = h(A) - h(B)$: it holds when A and B are related and $h(A) > h(B)$.*

The hypothesis makes three testable commitments. Grade is a property of a concept, so one $\mathbf{u}$ fitted on some edges must orient edges in unseen subtrees and transfer between ontologies. Grade is a grading, so h must be order-preserving along edges, additive along paths, and correlated with depth. And the two structures are jointly sufficient, so a classifier on $(R, \Delta h)$ must recover subsumption where either scalar fails and approach classifiers that see the whole embedding. We instantiate the last as the *two-factor model*, a logistic regression on $[R,\ \Delta h,\ R \cdot \Delta h]$ with $R \in \{R_{\mathrm{Incl}}, R_{\cos}\}$ and $R_{\cos}(A, B) = \cos(\mathbf{x}_A, \mathbf{x}_B)$. Counting the $k = 64$ coefficients that define $\mathbf{u}$, it has 68 hierarchy-supervised parameters; the interaction term lets the grade gap count only when the pair is related.

3 Experimental Setup

Data. We use the respiratory-disorder subtree of SNOMED CT (July 2026 International Edition): 4,253 concepts under *Disorder of respiratory system*, of whose internal *is-a* edges 7,153 serve as reference edges. A concept's non-hierarchical relationships define its curated attributes, exact (relation type, destination) pairs, 1,506 in all. Because 45% of them have a single positive concept, direction fitting uses the 105 attributes with at least 20 positive and negative concepts (HS20; HS50 with 51 attributes as a check), and inclusion experiments use a matched 104-attribute subset. For

transfer we use the 82,115 noun synsets and 75,831 hypernym edges of WordNet 3.0 (Miller, 1995). Details are in Appendix B.

Candidates and tasks. For each reference edge $C \lessdot P$ we form the forward pair, the reversed pair, a *sibling* (shares a direct parent, incomparable to C), an *unrelated* concept (neither comparable nor sibling), a *depth-matched non-ancestor* (same shortest root depth as P), and, when it exists, the *grandparent*. Random negatives, the setting of Xiong (2026), are almost always unrelated pairs; the reversed, sibling and depth-matched candidates are the hard negatives that separate direction from relatedness. Four tasks: *attribute recovery* (predict incidence from the embedding); *edge orientation* (which endpoint of a known edge is the parent); *two-hop subsumption*, the primary relation task (parents and grandparents positive, all else negative; 14,253 vs. 28,559); and *Hasse adjacency* (only the direct parent positive; 7,153 vs. 35,659).

Protocols. Attribute recovery uses a 90/10 random object split (3,827/426 concepts). *Parent-grouped five-fold cross-validation* keeps all candidates of an edge, and all edges sharing a parent, in one fold. *Complete-branch holdout* withholds five descendant branches in turn (respiratory tumor, upper-respiratory disease, lower-respiratory disease, respiratory infection, asthma), excluding every held-out concept from all fitting, including PCA, attribute directions and thresholds; this is the protocol under which a direction must generalize to a region of the ontology it has never seen. Branches overlap through multiple inheritance, so their standard deviations describe branch variation. *Cross-ontology transfer* fits $\mathbf{u}$ on all edges of one ontology and applies it to the other with no recalibration of sign, scale or threshold. Metrics are macro-averaged F1 and AUROC over attributes; pair accuracy (fraction of edges with $D(C, P) > 0$) and orientation AUROC; and AUROC and AUPRC for relation tasks, with per-negative-type AUROC. All thresholds and regularization strengths are selected on training data only.

Models and estimators. Llama-3.1-8B (Dubey et al., 2024) is primary; KoBioMed-Llama-3.1-8B (Cho et al., 2025) and Gemma-2-9B (Mesnard et al., 2024) replicate every principal experiment with identical data, splits and grids. Following Xiong (2026), token representations are mean-pooled within each synonym and averaged across synonyms. Attribute directions are regularized Fisher discriminants with Ledoit–Wolf covariances (Fisher, 1936; Ledoit & Wolf, 2004) and midpoint thresholds. The grade uses label-free PCA with $k = 64$ components, one global RMS scaling, and a no-intercept L2 logistic regression on mirrored differences, mapped back as $\mathbf{u} = \mathbf{P}_k^\top \mathbf{w}/s$ (Appendix E.1). Capacity-matched references are L2 logistic regressions on the attribute-profile pair $[\tilde{\mathbf{q}}_A; \tilde{\mathbf{q}}_B; \tilde{\mathbf{q}}_A \odot \tilde{\mathbf{q}}_B]$, where $\tilde{\mathbf{q}}$ is the standardized attribute projections and on PCA-k pairs $[\mathbf{z}_A; \mathbf{z}_B; \mathbf{z}_A \odot \mathbf{z}_B]$ ($3k + 1$ coefficients), with label-shuffled controls (Appendix G.1).

4 Results

4.1 Attributes Are Linearly Encoded but Underdetermine Direction

Half-spaces are real. Table 7 (Appendix C) reports attribute recovery on human-curated intents. Every attribute with adequate support is a linear half-space: directions fitted on 3,827 concepts classify the 426 held-out concepts at macro AUROC above 0.98 (macro F1 0.81–0.87) on both panels and all three encoders, and per-attribute AUROCs correlate at 0.97–0.99 between encoders. With profiles permuted across concepts, or the concept–attribute graph rewired under fixed degrees, the estimator still fits the training set but held-out AUROC falls to 0.49–0.52, and deleting attribute phrases from concept names leaves AUROC at 0.94–0.96 (Appendix C), so neither capacity nor lexical overlap explains the result.

The inclusion score measures relatedness. With a train-calibrated threshold, the released Incl reaches direct-parent macro F1 0.850 against matched non-ancestor negatives, close to the 0.863 ceiling from oracle binary attributes on the same panel; the original's fixed threshold of 0.6 gives 0.746 (Appendix D.1). Table 1 examines what the score captures. It separates direct from unrelated pairs (AUROC 0.888), but 71.0% of sibling pairs pass its threshold, direct-versus-sibling AUROC is 0.624, and 78.3% of true edges pass in *both* directions. On orientation it is near chance: forward exceeds reverse on 59.5% of edges (AUROC 0.556). The negative-evidence variant, which tests $Y_A^c \subseteq Y_B^c$, reaches 65.2%, but its directional difference correlates with the difference in attribute

Table 1: **The inclusion score measures relatedness.** Left: released-code Incl on 7,153 matched (direct, sibling, unrelated) triplets, Llama-3.1-8B; "Pass" is the fraction above the train-calibrated threshold 0.360. Right: AUROC of direct pairs against each candidate type and orientation of true edges, for the released score and its negative-evidence variant (Appendix D.3).

Pair type	Forward	Reverse	Pass
Direct (C, P)	0.661	0.608	86.2%
Sibling (C, S)	0.536	0.545	71.0%
Unrelated (C, U)	0.190	0.179	10.4%

	Incl	Negative-evidence
Direct vs. unrelated (AUROC)	0.888	0.867
Direct vs. sibling (AUROC)	0.624	0.664
Direction accuracy	59.5%	65.2%
Orientation AUROC	0.556	0.629

Table 2: **Curated attributes underdetermine direction.** Intent relations under the full 1,506-attribute oracle on all 7,352 internal direct edges and 41,532 sibling pairs, before and after closing attribute values along their own *is-a* order. Oracle recovery classifies direct parents with $|Y_C \cap Y_P|/|Y_P|$ and a train-selected threshold.

	Direct edges (C, P)			Sibling pairs		Oracle recovery	
Incidence	$Y_P \subsetneq Y_C$	$Y_P = Y_C$	Incomparable	Comparable	Incomparable	F1	AUROC
ExactFlat	26.7%	8.0%	65.3%	7.2%	90.6%	0.922	0.936
DestinationClosure	91.9%	8.1%	0.0%	11.8%	86.1%	0.996	0.996

counts at Spearman 0.771, consistent with Proposition 1(iii). Appendix D.2 approaches the same question from the representation side: given a known edge, Incl orients 59.5%, the signed attribute-count difference 64.3%, a logistic regression over the full 104-dimensional attribute profile 83.1%, and one over 64 PCA coordinates of the raw embedding above 94%. Orientation information is lost progressively from the embedding to the attribute profile to the aggregated score.

Attributes underdetermine direction. Table 2 uses the curated incidence directly, removing prediction error. A small panel loses identity information: the 51 best-supported attributes compress 4,253 concepts into 602 intents, and a quarter of sibling pairs share one (Appendix J). The full vocabulary resolves identities (3,635 intents) but does not make intents a partial order: strict containment holds on 26.7% of edges and 65.3% are incomparable, the situation of Proposition 2. Every incomparable edge involves a same-type value replacement, on 71.2% it is the *only* difference, and on 96.3% every replacement is a clean specialization of the parent's value, concentrated in *finding site* and *associated morphology*. Closing values along their own hierarchy restores $Y_P \subseteq Y_C$ on every edge and lifts oracle F1 from 0.922 to 0.996 without collapsing concepts (equal-intent siblings rise from 2.12% to 2.13%). Two conclusions follow. Direction is recoverable from attributes only if the value hierarchy is supplied, so a flat lattice cannot recover it in principle. And even the closed oracle is not a relation detector: 11.8% of sibling pairs become strictly comparable, so inclusion is necessary for subsumption but not sufficient.

4.2 A Single Direction Grades the Hierarchy

One direction orients 94% of edges. Table 3 shows that the orientation information the inclusion score loses is linearly recoverable from the embedding with few supervised coefficients. With 64 coefficients the grade correctly orients 94.2% of held-out edges under parent-grouped folds; 16 coefficients already give 85%, and 4,096 add only 2.5 points. Directions fitted in different folds agree at mean pairwise cosine 0.963. Under complete-branch holdout the learned direction still orients 90.3% of edges (87.8–93.8% across branches), while shuffled labels return to chance, a sign-selected random direction reaches 58%, and the first principal component 71%. The grade is thus aligned with but not identical to the dominant variance direction, nor to the mean child-to-parent offset (cosine 0.44 on held-out branches, Appendix E.3). All three encoders recover the same structure, 93–94% under parent-grouped folds and 90–92% under branch holdout, with the medically continued-pretrained model offering no consistent gain (Appendix I).

It behaves as a grading. The direction is trained only to orient direct edges; Table 11 (Appendix F) tests it on candidates and statistics it never saw under branch holdout. The grandparent result follows

Table 3: **One direction orients the hierarchy.** Orientation of known direct edges (mean ± s.d.). Top: supervised capacity under parent-grouped folds, Llama-3.1-8B. Middle: controls under complete-branch holdout. Bottom: other encoders under complete-branch holdout.

Direction	Pair accuracy	Orientation AUROC
Supervised capacity, parent-grouped folds		
PCA-16 (16 coefficients)	0.853 ± 0.019	0.927 ± 0.014
PCA-64 (64 coefficients)	0.942 ± 0.009	0.982 ± 0.004
PCA-256 (256 coefficients)	0.956 ± 0.008	0.989 ± 0.002
Raw-4096 (4,096 coefficients)	0.967 ± 0.006	0.993 ± 0.002
Controls, complete-branch holdout, PCA-64		
Shuffled-label logistic regression (20 repeats)	0.506 ± 0.023	0.508 ± 0.029
Random sign-selected direction (100 draws)	0.580 ± 0.013	0.602 ± 0.015
Embedding PC1, sign-selected	0.714 ± 0.049	0.760 ± 0.061
Learned $\mathbf{u}$	0.903 ± 0.022	0.964 ± 0.014
Other encoders, complete-branch holdout, learned $\mathbf{u}$		
KoBioMed-Llama-3.1-8B	0.900 ± 0.022	0.966 ± 0.012
Gemma-2-9B	0.922 ± 0.010	0.971 ± 0.003

Proposition 3(iii): in 84% of matched cases the gap to the grandparent exceeds the gap to the parent (AUROC 0.278), and the mean grandparent gap is 1.8 times the parent gap. This would count against an adjacency detector, but it is the expected behavior of a grading. Siblings sit at the same height: the parent is preferred over a sibling in 90% of cases, which requires only $h(S) > h(P)$ since $D(C, P) - D(C, S) = h(S) - h(P)$, and sibling gaps are centered at zero. The unrelated and depth-matched contrasts are weak (0.709, 0.676), as they should be: a grading orders every pair whether or not it is related, so the grade alone cannot tell whether an edge exists; that is the role of the relatedness factor. Finally, h correlates with depth without encoding it exactly: its Spearman correlation with depth on unseen nodes is positive in every branch (0.38–0.53), and among pairs with no ancestral relation the probability that the deeper one is graded higher rises from chance to 0.75 at a gap of three, then plateaus (Figure 3). The grade is a real-valued grading in the sense of Definition 1: an approximately order-preserving coordinate on the poset, not an exact encoding of depth.

It transfers across ontologies. A direction fitted on SNOMED CT respiratory disorders, with no WordNet label used to set its sign, scale or threshold, orients held-out WordNet branches (*animal*, *plant*, *artifact*, *cognition*) at AUROC 0.877 and accuracy 0.807; a direction fitted on the other WordNet branches reaches 0.929 and 0.858. On all 75,831 WordNet hypernym edges the SNOMED direction reaches 0.839. The reverse transfer reaches 0.790 on SNOMED branches and 0.813 on all SNOMED edges. Restricting to WordNet edges whose lemmas never occur in training lowers the transferred AUROC by only 0.02, so the signal is not lexical memorization (Appendix H). Within-ontology fitting is better on both sides, so the transferable component is shared but not the whole of $\mathbf{u}$; the grade is not specific to clinical text.

4.3 Relatedness and Grade Account for Subsumption

Two scalars recover subsumption. Table 4 tests the joint-sufficiency commitment. Every relatedness score alone, whether the released Incl, its symmetrization or raw cosine, stays near 0.6 AUROC: with reversed and sibling candidates among the negatives, relatedness is nearly uninformative. The grade gap alone reaches 0.857 under parent-grouped folds and 0.837 on unseen branches. Combining them improves on both: the cosine two-factor model reaches 0.919 and 0.887 with 68 hierarchy-supervised parameters. Under the strict protocol it matches the PCA-64 pair probe (0.886, 193 coefficients) and exceeds the PCA-512 probe (0.877, 1,537 coefficients) with smaller branch-to-branch variance; under parent-grouped folds it recovers 96% of the best probe's AUROC at $1/22$ of its parameters. Table 5 shows why the factors are complementary rather than redundant. Cosine alone is below chance against reversed and sibling candidates, because a grandparent is on average less similar to the child than a reversed or sibling pair, and it is strong only against unrelated and depth-matched ones. The grade alone nearly resolves reversal (0.976) and siblings (0.858) but is

Table 4: **Relatedness and grade account for most of the decodable subsumption signal.** Two-hop subsumption (parents and grandparents positive; reversed, sibling, unrelated and depth-matched candidates negative), Llama-3.1-8B, mean ± s.d. over five parent-grouped folds or five held-out branches. "Coef." counts hierarchy-supervised coefficients, including the 64 that define $\mathbf{u}$; dashes denote models not repeated under the stricter protocol.

		Parent-grouped folds		Held-out branches	
Method	Coef.	AUROC	AUPRC	AUROC	AUPRC
Released Incl	0	0.650 ± 0.017	0.420 ± 0.015	0.573 ± 0.021	0.375 ± 0.017
Symmetric R_{Incl}	0	0.632 ± 0.009	0.416 ± 0.005	0.572 ± 0.015	0.385 ± 0.009
Cosine $R_{\cos}$	0	0.638 ± 0.006	0.412 ± 0.003	0.609 ± 0.026	0.393 ± 0.012
Δh	64	0.857 ± 0.004	0.697 ± 0.005	0.837 ± 0.033	0.671 ± 0.067
$R_{\text{Incl}} + \Delta h$	68	0.907 ± 0.007	0.810 ± 0.014	0.847 ± 0.030	0.702 ± 0.053
$R_{\cos} + \Delta h$	68	0.919 ± 0.005	0.838 ± 0.008	$\mathbf{0.887 \pm 0.021}$	$\mathbf{0.787 \pm 0.043}$
Attribute-profile pair	313	0.879 ± 0.007	0.729 ± 0.018	–	–
PCA-64 pair	193	0.936 ± 0.005	0.869 ± 0.006	0.886 ± 0.033	0.777 ± 0.073
PCA-512 pair	1,537	$\mathbf{0.954 \pm 0.002}$	$\mathbf{0.898 \pm 0.004}$	0.877 ± 0.038	0.759 ± 0.071
PCA-64 pair, shuffled labels	193	0.508 ± 0.019	0.342 ± 0.019	0.498 ± 0.059	0.334 ± 0.034

Table 5: **Complementary failure modes.** AUROC of two-hop positives against each negative type under complete-branch holdout (Llama-3.1-8B, mean over five branches).

Method	Reverse	Sibling	Unrelated	Depth-matched
Cosine $R_{\cos}$	0.389	0.450	0.805	0.777
Δh	**0.976**	0.858	0.771	0.746
$R_{\text{Incl}} + \Delta h$	0.949	0.820	0.822	0.798
$R_{\cos} + \Delta h$	0.958	0.843	0.884	0.861
PCA-64 pair	0.933	**0.860**	**0.888**	**0.866**

weakest against unrelated and depth-matched candidates. The combined model retains the strengths of both and tracks the 193-coefficient reference within 0.02 on every contrast. All three encoders give two-factor AUROCs of 0.887–0.896 under branch holdout (Appendix I).

Capacity and adjacency. More capacity is not better. Two-hop AUROC of PCA pair probes peaks at PCA-512 under parent-grouped folds and falls to 0.896 at PCA-4096, and under branch holdout PCA-512 loses more than PCA-64 relative to its cross-validated value, because later principal coordinates carry branch-specific information (Appendix G.2). For the same reason parameter-free cosine transfers better than attribute-based relatedness and is our primary variant. On Hasse adjacency, where the grandparent is a negative, the two-factor model reaches 0.827 under parent-grouped folds and 0.796 under branch holdout against 0.814 and 0.777 for the PCA-64 probe, but its direct-versus-grandparent contrast is only 0.571 and 0.536 (Appendix G.3). This again reflects Proposition 3(iii): a grading orders ancestors by distance and cannot say whether one is a single edge away. The GRH is a hypothesis about subsumption, and adjacency requires information that neither factor provides. Two caveats apply: the references are linear pair probes, so a nonlinear probe may set a higher bar, and under parent-grouped folds a 0.035 gap to the best probe remains.

5 Conclusion

We tested the LRH on an ontology whose intents are human-curated and found that only part of it holds. Half-spaces are recovered reliably, but the inclusion score is invariant to every attribute the child has that the parent lacks and so measures relatedness far more than subsumption; better attributes cannot repair it, because SNOMED's attribute values are themselves hierarchical. A single linear direction of the embedding recovers the missing orientation: a grade that correctly orients 94% of held-out *is-a* edges, is additive along paths, and transfers between SNOMED CT and WordNet without retraining. The GRH says a concept hierarchy is realized by a relatedness lattice together with a linear grading; a 68-parameter model on these two scalars matches linear pair probes with an order of magnitude more parameters under subtree holdout.

AI USE STATEMENT

We used generative AI tools in three ways. In the early stage of the project, we used LLM assistants to survey and understand the literature on the Linear and Lattice Representation Hypotheses; every source they pointed to was read and verified by the authors, and all citations were checked manually. We also used them to assist in editing and polishing the text of the manuscript. In the final stage, we used Codex to organize the submitted code and to check that it contained no identifying or otherwise sensitive information; two authors then independently re-checked the cleaned code. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All experiments use publicly available resources: the SNOMED CT, WordNet 3.0, and the open-weight encoders Llama-3.1-8B, KoBioMed-Llama-3.1-8B and Gemma-2-9B. Appendix B specifies the cohort construction, the atomic-attribute definition, the attribute panels and the operational definitions of every candidate type and evaluation protocol, including the five held-out branches. Appendices C, D, E and G give the estimators, hyperparameter-selection rules and control constructions for attribute recovery, inclusion scores, the grade and the two-factor model, and Appendices H, I and J give the corresponding details for cross-ontology transfer, cross-model replication and the oracle analyses. Proofs of Propositions 1–3 are in Appendix A.

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.

L Bos and K Donnelly. Snomed-ct: The advanced terminology and coding system for ehealth. *Stud Health Technol Inform*, 121:279–290, 2006.

Hyeongmin Cho, Inhu Kim, Donghyoung Lee, Sanghwan Kim, Dalho Park, Inchul Kang, Kyul Kim, Jihoon Cho, and Jongbeom Park. Kobiomed-llama-3.1-8b. 2025. URL `https://huggingface.co/Lowenzahn/KoBioMed-Llama-3.1-8B`.

Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Ramakrishna Vedantam. Hyperbolic image-text representations. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

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.

Joshua Engels, Eric J. Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *International Conference on Learning Representations (ICLR)*, 2025. URL `https://arxiv.org/abs/2405.14860`.

Ronald A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7 (2):179–188, 1936.

Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 1646–1655, 2018.

Bernhard Ganter, Gerd Stumme, and Rudolf Wille. *Formal concept analysis: foundations and applications*, volume 3626. springer, 2005.

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.

Md Rakibul Haque, KM Arefeen Sultan, Tushar Kataria, and Shireen Elhabian. MedConcept: Unsupervised concept discovery for interpretability in medical VLMs. *arXiv preprint arXiv:2604.11868*, 2026.

Yuan He, Zhangdie Yuan, Jiaoyan Chen, and Ian Horrocks. Language models as hierarchy encoders. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

Nhat Minh Le, Ciyue Shen, Neel Patel, Chintan Shah, Darpan Sanghavi, Blake Martin, Alfred Eng, Daniel Shenker, Harshith Padigela, Raymond Biju, Syed Ashar Javed, Jennifer Hipp, John Abel, Harsha Pokkalla, Sean Grullon, and Dinkar Juyal. Learning biologically relevant features in a pathology foundation model using sparse autoencoders. *NeurIPS 2024 Workshop on Advancements in Medical Foundation Models*, 2024.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

Omer Levy, Steffen Remus, Chris Biemann, and Ido Dagan. Do supervised distributional methods really learn lexical inference relations? In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 970–976, 2015.

Mingjie Li, Rui Liu, Zeyi Shi, Mingfei Han, Lina Yao, Zhihui Li, Xiaojun Chang, Kilian M Pohl, Md Tauhidul Islam, and Lei Xing. Advancing in-context learning for efficient and stable medical report generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

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.

Sheng Liu, Haotian Ye, Lei Xing, and James Y. Zou. In-context vectors: Making in context learning more effective and controllable through latent space steering. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=dJTChKgv3a.

Sheng Liu, Haotian Ye, and James Y Zou. Reducing hallucinations in large vision-language models via latent space steering. In *International Conference on Learning Representations*, volume 2025, pp. 72402–72419, 2025.

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.

Tomás Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pp. 746–751. The Association for Computational Linguistics, 2013.

George A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11): 39–41, 1995.

Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models. *arXiv preprint arXiv:2309.00941*, 2023.

Kim Anh Nguyen, Maximilian Köper, Sabine Schulte im Walde, and Ngoc Thang Vu. Hierarchical embeddings for hypernymy detection and directionality. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 233–243, 2017.

Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Avik Pal, Max van Spengler, Guido Maria D'Amely di Melendugno, Alessandro Flaborea, Fabio Galasso, and Pascal Mettes. Compositional entailment learning for hyperbolic vision-language models. In *International Conference on Learning Representations (ICLR)*, 2025.

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.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. Language models as knowledge bases? In *EMNLP/IJCNLP (1)*, pp. 2463–2473. Association for Computational Linguistics, 2019.

Masaki Sakata, Benjamin Heinzerling, Takumi Ito, Sho Yokoi, and Kentaro Inui. Linear representations of hierarchical concepts in language models. *arXiv preprint arXiv:2604.07886*, 2026.

Gemma Team. Gemma. 2024. doi: 10.34740/KAGGLE/M/3301. URL `https://www.kaggle.com/m/3301`.

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. In *International Conference on Learning Representations (ICLR)*, 2016.

Ivan Vulić and Nikola Mrkšić. Specialising word vectors for lexical entailment. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 1134–1145, 2018.

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 *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=5K1FG92m5s`.

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 PROOFS

Proof of Proposition 1. Write $\mathbb{E}_m$ for the uniform average over M. Since W_B is binary, $\sum_m W_B(m)S_A(m)/\sum_m W_B(m) = \mathbb{E}_m[W_B S_A]/\mathbb{E}_m[W_B] = \mathbb{E}_m[S_A \mid W_B = 1]$, and $\mathbb{E}_m[W_B S_A] = \mathbb{E}_m[W_B]\,\mathbb{E}_m[S_A] + \mathrm{Cov}_m(W_B, S_A)$. Dividing by $\mathbb{E}_m[W_B] > 0$ gives equation 2. The same computation applies to any nonnegative weighting $w_B(m) \geq 0$ in place of $W_B(m)$, in particular to the softplus weights of the paper form, giving $\mathbb{E}_m[S_A] + \mathrm{Cov}_m(w_B, S_A)/\mathbb{E}_m[w_B]$. (i) The numerator of equation 1 sums $S_A(m)$ only over m with $W_B(m) = 1$, so changing $S_A(m)$ at any m with $W_B(m) = 0$, in particular at any attribute on which A is positive and B is not, leaves the score unchanged. (ii) is (i) applied to C and S. (iii) Substituting $S_A = W_A$ and identifying a binary function with its support gives $\mathrm{Incl}(A, B) = |W_A \cap W_B|/|W_B|$ and $\mathrm{Incl}(B, A) = |W_A \cap W_B|/|W_A|$; when $W_A \cap W_B \neq \emptyset$, the first exceeds the second iff $|W_A| > |W_B|$. □

Proof of Proposition 2. By hypothesis $(r, v') \in Y_p \setminus Y_c$ and $(r, v) \in Y_c \setminus Y_p$, so neither intent contains the other. A score that depends on (Y_c, Y_p) only through $|Y_c \cap Y_p|$, $|Y_c \setminus Y_p|$ and $|Y_p \setminus Y_c|$ takes the same value on every pair of intents with those counts, whether or not the pair is an edge; in the hard-threshold limit, equation 1 on exact intents equals $|Y_c \cap Y_p|/|Y_p|$ and is such a score. For the closed context, let $(r, a) \in Y_p^{\uparrow}$, so $(r, v') \in Y_p$ for some $v' \preceq_V a$. By assumption c keeps (r, v') or refines it to some (r, v) with $v \preceq_V v'$; in either case c carries (r, v'') with $v'' \preceq_V v' \preceq_V a$, hence $(r, a) \in Y_c^{\uparrow}$ by transitivity of $\preceq_V$. □

Proof of Proposition 3. (i) By linearity of $\mathbf{\Phi}$, $\mathbf{w}^\top \mathbf{\Phi}(\mathbf{x}_C - \mathbf{x}_P) = (\mathbf{\Phi}^\top \mathbf{w})^\top \mathbf{x}_C - (\mathbf{\Phi}^\top \mathbf{w})^\top \mathbf{x}_P = h(C) - h(P)$, and exchanging the arguments negates the difference. When $\mathbf{\Phi}$ is a centered and scaled PCA projection, $\mathbf{\Phi}(\mathbf{x}) = \mathbf{P}_k(\mathbf{x} - \boldsymbol{\mu})/s$, the centering cancels in the difference, so $\mathbf{\Phi}$ acts on differences as $\mathbf{P}_k/s$ and $\mathbf{u} = \mathbf{P}_k^\top \mathbf{w}/s$. (ii) $D(c, p) > 0$ on every cover is, by (i), $h(c) > h(p)$

Table 6: Attribute panels used in the experiments.

Panel	Selection rule	Dim.
ExactFlat-All	All exact atomic attributes observed in the cohort	1,506
Random-HS20	Balanced support of at least 20 and evaluable under the fixed random split	105
Random-HS50	Balanced support of at least 50 and evaluable under the fixed random split	51
Matched-HS20	Matched 104-attribute subset of Random-HS20 for attribute-profile comparisons	104
BranchTrain-HS20	Selected in each branch-holdout run from training-branch support only	62–107

on every cover, which is Definition 1; linearity of h is immediate. (iii) The sum telescopes to $h(g_0) - h(g_k)$; for $k \geq 2$ it has at least two summands, and if every summand is positive it exceeds its first term $h(c) - h(g_1)$, the gap to the direct parent. □

B Dataset Construction and Evaluation Conventions

Concept graph and text. We use the July 2026 release of the SNOMED CT International Edition and its active inferred relationship snapshot. Starting from `50043002 |Disorder of respiratory system|`, we retain the root and all descendants (4,253 concepts). An *is-a* edge is kept only when both endpoints belong to the cohort, so parents, ancestors and depth below refer to this induced graph, not to the full SNOMED CT hierarchy. Concept texts are the active English preferred and acceptable synonyms of the US English reference set, excluding fully specified names. Following Xiong (2026), texts are deduplicated within a concept, token representations are mean-pooled, and synonym representations are averaged to one embedding per concept.

Atomic attributes. A non-hierarchical relationship is represented as the exact pair $m = (\mathrm{typeId}, \mathrm{destinationId})$, with $I(g, m) = 1$ when the snapshot records such a relationship from g. *Is-a* relations are excluded from the attribute matrix. Destinations need not belong to the cohort; role-group identity is ignored, and repeated relationships with the same source, type and destination collapse to one incidence. An unrecorded relationship is an operational negative, not a clinical negation.

Attribute panels. The exact vocabulary is long-tailed: of 1,506 attributes, 684 (45.4%) have a single positive concept, 1,398 (92.8%) fewer than 20, and only 51 (3.4%) at least 50. Table 6 lists the panels used; support-based selection under the random object split is distinguished from selection using training branches only.

Candidate relations. For a direct edge from child c to parent p, the *direct* pair is (c, p) and the *reverse* is (p, c). A *sibling* shares at least one direct parent with c and is not comparable to it. A *grandparent* is reached by a path $c \to p \to g$. An *unrelated* candidate is neither comparable to c nor a sibling; a *depth-matched non-ancestor* additionally has the same shortest root depth as p. These are graph categories: *unrelated* does not imply semantic unrelatedness. Grandparents are positives for two-hop subsumption and negatives for Hasse adjacency. Under the parent-grouped protocol the adjacency task has 7,153 positives and 35,659 negatives; relabeling grandparents gives 14,253 positives and 28,559 negatives for two-hop subsumption. Because class priors differ, AUPRC is comparable only within a task and protocol.

Protocols. Attribute recovery uses a fixed 90/10 random split of concepts (3,827/426). Hierarchy analyses use five-fold cross-validation grouped by direct-parent identifier, with candidate groups of the same reference edge kept in one fold. The strictest protocol holds out five complete descendant branches in turn (respiratory-system tumor, upper-respiratory disease, lower-respiratory disease, respiratory infection, asthma): every concept of the branch is excluded from PCA, attribute directions, classifier parameters and thresholds, and edges crossing the boundary are omitted. Because of mul-

Table 7: Held-out recovery of exact SNOMED CT attributes, macro-averaged, for three encoders; and label-randomization controls on Matched-HS20 (Llama-3.1-8B, means over seeds).

		HS20 (105 attributes)		HS50 (51 attributes)	
Labels	Encoder	Macro F1	Macro AUROC	Macro F1	Macro AUROC
Original	Llama-3.1-8B	0.827	0.993	0.862	0.988
Original	KoBioMed-Llama-3.1-8B	0.842	0.993	0.874	0.986
Original	Gemma-2-9B	0.814	0.994	0.868	0.989
Profile permutation	Llama-3.1-8B	0.026	0.494	–	–
Degree-preserving edge swap	Llama-3.1-8B	0.033	0.522	–	–

tiple inheritance the branches overlap, so their standard deviations describe branch variation rather than independent folds. Identifiers, seeds and branch-specific counts are provided with the code.

C Linear Recovery of Medical Attributes

Estimator. For each attribute m we fit a regularized Fisher discriminant on training concepts only (Fisher, 1936; Ledoit & Wolf, 2004). With class means $\boldsymbol{\mu}_m^{\pm}$ and Ledoit–Wolf covariances $\boldsymbol{\Sigma}_m^{\pm}$,

$$\mathbf{d}_m = \mathrm{normalize}\Big[\big(\boldsymbol{\Sigma}_m^{+} + \boldsymbol{\Sigma}_m^{-} + 10^{-5}\mathbf{I}\big)^{-1}\big(\boldsymbol{\mu}_m^{+} - \boldsymbol{\mu}_m^{-}\big)\Big],$$

the threshold τ_m is the midpoint of the mean positive and negative training projections, and a held-out concept is positive when $\mathbf{x}_g^{\top}\mathbf{d}_m \geq \tau_m$. Metrics are computed per attribute and macro-averaged.

Held-out recovery. Table 7 reports macro F1 and AUROC; macro AUPRC is 0.912 on Random-HS20 and 0.925 on Random-HS50 for Llama-3.1-8B, and lies within 0.912–0.920 and 0.923–0.925 for the other encoders. HS50 has higher thresholded F1, HS20 slightly higher AUROC, so the result is not tied to one support cutoff.

Label-randomization controls. High-dimensional embeddings could separate arbitrary sparse labels on the training set. We therefore refit all directions and thresholds after two randomizations of Matched-HS20. *Profile permutation* (five seeds) permutes whole attribute-profile rows within each split, preserving every profile, its cardinality, marginal support and co-occurrence while breaking the concept–profile correspondence. *Degree-preserving edge swap* (three seeds) rewires the concept–attribute bipartite graph under fixed concept and attribute degrees, additionally destroying co-occurrence. Both remain almost perfectly separable on training concepts (train AUROC 0.999 against 1.000 for the true labels) but not on held-out ones: test AUROC is 0.494 ± 0.018 and 0.522 ± 0.018, and test AUPRC (0.032, 0.037) falls to the positive-label prevalence, against 0.912 for the true labels. Profile permutation is the stronger control because it keeps the empirical profiles intact; the small positive deviation under edge swap is not read as residual structure.

Lexical-deletion control. Attribute names can occur verbatim in descriptions (*inorganic-dust pneumoconiosis* contains *inorganic dust*). We re-embed every concept after deleting either the full attribute phrases or the attribute keywords, and compare with deleting the same number of random tokens, which controls for shorter and corrupted text (Table 8; ranges cover the two granularities). Random deletion alone lowers macro F1 substantially, so the full gap cannot be attributed to overlap; targeted deletion lowers F1 by a further 0.08, so overlap contributes, while AUROC remains at 0.94–0.96.

Table 8: HS20 attribute recovery after deleting attribute phrases or keywords from descriptions, against matched random deletion.

Description text	Macro F1	Macro AUROC
Original	0.823–0.825	0.992–0.993
Matched random deletion	0.629–0.634	0.956–0.966
Targeted attribute deletion	0.548–0.553	0.945–0.957

Table 9: Inclusion-based relation recovery on held-out child concepts. All non-fixed variants use thresholds selected on training-child pairs.

	Direct parent			Any ancestor		
Method	F1	AUROC	AUPRC	F1	AUROC	AUPRC
Paper form	0.5620	0.5975	0.6114	0.5305	0.5516	0.5580
Released Incl, fixed 0.6	0.7463	0.8656	0.7969	0.5275	0.6717	0.6571
Released Incl, calibrated	0.8502	0.8656	0.7969	0.6693	0.6717	0.6571
Baseline-centered L	**0.8567**	**0.8864**	**0.9130**	**0.6749**	**0.6960**	**0.7428**
Covariance C	0.8238	0.8566	0.8790	0.6424	0.6690	0.7097

D Attribute Profiles and Inclusion Scores

D.1 Inclusion implementations and threshold calibration

Let $\pi_C(m) = \mathbf{x}_C^\top \mathbf{d}_m$. The paper form of the inclusion score (Xiong, 2026) ℓ_2-normalizes the projection profile, $\widetilde{\boldsymbol{\pi}}_C = \boldsymbol{\pi}_C / \|\boldsymbol{\pi}_C\|_2$, and computes

$$\mathrm{Incl}_{\mathrm{paper}}(A, B) = \frac{\sum_m \mathrm{softplus}(\widetilde{\pi}_B(m))\, \sigma(\widetilde{\pi}_A(m))}{\sum_m \mathrm{softplus}(\widetilde{\pi}_B(m))}.$$

The released implementation instead centers each projection by its training threshold, $q_C(m) = \pi_C(m) - \tau_m$, and replaces the soft weight by the hard gate $W_B(m) = \mathbb{I}[q_B(m) > 0]$, giving $r(A, B) = \mathbb{E}_m[S_A(m) \mid W_B(m) = 1]$ with $S_A(m) = \sigma(q_A(m))$, the form in equation 1; a parent with no gated attribute receives score one. From the decomposition equation 2 we also evaluate the baseline-centered score $L(A, B) = r(A, B) - \mathbb{E}_m[S_A]$ and the covariance $C(A, B) = \mathbb{E}_m[W_B]\, L(A, B) = \mathrm{Cov}_m(W_B, S_A)$.

Table 9 uses Matched-HS20. Positives are direct child–parent pairs or transitive child–ancestor pairs, each with one deterministic non-ancestor negative; the child is in the training or test split and the candidate parent may be any concept. Except for the released fixed value 0.6, the relation threshold maximizes macro F1 on training-child pairs and is applied unchanged to test-child pairs (1,530 direct and 9,768 transitive pairs).

Calibration raises the released score's direct-parent F1 from 0.7463 to 0.8502 and transitive F1 from 0.5275 to 0.6693 without changing the ranking metrics; the selected thresholds are 0.3601 and 0.3226, so 0.6 does not transfer to this setting. Applying the hard rule $|Y_A \cap Y_B|/|Y_B|$ to the curated incidence of the same 104 attributes gives an attribute-based ceiling of 0.8627 (direct) and 0.7343 (transitive); the calibrated score is within 0.013 of the direct ceiling. Removing the child's activation baseline adds little; the raw covariance is lower, plausibly because its extra factor $\mathbb{E}_m[W_B]$ down-weights parents with sparse predicted profiles.

D.2 Relatedness and directional information

For each of the 7,153 reference edges we form a matched triplet of the true parent, a sibling and an unrelated candidate, and score every pair in both directions with the released Incl; no classifier is fitted, and the calibrated threshold 0.3601 is only a reference. Table 1 (main text) reports the means, pass rates and contrasts; the forward P10–P90 ranges are 0.235–0.960 (direct), 0.101–0.915 (sibling) and 0.038–0.371 (unrelated). The direct and sibling distributions overlap heavily while both are separated from unrelated pairs, and 78.3% of direct pairs pass the threshold in both directions.

Table 10: Orientation accuracy of known direct edges by representation (five-fold parent-grouped cross-validation).

Representation and model	Direction accuracy
Released Incl (no fitting)	59.5%
Signed difference of 104-dim. sigmoid attribute counts	64.3%
104-dim. sigmoid attribute profile + L2 logistic regression	83.1%
PCA-32 embedding + L2 logistic regression	92.5%
PCA-64 embedding + L2 logistic regression	94.2%
PCA-128 embedding + L2 logistic regression	94.6%
Raw-4096 embedding + L2 logistic regression	96.7%

Where the directional information lies. Fixing the task of orienting a known direct edge, we vary only the representation given to a restricted linear model (Table 10; five-fold parent-grouped cross-validation). The signed difference in predicted-positive attribute counts barely improves on Incl; a logistic regression over the full 104-dimensional sigmoid profile recovers much more, so the hand-designed aggregation discards information the attribute coordinates contain; and PCA coordinates of the raw embedding recover more still, indicating that the selected 104-attribute representation provides a less effective basis for orientation than the PCA representation. The results therefore identify an aggregation bottleneck and a further representation gap, rather than establishing two distinct stages of information loss.

D.3 Negative-evidence inclusion score

For hard sets, $Y_B \subseteq Y_A$ iff $Y_A^c \subseteq Y_B^c$, which motivates querying the attributes on which the child is negative:

$$r_-(A, B) = \frac{\sum_m \mathrm{softplus}(-q_A(m))\, \sigma(-q_B(m))}{\sum_m \mathrm{softplus}(-q_A(m))}.$$

This relaxation is not algebraically equivalent to the positive-evidence score, and an unrecorded attribute is only an operational negative. On the same triplets and with no new threshold (Table 1, main text), it improves direct-versus-sibling ranking and orientation but slightly reduces direct-versus-unrelated ranking. Its values saturate (mean forward scores 0.949, 0.945 and 0.939 for direct, sibling and unrelated pairs), and its directional difference correlates with the difference in positive attribute counts at Spearman $\rho = 0.771$, so the gain reflects cardinality asymmetry more than a calibrated relation score.

E Hierarchy-Potential Estimation and Controls

E.1 Estimator

Within each training fold, PCA is fitted without labels to the unique concept endpoints of the training edges; the first k components $\mathbf{P}_k$ are retained and divided by one global RMS value s computed on those concepts, which normalizes overall scale without whitening. Because centering cancels in a difference, $\mathbf{z}_C - \mathbf{z}_P = \mathbf{P}_k(\mathbf{x}_C - \mathbf{x}_P)/s$. A no-intercept L2 logistic regression is fitted to the mirrored examples $(\mathbf{z}_C - \mathbf{z}_P, 1)$ and $(\mathbf{z}_P - \mathbf{z}_C, 0)$, with the regularization strength selected on a parent-grouped validation split inside the fold. Its score $D(C, P) = \mathbf{w}^\top(\mathbf{z}_C - \mathbf{z}_P)$ equals $h(C) - h(P)$ with $\mathbf{u} = \mathbf{P}_k^\top \mathbf{w}/s$ (Proposition 3); $\mathbf{u}$ is not fitted separately but read off the PCA-space coefficients. Pair accuracy is the fraction of held-out edges with $D(C, P) > 0$; orientation AUROC contrasts forward and reversed orders. The appendix refers to h as the hierarchy potential.

E.2 Supervised capacity

For PCA-k the direction is restricted to the span of the first k principal components and the classifier has k coefficients; Raw-4096 fits one coefficient per embedding dimension. Table 3 (main text) reports PCA-16, PCA-64, PCA-256 and Raw-4096 under parent-grouped folds; PCA-32 gives 0.9245 ± 0.0127 accuracy and 0.9740 ± 0.0048 AUROC, and PCA-128 gives 0.9458 ± 0.0066 and

Table 11: Complete-branch holdout, Llama-3.1-8B, means over five branches. Left: the direct-parent score $D(C, P)$ against an alternative candidate's score for the same child; "Direct higher" is the paired probability that the direct score is larger. Right: Spearman correlation of h with graph depth on unseen nodes, and the probability that the deeper of two mutually incomparable concepts has the higher grade, by shortest-depth gap.

Alternative candidate	Direct higher	AUROC
Sibling	0.904	0.806
Unrelated	0.737	0.709
Depth-matched non-ancestor	0.709	0.676
Grandparent	0.157	0.278

Depth measure	ρ	Depth gap	Deeper higher
Shortest	0.381	0	0.502
Recursive mean	0.503	1	0.600
Longest	0.531	2	0.690
		3	0.750
		4+	0.753

0.9848 ± 0.0021. The signal saturates early, and PCA-64 is used as the canonical low-capacity potential throughout.

E.3 Direction stability

Directions fitted in different parent-grouped folds have pairwise cosine 0.963 on average (0.951–0.973); under complete-branch holdout the mean falls to 0.837 (0.667–0.957), so changing the training subdomain rotates the direction more than resampling does, while all pairs remain positively aligned. Compared with the mean child-to-parent offset $\bar{\boldsymbol{\delta}} = \mathbb{E}[\mathbf{x}_C - \mathbf{x}_P]$, the direction a difference-of-means estimator would return, $\mathbf{u}$ has cosine 0.594–0.709 (mean 0.651) with $\bar{\boldsymbol{\delta}}$ computed on training edges and 0.320–0.526 (mean 0.438) with $\bar{\boldsymbol{\delta}}$ computed on the held-out branch, so it is aligned with but far from identical to the mean offset.

E.4 Controls under held-out branches

Three controls are run under complete-branch holdout (Table 3, main text). *Shuffled labels* independently reverse the orientation of training edges and repeat the full fitting and regularization selection 20 times. *Random direction* draws 100 unit Gaussian directions in the training PCA-64 space. *PC1* is the first principal component of the raw training embeddings. The latter two have only their sign selected on training edges, which makes them more informative than unoriented axes while leaving the direction unsupervised. The learned direction reaches 87.82%–93.78% accuracy and 0.945–0.986 AUROC across branches; shuffled labels return to chance, and the sign-selected controls retain some signal but remain well below it.

F Structural Interpretation of the Hierarchy Potential

All results in this section use complete-branch holdout: each direction is fitted outside the evaluated branch and applied only to its nodes, on candidate types and graph statistics never used as training targets.

F.1 Hierarchical span rather than direct adjacency

Each held-out direct edge is paired with a sibling, an unrelated candidate, a depth-matched non-ancestor and, where it exists, a grandparent; Table 11 (left) reports the paired probability that the direct score is larger and the AUROC of direct against alternative scores, visualized in Figure 3a. Selecting the parent over a sibling requires only $h(S) > h(P)$, since $D(C, P) - D(C, S) = h(S) - h(P)$; it does not show that the score identifies the particular edge. The weak unrelated and depth-matched contrasts show that a scalar rank cannot determine whether two concepts are connected, and the grandparent reversal ($D(C, G) > D(C, P)$ in more than four fifths of matched cases) is the additivity of Proposition 3(iii).

Score distributions under parent-grouped folds. Under the parent-grouped protocol, direct forward pairs have mean $D = 5.19$ (median 4.36; 94.2% positive). Sibling gaps are centered at zero (mean 0.08, 52.1% positive) but with standard deviation 4.20, so siblings lie at similar but not identical heights; unrelated gaps are also centered at zero (mean 0.25, 51.3% positive) with a larger spread

Table 12: Grade differences between mutually non-ancestral concepts in held-out branches. Positive rate is the branch-mean probability that the deeper node has the higher grade; the last column averages branch-level medians of the standardized difference. Pair counts are totals across runs; the 4+ bin has eligible pairs in four branches.

Depth gap	Pairs	Deeper-positive rate	Median Δh_{std}
0	22,469	0.502	0.003
1	21,212	0.600	0.322
2	17,876	0.690	0.694
3	12,304	0.750	1.065
4+	6,748	0.753	1.168

(6.76), since an arbitrary concept may lie anywhere on the axis. Grandparent gaps have mean 9.51, about 1.8 times the direct mean, are positive in 98.9% of cases, and exceed the direct-parent score in 91.8% of matched pairs.

F.2 Association with graph depth on unseen nodes

Shortest and longest depth are the minimum and maximum root-to-node path lengths; recursive mean depth is $d_{\mathrm{mean}}(v) = 0$ at the root and $1 + |\mathrm{Pa}(v)|^{-1} \sum_{p \in \mathrm{Pa}(v)} d_{\mathrm{mean}}(p)$ otherwise, the expected path length when a parent is chosen uniformly at each step. Depth is never supplied to PCA or the classifier. Spearman correlation is computed over the unseen branch only; Table 11 (right) reports branch means, with branch ranges 0.266–0.484 (shortest), 0.401–0.631 (mean) and 0.414–0.601 (longest). All are positive in every branch and none is deterministic. Shortest depth is the weakest proxy, as expected under multiple inheritance, where the shortest route to the root need not lengthen along every edge; the stronger association with mean and longest depth indicates an accumulated hierarchical component rather than the minimum root distance.

F.3 Depth gradients among incomparable concepts

We sample mutually non-ancestral pairs within each held-out branch, grouped by shortest-depth gap, with at most four fixed-seed samples per deeper anchor and gap bin; equal-depth pairs are randomly oriented as a chance control. Potentials are standardized within each branch before subtraction because independently fitted directions differ in scale (Table 12; Figure 3b). Equal-depth pairs sit at chance, and both the sign consistency and the median standardized difference grow with the gap, but the overlap and the plateau at the largest gaps rule out an exact depth encoding.

G Two-Factor Models for Hierarchical Relation Recovery

G.1 Models

The two-factor model uses $\boldsymbol{\phi}_R(A, B) = [R(A, B),\ \Delta h(A, B),\ R(A, B)\Delta h(A, B)]$ with $R \in \{R_{\mathrm{Incl}}, R_{\cos}\}$; the three columns are standardized on the outer training portion and fed to an L2 logistic regression. Its 68 hierarchy-supervised coefficients are the 64 defining the PCA-64 direction plus three weights and an intercept; the count excludes the frozen encoder and, for R_{Incl}, the attribute directions. The attribute-profile reference uses standardized 104-dimensional continuous attribute projections and the pair representation $\boldsymbol{\phi}_{\mathrm{attr}}(A, B) = [\widetilde{\mathbf{q}}_A; \widetilde{\mathbf{q}}_B; \widetilde{\mathbf{q}}_A \odot \widetilde{\mathbf{q}}_B]$ (313 coefficients including the intercept). The embedding references fit PCA within the outer training portion and use $\boldsymbol{\phi}_k(A, B) = [\mathbf{z}_A^{(k)}; \mathbf{z}_B^{(k)}; \mathbf{z}_A^{(k)} \odot \mathbf{z}_B^{(k)}]$ ($3k + 1$ coefficients); they are capacity-matched empirical references, not upper bounds. Regularization is selected inside each training portion. The direction $\mathbf{u}$ is fitted from direct training edges and their reversed copies only; grandparents enter only the downstream classifier. Principal results are in Tables 4 and 5 (main text).

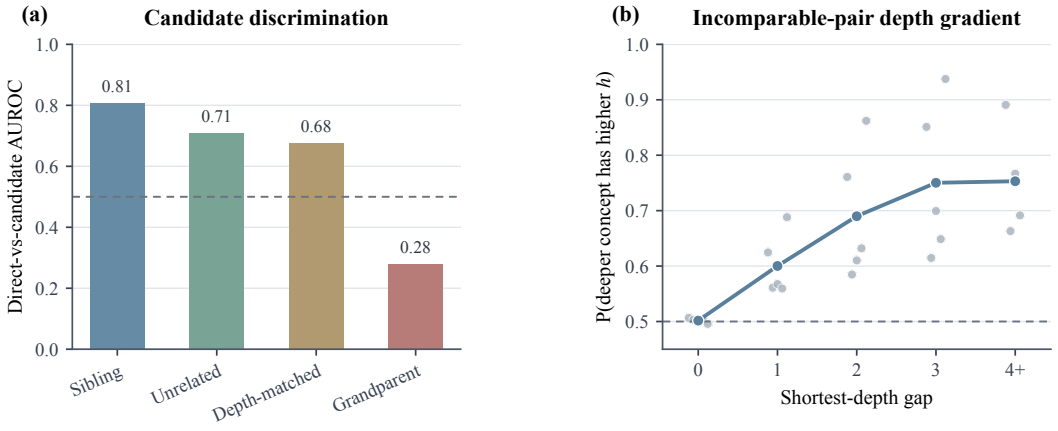


Figure 3: Structural interpretation of the hierarchy potential under complete-branch holdout. (a) Direct-parent scores against four candidate types; the dashed line is chance AUROC. (b) For mutually non-ancestral pairs, the probability that the deeper concept has the higher potential by shortest-depth gap. Gray points are individual branches, the blue line their mean, and the dashed line 0.5.

Table 13: PCA pair capacity under parent-grouped folds; each classifier has $3k + 1$ downstream coefficients.

PCA dimension	Coefficients	Hasse AUROC	Two-hop AUROC
16	49	0.769	0.867
32	97	0.805	0.920
64	193	0.814	0.936
128	385	**0.816**	0.943
256	769	0.812	0.949
512	1,537	0.800	**0.954**
1,024	3,073	0.777	0.951
2,048	6,145	0.711	0.944
4,096	12,289	0.536	0.896

G.2 Capacity and parameter efficiency

Table 13 gives the PCA pair ladder for both tasks under parent-grouped folds. Performance rises quickly at low dimension and is not monotone: Hasse adjacency peaks at PCA-128 and two-hop subsumption at PCA-512, and PCA-4096 collapses. Under complete-branch holdout PCA-512 falls below PCA-64 and the cosine two-factor model (Table 4), losing 0.077 against 0.050 relative to its cross-validated value; later principal coordinates carry branch-specific information that helps within a distribution and hurts across one. The low- and intermediate-dimensional models are therefore the references, and PCA-4096 is not treated as an upper bound.

G.3 Hasse adjacency

Table 14 reports the adjacency task, in which grandparents are negatives. The attribute-based two-factor model is best under parent-grouped folds and the cosine variant under branch holdout, where attribute directions must be refitted on shifted support. The binding contrast is direct parent versus grandparent: 0.571 for the attribute model under parent-grouped folds and 0.536 for the cosine model under branch holdout, the additivity of Section F.1. The two-factor representation is suited to ordered subsumption and does not by itself encode whether an ancestor is exactly one edge away.

Table 14: Hasse-adjacency AUROC (grandparents negative), mean ± standard deviation.

Method	Parent-grouped folds	Held-out branches
Symmetric R_{Incl}	0.728 ± 0.011	–
Cosine $R_{\cos}$	0.711 ± 0.009	0.704 ± 0.028
Δh	0.694 ± 0.006	–
$R_{\text{Incl}} + \Delta h$	$\mathbf{0.827 \pm 0.010}$	0.727 ± 0.024
$R_{\cos} + \Delta h$	0.807 ± 0.006	$\mathbf{0.796 \pm 0.010}$
PCA-64 pair	0.814 ± 0.010	0.777 ± 0.023

Table 15: Within- and cross-ontology orientation of direct edges, mean ± s.d. across four WordNet or five SNOMED CT held-out branches.

Test ontology	Direction source	Accuracy	AUROC
WordNet	WordNet, branch-held-out	0.858 ± 0.054	0.929 ± 0.036
WordNet	SNOMED CT, all edges	0.807 ± 0.068	0.877 ± 0.059
SNOMED CT	SNOMED CT, branch-held-out	0.903 ± 0.022	0.964 ± 0.014
SNOMED CT	WordNet, all edges	0.729 ± 0.036	0.790 ± 0.020

H Cross-Ontology Transfer between SNOMED CT and WordNet

WordNet 3.0 contributes 82,115 noun synsets and 75,831 direct hypernym edges after excluding instance nodes; each lemma is embedded with the SNOMED CT procedure and lemma vectors are averaged per synset. For each source ontology a PCA-64 orientation model is fitted as in Section E, with WordNet PCA fitted on a fixed-seed sample of at most 12,000 training-edge endpoints. The coefficient is mapped back to the shared 4,096-dimensional space and applied to the other ontology with no target label used to recalibrate, choose the sign, or select a threshold. Metrics are orientation accuracy and AUROC on known edges; they do not test parent recovery from arbitrary pairs.

H.1 Zero-shot transfer on held-out branches

For WordNet the complete branches rooted at *animal*, *plant*, *artifact* and *cognition* (4,033, 4,493, 10,733 and 3,945 test edges) are held out in turn; SNOMED CT uses its five branches. Within-ontology directions are fitted outside the target branch, cross-ontology directions on all edges of the other ontology, and boundary-crossing edges are excluded (Table 15). Simple branch means keep the largest branch from dominating; as before, branches may overlap, so standard deviations describe branch variation. Transfer is bidirectional but incomplete: a shared hierarchy-related component generalizes, while ontology-specific fitting remains beneficial.

H.2 Full-edge audit

Because branch means depend on the chosen roots, each full-source direction is also applied to every direct edge of the other ontology (Table 16). The signal persists, and the asymmetry seen in the branch means shrinks to about 2.5 accuracy points, so it is not an artifact of the holdout roots.

H.3 Lexically unseen WordNet edges

Lemmas can belong to several synsets, so holding out synsets does not make every test lemma unseen. Restricting to the 8,477 edges whose child and parent lemmas are absent from all training nodes of the corresponding run, with directions unchanged (Table 17), leaves both directions well above chance; the transferred AUROC moves from 0.877 to 0.858. Since the filter also changes the test composition, the difference is descriptive rather than an estimate of leakage.

Table 16: Cross-ontology orientation on all target direct edges (off-domain evaluations only).

Test ontology	Direction source	Edges	Accuracy	AUROC
WordNet	SNOMED CT	75,831	0.762	0.839
SNOMED CT	WordNet	7,153	0.737	0.813

Table 17: Orientation on WordNet edges whose child and parent lemmas are unseen in the corresponding training split, mean $\pm$ s.d. across four branches.

Direction source	Test edges	Accuracy	AUROC
WordNet, branch-held-out	8,477	0.853 ± 0.051	0.927 ± 0.034
SNOMED CT, all edges	8,477	0.788 ± 0.086	0.858 ± 0.078

I Cross-Model Replication

Llama-3.1-8B, KoBioMed-Llama-3.1-8B (Cho et al., 2025) and Gemma-2-9B (Team, 2024) use the same concepts, texts, labels, candidates, splits and grids; attribute directions, PCA, hierarchy directions and classifiers are fitted independently per encoder. Attribute recovery is in Table 7; per-attribute AUROC vectors on HS20 correlate at 0.970–0.990 between each pair of encoders, so the same attributes are easy or hard for every model. Table 18 gives orientation and two-factor results under branch holdout; the three two-factor models differ by at most 0.009 AUROC, and the medically continued-pretrained model gives no consistent advantage.

J Oracle Attributes and Destination-Value Closure

Oracle reference. Predicted profiles are replaced by the full 1,506-column curated incidence, $Y_C = \{m \mid I(C, m) = 1\}$, and direct parents are scored by $S(C \sqsubseteq P) = |Y_C \cap Y_P|/|Y_P|$ with a threshold selected on training children. This removes attribute selection and prediction error but keeps the vocabulary, incidence construction and inclusion rule, so it is an attribute-based reference rather than an unconstrained bound. The smaller Core-51 panel compresses the 4,253 concepts into 602 distinct intents, with 92.5% of concepts colliding with another and a quarter of sibling pairs sharing an intent; Full-1506 yields 3,635 intents.

Value closure. ExactFlat treats each $(\mathrm{typeId}, \mathrm{destinationId})$ as an independent column and cannot see that a child replaces a broad destination with a descendant. DestinationClosure applies $(r, d) \Rightarrow (r, a)$ for every ancestor a of d, activating only ancestor-valued columns already in the vocabulary, globally and before any pair is compared. Table 2 (main text) gives the intent relations on all 7,352 internal edges: strict containment rises from 26.66% to 91.89%, equal intents from 8.01% to 8.11%, incomparable intents fall from 65.27% to zero, and the four reverse containments (0.05%) disappear, so every edge satisfies $Y_P^{\mathrm{closure}} \subseteq Y_C^{\mathrm{closure}}$. Of the 4,799 incomparable ExactFlat edges, all involve a same-type destination change, 71.24% involve nothing else, and at the pair level 96.31% consist only of clean specializations in which every child value descends from a parent value; the changes concentrate in *finding site* (2,531 units) and *associated morphology* (2,134). Oracle recovery rises from F1 0.9215 and AUROC 0.9361 to 0.9961 on both.

Sibling control. Closure causes little additional collapse of intent profiles: among 41,532 sibling pairs the equal-intent rate moves from 2.1213% to 2.1261% and 86.07% remain incomparable, while the strictly comparable fraction rises from 7.24% to 11.80%; incidence density rises from 0.21% to 1.12% and the number of distinct intents changes from 3,635 to 3,627. Strict inclusion thus also occurs for some siblings, so inclusion is not a complete relation criterion, and because the relationships and the hierarchy come from the same ontology the analysis measures schema consistency, not discovery.

Table 18: Hierarchy-potential orientation and cosine two-factor two-hop subsumption under complete-branch holdout, mean ± s.d. across five branches.

	Hierarchy potential		Two-factor model	
Model	Accuracy	AUROC	AUROC	AUPRC
Llama-3.1-8B	0.903 ± 0.022	0.964 ± 0.014	0.887 ± 0.021	0.787 ± 0.043
KoBioMed-Llama-3.1-8B	0.900 ± 0.022	0.966 ± 0.012	0.896 ± 0.015	0.805 ± 0.033
Gemma-2-9B	0.922 ± 0.010	0.971 ± 0.003	0.893 ± 0.015	0.793 ± 0.038

K Related Work and Discussion

Linear representations and the geometry of hierarchy. That semantic attributes correspond to directions in embedding space dates to the offset regularities of word vectors (Mikolov et al., 2013) and now underlies probing, steering, and the causal formulation of the *Linear Representation Hypothesis* (Park et al., 2024; Gurnee & Tegmark, 2024; Nanda et al., 2023; Han et al., 2024; Liu et al., 2024); Engels et al. (2025) document the exception. Work on hierarchy has built on this base in three ways, each treating hierarchy as a relation between two concepts. Park et al. (2025) show that the vector from a child to its parent is orthogonal to the parent, a local constraint that yields no global ordering. Xiong (2026) lift attribute directions to half-spaces whose intersections give an inclusion score over attribute profiles. Sakata et al. (2026) learn a separate linear map from child to parent at each depth, treating depth as a variable but parameterizing it layer by layer. These works primarily formulate hierarchy as a relation to be computed between two concepts. Our results show that relation-supervised linear probing induces a single-concept score whose differences provide substantially stronger orientation information than the released inclusion score, which retains only a weak directional signal. The finding is compatible with the orthogonality of Park et al. (2025): child-to-parent offsets share a common component $\mathbf{u}$ that orients the edge, although we do not decompose the remainder. These results qualify the LRH rather than refute it. Half-spaces are recovered reliably, and attribute inclusion and embedding similarity provide complementary evidence about relatedness, whereas the hierarchy potential provides evidence about orientation; but the *is-a* poset of an ontology whose attribute values are themselves hierarchical is not the concept lattice of a flat context. In order-theoretic terms, a lattice with a rank function is a graded lattice, and the GRH posits an analogous real-valued grading for LLM concept hierarchies. The combination of one relatedness score and one grade difference offers a compact empirical approximation to subsumption, not an exact two-dimensional representation of the ontology, and a probe trained against random negatives measures only the relatedness half.

Generality as a scalar. Encoding generality in one coordinate has a long history in embeddings trained for it: order embeddings (Vendrov et al., 2016), Poincaré embeddings and entailment cones (Nickel & Kiela, 2017; Ganea et al., 2018), norm-based hypernymy models (Nguyen et al., 2017; Vulić & Mrkšić, 2018), hyperbolic vision-language models that read norm as generality (Desai et al., 2023; Pal et al., 2025), and hierarchy-tuned language models (He et al., 2024), which presuppose that off-the-shelf models encode hierarchy poorly. In all of these the hierarchy is written into the objective; we find that a relation-supervised linear probe reads such a coordinate from frozen representations that were never trained for it. A distributional reading suggests why: a general term occurs in more contexts and in the descriptions of many specific terms, a model trained to predict text has reason to represent this breadth, and breadth is naturally a scalar quantity. Levy et al. (2015) observed that supervised hypernymy classifiers learn a per-word "prototypical hypernymy" score rather than a relation and treated it as an artifact; we find the score is a coordinate of the representation, recoverable from frozen representations by a linear probe, stable across encoders and reusable across ontologies without retraining, and that the missing piece they identified, telling related from unrelated pairs, is exactly what the lattice supplies.

Concept structure in medical language models. Medical foundation models are probed for concept knowledge and decomposed with sparse autoencoders into interpretable features (Le et al., 2024; Haque et al., 2026), but the geometry of their concept hierarchies has not been examined. SNOMED CT serves here not as an application domain but as a large ontology whose curated defin-

ing attributes make the intents of the LRH available as ground truth; consistently, the medically continued-pretrained encoder offers no advantage over general-purpose ones (Section I).

What the grade is. Three observations constrain what h can be. It is a property of single concepts: one direction orders edges in unseen subtrees and in an unseen ontology, and directions fitted on different branches agree at cosine 0.84. It is not a memorized offset: it is only partially aligned with the mean child-to-parent difference and with the first principal component, each of which alone orients far fewer edges. And it behaves like depth without being depth: it correlates with depth at 0.4–0.5, grows with the depth gap even between unrelated concepts and then plateaus, while siblings scatter around zero. The grade is a real-valued grading in the sense of Definition 1, a specificity potential rather than an integer rank, which is what an ontology without a canonical unit-step rank can support.

Limitation. Three limitations apply. Curated intents exist for one clinical subtree, so the two halves of the GRH are tested jointly only on SNOMED CT, while the grade alone is also tested on WordNet. The grade is only an approximate real-valued grading: it does not encode adjacency, its correlation with depth is moderate, and cross-ontology transfer is incomplete. And we read the grade with linear probes on concept names from three 8–9B encoders; whether the model uses it causally, for instance under steering along $\mathbf{u}$, is untested.