TASK ENTROPY MUST BE PAID: WHEN NEGATIVE CURVATURE HELPS REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

When does negative curvature become a useful representation resource rather than an architectural preference? A hierarchy alone does not decide this: the same input can induce a fine task that must preserve exponentially many distinctions or a coarse task that collapses most of them. We formalize this task-side demand through *task entropy*, the packing growth of a loss- or evaluation-induced pseudometric. For packings whose evaluation contains an N-way margin-identification restriction, we prove an exact capacity law: with unrestricted coordinatewise L-Lipschitz scores, a radius-R representation supports exactly $P_{\mathcal{M}}(R, m/L)$ identities at margin m. This yields a packing-restricted margin-error bottleneck, Euclidean resource tradeoffs, and a hyperbolic curvature threshold. On a regular tree, Euclidean obstruction and hyperbolic realization are governed by the same exponential leaf-growth rate, while a coarse task on the same hierarchy has zero task entropy and a bounded Euclidean realization. On a controlled 800-node WordNet intervention, Poincaré embeddings reduce hierarchy-task stress in all five seeds ($4.47\% \pm 1.80\%$); frozen-embedding margin-error curves provide a post-training identification diagnostic. Five real networks retain both positive and negative cases, probing task dependence and boundary cases highlighted by the capacity analysis rather than a universal optimizer-performance law.

1 INTRODUCTION

Hyperbolic representation learning is commonly motivated by input growth: trees contain exponentially many nodes with depth, fixed-dimensional Euclidean balls have polynomial volume growth, and hyperbolic balls have exponential growth. This mismatch explains why hyperbolic spaces are effective for low-distortion tree and hierarchy embeddings and has motivated hyperbolic neural, graph, and metric-learning models (Nickel and Kiela, 2017; 2018; Sala et al., 2018; Sarkar, 2011; Ganea et al., 2018; Chami et al., 2019; Gu et al., 2019).

A learning objective, however, need not preserve every distinction in the input. On one fixed b-ary tree, leaf identification must keep b^T leaves distinguishable at depth T, whereas top-branch classification maps the same leaves to only b labels. The input is unchanged, but the representational demand is not. Recent empirical work similarly finds that hyperbolic gains depend on whether the downstream target preserves graph geometry (Naddeo et al., 2026), while hierarchical inputs alone need not yield a consistent advantage (Van Nooten and Daelemans, 2025).

This leaves a sharper model-selection question:

When do task-required distinctions exceed Euclidean capacity under controlled dimension, radius, margin, and readout regularity—and when can negative curvature close that deficit?

We separate *task demand* from *representation capacity*. A task induces a pseudometric d_T on evaluation-level states: states are far when the loss or evaluation requires them to remain distinguishable. Its packing curve counts how many such states coexist at a chosen resolution, and exponential growth of that count defines task entropy. For packings whose evaluation contains an N-way margin-identification restriction, the core result is exact: under unrestricted coordinatewise L-Lipschitz scores, the identification capacity of a radius-R representation at margin m equals $P_{\mathcal{M}}(R, m/L)$. Thus packing is not merely a loose upper bound; it is the exact identification capacity in the declared Lipschitz model. Figure 1 summarizes this task-to-capacity chain.

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.

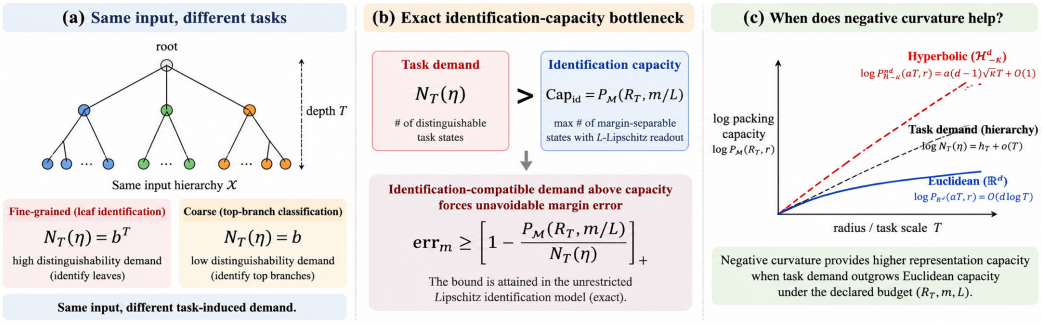

Figure 1: Task demand, exact margin-identification capacity, and geometry-dependent growth. (a) At the illustrated task resolution, the same hierarchy yields $N_T(\eta) = b^T$ leaf identities but only $N_T(\eta) = b$ top-branch identities. (b) For unrestricted coordinatewise L-Lipschitz scores, identification capacity equals $P_{\mathcal{M}}(R_T, m/L)$; if an identification-compatible packing exceeds this capacity, its packing-restricted margin error is unavoidable. (c) In the schematic regime $a(d-1)\sqrt{\kappa} > h > 0$, fixed-dimensional Euclidean log-packing grows as $O(d \log T)$ while hyperbolic log-packing grows linearly at rate $a(d-1)\sqrt{\kappa}$.

The consequences are explicit. In fixed-dimensional Euclidean space, linear radius and constant separation provide only subexponential packing, so positive task entropy must be paid through dimension, radius, margin, or readout regularity. A d-dimensional hyperbolic ball of curvature $-\kappa$ and radius aT supplies exponential packing rate $a(d-1)\sqrt{\kappa}$, yielding a curvature threshold. For regular-tree identification we construct hyperbolic prototypes at the required exponential scale, while the Euclidean bound forces another resource to grow. On the same tree, coarsening the task removes the asymptotic capacity demand.

The metric ingredients behind the exact characterization are elementary: margin correctness forces separation, and distance-to-prototype scores attain the converse. We do not claim a new packing inequality. The contribution is the operational task-indexed characterization, the exact margin-identification capacity it induces, and the resulting resource criterion for representation geometry. This differs from input-growth priors (Weber, 2020), empirical geometry–task alignment (Naddeo et al., 2026), and the distribution-level minimax separation of Rawal and Vishwanath (Rawal and Vishwanath, 2026). Their result is statistically stronger within its hierarchical learning model; ours is deterministic and representation-level, and applies only when the task evaluation contains the stated margin-identification requirement.

Contributions.

- We define task entropy from a loss- or evaluation-induced pseudometric and distinguish general task packing from the *identification-compatible* packings to which the margin theorem applies.
- We prove that metric packing is the *exact identification capacity* of the declared unrestricted coordinatewise Lipschitz model, yielding a tight error floor for the isolated identification problem, Euclidean resource tradeoffs, and an explicit hyperbolic threshold.
- We prove a same-input/different-task theorem and evaluate the framework on a controlled WordNet intervention and five real networks. Experiments use matched optimization/tuning budgets, report realized margins and radii, expose the full empirical margin-error curve, and retain negative and seed-sensitive cases.

2 Related Work

Hyperbolic representations and geometry–task alignment. Hyperbolic embeddings are standard for trees and hierarchies (Nickel and Kiela, 2017; 2018), with neural, graph, and mixed-curvature extensions (Ganea et al., 2018; Chami et al., 2019; Gu et al., 2019). Embedding theory explains low-distortion tree realizations and dimension–distortion tradeoffs (Bonk and Schramm, 2000; Sarkar,

2011; Sala et al., 2018); surveys cover vision and foundation-model applications (Mettes et al., 2024; He et al., 2025). Weber (Weber, 2020) selects geometric priors from input-neighborhood growth, whereas Naddeo et al. (Naddeo et al., 2026) explicitly study whether the downstream task is aligned with graph geometry. Our object is different: task packing is induced by what the evaluation must distinguish, so the same input may generate different capacity requirements. Recent applications and alternatives (Bui et al., 2025; Bläsius et al., 2025; Zhou, 2026) address complementary architectures and tasks.

Closest theory and learning capacity. Rawal and Vishwanath (Rawal and Vishwanath, 2026) prove an exponential Euclidean–hyperbolic sample-complexity separation for a specified hierarchical model under Lipschitz regularization, including a minimax lower bound. We instead characterize a deterministic representation–readout capacity, without a data-generating distribution or optimizer. Packing, covering, and fat-shattering quantities are classical learning-theoretic tools (Anthony and Bartlett, 1999; Shalev-Shwartz and Ben-David, 2014); our claim is not a new packing inequality, but an exact operational identification of task demand with representation packing under a declared margin and Lipschitz budget.

3 Problem Setup and Task Entropy

The purpose of the formalism is simple: specify *which target distinctions matter*, then compare their demand with the number of margin-separated states a representation can support.

3.1 Tasks, scale, and task packing

Let $\mathcal{X}$ be an ambient set of evaluation-level states. Typical elements are taxonomy nodes or labels, retrieval targets, and structured outputs; multiple raw examples may correspond to the same state when the evaluation treats them as equivalent. A *task family* is a sequence

$$\mathcal{T} = \{(S_T, d_T) : T = 1, 2, \ldots\}, \tag{1}$$

where $S_T \subset \mathcal{X}$ is finite and d_T is a pseudometric induced by the loss or evaluation. The scale T is the task's natural growth parameter—for example taxonomy depth, graph radius, sequence horizon, or evaluation resolution. No metric on raw inputs is assumed.

Definition 1 (Packing number)**.** For a metric or pseudometric space (S, d) and resolution $\eta > 0$, define

$$\mathrm{pack}(S, d, \eta) := \sup\Big\{ |Q| : Q \subseteq S \text{ finite},\ d(q, q') \geq \eta \text{ for all distinct } q, q' \in Q \Big\}. \tag{2}$$

The value may be $+\infty$. If S is finite, the supremum is a maximum.

Definition 2 (Task-packing rates and task entropy)**.** For a task family $\mathcal{T}$, let

$$N_T(\eta) := \mathrm{pack}(S_T, d_T, \eta), \qquad h_T(\eta) := \frac{1}{T} \log N_T(\eta). \tag{3}$$

Define

$$\underline{h}_{\mathcal{T}}(\eta) := \liminf_{T \to \infty} h_T(\eta), \qquad \overline{h}_{\mathcal{T}}(\eta) := \limsup_{T \to \infty} h_T(\eta). \tag{4}$$

When they agree, their common value $h_{\mathcal{T}}(\eta)$ is the *task entropy* at resolution η. We write $Q_T(\eta) = \{q_1, \ldots, q_{N_T(\eta)}\}$ for one maximizing packing when S_T is finite.

Positive lower rate means that the evaluation retains exponentially many distinguishable states. Because d_T may be a pseudometric, distinct states with $d_T = 0$ are treated as indistinguishable at every positive resolution and contribute only once to a positive-resolution packing. For example, a complete b-ary tree has b^T leaf identities, whereas a root-level b-class objective on the same leaves has $N_T(\eta) = b$ for $0 < \eta \leq 1$ and zero asymptotic task entropy.

A general task packing need not induce an identification problem; the compatibility condition required by the capacity theorem is stated below after the readout model.

Table 1: Examples of task-induced pseudometrics. Scale and resolution belong to the evaluation and are fixed before observing the learned representation.

Task	S_T	d_T	natural scale T
Hierarchy-aware prediction	labels/nodes	taxonomy distance	depth
Graph retrieval	query/target nodes	target path/relevance	radius/diameter
Structured prediction	outputs	edit/transport distance	horizon/resolution
Flat K-class prediction	labels	discrete $0/1$ metric	underlying task scale

3.2 Representation capacity and margin identification

Let $(\mathcal{M}, \rho)$ be a metric representation space and $f_T : S_T \to \mathcal{M}$ an encoder. A radius budget R_T means that $f_T(S_T)$ lies in some ball $B_{\mathcal{M}}(z_0, R_T)$. Define the center-free packing capacity

$$P_{\mathcal{M}}(R, r) := \sup_{z_0 \in \mathcal{M}} \operatorname{pack}(B_{\mathcal{M}}(z_0, R), \rho, r). \tag{5}$$

This center-free capacity takes values in $\mathbb{N} \cup \{+\infty\}$; the exact-capacity theorem below is understood in this extended sense. For the common linear-radius regime $R_T = aT$, define

$$g_{\mathcal{M}}(a, r) := \limsup_{T \to \infty} \frac{1}{T} \log P_{\mathcal{M}}(aT, r). \tag{6}$$

For fixed $a, r > 0$, $g_{\mathbb{R}^d}(a, r) = 0$ for $d \geq 1$. For hyperbolic space $\mathbb{H}^d_{-\kappa}$ with $d \geq 2$ and $\kappa > 0$,

$$g_{\mathbb{H}^d_{-\kappa}}(a, r) = a(d-1)\sqrt{\kappa}. \tag{7}$$

Appendix A.1 gives the corresponding finite-T packing asymptotics.

For an N-way identification problem, a score readout is $s = (s_1, \ldots, s_N) : \mathcal{M} \to \mathbb{R}^N$. We require every coordinate to be L-Lipschitz, equivalently

$$\max_j |s_j(z) - s_j(z')| \leq L\rho(z, z') \qquad \forall z, z' \in \mathcal{M}. \tag{8}$$

For a finite identification set $Q = \{q_1, \ldots, q_N\}$, define the margin error on Q by

$$\operatorname{err}_{m,Q}(f_T, s) := \frac{1}{N} \Big| \big\{ q_i \in Q : \exists j \neq i,\ s_i(f_T(q_i)) < s_j(f_T(q_i)) + m \big\} \Big|. \tag{9}$$

Thus $\operatorname{err}_{m,Q}$ is normalized by the identification set Q, not by the full ambient task set S_T. When the complete evaluation set is the identification set, we abbreviate it as err_m.

Identification-compatible packing. Let $Q = \{q_1, \ldots, q_N\} \subseteq S_T$. We say that the task evaluation admits an *m-margin identification restriction on Q* if its prediction on Q is represented by the score map above, with one distinct target coordinate per state, and q_i is counted as successful only when

$$s_i(f_T(q_i)) \geq s_j(f_T(q_i)) + m \qquad \forall j \neq i. \tag{10}$$

When this restriction is part of the task evaluation, we call *Q identification-compatible at margin m*. Instance identification and prototype/label retrieval are direct examples. Distance regression or metric fitting is not automatically identification-compatible; it inherits the bottleneck only through an explicit identification restriction of its evaluation.

4 The Entropy Bottleneck

How many identities can a radius-constrained representation support at prescribed margin and readout smoothness? In the unrestricted Lipschitz identification model, the answer is exact. Define $\operatorname{Cap}_{\mathrm{id}}(R; L, m)$ as the supremum of finite N for which there exist a center z_0, prototypes $z_1, \ldots, z_N \in B_{\mathcal{M}}(z_0, R)$, and coordinatewise L-Lipschitz scores satisfying $s_i(z_i) \geq s_j(z_i) + m$ for every i and $j \neq i$.

Theorem 3 (Exact Lipschitz identification capacity)**.** *For every metric space $(\mathcal{M}, \rho)$ and $L, m, R > 0$,*

$$\boxed{\mathrm{Cap}_{\mathrm{id}}(R; L, m) = P_{\mathcal{M}}\Big(R, \frac{m}{L}\Big).} \tag{11}$$

The proof has two directions. Margin correctness forces every pair of correctly identified prototypes to be at least m/L apart. Conversely, any m/L-packing $\{z_i\}$ is realized by the distance scores $s_i(z) = -L\rho(z, z_i)$. Thus representation packing is not merely an upper bound for this model; it is its exact identification capacity. The equality concerns *representation-space* capacity: an N-identity problem still uses N score coordinates, so the theorem does not claim constant output or computational complexity.

Corollary 4 (Task-induced bottleneck; tightness for isolated identification)**.** *Let $Q_T(\eta)$ be an identification-compatible task packing at margin m, with $N = N_T(\eta)$. If $f_T(Q_T(\eta))$ lies in a radius-R_T ball and the score coordinates are L-Lipschitz, then*

$$\mathrm{err}_{m, Q_T(\eta)}(f_T, s) \geq \left[1 - \frac{P_{\mathcal{M}}(R_T, m/L)}{N_T(\eta)}\right]_+. \tag{12}$$

Moreover, for the isolated N-state identification problem, over all encoders into an arbitrary radius-R_T ball and all unrestricted coordinatewise L-Lipschitz readouts,

$$\inf_{f,s} \mathrm{err}_m(f, s) = 1 - \frac{\min\{N, P_{\mathcal{M}}(R_T, m/L)\}}{N}. \tag{13}$$

The corollary is a deterministic representation-capacity statement. It does not say that every task pseudometric contains an identification reduction, nor that an optimizer will discover a capacity-achieving embedding when one exists.

4.1 Euclidean and hyperbolic consequences

For a Euclidean ball,

$$P_{\mathbb{R}^d}(R, r) \leq \left(1 + \frac{2R}{r}\right)^d. \tag{14}$$

This yields the following resource constraints.

Corollary 5 (Euclidean Lipschitz inflation)**.** *Assume $N_T(\eta) \geq e^{hT}$ for some $h > 0$ and $f_T(Q_T(\eta)) \subseteq B_2(0, aT) \subset \mathbb{R}^d$. If $\mathrm{err}_{m, Q_T(\eta)}(f_T, s) \leq \varepsilon < 1$, then*

$$L \geq \frac{m}{2aT}\left[((1-\varepsilon)e^{hT})^{1/d} - 1\right]_+. \tag{15}$$

For fixed d, linear radius, and fixed margin, the required Lipschitz scale grows exponentially in T whenever the lower bound is non-vacuous.

Corollary 6 (Hyperbolic packing threshold)**.** *Let $g = a(d-1)\sqrt{\kappa}$. Under fixed positive margin and fixed L:*

1. *if $g < \underline{h}_{\mathcal{T}}(\eta)$, then the margin error on $Q_T(\eta)$ tends to one;*
2. *if $g > \overline{h}_{\mathcal{T}}(\eta)$, the ambient hyperbolic packing capacity eventually exceeds task demand at the exponential scale, so the packing obstruction becomes vacuous.*

When $h_T(\eta) \to h$, the boundary is

$$a(d-1)\sqrt{\kappa} = h. \tag{16}$$

At equality, constant factors and the separation scale matter.

Corollary 7 (Euclidean resource tradeoff)**.** *Suppose $N_T(\eta) \geq \exp(hT - o(T))$. For a Euclidean representation in dimension d_T, radius R_T, fixed margin m, and Lipschitz budget L_T, non-vanishing margin accuracy on $Q_T(\eta)$ requires*

$$d_T \log\left(1 + \frac{2R_T L_T}{m}\right) \geq hT - o(T). \tag{17}$$

Thus $R_T = O(T)$ and $L_T = O(1)$ require $d_T = \Omega(T/\log T)$; fixed d_T and $L_T = O(1)$ require exponential radius; and fixed d_T with $R_T = O(T)$ requires exponential Lipschitz scale.

Corollary 8 (Polynomial-budget Euclidean impossibility)**.** *Let $N_T(\eta) \geq \exp(hT - o(T))$ for some $h > 0$. If*

$$R_T = T^{O(1)}, \qquad L_T = T^{O(1)}, \qquad d_T = o(T/\log T), \tag{18}$$

then $\mathrm{err}_{m,Q_T(\eta)}(f_T, s) \to 1$ for every fixed margin $m > 0$.

Negative curvature is therefore one capacity currency, not the only one: dimension, radius, margin, and readout regularity can substitute for it. The threshold is a capacity statement at fixed curvature; it does not treat increasing κ as computationally or numerically free, nor model curvature-dependent optimization difficulty.

4.2 A FINITE-SCALE CAPACITY GAP

When the ambient packing term is analytically available or has a certified bound, define

$$\Gamma_T(\mathcal{M}; \eta) := \frac{1}{T} \log N_T(\eta) - \frac{1}{T} \log P_{\mathcal{M}}(R_T, m/L_T). \tag{19}$$

Corollary 4 gives

$$\mathrm{err}_{m,Q_T(\eta)}(f_T, s) \geq 1 - e^{-T\Gamma_T(\mathcal{M};\eta)} \qquad \text{whenever } \Gamma_T(\mathcal{M}; \eta) > 0. \tag{20}$$

A positive gap gives a finite-scale, training-independent error lower bound; it forces $\mathrm{err}_{m,Q_T(\eta)} \to 1$ only when $T\Gamma_T(\mathcal{M}; \eta) \to \infty$. A non-positive gap merely removes this packing obstruction. Because $P_{\mathcal{M}}$ is an ambient-space capacity, we do not replace it empirically by packing of a learned finite point cloud.

5 WHEN CURVATURE PAYS THE RIGHT ENTROPY

For identification-compatible tasks, the bottleneck is necessary. To show that its curvature threshold is attainable at the relevant exponential scale, we now exhibit a concrete task whose Euclidean obstruction and hyperbolic construction are governed by the same exponential leaf-growth rate. These are independent statements; we do not claim minimax optimality or a tight Euclidean–hyperbolic equivalence.

Let $\mathcal{B}_{b,T}$ be the complete b-ary tree of depth T with unit edge lengths and leaf set $\mathcal{L}_T$. The leaf-identification task has $|\mathcal{L}_T| = b^T$ and task entropy $\log b$.

Theorem 9 (Tree-indexed identification: Euclidean obstruction and hyperbolic realization)**.** *Fix $b \geq 2$ and the leaf-identification task on $\mathcal{L}_T$.*

1. ***Euclidean lower bound.*** *If $f_T(\mathcal{L}_T) \subseteq B_2(0, aT) \subset \mathbb{R}^d$, then any L-Lipschitz readout that classifies a $(1-\varepsilon)$ fraction of leaves with margin m satisfies*

$$L \geq \frac{m}{2aT} \left[((1-\varepsilon)b^T)^{1/d} - 1 \right]_+ . \tag{21}$$

2. ***Hyperbolic realization.*** *If $a\sqrt{\kappa} > \log b$, there are constants C, T_0 and embeddings $\phi_T : \mathcal{L}_T \to \mathbb{H}^2_{-\kappa}$ with $\phi_T(\mathcal{L}_T) \subseteq B(o, aT)$ such that, for $T \geq T_0$,*

$$\min_{u \neq v} d_{\mathbb{H}}(\phi_T(u), \phi_T(v)) \geq 2 \left(a - \frac{\log b}{\sqrt{\kappa}} \right) T - C. \tag{22}$$

Hence the nearest-prototype scores $s_v(z) = -d_{\mathbb{H}}(z, \phi_T(v))$ are 1-Lipschitz and realize every fixed positive margin for all sufficiently large T.

The second part is a capacity construction, not a new low-distortion tree-embedding theorem. Index the b^T leaves arbitrarily and place their prototypes at equally spaced angles on the hyperbolic circle of radius aT. Their minimum angular separation is $\Theta(b^{-T})$, and the hyperbolic law of cosines gives (22). Both directions are governed by the same exponential leaf-growth rate b^T, but they are proved independently and do not establish minimax or tight Euclidean–hyperbolic optimality. The representation dimension is 2, but the identification readout still contains $N = b^T$ prototype scores;

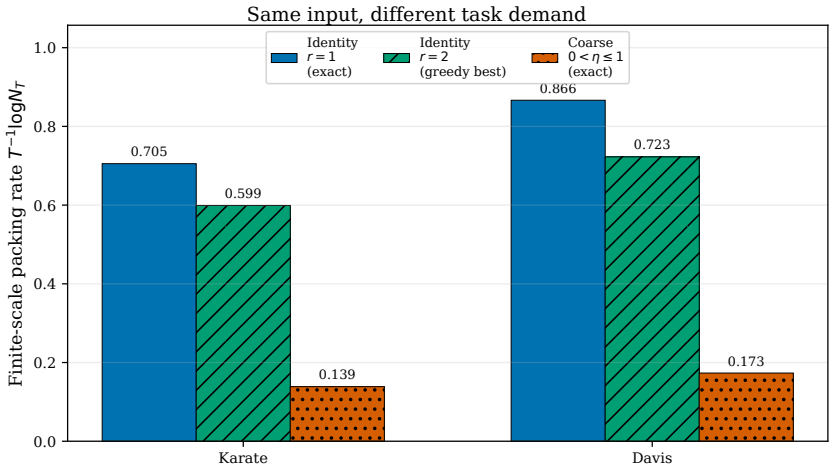

Figure 2: Same real input, different finite-scale task demand. On Karate and Davis, identity/retrieval uses shortest-path distance, whereas the coarse task uses the original two-class label pseudometric. Bars report $T^{-1}\log N_T$ where exact; the $r = 2$ identity bars use the best greedy lower estimate. This is a finite-scale diagnostic, not an estimate of a universal asymptotic rate.

capacity is not being hidden as constant output complexity. Classical results give stronger geometric guarantees when the full tree metric must be preserved (Bonk and Schramm, 2000; Sarkar, 2011); here only leaf distinguishability is needed.

A standard learning-capacity consequence is recorded separately in Appendix A.6: if a geometry forces the readout Lipschitz scale upward, the associated Lipschitz class is evaluated at correspondingly finer metric resolution. This observation is not needed for the deterministic bottleneck itself.

6 Same Input, Different Task

Input hyperbolicity cannot by itself determine the representation geometry required by a task.

Figure 2 gives a finite-scale analogue on two public graphs: the input is fixed while the task-induced packing demand changes sharply under label coarsening. The $r = 2$ identity bars use the best greedy lower estimate in place of exact N_T, as indicated in the legend; the coarse two-class packings are exact.

Theorem 10 (Same input, different capacity demand)**.** *For every $b \geq 2$, consider the same sequence of complete b-ary trees of depth T.*

1. *Leaf identification has task entropy $\log b$ and obeys the Euclidean exponential Lipschitz lower bound in Theorem 9.*
2. *Top-branch classification assigns one label to all leaves below each child of the root and uses the induced $0/1$ label pseudometric. For $0 < \eta \leq 1$, its task packing has size b for every T, so its asymptotic task entropy is zero. It is realized in $\mathbb{R}^b$ with constant radius, constant margin, and constant-Lipschitz linear scores, independently of T.*

The difference is not a change in the input graph but a change in which distinctions survive the loss. Euclidean dimension can grow, radius can expand, margins can shrink, or the readout can become more irregular; any of these resources can pay the same entropy. Conversely, sufficient hyperbolic packing does not guarantee successful optimization. The criterion is therefore a necessary-capacity diagnostic under declared budgets rather than an unconditional ordering of geometries.

Finite-scale use. For a finite dataset, one need not extrapolate an asymptotic exponent from a single sample. A practical task-side summary is the curve $\eta \mapsto h_T(\eta)$ in (3), evaluated over task-fixed resolutions. When the ambient geometry and budgets are analytically specified, (19) provides the

corresponding capacity obstruction. When they must be inferred from learned finite representations, we report task packing, margin, and radius separately instead of replacing $P_{\mathcal{M}}$ by packing of the observed points.

7 Empirical Evaluation

The experiments probe task dependence and expose the quantities that enter the capacity theorem. They do *not* hold the realized theoretical budgets (R_T, m, L_T) fixed across learned geometries. We match optimization and tuning budgets—data splits, dimension, seed set, training steps, and tuning-grid size—and report realized margins and radii separately. Stress or MAP differences are therefore descriptive geometry comparisons, not causal estimates of curvature alone.

For fixed test pairs $\mathcal{P}_{\text{test}}$, normalized stress is

$$\text{Stress} = \left(\frac{\sum_{(i,j)\in\mathcal{P}_{\text{test}}} (\widehat{d}_{ij} - d^{\star}_{ij})^2}{\sum_{(i,j)\in\mathcal{P}_{\text{test}}} (d^{\star}_{ij})^2} \right)^{1/2}. \tag{23}$$

Stress is a distance-fitting metric and is distinct from the packing-restricted margin error $\text{err}_{m,Q_T(\eta)}$ in Corollary 4. We report relative stress gain as $(\text{Stress}_E - \text{Stress}_H)/\,\text{Stress}_E$ and retrieval change as $\Delta\text{MAP} = \text{MAP}_H - \text{MAP}_E$.

7.1 Controlled WordNet intervention: same input, different tasks

We build a deterministic rooted tree from 800 WordNet noun synsets (Miller, 1995) and keep nodes, pair splits, dimension, training budget, and evaluation protocol fixed while changing only the target metric. The *hierarchy task* fits normalized tree distance. The *coarse task* maps the same nodes to $K = 12$ frozen equivalence classes and fits the corresponding $0/1$ pseudometric.

The task intervention changes finite-scale packing sharply. In the hierarchy experiment, r is simply the integer tree-distance instantiation of the theoretical resolution η; we evaluate $N_T(\eta)$ at $\eta = r$. Fifty greedy restarts give median lower estimates 800, 731.5, and 44 at $r = 1, 2, 3$, with the $r = 3$ range 30–46. These greedy values are conservative for the bottleneck because they can only underestimate the true maximum packing. For the coarse metric, $N_T(\eta) = 12$ exactly for $0 < \eta \le 1$.

We train $d = 5$ Euclidean and Poincaré embeddings with the same pair splits, five seeds, 1,500-step budget, and equal-size learning-rate tuning grids. On the hierarchy task, Poincaré reduces stress from 0.1521 ± 0.0009 to 0.1453 ± 0.0028, a relative gain of $4.47\% \pm 1.80\%$ that is positive in all five seeds; one-hop retrieval mean average precision (MAP) changes from 0.0411 to 0.0448. On the $K = 12$ coarse task, stress changes from 0.0999 ± 0.0052 to 0.0971 ± 0.0042 ($2.72\% \pm 5.71\%$), and prototype micro-accuracy changes from 0.9398 to 0.9440.

Because WordNet is trained by metric fitting rather than N-way margin identification, Figure 3(c) is a post-training identification diagnostic rather than a direct test of the theorem's antecedent. The curves are close, with Poincaré slightly lower over much of the intermediate-margin range but no uniform dominance, consistent with a capacity theorem that does not guarantee optimization success. The hierarchy-minus-coarse stress-gain difference is 1.76 ± 4.73 percentage points and positive in only three of five seeds; an unmerged depth-2 $K = 22$ sensitivity run reverses the mean cross-task ordering (-0.41 points). We therefore claim a reproducible task-packing intervention and a consistent hierarchy-task Poincaré stress advantage in the frozen $K = 12$ protocol, not a universal monotone relation between task entropy and empirical gain.

For the fixed prototype score $s_j(z) = -d(z, p_j)$, $L = 1$ exactly. On the hierarchy task, mean median prototype margins are 0.0807 (Euclidean) and 0.0809 (Poincaré), with mean maximum radii 0.7825 and 0.7295. Coarse-task margins, full margin-error curves, all seeds, and task-definition details are reported in Appendix C.

7.2 Cross-network learning and failure cases

We fit Euclidean and Poincaré representations to shortest-path targets on five public NetworkX graphs: Karate, Davis, Florentine, Les Misérables, and Kite. Every geometry–dimension pair uses

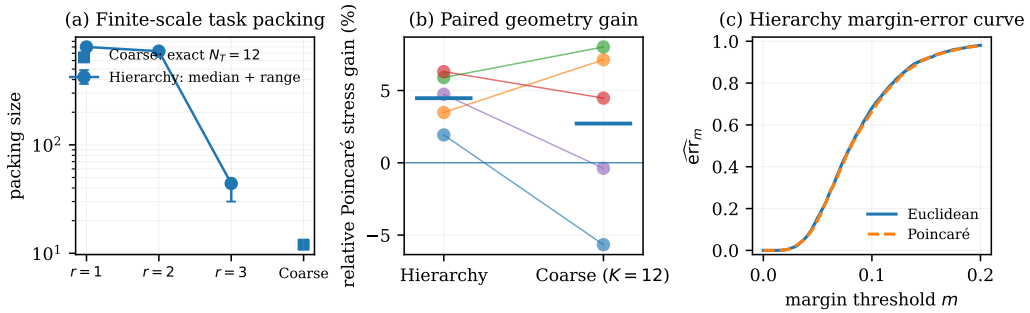

Figure 3: Controlled WordNet intervention on the same 800 nodes. (a) Finite-scale task packing; hierarchy values are 50-restart greedy lower estimates, while the coarse packing is exact. (b) Paired five-seed relative Poincaré stress gains. (c) The empirical hierarchy identification curve $m \mapsto \widehat{\mathrm{err}}_m$, computed directly from the frozen five-seed embeddings using the 1-Lipschitz score $s_j(z) = -d(z, p_j)$. The curves are descriptive learned-representation diagnostics and do not estimate ambient $P_{\mathcal{M}}$.

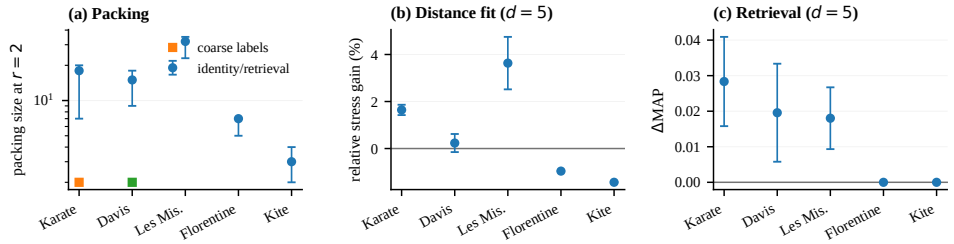

Figure 4: Cross-network evaluation at $d = 5$. (a) Median greedy packing at $r = 2$ with min–max restart ranges; Karate and Davis also show exact coarse-label packings. (b) Relative Poincaré stress gain and (c) change in neighbor-retrieval MAP, both as mean $\pm$ standard deviation over five paired seeds. Negative stress gains on Florentine and Kite are retained.

the same 250-step Adam budget and seeds 0–4. The main comparison uses $d = 5$; $d = 2$ is retained as a sensitivity study because several runs exhibit local-optimization outliers.

At $d = 5$, relative Poincaré stress gains are $+0.23\%$ (Davis), $+1.64\%$ (Karate), $+3.64\%$ (Les Misérables), -0.96% (Florentine), and -1.44% (Kite). MAP improves on Davis, Karate, and Les Misérables; both geometries reach MAP 1 on Florentine and Kite. Figure 4 reports positive and negative cases together with $r = 2$ greedy packing lower estimates.

We do not convert five heterogeneous graphs into a correlation claim. Greedy packing underestimates true demand, graph diameters are small, and increasing dimension itself changes the geometry comparison. The purpose is to establish the boundary condition central to the theory: negative curvature is not uniformly favorable even on graphs with nontrivial shortest-path structure. Appendix C reports the consolidated $d = 5$ summary, the $d = 2$ sensitivity, and full restart distributions.

8 Conclusion

The main lesson is a model-selection principle: compare geometry with the distinctions the *task* must preserve. Under a margin-identification restriction, unrestricted Lipschitz identification capacity equals metric packing capacity, yielding a bottleneck on the packed subset. Euclidean models may pay through dimension, radius, margin, or regularity; negative curvature is one alternative, not a universal winner. This capacity law does not predict optimization; fixed neural readouts and non-identification losses remain open for practical architectures and broader learning objectives alike.

REFERENCES

Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

Thomas Bläsius, Jean-Pierre von der Heydt, Maximilian Katzmann, and Nikolai Maas. Weighted embeddings for low-dimensional graph representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(15):15587–15595, 2025.

Mario Bonk and Oded Schramm. Embeddings of Gromov hyperbolic spaces. *Geometric and Functional Analysis*, 10:266–306, 2000.

Ngoc Bui, Menglin Yang, Runjin Chen, Leonardo Neves, Mingxuan Ju, Rex Ying, Neil Shah, and Tong Zhao. Learning along the arrow of time: Hyperbolic geometry for backward-compatible representation learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *PMLR*, pages 5842–5855, 2025.

Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. Hyperbolic graph convolutional neural networks. In *Advances in Neural Information Processing Systems*, 32, 2019.

Allison Davis, Burleigh B. Gardner, and Mary R. Gardner. *Deep South: A Social Anthropological Study of Caste and Class*. Univ of South Carolina Press, 2009.

Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. In *Advances in Neural Information Processing Systems*, 31, 2018.

Albert Gu, Frederic Sala, Beliz Günel, and Christopher Ré. Learning mixed-curvature representations in product spaces. In *International Conference on Learning Representations*, 2019.

Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference*, pages 11–15, 2008.

Neil He, Hiren Madhu, Ngoc Bui, Menglin Yang, and Rex Ying. Hyperbolic deep learning for foundation models: A survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6021–6031, 2025.

Pascal Mettes, Mina Ghadimi Atigh, Martin Keller-Ressel, Jeffrey Gu, and Serena Yeung. Hyperbolic deep learning in computer vision: A survey. *International Journal of Computer Vision*, 132(9):3484–3508, 2024.

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

Dionisia Naddeo, Jonas Linkerhägner, Nicola Toschi, Geri Skenderi, and Veronica Lachi. Hyperbolic graph neural networks under the microscope: The role of geometry-task alignment. *arXiv preprint arXiv:2602.01828*, 2026.

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

Maximilian Nickel and Douwe Kiela. Learning continuous hierarchies in the Lorentz model of hyperbolic geometry. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *PMLR*, pages 3779–3788, 2018.

John F. Padgett and Christopher K. Ansell. Robust action and the rise of the Medici, 1400–1434. *American Journal of Sociology*, 98(6):1259–1319, 1993.

Divit Rawal and Sriram Vishwanath. Minimax rates for hyperbolic hierarchical learning. *arXiv preprint arXiv:2601.20047*, 2026.

Frederic Sala, Christopher De Sa, Albert Gu, and Christopher Ré. Representation tradeoffs for hyperbolic embeddings. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *PMLR*, pages 4460–4469, 2018.

Rik Sarkar. Low distortion Delaunay embedding of trees in hyperbolic plane. In *Graph Drawing: 19th International Symposium*, volume 7034 of *LNCS*, pages 355–366. Springer, 2011.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Jens Van Nooten and Walter Daelemans. Jump to hyperspace: Comparing Euclidean and hyperbolic loss functions for hierarchical multi-label text classification. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4260–4273, 2025.

Melanie Weber. Neighborhood growth determines geometric priors for relational representation learning. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108 of *PMLR*, pages 266–276, 2020.

Wayne W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33(4):452–473, 1977.

Weilin Zhou. The natural geometry of code: Hyperbolic representation learning for program reasoning. In *International Conference on Learning Representations*, 2026.

Reproducibility statement

All assumptions for the theoretical statements are given in the main text, and complete proofs are provided in Appendix A. Appendix B specifies the finite-scale packing procedure, while Appendix C gives the WordNet construction, frozen task definitions, learning-rate selection rule, pair budgets, random seeds, stress definition, embedding objectives, margin/radius diagnostics, five-network restart protocol, and the reported aggregate results. Appendix D reports the additional theory-aligned and robustness experiments, including packing-restricted margin-error evaluation, the ten-seed WordNet extension, the controlled direct-identification capacity study, and the finite-precision sanity check, together with their fixed configurations and aggregation rules. The experiments use WordNet, public NetworkX graphs, and controlled synthetic identification tasks, and require only lightweight computation.

AI use statement

Generative AI tools were used only as assistive tools for writing, language editing, and figure preparation. Some schematic figures were drafted or polished with AI assistance; their scientific content, structure, labels, and conclusions were designed and verified by the authors. AI did not generate or infer any experimental data or results. The authors independently conducted the research, verified the theoretical statements and proofs, checked the cited literature, and reviewed all figures, values, and conclusions. The authors take full responsibility for the final content of the submission

A Proofs

A.1 Packing estimates

For a Euclidean ball $B_2(0,R) \subset \mathbb{R}^d$ with $d \geq 1$,

$$\operatorname{pack}(B_2(0,R), \|\cdot\|_2, r) \leq \left(1 + \frac{2R}{r}\right)^d. \tag{24}$$

This follows by placing disjoint $r/2$-balls around an r-packing and comparing volumes inside $B_2(0, R + r/2)$.

For hyperbolic space $\mathbb{H}^d_{-\kappa}$ with $d \geq 2$ and $\kappa > 0$,

$$\operatorname{vol}(B(R)) = C_{d,\kappa} \int_0^R \sinh^{d-1}(\sqrt{\kappa}\, t)\, dt = \exp\big\{(d-1)\sqrt{\kappa}\, R + O(1)\big\}. \tag{25}$$

Thus, for fixed $r > 0$,

$$\log \operatorname{pack}(B(R), d_{\mathbb{H}}, r) = (d-1)\sqrt{\kappa}\, R + O_{d,\kappa,r}(1). \tag{26}$$

If $r_T = m/L_T \to 0$ and $L_T = \exp(\beta T + o(T))$, local hyperbolic volume is Euclidean to first order at scale r_T, so the small-scale factor contributes $d\beta T + o(T)$ to the logarithm of the packing number. Therefore

$$\frac{1}{T} \log \operatorname{pack}(B_{\mathbb{H}^d_{-\kappa}}(aT), r_T) = a(d-1)\sqrt{\kappa} + d\beta + o(1), \tag{27}$$

while in Euclidean space the corresponding exponential rate is $d\beta + o(1)$.

A.2 Proof of Theorem 3 and Corollary 4

For the upper bound, suppose $z_1, \ldots, z_N$ in a radius-R ball are all margin-correct. For distinct i, j, margin correctness gives

$$s_i(z_i) - s_j(z_i) \geq m, \qquad s_i(z_j) - s_j(z_j) \leq -m. \tag{28}$$

Hence

$$\begin{aligned} 2m &\leq |[s_i(z_i) - s_j(z_i)] - [s_i(z_j) - s_j(z_j)]| && (29)\\ &\leq |s_i(z_i) - s_i(z_j)| + |s_j(z_i) - s_j(z_j)| && (30)\\ &\leq 2L\rho(z_i, z_j), && (31) \end{aligned}$$

so $\rho(z_i, z_j) \geq m/L$. The prototypes therefore form an (m/L)-packing, proving $\mathrm{Cap}_{\mathrm{id}}(R; L, m) \leq P_{\mathcal{M}}(R, m/L)$.

Conversely, let $\{z_1, \ldots, z_N\}$ be any (m/L)-packing in a radius-R ball and define

$$s_i(z) := -L\rho(z, z_i). \tag{32}$$

The reverse triangle inequality gives $|s_i(z) - s_i(z')| \leq L\rho(z, z')$. At z_i, $s_i(z_i) = 0$ while, for $j \neq i$, $s_j(z_i) = -L\rho(z_i, z_j) \leq -m$. Thus all N identities are margin-correct. Taking suprema proves (11).

For an identification-compatible task packing, the margin-correct images within $Q_T(\eta)$ form an (m/L)-packing. Since $\mathrm{err}_{m,Q_T(\eta)}$ is normalized by $N_T(\eta) = |Q_T(\eta)|$, this yields (12). To prove tightness, let $K = \min\{N, P_{\mathcal{M}}(R_T, m/L)\}$. Choose K separated prototypes and the distance scores above. If $K < N$, set every remaining score coordinate identically to $-m$ and map the remaining task states arbitrarily into the ball. The first K identities are therefore margin-correct, which gives

$$\inf_{f,s} \mathrm{err}_m(f, s) \leq 1 - \frac{K}{N}.$$

Corollary 4 gives the reverse inequality. Hence

$$\inf_{f,s} \mathrm{err}_m(f, s) = 1 - \frac{K}{N},$$

which proves (13).

A.3 Proofs of Corollaries 5–8

Equation (24) and Corollary 4 imply

$$\left(1 + \frac{2aTL}{m}\right)^d \geq (1 - \varepsilon)e^{hT}, \tag{33}$$

which rearranges to (15). For Corollary 6, set $R = aT$ in the hyperbolic packing asymptotics above. If m/L is fixed, the representation packing rate is $g = a(d-1)\sqrt{\kappa}$. If $g < \underline{h}_{\mathcal{T}}(\eta)$, then for some $\delta > 0$ and all sufficiently large T, $N_T(\eta) \geq \exp((g+\delta)T)$ while $P_{\mathcal{M}}(aT, m/L) = \exp(gT + o(T))$, so (12) tends to one. If $g > \overline{h}_{\mathcal{T}}(\eta)$, then the reverse exponential inequality holds eventually and the packing lower bound becomes vacuous. If instead $L_T = \exp(\beta T + o(T))$, a necessary exponential-rate condition for non-vanishing margin accuracy on $Q_T(\eta)$ is

$$a(d-1)\sqrt{\kappa} + d\beta \geq \underline{h}_{\mathcal{T}}(\eta). \tag{34}$$

For Corollary 7, applying (24) at radius R_T and separation m/L_T yields

$$\left(1 + \frac{2R_T L_T}{m}\right)^{d_T} \geq \exp(hT - o(T)), \tag{35}$$

which gives (17) after taking logarithms. Under the polynomial assumptions of Corollary 8, the logarithmic factor is $O(\log T)$ and $d_T = o(T/\log T)$, so the left-hand side of (17) is $o(T)$, a contradiction to non-vanishing margin accuracy on $Q_T(\eta)$.

A.4 Proof of the finite-scale gap bound

From Corollary 4, at most $P_{\mathcal{M}}(R_T, m/L_T)$ of the $N_T(\eta)$ states in $Q_T(\eta)$ can be margin-correct. Therefore

$$\mathrm{err}_{m,Q_T(\eta)}(f_T, s) \geq 1 - \frac{P_{\mathcal{M}}(R_T, m/L_T)}{N_T(\eta)} = 1 - e^{-T\Gamma_T(\mathcal{M};\eta)}, \tag{36}$$

whenever $\Gamma_T(\mathcal{M};\eta) > 0$.

A.5 Proof of Theorem 9

Part (1) follows from Corollary 5 with $N_T = b^T$.

For part (2), let $N = b^T$, set $R = aT$, and place the N prototypes at equally spaced angles

$$\theta_k = \frac{2\pi k}{N}, \qquad k = 0, \ldots, N-1, \tag{37}$$

on the hyperbolic circle of radius R in $\mathbb{H}^2_{-\kappa}$. Thus the minimum angular separation is $\Delta\theta = 2\pi b^{-T}$. For two points at the same radius, the hyperbolic law of cosines gives

$$\cosh(\sqrt{\kappa} d_{\mathbb{H}}) = 1 + \sinh^2(\sqrt{\kappa} R)(1 - \cos\Delta\theta). \tag{38}$$

For all sufficiently large T, $1 - \cos\Delta\theta \geq c_1 b^{-2T}$ and $\sinh^2(\sqrt{\kappa} aT) \geq c_2 e^{2a\sqrt{\kappa}T}$ for constants $c_1, c_2 > 0$. Hence

$$\cosh(\sqrt{\kappa} d_{\mathbb{H}}) \geq 1 + c_3 \exp\{2(a\sqrt{\kappa} - \log b)T\}. \tag{39}$$

Using $\mathrm{arcosh}(1 + x) \geq \log x - C_0$ for large x, we obtain

$$d_{\mathbb{H}} \geq 2\left(a - \frac{\log b}{\sqrt{\kappa}}\right) T - C \tag{40}$$

for a constant C independent of T. When $a\sqrt{\kappa} > \log b$, the minimum separation therefore grows linearly. Each score $s_v(z) = -d_{\mathbb{H}}(z, \phi_T(v))$ is 1-Lipschitz by the triangle inequality; at prototype $\phi_T(v)$, its margin over every other prototype equals the corresponding pairwise distance. Thus every fixed positive margin is achieved for all sufficiently large T.

A.6 Lipschitz readouts and fat-shattering

The next standard consequence is included only to connect the geometric resource tradeoff with a familiar learning-capacity quantity; it is not used in Theorem 3 or Corollary 4.

Proposition 11 (Lipschitz scale and fat-shattering)**.** *Let (Z, d) be a metric space and $\mathcal{G}_L = \{g : Z \to \mathbb{R} : \mathrm{Lip}(g) \leq L\}$. For every $\gamma > 0$,*

$$\mathrm{fat}_\gamma(\mathcal{G}_L) \geq \mathrm{pack}\left(Z, d, \frac{4\gamma}{L}\right). \tag{41}$$

Proof. Let $\{z_1, \ldots, z_M\}$ be a $4\gamma/L$-separated subset of Z. For any signs $\sigma_i \in \{-1, +1\}$, assign $y_i = 2\gamma\sigma_i$. Since

$$|y_i - y_j| \leq 4\gamma \leq L d(z_i, z_j), \tag{42}$$

the assignment is L-Lipschitz on the finite subset. The McShane extension theorem extends it to an L-Lipschitz function on all of Z, and $\sigma_i g(z_i) \geq 2\gamma$. Thus the points are γ-fat-shattered at threshold zero. Taking the largest separated subset proves (41).

A.7 Proof of Theorem 10

Leaf identification has b^T distinguishable states, so its task entropy is $\log b$ and Theorem 9 applies. For top-branch classification, let $y(v) = v_1 \in \{0, \dots, b-1\}$. The number of labels is b independent of T, hence

$$\lim_{T\to\infty} \frac{1}{T} \log b = 0. \tag{43}$$

Embed every leaf with first branch j as the standard basis vector $e_j \in \mathbb{R}^b$. Linear scores $s_j(z) = z_j$ classify all points with constant radius, constant margin, and constant Lipschitz constant.

B Estimating Finite-Scale Task Demand

For a finite sample S and a task metric or proxy distance d_T, compute the packing curve

$$\eta \mapsto \log \operatorname{pack}(S, d_T, \eta) \tag{44}$$

using exact packing for small graphs, greedy farthest-point selection, or approximate nearest-neighbor methods for larger data. When no single threshold is canonical, the curve over a task-relevant range of η is more informative than a single selected value. If a scale parameter is explicit, estimate the slope over the range used by the loss; for graph radius, one may also estimate shell growth or task-packing growth over radii.

The metric should be tied to the evaluation. Hierarchy-aware losses suggest label-tree distance; retrieval suggests relevance or target shortest-path distance; structured prediction suggests a target-space metric such as edit or transport distance. For a flat cross-entropy objective with a fixed small label set, the induced task entropy is that of the label set rather than the full input hierarchy.

C Experiment Details

C.1 WordNet-derived same-input experiment

We use 800 noun synsets from a deterministic rooted tree derived from WordNet (Miller, 1995). Starting at `entity.n.01`, breadth-first traversal is deterministic; when a synset has multiple hypernym parents, a fixed lexical tie-break selects one parent, producing a rooted spanning tree. The resulting tree has diameter $T = 7$. The hierarchy task uses normalized tree shortest-path distance. Train/validation/test pair sets are fixed and shared by the two geometries; the formal training runs use 30,000 train pairs, $d = 5$, Adam, 1,500 maximum steps, validation every 25 steps, and seeds 0–4. Each geometry receives the same two-value learning-rate tuning budget on seed 0, after which the selected rate is frozen: 0.03 for Euclidean and 0.015 for Poincaré.

All embedding experiments use the normalized-stress definition in Equation (23) on the fixed test-pair set.

For the coarse task, a direct depth-2 grouping produces too many small classes for the intended 5–20-class control. Before learning-rate selection and all formal five-seed runs, rare depth-2 branches and shallow nodes were deterministically merged into an OTHER class, giving $K = 12$ classes with sizes $365, 187, 63, 38, 27, 25, 20, 18, 17, 16, 15, 9$. This frozen definition is the one reported in the main paper. A separate unmerged depth-2 $K = 22$ sensitivity run yields a mean hierarchy-minus-coarse stress-gain difference of -0.41 percentage points; we disclose it because the cross-task gain comparison is sensitive to label granularity.

Task-H packing uses 50 greedy maximal-separated-set restarts for $\rho \in \{0.1, 0.2, 0.3, 0.4\}$ with $r = \max(1, \operatorname{round}(\rho T))$. Here the discrete threshold r is the experimental value of the theoretical resolution η. The distinct thresholds are $r = 1, 2, 3$, for which the median packing sizes are 800, 731.5, and 44; the corresponding ranges are 800–800, 671–754, and 30–46. For the $0/1$ coarse pseudometric, the exact packing is $K = 12$ for $0 < \eta \le 1$.

For a fixed prototype p_j, the score $s_j(z) = -d(z, p_j)$ is 1-Lipschitz in either geometry. On the hierarchy task, each learned node embedding is its own prototype; the mean median margin is 0.0807 (Euclidean) and 0.0809 (Poincaré), with mean fifth percentiles 0.0377 and 0.0379. On the coarse task, class medoids define the prototype readout; mean median margins are 0.9650 and 0.9594,

Table 2: WordNet-derived learned-representation diagnostics (mean $\pm$ standard deviation over five seeds where applicable). Relative stress gain is $(\mathrm{Stress}_E - \mathrm{Stress}_H)/\mathrm{Stress}_E$.

Task	Stress E/H	Rel. gain	Downstream E/H	Max radius E/H
Hierarchy	$.1521 \pm .0009$ / $.1453 \pm .0028$	$4.47\% \pm 1.80$	MAP .0411/.0448	.7825/.7295
Coarse $K = 12$	$.0999 \pm .0052$ / $.0971 \pm .0042$	$2.72\% \pm 5.71$	Acc. .9398/.9440	.8420/.8056

Table 3: Consolidated finite-scale empirical summary at $d = 5$. Packing values are task-side measurements; no cross-dataset scaling law is assumed.

Setting	Packing	Stress gain	Downstream change
WordNet hierarchy	44 med. (30–46), $r = 3$	$+4.47\% \pm 1.80$	MAP .0411 $\rightarrow$ .0448
WordNet coarse	12 exact	$+2.72\% \pm 5.71$	Acc. .9398 $\rightarrow$ .9440
Karate	18 med. (7–20), $r = 2$	$+1.64\%$	ΔMAP $+.0284$
Davis	15 med. (9–18), $r = 2$	$+0.23\%$	ΔMAP $+.0196$
Les Misérables	32 med. (23–35), $r = 2$	$+3.64\%$	ΔMAP $+.0180$
Florentine	7 med. (5–7), $r = 2$	-0.96%	ΔMAP .0000
Kite	3 med. (2–4), $r = 2$	-1.44%	ΔMAP .0000

while fifth percentiles are -0.0674 and -0.0103. These values are representation diagnostics, not out-of-sample generalization guarantees. Table 2 collects the corresponding learned-representation measurements in one place.

Figure 5 reports the full empirical margin-error curves computed from the frozen embeddings, without retraining or post-hoc threshold selection. The hierarchy curve uses each node as its identity prototype; the coarse curve uses the frozen class-medoid rule.

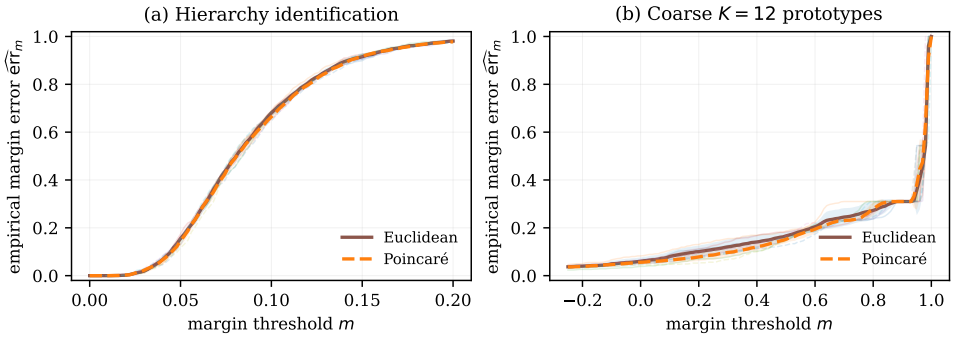

Figure 5: Full five-seed empirical margin-error curves. Thin seed-level variation is summarized by the mean and one-standard-deviation bands. These curves measure $\widehat{\mathrm{err}}_m$ for learned representations; they do not substitute observed-point packing for the ambient capacity $P_{\mathcal{M}}$.

C.2 Five-network robustness study

We use the NetworkX implementations (Hagberg et al., 2008) of Karate Club (Zachary, 1977), Davis Southern Women (Davis et al., 2009), Florentine Families (Padgett and Ansell, 1993), Les Misérables, and Krackhardt Kite. All five graphs are connected. Their node counts are $34, 32, 15, 77, 10$ and diameters are $5, 4, 5, 5, 4$, respectively. Identity/retrieval distance is unweighted shortest-path distance. Karate uses the club split and Davis uses the bipartite type as coarse labels; no artificial coarse labels are assigned to the other graphs.

For each nominal $\rho \in \{0.1, 0.2, 0.3, 0.4\}$, we set $r = \max(1, \lfloor \rho T + 0.5 \rfloor)$ and run 50 deterministic greedy packing restarts. Since the graph diameters are only four or five, these nominal resolutions collapse to $r = 1$ or $r = 2$. The $r = 1$ packing equals the number of nodes exactly. The nontrivial $r = 2$ restart summaries are reported in Table 4.

Table 4: Greedy packing stability at $r = 2$. The best value is an empirical lower bound on the true packing number.

Graph	Mean ± SD	Median	Best (min–max)
Karate	16.82 ± 3.71	18	20 (7–20)
Davis	14.42 ± 2.44	15	18 (9–18)
Florentine	6.66 ± 0.63	7	7 (5–7)
Les Misérables	31.76 ± 2.72	32	35 (23–35)
Kite	3.32 ± 0.65	3	4 (2–4)

Euclidean and Poincaré embeddings are optimized for 250 Adam steps with matched optimization budgets, dimensions $d \in \{2, 5\}$, and seeds 0–4. All 100 runs complete without numerical failure. The $d = 5$ results used for the main-text robustness summary are reported in Table 5.

Table 5: Five-network results at $d = 5$ (means over five seeds). Positive relative stress gain favors Poincaré.

Graph	Stress E/H	Rel. gain	MAP E/H	ΔMAP
Davis	.1793/.1789	+0.23%	.6666/.6861	+.0196
Karate	.1110/.1092	+1.64%	.8716/.9000	+.0284
Les Misérables	.1171/.1128	+3.64%	.7485/.7665	+.0180
Florentine	.0788/.0796	−0.96%	1.000/1.000	.0000
Kite	.0481/.0488	−1.44%	1.000/1.000	.0000

At $d = 2$, several runs are seed-sensitive: in particular, Florentine, Kite, and Les Misérables contain local-optimization outliers, and Les Misérables changes the sign of its mean stress comparison relative to its median. We therefore use the $d = 2$ results only as a sensitivity analysis rather than a geometry ranking. The fixed-step budget is intentionally uniform across geometries; some $d = 2$ runs still improve near step 250, which is another reason not to interpret small mean differences as definitive wins or losses.

C.3 Supplementary figures and dimension sensitivity

Figure 6 shows the full five-network packing profiles rather than only the $r = 2$ summary used in the main text. Because the graph diameters are four or five, multiple nominal ρ values collapse to the same integer threshold; repeated points should not be interpreted as independent resolutions. Figure 2 in the main text isolates the same-input task intervention on Karate and Davis.

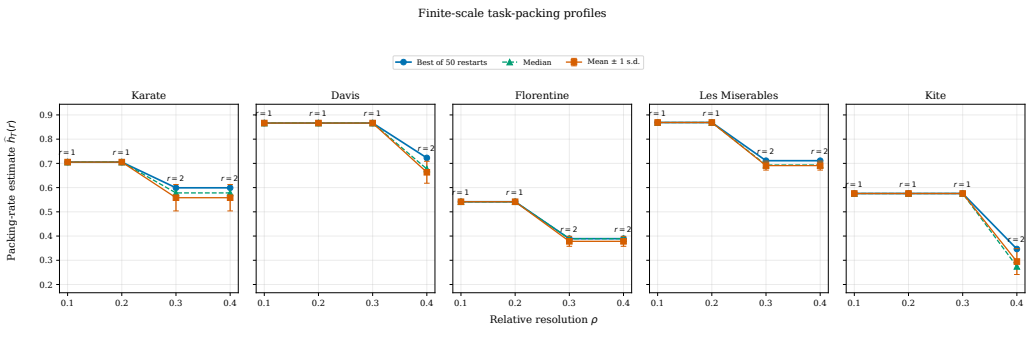

Figure 6: Five-network finite-scale packing profiles from 50 greedy restarts. The labels above points show the rounded integer separation r. Best values are lower bounds on the true packing number; restart spread is reported rather than hidden.

At $d = 2$, local-optimization variability is visibly larger. Table 6 reports the full mean ± standard deviation rather than selecting favorable seeds. Davis and Karate retain small positive stress gains;

Les Misérables has a negative mean stress gain even though its MAP change is positive. This dimension sensitivity is one empirical example of the resource substitution predicted by Corollary 7.

Table 6: Five-network sensitivity at $d = 2$ (mean $\pm$ standard deviation over five seeds). Positive stress gain favors Poincaré.

Graph	Stress E/H	Rel. stress gain	MAP E/H
Davis	$.2923 \pm .0052$ / $.2873 \pm .0063$	$+1.70\%$	$.4950 \pm .0127$ / $.4941 \pm .0111$
Karate	$.2083 \pm .0082$ / $.2053 \pm .0090$	$+1.47\%$	$.5677 \pm .0213$ / $.5761 \pm .0315$
Florentine	$.1496 \pm .0338$ / $.1484 \pm .0331$	$+0.80\%$	$.8456 \pm .0454$ / $.8743 \pm .0243$
Kite	$.0949 \pm .0410$ / $.0928 \pm .0371$	$+2.25\%$	$.9854 \pm .0326$ / $.9844 \pm .0349$
Les Misérables	$.2577 \pm .0122$ / $.2665 \pm .0257$	-3.40%	$.5137 \pm .0141$ / $.5212 \pm .0243$

D Additional Theory-Aligned and Robustness Experiments

This appendix reports four complementary checks. First, we evaluate the margin-error object on frozen task states rather than on the full learned point cloud. Second, we extend the WordNet runs from five to ten paired optimization seeds. Third, we study a controlled direct-identification problem in which task demand, radius, dimension, margin, and the readout Lipschitz constant are explicit. Finally, we separate the asymptotic hyperbolic construction from numerical cancellation in finite precision. None of these experiments estimates the ambient packing number $P_{\mathcal{M}}$ from learned embeddings, and none is used as an empirical proof of the exact capacity result.

D.1 Packing-Restricted Margin Error on Frozen WordNet Embeddings

For the hierarchy task, we keep the previously frozen tree-distance threshold $r = 3$ and the same 50 greedy separated-set restarts. The resulting subsets contain 30–46 states (median 44). Because the greedy procedure need not return a maximum packing, we treat these sets as fixed packing lower-estimate subsets rather than as exact maximizers. For a frozen embedding and a subset $\widetilde{Q}$, we use the distance score

$$s_j(z) = -d(z, z_j),$$

which is 1-Lipschitz in each coordinate, and evaluate

$$\widehat{\mathrm{err}}_{m,\widetilde{Q}} = \frac{1}{|\widetilde{Q}|} \sum_{i \in \widetilde{Q}} \mathbf{1} \left\{ \min_{j \in \widetilde{Q},\, j \neq i} d(z_i, z_j) < m \right\}.$$

Within each optimization seed, we first average the error over the 50 frozen separated sets and then report the mean and sample standard deviation across ten seeds. For the coarse task, the task states are the 12 frozen class medoids, each with equal weight; its error is therefore normalized over 12 class states rather than over the 800 original nodes.

Figure 7 shows a small but nonuniform separation between the two geometries. On the hierarchy task, at $m = 0.15$ the mean error is 0.159 ± 0.022 for Euclidean embeddings and 0.144 ± 0.010 for Poincaré embeddings; at $m = 0.20$ the corresponding values are 0.333 ± 0.023 and 0.319 ± 0.011. The per-seed ordering is not uniform: the Euclidean-minus-Poincaré error difference is positive in 8/10 seeds at $m = 0.15$ and 7/10 seeds at $m = 0.20$. For the 12-state coarse task, at $m = 0.50$ the mean errors are 0.350 ± 0.123 and 0.308 ± 0.088, respectively, but the seedwise difference is positive in only 4/10 seeds, negative in 2/10, and tied in 4/10. Thus the packing-restricted curves are useful as a direct finite-sample diagnostic of the margin-error object, while they do not support a uniform geometry ranking.

D.2 Ten-Seed WordNet Robustness

We next extend the frozen WordNet protocol from seeds 0–4 to seeds 0–9. The node set, pair split, dimension $d = 5$, Poincaré curvature -1, 30,000 training pairs, learning rates, validation schedule, and maximum 1,500 steps are unchanged. All 40 task–geometry–seed runs completed with status `ok`. A replay audit of the archived runs under the current environment passed all 20 checks, with maximum absolute numerical discrepancy 1.11×10^{-16}.

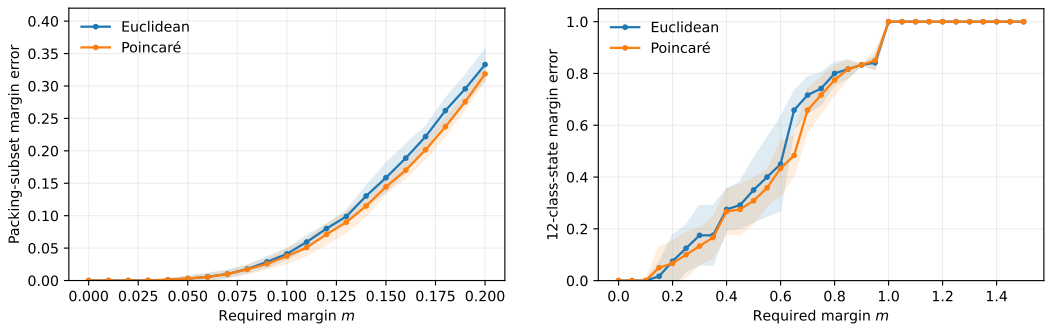

Figure 7: Packing-restricted margin diagnostics on frozen WordNet embeddings. Left: hierarchy task with $r = 3$; each seed averages over the same 50 frozen separated-set restarts, and the band is ± 1 sample SD across ten optimization seeds. Right: 12 frozen coarse class states with equal class-state weight, again shown as mean ± 1 sample SD across ten seeds. These curves evaluate learned representations on frozen task states; they do not estimate the ambient packing capacity.

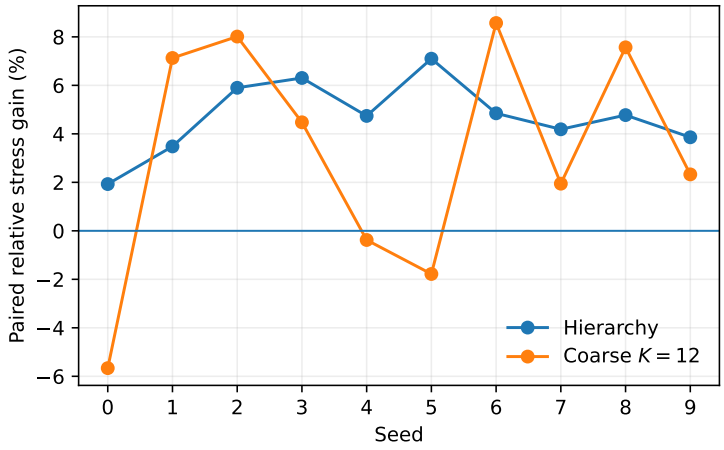

Figure 8: Per-seed relative stress gain of Poincaré over Euclidean embeddings under the frozen ten-seed protocol. Hierarchy gains are positive in all ten seeds, whereas the coarse task remains substantially more variable.

Table 7: Ten-seed WordNet robustness. Values are mean ± sample SD across paired seeds. Relative stress gain is computed per seed as $(\mathrm{Stress}_E - \mathrm{Stress}_H)/\mathrm{Stress}_E$ before aggregation.

Task	Euclidean stress	Poincaré stress	Rel. gain	Positive	Downstream metric (E / H)
Hierarchy	0.15241 ± 0.00097	0.14523 ± 0.00246	$4.71\% \pm 1.49\%$	$10/10$	MAP: 0.04040 ± 0.00119 / 0.04430 ± 0.00194
Coarse $K = 12$	0.10106 ± 0.00576	0.09761 ± 0.00331	$3.22\% \pm 4.79\%$	$7/10$	Acc.: 0.94125 ± 0.01242 / 0.94513 ± 0.00927

The hierarchy result is stable under additional optimizer randomness: all ten seeds have positive relative stress gain, and Poincaré MAP exceeds Euclidean MAP in all ten seeds. The coarse task remains less stable: its stress gain is positive in 7/10 seeds and its micro-accuracy difference is positive in 5/10 seeds. The paired hierarchy-minus-coarse stress-gain difference is only 1.49 ± 4.60 percentage points and is positive in 6/10 seeds. We therefore use the ten-seed results as an optimization-robustness check, not as evidence that hierarchy must benefit more than the coarse task.

Table 8: Ten-seed geometric diagnostics on WordNet. Values are mean ± sample SD across seeds. These quantities describe the learned embeddings and are not estimates of ambient packing capacity.

Task	Geometry	Median margin	5th-pct. margin	Max radius
Hierarchy	Euclidean	0.08066 ± 0.00173	0.03726 ± 0.00195	0.78057 ± 0.00662
Hierarchy	Poincaré	0.08044 ± 0.00179	0.03680 ± 0.00277	0.72940 ± 0.00268
Coarse $K = 12$	Euclidean	0.96469 ± 0.00909	-0.08705 ± 0.13296	0.84325 ± 0.01498
Coarse $K = 12$	Poincaré	0.95867 ± 0.01122	-0.03444 ± 0.10370	0.80988 ± 0.01635

D.3 Controlled Direct-Identification Capacity Transition

To isolate the capacity mechanism from the WordNet optimizer and task construction, we use a synthetic b-ary identification problem with $b = 3$, depths $T \in \{3, 4, 5, 6\}$, and $N = 3^T$ identities. We fix the required margin to $m = 0.5$ and use distance scores $s_i(z) = -d(z, p_i)$, hence $L = 1$. The radius is $R = aT$ with $a \in \{0.5, 1, 1.5\}$. Euclidean runs use $d \in \{2, 4, 8, 16\}$; hyperbolic runs use $\mathbb{H}^2_{-1}$. Each cell is repeated for ten seeds with the same maximum training budget. Across all 600 runs, 530 reach zero loss before the maximum step, 70 reach the 1,500-step limit, and no run exhibits nonfinite values or radius violations.

For Euclidean space, we mark a cell as *certified impossible* only when the standard packing upper bound

$$P_{\mathbb{R}^d}(R, m) \le \left(1 + \frac{2R}{m}\right)^d$$

is already smaller than N. This is deliberately one-sided: cells not certified impossible may still be difficult or infeasible. For the hyperbolic construction, we separately mark cells in which the explicit circle construction has minimum pairwise separation at least m.

The controlled experiment shows the predicted resource dependence at finite scale. At $R = T$ and $d = 2$, Euclidean mean errors are 0, 0, 0.039 ± 0.080, and 1.000 for $T = 3, 4, 5, 6$, whereas hyperbolic $d = 2$ has zero error for all four depths. At $T = 6$, the Euclidean upper bound is $625 < N = 729$, yielding the certified lower bound $1 - 625/729 = 0.143$, while the explicit hyperbolic circle construction has minimum separation $1.572 > m$ and the empirical error is zero. This comparison should not be read as a universal Euclidean–hyperbolic ranking. With $R = 1.5T$, both $d = 2$ geometries achieve zero error over the tested depths; with $R = 0.5T$, both low-dimensional geometries eventually fail. Euclidean dimensions $d = 4, 8, 16$ achieve zero error throughout the tested grid, including $R = 0.5T$. Thus radius, dimension, and curvature can each pay part of the identification demand.

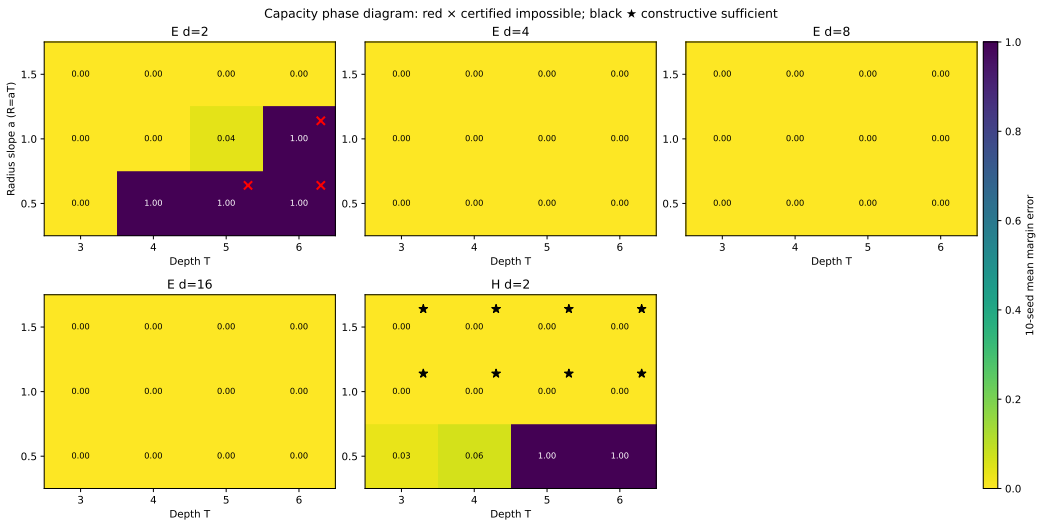

Figure 9: Controlled capacity phase diagram. Numbers are ten-seed mean margin errors. Crosses mark Euclidean cells that are certified impossible by the packing upper bound; stars mark hyperbolic cells for which the explicit circle construction is sufficient at the required margin. The empirical optimizer is reported separately from these certificates.

D.4 Finite-Precision Sanity Check

The asymptotic circle construction involves exponentially small angular gaps. We therefore test whether the apparent finite-depth floor comes from the geometry or from a numerically unstable formula. For a binary tree, let

$$\Delta\theta_T = 2\pi\, 2^{-T}, \qquad T = 5, \ldots, 60.$$

We compare the naive computation $1 - \cos(\Delta\theta_T)$ with the algebraically equivalent stable computation $2\sin^2(\Delta\theta_T/2)$ in float32 and float64, using 120-digit arithmetic as the reference.

The naive angular gap first rounds to zero at $T = 15$ in float32 and $T = 30$ in float64; from those depths onward it also produces margin-decision mismatches. In contrast, the stable formula has no rounded-to-zero gap, nonfinite value, or margin-decision mismatch through $T = 60$ in either precision. This experiment does not claim numerical stability at arbitrary depth; it only shows that the early floor obtained from direct subtraction is avoidable and should not be interpreted as a limitation of the underlying metric construction.

Reproducibility and interpretation. The experiments were run with Python 3.10.20, PyTorch 2.5.1+cu121, CUDA 12.1, and an NVIDIA GeForce RTX 3090. The WordNet experiments retain the frozen $K = 12$ coarse labels, $d = 5$, and Poincaré curvature -1. The controlled experiments use only the preregistered grids described above. Failed or maximum-step runs are retained rather than filtered. Error bands in this appendix are descriptive sample standard deviations over optimization seeds, not dataset-level confidence intervals. The previously reported $K = 22$ sensitivity remains a separate task-definition check and is not overwritten by the $K = 12$ ten-seed extension.

E Additional Scope Discussion

The exact identification theorem is deliberately scoped to a specific Lipschitz readout model, allowing geometric capacity to be separated from optimization and statistical effects. Its conclusions therefore characterize the capacity required to realize task-induced distinctions rather than universal sample-complexity or optimization guarantees for arbitrary hyperbolic architectures. Within this scope, curvature is one of several geometric resources: Euclidean dimension, radius, margin, and readout regularity can modify the same capacity balance. Finite-scale packing values are likewise defined relative to the chosen task pseudometric and resolution, making empirical comparisons

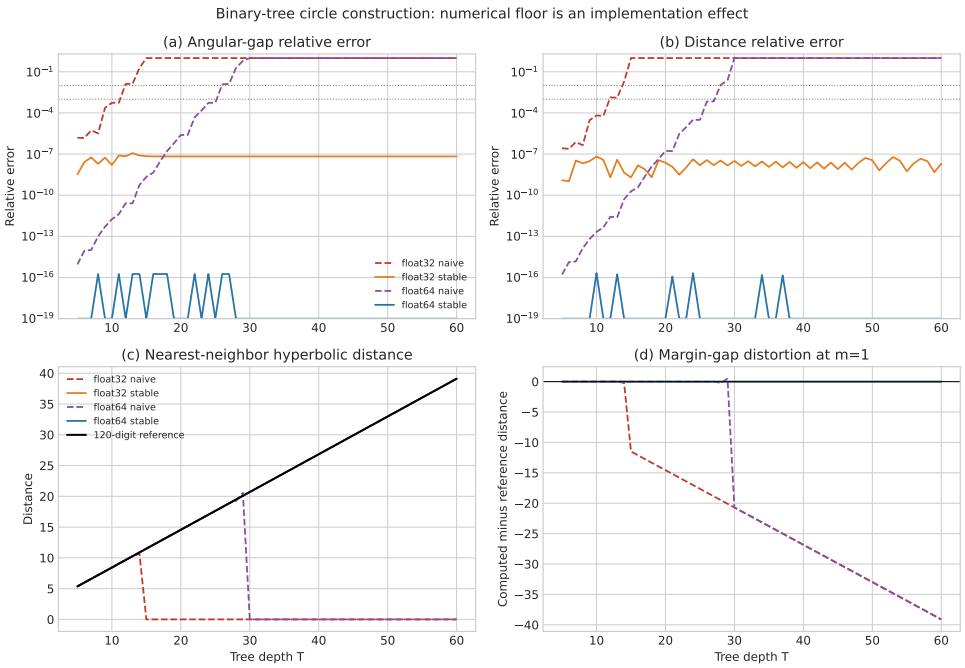

Figure 10: Finite-precision sanity check for the binary-tree circle construction. The naive $1 - \cos(\Delta\theta)$ computation loses the angular gap at $T = 15$ in float32 and $T = 30$ in float64. The stable $2\sin^2(\Delta\theta/2)$ computation remains nonzero and produces no margin-decision mismatch through $T = 60$. The observed floor is therefore an implementation effect for this computation, not a mathematical capacity boundary.

most informative when these choices are fixed and reported explicitly. The coarse-label sensitivity in WordNet and the dimension-dependent five-network results are consistent with this resource-dependent view and help delineate the regime in which the capacity analysis is informative.