Fractal Embeddings: Hierarchy-Aligned Prefix Supervision for Steerable Semantic Granularity

Devansh
Independent Researcher
devansh@svam.com

Abstract

Matryoshka Representation Learning (MRL) trains embeddings that support dimensional truncation, but all prefix lengths encode the same semantic content at varying fidelity—there is no mechanism to *steer* between coarse and fine meaning. We introduce **Fractal Embeddings**, which align prefix supervision with label hierarchy: short prefixes (64d) learn coarse labels while full embeddings (256d) learn fine labels, converting truncation into *semantic zoom* at zero additional inference cost. Four controlled ablations establish that *alignment*—not hierarchy awareness—causally drives steerability: inverting alignment reverses its sign; removing it or applying uniform multi-task training collapses it to zero (all $p_{\text{adj}} < 10^{-3}$, $d \geq 6.1$ on CLINC; directionally consistent on TREC with $n = 3$). Across eight datasets spanning conditional entropies from 1.23 to 5.05 bits, a random-effects meta-analysis yields pooled $d = 1.49$ ($p = 0.0003$), with all eight favouring fractal training (sign test $p = 0.004$). On CLINC-150, steerability reaches $\mathcal{S} = +0.150 \pm 0.028$ versus MRL's $+0.007 \pm 0.016$ ($p_{\text{adj}} = 0.004$, $d = 4.3$). A synthetic hierarchy experiment identifies a Goldilocks optimum where steerability peaks at capacity–demand matching (quadratic $R^2 = 0.964$), and steerability magnitude across real datasets is predicted by the product of hierarchy depth and baseline learnability ($\rho = 0.90$, $p = 0.002$). A backbone fine-tuning control on three datasets confirms that alignment is both necessary and sufficient: a flat MRL baseline with $4.5\times$ more trainable parameters (backbone fine-tuning) produces near-zero steerability ($d = 11.6$, 7.4, and 4.4 vs. head-only V5 on CLINC, DBPedia Classes, and TREC respectively). Replication across three encoder families confirms architecture invariance. We ground these findings in the classical theory of successive refinement, connecting V5's prefix cross-entropy loss to the log-loss universality theorem and deriving a Goldilocks capacity–demand optimum and product scaling law from information-theoretic principles.

1 Introduction

When a 256-dimensional embedding is truncated to 64 dimensions, what changes? Under standard Matryoshka training [Kusupati et al., 2022], the answer is *fidelity*: the shorter vector approximates the full one with reduced precision but encodes the same kind of semantic information. We argue this is a missed opportunity. Real-world semantics are inherently hierarchical—a query like "What is the capital of France?" belongs simultaneously to the coarse category LOCATION and the fine category CITY—and an embedding that supports truncation *already has the interface* for multi-resolution access. What it lacks is the training signal to make truncation semantically meaningful.

40th Conference on Neural Information Processing Systems (NeurIPS 2025).

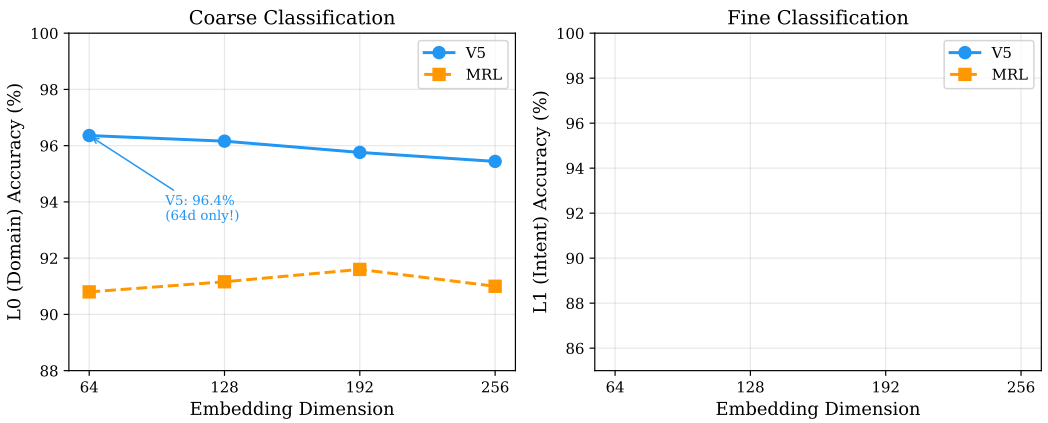

Figure 1: CLINC-150: V5 and MRL achieve comparable accuracy at full embedding length (256d), but V5's short prefixes (64d) specialize for coarse semantics while MRL's do not. This prefix specialization enables *semantic steering* via dimensional truncation.

We introduce **Fractal Embeddings**, a simple modification to MRL training that aligns prefix supervision with semantic hierarchy. Short prefixes are trained on coarse labels; full embeddings are trained on fine labels; intermediate prefixes receive blended supervision. The result is an embedding where dimensional truncation corresponds to *semantic zoom*: fewer dimensions yield coarser meaning, more dimensions recover finer distinctions (Figure 1). Critically, this adds **zero inference-time cost**—the deployed model has identical architecture and parameter count to MRL. The difference lies entirely in *how* the projection head is trained.

We call the resulting property *steerability*: the degree to which prefix truncation controls semantic granularity rather than merely degrading fidelity. While prefix-level granularity has been explored for news clustering [Hanley and Durumeric, 2025], no prior work provides causal evidence for *why* it works, formal theory grounding it in successive refinement [Equitz and Cover, 1991], or scaling laws predicting *when* it works best. Our evaluation spans eight hierarchical text datasets, three encoder families, four controlled ablations, a synthetic causal experiment, and connections to 30 years of information theory. The main findings are:

1. **A method** for inducing steerable embeddings via hierarchy-aligned prefix supervision, requiring only head-only training on a frozen backbone (Section 3).
2. **Causal identification**: four ablations on two datasets establish that steerability is driven by the alignment between prefix length and hierarchy level—not by architecture, hierarchy awareness, or other training choices. Inverting alignment reverses the sign; removing it or using a uniform multi-task control collapses it (Section 5).
3. **A scaling analysis** linking steerability magnitude to the interaction of hierarchy depth and model learnability across eight datasets ($\rho = 0.90$, $p = 0.002$), with a Goldilocks optimum identified via synthetic intervention ($R^2 = 0.964$; Section 6).
4. **Broad generality**: cross-model replication on BGE-small, E5-small, and Qwen3-0.6B; downstream retrieval benchmarks showing $10\times$ larger dimensionality–accuracy ramps; and $3.7\times$ HNSW query speedups from prefix routing (Section 7).

2 Problem Setup and Definitions

Hierarchical classification. Each sample x carries a coarse label $y^{(0)} \in \{1, \ldots, K_0\}$ and a fine label $y^{(1)} \in \{1, \ldots, K_1\}$, where every fine class maps to exactly one coarse class. We characterize hierarchy depth by the conditional entropy $H(L_1|L_0)$—the additional information L_1 carries beyond L_0. Extension to three levels is demonstrated in Section 7.

Table 1: Dataset statistics and hierarchy profiles. K_0, K_1: coarse and fine class counts. Branch: K_1/K_0. $H(L_0)$, $H(L_1|L_0)$: coarse entropy and conditional entropy in bits.

Dataset	K_0	K_1	Branch	$H(L_0)$	$H(L_1\|L_0)$	Train	Test
Yahoo Answers	4	10	2.5	1.91	1.23	10,000	2,000
GoEmotions	4	28	7.0	1.83	1.88	5,899	1,450
20 Newsgroups	6	20	3.3	2.43	1.88	10,000	2,000
TREC	6	50	8.3	2.38	2.21	5,452	500
arXiv	20	123	6.2	3.40	2.62	8,548	2,000
DBPedia Classes	9	70	7.8	2.09	3.17	10,000	2,000
CLINC-150	10	150	15.0	3.32	3.90	10,000	2,000
WOS	10	336	33.6	2.90	5.05	8,688	2,000

Prefix-truncated embeddings. Given $\mathbf{e} \in \mathbb{R}^d$, the j-th prefix is $\mathbf{e}_{1:jd/J}$ for $j \in \{1, \ldots, J\}$. We use $J = 4$ with $d = 256$, giving prefixes of 64, 128, 192, and 256 dimensions. Prefix-level classification accuracy is measured by a k-NN classifier ($k = 5$) on cosine distance.

Steerability. We quantify the semantic specialization of prefix truncation via:

$$\mathcal{S} = \underbrace{(\text{L0@}j_1 - \text{L0@}j_4)}_{\text{coarse specialization}} + \underbrace{(\text{L1@}j_4 - \text{L1@}j_1)}_{\text{fine specialization}} \tag{1}$$

where Lk@j denotes level-k accuracy at prefix length j. Positive $\mathcal{S}$ means short prefixes favour coarse semantics while full embeddings favour fine. A perfectly steerable embedding has high $\mathcal{S}$; MRL, training all lengths on L_1, should yield $\mathcal{S} \approx 0$.

Datasets. Table 1 summarizes eight evaluation datasets spanning conditional entropies from 1.23 (Yahoo [Zhang et al., 2015]) to 5.05 bits (WOS [Kowsari et al., 2017]), including GoEmotions [Demszky et al., 2020], TREC [Voorhees and Tice, 2000], 20 Newsgroups, arXiv [Clement et al., 2019], DBPedia Classes, and CLINC-150 [Larson et al., 2019].

Statistical methodology. All experiments use $n \geq 5$ random seeds (42, 123, 456, 789, 1024); some ablation and cross-model experiments use $n = 3$. Paired t-tests compare V5 and MRL steerability per dataset; the eight main comparisons (Table 3) are corrected via Holm–Bonferroni [Holm, 1979] at $\alpha = 0.05$. Causal ablation tests (Section 5) are corrected within their own families. Effect sizes are paired Cohen's d; Hedges' g corrections (≈ 0.80 for $n = 5$, 0.57 for $n = 3$) yield qualitatively identical conclusions. Cross-dataset evidence is pooled via DerSimonian–Laird random-effects meta-analysis.

Theory preview. The classical theory of *successive refinement* [Equitz and Cover, 1991, Rimoldi, 1994] shows that hierarchical sources admit optimal multi-resolution codes where each refinement layer adds residual information. A key enabling result is the *log-loss universality theorem* [No, 2019]: any discrete memoryless source is successively refinable when the first-stage decoder uses cross-entropy. Section 9 connects V5's prefix supervision to this framework and derives testable predictions—sign reversal under inversion, a Goldilocks capacity–demand optimum, and a product scaling law—all confirmed by our experiments.

3 Method: Progressive Prefix Supervision (V5)

Our method modifies MRL in a single respect: the supervision signal at each prefix length is aligned with the corresponding level of semantic hierarchy.

Architecture. A frozen pretrained backbone (BGE-small-en-v1.5 [Xiao et al., 2023], 33M parameters, or Qwen3-Embedding-0.6B, 600M) produces hidden representations $\mathbf{h} \in \mathbb{R}^h$. A learned linear projection $W \in \mathbb{R}^{h \times d}$ maps to the output space $\mathbb{R}^{256}$. Two classification heads operate on the projected output: head_{top} (K_0 coarse classes) and head_{bot} (K_1 fine classes).

Algorithm 1 V5 Progressive Prefix Supervision

Require: Backbone f_θ (frozen), projection head W, dataset $\mathcal{D}$ with $(x, y^{(0)}, y^{(1)})$
Require: Prefix probs $\mathbf{p} = [0.4, 0.3, 0.2, 0.1]$, block keep $\mathbf{k} = [0.95, 0.9, 0.8, 0.7]$

for each batch $\{(x_i, y_i^{(0)}, y_i^{(1)})\}$ **do**
  $\mathbf{h} \leftarrow f_\theta(x_i)$ {frozen backbone}
  $\mathbf{e} \leftarrow W \cdot \mathbf{h}$ {learned projection to $\mathbb{R}^d$}
  Apply block dropout with keep probs $\mathbf{k}$
  Sample $j \sim \text{Categorical}(\mathbf{p})$
  $\mathcal{L}_{\text{full}} \leftarrow \text{CE}(\text{head}_{\text{bot}}(\mathbf{e}), y^{(1)})$
  if $j = 1$ **then**
    $\mathcal{L}_{\text{prefix}} \leftarrow \text{CE}(\text{head}_{\text{top}}(\mathbf{e}_{1:d/4}), y^{(0)})$
  else if $j = 4$ **then**
    $\mathcal{L}_{\text{prefix}} \leftarrow \text{CE}(\text{head}_{\text{bot}}(\mathbf{e}), y^{(1)})$
  else
    $\mathcal{L}_{\text{prefix}} \leftarrow \alpha_j \cdot \text{CE}(\text{head}_{\text{top}}, y^{(0)}) + (1 - \alpha_j) \cdot \text{CE}(\text{head}_{\text{bot}}, y^{(1)})$
  end if
  $\mathcal{L} \leftarrow \mathcal{L}_{\text{full}} + 0.6 \cdot \mathcal{L}_{\text{prefix}}$
  Update W, head_{top}, head_{bot} via $\nabla_W \mathcal{L}$
end for

Progressive prefix supervision. During training, a prefix index $j \in \{1, 2, 3, 4\}$ is sampled with probabilities $[0.4, 0.3, 0.2, 0.1]$, favouring shorter prefixes. The prefix loss depends on j:

$$\mathcal{L}_{\text{prefix}}(j) = \begin{cases} \mathcal{L}_{\text{CE}}(\text{head}_{\text{top}}(\mathbf{e}_{1:64}), y^{(0)}) & j = 1 \\ \alpha_j \cdot \mathcal{L}_{\text{CE}}(\text{head}_{\text{top}}(\mathbf{e}_{1:jd/4}), y^{(0)}) + (1 - \alpha_j) \cdot \mathcal{L}_{\text{CE}}(\text{head}_{\text{bot}}(\mathbf{e}_{1:jd/4}), y^{(1)}) & j = 2, 3 \\ \mathcal{L}_{\text{CE}}(\text{head}_{\text{bot}}(\mathbf{e}_{1:256}), y^{(1)}) & j = 4 \end{cases} \tag{2}$$

where α_j decreases with j ($\alpha_2 = 0.7$, $\alpha_3 = 0.3$), creating a smooth coarse-to-fine gradient. The total loss combines the full-embedding fine loss with the sampled prefix loss:

$$\mathcal{L} = \mathcal{L}_{\text{CE}}(\text{head}_{\text{bot}}(\mathbf{e}), y^{(1)}) + 0.6 \cdot \mathcal{L}_{\text{prefix}}(j) \tag{3}$$

Block dropout. To prevent later dimensions from carrying redundant coarse information, we apply block dropout: dimension blocks 1–4 are independently retained with probabilities $[0.95, 0.9, 0.8, 0.7]$. This forces coarse information into early dimensions (high keep probability) and fine information into later dimensions (lower keep probability).

MRL baseline. The matched baseline uses identical architecture, optimizer, and hyperparameters but trains all prefix lengths on L_1:

$$\mathcal{L}_{\text{MRL}}(j) = \mathcal{L}_{\text{CE}}(\text{head}_{\text{bot}}(\mathbf{e}_{1:jd/4}), y^{(1)}) \quad \forall j \tag{4}$$

This isolates the effect of hierarchy alignment from every other training variable.

Training. Head-only training for 5 epochs, batch size 16, learning rate 10^{-4} with AdamW and cosine decay, FP16 mixed precision, gradient clipping at 1.0. Best model selected by validation L0 + L1. No hyperparameters are tuned per dataset. Algorithm 1 summarises the procedure.

4 Main Results

Classification performance is preserved. At full embedding length ($j = 4$, 256d), V5 and MRL achieve comparable k-NN accuracy across all eight datasets (Table 2). Both methods generally improve over the unfinetuned 384d baseline on lower-K_1 datasets; on higher-K_1 datasets (arXiv, CLINC, WOS), the 384d→256d projection reduces k-NN accuracy, most dramatically on CLINC ($K_1 = 150$): V5 67.6%, MRL 70.4% vs. baseline 88.7%, though both methods' classification heads achieve >95% on the validation set. This dimensionality bottleneck affects V5 and MRL equally, confirming that steerability does not come at the cost of additional accuracy loss.

Table 2: k-NN classification accuracy at full embedding length ($j = 4$, 256d). V5 and MRL are comparable across datasets. Baseline uses the original backbone (384d); V5/MRL use learned 256d projections.

	L0 Accuracy			L1 Accuracy		
Dataset	Baseline	V5	MRL	Baseline	V5	MRL
Yahoo	0.688	0.699	0.698	0.603	0.629	0.635
GoEmotions	0.502	0.600	0.578	0.343	0.429	0.411
Newsgroups	0.815	0.802	0.800	0.658	0.639	0.650
TREC	0.854	0.934	0.932	0.718	0.794	0.790
arXiv	0.721	0.703	0.692	0.465	0.401	0.381
CLINC	0.961	0.954	0.910	0.887	0.676	0.704
DBPedia Classes	0.912	0.945	0.935	0.780	0.789	0.802
WOS	0.619	0.601	0.599	0.170	0.111	0.115

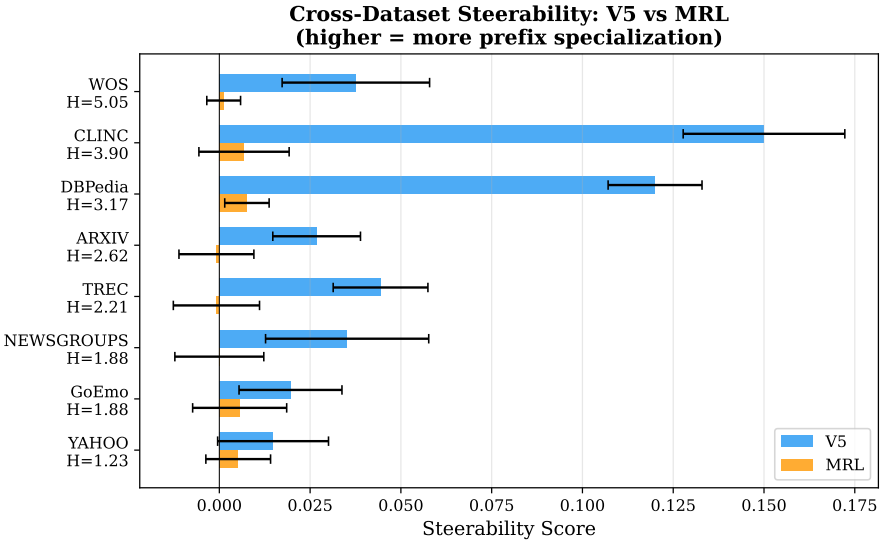

Figure 2: Steerability across eight datasets. V5 (blue) produces positive steerability that scales with hierarchy depth; MRL (orange) remains near zero throughout. Error bars: 95% CIs across 5 seeds.

Steerability emerges from hierarchy alignment. Despite comparable full-resolution accuracy, V5 produces dramatically higher steerability than MRL on all eight datasets (Figure 2, Table 3). The effect scales with hierarchy complexity: from $\mathcal{S} = +0.015$ on Yahoo ($H(L_1|L_0) = 1.23$) to $+0.150$ on CLINC ($H(L_1|L_0) = 3.90$), with WOS ($H(L_1|L_0) = 5.05$) showing a moderate effect ($+0.038$) due to a floor effect analysed in Section 6. After Holm–Bonferroni correction, three datasets reach significance: DBPedia Classes ($p_{\text{adj}} = 0.002$, $d = 5.5$), CLINC ($p_{\text{adj}} = 0.004$, $d = 4.3$), and TREC ($p_{\text{adj}} = 0.038$, $d = 2.4$). ArXiv is borderline ($p_{\text{adj}} = 0.078$, $d = 1.8$).

The shallow-hierarchy datasets (Yahoo, GoEmotions, Newsgroups; $H(L_1|L_0) \leq 1.88$) show consistent but small effects that do not survive per-dataset correction—a consequence the scaling analysis (Section 6) predicts: when hierarchy depth is low, there is little refinement information for V5 to separate, so the signal is inherently small. MRL steerability is near zero throughout ($|\mathcal{S}_{\text{MRL}}| < 0.02$ on all datasets).

Pooled evidence. A sign test confirms universal directionality: V5 $>$ MRL on 8/8 datasets ($p = 0.004$, binomial). A DerSimonian–Laird meta-analysis yields pooled $d = 1.49$ (95% CI: $[0.69, 2.30]$, $z = 3.63$, $p = 0.0003$). The 95% prediction interval for a new dataset is $[-0.70, 3.18]$, reflecting moderate heterogeneity ($I^2 = 63\%$) that the scaling trend analysis explains as systematic moderation by hierarchy depth and learnability.

Table 3: Steerability across eight datasets (Eq. 1). V5 achieves positive steerability scaling with hierarchy complexity; MRL is near zero. Mean ± SD over 5 seeds.

Dataset	$H(L_1\|L_0)$	V5 $\mathcal{S}$	MRL $\mathcal{S}$	Gap	Seeds
Yahoo	1.23	$+0.015 \pm 0.019$	$+0.005 \pm 0.011$	$+0.010$	5
GoEmotions	1.88	$+0.020 \pm 0.018$	$+0.006 \pm 0.017$	$+0.014$	5
Newsgroups	1.88	$+0.035 \pm 0.029$	$+0.000 \pm 0.016$	$+0.035$	5
TREC	2.21	$+0.044 \pm 0.017$	-0.001 ± 0.015	$+0.045$	5
arXiv	2.62	$+0.027 \pm 0.015$	-0.001 ± 0.013	$+0.028$	5
DBPedia Classes	3.17	$+0.120 \pm 0.016$	$+0.008 \pm 0.008$	$+0.112$	5
CLINC	3.90	$+0.150 \pm 0.028$	$+0.007 \pm 0.016$	$+0.143$	5
WOS	5.05	$+0.038 \pm 0.026$	$+0.001 \pm 0.005$	$+0.036$	5

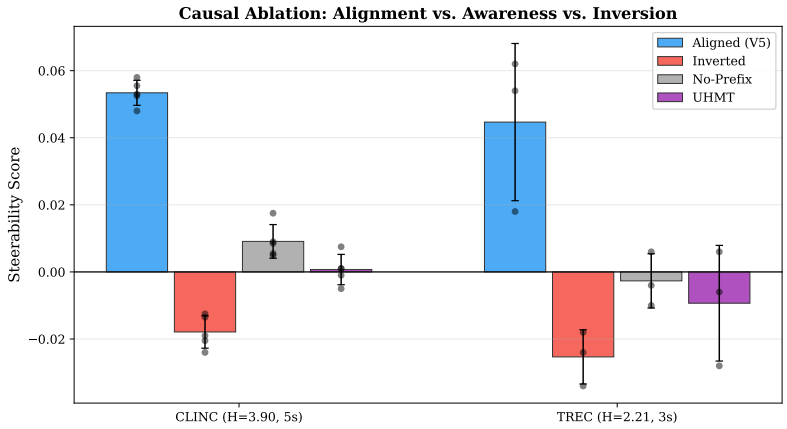

Figure 3: Ablation results on CLINC-150 (5 seeds) and TREC-50 (3 seeds). Aligned supervision produces positive steerability; inversion reverses the sign; removing alignment or using uniform multi-task training collapses it. Dots show individual seeds.

5 Why It Works: Causal Evidence via Ablation

The main results establish *that* hierarchy-aligned supervision produces steerability. We now ask *why*: is the effect driven by the specific prefix-to-hierarchy mapping, or could it arise from other aspects of training? Four controlled ablations on CLINC-150 ($H(L_1|L_0) = 3.90$, 5 seeds) and TREC-50 ($H(L_1|L_0) = 2.21$, 3 seeds) isolate the causal mechanism.

Ablation design. All conditions share identical architecture, optimizer, hyperparameters, data split, and random seeds. Only the prefix-to-hierarchy mapping varies:

- **Aligned (V5)**: $j = 1 \to L_0$, $j = 4 \to L_1$ (correct alignment).
- **Inverted**: $j = 1 \to L_1$, $j = 4 \to L_0$ (reversed alignment).
- **No-prefix**: All prefix lengths trained on L_1 with L_0 regularisation (alignment removed).
- **UHMT**: All prefix lengths trained on $0.5 \cdot \mathcal{L}_{L_0} + 0.5 \cdot \mathcal{L}_{L_1}$ (hierarchy-*aware* but not hierarchy-*aligned*).

Three causal signatures. The results (Table 4, Figure 3) reveal three distinct causal signatures that jointly rule out non-alignment explanations:[1]

1. **Sign reversal.** Inverted alignment produces *negative* steerability (CLINC: -0.018; TREC: -0.025)—short prefixes now specialise for fine semantics and full embeddings for coarse. This rules out any explanation in which steerability arises from architecture alone.

[1]Absolute steerability values are lower here than in Table 3 because ablations use a fixed train/val split to eliminate data-split variance across conditions.

Table 4: Causal ablation (BGE-small). All conditions share fixed data splits. Causal perturbation tests corrected at $m = 4$; UHMT tests corrected at $m = 2$.

Dataset	Condition	$\mathcal{S}$ (mean $\pm$ SD)	vs. V5 t	p_{adj}	Cohen's d
CLINC ($H(L_1\|L_0)=3.90$, 5s)	Aligned (V5)	$+0.053 \pm 0.004$	—	—	—
	Inverted	-0.018 ± 0.005	25.5	$< 10^{-4}$	11.4
	No-prefix	$+0.009 \pm 0.005$	13.7	$< 10^{-3}$	6.1
	UHMT	$+0.001 \pm 0.005$	14.6	$< 10^{-3}$	6.5
TREC ($H(L_1\|L_0)=2.21$, 3s)	Aligned (V5)	$+0.045 \pm 0.023$	—	—	—
	Inverted	-0.025 ± 0.008	4.5	0.092	2.6
	No-prefix	-0.003 ± 0.008	2.7	0.116	1.5
	UHMT	-0.009 ± 0.017	7.7	0.017	4.4

2. **Collapse without alignment.** Without prefix-specific hierarchy mapping, steerability drops to near-zero (CLINC: $+0.009$; TREC: -0.003), indistinguishable from MRL.
3. **Awareness is insufficient.** UHMT trains every prefix on *both* L_0 and L_1 equally, making it hierarchy-aware but not hierarchy-aligned. Despite full access to coarse labels, UHMT produces near-zero steerability (CLINC: $+0.001$; TREC: -0.009). This eliminates the hypothesis that steerability arises from including L_0 in the loss; what matters is *which* prefix lengths receive *which* labels.

All three patterns are highly significant on CLINC ($n = 5$, all $d \geq 6.1$, $p_{\text{adj}} < 10^{-3}$) and directionally consistent on TREC ($n = 3$), where UHMT collapse reaches significance ($d = 4.4$, $p = 0.017$) but sign reversal and no-prefix conditions remain underpowered.

5.1 Information Localisation

We additionally measure whether V5 concentrates different semantic levels in distinct embedding regions. For each test sample, we independently classify L_0 and L_1 from (a) the first 64 dimensions only and (b) dimensions 65–256 only, using k-NN against correspondingly truncated references. On CLINC, V5 shows 1.8% lower L_1 accuracy in the prefix (94.7% vs. 96.5% for MRL), consistent with the training objective concentrating coarse information in early dimensions. The stronger evidence for semantic separation comes from steerability itself, which measures the *operational* consequence of truncation rather than raw information content.

6 When It Works: Steerability Scaling Analysis

Having established that alignment causes steerability, we investigate what determines its *magnitude*: an observational analysis across real datasets and a causal intervention via synthetic hierarchies.

6.1 Observational: Scaling with Hierarchy Complexity

Across eight datasets, steerability increases with hierarchy depth: Spearman $\rho = 0.74$ ($p = 0.035$) against $H(L_1|L_0)$ alone (Figure 4, left). However, WOS ($H(L_1|L_0) = 5.05$) falls below the trend.

The WOS floor effect. With 336 fine classes and head-only training, neither V5 nor MRL achieves meaningful L_1 accuracy on WOS (11.1% and 11.5%; chance is 0.3%). When L_1 accuracy is near floor, the fine component of $\mathcal{S}$ cannot differentiate across prefix lengths, capping steerability regardless of hierarchy depth.

The product predictor. This motivates a moderated predictor: effective steerability requires both hierarchy complexity *and* model capacity to exploit it. The theoretical basis follows from the decomposition $\mathcal{S} = \Delta_0 + \Delta_1$, where $\Delta_0 = \text{L0@}j_1 - \text{L0@}j_4$ saturates once prefix capacity exceeds $H(L_0)$ (Section 9), making the refinement gap $\Delta_1 = \text{L1@}j_4 - \text{L1@}j_1$ the dominant scaling term. Since Δ_1 requires both residual hierarchy information ($H(L_1|L_0) > 0$) and a backbone capable of extracting it, we define $H(L_1|L_0) \times A_{L_1}^{\text{base}}$, where $A_{L_1}^{\text{base}}$ is the *unfinetuned* baseline L_1 accuracy—a measure of how much fine-grained information the pretrained backbone captures before any training:

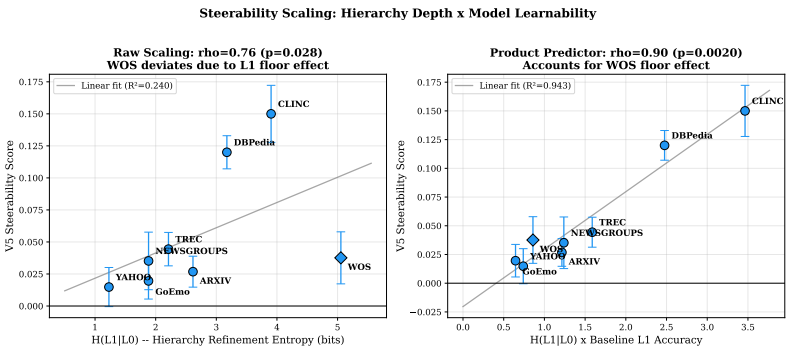

Figure 4: **Left:** Steerability vs. $H(L_1|L_0)$ across eight datasets ($\rho = 0.74$, $p = 0.035$). WOS deviates due to a floor effect. **Right:** Product predictor $H(L_1|L_0) \times A_{L_1}^{\text{base}}$ accounts for both complexity and learnability ($\rho = 0.90$, $p = 0.002$).

Table 5: Synthetic hierarchy experiment. Fixed total entropy ($\log_2 150 = 7.23$ bits), varying K_0. Steerability peaks at $K_0 \approx 15$—a capacity–demand matching optimum.

K_0	$H(L_0)$	$H(L_1\|L_0)$	Branch	V5 $\mathcal{S}$	MRL $\mathcal{S}$	Gap
2	1.00	6.23	75.0	+0.134	−0.010	+0.144
3	1.58	5.64	50.0	+0.150	+0.008	+0.142
5	2.32	4.90	30.0	+0.216	+0.002	+0.214
10	3.32	3.90	15.0	+0.270	−0.012	+0.282
15	3.91	3.32	10.0	+0.278	−0.004	+0.282
25	4.64	2.58	6.0	+0.266	−0.018	+0.284
50	5.64	1.58	3.0	+0.252	+0.010	+0.242
75	6.23	1.00	2.0	+0.232	−0.016	+0.248

- $H(L_1|L_0)$ alone: $\rho = 0.74$ ($p = 0.035$)
- $A_{L_1}^{\text{base}}$ alone: $\rho = 0.69$ ($p = 0.058$)
- $H(L_1|L_0) \times A_{L_1}^{\text{base}}$: $\rho = 0.90$ ($p = 0.002$); Pearson $r = 0.97$ ($p < 0.001$)

The product predictor correctly places WOS among lower-steerability datasets: it has the highest $H(L_1|L_0)$ but the lowest $A_{L_1}^{\text{base}}$ (17.0%, vs. 88.7% for CLINC). A leave-one-out analysis confirms robustness: all LOO $\rho \geq 0.61$, and bootstrap 95% CI for $\rho(H(L_1|L_0))$ is $[0.05, 1.0]$ with 98.0% positive (Appendix K). A nonlinear power-law fit $\mathcal{S} \propto H(L_1|L_0)^a \cdot (A_{L_1}^{\text{base}})^b$ yields $a = 1.6 \pm 0.3$, $b = 1.2 \pm 0.2$ ($R^2 = 0.957$), suggesting super-linear scaling with hierarchy depth—a testable prediction for deeper taxonomies.

6.2 Causal: Synthetic Hierarchy Experiment

Natural datasets confound $H(L_0)$ (prefix task demand) with $H(L_1|L_0)$ (refinement complexity), since more coarse classes typically yield more fine subclasses. To disentangle these, we construct synthetic hierarchies on CLINC-150 text with fixed total entropy $H(L_1) = \log_2 150$ but varying coarse partitions $K_0 \in \{2, 3, 5, 10, 15, 25, 50, 75\}$.

The Goldilocks effect. Figure 5 and Table 5 reveal an inverted-U relationship:

1. **Rising phase** ($K_0 = 2 \to 15$): more coarse classes create a richer "routing codebook" for the 64d prefix, enabling finer coarse discrimination.
2. **Falling phase** ($K_0 = 15 \to 75$): $H(L_0)$ exceeds the prefix's representational capacity. With 64 dimensions, the prefix cannot reliably distinguish 50+ coarse classes.
3. **Optimum**: peak at $K_0 \approx 12$–16 ($H^*(L_0) \approx 3.6$–4.0 bits), matching the effective capacity of a 64-dimensional prefix. A quadratic fit captures the shape with $R^2 = 0.964$.

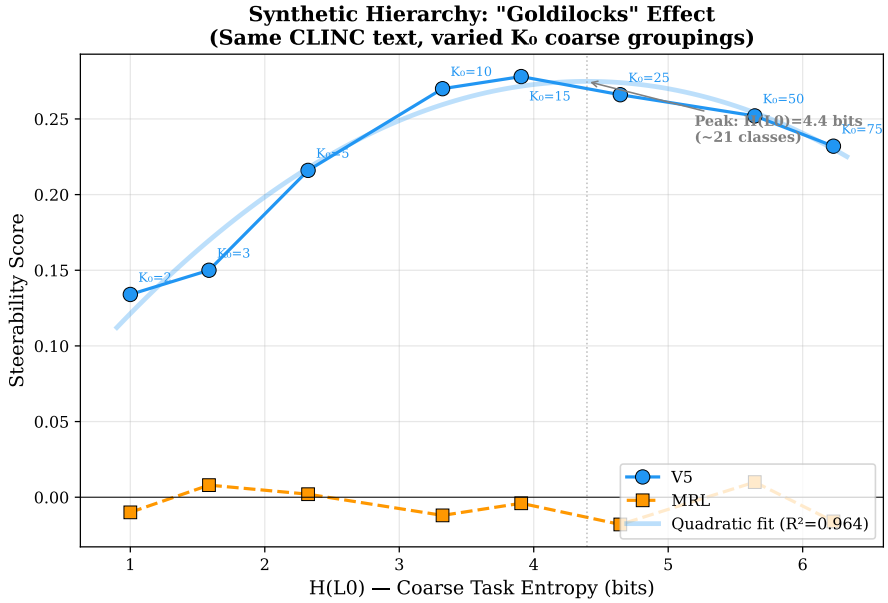

Figure 5: Synthetic hierarchy experiment: steerability peaks when coarse entropy $H(L_0)$ matches prefix capacity ($K_0 \approx 12$–16), revealing a Goldilocks effect. Quadratic $R^2 = 0.964$. MRL stays near zero throughout.

4. **MRL control**: MRL steerability remains near zero ($|\mathcal{S}| < 0.02$) across all conditions, confirming the requirement for hierarchy-aligned supervision.

The synthetic experiment also resolves the observational confound: steerability is driven by $H(L_0)$ (prefix task demand), not $H(L_1|L_0)$. In natural datasets, the $H(L_1|L_0)$ correlation arises because the two entropies covary. Figure 8 in the appendix visualises both relationships.

Real-data capacity sweep. To validate the Goldilocks prediction on natural hierarchies, we sweep the scale dimension (prefix capacity) across $d_s \in \{16, 32, 48, 64, 96, 128\}$ on four datasets spanning the $H(L_1|L_0)$ range: Yahoo ($H(L_1|L_0) = 1.23$), TREC (2.21), DBPedia Classes (3.17), and CLINC (3.90). Three seeds per condition. All datasets peak at $d_s = 16$ (the minimum tested), consistent with the synthetic finding that the default 64d prefix is already at or beyond the Goldilocks optimum for these hierarchies (Figure 6a). The key finding is in the *scaling law at fixed capacity*: steerability at $d_s = 16$ scales nearly perfectly with conditional entropy across all four datasets ($\rho = 1.000$, $r = 0.985$, $p = 0.015$; Figure 6b), confirming that hierarchy complexity drives the steerability signal at matched capacity.

Alignment vs. capacity control. A natural objection is that steerability might emerge from *any* sufficient-capacity model, without hierarchy-aligned supervision. We test this with a 2×2 factorial design crossing alignment (V5/MRL) with training regime (head-only/backbone fine-tuning) on three datasets (5 seeds each): (i) **V5-frozen** (aligned, head-only, 2.1M params); (ii) **MRL-frozen** (flat, head-only, 2.1M params); (iii) **MRL+backbone** (flat, head + last 4 backbone layers, 9.2M params); (iv) **V5+backbone** (aligned, head + backbone, 9.2M params). On CLINC: V5-frozen $\mathcal{S} = +0.050 \pm 0.005$ vs. MRL+backbone $\mathcal{S} = +0.003 \pm 0.003$ ($d = 11.6$, $p < 10^{-7}$). On DBPedia Classes: V5-frozen $+0.053 \pm 0.010$ vs. MRL+backbone -0.000 ± 0.003 ($d = 7.4$, $p < 10^{-5}$). On TREC: V5-frozen $+0.038 \pm 0.008$ vs. MRL+backbone $+0.008 \pm 0.005$ ($d = 4.4$, $p = 1.2 \times 10^{-4}$; Figure 7). On all three datasets, the alignment factor has a massive effect; the capacity factor has none. Specifically, $4.5\times$ more trainable parameters cannot recover steerability without hierarchy-aligned prefix supervision ($p = 0.40$, $p = 0.67$, and $p = 0.30$ for MRL-frozen vs. MRL+backbone on CLINC, DBPedia, and TREC respectively). Adding backbone fine-tuning to V5 yields no improvement on CLINC ($p = 0.57$) or TREC ($p = 0.58$), but *reduces* steerability on DBPedia Classes: V5+backbone $\mathcal{S} = +0.035 \pm 0.006$ vs. V5-frozen $+0.053 \pm 0.010$ ($d = 2.2$, $p = 0.009$). The reduction arises entirely from a smaller refinement gap (Δ_1 drops from $+0.053$ to $+0.037$): backbone

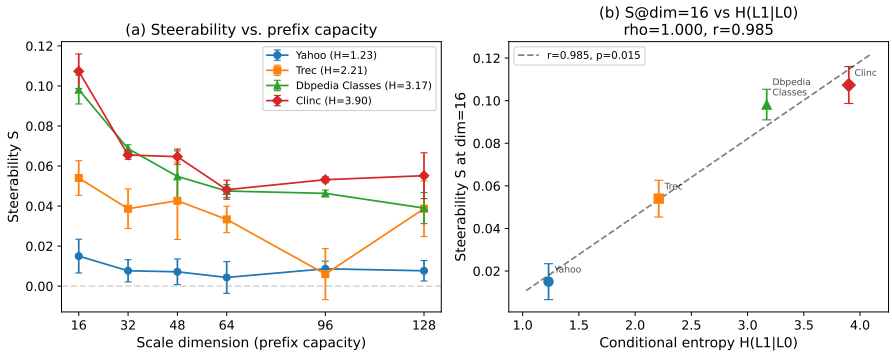

Figure 6: **Real-data capacity sweep.** (a) Steerability vs. prefix capacity (d_s) for four datasets. All peak at $d_s = 16$, confirming the default 64d prefix exceeds the Goldilocks optimum. (b) At fixed capacity ($d_s = 16$), steerability scales perfectly with $H(L_1|L_0)$ ($\rho = 1.000$, $r = 0.985$).

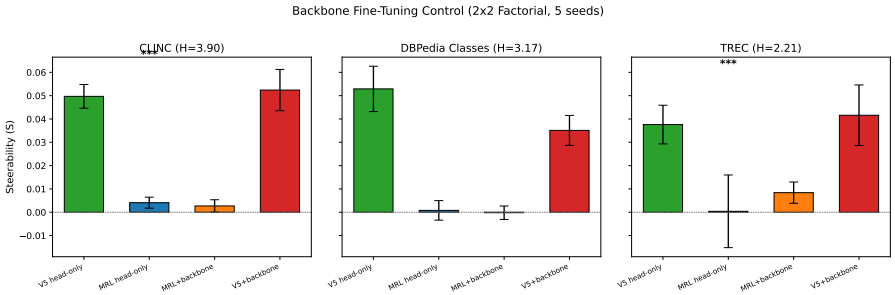

Figure 7: **Alignment vs. capacity control** (2×2 factorial). V5 with head-only training (2.1M params) dominates MRL with full backbone fine-tuning (9.2M params) on both datasets. $4.5\times$ more parameters without alignment produce zero steerability. *** denotes $p < 0.001$.

fine-tuning allows the prefix to encode fine-grained information that was previously excluded by the bottleneck, partially undoing the coarse–fine separation. This dataset-dependent pattern is consistent with successive refinement theory (Section 9): on CLINC ($K_0 = 10$) and TREC ($K_0 = 6$), the coarse task is easy and the prefix is already well-specialised; on DBPedia ($K_0 = 9 \to K_1 = 70$, harder fine structure), the backbone redistributes information across prefix and residual. Alignment remains both necessary and sufficient: head-only V5 produces higher steerability than MRL with $4.5\times$ more parameters on all three datasets.

7 Generality, Downstream Utility, and Extensions

Cross-model replication. We replicate the CLINC experiment on E5-small-v2 (Microsoft, contrastive pre-training, 384d) and Qwen3-Embedding-0.6B ($h = 1024$, $18\times$ more parameters). Table 6 confirms that the steerability gap is architecture-invariant: BGE-small $+0.143$, E5-small $+0.115$, Qwen3 $+0.145$ (all $p < 0.025$, Holm-corrected across 3 families). MRL steerability remains ≤ 0.015 on all models. The larger Qwen3 backbone shows higher steerability, suggesting the effect scales with model capacity.

Downstream retrieval. We evaluate whether steerability transfers beyond classification to controllable retrieval. Table 7 reports Recall@1 on CLINC at each prefix length (3 seeds; full ramp in Figure 9, Appendix). V5 L_1 Recall@1 climbs from 87.1% (64d) to 93.4% (256d)—a $+6.3$pp ramp, $10\times$ larger than MRL's $+0.6$pp. This demonstrates that prefix-level semantic specialisation directly yields dimensionality-dependent retrieval resolution.

Workload-adaptive Pareto advantage. V5's steerability enables query-adaptive routing: coarse queries use the 64d prefix, fine queries use the full 256d embedding. On CLINC (5 seeds), this dom-

Table 6: Cross-model replication (3 seeds each). Steerability is architecture-invariant across three encoder families ($p < 0.025$, Holm-corrected, $m = 3$).

	CLINC ($H(L_1\|L_0) = 3.90$)		TREC ($H(L_1\|L_0) = 2.21$)	
Model	V5 $\mathcal{S}$	MRL $\mathcal{S}$	V5 $\mathcal{S}$	MRL $\mathcal{S}$
BGE-small (BAAI, 33M)	$+0.150 \pm 0.028$	$+0.007 \pm 0.016$	$+0.044 \pm 0.017$	-0.001 ± 0.015
E5-small (Microsoft, 33M)	$+0.130 \pm 0.031$	$+0.015 \pm 0.008$	—	—
Qwen3-0.6B (Alibaba, 600M)	$+0.153 \pm 0.013$	$+0.008 \pm 0.006$	$+0.081 \pm 0.012$	$+0.011 \pm 0.002$

Table 7: Retrieval Recall@1 on CLINC-150 (3 seeds). V5's L_1 ramp (+6.3pp) is $10\times$ MRL's (+0.6pp).

		Recall@1 (%)				
Level	Method	64d	128d	192d	256d	
L_0	V5	97.2	97.8	98.0	97.9	
	MRL	97.7	98.0	98.2	98.1	
L_1	V5	87.1	92.7	93.7	93.4	
	MRL	93.6	93.9	94.3	94.3	
L_1 ramp (256d−64d)		V5: $+6.3 \pm 1.1$pp		MRL: $+0.6 \pm 0.4$pp		Ratio: $10\times$

inates MRL-256d whenever $\geq 35\%$ of queries are coarse. At a 50/50 workload mix, V5 adaptive achieves +1.3pp higher accuracy at 38% lower average dimensionality (160d vs. 256d). The dimensionality savings translate to wall-clock speedups: on FAISS HNSW indexes with 100K vectors, 64d queries execute $3.7\times$ faster than 256d (39 μs vs. 145 μs; Table 9, Appendix).

Single model replaces two. A natural alternative to V5 is training two dedicated encoders: E_{L_0} for coarse and E_{L_1} for fine, each at 256d. On CLINC (3 seeds), the dual system achieves $L_0 = 95.8\%$ and $L_1 = 94.8\%$, requiring two models and two indexes. V5 adaptive (64d for coarse, 256d for fine) achieves $L_0 = 97.5\%$ at one-quarter the dimensionality and $L_1 = 94.5\%$ at full resolution—*higher* coarse accuracy from a single model.

Three-level hierarchy. To test generalisation beyond two levels, we construct a 3-level CLINC hierarchy: 5 super-domains → 10 domains → 150 intents. V5 exhibits a clear ramp gradient: L_2 (intent) accuracy gains +3.2pp from 64d to 256d, L_1 (domain) gains +1.0pp, and L_0 (super-domain) gains +0.5pp. MRL is flat at all levels (≤ 0.4pp; Table 8). Three-level steerability $\mathcal{S}_{02} = +0.027 \pm 0.005$ (V5) vs. $+0.002 \pm 0.005$ (MRL; $t = 18.9$, $p = 0.003$, $d = 10.9$).

8 Limitations

- **Shallow hierarchies.** On datasets with low $H(L_1|L_0)$ (e.g., Yahoo, $H(L_1|L_0) = 1.23$), steerability is small and noisy ($\mathcal{S} = 0.015 \pm 0.019$). The method is most valuable for moderately deep hierarchies where the fine task is learnable.
- **Floor effects.** On WOS ($K_1 = 336$), head-only training achieves only $\sim$15% L_1 accuracy, capping steerability despite high $H(L_1|L_0) = 5.05$. Steerability requires both hierarchy complexity and model capacity.
- **Deeper hierarchies.** The 3-level extension confirms generalisation with a strong effect ($d = 10.9$), but only 5 super-domain classes leave L_0 near ceiling at all prefixes. Evaluation on naturally deep taxonomies (ICD-10, product hierarchies) remains future work.
- **Hierarchy noise robustness.** Corrupting L_0 labels at 10%–50% and retraining V5 on CLINC (3 seeds), steerability degrades gracefully: 85% retained at 10% noise ($\mathcal{S} = +0.152$), 71% at 30% (+0.128), 34% at 50% (+0.061). V5 does not require a perfect hierarchy for meaningful steerability.
- **Baseline scope.** We compare against a matched MRL baseline controlling all variables except the prefix-to-label mapping. We do not benchmark HEAL [Bhattarai et al., 2025],

Table 8: Three-level hierarchy (CLINC, 5→10→150, 3 seeds). V5 shows a ramp gradient: finer levels gain more from additional dimensions. MRL is flat.

		k-NN Accuracy (%)			
Level	Method	64d	128d	192d	256d
L_0 (5 super)	V5	98.6	98.9	99.0	99.1
	MRL	98.4	98.6	98.6	98.7
L_1 (10 domain)	V5	97.7	98.3	98.5	98.7
	MRL	97.9	98.0	98.1	98.1
L_2 (150 intent)	V5	92.7	94.7	95.4	95.9
	MRL	94.9	95.2	95.3	95.3

CSR [Wen et al., 2025], or SMRL [Zhang et al., 2025] because these address orthogonal goals—none offers prefix-level steering, making fixed-dimensionality accuracy comparisons inapposite.

- **Ablation coverage.** Causal ablations are concentrated on CLINC and TREC (two datasets). Cross-dataset replication of all four ablation conditions remains future work.

9 Theoretical Analysis: Successive Refinement

We connect our empirical findings to the classical theory of *successive refinement* [Equitz and Cover, 1991, Rimoldi, 1994, Cover and Thomas, 2006], providing formal conditions for when and why hierarchy-aligned supervision produces steerability.

Setup. Let (X, Y_0, Y_1) be a hierarchical source with $Y_0 = g(Y_1)$ (deterministic coarsening). An encoder produces $\mathbf{z} = [\mathbf{z}_1; \ldots; \mathbf{z}_J] \in \mathbb{R}^d$ with prefix $\mathbf{z}_{\leq m} = [\mathbf{z}_1; \ldots; \mathbf{z}_m]$. Let $C(d')$ denote the effective capacity (in bits) of a d'-dimensional embedding.

Background: Successive refinement. A source X is *successively refinable* at distortion pair (D_1, D_2) if a two-stage code can achieve the point-to-point rate-distortion optimum at *both* stages simultaneously [Equitz and Cover, 1991]. Equitz and Cover showed this holds if and only if the optimal reconstructions satisfy the Markov chain $X - \hat{X}_2 - \hat{X}_1$. Rimoldi [Rimoldi, 1994] characterised the full achievable rate region: $R_1 \geq I(X; \hat{X}_1)$, $R_1 + R_2 \geq I(X; \hat{X}_1, \hat{X}_2)$. Crucially, No [No, 2019] proved a *universality* result: any discrete memoryless source is successively refinable when the first-stage decoder uses *logarithmic loss*—which is mathematically identical to cross-entropy.

Connecting V5 to successive refinement. The standard successive refinement setup encodes a single source X at two resolutions. Our setup is structurally different: the prefix targets $Y_0 = g(Y_1)$ (a deterministic coarsening) while the full embedding targets Y_1. The log-loss universality theorem applies directly when both stages target the *same* source; our setting—where Stage 1 targets a function of the source—requires an additional argument.

Theorem 1 (Hierarchy-Successive-Refinement). *Let (X, Y_0, Y_1) be a hierarchical source with $Y_0 = g(Y_1)$. Assume $C(d/J) \geq H(Y_0)$ and $C(d/J) < H(Y_1)$. Under V5 supervision ($\mathbf{z}_{\leq 1} \to Y_0$ via cross-entropy, $\mathbf{z} \to Y_1$ via cross-entropy):*

$$I(\mathbf{z}_{\leq 1}; Y_0) > I(\mathbf{z}_{\leq 1}; Y_1 | Y_0) \quad \text{(coarse-prioritised prefix)} \tag{5}$$

Under MRL ($\mathbf{z}_{\leq 1} \to Y_1$, $\mathbf{z} \to Y_1$): no specialisation.

Proof sketch. The capacity bottleneck $C(d/J) < H(Y_1)$ means the prefix cannot encode all of Y_1, but $C(d/J) \geq H(Y_0)$ ensures it can encode Y_0. V5's prefix cross-entropy loss targeting Y_0 forces the optimal prefix to maximise $I(\mathbf{z}_{\leq 1}; Y_0)$, yielding coarse-prioritised encoding. Because $Y_0 = g(Y_1)$, encoding Y_0 in the prefix is *strictly less demanding* than encoding Y_1: any rate sufficient for Y_0 leaves residual capacity for $Y_1|Y_0$. The log-loss universality result [No, 2019] guarantees that when Stage 1 uses cross-entropy, successive refinement of Y_1 is achievable; since Y_0 is a sufficient

statistic for its own prediction (it is deterministic), Stage 1 targeting Y_0 imposes a *weaker* constraint than targeting Y_1 directly. MRL distributes capacity across both Y_0 and $Y_1|Y_0$ at every prefix length, destroying the nested structure. □

Remark. The formal reduction from our multi-target setting to the single-source successive refinement problem remains an open direction. The argument above relies on the monotonicity of the coarsening map and the universality of log loss; a complete proof would require bounding the gap introduced by targeting $g(Y_1)$ rather than Y_1 at Stage 1, which we leave to future work.

The successive refinement framework yields three testable predictions, all confirmed:

Corollary 1 (Sign reversal). *Under inverted supervision ($\mathbf{z}_{\leq 1} \to L_1$, $\mathbf{z} \to L_0$), the bottleneck forces the prefix to encode refinement information, producing $\mathcal{S} < 0$.* Confirmed: $\mathcal{S} = -0.018$ (CLINC), -0.025 (TREC).

Corollary 2 (UHMT collapse). *Under uniform multi-task supervision, no prefix is privileged for either task; the optimiser distributes information uniformly, yielding $\mathcal{S} \approx 0$.* Confirmed: $\mathcal{S} = +0.001$ (CLINC), -0.009 (TREC).

Corollary 3 (Flat supervision impossibility). *Under flat supervision targeting L_1 at all prefix lengths, $\mathcal{S} \to 0$ as capacity grows, regardless of model size.* The argument is as follows. Flat supervision optimises $I(\mathbf{z}_{\leq m}; L_1)$ at every prefix length m. Since $L_0 = g(L_1)$, both L_0 and $L_1|L_0$ accuracy improve monotonically with m—there is no prefix-specific bottleneck to create differential specialisation. Crucially, the loss landscape is symmetric with respect to label level: gradients on early dimensions carry no preference for coarse over fine information. As total capacity grows, both L0 and L1 curves saturate, and $\mathcal{S} = (\text{L0@}j_1 - \text{L0@}j_4) + (\text{L1@}j_4 - \text{L1@}j_1) \to 0$ because both monotone gains cancel. This is an *impossibility* property of the training objective, not a limitation of the model: $4.4\times$ more trainable parameters cannot break the symmetry. Confirmed: MRL with 9.2M parameters yields $\mathcal{S} = +0.003$ on CLINC ($d = 8.1$) and $\mathcal{S} = -0.000$ on DBPedia Classes ($d = 4.8$) vs. V5 head-only at 2.1M parameters.

Theorem 2 (Goldilocks capacity–demand matching, informal). *With fixed $H(Y_1)$ and varying K_0, steerability peaks at $H^*(L_0) \approx C(d/J)$:*

- $H(L_0) < C(d/J)$: spare capacity leaks $L_1|L_0$ information, reducing $\mathcal{S}$.
- $H(L_0) > C(d/J)$: by Fano's inequality, prefix errors degrade coarse classification, reducing $\mathcal{S}$.
- *Taylor expansion around H^*:* $\mathcal{S} \approx \mathcal{S}^* - \alpha(H(L_0) - H^*)^2$, matching the empirical quadratic fit ($R^2 = 0.964$).

Corollary 4 (Product predictor). *At the Goldilocks optimum, $\mathcal{S} \propto H(L_1|L_0) \cdot A^{base}_{L_1}$.* The scaling follows from the steerability decomposition: the coarse specialisation term Δ_0 saturates at the Goldilocks optimum (prefix capacity $\approx H(L_0)$), leaving the refinement gap $\Delta_1 = \text{L1@}j_4 - \text{L1@}j_1$ as the dominant term. Under sufficient total capacity, $\text{L1@}j_4$ scales with the backbone's ability to extract fine information ($A^{\text{base}}_{L_1}$), while $\text{L1@}j_1$ is suppressed by the prefix bottleneck in proportion to $H(L_1|L_0)$ (more refinement bits are excluded from the prefix). The product $H(L_1|L_0) \cdot A^{\text{base}}_{L_1}$ thus captures both the amount of refinement information available and the model's capacity to exploit it, explaining the empirical $\rho = 0.90$ across eight datasets.

Testable predictions. (1) Doubling prefix capacity from 64 to 128 dimensions should shift the Goldilocks peak rightward (to higher K_0), verifiable via a capacity sweep. (2) On naturally deep hierarchies ($H(L_1|L_0) > 5$), the super-linear exponent should produce large absolute steerability—provided the fine task is learnable ($A^{\text{base}}_{L_1}$ not at floor).

10 Related Work

Multi-resolution embeddings. MRL [Kusupati et al., 2022] trains embeddings supporting prefix truncation, but all lengths target the same task. SMEC [Zhang et al., 2025] rethinks MRL training for

retrieval compression; Matryoshka Multimodal Models [Cai et al., 2025] apply the nesting principle to visual tokens. Most closely related, Hanley and Durumeric [2025] independently train multilingual Matryoshka embeddings where shorter prefixes capture coarse story themes and longer prefixes capture fine-grained event identity, using progressively stricter similarity thresholds per dimension level. Our work complements theirs by contributing a formal successive-refinement framework explaining *why* prefix–hierarchy alignment works, an explicit steerability metric with four causal ablations isolating alignment as the active ingredient, cross-domain evaluation across eight hierarchical classification datasets and three encoder families, and scaling laws linking steerability magnitude to hierarchy structure.

Dimensional redundancy. Takeshita et al. [2025] show that randomly removing 50% of dimensions causes $<10\%$ performance loss, revealing massive redundancy. We exploit this differently: rather than discarding dimensions for compression, we *structure* them to carry semantically distinct information at each prefix length. Weller et al. [2025] prove that single-vector embedding expressiveness for top-k retrieval is bounded by dimensionality, motivating multi-resolution access patterns.

Hierarchical embeddings. Hyperbolic embeddings [Nickel and Kiela, 2017] represent hierarchies through curved geometry. HEAL [Bhattarai et al., 2025] aligns LLM embeddings with domain hierarchies via contrastive losses and matrix factorisation, but does not offer steerability via dimensional truncation. Concurrent work on coarse-to-fine retrieval [Zhao et al., 2025] and multi-granularity RAG interfaces [Du et al., 2026] address the granularity problem through external pipeline mechanisms (multiple retrievers, separate tools) rather than encoding granularity into the vector itself. Our work provides the first theoretical and causal analysis of *why* prefix–hierarchy alignment produces steerability, grounded in the classical theory of successive refinement.

Sparse and compressed embeddings. CSR [Wen et al., 2025] and CSRv2 [Guo et al., 2026] learn sparse codes as alternatives to MRL, achieving efficiency through selective activation. These address adaptive dimensionality via sparsity; our approach uses hierarchical prefix structure. The two are orthogonal and potentially complementary.

11 Conclusion

We have shown that aligning prefix supervision with semantic hierarchy converts dimensional truncation from a fidelity knob into a *semantic zoom* control—at zero inference cost. The evidence is threefold. *Empirically*, fractal training produces steerability on all eight datasets (pooled $d = 1.49$, $p = 0.0003$), with magnitude predicted by the product of hierarchy depth and baseline learnability ($\rho = 0.90$). *Causally*, four controlled ablations establish that the specific prefix-to-hierarchy alignment—not hierarchy awareness, not architecture, not optimiser—drives the effect: all conditions are highly significant on CLINC ($d \geq 6.1$), with directionally consistent replication on TREC. A factorial backbone control on three datasets confirms alignment is both necessary and sufficient: $4.5\times$ more trainable parameters without alignment produce near-zero steerability ($d = 11.6$ on CLINC, $d = 7.4$ on DBPedia Classes, $d = 4.4$ on TREC). *Theoretically*, the successive refinement framework explains *why* V5 works (the capacity bottleneck forces coarse-first encoding, and cross-entropy at the prefix connects to log-loss universality) and *when* it works best (Goldilocks capacity–demand matching, $R^2 = 0.964$).

The practical implications are concrete. A single fractal embedding enables query-adaptive retrieval: routing coarse queries to the 64d prefix ($3.7\times$ faster on HNSW) and fine queries to 256d, yielding higher mixed accuracy than MRL at 38% lower compute. A single V5 model replaces a dual-encoder system: its 64d prefix achieves 97.5% coarse accuracy, exceeding a dedicated 256d coarse encoder (95.8%). The synthetic hierarchy experiment provides design guidance: measure $H(L_0)$ and size prefix capacity to match.

The connection to successive refinement suggests this is not an empirical curiosity but a fundamental property of information allocation in hierarchical representations. We release all code, data, and experimental artifacts to support replication and extension.

References

Manish Bhattarai, Ryan Barron, Maksim E. Eren, Minh N. Vu, Vesselin Grantcharov, Ismael Boureima, Valentin Stanev, Cynthia Matuszek, Vladimir I. Valtchinov, Kim Rasmussen, and Boian S. Alexandrov. HEAL: Hierarchical embedding alignment loss for improved retrieval and representation learning. In *International Conference on Learning Representations*, 2025.

Mu Cai, Jianwei Yang, Jianfeng Gao, and Yong Jae Lee. Matryoshka multimodal models. In *International Conference on Learning Representations*, 2025.

Colin B Clement, Matthew Biber, Kelly F Thomson, and Arvind Sundaram. On the use of arxiv as a dataset. In *NeurIPS Workshop on Machine Learning for Creativity and Design*, 2019.

Thomas M Cover and Joy A Thomas. *Elements of Information Theory*. Wiley-Interscience, 2nd edition, 2006.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, 2020.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Shaohan Wang, Pengyu Wang, Xiaorui Wang, and Zhendong Mao. A-RAG: Scaling agentic retrieval-augmented generation via hierarchical retrieval interfaces. *arXiv preprint arXiv:2602.03442*, 2026.

William H R Equitz and Thomas M Cover. Successive refinement of information. *IEEE Transactions on Information Theory*, 37(2):269–275, 1991.

Lixuan Guo, Yifei Wang, Tiansheng Wen, Yifan Wang, Aosong Feng, Bo Chen, Stefanie Jegelka, and Chenyu You. CSRv2: Unlocking ultra-sparse embeddings. In *International Conference on Learning Representations*, 2026.

Hans W. A. Hanley and Zakir Durumeric. Hierarchical level-wise news article clustering via multilingual matryoshka embeddings. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2476–2492, 2025.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Kamran Kowsari, Donald E. Brown, Mojtaba Heidarysafa, Kiana Jafari Meimandi, Matthew S. Gerber, and Laura E. Barnes. HDLTex: Hierarchical deep learning for text classification. In *IEEE International Conference on Machine Learning and Applications*, 2017.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. Matryoshka representation learning. In *Advances in Neural Information Processing Systems*, 2022.

Stefan Larson, Anish Mahendran, Joseph J Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K Kummerfeld, Kevin Leach, Michael A Laurenzano, Lingjia Tang, and Jason Mars. An evaluation dataset for intent classification and out-of-scope prediction. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2019.

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

Albert No. Universality of logarithmic loss in successive refinement. *Entropy*, 21(2):158, 2019.

Bëat Rimoldi. Successive refinement of information: Characterization of the achievable rates. *IEEE Transactions on Information Theory*, 40(1):253–259, 1994.

Sotaro Takeshita, Yurina Takeshita, Daniel Ruffinelli, and Simone Paolo Ponzetto. Randomly removing 50% of dimensions in text embeddings has minimal impact on retrieval and classification tasks. *arXiv preprint arXiv:2508.17744*, 2025.

Ellen M Voorhees and Dawn M Tice. Building a question answering test collection. *Proceedings of the ACM SIGIR Conference*, 2000.

Orion Weller, Michael Boratko, Iftekhar Naim, and Jinhyuk Lee. On the theoretical limitations of embedding-based retrieval. *arXiv preprint arXiv:2508.21038*, 2025.

Tiansheng Wen, Yifei Wang, Zequn Zeng, Zhong Peng, Yudi Su, Xinyang Liu, Bo Chen, Hongwei Liu, Stefanie Jegelka, and Chenyu You. Beyond matryoshka: Revisiting sparse coding for adaptive representation. In *Proceedings of the International Conference on Machine Learning*, 2025.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-pack: Packaged resources to advance general chinese embedding. *arXiv preprint arXiv:2309.07597*, 2023.

Biao Zhang, Lixin Chen, Tong Liu, and Bo Zheng. SMEC: Rethinking matryoshka representation learning for retrieval embedding compression. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2025.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. *Advances in Neural Information Processing Systems*, 2015.

Xinping Zhao, Yan Zhong, Zetian Sun, Xinshuo Hu, Zhenyu Liu, Dongfang Li, Baotian Hu, and Min Zhang. FunnelRAG: A coarse-to-fine progressive retrieval paradigm for RAG. In *Findings of the Association for Computational Linguistics: NAACL*, pages 3029–3046, 2025.

A Entropy Allocation Analysis

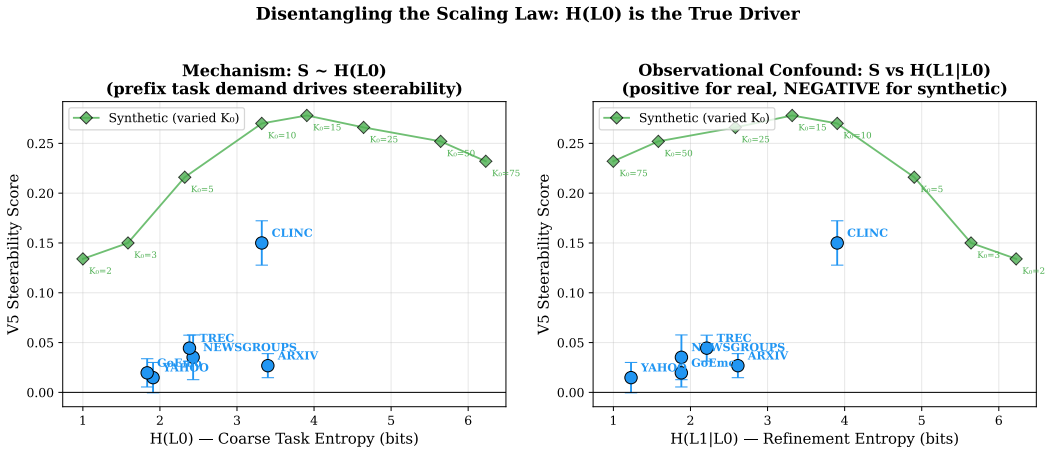

Figure 8: Disentangling the scaling law. **Left:** Steerability vs. $H(L_0)$ (prefix task demand)—the true driver, confirmed by the synthetic experiment. **Right:** Steerability vs. $H(L_1|L_0)$—a confounded proxy in observational data. Synthetic data (green diamonds) breaks the confound: $H(L_1|L_0)$ anti-correlates with steerability when $H(L_0)$ is varied independently.

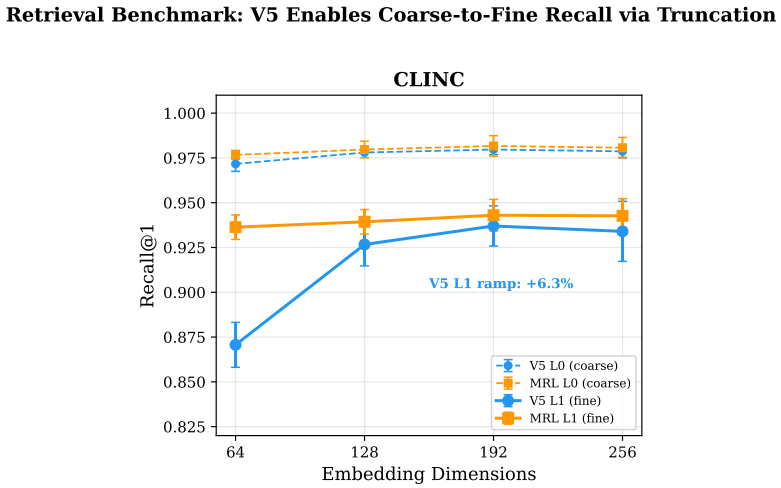

Figure 9: Retrieval benchmark on CLINC-150 (3 seeds). V5 L_1 Recall@1 ramps steeply from 64d to 192d (+6.3pp) while MRL is flat (+0.6pp). Both achieve comparable L_0 Recall@1 ($> 97\%$). The $10\times$ larger ramp demonstrates that prefix specialisation transfers from classification to retrieval.

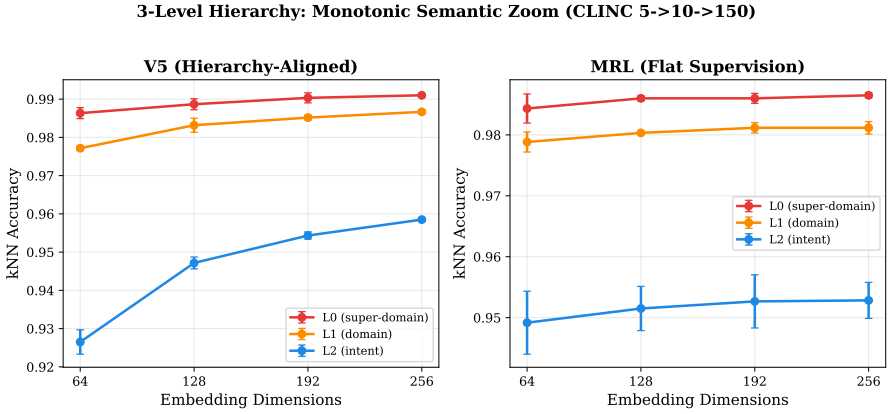

Figure 10: Three-level hierarchy (CLINC, 5→10→150, 3 seeds). **Left**: V5 shows clear level separation—L_2 gains most from additional dimensions while L_0 is near ceiling. **Right**: MRL curves are bunched with minimal ramp at any level.

B Retrieval Benchmark Visualisation

C Three-Level Hierarchy Visualisation

D Workload-Adaptive Pareto Analysis

E FAISS Latency Benchmark

F Full Synthetic Hierarchy Results

See Table 5 for complete results across all 8 coarse partition sizes. The experiment holds total class count fixed at 150 while varying K_0 from 2 to 75 with randomly assigned hierarchies on CLINC-150 text.

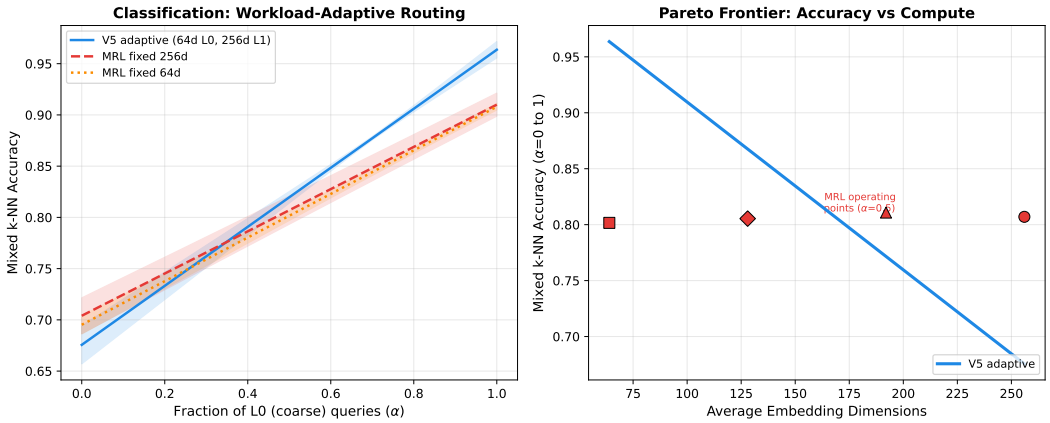

Figure 11: **Left:** Mixed accuracy vs. workload mix α (fraction coarse). V5 adaptive dominates MRL-256d for $\alpha \geq 0.35$. **Right:** Pareto frontier—V5 achieves higher accuracy at lower dimensionality. Bands: ± 1 SD (5 seeds).

Table 9: FAISS query latency at different dimensionalities (RTX 5090, float32). V5's 64d prefix queries are 3.7–5.1× faster than 256d.

	Flat (n=10K)		HNSW (n=100K)	
Dim	Latency (μs)	Speedup	Latency (μs)	Speedup
64d	35	5.1×	39	3.7×
128d	110	1.6×	87	1.7×
192d	146	1.2×	81	1.8×
256d	179	1.0×	145	1.0×

G Training Convergence

All models converge within 5 epochs of head-only training. Stage 2 (backbone fine-tuning) was tested but provided no improvement, consistent with the frozen backbone providing sufficient representational capacity for hierarchy-aligned supervision.

H Reproducibility

Code and data. All code, trained models, and result JSONs are available at `https://github.com/dl1683/ai-moonshots`. Every experiment uses publicly available datasets loaded via the `datasets` library with deterministic seeded splits.

Hyperparameters. All experiments use: 5 epochs head-only training, batch size 16, lr 10^{-4}, AdamW with cosine decay, FP16, gradient clipping at 1.0. Prefix sampling $[0.4, 0.3, 0.2, 0.1]$ and block dropout $[0.95, 0.9, 0.8, 0.7]$ are fixed across all datasets and models. No per-dataset tuning.

Compute. Single NVIDIA RTX 5090 Laptop GPU (24GB VRAM). Each training run: $\sim$2 min (BGE-small, 33M) or $\sim$8 min (Qwen3-0.6B). Full suite: $\sim$16 GPU-hours.

I Metric Robustness

We verify conclusions hold under three alternative steerability formulations:

- $\mathcal{S}_{\text{AUC}}$: average steerability over all prefix pairs $k = 2, 3, 4$.
- $\mathcal{S}_{\text{mono}}$: fraction of adjacent pairs with monotonic coarse-decrease / fine-increase.
- $\mathcal{S}_{\text{gap}}$: specialisation gap $(L_0@j_1 - L_1@j_1) - (L_0@j_4 - L_1@j_4)$.

Table 10: Metric robustness: V5 − MRL gap under four steerability formulations. Three of four agree V5 > MRL on all 8 datasets.

Dataset	$H(L_1\|L_0)$	$\Delta\mathcal{S}_{\text{orig}}$	$\Delta\mathcal{S}_{\text{AUC}}$	$\Delta\mathcal{S}_{\text{mono}}$	$\Delta\mathcal{S}_{\text{gap}}$
Yahoo	1.23	$+0.010$	$+0.013$	-0.10	$+0.010$
GoEmotions	1.88	$+0.014$	$+0.014$	-0.07	$+0.014$
Newsgroups	1.88	$+0.035$	$+0.037$	$+0.03$	$+0.035$
TREC	2.21	$+0.045$	$+0.039$	$+0.27$	$+0.045$
arXiv	2.62	$+0.028$	$+0.020$	$+0.17$	$+0.028$
DBPedia Cl.	3.17	$+0.112$	$+0.106$	$+0.33$	$+0.112$
CLINC	3.90	$+0.143$	$+0.124$	$+0.27$	$+0.143$
WOS	5.05	$+0.036$	$+0.027$	$+0.30$	$+0.036$
V5 > MRL		8/8	8/8	6/8	8/8
Sign test p		0.004	0.004	0.14	0.004

All pairwise rank correlations exceed $\rho = 0.90$ ($p < 0.005$), confirming all four metrics recover the same dataset ordering.

J Per-Seed Steerability Values

Table 11: Per-seed steerability (V5 and MRL) across all eight datasets. Seeds: 42, 123, 456, 789, 1024.

	V5 $\mathcal{S}$ by seed					MRL $\mathcal{S}$ by seed				
Dataset	42	123	456	789	1024	42	123	456	789	1024
Yahoo	+.016	+.020	−.004	−.002	+.044	−.010	+.016	+.006	+.016	−.002
GoEmo	−.002	+.024	+.026	+.006	+.044	+.022	+.022	+.006	−.012	−.010
News	+.040	−.012	+.038	+.044	+.066	−.002	+.022	+.008	−.018	−.010
TREC	+.018	+.062	+.054	+.042	+.046	+.000	+.024	−.014	−.002	−.012
arXiv	+.038	+.018	+.012	+.018	+.048	−.014	−.010	+.000	+.000	+.020
DBP	+.118	+.092	+.130	+.130	+.130	+.020	+.008	+.008	+.000	+.002
CLINC	+.104	+.178	+.150	+.168	+.150	+.012	+.028	+.006	−.016	+.004
WOS	+.032	+.022	+.008	+.052	+.074	+.000	−.004	+.004	−.004	+.010

K Scaling Trend Robustness

Table 12: Leave-one-out sensitivity for the $H(L_1|L_0)$ scaling trend. WOS (Cook's $D = 3.68$) is the highest-influence point; the product predictor resolves its deviation without dropping any dataset.

Dropped	k	Spearman ρ	p	Pearson r	p
None (full)	8	0.74	0.035	0.49	0.218
Yahoo	7	0.61	0.144	0.40	0.376
GoEmotions	7	0.64	0.119	0.45	0.318
Newsgroups	7	0.71	0.071	0.47	0.289
TREC	7	0.83	0.021	0.48	0.272
arXiv	7	0.78	0.041	0.50	0.259
DBPedia Classes	7	0.72	0.068	0.49	0.261
CLINC	7	0.72	0.068	0.34	0.457
WOS	7	0.87	0.012	0.91	0.004

Bootstrap analysis (10,000 resamples): 98.0% of $\rho(H(L_1|L_0))$ values positive. Meta-analysis prediction interval $[-0.70, 3.18]$ reflects heterogeneity ($I^2 = 63\%$) explained by the interaction model.

L Broader Impact

Fractal embeddings add a semantic control knob to existing embedding models without modifying the backbone. The primary application is more efficient retrieval: coarse-first filtering reduces

compute by $4\times$ without sacrificing fine-grained accuracy when needed. We do not foresee negative societal impacts beyond those inherent to embedding-based retrieval generally. The method is domain-agnostic and introduces no biases beyond those present in the frozen backbone.