

Peeling the Matryoshka: Extracting Ontological Structure from Retrieval Embeddings

Eunchan Kim

Hanyang University

eunchan@hanyang.ac.kr

Inwhae Joe

Hanyang University

iwjoe@hanyang.ac.kr

Abstract

Taxonomy construction is essential for knowledge-intensive NLP yet remains costly: clustering-based methods (i) need careful per-corpus tuning and yield unstable hierarchies, while LLM-based pipelines (ii) depend on expensive task-specific training that does not transfer across domains. We report an empirical finding that bears on both: in Matryoshka Representation Learning (MRL) embeddings, where each leading prefix is itself a usable, coarser vector, prefix length aligns with the coarse-to-fine granularity of gold taxonomies; shuffled- and reverse-dimension controls isolate the alignment to the learned early prefix dimensions. The pattern holds across three architecturally distinct MRL models (full statistics in §3). To test whether this regularity is usable, we introduce PEEL (Peeling Embeddings to Extract Latent ontologies): a minimal probe over these frozen embeddings, trained once with an order-embedding objective that represents hierarchy as vector containment; a minimum spanning arborescence then decodes taxonomic DAGs. Using only a concept vocabulary (no documents), PEEL recovers taxonomic DAGs with high structural fidelity on two OLLM benchmarks, transferring to a new domain with no target-edge supervision. The end-to-end OLLM pipeline—which jointly discovers concepts and induces structure from document corpora—is reported as a cross-task reference (§5.3). Pre-trained retrieval embeddings encode a decodable granularity gradient; a minimal probe renders it as explicit parent-child (*is-a*) structure.

1 Introduction

Domain taxonomies, hierarchical structures of concepts linked by taxonomic (*is-a*) relations, are a fundamental form of ontological structure and the backbone of structured knowledge systems, from biomedical terminologies such as MeSH

(Lipscomb, 2000) to web-scale category hierarchies like Wikipedia (Ponzetto and Strube, 2007) and academic classification schemes. Throughout, *ontological structure* refers to this taxonomic (*is-a*) backbone, following OLLM’s usage (Lo et al., 2024). Yet constructing them remains one of NLP’s most labor-intensive tasks.

Automated taxonomy construction has been approached from two directions. Embedding-based methods range from flat clustering of term vectors to recursive or geometry-based schemes (e.g., hyperbolic and box embeddings); the former lose hierarchy, the latter require dedicated, hierarchy-supervised embedding spaces. LLM-based methods instead generate hierarchy explicitly, yet demand extensive fine-tuning or prompt engineering per domain. Meanwhile, structural probing has shown that pre-trained representations encode rich linguistic structure recoverable via simple learned projections (Hewitt and Manning, 2019), but has focused on syntax, not taxonomic hierarchy. An alternative remains unexplored: do embeddings trained purely for retrieval already encode hierarchical signal exploitable for taxonomy construction?

We find that they do: specifically, embeddings trained with Matryoshka Representation Learning (MRL; Kusupati et al., 2022), whose nested prefix structure makes truncation a principled resolution control. Every prefix of the full vector is itself a useful representation, concentrating coarse semantic information in early dimensions. We observe that this coarse-to-fine organization aligns with gold taxonomic structure: across three open-source MRL models spanning distinct encoder families and scales (nomic, 137M, BERT; jina-v3, 572M, XLM-R; Qwen3-Embedding, 600M, decoder), MRL prefix truncation consistently outperforms both a shuffled-dimension control (ordering destroyed, content preserved) and a last- k reverse control (tail dimensions), the advantage con-

centrated in the learned early prefix dimensions. A supporting ANOVA finds the label separation effect large at the deepest taxonomic level, consistent across all model-by-dataset sequences (Table 1; full statistics in §3), indicating the structure is not specific to a single encoder family.

We present PEEL (**P**eeling **E**mbdings to **E**xtract **L**atent ontologies), a lightweight pipeline that exploits this signal. Pre-trained embeddings mix taxonomic structure with retrieval-specific features; PEEL trains a 98K-parameter probe with an order-embedding loss (Vendrov et al., 2016) to project frozen retrieval embeddings into a space where coordinate-wise dominance reflects parent-child relations. A Chu-Liu/Edmonds minimum spanning arborescence (Chu and Liu, 1965; Edmonds, 1967) then extracts a threshold-free taxonomy tree, expanded into a DAG by greedily adding lowest-violation edges up to a budget.

On the OLLM benchmarks (Lo et al., 2024), PEEL—using only the concept vocabulary, no documents—performs taxonomy structure induction over a given concept set, recovering taxonomic DAGs with high structural fidelity (Graph F1 .885/.931 on Wikipedia and arXiv; cross-task comparison with the document-fed OLLM pipeline in §5.3).

Our contributions are:

1. **Discovery.** An empirical finding that pre-trained MRL embeddings encode taxonomic structure in their learned early prefix dimensions, isolated by shuffled- and reverse-dimension controls and replicated across three architecturally distinct open-source MRL models (§3).
2. **Demonstration.** PEEL, a minimal probe decoding taxonomy DAGs via minimum spanning arborescence without task-specific embedding training (§4).
3. **Evaluation.** On two OLLM benchmarks, PEEL recovers taxonomic DAGs from a concept vocabulary alone with high structural fidelity on the primary structure-sensitive metrics (Graph F1 and Cont.F1); we report the document-fed OLLM pipeline—which jointly discovers concepts *and* induces structure from documents—as a cross-task reference (§5).

2 Related Work

2.1 Taxonomy Construction

Automated taxonomy construction has evolved through several paradigms. *Pattern-based* methods, originating with Hearst patterns (Hearst, 1992) and revisited at scale by Roller et al. (2018), extract hypernymy from lexico-syntactic cues but suffer from low recall. *Statistical* approaches integrate heterogeneous evidence to induce taxonomies with supervised classifiers (Snow et al., 2006) or structured prediction (Bansal et al., 2014). Embedding-based methods such as TaxoGen (Zhang et al., 2018) cluster term embeddings recursively.

More recently, large language models have been applied directly. OLLM (Lo et al., 2024) fine-tunes an LLM to generate ontology subgraphs end-to-end; Chain-of-Layer (Zeng et al., 2024) decomposes taxonomy induction into a layer-by-layer prompting procedure; TaxoLLaMA (Moskvoret-skii et al., 2024) unifies taxonomy tasks via instruction tuning; and TaxoAdapt (Kargupta et al., 2025) adapts LLM-generated taxonomies to evolving corpora.

These methods share a common pattern: they either *learn* taxonomy structure through task-specific training or *generate* it via prompted LLMs. LLMs4OL (Babaei Giglou et al., 2024) organises ontology learning into three tasks—term typing, taxonomy discovery, and non-taxonomic relation extraction—and PEEL targets taxonomy discovery: recovering *is-a* pairs over a given set of types. Whether pre-trained embeddings—trained solely for retrieval—already encode exploitable taxonomic structure has remained unexplored.

2.2 Hierarchical Representation Learning

Order Embeddings (Vendrov et al., 2016) encode hierarchy directly in embedding geometry, modelling asymmetric *is-a* relations via coordinate-wise inequalities: a parent’s coordinates are no larger than its child’s in every dimension. Later work extends this ordering to density-based formulations (Athiwaratkun and Wilson, 2018). Poincaré Embeddings (Nickel and Kiela, 2017) use hyperbolic space to capture tree-like hierarchies.

Most closely related to our approach, HiTs (He et al., 2024) retrains pre-trained sentence transformers in Poincaré space with a centripetal loss to produce a hierarchy encoder. While both HiTs

and PEEL map pre-trained language models to hierarchical structure, they differ fundamentally: HiTs *injects* hierarchy by retraining the entire encoder (22M–109M parameters) with a hierarchy-aware objective, whereas PEEL *discovers* hierarchy latent in frozen retrieval embeddings via a 98K-parameter probe.

2.3 Structural Probing of Pre-trained Representations

What linguistic structure pre-trained representations encode has been studied extensively through probing classifiers (Belinkov and Glass, 2019; Conneau et al., 2018; Tenney et al., 2019). Most directly relevant, Hewitt and Manning (2019) demonstrate that a linear transformation of BERT representations recovers parse tree distances, establishing the *structural probe* paradigm: a simple learned projection can extract complex structure from frozen representations. Subsequent work has examined what probing accuracy actually reveals: Hewitt and Liang (2019) propose control tasks to measure probe selectivity, and Voita and Titov (2020) frame probing as compression, arguing that a probe’s minimum description length—not just accuracy—determines whether structure is genuinely encoded.

Crucially, existing structural probes target *syntactic* structure (dependency and constituency trees). PEEL extends this paradigm to *taxonomic* structure: we show that a learned projection of retrieval embeddings, trained with an order embedding loss, recovers parent–child relations in concept hierarchies—a qualitatively different structural signal from syntax.

2.4 Matryoshka Representation Learning

Matryoshka Representation Learning (MRL; Kusupati et al., 2022) trains embeddings so that every prefix of the full vector is itself a valid representation, producing a nested family of $O(\log d)$ representations from a single forward pass. Originally designed for efficient retrieval at variable compute budgets, MRL models are now widely deployed in production systems. Its authors already noted that this nesting “naturally captures the underlying hierarchy of the class labels” in a two-level ImageNet superclass setting (their §5).

Hanley and Durumeric (2025) instead engineer it: they train custom Matryoshka embeddings with a modified contrastive loss that explicitly encodes news similarity hierarchies into different dimen-

sion ranges, an approach we term *hierarchy injection*. We establish the complementary result: standard MRL models—trained only for retrieval, with no hierarchical supervision—already exhibit prefix structure that tracks depth in a multi-level gold concept taxonomy, an alignment that §3 isolates with shuffled- and reverse-dimension controls and replicates across three architecturally distinct encoders. This *hierarchy discovery* requires no custom training and applies to any MRL-capable encoder.

Summary. PEEL sits at the intersection of four threads: it applies the structural probe paradigm (§2.3) with order embedding geometry (§2.2) to frozen MRL representations (§2.4) for taxonomy construction (§2.1), requiring neither task-specific embedding training nor LLM generation.

3 MRL Prefixes Encode Taxonomic Structure

As noted in §1, MRL embeddings concentrate coarse semantic information in early dimensions by design (Kusupati et al., 2022). We hypothesize that this coarseness aligns with *taxonomic structure*: shorter prefixes capture broad category distinctions while longer prefixes preserve finer-grained structure.

Experimental design. We evaluate kNN label purity ($k=10$) across three MRL models spanning distinct encoder families and scales: Nomic-embed-text-v1.5 (Nussbaum et al., 2025) (137 M, BERT), jina-embeddings-v3 (Sturua et al., 2024) (572 M, XLM-R), and Qwen3-Embedding (Zhang et al., 2025) (600 M, decoder), on two taxonomic benchmarks. We compare MRL prefix truncation against two controls: (i) a *shuffled* control that randomly permutes all dimensions before truncation, preserving total information content while destroying the learned dimension ordering (cf. control tasks; Hewitt and Liang, 2019), and (ii) a *reverse* control that truncates the last k dimensions, excluding the front-loaded learned structure.

Results. MRL prefix truncation outperforms *both* controls (the order-destroying shuffled control and the last- k tail-dimension reverse control) across all three architecturally distinct encoder families, with the advantage concentrated in the early, truncated prefix dimensions (Figure 1). At full dimensionality the prefix coincides with the complete embedding, so all three coincide by

Model	Data	Depth 1	Depth 2	Depth 3
Nomic (137M)	wiki	.058	.159	.364
jina-v3 (572M)	wiki	.067	.181	.400
Qwen3 (600M)	wiki	.043	.131	.337
Nomic (137M)	arXiv	.039	.122	—
jina-v3 (572M)	arXiv	.078	.205	—
Qwen3 (600M)	arXiv	.039	.131	—

Table 1: ANOVA effect sizes (η^2) for kNN label purity by taxonomy depth, reported as a supporting cross-model consistency check. arXiv has max depth 2. The depth ordering holds in every row. Because gold class counts differ sharply across depths (Wikipedia: 33/875/7124 at Depth 1/2/3), these values index consistency, *not* an absolute cross-depth magnitude. Cohen’s benchmarks: .01 small, .06 medium, .14 large (Cohen, 1988).

construction; the separation therefore reflects the *learned early-prefix ordering*. That the advantage holds for every model against *both* a randomly permuted and a tail-truncated control isolates the signal to this learned ordering. This directional consistency holds for every model (one-sided sign test; Appendix D). As supporting triangulation, we quantify how strongly taxonomic labels separate at each depth via one-way ANOVA over every embedding dimension, reporting the mean η^2 (Table 1): on Wikipedia, effects are small-to-medium at Depth 1 (.04–.07), medium at Depth 2 (.13–.18), and Cohen’s-large at Depth 3 (.34–.40); the shallower arXiv taxonomy (two gold levels) shows the same direction (Depth 1 .04–.08, Depth 2 .12–.20). The depth ordering of effect sizes holds without exception across all six model×dataset sequences. This η^2 comparison serves only as a supporting cross-model consistency check.

The signal’s concentration in early dimensions, with higher dimensions adding little over the controls, suggests that full embeddings mix taxonomic structure with retrieval-specific features that dilute hierarchy. A learned projection trained to preserve parent–child ordering should therefore separate these components. A training-free pre-probe check is reported in Appendix E. §4 operationalizes this insight.

4 Method

PEEL operationalizes the observation from §3—that MRL prefixes encode taxonomic structure—through three stages (Figure 2): (1) projecting frozen embeddings into an order-embedding space (Vendrov et al., 2016) via a probe trained once on source-domain gold edges, (2) extract-

ing a threshold-free taxonomic tree via Chu-Liu/Edmonds minimum spanning arborescence (Chu and Liu, 1965; Edmonds, 1967), and (3) expanding the tree into a DAG by adding low-violation edges. Each component is deliberately simple: §3 establishes that the hierarchical signal is already present in pre-trained embeddings; PEEL demonstrates that it is operationally exploitable with simple machinery.

4.1 Order-Embedding Probe

Pre-trained MRL embeddings mix taxonomic signal with retrieval-specific features in overlapping dimensions (§3). Following the structural probing paradigm (Hewitt and Manning, 2019), we train a learned projection to separate these components:

$$\mathbf{f} = \text{ReLU}(W\mathbf{e} + \mathbf{b}), \quad W \in \mathbb{R}^{d_{\text{out}} \times D} \quad (1)$$

We take this paradigm as conceptual inspiration—a minimal probe can surface latent structure—and PEEL serves as a demonstration that the structure §3 establishes is operationally exploitable, not as a diagnostic claim about the embeddings. ReLU projects outputs into the non-negative orthant $\mathbb{R}_+^{d_{\text{out}}}$ required by the order-embedding framework (Vendrov et al., 2016). We set $d_{\text{out}}=128$ based on ablation (Table 4).

We train the projection using order-embedding loss (Vendrov et al., 2016):

$$\mathcal{L} = \mathcal{L}_+ + \mathcal{L}_- \quad (2)$$

$$\mathcal{L}_+ = \frac{1}{|\mathcal{P}|} \sum_{(p,c) \in \mathcal{P}} \|\max(\mathbf{0}, \mathbf{f}_p - \mathbf{f}_c)\|^2 \quad (3)$$

$$\mathcal{L}_- = \frac{1}{|\mathcal{N}|} \sum_{(p,n) \in \mathcal{N}} \max(0, m - \|\max(\mathbf{0}, \mathbf{f}_p - \mathbf{f}_n)\|^2) \quad (4)$$

where $\mathcal{L}_+$ encourages $\mathbf{f}_p \leq \mathbf{f}_c$ for positive parent–child pairs $\mathcal{P}$, and $\mathcal{L}_-$ pushes negatives apart by margin m ($\mathcal{N}$ = negative pairs). Because §3 isolates this signal using controls that involve no probe training, the learned projection recovers a granularity ordering the embeddings already encode rather than imposing it through supervision.

We choose order embeddings for two reasons: (1) they directly encode the asymmetric is-a relation, unlike symmetric contrastive losses; and (2) the coordinate-wise ordering constraint yields the violation score used for tree extraction (§4.3).

4.2 Random Truncation Training

Datasets differ in domain and structural profile, so no single truncation captures the full taxonomic

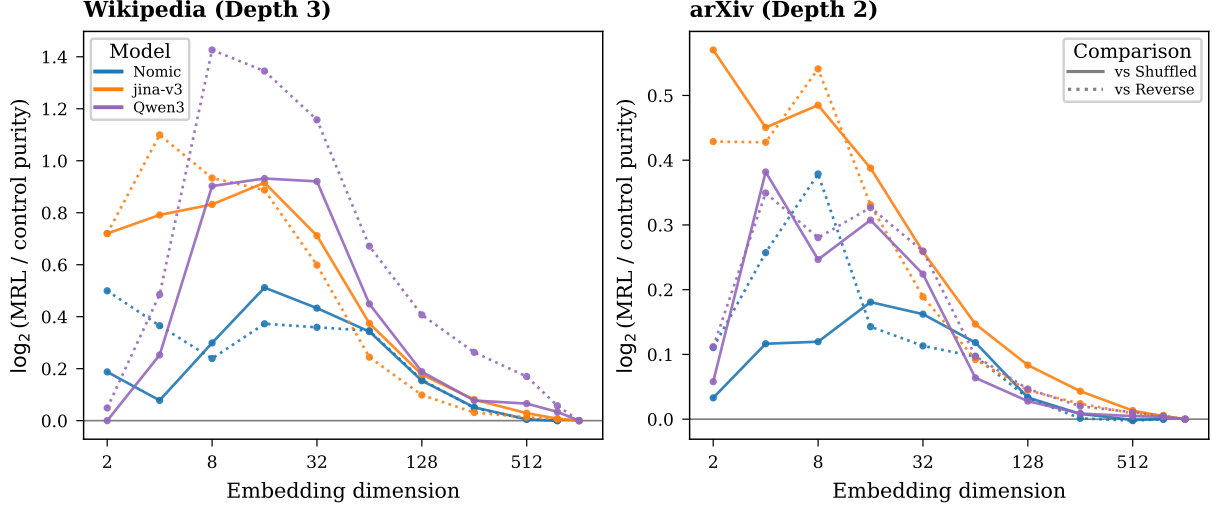


Figure 1: Relative advantage of MRL prefix truncation over two controls, by embedding dimension, on Wikipedia and arXiv at the deepest gold taxonomic level. $y = \log_2(\text{MRL} \div \text{control kNN purity})$; $y=0$ is parity with the control. Solid = vs. shuffled control (dimension ordering destroyed); dotted = vs. reverse control (last- k tail dimensions). Colors denote three architecturally distinct MRL models (Nomic, jina-v3, Qwen3). All curves rise above parity and peak within the early dimensions; at full dimensionality the prefix equals the complete embedding, so the controls coincide by construction.

structure. We exploit MRL’s prefix-truncation property: at each training step, we randomly sample $d \in \{64, 128, 256, 384, 512, 768\}$, truncate all embeddings to d dimensions, and zero-pad back to $D=768$ before projection. This produces a single model that handles all resolutions; at inference, we select the best dimensionality per dataset. Training at a single truncation length would tie the probe to that resolution’s feature distribution; random truncation instead forces it to learn order-embedding relationships that hold across all resolutions simultaneously (Table 3).

4.3 Tree Extraction via Chu-Liu/Edmonds

Given projected embeddings $\mathbf{f}$, we define the violation score between nodes i and j as:

$$v(i, j) = \sum_{k=1}^{d_{\text{out}}} [\max(0, f_k(i) - f_k(j))]^2 \quad (5)$$

A low $v(i, j)$ indicates that $\mathbf{f}_i \leq \mathbf{f}_j$ coordinate-wise, suggesting i is a parent of j .

We construct a directed graph weighted by violation scores and extract the minimum spanning arborescence (MSA) via the Chu-Liu/Edmonds algorithm (Chu and Liu, 1965; Edmonds, 1967), producing an optimal rooted tree with exactly $N-1$ edges. CLE requires no threshold, unlike containment-ratio or edge-weight thresholding approaches.

4.4 DAG Expansion

A tree contains exactly $N-1$ edges, which may undercount the true is-a relations in a polyhierarchical ontology. To recover additional edges, we rank all non-tree node pairs by violation score and greedily add the lowest-violation pairs while maintaining acyclicity, up to an edge budget n_{expand} (§5.1).

5 Experiments

5.1 Setup

Datasets. We evaluate on two datasets from the OLLM benchmarks (Lo et al., 2024), which pair document corpora with gold taxonomies. The two datasets span distinct domains and structural profiles. Wiki-ol provides 8,033 concept terms with a gold taxonomy of 14,673 edges at depth 3. ArXiv-ol contains 61 categories in a shallow hierarchy of depth 2.

Evaluation Metrics. We report five metrics from the OLLM evaluation suite (Lo et al., 2024): Lit.F1 (literal edge overlap, OLLM’s sanity-check reference), Fuzzy F1 (thresholded edge matching; all-MiniLM-L6-v2, $\tau=0.436$), Cont.F1 (threshold-free Hungarian edge matching, scored by the weaker of the two endpoint similarities), GraphF1 (Hungarian node matching over SG-Conv neighbourhood embeddings), and Motif Dist. (triadic-census total-variation divergence). We designate GraphF1 and Cont.F1 as the primary

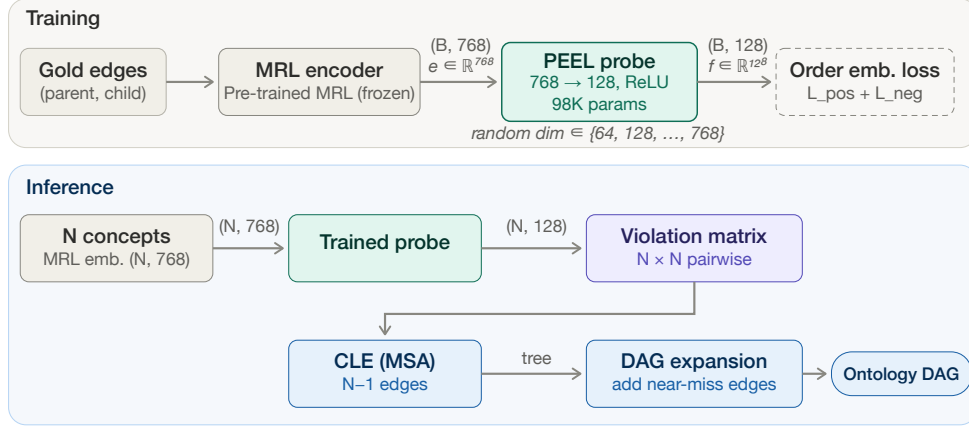


Figure 2: The PEEL pipeline. **Training** (top): frozen MRL embeddings are projected through a probe (98K parameters) trained with order embedding loss under random dimensional truncation. **Inference** (bottom): the trained probe maps N concept embeddings to a 128-d order-embedding space; a pairwise violation matrix is constructed and resolved by Chu-Liu/Edmonds’ algorithm into a threshold-free tree, then expanded into a directed acyclic graph via near-miss edges.

structure-sensitive metrics: they capture global structural similarity and edge-endpoint correspondence, and are robust to surface-level label variation.

Baselines. All baselines are from Lo et al. (2024) with reported numbers: Memorisation (training edges verbatim), Hearst Patterns, Zero-shot and Finetune (Mistral-7B), and the full OLLM pipeline. PEEL receives only the gold concept vocabulary; all LLM-based baselines (Zero-shot, Finetune, OLLM) additionally consume each article’s title and abstract per the OLLM protocol (Lo et al., 2024, Figs. 5–6), and Hearst extracts from the source corpus. For arXiv-ol, OLLM (transfer) fine-tunes on 2,048 additional arXiv samples; PEEL uses no arXiv training data (zero-shot).

Implementation. PEEL uses Nomic-embed-text-v1.5 (Nussbaum et al., 2025) as the frozen encoder (768-dimensional). The probe is a single linear layer (768→128) followed by ReLU, totaling 98K trainable parameters. We train with Order Embedding loss (margin $m=4.0$) using Adam (learning rate 5×10^{-4} , batch size 512) on wiki-ol gold parent–child edges for 10 epochs on a single GPU (~ 4 seconds).¹

Training and inference protocol. The probe is trained on wiki-ol gold parent–child edges, with the training graph split 80%/10%/10% into train/validation/holdout (edge-level split); each

epoch we track validation loss and edge-ranking AUROC, retaining the lowest-loss checkpoint over the 10-epoch schedule. Negative pairs use a 10:1 negative-to-positive ratio, mixing 50% random nodes, 30% sibling distractors (other children of the same parent), and 20% reversed pairs. For tree extraction, the virtual root connects to every node at edge weight equal to the 95th percentile of positive violation scores. Wiki-ol inference hyperparameters (prefix truncation $d=512$, DAG expansion budget $n_{\text{expand}}=15,000$, CLE candidate count $k=10$) are chosen on wiki-ol validation performance; Table 3 reports per- d sensitivity. For arxiv-ol, the wiki-ol-trained probe is applied unchanged, with inference settings ($d=768$, $k=20$, tree-only) following from the dataset’s compact 61-concept structure rather than from arxiv-ol-specific tuning; no arxiv-ol data is used for training, validation, or hyperparameter selection.

5.2 Structural Fidelity from a Concept Vocabulary

PEEL (a 98K-parameter probe receiving *only* the concept vocabulary) recovers taxonomic DAGs with GraphF1 .885 on wiki-ol (supervised, edge-level 80/10/10 split) and .931 on arXiv-ol (zero-shot transfer from the wiki-ol-trained probe). OLLM, which consumes the document corpus and fine-tunes Mistral-7B to jointly discover concepts and induce structure, scores .644 and .633 on the same benchmarks as a cross-task reference (§5.3). Table 2 reports all five metrics.

The probe trained on 80% of wiki-ol gold edges (§5.1) recovers .885 GraphF1 across the

¹Code and data: <https://github.com/Eunchan24/PEEL>

Dataset	Method	Input	Lit.F1↑	Fuz.F1↑	Cont.F1↑	Grph.F1↑	Motif.D↓
<i>Concept-only methods (no document access)</i>							
wiki-ol	Memorisation	training edges	.134	.837	.314	.419	.063
	PEEL (ours)	concepts only	.111	.877	.551	.885	.000
<i>Document-using methods (corpus + concepts)</i>							
wiki-ol	Hearst Patterns	corpus + concepts	.003	.538	.350	.544	.163
	Zero-shot (M-7B)	title + abstract	.007	.871	.455	.639	.341
	Finetune (M-7B)	title + abstract	.124	.884	.470	.588	.050
	OLLM	title + abstract	.093	.915	.500	.644	.080
arXiv-ol	OLLM (transfer)	title + abstract	.040	.570	.357	.633	.097
	PEEL (zero-shot)	concepts only	.298	.748	.604	.931	.011

Table 2: Main results on OLLM benchmarks, grouped by input modality. LLM-based baselines additionally consume each article’s title and abstract per the OLLM protocol (Lo et al., 2024, Figs. 5–6); OLLM (transfer) fine-tunes Mistral-7B on 2,048 additional arXiv samples. PEEL rows use a gold concept vocabulary and a probe trained on wiki-ol gold edges (§5.1), the arXiv-ol row zero-shot; OLLM instead discovers concepts and structure from documents, so the OLLM scores are a cross-task reference, not a ranking. Motif Dist. values below 10^{-4} are reported as .000.

full 8,033-concept structure, generalizing the taxonomic signal beyond the training-edge-replay floor (Memorisation: .419).

Wiki-ol. On wiki-ol’s 8,033 concepts, Cont. F1 rises from .500 (OLLM) to PEEL’s .551. On the two label-string metrics (Lit. F1, Fuzzy F1), LLM label-generation systems hold an expected advantage; Memorisation tops Lit. F1 (.134) by replaying training edges verbatim, and OLLM leads Fuzzy F1 (.915) via title-string overlap.

ArXiv-ol (zero-shot transfer). The wiki-ol-trained probe transfers zero-shot to arXiv-ol’s 61 categories: PEEL attains Graph F1 .931 and Cont. F1 .604; OLLM, which additionally fine-tunes Mistral-7B on 2,048 arXiv samples, scores .633 and .357.

5.3 Cross-Task Reference

PEEL performs structure induction from a concept vocabulary alone. OLLM addresses a broader task—joint concept discovery *and* structure induction—and, like the Mistral-7B baselines, consumes article text. The comparison is therefore asymmetric in both task scope and input: PEEL is given strictly less. We report it for reference, not as a same-conditions claim.

5.4 Ablation Studies

Embedding dimension. Table 3 varies the input prefix truncation d with probe output fixed at $d_{\text{out}}=128$. Cont. F1 peaks at $d=512$ (.551) and drops at full dimensionality (.479 at $d=768$), matching §3’s coarse-to-fine prefix gradient. We select $d=512$ on validation Cont. F1; Graph F1

stays in .879–.906 across dimensions, peaking at $d=768$. Even with $d_{\text{out}}=32$ (24K parameters; Appendix A, Table 4), PEEL yields Graph F1 .870 and Cont. F1 .517.

Multi-resolution decomposition. We extract Chu-Liu/Edmonds trees at each truncation point. Edge-set Jaccard between all dimension pairs stays below .18, with adjacent dimensions ($d=512$ vs. $d=768$: .178) sharing more edges than distant ones ($d=64$ vs. $d=768$: .018). Across all five dimensions, 99% of nodes change parents: prefix truncation yields *non-nested decompositions*, each prefix length inducing a qualitatively distinct hierarchy across resolutions. At the category level, the effect is vivid: the *Mathematics* subtree grows from 1 child at $d=64$ to 19 at $d=256$ and 68 at $d=768$, illustrating how increasing prefix dimension progressively unfolds fine-grained taxonomic structure (Figure 3).

6 Discussion

What the probe reveals about embedding structure. A 98K-parameter probe recovers taxonomic DAGs (Graph F1 .885 on Wikipedia, .931 on arXiv) from *frozen* retrieval embeddings, showing that the prefix-ordering regularity §3 isolates is decodable into explicit parent–child structure without retraining the encoder. The coarse-to-fine gradient isolated by shuffled- and reverse-dimension controls (§3) implies that MRL’s nested-prefix training implicitly organises semantic granularity along the dimension axis, a structural regularity that, to our knowledge, has not been previously documented in standard MRL encoders trained purely for retrieval.

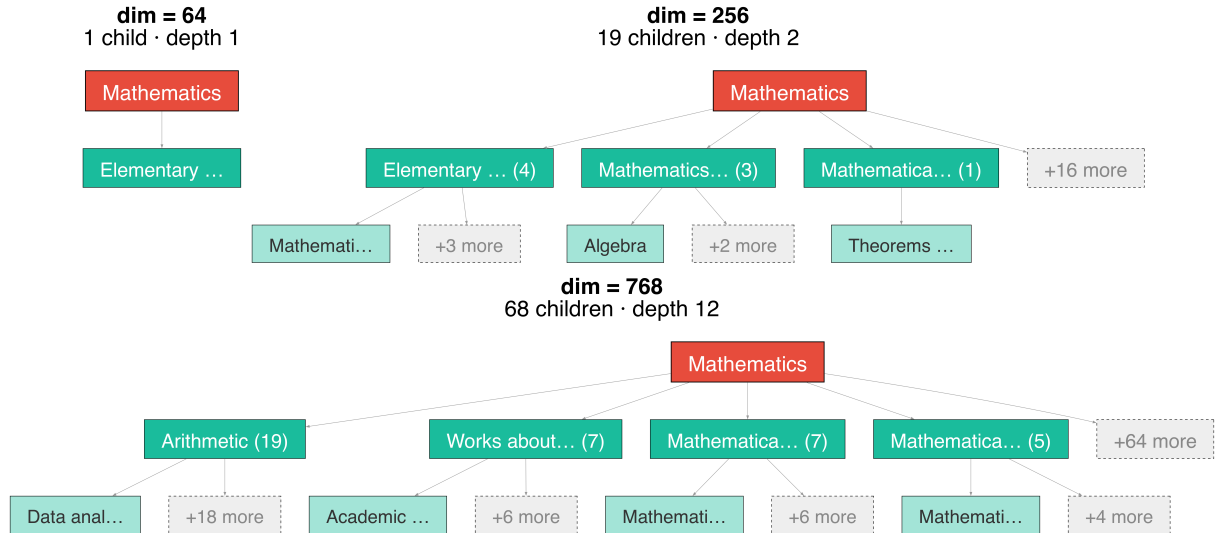


Figure 3: Mathematics subtree extracted at three prefix dimensions from the trained probe: at $d=64$, only one child is recovered; at $d=256$, 19 children emerge across depth 2; at $d=768$, 68 children span 12 levels of depth. Increasing prefix dimension progressively unfolds fine-grained hierarchical structure.

Euclidean order embeddings vs. hyperbolic alternatives. PEEL’s Euclidean order-embedding probe avoids the encoder-retraining required by hyperbolic alternatives such as HiTs (He et al., 2024), achieving structural fidelity with no hierarchy-aware encoder training (geometry comparison in Appendix C).

Each prefix length captures a different hierarchy. §5.4 reports that CLE trees at different prefix lengths yield non-nested, complementary decompositions. This multi-resolution property—extracting multiple taxonomic granularities from a single forward pass over a frozen retrieval embedding, with no retraining or per-resolution recomputation—directly leverages MRL’s nested-prefix design. The kNN purity gradient (§3) and the prefix-level tree disjointness are complementary: the former is a local neighbourhood measure, the latter a global parent–child topology, so prefixes can share neighbourhoods yet decode distinct global hierarchies.

Structure induction as a modular component. PEEL performs structure induction *from a concept vocabulary alone*, isolating it as a separable subproblem: a 98K-parameter probe over frozen embeddings recovers taxonomic DAGs at Graph F1 .885 and .931 with no document access. Any concept-discovery front-end—an LLM extractor, a corpus miner, or a curated vocabulary—can supply that vocabulary. OLLM, which couples concept discovery and structure induction in a single document-fed pipeline, addresses a broader

task and serves as a cross-task reference (§5.3). Pairing PEEL with any such front-end yields an end-to-end taxonomy pipeline without retraining the structure-induction stage.

7 Conclusion

Pre-trained Matryoshka Representation Learning (MRL) embeddings organise semantic granularity along the dimension axis; shuffled- and reverse-dimension controls isolate this gradient to the learned early prefix dimensions across three architecturally distinct MRL models (§3). Exploiting this signal, PEEL (a 98K-parameter probe over frozen embeddings, applied to a concept vocabulary alone) recovers taxonomic DAGs with high structural fidelity (Graph F1 .885 on Wikipedia, .931 on arXiv); the document-fed OLLM pipeline is reported as a cross-task reference (§5.3).

A retrieval-efficiency objective thus yields a multi-resolution taxonomic geometry that PEEL can decode.

Integrating PEEL with corpus-based concept discovery for end-to-end taxonomy construction is a natural direction for future work.

Limitations

Our work has the following limitations. (1) Our evaluation covers two domains from the OLLM benchmarks (Lo et al., 2024) spanning distinct structural profiles (deep polyhierarchy and shallow tree). PEEL transfers zero-shot from wiki-ol to arXiv-ol (§5.2). Our validated transfer covers target ontologies no deeper than the source

domain; how far this reach extends is genuinely open. Extension to specialised domains such as biomedicine and law is a natural direction for future work. (2) We adopt the five structural metrics established by the OLLM benchmark (Lo et al., 2024) (Lit F1, Fuzzy F1, Cont F1, Graph F1, Motif Distance), which evaluate taxonomy structure directly. Label realisation quality and downstream utility (e.g., retrieval, question answering) are outside this paper’s scope. (3) The resolution-control analysis in §3 relies on MRL’s nested prefix-truncation property. The probe itself (§4) operates on any fixed-dimensional embedding. (4) PEEL assumes a gold concept vocabulary and source-domain training edges; transfer requires no target-domain edges.

Ethics Statement

This work extracts taxonomic structure from publicly available pre-trained embedding models and evaluates on published benchmarks (Lo et al., 2024). No human subjects, private data, or crowd-sourced annotations are involved. The taxonomies we construct reflect biases present in the pre-training corpora of the underlying MRL models; users deploying PEEL in sensitive domains should audit the resulting hierarchies for undesirable biases or omissions. Our method’s low computational cost (seconds on a single GPU) reduces the environmental footprint relative to approaches that require fine-tuning large language models.

References

- Ben Athiwaratkun and Andrew Gordon Wilson. 2018. Hierarchical density order embeddings. In *International Conference on Learning Representations (ICLR)*.
- Hamed Babaei Giglou, Jennifer D’Souza, Sameer Sadruddin, and Sören Auer. 2024. [LLMs4OL 2024 datasets: Toward ontology learning with large language models](#). *Open Conference Proceedings*, 4:17–30.
- Mohit Bansal, David Burkett, Gerard de Melo, and Dan Klein. 2014. Structured learning for taxonomy induction with belief propagation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1041–1051.
- Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics (TACL)*, 7:49–72.
- Yoeng-Jin Chu and Tseng-Hong Liu. 1965. On the shortest arborescence of a directed graph. *Scientia Sinica*, 14:1396–1400.
- Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd edition. Lawrence Erlbaum Associates.
- Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $\mathbb{R}^d$ vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2126–2136.
- Jack Edmonds. 1967. Optimum branchings. *Journal of Research of the National Bureau of Standards B*, 71B(4):233–240.
- Hans W. A. Hanley and Zakir Durumeric. 2025. Hierarchical level-wise news article clustering via multilingual matryoshka embeddings. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Yuan He, Zhangdie Yuan, Jiaoyan Chen, and Ian Horrocks. 2024. Language models as hierarchy encoders. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING)*, pages 539–545.
- John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743.
- John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.
- Priyanka Kargupta, Nan Zhang, Yunyi Zhang, Rui Zhang, Prasenjit Mitra, and Jiawei Han. 2025. TaxoAdapt: Aligning LLM-based multidimensional taxonomy construction to evolving research corpora. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. 2022. Matryoshka representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Carolyn E. Lipscomb. 2000. Medical subject headings (MeSH). *Bulletin of the Medical Library Association*, 88(3):265–266.

Andy Lo, Albert Q. Jiang, Wenda Li, and Mateja Jamnik. 2024. End-to-end ontology learning with large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Viktor Moskvoretskii, Ekaterina Neminova, Alina Lobanova, Alexander Panchenko, and Irina Nikishina. 2024. TaxoLLaMA: WordNet-based model for solving multiple lexical semantic tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

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

Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. 2025. Nomic Embed: Training a reproducible long context text embedder. *Transactions on Machine Learning Research*.

Simone Paolo Ponzetto and Michael Strube. 2007. Deriving a large scale taxonomy from Wikipedia. In *Proceedings of the 22nd National Conference on Artificial Intelligence (AAAI)*, pages 1440–1445.

Stephen Roller, Douwe Kiela, and Maximilian Nickel. 2018. Hearst patterns revisited: Automatic hypernym detection from large text corpora. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 358–363.

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2006. Semantic taxonomy induction from heterogeneous evidence. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)*, pages 801–808.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, and Han Xiao. 2024. jina-embeddings-v3: Multilingual embeddings with task LoRA. *arXiv preprint arXiv:2409.10173*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4593–4601.

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

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196.

Configuration	Cont.F1↑	Grph.F1↑
<i>Input prefix truncation d ($d_{out}=128$ fixed)</i>		
$d = 64$	.463	.879
$d = 128$	.485	.880
$d = 256$	.519	.879
$d = 384$	.537	.880
$d = 512^*$	.551	.885
$d = 768$	.479	.906
OLLM (reference)	.500	.644

Table 3: Input prefix truncation d ablation on wiki-ol. * denotes the Table 2 configuration. OLLM (reference) included for the same metrics.

Qingkai Zeng, Yuyang Bai, Zhaoxuan Tan, Shangbin Feng, Zhenwen Liang, Zhihan Zhang, and Meng Jiang. 2024. Chain-of-Layer: Iteratively prompting large language models for taxonomy induction from limited examples. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM)*.

Chao Zhang, Fangbo Tao, Xiusi Chen, Jiaming Shen, Meng Jiang, Brian Sadler, Michelle Vanni, and Jiawei Han. 2018. TaxoGen: Unsupervised topic taxonomy construction by adaptive term embedding and clustering. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 Embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

A Additional Ablation Studies

Table 3, referenced from §4 and §5, reports PEEL’s primary ablation—input prefix truncation d . Figure 4 visualises this sweep, showing PEEL’s Graph F1 and Cont. F1 relative to the OLLM reference baselines across the tested dimensions.

Probe output dimension. Table 4 varies d_{out} with input dimension fixed at $d=512$. Even at $d_{out}=32$ (24K parameters), PEEL achieves Graph F1 .870 and Cont. F1 .517. Performance improves monotonically with d_{out} , confirming that additional projection capacity helps but is not strictly necessary.

Edge budget sensitivity. Table 5 varies the DAG expansion budget n_{expand} with all other hyperparameters fixed. Graph F1 remains above .868 across all tested budgets (10K–25K). Cont. F1 peaks at $n_{expand}=15,000$, the configuration used in Table 2. The stability across budgets confirms

d_{out}	Params	Cont. F1 $\uparrow$	Graph F1 $\uparrow$
32	24K	.517	.870
64	49K	.544	.878
128*	98K	.551	.885
OLLM (ref)	7B	.500	.644

Table 4: Probe output dimension sweep on wiki-ol ($d=512$ fixed). * denotes the Table 2 configuration.

n_{expand}	Cont. F1 $\uparrow$	Graph F1 $\uparrow$	Motif Dist. $\downarrow$
10,000	.503	.899	.000
12,000	.533	.893	.000
15,000*	.551	.885	.000
20,000	.500	.876	.001
25,000	.451	.868	.001
OLLM (ref)	.500	.644	.080

Table 5: DAG expansion budget (n_{expand}) sensitivity on wiki-ol. * denotes the Table 2 configuration. Graph F1 is stable across all tested budgets.

that n_{expand} is not a sensitive hyperparameter; Table 6 reports the corresponding literal edge precision and recall.

B Qualitative Analysis

Dimensional sensitivity of tree topology. CLE trees extracted at different dimensions exhibit structurally distinct topologies (e.g., 10 top-level categories at $d=64$ vs. 17 at $d=512$ on arXiv-ol), consistent with the dimensional sensitivity observed in Figure 1. Figures 5 and 6 extend the *Mathematics* example from Figure 3 to two additional categories, confirming that prefix-dependent unfolding of fine-grained structure generalizes across diverse subtrees.

Edge overlap vs. structural similarity. On arXiv-ol, PEEL reproduces 18 of 60 gold edges exactly (30%), while achieving Graph F1 .931. This gap reflects the difference between exact edge matching and SGConv-based structural similarity: PEEL captures global topology even when individual parent-child assignments differ from the gold standard.

C Hyperbolic Comparison: Order Embeddings vs. Poincaré Approaches

PEEL’s probe operates in Euclidean space with coordinate-wise dominance constraints (Vendrov et al., 2016), rather than in hyperbolic space (Nickel and Kiela, 2017). This appendix details the comparison referenced in §6.

n_{expand}	Lit-P $\uparrow$	Lit-R $\uparrow$
tree only	16.9	9.3
10,000	14.6	10.0
12,000	12.8	10.5
15,000*	11.0	11.2
20,000	9.2	12.6
25,000	8.0	13.7

Table 6: Literal edge precision and recall by expansion budget on wiki-ol (%). * denotes the Table 2 configuration; the reported setting is balanced rather than recall-maximizing. See §5.1 for the primary structure-sensitive metrics.

Hyperbolic encoders. Poincaré-based approaches such as HiTs (He et al., 2024) retrain the entire encoder (22M–109M parameters) with hierarchy-aware losses, exploiting the exponential volume growth of hyperbolic space to capture tree-like hierarchies with low distortion.

Euclidean order embeddings. PEEL trains only a 98K-parameter projection on top of frozen MRL embeddings using an order-embedding loss, requiring no hierarchy-aware pre-training of the encoder. The coordinate-wise dominance constraint directly encodes the asymmetric is-a relation, and the violation score derived from this constraint provides the edge weight for CLE tree extraction (§4.3).

Comparison summary. Both approaches map pre-trained language models to hierarchical structure, but differ on two axes: (i) *trainable parameters*—22M–109M (HiTs) versus 98K (PEEL), a three-order-of-magnitude reduction; (ii) *training objective*—hierarchy-aware contrastive loss applied to the entire encoder (HiTs) versus order-embedding loss applied only to a lightweight probe over frozen embeddings (PEEL). A direct head-to-head benchmark comparison between HiTs and PEEL on the OLLM taxonomies is left for future work.

D Directional-Consistency Counts

Table 7 reports, for each model and benchmark, the number of prefix $\times$ depth conditions in which MRL prefix truncation exceeds the shuffled-dimension control, with the combined-benchmark one-sided sign test (null $p_0=0.5$). Every per-model p is below 10^{-6} . The four non-wins in the Nomic-arXiv cell are the full-dimensionality conditions, where the prefix coincides with the complete embedding and MRL, shuffled, and reverse

Model	Wiki	arXiv	Comb.	one-sided p
Nomic	26/30	16/20	42/50	5.8×10^{-7}
jina-v3	30/33	20/22	50/55	1.1×10^{-10}
Qwen3	29/33	20/22	49/55	9.1×10^{-10}

Table 7: Per-condition MRL > shuffled win counts (prefix $\times$ depth) and combined one-sided sign test, by model and benchmark. All per-model $p < 10^{-6}$.

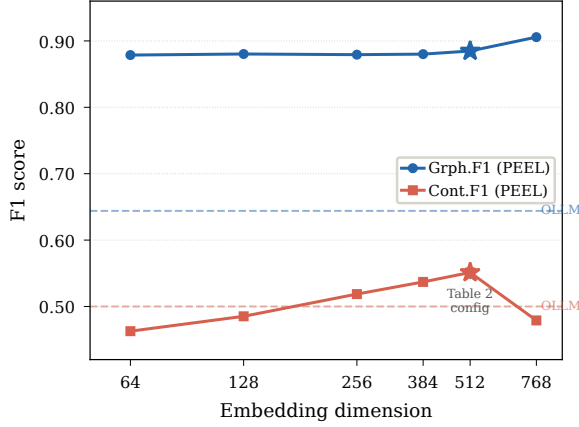


Figure 4: PEEL performance by embedding dimension on wiki-ol ($d_{\text{out}}=128$ fixed). * marks the Table 2 configuration (dim=512). Dashed lines show OLLM baselines.

are identical by construction; the cell remains significant (16/20, $p=5.9 \times 10^{-3}$).

E Raw Geometry Before the Probe

Before any probe training, the frozen embeddings already recover which of two concepts is more general: a training-free measure over raw prefixes scores 69.1% on gold parent–child pairs and 91.3% on ancestor–descendant pairs at depth gap ≥ 2 , while equal-granularity siblings sit at chance (50.0%). Explicit, coordinate-aligned *is-a* containment remains the probe’s contribution: order violation on the raw axes is at chance under a random-rotation null.

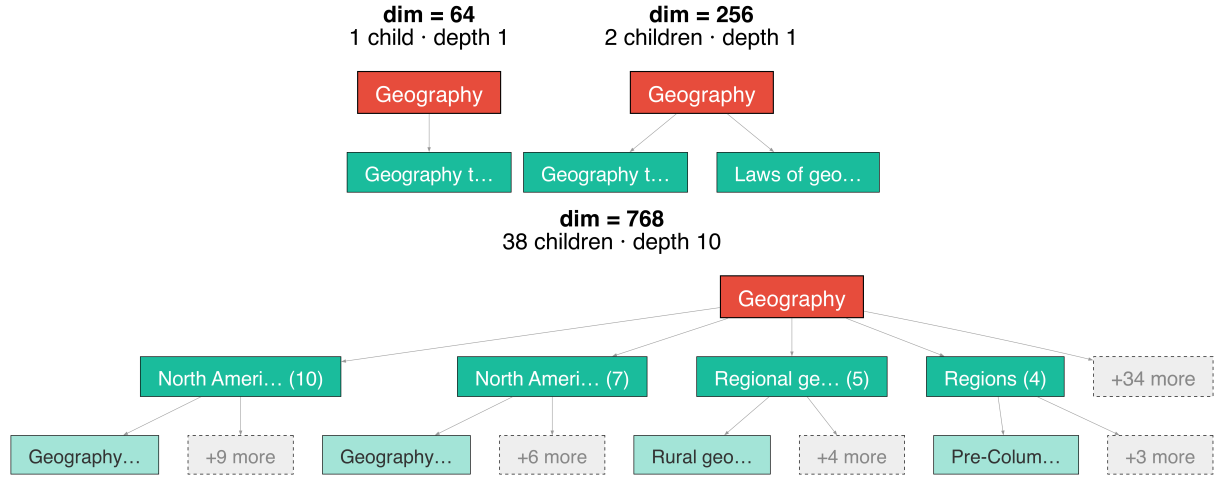


Figure 5: *Geography* subtree at three prefix dimensions. At $d=64$, only 1 child is recovered at depth 1; at $d=256$, 2 children remain at depth 1; at $d=768$, 38 children span depth 10. Compare with Mathematics (Figure 3).

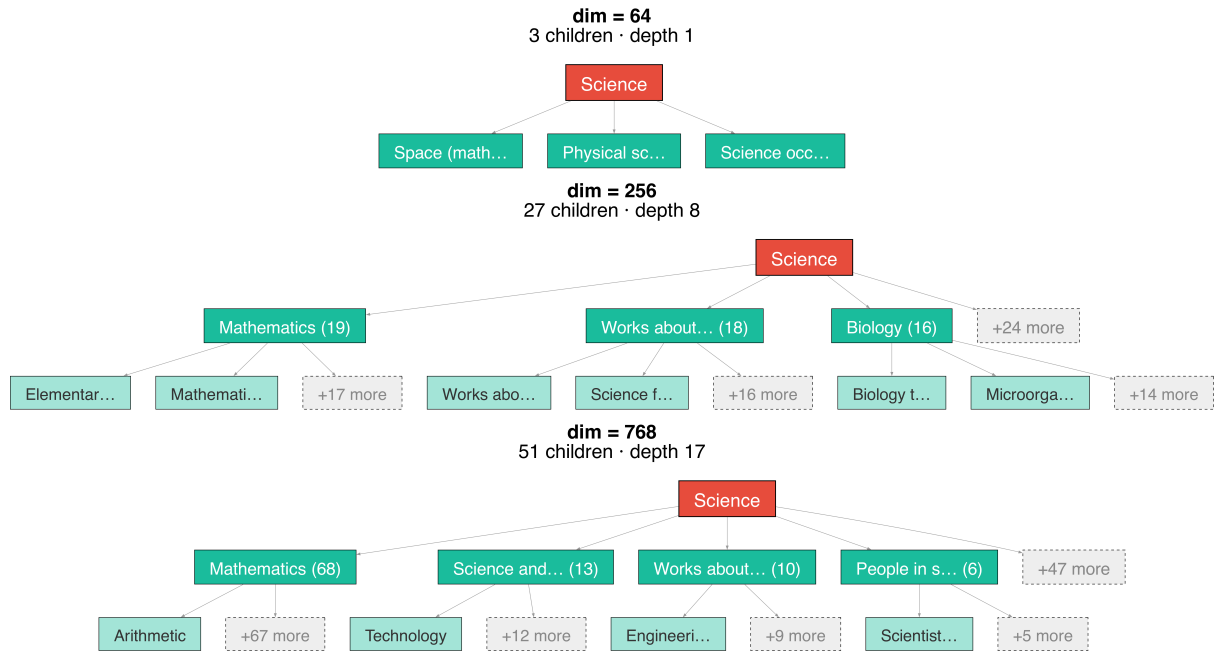


Figure 6: *Science* subtree at three prefix dimensions. At $d=64$, 3 children appear at depth 1; at $d=256$, 27 children span depth 8; at $d=768$, 51 children span depth 17. Together with Geography (Figure 5) and Mathematics (Figure 3), these examples confirm that prefix truncation induces qualitatively distinct hierarchical decompositions across diverse categories, consistent with the low inter-dimension Jaccard similarity ($< .18$) reported in §5.4.