One Corpus, Many Hierarchies: Generating Diverse Taxonomies via Recursive Orthogonalization

Anonymous ACL submission

Abstract

Taxonomies are essential for organizing large and heterogeneous document collections into meaningful hierarchical structures. However, most existing taxonomy generation methods focus on producing a single hierarchy, even when the same corpus can be interpreted from multiple valid perspectives. To address this gap, we propose *Recursive Orthogonalization Of Taxonomies* (ROOT), an unsupervised and structurally rigorous framework that generates a set of K diverse, semantically labeled taxonomies from a document collection. This algorithm leverages the geometry of the document embedding space with LLM usage limited to node labeling. ROOT accumulates taxonomies that together span the latent semantic space of the corpus from genuinely different viewpoints while ensuring maximum coverage and maintaining competitive within-taxonomy coherence and human-judged interpretability. To the best of our knowledge, no prior research has explored the generation of diverse taxonomies from a set of documents. We evaluate ROOT on two published scholarly taxonomy-generation datasets and introduce a strong evaluation suite for assessing the diversity, semantics, structure, and node-label quality of the generated taxonomies.

1 Introduction

Researchers are increasingly overwhelmed by the accelerating volume of published work (Hanson et al., 2024). As scientific and technical literature continues to expand across domains (Reisz et al., 2022), organizing large document collections into meaningful and navigable structures has become increasingly important. Taxonomies provide one such structure: they group documents or concepts into hierarchical categories that capture both topical similarity and semantic abstraction. Because of this, taxonomies have long served as effective knowledge organization tools (Wang et al., 2017) and have been shown to improve a wide range of Natural Language Processing (NLP) and Information Retrieval (IR) applications, including question answering (Yang et al., 2017), textual entailment (Geffet and Dagan, 2005), personalized recommendation (Zhang et al., 2014), and query understanding (Hua et al., 2017).

A key limitation of many existing taxonomy construction methods (Lahiri et al., 2025; Zhu et al., 2025) is that they tend to produce a single dominant hierarchy, even when the underlying corpus admits multiple valid organizational views. Existing approaches to taxonomy construction offer no satisfactory solution to this problem. Deterministic methods produce one hierarchy per run and running them repeatedly yields variations on the same dominant theme rather than orthogonal perspectives. LLM-based methods that prompt for outlines are sensitive to the framing of the prompt but do not provide a systematic mechanism for diversity.

To address this gap, we propose ROOT (**R**ecursive **O**rthogonalization **O**f **T**axonomies), a framework that addresses diversity at the geometric level. The central insight is as follows: a hierarchical partitioning of documents is geometrically encoded by the sequence of split directions, namely the eigenvectors of local scatter matrices used to descend from the root to the leaves. The subspace spanned by these directions is precisely the part of embedding space that the taxonomy "used up". Before constructing the next taxonomy, ROOT projects all document embeddings onto the null space of this used subspace, preventing subsequent taxonomies from rediscovering the same geometric structure. Diversity is thus not imposed through a post-hoc penalty or heuristic, but maintained as a structural invariant during generation.

Building on this formulation, we make three main contributions.[1] First, We propose a geometric diversity mechanism (ROOT) for hierarchical tax-

[1] Code and data will be released upon acceptance.

onomy generation based on iterative null-space projection of document embeddings, providing a formal guarantee that successive taxonomies capture orthogonal semantic dimensions. Our proposed algorithm relies mainly on document embeddings to generate diverse taxonomy trees and uses a LLM only to verbalize the node-level document clusters.

Second, we propose an evaluation suite that measures the diverse taxonomy outputs along different axes: semantic diversity (whether different trees partition papers differently), taxonomy semantics (whether each tree's clusters are internally coherent and externally distinct), structural diversity (whether trees vary in shape, branching, and depth rather than being rigid copies) along with LLM-as-a-judge and human evaluation for external validation of paper placement and holistic quality.

Third, we show empirically that ROOT creates substantially diverse taxonomy trees when compared with baselines while maintaining taxonomy semantics, structure and label quality.

2 Related Work

For decades, Natural Language Processing (NLP) has approached taxonomy learning by leveraging semantic relationships within text (Hearst, 1992; Pantel and Pennacchiotti, 2006; Suchanek et al., 2006; Ponzetto and Strube, 2011; Rios-Alvarado et al., 2013; Dietz et al., 2012; Liu et al., 2012; Diederich and Balke, 2007; Wang et al., 2010; Kang et al., 2016; Kozareva and Hovy, 2010; Velardi et al., 2013).

Automated taxonomy generation has garnered recent attention with the advent of LLMs. Models like TAXOALIGN (Lahiri et al., 2025) and (Vu et al., 2025) have harnessed the power of LLMs for taxonomy generation while models like HiGTL (Hu et al., 2025) have advanced the field by introducing graph- and clustering-based techniques. TaxoAdapt (Kargupta et al., 2025) dynamically constructs an LLM-enhanced, corpus-specific taxonomy using classification-based expansion signals. (Zhu et al., 2025) proposed a context-aware hierarchical taxonomy generation framework that integrates LLM-guided multi-aspect encoding with dynamic clustering. However, while the majority of these methods produce only a single-perspective taxonomy tree, our proposed algorithm ROOT distinguishes itself by generating diverse taxonomies that reflect multiple perspectives.

3 Methodology

We propose a structurally rigorous, unsupervised framework for generating diverse, multi-perspective hierarchical classifications over large document collections. Our approach, **Recursive Orthogonalization Of Taxonomies (ROOT)**, shifts the burden of diversity enforcement from brittle autoregressive prompting to the continuous geometry of the document embedding space.

3.1 Formal Problem Statement

Let $\mathcal{X} = \{x_1, \ldots, x_N\}$ be a collection of N scientific documents. Our objective is to generate a set of K distinct, human-interpretable taxonomies, $\mathbb{T} = \{T_1, \ldots, T_K\}$. Each $T_k \in \mathbb{T}$ is a hierarchical tree, where internal nodes v represent latent classification intents and leaf nodes contain documents.

A valid generation must satisfy four constraints. First, **Intra-Taxonomy Coherence** dictates that documents sharing a subtree must exhibit high semantic similarity relative to that branch's specific intent. Second, **Exhaustive Coverage** requires that every document in the collection is assigned to exactly one leaf node, ensuring the entire corpus is organized without any omissions or duplicates. Third, **Inter-Taxonomy Diversity** mandates that for any pair $T_a, T_b \in \mathbb{T}$ $(a \neq b)$, the structural routing decisions and underlying continuous intents must be mathematically orthogonal, yielding genuinely multi-faceted perspectives. Finally, **Semantic Interpretability** ensures every internal node v is automatically assigned a discrete, human-readable string L_v articulating the distinguishing principle separating its child branches.

3.2 ROOT Framework

To satisfy the constraints of exhaustive coverage and inter-taxonomy diversity, ROOT iteratively generates tree structures while mathematically eradicating the continuous spatial history of prior classifications. Let $f_\theta : \mathcal{X} \to \mathbb{R}^d$ be a pre-trained dense document encoder. For each document $x_i \in \mathcal{X}$, we extract the raw dense representation $e_i = f_\theta(x_i)$. To establish a proper geometric origin for variance-driven clustering, we compute the global mean centroid $\mu = \frac{1}{N}\sum_{j=1}^{N} e_j$ and initialize the mean-centered base embeddings:

$$h_i^{(1)} = e_i - \mu \tag{1}$$

These centered vectors are stacked to form the initial global embedding matrix $H^{(1)} \in \mathbb{R}^{N \times d}$. For

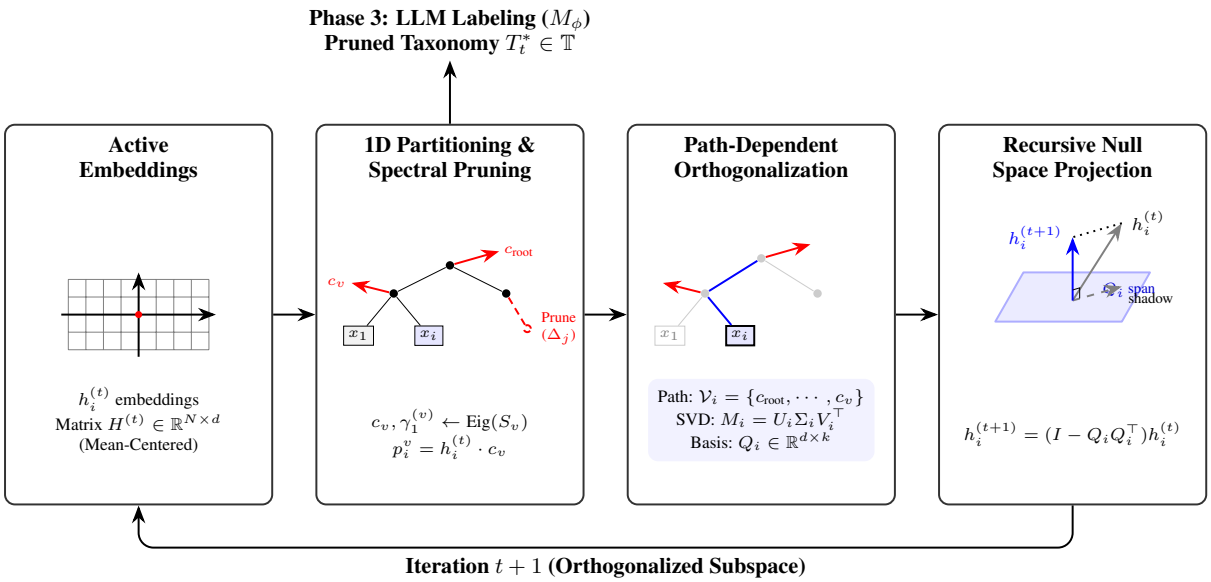

Figure 1: Architectural overview of ROOT. The framework iteratively alternates between generating a structural tree via 1D partitioning (Phase 1) and dynamic spectral pruning (Phase 2). The pruned nodes are contrastively labeled by an LLM (Phase 3). Finally, by tracing a document's path and extracting an exact orthonormal basis via SVD, we project its embedding into the null space of its specific classification history (Phase 4), mathematically guaranteeing diversity in the subsequent iteration.

each iteration t from 1 to K, ROOT executes the following sequential phases (summarized in Algorithm 1).

Phase 1: Single-Intent 1D Partitioning To construct the initial tree T_t, we employ an adaptive, variance-driven divisive clustering strategy. For any internal node v containing a subset of documents S_v, let $H_v^{(t)} \in \mathbb{R}^{|S_v| \times d}$ denote their corresponding embeddings. We compute the unnormalized within-cluster scatter matrix $S_v^{scatter} = \sum_{i \in S_v} (h_i^{(t)} - \mu_v)(h_i^{(t)} - \mu_v)^\top$, where μ_v is the local mean centroid of the node. By applying Eigenvalue Decomposition, we extract the dominant eigenpair $(c_v, \gamma_1^{(v)})$.

The eigenvector $c_v \in \mathbb{R}^d$ points in the direction of maximum variance, mathematically defining the latent classification intent. The intuition here is rooted in classical taxonomy design: at any given hierarchical level, a valid taxonomy divides objects based on a singular, dominant distinguishing feature rather than a composite of confounding variables. Standard high-dimensional clustering entangles multiple axes of variance, often yielding semantically impure and uninterpretable splits. To preserve semantic purity, we isolate this singular dominant feature by projecting the embeddings strictly onto the 1D principal axis, yielding scalar projections $p_i^v = h_i^{(t)} \cdot c_v$ for all $i \in S_v$.

Crucially, this 1D projection unlocks exact global optimality for the clustering step. While standard multi-dimensional k-means is NP-hard (Aloise et al., 2009) and relies on heuristic approximations that fall into local minima (e.g., Lloyd's algorithm), partitioning points in a single dimension admits a polynomial-time Dynamic Programming (DP) solution that mathematically guarantees the globally optimal partition minimizing the within-cluster sum of squares exactly (Grønlund et al., 2017).

To determine the optimal branching factor k_{opt} for a node v, we evaluate a candidate range $k \in [2, K_{\max}]$. For each k, we apply the exact 1D DP algorithm to the scalar projections p_i^v. We select the k_{opt} that maximizes the Silhouette score, thereby dynamically calibrating the tree's width to the data's natural semantic grouping. This recursive partitioning halts at a computational maximum depth D_{max}.

Phase 2: Post-Hoc Spectral Pruning To dynamically contour the taxonomy depth, we rely on the spectral topology of the unpruned tree T_t. This contouring is mathematically grounded in the following property of our variance-driven splits:

Claim 1 (Monotonicity of Spectral Variance): *Let v be a parent node partitioned into a set of child nodes C. Because the unnormalized scatter*

Algorithm 1: ROOT: Recursive Orthogonalization Of Taxonomies

Input: $\mathcal{X}=\{x_i\}_{i=1}^{N}$, encoder f_θ, LLM M_ϕ, target K, limits D_{min}, D_{max}
Output: Diverse taxonomies $\mathbb{T}=\{T_1,\ldots,T_K\}$ with labeled nodes

$h_i^{(1)} \leftarrow f_\theta(x_i)\ \forall i$; form mean-centered $H^{(1)}$; $\mathbb{T} \leftarrow \emptyset$;
for $t = 1$ **to** K **do**
  /* Phase 1: Single-Intent 1D Partitioning */
  $Q \leftarrow [\mathrm{root}(\mathcal{X})]$;
  while $Q \neq \emptyset$ **do**
    Pop node v at depth d_v from Q;
    if $d_v < D_{max}$ *and* $|v| > \delta_{\min}$ **then**
      $c_v,\ \gamma_1^{(v)} \leftarrow \mathrm{DomEigPair}(\mathrm{Scatter}(H_v^{(t)}))$;
      $p_i^v \leftarrow h_i^{(t)} \cdot c_v\ \forall i \in S_v$; // Project to 1D
      $k^* \leftarrow \arg\max_k \mathrm{Silhouette}(p_i^v, k)$;
      Partition S_v into k^* children; push to Q;
    end
  end
  /* Phase 2: Post-Hoc Spectral Pruning */
  foreach *path* $(v_0,\ldots,v_L)$ *in* T_t *with* $L \geq D_{min}$ **do**
    $\Delta_j \leftarrow (\gamma_1^{(v_{j-1})} - \gamma_1^{(v_j)})/\gamma_1^{(v_{j-1})}$;
    $j^* \leftarrow \arg\max_{j \geq D_{min}} \Delta_j$;
    Make v_{j^*} a terminal leaf in T_t^*;
  end
  /* Phase 3: Contrastive Generative Labeling */
  foreach *internal node* $v \in T_t^*$ **do**
    $E_v \leftarrow \{Q_0, Q_{25}, Q_{50}, Q_{75}, Q_{100},\}$ w.r.t. p_i^v;
    $L_v \leftarrow M_\phi$("Contrast E_v"); label $v \leftarrow L_v$;
  end
  $\mathbb{T} \leftarrow \mathbb{T} \cup \{T_t^*\}$;
  /* Phase 4: Path Orthogonalization & Null-Space Projection */
  foreach $i \in \{1,\ldots,N\}$ **do**
    $Q_i \leftarrow \mathrm{OrthoBasis}(\{c_v \mid v \in P(i)\})$;
    $h_i^{(t+1)} \leftarrow (I - Q_i Q_i^\top)\, h_i^{(t)}$; // Erase path history
  end
  $H^{(t+1)} \leftarrow \mathrm{MeanCenter}(H^{(t+1)})$;
end
return $\mathbb{T}$;

matrix S_v measures total absolute variance and is additive across disjoint subsets, the total scatter of the parent strictly bounds the scatter of any individual child $c \in C$. Consequently, the maximum variance along any principal axis must decrease after a partition, mathematically guaranteeing that the dominant eigenvalue strictly decreases monotonically along any downward path in the tree: $\gamma_1^{(c)} < \gamma_1^{(v)}$.

Motivation for Spectral Truncation: Because γ_1 strictly tracks the total unresolved semantic variance within a node, its monotonic decrease provides an optimal stopping metric. A massive drop between a parent and child ($\gamma_1^{(v_{j-1})} \gg \gamma_1^{(v_j)}$) indicates that the split resolved a major semantic ambiguity. Conversely, a negligible drop implies the partition artificially forced a split within an already homogeneous space.

Therefore, we evaluate the structural resolution by tracking the relative spectral drop along every unique path $P = (v_0, \ldots, v_L)$ for each parent-child step j:

$$\Delta_j = \frac{\gamma_1^{(v_{j-1})} - \gamma_1^{(v_j)}}{\gamma_1^{(v_{j-1})}} \tag{2}$$

To prevent premature truncation of foundational taxonomy levels, we enforce a minimum depth limit D_{min}. We isolate the optimal semantic stopping point—the structural "sweet spot"—by identifying the split that resolved the highest relative variance: $j^* = \arg\max_{j \geq D_{min}} \Delta_j$. We prune the branch strictly at v_{j^*}, yielding the structurally optimal tree T_t^*.

Phase 3: Contrastive Generative Labeling While structural decisions are driven by continuous geometry, human interpretability requires discrete symbolic labels. Generative language models struggle with full-corpus taxonomy induction but excel at highly constrained contrastive reasoning. For each internal node $v \in T_t^*$, we leverage the 1D projections to isolate a representative quintet of documents located exactly at the 0th, 25th, 50th, 75th, and 100th percentiles along the intent axis c_v. This uniformly distributed subset, $E_v = \{x_{0\%}, x_{25\%}, x_{50\%}, x_{75\%}, x_{100\%}\}$, captures the complete gradient of semantic variance across the split. We pass E_v to a Language Model (M_ϕ) prompted to analyze this continuous spectrum and articulate the singular distinguishing principle. The output string L_v serves as the node's discrete label.

Phase 4: Path-Dependent Orthogonalization & Null Space Projection This phase is the theoretical engine of ROOT, mathematically guaranteeing inter-taxonomy diversity. Instead of relying on generative heuristics to avoid redundancy, we explicitly enforce structural novelty by geometrically erasing the variance axes that governed prior classifications.

Because a document only "experiences" the specific semantic splits along its individual trajectory from root to leaf, erasing global tree intents would indiscriminately destroy orthogonal information. Thus, for each document i, we trace its unique path $P(i)$ in T_t^* and accumulate strictly the local principal components it directly encountered: $\mathcal{V}_i = \{c_v \mid v \in P(i)\}$.

Since these local intent vectors are computed at varying depths on shrinking data subsets, they are rarely perfectly orthogonal. Sequential subtraction would therefore cause mathematical interference

and distortion. To resolve this, we compute the Singular Value Decomposition (SVD) of the stacked vectors in $\mathcal{V}_i$. Extracting the non-zero columns of the left singular vectors yields Q_i, an exact orthonormal basis defining the document's historical classification subspace.

Finally, we project the document's embedding into the null space of this historical subspace:

$$h_i^{(t+1)} = \left(I - Q_i Q_i^\top\right) h_i^{(t)} \tag{3}$$

The rejection matrix $(I - Q_i Q_i^\top)$ acts as a geometric filter, stripping the embedding of any variance that aligns with its past intents. By applying this path-dependent erasure to all documents and re-centering the global matrix $H^{(t+1)}$, the residual representations contain exclusively unexploited semantic features. This mathematically guarantees that the next iteration evaluates the corpus from a strictly orthogonal, fundamentally novel perspective.

3.3 Theoretical Analysis of Diversity

The primary contribution of the ROOT framework is its ability to move beyond heuristic prompting to mathematically guarantee inter-taxonomy diversity. We formalize this guarantee through two geometric properties of the recursive null space projection.

Claim 2 (Strict Orthogonality of Sequential Intents): *Let $Q_i \in \mathbb{R}^{d \times k}$ be the exact orthonormal basis representing the historical classification intents for document i up to iteration t. The projected embedding $h_i^{(t+1)}$ is perfectly orthogonal to any intent contained in the column space of Q_i.*

Proof: The updated embedding is defined as the residual of the projection: $h_i^{(t+1)} = (I - Q_i Q_i^\top) h_i^{(t)}$. To determine its alignment with the historical subspace, we project it back onto the basis Q_i:

$$\begin{aligned} Q_i^\top h_i^{(t+1)} &= Q_i^\top \left(I - Q_i Q_i^\top\right) h_i^{(t)} \\ &= \left(Q_i^\top - Q_i^\top Q_i Q_i^\top\right) h_i^{(t)} \end{aligned}$$

Because Q_i is an orthonormal basis by construction via SVD, $Q_i^\top Q_i = I$. Therefore:

$$= \left(Q_i^\top - I Q_i^\top\right) h_i^{(t)} = \mathbf{0}$$

This zero-vector result mathematically guarantees that the 1D partitioning algorithm in iteration $t+1$ cannot reuse any semantic features that previously dictated the structural placement of document i. The subsequent taxonomy is strictly forced to discover an entirely novel perspective.

Claim 3 (Monotonic Depletion of Semantic Variance): *The Euclidean norm (magnitude) of any document's embedding strictly non-increases across iterations: $\|h_i^{(t+1)}\|_2 \leq \|h_i^{(t)}\|_2$.*

Proof: The null space projection decomposes the original embedding $h_i^{(t)}$ into two perfectly orthogonal vectors: the historical "shadow" $Q_i Q_i^\top h_i^{(t)}$ and the residual $h_i^{(t+1)}$. We can reconstruct the original vector as their sum:

$$h_i^{(t)} = Q_i Q_i^\top h_i^{(t)} + h_i^{(t+1)} \tag{4}$$

Because these two components are strictly orthogonal, we apply the Pythagorean theorem in $\mathbb{R}^d$:

$$\|h_i^{(t)}\|_2^2 = \|Q_i Q_i^\top h_i^{(t)}\|_2^2 + \|h_i^{(t+1)}\|_2^2 \tag{5}$$

Since the squared norm of any real vector is non-negative ($\|Q_i Q_i^\top h_i^{(t)}\|_2^2 \geq 0$), it must hold that $\|h_i^{(t+1)}\|_2^2 \leq \|h_i^{(t)}\|_2^2$.

This bounded energy reduction ensures that the generation process does not artificially inflate the vector space. Instead, ROOT systematically strips away the specific semantic features—the underlying meanings and classification intents—that have already been utilized in previous taxonomies. The document vectors physically shrink across iterations because they contain progressively less unexplained information. This mathematically forces the generation process to naturally halt when all distinguishing perspectives have been exhausted and there is no meaningful variation left to cluster.

4 Baselines

We compare our proposed method with three baselines. Detailed descriptions of each baseline are provided in the Appendix A.

Repeated Hierarchical Clustering: Generates diverse taxonomies via repeated random restarts of agglomerative clustering over noise-perturbed sentence-transformer embeddings, with an LLM labeling the resulting nodes.

Aspect-based GMM: Adapted from (Zhu et al., 2025), this method prompts an LLM to produce K distinct semantic aspects of the corpus, then builds a separate GMM-based recursive taxonomy tree for each aspect's embedding space.

LLM Prompting: Delegates both organization and labeling entirely to an LLM.

Method	TaxoBench-CS			CS-TaxoBench		
	DW-Jaccard	wARI	LabelUniq	DW-Jaccard	wARI	LabelUniq
ROOT	0.1613 ±0.046	**0.0215 ±0.031**	0.1048 ±0.027	0.1813 ±0.062	**0.0371 ±0.050**	0.1043 ±0.030
LLM	0.2798 ±0.111	0.0355 ±0.050	0.1217 ±0.036	0.2492 ±0.114	0.0386 ±0.066	0.1148 ±0.034
Agglom. Clustering	**0.1280 ±0.078**	0.0942 ±0.091	0.0627 ±0.020	**0.1768 ±0.147**	0.1575 ±0.159	0.0672 ±0.023
Aspect-based GMM	0.3416 ±0.085	0.1927 ±0.146	0.0787 ±0.025	0.3510 ±0.096	0.2169 ±0.141	0.0818 ±0.033

Table 1: Diversity Evaluation Results

5 Experimental Setup

We use two published datasets on taxonomy generation CS-TaxoBench (Lahiri et al., 2025) and TaxoBench-CS (Zhu et al., 2025). These datasets were mainly constructed for single taxonomy generation, but we use the papers collected by them for our task of diverse taxonomy generation.

We use Llama-3.1-8B-Instruct in all the experiments. We use the BGE model for all the embedding experiments and we use the SPECTER model (Cohan et al., 2020) for all the encodings during the testing phase. Details about the datasets and experimental settings are in Appendix B.

6 Diversity Evaluation

We measure whether the generated K taxonomy trees discover genuinely different ways to organize the same set of papers. To do so, we compare pairs of trees by slicing them at each depth into flat clusterings and measuring agreement at both coarse and fine scales. Detailed formulations and derivations are provided in the Appendix.

Both metrics use a log-decaying depth weight, giving higher importance to top-level splits, as coarse-grained structure drives finer-grained clustering decisions downstream.

6.1 Depth-Weighted Jaccard

At each depth d, we compute the Jaccard index between the co-clustering sets of two trees — the sets of paper pairs grouped together at that level. These per-depth scores are then aggregated using the log-decay weights. A low Depth-Weighted Jaccard indicates that the trees genuinely disagree about which papers should be grouped together, especially at the coarsest levels.

6.2 Depth-Weighted Adjusted Rand Index

At each depth d, we compute the Adjusted Rand Index (ARI) (Warrens and Van Der Hoef, 2022) between the flat clusterings of two trees, correcting for chance agreements, respectively. These scores are aggregated with the same log-decay weights. A wARI near 0 indicates that the trees' partitions are statistically independent, confirming that each tree discovers a structurally distinct organization.

6.3 Label Uniqueness (LabelUniq)

For each pair of trees, we compute a symmetric nearest-neighbour overlap between their label sets using cosine similarity of label embeddings, then subtract the mean overlap from unity. A high Label Uniqueness score indicates that each tree employs a genuinely distinct vocabulary of topic labels.

6.4 Taxonomy Diversity: Results

Table 1 reports diversity evaluation results. ROOT achieves near-zero wARI on both datasets with low standard deviation, indicating that its trees are close to statistically independent and that this behavior is stable across runs — outperforming both LLM and Aspect-based GMM baselines on this metric. While Agglomerative Clustering attains the lowest DW-Jaccard, its substantially higher wARI suggests its apparent diversity stems from low-level mismatches rather than genuinely independent hierarchical organization. Aspect-based GMM produces the least diverse trees overall, with the highest DW-Jaccard and wARI across both datasets. The LLM baseline achieves the highest Label Uniqueness due to explicit prompting for distinct labels, though ROOT ranks second among all methods on this metric. Overall, ROOT yields the most structurally principled diversity, with chance-corrected disagreement between trees that is consistently stronger than all baselines.

7 Taxonomy Semantics

The metrics described here judge the semantic quality of each individual tree in isolation, without comparing trees to each other. Detailed formulations and derivations are provided in the Appendix D.

Method	TaxoBench-CS			CS-TaxoBench		
	Cohesion	Separation	Node Precision	Cohesion	Separation	Node Precision
ROOT	0.8118 ±0.028	0.0870 ±0.017	0.0976 ±0.048	**0.7718 ±0.036**	0.1168 ±0.023	**0.0933 ±0.047**
LLM	0.8008 ±0.031	0.0455 ±0.019	0.0760 ±0.049	0.7547 ±0.050	0.0691 ±0.038	0.0788 ±0.051
Agglom. Clustering	**0.8161 ±0.026**	**0.0902 ±0.035**	**0.1241 ±0.044**	0.7562 ±0.037	**0.1397 ±0.057**	0.0562 ±0.039
Aspect-based GMM	0.8112 ±0.028	0.0408 ±0.013	0.0809 ±0.049	0.7707 ±0.048	0.0602 ±0.022	0.0691 ±0.039

Table 2: Taxonomy Semantic quality metrics across methods and datasets.

7.1 Within-Cluster Cohesion

Cohesion measures how semantically homogeneous the documents within each subtree are, computed as the size-weighted mean pairwise cosine similarity across all nodes. Size-weighting ensures that large, semantically incoherent subtrees are penalized more heavily than small, trivially pure leaf clusters.

7.2 Inter-Cluster Separation

Separation measures whether sibling nodes at each branching point represent genuinely distinct sub-topics, computed as the size-weighted mean pairwise cosine distance between sibling subtree centroids. Higher separation indicates that the algorithm's splits carve out semantically well-differentiated regions of the embedding space.

7.3 Node Precision

Node Precision evaluates whether each paper genuinely belongs under its assigned label using a three-step NLI-based protocol — a recall check, an individual relevance check, and a leave-one-out check — to distinguish truly irrelevant papers from merely marginal ones. A paper is deemed irrelevant only if it does not individually support the node label yet the remaining cluster still represents it, following the citation precision metric of (Gao et al., 2023).

7.4 Taxonomy Semantics: Results

Table 2 reports structural quality results. All four methods achieve high cohesion scores (0.75–0.82), indicating that documents within each subtree are semantically similar regardless of construction strategy. Agglomerative Clustering leads marginally on TaxoBench-CS while ROOT tops CS-TaxoBench; the LLM baseline consistently trails, suggesting free-form groupings introduce slightly more within-subtree semantic noise. For separation, ROOT and Agglomerative Clustering substantially outperform LLM and Aspect-based GMM on both datasets, with Agglomerative Clustering ranking highest overall — expected, as clustering algorithms directly optimise for inter-cluster distance. Aspect-based GMM consistently achieves the lowest separation, indicating weak semantic distinction between sibling subtrees at each branching point. For Node Precision, Agglomerative Clustering leads on TaxoBench-CS (0.1241) but collapses to last place on CS-TaxoBench (0.0562), suggesting it overfits to one corpus and does not generalise well. ROOT, by contrast, ranks first or second on both datasets (0.0976 / 0.0933), demonstrating more robust label assignment across different domain distributions.

8 Structural Evaluation

8.1 Normalised Tree Edit Distance

This metric captures topological differences between trees by stripping each tree to its bare parent-child skeleton and computing the Tree Edit Distance (TED) — the minimum number of node insertions, deletions, or matches needed to transform one tree into another. Raw TED is normalised by the sum of node counts in both trees to account for size differences, and the final score is the mean normalised TED across all tree pairs.

Method	TaxoBench-CS	CS-TaxoBench
ROOT	**0.3061 ±0.068**	**0.2890 ±0.067**
LLM	0.2851 ±0.133	0.2757 ±0.137
Agglom. Clustering	0.1531 ±0.034	0.1509 ±0.039
Aspect-based GMM	0.0440 ±0.043	0.0536 ±0.042

Table 3: Taxonomy Structure (TED) results

8.2 Taxonomy Structure: Results

Table 3 reports structural diversity results. ROOT produces the highest topological variety across runs on both datasets (0.3061 / 0.2890), with LLM following closely (0.2851 / 0.2757). Agglomerative Clustering drops substantially (0.1531 / 0.1509),

Metric	TaxoBench-CS		CS-TaxoBench	
	ROOT	ROOT(noOrtho)	ROOT	ROOT(noOrtho)
DW-Jaccard ↓	$\mathbf{0.1613 \pm 0.046}$	1.0000 ± 0.000	$\mathbf{0.1813 \pm 0.062}$	0.9963 ± 0.061
wARI ↓	$\mathbf{0.0215 \pm 0.031}$	1.0000 ± 0.000	$\mathbf{0.0371 \pm 0.050}$	0.9963 ± 0.061
LabelUniq ↑	$\mathbf{0.1048 \pm 0.027}$	0.0000 ± 0.000	$\mathbf{0.1043 \pm 0.030}$	0.0000 ± 0.000
TED ↑	$\mathbf{0.3061 \pm 0.068}$	0.0000 ± 0.000	$\mathbf{0.2890 \pm 0.067}$	0.0000 ± 0.000
Cohesion ↑	0.8118 ± 0.028	$\mathbf{0.8240 \pm 0.026}$	0.7718 ± 0.036	$\mathbf{0.7884 \pm 0.035}$
Separation ↑	$\mathbf{0.0870 \pm 0.017}$	0.0848 ± 0.022	$\mathbf{0.1168 \pm 0.023}$	0.1156 ± 0.028
Node Precision ↑	$\mathbf{0.0976 \pm 0.048}$	0.0918 ± 0.060	$\mathbf{0.0933 \pm 0.047}$	0.0888 ± 0.061

Table 4: Ablation results for ROOT and ROOT(noOrtho) on TaxoBench-CS and CS-TaxoBench datasets.

while Aspect-based GMM is nearly deterministic — its TED is close to zero on both datasets (0.0440 / 0.0536), indicating near-identical tree shapes across runs.

9 LLM-Based Qualitative Evaluation

The LLM judge is presented with the complete hierarchical outline of all K trees and asked to rate the system on three dimensions using integer scores from 1 to 10. We ran this evaluation on the TaxoBench-CS dataset using GPT-4o.

The judge evaluates three criteria: *Diversity:* Rates how different the K trees are from each other in terms of both structure and groupings; *Structure:* Rates how evenly papers are distributed across leaf clusters; and, *Label Quality:* Rates the informativeness and mutual exclusivity of the node labels.

Method	Diversity	Structure	Label Quality
ROOT	**7.8205 ±0.384**	7.0769 ±0.432	6.8013 ±0.571
LLM (baseline)	7.8012 ±0.355	7.1923 ±0.852	6.8077 ±0.885
Aspect-GMM	7.0192 ±0.674	7.2244 ±0.944	6.3013 ±0.887
Agglom. Clust.	7.4258 ±0.494	**7.3269 ±0.720**	**6.9038 ±0.664**

Table 5: Results for LLM-as-a-judge

Table 5 reports the results. ROOT leads on Diversity (7.82) but is marginally weaker on Structure (7.08), while Agglomerative Clustering tops both Structure (7.33) and Label Quality (6.90) at the cost of lower diversity.

10 Human Evaluation

Method	Diversity	Structure	Label Quality
ROOT	**6**	5.13	6.07
LLM (baseline)	5.27	5.33	**6.1**
Aspect-GMM	4.27	**5.5**	5.57
Agglom. Clust.	4.67	5.2	5.2

Table 6: Results of Human Evaluation

We gave 10 taxonomy tree sets to reseachers in AI/ML, 5 from each dataset, with the guidelines as in Appendix E. The human evaluation confirms the same broad pattern as the LLM-as-a-judge, but with lower absolute scores. The label quality results show a subtle difference between automatic and human judgment. The LLM judge prefers Agglomerative Clustering for labels, while humans slightly prefer the LLM baseline and ROOT. That means label quality is somewhat subjective, and the method that looks best to the model judge is not always the one humans prefer.

11 Ablation Study

We run an ablation study to explore the effect of ROOT without the Orthogonalisation step. We run Phases 1-3 for tree construction K times on the same (unchanged) embedding matrix. Table 4 reports ablation results. Removing orthogonalisation (ROOT(noOrtho)) completely collapses taxonomic diversity — TED and LabelUniq drop to exactly 0, while DW-Jaccard and wARI hit their worst possible values ($\approx$1.0) on both datasets, confirming that orthogonalisation is the source of diversity rather than algorithmic stochasticity. ROOT(noOrtho) marginally outperforms ROOT on cohesion, while ROOT leads on separation and node precision, though by narrow margins.

12 Conclusion

We presented ROOT, a geometric framework for generating diverse taxonomy trees via iterative null-space projection of document embeddings. We also proposed a comprehensive evaluation pipeline covering semantic diversity, taxonomy semantics, structural diversity, LLM-as-a-judge, and human evaluation. Empirical results show that ROOT consistently produces more diverse taxonomies than all baselines while maintaining competitive semantics and label quality.

Limitations

ROOT relies on dense document embeddings, so its diversity guarantees are bounded by the geometry of the embedding space — if the embedding model fails to capture certain semantic distinctions, null-space projection cannot recover them. We use LLMs for coding although with thorough scrutiny of every code. Furthermore, while ROOT uses an LLM only for node verbalization, the quality of node labels remains sensitive to the choice of LLM and prompt design. Finally, our evaluation pipeline, though multi-faceted, relies in part on LLM judges and human annotators, both of which introduce their own biases and variability.

References

Daniel Aloise, Amit Deshpande, Pierre Hansen, and Preyas Popat. 2009. NP-hardness of Euclidean sum-of-squares clustering. *Machine Learning*, 75(2):245–248.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. SPECTER: Document-level Representation Learning using Citation-informed Transformers. In *ACL*.

Jörg Diederich and Wolf-Tilo Balke. 2007. The semantic growbag algorithm: automatically deriving categorization systems. In *Proceedings of the 11th European Conference on Research and Advanced Technology for Digital Libraries*, ECDL'07, page 1–13, Berlin, Heidelberg. Springer-Verlag.

Emmanuelle-Anna Dietz, Damir Vandic, and Flavius Frasincar. 2012. Taxolearn: A semantic approach to domain taxonomy learning. In *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 58–65.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

Maayan Geffet and Ido Dagan. 2005. The distributional inclusion hypotheses and lexical entailment. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 107–114, Ann Arbor, Michigan. Association for Computational Linguistics.

Allan Grønlund, Kasper Green Larsen, Alexander Mathiasen, Jesper Sindahl Nielsen, Stefan Schneider, and Mingzhou Song. 2017. Fast exact k-means, k-medians and bregman divergence clustering in 1d. *arXiv preprint arXiv:1701.07204*.

Mark A. Hanson, Pablo Gómez Barreiro, Paolo Crosetto, and Dan Brockington. 2024. The strain on scientific publishing. *Quantitative Science Studies*, 5(4):823–843.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*.

Yuntong Hu, Zhuofeng Li, Zheng Zhang, Chen Ling, Raasikh Kanjiani, Boxin Zhao, and Liang Zhao. 2025. Taxonomy tree generation from citation graph. *Preprint*, arXiv:2410.03761.

Wen Hua, Zhongyuan Wang, Haixun Wang, Kai Zheng, and Xiaofang Zhou. 2017. Understand short texts by harvesting and analyzing semantic knowledge. *IEEE Transactions on Knowledge and Data Engineering*, 29(3):499–512.

Yong-Bin Kang, Pari Delir Haghigh, and Frada Burstein. 2016. TaxoFinder: A Graph-Based Approach for Taxonomy Learning . *IEEE Transactions on Knowledge & Data Engineering*, 28(02):524–536.

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 (Volume 1: Long Papers)*, pages 29834–29850, Vienna, Austria. Association for Computational Linguistics.

Zornitsa Kozareva and Eduard Hovy. 2010. A semi-supervised method to learn and construct taxonomies using the web. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1110–1118, Cambridge, MA. Association for Computational Linguistics.

Avishek Lahiri, Yufang Hou, and Debarshi Kumar Sanyal. 2025. TaxoAlign: Scholarly taxonomy generation using language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30203–30223, Suzhou, China. Association for Computational Linguistics.

Xueqing Liu, Yangqiu Song, Shixia Liu, and Haixun Wang. 2012. Automatic taxonomy construction from keywords. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, page 1433–1441, New York, NY, USA. Association for Computing Machinery.

Patrick Pantel and Marco Pennacchiotti. 2006. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 113–120, Sydney, Australia. Association for Computational Linguistics.

Simone Paolo Ponzetto and Michael Strube. 2011. Taxonomy induction based on a collaboratively built knowledge repository. *Artificial Intelligence*, 175(9):1737–1756.

Niklas Reisz, Vito D P Servedio, Vittorio Loreto, William Schueller, Márcia R Ferreira, and Stefan Thurner. 2022. Loss of sustainability in scientific work. *New Journal of Physics*, 24(5):053041.

Ana B. Rios-Alvarado, Ivan Lopez-Arevalo, and Victor J. Sosa-Sosa. 2013. Learning concept hierarchies from textual resources for ontologies construction. *Expert Systems with Applications*, 40(15):5907–5915.

Fabian M. Suchanek, Georgiana Ifrim, and Gerhard Weikum. 2006. Combining linguistic and statistical analysis to extract relations from web documents. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, page 712–717, New York, NY, USA. Association for Computing Machinery.

Paola Velardi, Stefano Faralli, and Roberto Navigli. 2013. OntoLearn reloaded: A graph-based algorithm for taxonomy induction. *Computational Linguistics*, 39(3):665–707.

Binh Vu, Rashmi Govindraju Naik, Bao Khanh Nguyen, Sina Mehraeen, and Matthias Hemmje. 2025. Automated taxonomy construction using large language models: A comparative study of fine-tuning and prompt engineering. *Eng*, 6(11).

Chengyu Wang, Xiaofeng He, and Aoying Zhou. 2017. A short survey on taxonomy learning from text corpora: Issues, resources and recent advances. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1190–1203, Copenhagen, Denmark. Association for Computational Linguistics.

Wei Wang, Payam Mamaani Barnaghi, and Andrzej Bargiela. 2010. Probabilistic topic models for learning terminological ontologies. *IEEE Transactions on Knowledge and Data Engineering*, 22(7):1028–1040.

Matthijs J Warrens and Hanneke Van Der Hoef. 2022. Understanding the adjusted rand index and other partition comparison indices based on counting object pairs. *Journal of Classification*, 39(3):487–509.

Shuo Yang, Lei Zou, Zhongyuan Wang, Jun Yan, and Ji-Rong Wen. 2017. Efficiently answering technical questions — a knowledge graph approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Yuchen Zhang, Amr Ahmed, Vanja Josifovski, and Alexander Smola. 2014. Taxonomy discovery for personalized recommendation. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, page 243–252, New York, NY, USA. Association for Computing Machinery.

Kun Zhu, Lizi Liao, Yuxuan Gu, Lei Huang, Xiaocheng Feng, and Bing Qin. 2025. Context-aware hierarchical taxonomy generation for scientific papers via LLM-guided multi-aspect clustering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15616–15634, Suzhou, China. Association for Computational Linguistics.

A Baselines

We compare our proposed method with three different baselines, each of which is designed to generate different taxonomy trees from a given corpus of papers.

Repeated Hierarchical Clustering: This baseline generates diversity through repeated random restarts of hierarchical clustering. Papers are encoded once using a sentence-transformer. Then, for each of K runs, two sources of randomness are injected: (a) small isotropic Gaussian noise is added to the embeddings and re-normalized, and (b) a distinct random seed is used. We apply ward-linkage agglomerative clustering to each perturbed embedding matrix and cut at two levels (coarse and fine) to produce a fixed three-level taxonomy (root → categories → subcategories → papers). Each run's dendrogram is cut to yield a preset number of top-level categories and subcategories. An LLM then labels every internal and leaf node, with prompts that enforce specificity and penalize generic labels.

Aspect-based GMM: We adapt this baseline from the framework proposed by (Zhu et al., 2025). The pipeline generates diverse taxonomies by viewing the same corpus through different semantic "lenses." First, an LLM (Llama-3.1-8B-Instruct) is prompted with the survey topic and a sample of papers to produce K distinct semantic aspects — short noun phrases capturing dimensions of variation (e.g., "compression technique," "evaluation methodology"). For each aspect, the LLM writes a one-sentence, aspect-specific summary of every paper. These summaries are then independently embedded using a sentence-transformer (BGE). Within each aspect's embedding space, a recursive tree is built top-down: at every node, Gaussian Mixture Models are fitted for varying numbers of components and the best is selected via BIC, the documents are split according to the GMM assignments, and the process recurses until a depth or minimum-leaf-size constraint is reached. Finally, the LLM labels each internal cluster node. Each aspect yields one complete taxonomy, so K aspects produce K structurally distinct taxonomies — di-

versity arises from the fact that each taxonomy operates on a different semantic projection of the papers.

LLM Prompting This baseline delegates both the organizational logic and the labeling entirely to the LLM, with no embedding or clustering step. For the first taxonomy, the model is prompted with the full list of paper titles and abstract snippets along with strict structural constraints (exact tree depth, branching factor limits, minimum leaf size, exhaustive paper assignment) and asked to output a taxonomy tree. For each subsequent taxonomy (2 through K), the prompt is augmented with the full text of all previously generated taxonomies and an explicit instruction to produce a substantially different organization — using a different grouping criterion.

B Dataset details

TaxoBench-CS (Zhu et al., 2025) contains 156 taxonomies covering approximately 11.6k scientific papers across multiple research domains.

CS-TaxoBench (Lahiri et al., 2025) consists of 460 taxonomies extracted from human-written survey papers along with an additional test set of 80 taxonomies curated from conference survey papers.

B.1 Experimental Setup Details

The ROOT implementation uses the following experimental hyperparameters. The number of diverse taxonomies to generate is $K = 3$, with a maximum tree depth of $D_{\max} = 3$ and a minimum depth threshold of $D_{\min} = 2$ for spectral pruning. Nodes are only split further if they contain more than `min_leaf_size` $= 4$ documents, and the maximum number of children per split is capped at $K_{\max} = 6$, with the optimal number of children k_{opt} selected automatically via silhouette score over the range $[2, K_{\max}]$. Documents are encoded using the `BAAI/bge-base-en-v1.5` sentence transformer with a batch size of 32. Taxonomy node labels are generated by a `meta-llama/Llama-3.1-8B-Instruct` language model with `max_new_tokens` $= 20$ and greedy decoding (`do_sample` = False); internal node prompts sample papers at five quantile positions (Q_0, Q_{25}, Q_{50}, Q_{75}, Q_{100}) along the projection axis.

C Diversity Evaluation

The main aim of this paper is to is that each tree will discover a genuinely different way to organize a set of given papers. We aim to measure whether our proposed algorithm and the baseline algorithms generate significantly different K taxonomy trees or not. We compare pairs of trees and measure how much they agree or disagree about which papers should be grouped together. The key mechanism is the "flat partition at depth d": for any tree, we can slice it horizontally at depth d to get a flat clustering, where every paper is assigned to whichever node sits at depth d on its root-to-leaf path (or to its leaf node if the path is shorter than d). By comparing these flat clusterings at every depth level, we capture disagreement at both coarse and fine scales. We measure the Depth-Weighted Jaccard and the Depth-Weighted ARI to check the diversity of the generated taxonomy trees. Both metrics use the log-decaying depth weight:

$$w_d = \frac{1}{\ln(1+d)}, \qquad d = 1, 2, \ldots, D_{\max}, \quad (6)$$

giving $w_1 \approx 1.443$ (maximum weight at the root's children) and smooth decay thereafter. The rationale behind giving more weightage to the top level is because the splits generated by them are the most important and further clustering is dependent on how well the tree is defined at the top levels.

C.1 Depth-Weighted Jaccard

Depth-Weighted Jaccard measures the overlap between two trees' grouping decisions at every level of the hierarchy, using the co-clustering set as its building block. At each depth d, for a given tree, we identify all pairs of papers that end up in the same cluster when the tree is cut at that depth. This "co-clustering set" captures the tree's grouping decision at that granularity level. At depth 1, the co-clustering set reflects the coarsest split: which papers are on the same side of the root's partition? At the maximum depth, it reflects the finest-grained leaf assignments. The Jaccard index between two trees at depth d is the size of the intersection of their co-clustering sets divided by the size of their union. We define the co-clustering set at depth d as $C_t^{(d)} = \{(i,j) : \pi_t^{(d)}(i) = \pi_t^{(d)}(j)\}$, i.e. all document pairs that land in the same cluster when tree T_t is cut at depth d. The Jaccard index at depth d

between trees T_s and T_t is:

$$J^{(d)}(T_s, T_t) = \frac{|C_s^{(d)} \cap C_t^{(d)}|}{|C_s^{(d)} \cup C_t^{(d)}|}. \quad (7)$$

The depth-weighted Jaccard aggregates with the log-decay weights from Equation (6):

$$\mathrm{DWJ}(T_s, T_t) = \frac{\sum_{d=1}^{D_{\max}} w_d \cdot J^{(d)}(T_s, T_t)}{\sum_{d=1}^{D_{\max}} w_d}. \quad (8)$$

The global score averages over all tree pairs. A low Depth-Weighted Jaccard means the trees genuinely disagree about which papers should be grouped together, especially at the most important coarse-grained level.

C.2 Depth-Weighted Adjusted Rand Index

The Adjusted Rand Index is a classical clustering comparison metric. If we take two completely random partitions of the same papers, their Jaccard overlap will be some positive number (because random partitions inevitably share some co-clustered pairs), but their ARI will be approximately 0. An ARI of 1 means the two clusterings are identical; an ARI of 0 means they agree no more than random partitions would; and negative values (rare in practice) mean they agree less than random.

At each depth d, the metric cuts both trees into flat clusterings and computes the ARI between them. It then applies the same log-decay depth weights as Jaccard and averages across all depths. The final score is the mean across all tree pairs. At each depth d, compute $ARI\big(\pi_s^{(d)}, \pi_t^{(d)}\big)$ and aggregate with the same depth weights:

$$\mathrm{wARI}(T_s, T_t) = \frac{\sum_{d=1}^{D_{\max}} w_d \cdot ARI\big(\pi_s^{(d)}, \pi_t^{(d)}\big)}{\sum_{d=1}^{D_{\max}} w_d}. \quad (9)$$

A wARI near 0 means the trees' partitions are statistically independent — the null-space projection successfully forced each tree to discover a different organisational structure.

C.3 Label Uniqueness (LabelUniq)

It checks whether different trees produce genuinely different label vocabularies. The metric works by computing, for each pair of trees, a symmetric nearest-neighbour overlap. For every label in Tree A, it finds the most similar label in Tree B (by cosine similarity of their embeddings) and records that similarity. It does the same in the reverse direction (every label in Tree B matched to its nearest neighbour in Tree A) and averages. This gives the mean "best-case overlap" between the two trees' label sets. The final score is 1 minus the average overlap across all tree pairs. A high Label Uniqueness score means each tree uses a genuinely different vocabulary of topic labels.

D Taxonomy Semantics

The metrics described here judge the semantic quality of each individual tree in isolation, without comparing trees to each other.

D.1 Within-Cluster Cohesion

Cohesion measures how semantically homogeneous the documents within each subtree are. At each node, we gathers all documents that belong to that node's subtree and compute the cosine similarity between every pair of those documents using their dense embeddings, and take the average. Intuitively, a high-quality cluster should contain papers that are semantically similar in the embedding space. if $|S_v|$ is the number of documents in that subtree for each node v in the tree and $|S_v| \geq 2$, we compute the mean pairwise cosine similarity among all documents in its subtree:

$$\mathrm{Coh}(v) = \frac{1}{\binom{|S_v|}{2}} \sum_{\substack{i<j \\ i,j \in S_v}} \cos(h_i, h_j). \quad (10)$$

The per-node scores are combined into a single tree-level score using size-weighted averaging: each node's cohesion is multiplied by the number of papers in its subtree before averaging. Without weighting, a tiny subtree containing 2 highly similar papers would contribute as much as a broad subtree containing 500 papers, which would overemphasize small, potentially trivial clusters. Size weighting prevents the metric from being artificially inflated by many tiny pure leaves. Also, large internal nodes represent major semantic themes and errors at those nodes affect many documents simultaneously. Therefore, incoherent large subtrees should be penalized more heavily than incoherent micro-clusters.

$$\mathrm{Cohesion}(T_t) = \frac{\sum_{v \in V(T_t)} |S_v| \cdot \mathrm{Coh}(v)}{\sum_{v \in V(T_t)} |S_v|}. \quad (11)$$

Higher values indicate stronger semantic consistency within clusters.

D.2 Inter-Cluster Separation

Separation measures whether sibling nodes in the tree actually represent distinct sub-topics. The metric works at every branching point in the tree — every internal node where the algorithm decided to split a collection of papers into two or more sub-groups. At each such node, it first computes the centroid of each child's sub-collection by averaging the embeddings of all papers that belong to that child's subtree. These centroids represent the "average semantic direction" of each sub-group. It then measures the cosine distance between every pair of sibling centroids and averages those distances. For every internal node v with children $\mathrm{ch}(v) = \{c_1, \ldots, c_m\}$, compute the centroid of each child's subtree and measure the mean pairwise cosine *distance* between sibling centroids:

$$\mathrm{Sep}(v) \;=\; \frac{1}{\binom{m}{2}} \sum_{j<k} \big(1 - \cos(\mu_{c_j},\, \mu_{c_k})\big). \quad (12)$$

The per-branching-point scores are aggregated with size-weighting (the number of papers at each branching node) and averaged across all K trees.

D.3 Node Precision

Node Precision evaluates whether each paper genuinely belongs under the label it has been assigned to, using a careful three-step protocol designed to distinguish truly irrelevant papers from papers that are merely marginal. To evaluate whether taxonomy labels accurately reflect the semantic content of their assigned papers, we adopt this metric inspired by the citation precision metric from (Gao et al., 2023).

For each eligible node v (non-empty label, $|S_v| \geq 2$), a three-step protocol is applied to every paper $x_i \in S_v$:

1. **Recall check.** Does the full set of papers in S_v, taken collectively, represent the topic L_v? If no, $\mathrm{NP}(v) = 0$.

2. **Individual relevance $\varphi_a(i)$.** Does paper x_i individually support L_v?

3. **Leave-one-out $\varphi_b(i)$.** If $\varphi_a(i) = 0$: does the remaining set $S_v \setminus \{x_i\}$ still represent L_v?

A paper is deemed *irrelevant* iff $\varphi_a(i) = 0$ and $\varphi_b(i) = 1$ (it does not support the label, but the cluster survives without it). Node-level precision is:

$$\mathrm{NP}(v) = \frac{|S_v| - |\{i \in S_v : \text{irrelevant}\}|}{|S_v|}. \quad (13)$$

We use a natural language inference (NLI) model as the entailment function ϕ acting as a semantic judge. For our experiments, we use the cross-encoder/nli-deberta-v3-base model as the NLI model.

E Human Annotation Guidelines

Task. You will evaluate automatically generated taxonomies that organise research papers into hierarchical categories. For each item, you are shown K taxonomy trees produced by a single system for a given topic. Your task is to rate the output on three dimensions using integer scores from 1 to 10 (higher is better).

Dimensions.

1. **Diversity** — How different are the K trees from each other?
   1–2: nearly identical trees; *9–10:* highly varied structures and groupings.

2. **Structure** — How evenly are papers distributed across leaf clusters?
   1–2: one cluster holds almost all papers; *9–10:* near-uniform spread.

3. **Label Quality** — Are the node labels informative and non-redundant?
   1–2: vague or synonymous labels (e.g. "Other", "Misc"); *9–10:* crisp, distinct, and descriptive labels.

Procedure.

1. Examine all K trees: expand nodes, read labels, and note paper counts.

2. Compare across trees before scoring Diversity.

3. Inspect leaf-level paper counts for Structure.

4. Check whether sibling labels overlap for Label Quality.

5. Submit your three scores. Use the notes field for any edge-case observations.

Key Rules. Use the full 1–10 scale; rate each dimension independently; do not let the variant name influence your judgement; and skip any item whose topic falls outside your expertise rather than guessing.

F Our Method – Prompts

F.1 Prompt for ROOT algorithm

System Prompt:

> You are a research taxonomy expert who creates SPECIFIC, DISCRIMINATIVE labels for clusters of academic papers. STRICT RULES — violating any rule means your answer is wrong: 1. Respond with ONLY the label. No explanation, no commentary, no punctuation at the end, no quotes. 2. The label MUST be 3–8 words. 3. NEVER reuse the parent node's label verbatim or in paraphrase. 4. The label should represent the closest concrete sub-task formulation, training paradigm, architectural component, application domain, or a sub-task/sub-domain of the parent label. 5. NEVER use these alone as a label: 'Techniques', 'Methods', 'Approaches', 'Systems', 'Models', 'Advances', 'Overview', 'Survey', 'Analysis', 'Framework', 'Architecture'. 5. ALWAYS name at least one SPECIFIC item: a concrete sub-task formulation, training paradigm, architectural component, or application domain.You will receive a...

Internal Node Prompt:

> PARENT NODE LABEL: {parent_label}
> Your label must be MORE SPECIFIC than the parent. Do NOT repeat or paraphrase it.
> Below are papers from this cluster ordered from one extreme of the splitting axis to the other:
> [Extreme-A (one end of axis)] paper text
> [Lower-Quartile (25%)] paper text
> [Midpoint (50%)] paper text
> [Upper-Quartile (75%)] paper text
> [Extreme-B (other end of axis)] paper text
> Good label examples (specific, discriminative): 'Autoregressive WaveNet-Style Neural Vocoders' 'Graph Convolutional Networks for ABSA' 'Sparse and Linear Self-Attention Variants' 'Cross-Lingual Transfer via Multilingual Pretraining' 'Reinforcement Learning from Human Feedback' 'Aspect Sentiment Triplet Extraction' Bad labels (too generic, forbidden): 'Natural Language Processing Techniques' 'Deep Learning Approaches' 'Speech Synthesis Methods' 'Transformer Model Improvements' (anything that just restates the parent or survey topic)
> What SPECIFIC sub-task, method, architecture, or technique UNIFIES the papers in this cluster shown above? Respond with ONLY the label (3–8 words).