TaxoAdapt: Aligning LLM-Based Multidimensional Taxonomy Construction to Evolving Research Corpora

Anonymous ACL submission

Abstract

The rapid evolution of scientific fields introduces challenges in organizing and retrieving scientific literature. While expert-curated taxonomies have traditionally addressed this need, the process is time-consuming and expensive. Furthermore, recent automatic taxonomy construction methods either (1) overrely on a specific corpus, sacrificing generalizability, or (2) depend heavily on the general knowledge of large language models (LLMs) contained within their pre-training datasets, often overlooking the dynamic nature of evolving scientific domains. Additionally, these approaches fail to account for the multi-faceted nature of scientific literature, where a single research paper may contribute to multiple *dimensions* (e.g., methodology, new tasks, evaluation metrics, benchmarks). To address these gaps, we propose **TaxoAdapt**, a framework that dynamically adapts an LLM-generated taxonomy to a given corpus across multiple dimensions. TaxoAdapt performs iterative hierarchical classification, expanding both the taxonomy width and depth based on corpus' topical distribution. We demonstrate its state-of-the-art performance across a diverse set of computer science conferences over the years to showcase its ability to structure and capture the evolution of scientific fields. As a multidimensional method, TaxoAdapt generates taxonomies that are 26.51% more granularity-preserving and 50.41% more coherent than the most competitive baselines judged by LLMs.

1 Introduction

Driven by increased research interest and accessibility, the rapid proliferation of scientific literature and subsequent creation of new branches of knowledge (e.g., the rise of generative models in the last five years) has made organizing and retrieving domain-specific knowledge increasingly challenging (Bornmann et al., 2021; Aggarwal et al., 2022). Taxonomies enhance data organization, support search engines, capture semantic relationships, and aid discovery. While expert-curated and crowdsourced taxonomies have traditionally structured topics into hierarchies (e.g., text classification → spam detection), manual curation is time-consuming and struggles to keep pace with rapidly evolving fields (Bordea et al., 2016; Jurgens and Pilehvar, 2016).

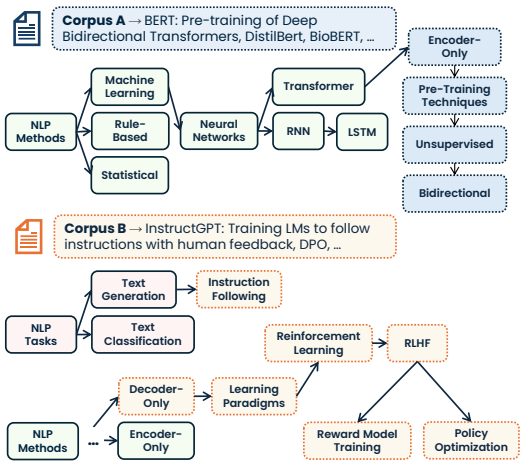

Figure 1: Each paper within a corpus contributes to different dimensions of scientific literature. We show how corpora from different eras of NLP (e.g., BERT-era; RLHF-era) can influence their respective dimension-specific taxonomies (we highlight certain subtrees).

Prior efforts in automating taxonomy construction (ATC) fall into two categories: *corpus-driven* methods that *extract* topics and relationships directly from text, and *LLM-based* approaches which *generate* taxonomies based on pre-existing knowledge. While corpus-driven methods effectively capture meaningful, domain-specific topics, they rely on *rigid approaches* that are restricted to only terms within the corpus vocabulary and lack extensive background knowledge, given their pre-LLM origins (Liu et al., 2012; Shen et al., 2018; Shang et al., 2020; Zhang et al., 2018). Conversely, LLM-based methods generate large-scale, general-purpose taxonomies but currently lack mechanisms to align

them with specialized knowledge, solely relying on their background knowledge of domains and their key topics (Chen et al., 2023; Shen et al., 2024; Zeng et al., 2024; Sun et al., 2024).

Moreover, as of now, *both approaches* overlook the *multidimensional* nature of scientific literature. A research paper may study and/or contribute to multiple aspects of the scientific method (tasks, methods, applications, etc.), based on which we could organize papers differently. When new knowledge emerges, we must adapt existing taxonomies. For example, in Figure 1, InstructGPT (Ouyang et al., 2022) introduces both "Instruction Following" as a novel NLP *task* and "Reinforcement Learning with Human Feedback" (RLHF) as an NLP *method*, highlighting the limitations of uni-dimensional taxonomies. Limiting ATC design to the task dimension is a *critical oversight*— obscuring the broader, evolving impacts of research. Ultimately, both corpus and LLM-based methods fail to provide a multidimensional view of scientific literature. To address these gaps, we propose **TaxoAdapt**, a framework that dynamically grounds LLM-based taxonomy construction to scientific corpora across multiple dimensions. TaxoAdapt operates on three core principles:

Knowledge-augmented expansion leads to specialized, relevant taxonomies. State-of-the-art LLMs struggle to accurately model specialized taxonomies in domains like computer science (Sun et al., 2024), particularly leaf-level entities. Existing LLM-based methods require pre-defined entity sets or are limited to entity-level context for taxonomy construction (Zeng et al., 2024; Chen et al., 2023), critically limiting the degree of domain-specific knowledge which they can exploit. Alternatively, TaxoAdapt leverages document-level reasoning; by using each paper's title and abstract, it identifies *which* dimensions a paper contributes to (e.g., methods, datasets) and *how*. For example, as shown in Figure 1, when expanding the "Transformer" node under NLP methods, TaxoAdapt selectively analyzes papers centered on Transformer-based architectures (e.g., BERT)– helping to derive subcategories like "Encoder-Only". Unlike mining important entities, this document-grounded approach enhances taxonomic precision by *aligning expansion with corpus knowledge specific to each dimension, layer, and node*.

Hierarchical text classification provides crucial signals for targeted exploration. Scientific fields evolve rapidly, with new subdomains emerging and existing ones merging or fading (Singh et al., 2022). Figure 1 illustrates this: Corpus A (2018–2022) emphasizes BERT-like encoders, while Corpus B (2022–present) highlights "RLHF" as a training method and "Instruction Following" as a key task behind InstructGPT and its successors. LLM-generated taxonomies often overlook such trends, favoring concepts broadly represented within the training data (e.g., high-level tasks like text classification). To address this, TaxoAdapt dynamically adapts the taxonomy by employing hierarchical text classification to determine *which nodes should be expanded and how*. A node with a high density of papers (e.g., RLHF) indicates further exploration and warrants ***depth expansion*** (e.g., Reward Model Training, Policy Optimization). Conversely, if a node has many unmapped papers (e.g., if "Decoder-Only" did not exist under "Transformer"), it signals parallel research to existing children (e.g., "Encoder-Only"), necessitating ***width expansion***. Nodes with minimal presence in the corpus (e.g., LSTMs) will consequently not be explored further.

Taxonomy-aware clustering enables meaningful expansion. Multiple factors determine which entities should be used to expand a given node: (1) maintaining hierarchical, granular relationships (e.g., identify a *dimension-specific child* of "Transformer" and a *sibling* of "Encoder-Only"), (2) prioritizing presence within the corpus, and (3) minimizing redundancy. Recently, LLMs have shown strong entity clustering abilities (Viswanathan et al., 2023; Zhang et al., 2023). Thus, TaxoAdapt utilizes its knowledge of the dimension, layer, and papers mapped to the specific node being expanded to determine granularity-consistent candidate entities. It then utilizes this information to *guide the clustering* of the candidate entities, maximizing coverage while minimizing redundancy during expansion.

Overall, **TaxoAdapt** aligns the multidimensional taxonomy generation (and expansion) process to a corpus. We summarize our contributions below:

- To the best of our knowledge, **TaxoAdapt** is the *first* framework to ground LLM-based taxonomy construction to a corpus and study this task from multiple dimensions.
- We propose a novel classification-based expansion and clustering framework for targeted, meaningful corpus exploration.
- Through quantitative experiments and real-

world case studies, we show that TaxoAdapt outperforms baselines in taxonomic coverage, granular-consistency, and adaptability to emerging research trends.

Reproducibility: Our dataset and code will be provided upon paper acceptance.

2 Related Works

Prior research on taxonomy construction can be broadly categorized into three types: manual, corpus-driven, and LLM-based methods.

Manual Curation. Previous works (Bordea et al., 2016; Jurgens and Pilehvar, 2016; Yang et al., 2013) focused on extracting hand-crafted taxonomies from candidate nodes or designing systems to support the creation of human-assisted taxonomies. These taxonomies involve mostly manual work, making them expensive both during the creation process and for future maintenance, especially given the rapid evolution of scientific fields. Thus, ATC is highly needed.

Corpus-driven Methods. A line of research (Lu et al., 2024; Lee et al., 2022a,b; Zhang et al., 2018; Huang et al., 2020) employed clustering to extract entities and their relationships from the corpus, identifying semantically coherent concept terms to complete a given seed taxonomy. Alternatively, NetTaxo (Shang et al., 2020) leveraged the metadata of corpus documents as additional signals to construct taxonomies from scratch. Without clustering, HiExpan (Shen et al., 2018) utilized a relation extraction module to perform depth expansion. Although these approaches maintain a high degree of specificity to the corpus, their lack of LLM usage limits access to broader background knowledge, which is crucial for preserving hierarchical and granular node relationships.

LLM-based Methods. Many recent works explore the potential of leveraging LLMs for taxonomy expansion or construction. Researchers aimed to answer whether LLMs are good replacement of traditional taxonomies and knowledge graphs, and they found that LLMs still could not capture the highly specialized knowledge of taxonomies and leaf-level entities well (Sun et al., 2024). In terms of LLM usage, prompting without explicit fine-tuning on any data outperformed fine-tuning-based methods (Chen et al., 2023). TaxoInstruct (Shen et al., 2024) unified three relevant tasks (entity set expansion, taxonomy expansion, and seed-guided taxonomy construction) by unleashing the instruction-following capabilities of LLMs. Although different iterative prompting approaches (Zeng et al., 2024; Gunn et al., 2024) have been proposed, there does not exist an LLM-based method that aligns well with the evolving scientific corpus to the best our knowledge. This reinforces our motivation of designing TaxoAdapt.

3 Methodology

As shown in Figure 2, TAXOADAPT aims to align LLM taxonomy generation to a specific corpus, improving adaptability to evolving research corpora.

3.1 Preliminaries

3.1.1 Problem Formulation

We assume that as input, the user provides a topic t (e.g., natural language processing), a set of dimensions D (e.g., tasks, datasets, methods, evaluation metrics), and a scientific corpus P. We assume that each paper $p \in P$ is relevant to t and studies at least one $d \in D$. TaxoAdapt aims to output a set of $|D|$ taxonomies $T_{d \in D}$, maximizing the quantity of papers $p \in P$ mapped across all nodes $n_d \in T_d$. The topic t and dimension $d \in D$ form the root topic n_0 of each taxonomy T_d (e.g., "*natural language processing tasks*"). In order to provide an additional level of flexibility, we define each taxonomy as a directed acyclic graph (DAG) since certain nodes may have two parents (e.g., the scientific question answering (QA) task may be placed under both "question_answering" and "scientific_reasoning").

3.1.2 Initial LLM-Based Taxonomy Construction

Recent works (Chen et al., 2023; Sun et al., 2024; Zeng et al., 2024; Shen et al., 2024) have explored leveraging LLMs for taxonomy construction, showing their potential for generating high-level, general-purpose taxonomies (although, these are not guaranteed to be representative of a specific corpus). Given the difficulty of acquiring expert-curated taxonomies across multiple domains and the lack of methods addressing taxonomy construction across multiple dimensions, we utilize an LLM to generate $|D|$ initial single-level taxonomies ($T_{d \in D}$) for **TaxoAdapt** to expand. This allows us to demonstrate TaxoAdapt's effectiveness while minimizing user input requirements. Nonetheless, this taxonomy can also be replaced by any specific taxonomy which the user desires.

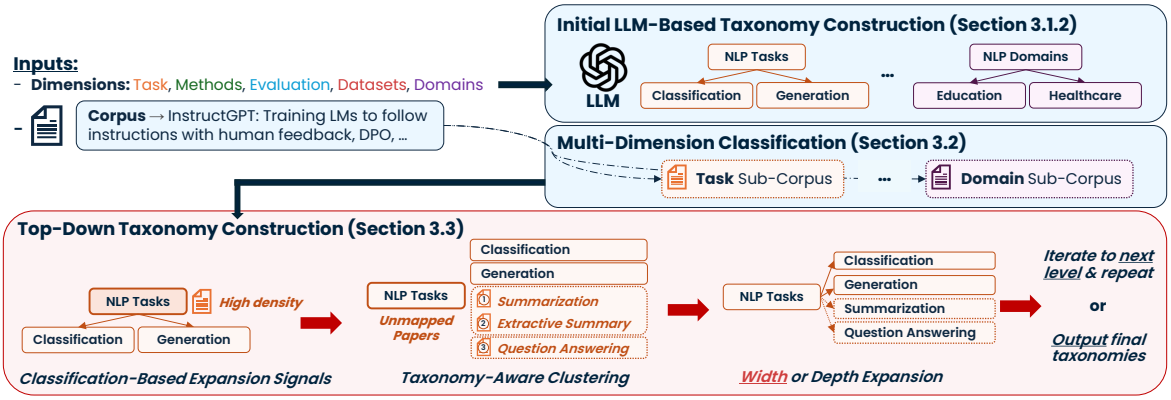

Figure 2: We propose **TAXOADAPT**, a framework which dynamically constructs a LLM-enhanced, corpus-specific taxonomy using classification-based expansion signals. The diagram demonstrates a *width* expansion example, but the same logic is applied to depth expansion (simply without the additional sibling context).

3.1.3 Taxonomy Expansion

Taxonomy expansion involves *both* **depth** and **width** expansions of a provided taxonomy, T_d. We formally define these below:

Definition 1 (DEPTH EXPANSION) *Expanding a leaf node $n_{i,d} \in T_d$ by identifying a set of child entities $n^i_{j,d} \in N^i_d$, which topically falls under $n_{i,d}$ and contains equally granular entities (e.g., $n^i_{1,d}$ and $n^i_{2,d}$ should be equally topically specific).*

Definition 2 (WIDTH EXPANSION) *Expanding the children of a non-leaf node $n_{i,d}$, where its existing children $n^i_{j,d} \in N^i_d$ represent an incomplete set of entities that need to be further completed by additional, unique sibling nodes, $n'^i_d \in N'^i_d$. N'^i_d and N^i_d are non-overlapping and at the same level of granularity.*

Note that we do not assume a user-provided set of entities for either, which has historically been the case (Zeng et al., 2024; Shen et al., 2018).

3.2 Multi-Dimension Classification

Scientific literature is inherently multifaceted, with individual papers often contributing to multiple aspects of a domain– such as tasks, methodologies, and datasets. Thus, we must construct a *set* of taxonomies $T_{d \in D}$ that captures the diverse aspects of scientific knowledge. TaxoAdapt seeks to *align* taxonomy T_d's construction with the dimension-specific contributions featured within a corpus. Thus, we study if and how to minimize the noise present from papers that do not make any contributions towards dimension d. For example, a paper that only proposes a new text classification dataset, but still utilizes standard F1-metrics would introduce noise for constructing the "evaluation method" taxonomy and consequently, may be omitted. To explore this, we partition the corpus based on the dimensions each paper contributes to before we perform taxonomy expansion.

We treat this task as a multi-label classification problem. Recent works have shown that LLMs are successful at fine-grained classification in a multitude of domains (Zhang et al., 2024b,a). Thus, we prompt the LLM to classify the paper p, where in-context, we provide the dimension options and their definitions. We define each dimension $d \in D$ with respect to the type of contribution we would expect a paper $p_{i,d}$ to make. By default, we assume each paper always falls under the task dimension. We make this assumption because every work has a contribution that is aligned to a specific goal/task. Ultimately, we utilize the output labels for each paper $p \in P$ in order to partition the corpus P into $|D|$ potentially overlapping subsets: $P_d \subseteq P$. Our definitions are summarized below (full-length version in Appendix B):

- **Task:** Assume all papers are aligned to a task(s).
- **Methodology:** A paper that introduces, explains, or refines a method or approach.
- **Datasets:** Introduces a new dataset.
- **Evaluation Methods:** A paper that assesses the performance, limitations, or biases of models, methods, or datasets.
- **Real-World Domains:** A paper that demonstrates the use of techniques to solve real-world problems or address specific domain challenges.

3.3 Top-Down Taxonomy Construction

An LLM-generated taxonomy may not sufficiently capture all the topics within a corpus, especially in

emerging research areas. These areas are underrepresented in the LLMs' general-purpose background knowledge but are highly represented within the input corpus (e.g., the node "*RLHF*" in Figure 1). Given that domain-specific trends are continually evolving in scientific literature, we must ensure that both the *depth* and *breadth* of the underlying research landscape are accurately represented.

To determine which nodes require deeper exploration, we employ hierarchical classification. Adapting an LLM-based text classification model (Zhang et al., 2024b), we enrich the taxonomy nodes (e.g., by adding keywords) to support top-down classification from $n_{i,d}$ to $n^i_{j,d}$. Specifically, given a dimension-specific paper p mapped to $n_{i,d}$, we adapt this model to determine whether p (based on its title and abstract) maps to any child node $n^i_{j,d} \in N^i_d$ via multi-label classification using node labels and descriptions. We define $n_{i,d}$'s **density** $\rho(n_{i,d})$ as the *number of papers* $|P_{i,d}|$ mapped to it, leveraging $\rho(n_{i,d})$ to decide whether its children (or lack thereof) should be expanded.

3.3.1 Depth & Width Expansion Signals

When many papers accumulate at a given leaf node $n_{i,d}$, as indicated by a high value of $\rho(n_{i,d})$, it suggests that the topic represented by $n_{i,d}$ is being explored in greater depth within the corpus– which the current taxonomy does not adequately reflect. Longer taxonomy paths signify popular research topics within the corpus. Figure 1 illustrates this: the path to "*bidirectional*" is significantly deeper than to "*rule-based*", reflecting the rise of bidirectional pre-trained language models in Corpus A and the subsequent decline of rule-based methods. In this scenario, if $\rho(n_{i,d}) \geq \delta$ (user-specified threshold), TaxoAdapt performs **depth expansion** (Definition 1) by identifying a set of child entities N^i_d that partition the topic into finer, granularity-consistent subtopics. For instance, as shown in Figure 1, if $\rho($"*encoder-only*"$) \geq \delta$, this warrants further decomposition– such as deepening the path to include "*pre-training techniques*"– to capture the ongoing, specialized research in that area.

A *complementary* signal is provided by the ***unmapped*** **density** $\tilde{\rho}(n_{i,d})$ of a non-leaf node. This arises when a node $n_{i,d}$ has a significant number of papers mapped to it (a high $\rho(n_{i,d})$) that are not allocated to any of its existing child nodes N^i_d.

Definition 3 (UNMAPPED DENSITY) *Let $P_{i,d}$ denote the set of all papers associated with node $n_{i,d}$, and let $n_{j,d} \in N^i_d$ denote the set of children under node $n_{i,d}$. The unmapped density is then given by:*

$$\tilde{\rho}(n_{i,d}) = \left| P_{i,d} - \bigcup_{j=0}^{|N^i_d|} P_{j,d} \right| \quad (1)$$

If $\tilde{\rho}(n_{i,d})$ exceeds a predefined threshold τ, this indicates that a significant portion of the corpus within $n_{i,d}$ is not adequately represented by its current children. In such cases, TaxoAdapt initiates **width expansion** by generating additional, non-overlapping sibling nodes $n'^i_{j,d} \in N'^i_d$ to cover the underrepresented research areas. For instance, the "*decoder-only*" node in Figure 1, where a high $\tilde{\rho}($*"NLP Methods"*$)$ signaled that the single "*encoder-only*" node did not adequately capture the surge in decoder-only architectures. Once node $n_{i,d}$ is triggered for either depth or width expansion, TaxoAdapt determines the new set of child entities N'^i_d through a pseudo-label clustering procedure (Section 3.3.2).

3.3.2 Taxonomy-Aware Clustering

Assuming that node $n_{i,d}$ has been marked for expansion, we must identify a set of child entities (N'^i_d if $n_{i,d}$ is a leaf node, otherwise N^i_d) which satisfy the following criteria:

1. *Maintaining* the hierarchical, granular *relationships* which currently exist within the taxonomy (parent-child and sibling-sibling relationships).
2. *Maximizing presence* within either the set of unmapped papers $\tilde{\rho}(n_{i,d})$ (width expansion), or $\rho(n_{i,d})$ (depth expansion).
3. *Minimizing redundancy* between the child entities $N^i_d \cup N'^i_d$.

Subtopic Pseudo-Labeling. In order to maintain the hierarchical relationships within the taxonomy, we utilize the LLM to generate dimension and granularity-preserving pseudo-labels based on each paper $p_{i,d} \in P_{i,d}$'s title and abstract. We prompt the LLM to determine its dimensional subtopic relative to $n_{i,d}$ as its parent ($n_{i,d}$'s label, dimension, description, and path of ancestors) and $n_{i,d}$'s existing children, if any.

Subtopic Clustering. Given that each paper is represented by its corresponding pseudo-label, clustering these pseudo-labels allows us to *maximize* the number of papers ($\tilde{\rho}(n_{i,d})$ or $\rho(n_{i,d})$) represented. Moreover, effective clustering inherently minimizes redundancy as it aims to

produce distinct, non-overlapping sets of papers. We specifically exploit LLM's clustering abilities (Viswanathan et al., 2023; Zhang et al., 2023) as this allows us to easily integrate dimension and granularity-specific information into the context and preserve these features within our clusters. Including the same context provided during Subtopic Pseudo-Labeling, in addition to the complete list of paper-subtopic pseudo-labels, we prompt an LLM to determine the primary sub-[dimension] topic clusters (e.g., sub-task, sub-methodology) that would best encompass the list of pseudo-labels, providing a label and description for each cluster. These generated clusters consequently form $N_d'^i$ if $n_{i,d}$ is a leaf node (depth expansion) and otherwise N_d^i (width expansion).

We iteratively classify, identify expansion signals, and perform taxonomy-aware clustering level-by-level. We provide the full top-down taxonomy construction algorithm in Algorithm 1 (Appendix D). Ultimately, this process ends when either no nodes are signaled for expansion or the maximum taxonomy depth is reached—outputting our final $T_d, \forall d \in D$.

4 Experimental Design

We explore **TAXOADAPT**'s performance using a hybrid of both open (`Llama-3.1-8B-Instruct`) and closed source (`GPT-4o-mini`) models. We do this to showcase how we can optimize the cost of the classification and pseudo-labeling steps (both run on `Llama`) while not needing to sacrifice performance. We construct initial, deterministic single-level taxonomies using `GPT-4o-mini` (Section 3.1.2). For all other modules of our framework, we sample from the top 1% of the tokens and set the temperature to 0.1. We set the density threshold $\delta = 40$ papers and the maximum depth $l = 2$.

4.1 Dataset

In order to evaluate TAXOADAPT's abilities to adapt to different corpora and reflect evolving research topics, we select several conferences spanning different subdomains within computer science. These conferences and their respective sizes are shown in Table 1, where we collect the title and abstract for each paper. We choose to explore our method specifically within computer science such that our dimensions can remain consistent across all conferences: task, methodology, dataset, evaluation methods, and real-world domains. We also include one conference from two different years (e.g., EMNLP'22 and EMNLP'24) in order to showcase how our method reflects the evolution of its respective field.

Table 1: Topic t and number of papers (size) per dataset.

Conference	Size	Topic t
EMNLP 2022	828	Natural Language Processing
EMNLP 2024	2954	
ICRA 2020	1000	Robotics
ICLR 2024	2260	Deep Learning
Total Papers	7,042	

4.2 Baselines

TaxoAdapt aligns LLM-based taxonomy construction to a specialized, multidimensional corpus. Consequently, we choose to compare our method with both *corpus-driven* and *LLM-based* approaches. Note that all LLM-based baselines utilize `GPT-4o-mini` as their underlying model. We provide detailed information on each baseline in Appendix A.

1. ***LLM-Only*** $\rightarrow$ **Chain-of-Layer (Zeng et al., 2024):** Given a set of entities, solely relies on an LLM (***no corpus***) to select relevant candidate entities for each taxonomy layer and construct the taxonomy from top to bottom.
2. ***LLM + Corpus*** $\rightarrow$ **Prompting-Based:** An iterative baseline which prompts the LLM to identify relevant papers to the dimension, child nodes, and their corresponding papers.
3. ***Corpus-Only*** $\rightarrow$ **TaxoCom (Lee et al., 2022a):** A corpus-driven, handcrafted taxonomy completion framework that clusters terms from the input corpus to recursively expand a handcrafted seed taxonomy.
4. ***No-Dim*** and ***No-Clustering*** are TaxoAdapt ablations which remove the dimension-specific partitioning and subtopic clustering respectively.

4.3 Evaluation Metrics

We design a thorough automatic evaluation suite using `GPT-4o` and `GPT-4o-mini` to determine the quality of our generated taxonomies, using both node-level and taxonomy-level metrics. For each judgment, we ask the LLM to provide additional rationalization (all prompts are in Appendix E):

- **(*Node-Wise*) Path Granularity:** Does the path to node $n_{i,d}$ preserve the hierarchical relationships between its entities (is each child n_j^i more

Table 2: Comparison of models on all datasets, averaged across all dimensions. All values are normalized and scaled by 100. The highest scores for each metric are **bolded**, and the second-highest scores are marked with a †.

Models	EMNLP'22					EMNLP'24				
	Path	Sib	Dim	Rel	Cover	Path	Sib	Dim	Rel	Cover
Chain-of-Layers	46.87	67.67	94.61	77.65	50.54	49.56	67.67†	92.56†	82.13	48.66
With-Corpus LLM	66.14	33.93	88.82	72.87	39.35	49.51	29.74	83.56	84.13†	39.20
TaxoCom	23.85	33.89	89.81	**91.31**	**64.53**	13.89	59.42	86.97	**95.96**	64.81
TaxoAdapt	81.09	**82.92**	**100.00**	82.69†	55.81†	83.04	**77.86**	98.04	88.76†	60.29†
- *No Dim*	**88.47**	82.30	99.49	81.46	62.26	**89.98**	76.97	**99.05**	86.23	**66.42**
- *No Clustering*	76.45†	69.33†	98.49†	81.63	50.38	65.15†	62.15	92.31	80.22	60.80

Models	ICRA'20					ICLR'24				
	Path	Sib	Dim	Rel	Cover	Path	Sib	Dim	Rel	Cover
Chain-of-Layers	52.92	43.46	95.06	95.00	55.96	40.75	43.16	95.92	69.66	48.50
With-Corpus LLM	74.58	32.54	97.34	94.18	45.50	70.44	29.70	88.37	67.78	33.62
TaxoCom	43.05	54.21	99.06†	96.28†	60.75†	30.00	67.00	91.27	**86.88**	56.25†
TaxoAdapt	86.69	**91.59**	**100.00**	**97.82**	52.09	78.93†	**81.47**	99.62†	71.99†	53.96
- *No Dim*	**91.82**	89.59†	**100.00**	92.95	**67.97**	**86.32**	76.45†	**100.00**	69.45	**62.54**
- *No Clustering*	87.74†	85.76	**100.00**	93.97	50.86	65.69	67.85	93.13	68.56	54.60

specific than the parent $n_{i,d}$)? Scored 0 or 1 by `GPT-4o`.

- **(*Level-Wise*) Sibling Coherence:** Determine whether a set of siblings $n_j \in N^i$ of parent node $n_{i,d}$ form a coherent set with the same level of specificity and granularity. Scored from 0 to 1 by `GPT-4o`.
- ***(Node-Wise)* Dimension Alignment:** Is the node $n_{i,d}$ relevant to the *dimension* d of the root topic t? Scored 0 or 1 by `GPT-4o`.
- **(*Node-Wise*) Paper Relevance:** Is the node $n_{i,d}$ relevant to at least 5% of the corpus? Scored 0 or 1 per node by `GPT-4o-mini` (due to longer paper context and thus, cost). Final score is averaged across all nodes.
- ***(Level-Wise)* Coverage:** Given a set of siblings $n_j \in N^i$ of parent node $n_{i,d}$, determine what portion of relevant papers of $n_{i,d}$ are covered by (relevant to) at least one node in the siblings. Scored by `GPT-4o-mini` (due to longer paper context and thus, cost).

In addition to this automatic evaluation, we also conduct a supplementary human evaluation for these evaluation metrics. We provide the LLM-human agreement analysis in Appendix C.

Table 3: Standard deviation of model performance across all **datasets *and* dimensions**.

Models	Path	Sib	Dim	Rel	Cover
Chain-of-Layers	0.078	0.109	0.008	0.043	0.005
With-Corpus LLM	0.054	0.036	0.010	0.027	**0.004**
TaxoCom	0.041	0.035	0.039	**0.016**	0.022
TaxoAdapt	**0.027**	**0.021**	**0.007**	0.043	0.015

5 Experimental Results

Overall Performance & Analysis. Table 2 shows the performance of **TAXOADAPT** compared with the baselines on a wide variety of node, level, and taxonomy-wise metrics. From the results, we can see that TaxoAdapt's taxonomies are 26.51% ***more granularity-preserving***, 50.41% ***more coherent***, 5.16% ***more dimension-specific***, 5.18% ***more relevant*** to the corpus, and 9.07% ***more representative*** of the corpus, compared to the most competitive baseline across all datasets and dimensions. These results indicate that TaxoAdapt is *significantly better* at aligning to a corpus across multiple dimensions, *while still greatly improving the structural integrity of the constructed taxonomies*. Based on our thorough set of experiments, we are able to draw several interesting insights:

TAXOADAPT constructs well-balanced, cohesive taxonomies. We observe that the baselines tend to generate significantly imbalanced taxonomies, where several of the nodes have only a *single child*. Furthermore, each level tends to have an *uncohesive mixture of granularities* (e.g., "*Sentiment Analysis*", "*Emotion Detection*" as siblings). This is especially the case for TaxoCom, which has a significantly low path granularity while having the highest relevance and coverage score. This is due to it selecting highly coarse-grained nodes (e.g., *NLP tasks* $\rightarrow$ *significant improvements* $\rightarrow$ *closed source, out of domain, text based,* ...). In contrast, TaxoAdapt preserves the hierarchical relationships between the topics of taxonomy with

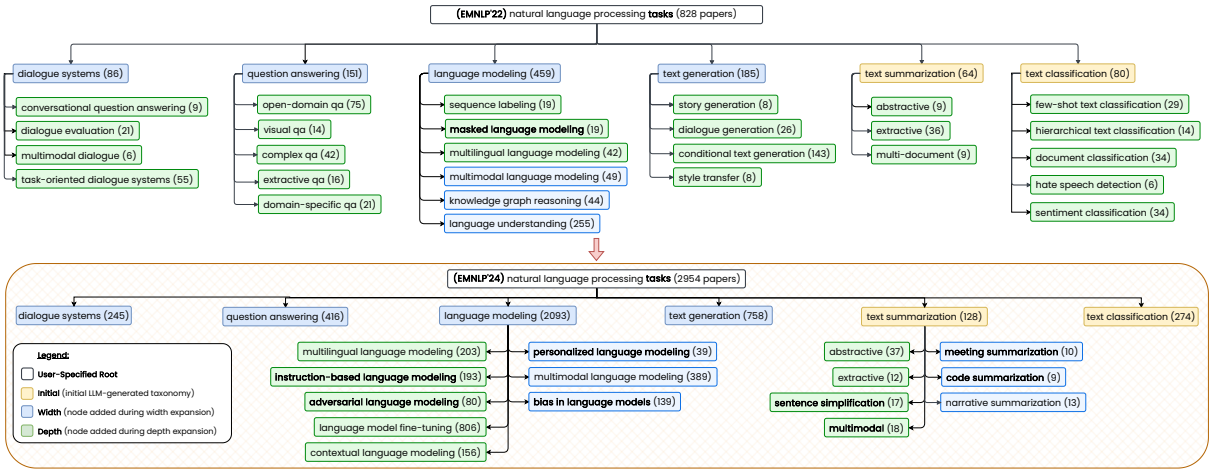

Figure 3: We show the evolution of NLP Tasks from EMNLP'22 to EMNLP'24. We ***highlight specific subtrees***, emphasizing nodes which reflect the most interesting research trends. We also show the number of papers that TaxoAdapt maps to each of the nodes (Section 3.3) in parentheses.

cohesive sets of children for each non-leaf node, where the children $n_j^i \in N^i$ of node n_i have high relevance and coverage of n_i's corresponding set of papers P_i. Furthermore, each child node n_j^i is relevant to at least 5% of the papers within the corpus P, reflected in increased path granularity, sibling cohesiveness, and coverage scores shown in Table 2. We can attribute these gains to TaxoAdapt's hierarchical classification and taxonomy-aware clustering steps based on the lower performance of ablation, *No Clustering*. We also note that *TaxoAdapt primarily uses* `Llama-3.1-8B` *as its backbone model for classification and clustering*, which is a significantly weaker model than the baselines' complete dependence on `GPT-4o-mini`.

TaxoAdapt is robust to different research dimensions. In addition to each of TaxoAdapt's nodes $n_{i,d} \in T_d$ better reflecting its corresponding dimension (***Dim***), TaxoAdapt exhibits robustness to the different research dimensions. Specifically, Table 3 showcases the standard deviation of each model's scores averaged across all dimensions and datasets. We observe that TaxoAdapt features the ***lowest standard deviations*** across all granularity metrics, while simultaneously scoring the highest for each (Table 2). We further explore this finding through ablation "*No-Dim*", which removes the initial dimension-specific partitioning of the corpus P into $P_{d \in D} \subset P$ (Section 3.2). We observe that partitioning the corpus improves granularity, but also negatively impacts relevance and coverage– only a narrowed, dimension-specific pool is considered *relevant* for dimension-specific taxonomy construction.

TaxoAdapt constructs taxonomies which reflect evolving research. In Figure 3, we demonstrate how TaxoAdapt's taxonomies adapt to corpora from different eras of natural language processing research (EMNLP'22 $\rightarrow$ EMNLP'24). We showcase the task dimension, where due to the rapid increase in EMNLP submissions and accepted papers, features more nodes overall (**EMNLP'22:** 62 nodes; **EMNLP'24:** 99 nodes). Furthermore, between the two conference years, we see certain nodes fall in research presence (e.g., *masked language modeling*) and others significantly rise (e.g., *language modeling, instruction-based language models, bias in language models*). We also see certain research trends start to arise as a result of performing ***width*** expansion based on initially unmapped papers (e.g., *personalized language models*). Overall, Figure 3 demonstrates the power of ***considering classification-based signals*** for ***knowledge-augmented expansion***.

6 Conclusion

We introduce **TaxoAdapt**, a novel framework for constructing multidimensional taxonomies aligned with evolving research corpora using LLMs. TaxoAdapt dynamically adapts to corpus-specific trends and research dimensions. Our comprehensive experiments demonstrate that TaxoAdapt significantly outperforms existing methods in granularity preservation, dimensional specificity, and corpus relevance. These results highlight TaxoAdapt's capabilities as a scalable, multidimensional, and dynamically adaptive method for organizing scientific knowledge in rapidly evolving domains.

7 Limitations

TaxoAdapt relies on LLMs to classify papers into specific dimensions. Although existing works have shown the success of LLMs on fine-grained classification, this classification relies on the parametric knowledge of LLMs, which could be a limitation when LLMs' knowledge becomes outdated. For example, when a dataset paper proposes a new benchmark that has the same (or similar) name as an existing methodology, LLMs might incorrectly assign it to the methodology dimension. However, this is a rare edge case, and TaxoAdapt already generates more dimension-specific taxonomies than baselines as discussed above.

Moreover, although we show comprehensive experiments on corpus across various computer science conferences, it would be a nice extension to run TaxoAdapt on corpus outside of the computer science domain such as healthcare and chemistry.

References

Karan Aggarwal, Maad M Mijwil, Abdel-Hameed Al-Mistarehi, Safwan Alomari, Murat Gök, Anas M Zein Alaabdin, Safaa H Abdulrhman, et al. 2022. Has the future started? the current growth of artificial intelligence, machine learning, and deep learning. *Iraqi Journal for Computer Science and Mathematics*, 3(1):115–123.

Georgeta Bordea, Els Lefever, and Paul Buitelaar. 2016. Semeval-2016 task 13: Taxonomy extraction evaluation (texeval-2). In *Proceedings of the 10th international workshop on semantic evaluation (semeval-2016)*, pages 1081–1091.

Lutz Bornmann, Robin Haunschild, and Rüdiger Mutz. 2021. Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*, 8(1):1–15.

Boqi Chen, Fandi Yi, and Dániel Varró. 2023. Prompting or fine-tuning? a comparative study of large language models for taxonomy construction. In *2023 ACM/IEEE International Conference on Model Driven Engineering Languages and Systems Companion (MODELS-C)*, pages 588–596. IEEE.

Michael Gunn, Dohyun Park, and Nidhish Kamath. 2024. Creating a fine grained entity type taxonomy using llms. *Preprint*, arXiv:2402.12557.

Jiaxin Huang, Yiqing Xie, Yu Meng, Yunyi Zhang, and Jiawei Han. 2020. Corel: Seed-guided topical taxonomy construction by concept learning and relation transferring. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1928–1936.

David Jurgens and Mohammad Taher Pilehvar. 2016. Semeval-2016 task 14: Semantic taxonomy enrichment. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 1092–1102.

Dongha Lee, Jiaming Shen, Seongku Kang, Susik Yoon, Jiawei Han, and Hwanjo Yu. 2022a. Taxocom: Topic taxonomy completion with hierarchical discovery of novel topic clusters. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 2819–2829, New York, NY, USA. Association for Computing Machinery.

Dongha Lee, Jiaming Shen, Seonghyeon Lee, Susik Yoon, Hwanjo Yu, and Jiawei Han. 2022b. Topic taxonomy expansion via hierarchy-aware topic phrase generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1687–1700, Abu Dhabi, United Arab Emirates. 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*, pages 1433–1441.

Yuyin Lu, Hegang Chen, Pengbo Mao, Yanghui Rao, Haoran Xie, Fu Lee Wang, and Qing Li. 2024. Self-supervised topic taxonomy discovery in the box embedding space. *Transactions of the Association for Computational Linguistics*, 12:1401–1416.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Jingbo Shang, Xinyang Zhang, Liyuan Liu, Sha Li, and Jiawei Han. 2020. Nettaxo: Automated topic taxonomy construction from text-rich network. In *Proceedings of the web conference 2020*, pages 1908–1919.

Jiaming Shen, Zeqiu Wu, Dongming Lei, Chao Zhang, Xiang Ren, Michelle T Vanni, Brian M Sadler, and Jiawei Han. 2018. Hiexpan: Task-guided taxonomy construction by hierarchical tree expansion. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2180–2189.

Yanzhen Shen, Yu Zhang, Yunyi Zhang, and Jiawei Han. 2024. A unified taxonomy-guided instruction tuning framework for entity set expansion and taxonomy expansion. *arXiv preprint arXiv:2402.13405*.

Chakresh Kumar Singh, Emma Barme, Robert Ward, Liubov Tupikina, and Marc Santolini. 2022. Quantifying the rise and fall of scientific fields. *PloS one*, 17(6):e0270131.

Yushi Sun, Hao Xin, Kai Sun, Yifan Ethan Xu, Xiao Yang, Xin Luna Dong, Nan Tang, and Lei Chen. 2024. Are large language models a good replacement of taxonomies? *arXiv preprint arXiv:2406.11131*.

Vijay Viswanathan, Kiril Gashteovski, Carolin Lawrence, Tongshuang Wu, and Graham Neubig. 2023. Large language models enable few-shot clustering. *arXiv preprint arXiv:2307.00524*.

Hui Yang, Alistair Willis, David Morse, and Anne de Roeck. 2013. Literature-driven curation for taxonomic name databases. In *Proceedings of the Joint Workshop on NLP&LOD and SWAIE: Semantic Web, Linked Open Data and Information Extraction*, pages 25–32, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

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*, pages 3093–3102.

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*, pages 2701–2709.

Yazhou Zhang, Mengyao Wang, Chenyu Ren, Qiuchi Li, Prayag Tiwari, Benyou Wang, and Jing Qin. 2024a. Pushing the limit of llm capacity for text classification. *arXiv preprint arXiv:2402.07470*.

Yunyi Zhang, Ruozhen Yang, Xueqiang Xu, Rui Li, Jinfeng Xiao, Jiaming Shen, and Jiawei Han. 2024b. Teleclass: Taxonomy enrichment and llm-enhanced hierarchical text classification with minimal supervision. *arXiv preprint arXiv:2403.00165*.

Yuwei Zhang, Zihan Wang, and Jingbo Shang. 2023. Clusterllm: Large language models as a guide for text clustering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13903–13920.

A Baselines

Our primary motivation for TaxoAdapt is to demonstrate its capabilities of aligning the LLM-based taxonomy construction to a specialized, multidimensional corpus. Consequently, we choose to compare our method with both *corpus-driven* and *LLM-based* approaches. Note that all LLM-based baselines utilize `GPT-4o-mini` as their underlying model.

1. ***LLM-Only*** $\rightarrow$ **Chain-of-Layer (Zeng et al., 2024):** A method which is provided a set of entities and solely relies on an LLM (***no corpus***) to select relevant candidate entities for each taxonomy layer and gradually build the taxonomy from top to bottom. We adapt this method to use an LLM to suggest entities based on the root topic t and dimension d.

2. ***LLM + Corpus*** $\rightarrow$ **Prompting-Based:** Given that no methods currently exist which guide LLM taxonomy construction based on a corpus, we design our own prompting-based baseline. Specifically, we conduct an iterative process, where we first ask the LLM to identify relevant papers to the dimension, relevant child nodes, and their corresponding papers. We continue this process until the maximum depth is reached.

3. ***Corpus-Only*** $\rightarrow$ **TaxoCom (Lee et al., 2022a):** A corpus-driven taxonomy completion framework that clusters terms from the input corpus to recursively expand a handcrafted seed taxonomy. We use the same single-level taxonomy from Section 3.1.2 as the seed input, but modify the label names to similar concepts if they do not already exist within the corpus.

B Dimension Type Definitions

We define each of our selected dimensions below:

- **Task:** We assume all papers are associated with a task.
- **Methodology:** A paper that ***introduces, explains, or refines a method or approach***, providing theoretical foundations, implementation details, and empirical evaluations to advance the state-of-the-art or solve specific problems.
- **Datasets:** ***Introduces a new dataset***, detailing its creation, structure, and intended use, while providing analysis or benchmarks to demonstrate its relevance and utility. It focuses on advancing research by addressing gaps in existing datasets/performance of SOTA models or enabling new applications in the field.
- **Evaluation Methods:** A paper that ***assesses the performance, limitations, or biases of models, methods, or datasets*** using systematic experiments or analyses. It focuses on benchmarking, comparative studies, or proposing new evaluation metrics or frameworks to provide insights and improve understanding in the field.
- **Real-World Domains:** A paper which demonstrates the use of techniques to solve specific, ***real-world problems or address specific domain challenges***. It focuses on practical implementation, impact, and insights gained from applying methods in various contexts. Examples include: product recommendation systems, medical record summarization, etc.

C LLM-Human Agreement Analysis

Since our automatic evaluation suite is mainly using `GPT-4o` and `GPT-4o-mini`, we conduct a small-scale human evaluation to test the reliability of our metrics. Using EMNLP'24, one human evaluator is responsible for validating the LLMs evaluation output on the task dimension of TaxoAdapt's taxonomy. We show the consensus percentage (the percentage of cases where both the LLM and the human evaluator agree on an instance) on path granularity, sibling coherence, and dimension alignment metrics as defined in Section 4.3. For path granularity, we select 30 random paths from TaxoAdapt's taxonomy and let the human evaluator make independent judgment about the hierarchical relationships between entities (scored 0 or 1 by the evaluator). Similarly, we select 10 random sets of siblings with respect to parent nodes for the evaluator to judge sibling coherence (scored 0.67 or 1 by the evaluator for reasonable or strongest coherence), and 30 random nodes are studied about their alignment to the task dimension (scored 0 or 1 by the evaluator). As for (node-wise) paper relevance and (level-wise) coverage metrics, since they are about evaluating node-paper relevance, we randomly select 16 node-paper pairs (8 pairs are considered relevant while the other 8 are considered irrelevant by `GPT-4o-mini`) for the evaluator to judge relevance in order to validate these two metrics.

Consensus percentage is shown in Table 4. The agreement percentages between the LLMs and the human evaluator range from 70% to 90%, indicating strong overall agreement. Thus, this human evaluation reinforces the validity of our metrics, so we decide to use them as our automatic evaluation metrics.

D Top-Down Taxonomy Expansion Algorithm

We include the high-level algorithm of TAXOADAPT in Algorithm 1.

Algorithm 1 Top-Down Taxonomy Expansion

Require: Topic t, Dimension $d \in D$, Corpus P, density_thresh = δ, max_depth=l

$T_d \in T$ = initialize_taxonomy(t, D) {T.depth $= 0$}
$P_d \subseteq P \leftarrow$ multi_dim_class(t, D) {Section 3.2}
q = queue$(\forall T_d \in T)$
while $len(q) > 0$ and T.depth $\leq l$ **do**
  $n_{i,d} \leftarrow pop(q)$
  if isLeaf$(n_{i,d})$ **then**
    $n^i_{j,d} \in N^i_d \leftarrow$ expand_depth$(n_{i,d}, t)$ {Section 3.3.2}
    q.append$(n_{i,d})$
  else
    classify_children$(n_{i,d}, t, d)$ {Section 3.3.1}
    if $\tilde{\rho}(n_{i,d}) > \delta$ **then**
      $n'^i_{j,d} \in N'^i_d \leftarrow$ expand_width$(n_{i,d}, t)$ {Section 3.3.2}
      if $|N'^i_d| > 0$ **then**
        classify_children$(n_{i,d}, t, d)$
      for $n^i_{j,d} \in N^i_d$ **do**
        if $n^i_{j,d}$.level $< l$ **and** $\rho(n^i_{j,d}) > \delta$ **then**
          q.append$(n^i_{j,d})$
return T

E LLM Evaluation Prompts

As described in Section 4.3, we show the LLM prompt that we use to generate evaluation output for computing automatic metrics in Figure 4.

Table 4: Consensus percentages of path granularity, sibling coherence, dimension alignment, and node-paper relevance between LLMs and the human evaluator.

Granularity	Coherence	Alignment	Relevance
0.900	0.700	0.700	0.875

```
path_granularity = (
            "Scientific concepts are naturally organized in multi-dimensional taxonomic structures, with more specific concepts being the children of a
broader research topic.\n\n"
            f"Given the root topic: '{root}', decide whether this path from the scientific concept taxonomy has good granularity: '{path}' Check whether the
child node is a more specific subaspect of the parent node. \n\n"
            "Output options: '<good granularity>' or '<bad granularity>'. Do some simple rationalization before giving the output if possible."
        )


sibling_coherence = (
            f"You are determining the coherence of a set of {root} subtopics of the parent topic {parent}.\n\nThe parent topic is: {parent}.\n\nThe set of
siblings, which are child subtopics of the parent, are: '{', '.join(siblings)}'\n\nEvaluate the overall coherence of the sibling set based on their
collective specificity and granularity relative to {parent}. Use the following scoring criteria:\n\n"
            f"Score=<no_sibling_coherence>: The set is highly inconsistent or incoherent (only one subtopic), with most topics significantly misaligned in
specificity relative to the parent.\n"
            f"Score=<weak_sibling_coherence>: The set shows considerable inconsistency, with several topics deviating noticeably from the expected level of
specificity.\n"
            f"Score=<reasonable_sibling_coherence>: The set is generally coherent, with only minor inconsistencies in specificity among the topics.\n"
            f"Score=<strong_sibling_coherence>: The set is fully coherent, with all topics properly matching the expected level of specificity and
granularity for the parent.\n"
            "Output options: '<no_sibling_coherence>', '<weak_sibling_coherence>', '<reasonable_sibling_coherence>', or '<strong_sibling_coherence>'. Do some
simple rationalization before giving the output if possible."
        )

dimension_alignment = (
            "Scientific concepts are naturally organized in multi-dimensional taxonomic structures, with more specific concepts being the children of a
broader research topic.\n\n"
            f"Given the root topic: '{root}', decide whether this node from a taxonomy is relevant to the {dim} aspect of the root topic: '{node}'\n\n"
            "Output options: '<relevant>' or '<irerelevant>'. Do some simple rationalization before giving the output if possible."
        )

paper_relevance = ("Scientific concepts are naturally organized in a multi-dimensional taxonomic structure, with more specific concepts being the children of
a broader research topic.\n\n"
                f"Given the root topic: '{root}', here is one of its subtopics: {node_name} and these are some papers: {index_papers}\n\n"
                "Provide a list of paper IDs that are relevant to this subtopic.\n\n"
                "Output options: '<rel_paper> ID1, ID2, ... </rel_paper>'. Do some rationalization before outputting the list of relevant paper IDs.")
```

Figure 4: LLM evaluation prompts used to compute path granularity, sibling coherence, dimension alignment, paper relevance, and coverage.