SCIENCE HIERARCHOGRAPHY: Hierarchical Organization of Science Literature

Anonymous ACL submission

Abstract

Scientific knowledge is growing rapidly, making it difficult to track progress and high-level conceptual links across broad disciplines. While tools like citation networks and search engines help retrieve related papers, they lack the abstraction needed to capture the *density* and structure of activity across subfields.

We motivate SCIENCE HIERARCHOGRAPHY, the goal of organizing broad swaths of scientific literature into a high-quality hierarchical structure that spans multiple levels of abstraction—from broad domains to specific studies. Such a representation can provide insights into which fields are well-explored and which are under-explored. To achieve this goal, we develop a hybrid approach that combines efficient embedding-based clustering with LLM-based prompting, striking a balance between *scalability* and *semantic precision*. Compared to LLM-heavy methods like iterative tree construction, our approach achieves superior quality-speed trade-offs. Our hierarchies capture different dimensions of research contributions, reflecting the interdisciplinary and multifaceted nature of modern science. We evaluate its utility by measuring how effectively an LLM-based agent can navigate the hierarchy to locate target papers. Results show that our method improves interpretability and offers an alternative pathway for exploring scientific literature beyond traditional search methods.

1 Introduction

The pace of scientific publishing is accelerating (Ware and Mabe, 2015), but this growth is uneven across fields (Hope et al., 2023). Some areas attract dense research activity, while others remain underexplored. This raises a natural question:

How do we understand the distribution of scientific efforts across different sub-areas?

Answering this question is essential *for both academic and policy stakeholders*. A clearer view of how research efforts are distributed enables institutions to spot emerging or neglected areas, prioritize strategic hiring and future agendas. For policymakers, it supports more informed funding decisions, ensuring that critical but underexplored domains receive the attention and resources they deserve.

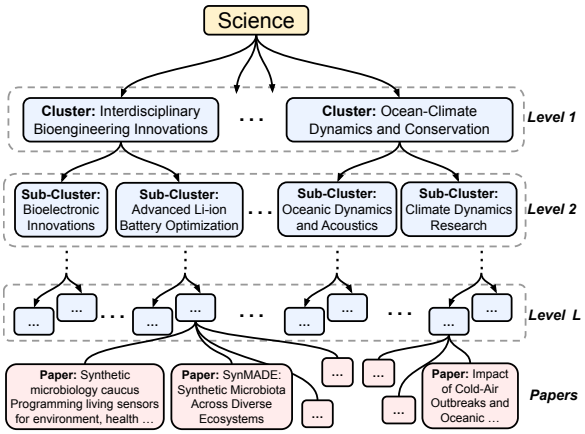


Figure 1: An example of SCIENCE HIERARCHOGRAPHY illustrates how scholarly work can be organized hierarchically—from broad research domains at the top, through increasingly specific sub-clusters, down to individual papers at the lowest level. Critically, this structure must be inferred automatically and at scale.

Conventional tools like Google Scholar are designed as retrieval engines, optimized to return a handful of papers that match a specific query. They offer little in the way of a comprehensive or structured view of the broader scientific landscape. Similarly, while modern LLM-based assistants can surface related works (seen during pretraining or via their retrieval tools), they fall short in offering a broad, bird's-eye perspective on scientific progress.

Addressing this challenge requires abstraction: a way to generalize over research problems and techniques and to connect broad scientific areas to specific papers via intermediate categories. At one end, we have high-level domains (e.g., physics, AI); at the other, individual papers. Between them lie a latent spectrum of subfields and methodological clusters. What's missing is a data structure that captures all these abstraction levels.

We propose building large-scale hierarchical representations of scientific literature, which we call SCIENCE HIERARCHOGRAPHY. A well-designed hierarchy provides a macro-level view of scientific progress, revealing how research is distributed across methods and application areas. This helps researchers spot emerging trends and gaps, and supports policymakers and institutions in making more strategic resource decisions. It also offers a new way to explore the literature—complementing traditional search by allowing users to navigate science through conceptual hierarchies.

How should scholarly work be represented? A central challenge in building a scientific hierarchy is defining what each node represents. Research papers often span multiple topics (e.g., *reinforcement learning for medical imaging* or *deep learning for oceanography*). To capture this complexity, we develop a prompting strategy that decomposes papers into key *contribution* types—such as the *problems addressed* and *techniques used* (§3.2). For each fixed contribution type, we construct a corresponding hierarchical structure, ensuring that papers are organized into meaningful, coherent categories.

What construction strategies balance scalability and quality? To address this, we introduce SCYCHIC (pronounced "psychic"), a new method for building high-quality hierarchical structures of scientific literature. SCYCHIC integrates fast embedding-based clustering with LLM prompting, combining the efficiency of embeddings with the semantic precision of language models (§4.1).

How can we evaluate the quality of a scientific hierarchy? Scientific hierarchies lack a fixed ground truth—they evolve over time as research landscapes shift. We therefore adopt an *evaluation-through-utilization* approach, measuring *whether an information seeker (human or AI) can efficiently locate specific content* (e.g., child nodes) by navigating the hierarchy from the root. This evaluation hinges on the idea that a good hierarchy enables rapid information discovery, even though its utility extends well beyond search alone (§5.2).

What did our empirical results show? Our approach achieves the best trade-off between quality and speed when compared to LLM-heavy methods like iterative tree construction or pruning. Extensive experiments show that SCYCHIC consistently produces higher-quality hierarchies than a broad set of baselines (§5.4). Validation on a 10K-paper dataset further confirms its strong accuracy and scalability for large-scale use.

Contributions: (1) We introduce the goal of constructing large-scale, abstract hierarchies of scientific literature to reveal how scholarly efforts are distributed across research areas. (2) We propose a utilization-based evaluation framework that measures how effectively users can discover information by traversing the hierarchy. (3) We present SCYCHIC, a new method that combines fast embedding-based clustering with LLM prompting to build high-quality, multidimensional hierarchies. Extensive experiments show that SCYCHIC outperforms baseline approaches, offering a more structured and bird's-eye view of scientific progress.

2 Related Work

Taxonomy induction: The field of taxonomy induction has progressed from early pattern-based techniques to modern LLM-augmented methods. Seminal work by Hearst (1992) introduced the use of hand-crafted hyponym patterns for extracting is-a relationships. Subsequent research expanded on this using statistical methods and large-scale information extraction to identify hypernym-hyponym structures (Pantel and Pennacchiotti, 2006; Yang and Callan, 2009; Girju et al., 2006).

Recent advances incorporate LLMs prompting to enhance taxonomy construction. For example, Wan et al. (2024); Zeng et al. (2024a); Chen et al. (2023); Zeng et al. (2024b) apply zero-/few-shot reasoning and ensemble ranking, while others explore open-ended, vocabulary-free taxonomy creation (Gunn et al., 2024), self-supervised expansion in low-resource domains (Mishra et al., 2024), and graph-based methods leveraging metadata and citations (Cong et al., 2024; Sas and Capiluppi, 2024; Shen et al., 2024). Optimization and in-context learning have also shown promise (Hu et al., 2024b; Shi et al., 2024; Xu et al., 2025; Jain and Espinosa Anke, 2022; Chen et al., 2021).

Our work differs in scope, scale, and methodological design. TaxoGen (Zhang et al., 2018) extracts term-level hierarchies by clustering frequent noun phrases from text corpora; however, when applied to academic abstracts, it tends to surface generic terms (e.g., "new method," "novel framework") rather than meaningful concepts suitable for organizing scholarly work. Hu et al. (2025) constructs topic-level concept hierarchies from citation graphs, requiring supervised training on taxonomies extracted from existing literature reviews, making it inapplicable to organizing arbitrary pa-

System	# of Levels	Node content	Node granularity	Assigned by	Purpose	Public
Web of Science	One	Research areas	One keyword	Editor	Indexing	No
Scopus	Three	Research areas	One keyword	Editor	Indexing	Yes
arXiv Taxonomy	Two	Research areas	One keyword	Authors	Indexing	Yes
PubMed MeSH	Multiple	Medical headings	One keyword	Indexer	Indexing	Yes
OpenAlex	Four	Research areas	Multiple keyword	Algorithms	Indexing	Yes
Microsoft Academic Graph	Multiple	Research areas	Multiple keywords	Algorithms	Indexing	Discontinued
SCIENCE HIERARCHOGRAPHY (Ours)	Multiple (by designer)	Rich contribution descriptions	Science contribution summary (many tokens)	Algorithms	Exploratory Analysis	Yes

Table 1: Comparison of hierarchical resources for organizing scientific literature, ordered by hierarchy depth. Conventional systems are built for indexing, relying on fixed, shallow taxonomies with keyword-based nodes and human-assigned labels. In contrast, SCIENCE HIERARCHOGRAPHY supports deeper, designer-controlled hierarchies with rich natural-language summaries, enabling more flexible and exploratory analysis of scientific work.

per collections without pre-existing review structures. In contrast, we focus on scaling taxonomy induction for the domain of scholarly literature—a setting that presents greater challenges than typical setups (e.g., entity hierarchy) due to the complexity, size, and evolving nature of scientific content.

The closest works are Oarga et al. (2024), which build domain-specific hierarchies (e.g., Chemical) using iterative LLM refinement, and Zhu et al. (2025), which organizes survey-based collections of fewer than 100 papers. Our objectives require fundamentally different algorithmic strategies and operate without access to ground truth labels.

Structured representation of science: As science grows at an unprecedented rate (Teufel et al., 1999; Pertsas and Constantopoulos, 2017; Constantin et al., 2016; Fisas et al., 2016; Liakata et al., 2010), numerous frameworks have emerged to structure this information through knowledge graphs and taxonomies (Fathalla et al., 2017; Jaradeh et al., 2019; Oelen et al., 2020; Vogt et al., 2020; Soldatova and King, 2006). Recent work includes prompt-based topic modeling (Pham et al., 2024), iterative taxonomy construction that incorporates object properties and graph mining (Cui et al., 2024; Marchenko and Dvoichenkov, 2024), and hybrid approaches that combine curated ontologies with data-driven maps (Zimmermann et al., 2024). Our work builds on these efforts by constructing a high-quality hierarchical structure tailored to scientific literature, in three key ways. The prior work: (1) Produces shallow hierarchies, typically only one or two levels deep; (2) Uses cluster labels based on keywords, whereas ours are derived from natural language summaries of papers; (3) Depends heavily on manual effort, while our pipeline is fully automated.

In Table 1 we summarize the differences with existing hierarchical resources. Most prior systems are limited to fixed depth and rely on manually assigned labels for indexing—a process often prone to bias (Hadfield, 2020). For example, Scopus employs a hierarchy (ASJC codes) assigned at the *journal* level, so paper-level classifications are inherited rather than content-derived. In contrast, our approach supports deeper, algorithmically generated hierarchies with semantically rich node descriptions. This enables a more flexible representation of scientific knowledge. We provide a detailed comparison with existing systems in §A.

3 SCIENCE HIERARCHOGRAPHY: Towards Hierarchy of Scholarly Work

We first define the problem (§3.1), then discuss representations of scientific contributions (§3.2) and hierarchy depth (§3.3).

3.1 Formal Problem Statement

We define the task of SCIENCE HIERARCHOGRAPHY as an inference problem where the **input** is a large set of scientific papers: $P = \{p_1, p_2, \ldots, p_n\}$. The goal is to infer a *hierarchical structure* (i.e., a tree) for a specific contribution type (e.g., problem statement) of a collection of papers. The nodes of this tree are the atomic concepts representing scholarly ideas or goals. The edge (relations connecting two nodes) encode whether one node is a specific version of another node (i.e., "isA" relationships) which defines a hierarchical link between node pairs, indicating a child node is a subclass of its more abstract parent node (e.g., "RLHF isA RL" means "RLHF" is a type of "RL"). The specific papers P are the nodes of this tree. The overall hierarchy represents levels of specificity and abstraction, with nodes closer to the root representing broader topics. Broader topics are at the upper levels, while more specific subtopics and individual papers are at the lower levels.

Why a tree structure? A tree offers a clear, interpretable way to capture hierarchical relations among scientific ideas, showing how concepts specialize or generalize without cycles or ambiguity. Each node inherits meaning from its ancestors, tracing the progression from broad themes to concrete contributions. We build *one tree per contribution type* (e.g., problem, method; §3.2), forming a **forest of hierarchies** that reflects the multi-faceted and interdisciplinary nature of scientific research.

3.2 Decomposing Papers to Contributions

A central challenge is how to represent the content of scholarly work within hierarchy nodes. Scientific papers are idea-dense, often combining broad goals, specific problems, and technical methods. To capture this complexity, we extract structured representations that disentangle these distinct aspects (D'Souza and Auer, 2020). This also mitigates the issue of input length: papers typically range from 4 to 10 pages (5K to 10K tokens), making full-document processing across large corpora infeasible and costly for LLMs.

We use an LLM (`gpt-4o`) to preprocess each paper (title and abstract), use the paper's title and abstract as the input, and request the LLM to break them down into a **pre-defined set of contributions**, akin to prior work (Hope et al., 2017; Chan et al., 2018) that mines "problem schema" from existing documents. We consider the following contribution types: (1) **problem statement** (the problem addressed), (2) **solution** (the technical approach used), (3) **result** (the key finding), and (4) **topic** (the overarching themes). (See §C for prompts and examples). We note that each contribution may include additional dimensions (sub-contributions). For instance, a "result" encompasses both the "outcome" and its "potential impact." In total, this yields $C = 11$ sub-contributions per paper.

3.3 Choosing a Hierarchy Depth

While the ideal number of hierarchy layers is ultimately empirical, we can build useful intuition from the structure of a near-balanced tree. For a tree with branching factor b and depth L, the total number of nodes is roughly $O(b^L)$. To organize C contributions, the number of nodes should scale with C, implying a depth of $L = O(\log_b C)$. In practice, we use $L = 3$ for a 2K-paper corpus and $L = 4$ for 10K papers, consistent with this logarithmic scaling. Extrapolating further, corpora of 10^7 papers would likely require depths of $L = 6$ or 7.

4 Tackling SCIENCE HIERARCHOGRAPHY

We present algorithms to address our proposed goal. We start with our main method, SCYCHIC (§4.1), explore its special cases (§4.2), and then describe alternative baselines that rely more heavily on LLMs (FLMSCI; §4.3). While all approaches leverage LLMs to some extent, they differ significantly in their reliance on them: some require many calls (linear or quadratic in the number of papers), while others are more efficient (e.g., logarithmic). Since our goal is to scale to a large number of papers, minimizing LLM usage is critical. Our objective is to identify a method that yields the highest-quality hierarchy with the lowest LLM overhead, balancing quality, latency, and cost.

Algorithm 1 SCYCHIC algorithm

Require: Set of papers $P = \{p_1, p_2, \ldots, p_n\}$, `embedder`, `clusterer`, `summarizer`, num of layers L, target cluster sizes $(k_1, k_2, \ldots, k_L)$

Initialization: For each paper $p_i \in P$, using `embedder` embed their selected components to form $\mathbb{R}^{d \times |C'|}$.
for layer $l = 1$ to $\lfloor L/2 \rfloor$ **do** ▷ Top-down phase
  if $l = 1$ **then**
    Apply `clusterer` to divide papers into k_1 clusters
  else
    for each cluster from layer $l - 1$ **do**
      Apply `clusterer` to divide into subclusters
  Use `summarizer` to generate summaries for clusters
for each cluster τ at level $\lfloor L/2 \rfloor$ **do** ▷ Bottom-up phase
  for layer $l = L$ to $\lfloor L/2 \rfloor + 1$ **do**
    if $l = L$ **then**
      Collect the embeddings of papers within τ.
    else
      Apply `embedder` on summaries of cluster $l+1$
    Apply `clusterer` to form higher-level clusters
    Use `summarizer` to generate summaries
return Hierarchical structure

4.1 SCYCHIC: Alternating Between Clustering and Summarization

Overview: Our method builds each contribution-type hierarchy through two complementary stages: a *top-down* phase that clusters paper embeddings into progressively finer subgroups and summarizes each cluster, followed by a *bottom-up* phase that embeds and reclusters the generated summaries to form higher-level abstractions. The combined process yields coherent, interpretable hierarchies that capture both fine-grained and global structure.
Ingredients: This approach is based on the following design choices: (1) access to `embedder`, a neural model that converts a description into a d-dimensional vector, (ideally) capturing its semantic meaning; (2) a clustering algorithm `clusterer` that, given the hyperparameter k, generates k clus-

ters; (3) a contribution type (e.g., problem definition) and its dimensions C' extracted per paper as detailed in §3.2 which determines the focus of the node descriptions; (4) `summarizer`, an LLM that generates a summary description which (ideally) provides a more abstract description of a collection of node descriptions; and (5) the total number of hierarchy layers L and target number of clusters in each layer $(k_1, k_2, \ldots, k_L)$.

Specifically, for `embedder` we use `gte-Qwen2-7B-instruct`, for our `summarizer` we use Llama-3.3-70B-Instruct (Grattafiori et al., 2024), and for `clusterer`, we apply k-means clustering. (further details in §G.)

Initialization: The approach begins by embedding each paper. For each paper p_i, we embed each component in C': $\texttt{embedder}(c_j^i) \in \mathbb{R}^d$, where $j \in C'$. This process results in $|C'|$ embeddings per paper. We concatenate these embeddings, yielding $\mathbb{R}^{d.|C'|}$ embeddings per paper. We now present the main algorithm consisting of two phases:

Phase 1: Top-down: We begin with a **top-down** strategy that recursively partitions the paper set through the upper half of the hierarchy ($l \in [1, \lfloor L/2 \rfloor]$). At the first level, all papers are clustered into k_1 groups using their embeddings. Each cluster is then processed independently—papers within a cluster are reclustered using `clusterer` to form finer subgroups. The number of subclusters assigned to each parent cluster scales linearly with its paper count, ensuring denser regions of the corpus receive finer resolution. This recursive subdivision continues until level $\lfloor L/2 \rfloor$, producing a coarse-to-fine hierarchy. At each level, `summarizer` to generate abstracted summaries for each of the clusters based on the clustered papers' titles and abstracts. The generated cluster description follows the same structure or style as the input descriptions. For example, if the inputs are statements about problem categories, the summaries are also in the same style, but more abstract.

Phase 2: Bottom-up: We switch to a **bottom-up** strategy to construct the remaining levels ($\lfloor L/2 \rfloor + 1$ through L). To form clusters for bottom-level (layer L), we apply `clusterer` to the paper embeddings within each sub-cluster within level-$\lfloor L/2 \rfloor$ (the lowest level clustering obtained from the top-down approach). We then use the `summarizer` to create an abstracted description for each cluster. We repeat this process for all layers from L to $\lfloor L/2 \rfloor + 1$. To build layer l, we start by embedding the generated cluster summaries from the level below $l-1$ using `embedder`, similar to how we embedded the papers. We then run `clusterer` on these new embeddings and generate abstracted summaries for the clusters to group these summaries into higher-level clusters. This bottom-up aggregation continues until we connect with the previously constructed level $\lfloor L/2 \rfloor$ clusters.

Rationale behind the hybrid design: The hybrid approach merges the strengths of top-down and bottom-up strategies. A bottom-up method may create less coherent top-level clusters. The top-down approach ensures high-quality top-level clusters but doesn't utilize the abstracted summaries from `summarizer` used by bottom-up clustering. By combining both methods, the hybrid design achieves robust and effective clustering. Our empirical results in §5.4 demonstrate this approach's strength by balancing quality and scalability.

4.2 Top-down and Bottom-up Baselines

In §5.4, we will examine two special cases of SCYCHIC: (1) only a top-down strategy and (2) only a bottom-up approach. These variants help evaluate the strengths and limitations of each method.

4.3 Pure LLM-based Baselines

We introduce baselines that heavily utilize LLM calls, based on the hypothesis that LLMs can make high-quality local decisions, collectively forming a robust global structure. The potential cost here is the need to make *many* LLM calls. We refer to these baselines as FLMSCI (pronounced "flimsy") and present two variants below. For both methods, we use `gpt-4o` to extract the contributions (§3.2), and `Llama-3.3-70B-Instruct` to place them into the hierarchy.

Initializing a Seed Hierarchy: The first step involves creating a seed hierarchy, starting with the hierarchy of sciences from the Wikipedia page on branches of science[1] and refined through several adjustments detailed in §D.

FLMSCI (parallel): parallel addition of contributions: This approach expands the seed hierarchy in parallel using a small number of LLM calls. All unique contributions extracted from papers are first collected and divided into batches of 100 (to fit within the LLM's context window). A multi-threaded program then assigns each batch to a separate thread, where the LLM adds those contributions to a cloned copy of the seed hierar-

[1] en.wikipedia.org/wiki/Branches_of_science

chy. Finally, the cloned hierarchies are merged (via a Python script rather than additional LLM calls) into a single unified structure.

FLMSCI (incremental): Incremental tree expansion: This method builds the hierarchy iteratively by adding one contribution at a time through layer-by-layer prompting. Starting from the root, the model navigates the tree and performs one of four actions: (a) *Go down*: move to a lower-level node; (b) *Add sibling*: insert a new node at the same level; (c) *Make parent*: create a new parent node; or (d) *Discard*: ignore the contribution if no suitable location exists (Fig. 11). Available actions depend on the current position in the tree. To avoid placing detailed contributions too high in the tree, we disable node-creation actions (b, c) above layer 3. When reaching a leaf node, the *Go down* action (a) is also unavailable. Pilot studies revealed frequent early-layer errors due to broad category labels; to mitigate this, we replaced top-level labels with descriptive definitions (Fig. 10), improving contextual understanding and placement accuracy.

4.4 Computational Complexity of Approaches

A major scalability bottleneck in hierarchy construction is the number of LLM calls. Let C be the number of contributions (§3.2), b the branching factor, and $L = O(\log_b C)$ the maximum depth for a near-balanced tree (§3.3). SCYCHIC requires $O(C/b)$ LLM calls for both top-down and bottom-up variants. Among the LLM-based baselines discussed in §4.3, FLMSCI (parallel) makes $O(C/l)$ calls (with l as batch size), offering lower complexity but at the cost of reduced quality. In contrast, FLMSCI (incremental) achieves higher accuracy but requires $O(C \log_b C)$ LLM calls due to root-to-leaf traversals during insertion. Empirically, the difference in LLM usage is significant: in our 2K-paper setup, FLMSCI (incremental) makes 61K calls compared to just 322 for SCYCHIC (Table 4).

Approach	# of LLM calls
SCYCHIC	$O(C/b)$
FLMSCI (parallel)	$O(C/l)$
FLMSCI (incremental)	$O(C \log_b C)$

Table 2: Computational complexity of hierarchy construction methods measured by LLM calls, with C = contributions, b = branching factor, and l = batch size.

5 Experimental Setup and Results

We describe our experimental setup, including the diverse paper collection used for our experiments (§5.1) and the evaluation framework (§5.2).

5.1 Collection of Science Papers

We compile a collection of scientific papers spanning domains such as computer science, neuroscience, biology, oceanography, and their interdisciplinary intersections. Our initial analysis focuses on a smaller set of approximately 2K papers (referred to as **SciPile**), allowing for rapid iteration over design choices and assessment of scalability. We then extend our analysis to a larger collection of $10K$ papers, referred to as **SciPileLarge**. Details on data collection and filtering are provided in §F.

5.2 Evaluation as Utilization

Ideally, hierarchy quality would be evaluated against a gold standard—but no such reference exists, and scientific literature continually evolves. As a result, we adopt an evaluation framework based on *utilization*, independent of fixed ground truth.

We assess hierarchy quality by measuring *how well it supports navigation and content discovery*. Specifically, we use an LLM-based agent to locate target papers via tree traversal, tracking accuracy at each level and across the full hierarchy. A stronger hierarchy should better capture conceptual relationships and improve information-seeking efficiency. While our evaluation focuses on retrieval, the hierarchy's utility extends beyond that.

Our evaluation design involves two choices: (a) queries and (b) an evaluation model. For (a), we sample paper titles and abstracts. Although we considered generating language questions from papers, pilot studies showed both approaches yield similar results, so we use the simpler method. For (b), we use `Qwen2.5-32b-instruct`, which performed closest to GPT-4 among open models (§B.1).

The process starts at the root: given a query and cluster descriptions (Fig. 2), the LLM selects the most relevant cluster. If it contains the target paper, traversal continues recursively through subclusters until the correct paper-level node is reached. We report two metrics: **Strict-Acc**, the fraction of cases where the model finds the target node, and **L1-Acc**, which measures how often it correctly identifies the top-level subtree containing the target.

Validation: We also validate the reliability of our LLM-based evaluation through both (a) *human*

Method	Accuracy (%)		LLM Cost		Hierarchy Structure		
	Strict-Acc ↑	L1-Acc ↑	Avg. # of Input Tokens ↓	# of Calls ↓	Depth	Avg. Branching Factor	Max. Branching Factor
Contributions type: Problem Statement							
SCYCHIC	**43.7** ± 6.5	85.8 ± 4.2	7451				26
↳Top-down	41.5 ± 8.2	**86.5** ± 5.6	8990	1572	4	8	30
↳Bottom-up	26.2 ± 5.4	41.9 ± 4.0	5924				26
Contributions type: Solution Statement							
SCYCHIC	**24.7** ± 4.8	**65.8** ± 2.5	7653				28
↳Top-down	22.4 ± 3.5	52.3 ± 3.0	4032	1572	4	8	26
↳Bottom-up	23.9 ± 3.3	51.3 ± 3.1	6150				28
Contributions type: Results Statement							
SCYCHIC	**27.6** ± 4.6	**69.8** ± 2.1	6457				30
↳Top-down	19.7 ± 4.0	54.0 ± 3.3	5380	1572	4	8	30
↳Bottom-up	23.6 ± 2.7	55.2 ± 2.9	4731				28

Table 3: Evaluation results of SCYCHIC and the corresponding baselines on the 10K (**SciPileLarge**) dataset. SCYCHIC maintains high accuracy across all contribution types, proving the rationale behind our hybrid design. The *problem statement* contribution type consistently yields the most accurate hierarchies, indicating this contribution type contributes most for hierarchy construction. Results on the 2K (**SciPile**) dataset can be found in §H.

assessment and (b) evaluation based on existing human-annotated hierarchy from ORKG (§B.2).

Method	Strict-Acc (%) ↑	L1-Acc (%) ↑	# of Calls ↓
Topic contributions			
SCYCHIC	**14.9** ± 2.7	**65.7** ± 4.4	322
↳Top-down	14.5 ± 4.7	62.5 ± 7.4	322
↳Bottom-up	13.9 ± 5.3	54.4 ± 12.7	322
↳FLMSCI (par)	4.0 ± 2.8	32.0 ± 6.3	226
↳FLMSCI (inc)	**18.0** ± 5.3	**91.0** ± 4.0	61K

Table 4: Evaluations results for SCYCHIC, FLMSCI (**par**allel) and FLMSCI(**inc**remental) when using *Topic* as the contribution type. All methods exhibit low Strict-Acc ($\leq 18.0\%$), underscoring the difficulty of the task. While FLMSCI (inc) achieves the highest accuracy, it requires approximately 200× more LLM calls than other methods. In contrast, SCYCHIC strikes a balance between performance and efficiency, achieving competitive accuracy (14.9% Strict-Acc, 65.7% L1-Acc) with substantially lower computing cost. Full results in §H.2.

5.3 Experiment Design

We conduct a series of experiments to evaluate our method (SCYCHIC, §4.1) against the baselines (§4.2, 4.3) using the proposed evaluation protocol. Hyperparameter settings for SCYCHIC are detailed in §G. The experiments are organized as follows: (1) We first compare all methods on the simplest contribution type ("topic") in Table 4. Due to the high computational cost, LLM-based baselines are evaluated only in this setting. (2) We then evaluate performance on more complex contributions (problem, solution, and results) using both **SciPile** (Table 10) and **SciPileLarge** (Table 3). Each results table also reports *LLM Cost* (average input tokens and number of calls) and *Hierarchy Structure* (average depth and branching factor).

5.4 Empirical Findings

SCYCHIC outperforms its special-case baselines. As shown in Table 3, SCYCHIC achieves higher Level-1 accuracy than the top-down and bottom-up baselines, while maintaining comparable Strict-Acc. Similar trends hold across other contribution types in both 2K and 10K results , highlighting its effectiveness. Notably, these gains are achieved with a similar number of tokens and LLM calls, underscoring SCYCHIC's compute efficiency.

LLM-based approaches can be expensive. While FLMSCI slightly outperforms SCYCHIC in strict accuracy in Table 4, it does so at the cost of a *massive* increase in LLM calls—making it impractical at scale. As a result, despite its strong performance, FLMSCI (incremental) simply doesn't scale.

SCYCHIC scales to larger paper corpus. For our $10K$ paper dataset **SciPileLarge**, due to the significant increase (×5) in corpus size, we extend the hierarchy to four layers (versus three previously). Notably, SCYCHIC achieved even higher L1-Acc (86.5%) on **SciPileLarge** compared to our smaller dataset **SciPile**. This improvement likely stems from the enhanced quality of our expanded dataset, which has more strict filtering mechanisms. While the Strict-Acc showed a minor decrease compared to results on **SciPile**, it remained at a satisfactory level. Collectively, these results provide compelling evidence that our method scales successfully to substantially larger paper corpora.

SCYCHIC outperforms existing taxonomies and

Method	Accuracy (%)		Hierarchy Structure	
	Strict-Acc	L1-Acc	Avg. BF	Max. BF
Contributions type: Problem Statement				
SCYCHIC	**43.7** ± 6.5	85.8 ± 4.2	8	26
TaxoAdapt	37.9 ± 3.9	82.5 ± 2.6	151.43	1405
Contributions type: Solution Statement				
SCYCHIC	**24.7** ± 4.8	65.8 ± 2.5	8	28
TaxoAdapt	14.3 ± 9.5	41.5 ± 4.6	105.39	732

Table 5: Comparison of SCYCHIC and TaxoAdapt on **SciPileLarge**. BF denotes Branching Factor. SCYCHIC outperforms TaxoAdapt in both settings.

Method	Accuracy (%)		Max. BF
	Strict-Acc	L1-Acc	
SCYCHIC (P)	**43.7** ± 6.5	**85.8** ± 4.2	26
SCYCHIC (S)	24.7 ± 4.8	65.8 ± 2.5	28
SCYCHIC (R)	27.6 ± 4.6	69.8 ± 2.1	30
OpenAlex	20.4 ± 3.2	78.8 ± 7.0	313

Table 6: Comparison of SCYCHIC and OpenAlex on **SciPileLarge**. BF denotes Branching Factor. P, S, R denote *Problem*, *Solution*, and *Results* Statement respectively. A detailed case study in §B.3 further illustrates OpenAlex's limitations in capturing paper semantics.

related methods. We compare our hierarchy against two baselines in Table 5 and 6: OpenAlex, a large-scale 4-level taxonomy, and TaxoAdapt (Kargupta et al., 2025), which iteratively expands LLM-generated taxonomies into DAG-structured hierarchies across multiple dimensions. For OpenAlex, we retrieve existing labels for **SciPileLarge** to form the hierarchy, which achieves only 20.4% Strict Accuracy, lower than SCYCHIC across all attribute types, especially *problem*. For TaxoAdapt, we compare its task and methodologies dimensions with our `problem` and `solution` attributes. Since TaxoAdapt assigns multiple memberships per paper, we count a paper as correct if the evaluator finds any correct path. While TaxoAdapt outperforms the bottom-up baseline, SCYCHIC consistently achieves better performance across settings.

5.5 Additional Analyses

We briefly cover the summary of additional analyses that are omitted from the main text due to space constraints. (Details are presented in the appendix.)

Detailed prompts significantly improve hierarchy quality. To demonstrate this, we compare two prompt types. The first is a "detailed" prompt—carefully curated with comprehensive instructions and reminders, which we use for all main experiments in this paper. The second is a "simplified" prompt containing only the core task description. The results confirm that the detailed prompt substantially outperforms the simplified version across all scenarios. More details are in §H.4.

Embedding quality varies a lot across models. We evaluate three models for `embedder`—Qwen's gte-Qwen2-7B-instruct (Li et al., 2023), OpenAI's text-embedding-3-large, and text-embedding-ada-002. The first two perform similarly, whereas text-embedding-ada-002 produces markedly weaker results. We select gte-Qwen2-7B-instruct for its strong balance of performance and its practical value as an open-weight model for reproducible research. The experimental results are in §H.3.

Quality diagnostics confirm the reliability of the hierarchies. We further analyze cluster coherence by examining citation patterns within and across clusters. Out of 3,056 total citations, 2,587 (84.7%) occur between papers in the same cluster, while the remaining 469 (15.3%) are inter-cluster citations. The visualization and examples of inter-cluster citations can be found in §I.

5.6 Sample Visualization of the Hierarchy

In §J we show a slice of the final hierarchy generated by SCYCHIC on the **SciPileLarge**. The original hierarchy has 4 levels, using *problem* contribution. Due to space constraints, this slice shows only two levels of clusters above the individual papers.

6 Discussion and Conclusion

Future applications: Our work opens several promising directions for future research. One key opportunity is to use the constructed hierarchies as tools for exploratory analysis across scientific domains. They can aid academic institutions and funding bodies in identifying emerging trends and underexplored areas, and can be adapted for domain-specific analyses that capture the unique structure of individual fields. This approach not only deepens our understanding of scientific progress but also provides a new lens for organizing the vast and growing body of scholarly work.

Conclusions: We introduced SCIENCE HIERARCHOGRAPHY, a scalable framework for hierarchical summarization of scientific literature that reveals how research effort is distributed. Our method, SCYCHIC, combines LLMs with efficient algorithms to balance quality and scale. We hope this helps researchers navigate the scientific landscape and support informed resource allocation.

Limitations

Although we evaluated our pipeline on $10K$ papers, this is still far from the true scale of scientific literature. We hope future work will enhance our approach to handle more realistic scales. Additionally, while our evaluation framework shows potential for efficient information discovery, it may have its own weaknesses and biases. Integrating human verification into the assessment process could help ensure the quality and reliability of the inferred hierarchies.

Ethics Statement

In our work, all data and models are accessed via licenses that grant us free and open access for research purposes. Expert annotations are provided by the paper's authors, who have contributed their efforts without compensation. We have not observed any harmful content in either the scholarly papers or the content generated by LLMs. On the other hand, since our resulting hierarchy reflects the distribution of scientific efforts across various fields, it offers a detailed map of where research activity is concentrated and where it is lacking. This nuanced view can guide decision-makers—such as government agencies and academic institutions—in making more informed choices about resource allocation. By highlighting underexplored yet promising areas alongside well-established fields, the hierarchy helps ensure that funding, support, and strategic initiatives are distributed more equitably. Ultimately, this balanced approach can foster innovation and drive progress in areas that might otherwise be overlooked, leading to a more inclusive and socially beneficial advancement of science.

References

Devichand Budagam, Sankalp KJ, Ashutosh Kumar, Vinija Jain, and Aman Chadha. 2024. Hierarchical prompting taxonomy: A universal evaluation framework for large language models.

Joel Chan, Joseph Chee Chang, Tom Hope, Dafna Shahaf, and Aniket Kittur. 2018. Solvent: A mixed initiative system for finding analogies between research papers. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–21.

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)*.

Catherine Chen, Kevin Lin, and Dan Klein. 2021. Constructing taxonomies from pretrained language models. In *Conference of the North American Chapter of the Association for Computational Linguistics* (NAACL).

Tianji Cong, Fatemeh Nargesian, Junjie Xing, and H. V. Jagadish. 2024. Openforge: Probabilistic metadata integration.

Alexandru Constantin, Silvio Peroni, Steve Pettifer, David Shotton, and Fabio Vitali. 2016. The document components ontology (doco). *Semantic web*, 7(2):167–181.

Wentao Cui, Meng Xiao, Ludi Wang, Xuezhi Wang, Yi Du, and Yuanchun Zhou. 2024. Automated taxonomy alignment via large language models: bridging the gap between knowledge domains.

Ingetraut Dahlberg. 1993. Knowledge organization: its scope and possibilities.

Jairo Diaz-Rodriguez. 2025. k-llmmeans: Summaries as centroids for interpretable and scalable llm-based text clustering.

Jennifer D'Souza and Sören Auer. 2020. Nlpcontributions: An annotation scheme for machine reading of scholarly contributions in natural language processing literature.

Said Fathalla, Sahar Vahdati, Sören Auer, and Christoph Lange. 2017. Towards a knowledge graph representing research findings by semantifying survey articles. In *Research and Advanced Technology for Digital Libraries: 21st International Conference on Theory and Practice of Digital Libraries, TPDL 2017, Thessaloniki, Greece, September 18-21, 2017, Proceedings 21*, pages 315–327. Springer.

Beatriz Fisas, Francesco Ronzano, and Horacio Saggion. 2016. A multi-layered annotated corpus of scientific papers. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3081–3088.

Roxana Girju, Adriana Badulescu, and Dan Moldovan. 2006. Automatic discovery of part-whole relations.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Michael Gunn, Dohyun Park, and Nidhish Kamath. 2024. Creating a fine grained entity type taxonomy using llms.

Ruth M. Hadfield. 2020. Delay and bias in PubMed medical subject heading (MeSH®) indexing of respiratory journals. *bioRxiv*. Preprint.

Yuan He, Moy Yuan, Jiaoyan Chen, and Ian Horrocks. 2024. Language models as hierarchy encoders. In *Advances in Neural Information Processing Systems* (NeurIPS).

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *International Conference on Computational Linguistics* (COLING).

Tom Hope, Joel Chan, Aniket Kittur, and Dafna Shahaf. 2017. Accelerating innovation through analogy mining. In *ACM Conference Knowledge Discovery and Data Mining* (KDD), pages 235–243.

Tom Hope, Doug Downey, Daniel S Weld, Oren Etzioni, and Eric Horvitz. 2023. A computational inflection for scientific discovery. *Communications of the ACM*, 66(8):62–73.

Yujia Hu, Tuan-Phong Nguyen, Shrestha Ghosh, and Simon Razniewski. 2024a. Gptkb: Comprehensively materializing factual llm knowledge.

Yuntong Hu, Zhuofeng Li, Zheng Zhang, Chen Ling, Raasikh Kanjiani, Boxin Zhao, and Liang Zhao. 2024b. Hireview: Hierarchical taxonomy-driven automatic literature review generation.

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.

Devansh Jain and Luis Espinosa Anke. 2022. Distilling hypernymy relations from language models: On the effectiveness of zero-shot taxonomy induction. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*.

Mohamad Yaser Jaradeh, Allard Oelen, Kheir Eddine Farfar, Manuel Prinz, Jennifer D'Souza, Gábor Kismihók, Markus Stocker, and Sören Auer. 2019. Open research knowledge graph: next generation infrastructure for semantic scholarly knowledge. In *Proceedings of the 10th international conference on knowledge capture*, pages 243–246.

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.

Uri Katz, Mosh Levy, and Yoav Goldberg. 2024. Knowledge navigator: Llm-guided browsing framework for exploratory search in scientific literature. In *AAAI*.

Michelle S. Lam, Janice Teoh, James A. Landay, Jeffrey Heer, and Michael S. Bernstein. 2024. Concept induction: Analyzing unstructured text with high-level concepts using lloom. In *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI 2024, Honolulu, HI, USA, May 11-16, 2024*, pages 766:1–766:28. ACM.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Maria Liakata, Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2010. Corpora for the conceptualisation and zoning of scientific papers.

Jesús Lovón-Melgarejo, Jose G. Moreno, Romaric Besançon, Olivier Ferret, and Lynda Tamine. 2023. Probing pretrained language models with hierarchy properties.

Oleksandr Marchenko and Danylo Dvoichenkov. 2024. Taxorankconstruct: A novel rank-based iterative approach to taxonomy construction with large language models.

George Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Sahil Mishra, Ujjwal Sudev, and Tanmoy Chakraborty. 2024. Flame: Self-supervised low-resource taxonomy expansion using large language models.

Viktor Moskvoretskii, Alexander Panchenko, and Irina Nikishina. 2024. Are large language models good at lexical semantics? a case of taxonomy learning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*.

Alexandru Oarga, Matthew Hart, Andres M Bran, Magdalena Lederbauer, and Philippe Schwaller. 2024. Scientific knowledge graph and ontology generation using open large language models. In *Neurips 2024 Workshop Foundation Models for Science: Progress, Opportunities, and Challenges*.

Allard Oelen, Mohamad Yaser Jaradeh, Markus Stocker, and Sören Auer. 2020. Generate fair literature surveys with scholarly knowledge graphs. In *Proceedings of the ACM/IEEE joint conference on digital libraries in 2020*, pages 97–106.

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*.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. 2025. The geometry of categorical and hierarchical concepts in large language models. In *International Conference on Learning Representations* (ICLR).

Vayianos Pertsas and Panos Constantopoulos. 2017. Scholarly ontology: modelling scholarly practices. *International Journal on Digital Libraries*, 18:173–190.

Chau Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2024. TopicGPT: A prompt-based topic modeling framework. In *Conference of the North American Chapter of the Association for Computational Linguistics* (NAACL).

Jason Priem, Heather Piwowar, and Richard Orr. 2022. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *Preprint*, arXiv:2205.01833.

Cezar Sas and Andrea Capiluppi. 2024. Automatic bottom-up taxonomy construction: A software application domain study.

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

Jingchuan Shi, Hang Dong, Jiaoyan Chen, Zhe Wu, and Ian Horrocks. 2024. Taxonomy completion via implicit concept insertion. In *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, pages 2159–2169. ACM.

Larisa N Soldatova and Ross D King. 2006. An ontology of scientific experiments. *Journal of the royal society interface*, 3(11):795–803.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Conference on Artificial Intelligence* (AAAI).

Simone Teufel, Jean Carletta, and Marc Moens. 1999. An annotation scheme for discourse-level argumentation in research articles. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 110–117.

Vijay Viswanathan, Kiril Gashteovski, Kiril Gashteovski, Carolin Lawrence, Tongshuang Wu, and Graham Neubig. 2024. Large language models enable few-shot clustering.

Lars Vogt, Jennifer D'Souza, Markus Stocker, and Sören Auer. 2020. Toward representing research contributions in scholarly knowledge graphs using knowledge graph cells. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, pages 107–116.

Mengting Wan, Tara Safavi, Sujay Kumar Jauhar, Yujin Kim, Scott Counts, Jennifer Neville, Siddharth Suri, Chirag Shah, Ryen W. White, Longqi Yang, Reid Andersen, Georg Buscher, Dhruv Joshi, and Nagu Rangan. 2024. Tnt-llm: Text mining at scale with large language models. In *ACM Conference Knowledge Discovery and Data Mining* (KDD).

Zihan Wang, Jingbo Shang, and Ruiqi Zhong. 2023. Goal-driven explainable clustering via language descriptions. In *Conference on Empirical Methods in Natural Language Processing* (EMNLP).

Zilong Wang, Hao Zhang, Chun-Liang Li, Julian Martin Eisenschlos, Vincent Perot, Zifeng Wang, Lesly Miculicich, Yasuhisa Fujii, Jingbo Shang, Chen-Yu Lee, and Tomas Pfister. 2024. Chain-of-table: Evolving tables in the reasoning chain for table understanding. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Mark Ware and Michael Mabe. 2015. The stm report: An overview of scientific and scholarly journal publishing.

Michael Wolfman, Donald Dunagan, Jonathan Brennan, and John Hale. 2024. Hierarchical syntactic structure in human-like language models. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*.

Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Qili Zhu. 2012. Probase: a probabilistic taxonomy for text understanding.

Hongyuan Xu, Yuhang Niu, Yanlong Wen, and Xiaojie Yuan. 2025. Compress and mix: Advancing efficient taxonomy completion with large language models. In *THE WEB CONFERENCE 2025*.

Hui Yang and Jamie Callan. 2009. A metric-based framework for automatic taxonomy induction. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*.

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

Qingkai Zeng, Yuyang Bai, Zhaoxuan Tan, Zhenyu Wu, Shangbin Feng, and Meng Jiang. 2024b. Codetaxo: Enhancing taxonomy expansion with limited examples via code language prompts. *CoRR*, abs/2408.09070.

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. *Preprint*, arXiv:1812.09551.

Yuwei Zhang, Zihan Wang, and Jingbo Shang. 2023. ClusterLLM: Large language models as a guide for text clustering. In *Conference on Empirical Methods in Natural Language Processing* (EMNLP).

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. *arXiv preprint arXiv:2509.19125*.

Johannes Zimmermann, Dariusz Wiktorek, Thomas Meusburger, Miquel Monge-Dalmau, Antonio Fabregat, Alexander Jarasch, Günter Schmidt, Jorge S Reis-Filho, and T Ian Simpson. 2024. The ontoverse: Democratising access to knowledge graph-based data through a cartographic interface.

A Additional Related Work

We include additional related work here because of the space limitation in the main text.

Clustering with LLMs: Recent advances in clustering methodologies augmented by LLMs have demonstrated effective ways to generate interpretable groupings of text. For example, (Viswanathan et al., 2024; Katz et al., 2024) apply few-shot clustering and thematic grouping to partition scientific literature into meaningful subtopics, while (Zhang et al., 2023; Wang et al., 2023) further refine these techniques by aligning clustering outcomes with natural language explanations and user intent. Other recent work iteratively refines cluster representations by replacing cluster centroids or summary points with LLM-generated natural language descriptions and inclusion criteria, thereby inducing more abstract, interpretable concepts over multiple clustering rounds (Lam et al., 2024; Diaz-Rodriguez, 2025). While these approaches improve clustering quality by using LLMs at various stages, they mostly result in flat groupings rather than hierarchical structures. Our approach builds on this by using LLMs to cluster documents and organizing these clusters into a structured hierarchy.

Structured knowledge in LLMs: Prior work has explored how LLMs internalize hierarchical knowledge. For example, (He et al., 2024; Lovón-Melgarejo et al., 2023; Park et al., 2025) extend the linear representation hypothesis to reveal that LLMs encode categorical concepts as polytopes, with hierarchical relationships reflected as orthogonal directions. Other works such as (Wolfman et al., 2024) and (Budagam et al., 2024) examine the benefits of explicit hierarchical syntactic structures and prompting frameworks for guiding LLM performance, while (Moskvoretskii et al., 2024) and (Hu et al., 2024a) focus on constructing and materializing large-scale structured knowledge bases about entities and events. In line with the same aspirations, our work explores the use of hierarchical structures to organize scientific literature.

Structured knowledge representation: Understanding and organizing knowledge is a fundamental pursuit in both artificial and human intelligence (Dahlberg, 1993). Abstraction hierarchies, such as WordNet for lexical semantics (Miller, 1995), ConceptNet for commonsense reasoning (Speer et al., 2017), and Probase for large-scale concept representation (Wu et al., 2012), have proven to be powerful tools for structuring information. Similarly, modern tabular reasoning leverages structured representations to facilitate systematic inference and knowledge retrieval, demonstrating that such structure remains crucial (Wang et al., 2024).

Comparison with existing hierarchical systems: Web of Science maintains a flat (one-level) collection of 250 research fields, which is useful for **categorization**. Given its flat structure, it is not a hierarchical structure. There are no parent-child relationships or summaries connecting broader and narrower concepts. These are best understood as only labels, not nodes in a multi-level taxonomy. Scopus uses a **fixed-depth** (2-layer) hierarchy based on research field names (ASJC Codes). Importantly, these codes are assigned at the journal level rather than to individual papers. Papers inherit classifications from their publishing journals, meaning the hierarchy is not derived from the actual paper content. PubMed MeSH terms provides hierarchical labeling for PubMed publications, but it functions at the level of **keywords (few tokens)** rather than leveraging the full richness of natural language from science papers. Crucially, it is organized around a fixed set of controlled terms rather than the actual semantic content of the papers, limiting its suitability for constructing dynamic or corpus-specific hierarchies. Additionally, because MeSH is manually curated, it introduces indexing delays—papers are only labeled after publication—and is subject to human bias, as noted by (Hadfield, 2020). Microsoft Academic Graph (MAG), though **discontinued** in 2021, offered a rich graph-based structure connecting papers and authors. Its hierarchical classification derived primarily from citation patterns and machine learning clustering rather than semantic paper content, which limited cluster interpretability.

B Evaluation Framework

We provide more context on our evaluation. As discussed in §5.2, we use randomly-sampled papers (title/abstract) as a query. The evaluator LLM goes through the hierarchy, starting from the root node and iteratively selects the relevant nodes to traverse. The prompt for each decision is shown in Fig.2.

Evaluation Framework

You are a scientist expert in taxonomy. Please read the following paper title and abstract.

Your task is to choose the next cluster/topic (while considering the current path) in the taxonomy that has the best chance of containing this paper.

Paper Title: {paper_title}

Paper Abstract: {paper_content}

Current Path: {path_so_far}

Choose from this cluster/topic list (MUST pick one):

{cluster_descriptions}

Required Response Format:

Cluster ID: [EXACT ID from the list] or Topic: [EXACT Category Name from the list]

Figure 2: Prompt used for Evaluation

B.1 Pilot Experiment for Evaluator Choice

One question is, **which LLM should we use for evaluation?** As discussed in §5.2, we chose `Qwen2.5-32b-instruct` for its strong instruction-following capabilities. In pilot experiments, Qwen showed a high consistency against GPT4 score, compared to other open-weight models. Here's a summary of that experiment: We evaluated one of the hierarchies produced by SCYCHIC using different models, including GPT-4. Assuming GPT-4 has the highest accuracy, we sought alternative models with the greatest consistency against it, as frequent evaluations with GPT4 are costly. Fig.7 presents the results. As it can be observed, Llama has the highest agreement, but we suspect bias since the hierarchy was also constructed with Llama. To avoid this, we selected the next best model, Qwen2.5-32b-instruct, for evaluation.

Evaluator LLLM	Agreement with GPT4
GPT-3.5	39.6
GPT4-mini	59.2
Gemma3-24b-it	62.1
Qwen2.5-32b-instruct	66.5
Llama 3.3 70B	72.4

Table 7: Agreement of different evaluator LLMs against GPT4.

B.2 Validation of LLM-based Evaluation

As we discuss in §5.2, to validate our evaluation framework, a Computer Science PhD student analyzes 200 error cases (50 cases per layer). For each case, the annotator determines whether the error comes from the LLM evaluator or from the hierarchy itself. The analysis reveals three types of cases. First, only 9 cases (4.5%) are clear evaluator errors. Second, in 39 cases (18.5%), both the evaluator's choice and the

hierarchy path are reasonable, which is expected for interdisciplinary works. Third, in the remaining 152 cases (77%), the evaluator agrees with the human annotator. These results confirm the reliability of our LLM-based evaluation approach.

To further validate our LLM-based evaluation approach, we downloaded the annotations from the Open Research Knowledge Graph (ORKG). On this website, papers are curated entirely by volunteers who are strongly familiar with the topics of the papers. We use the subset of the ORKG data focused on the Engineering domain. This led to a collection of 4.4K papers that are organized in a 2-layer hierarchy. Treating this data as a high-quality hierarchy, the question is whether our evaluation would assign it a high score. We ran our evaluation experiment with `Qwen2.5-32B-Instruct` as the LLM-as-a-judge. Similar to our setup from the paper, we use paper title/abstract as queries, and require the evaluator to traverse the hierarchy by incrementally making the most appropriate choice between all possible cluster candidates. Our results show that the evaluator model has an accuracy of 83% (i.e., in 83% of the runs it identified the correct paper). This indicates that our evaluation metric is able to assign a high score to a good hierarchy.

B.3 Comparison with OpenAlex taxonomy

To provide independent validation of our hierarchy's quality, we compare against OpenAlex (Priem et al., 2022)'s taxonomy. OpenAlex is a fully open catalog of scholarly works that succeeded Microsoft Academic Graph (MAG) after its discontinuation in 2021. It provides a 4-level hierarchical taxonomy which is widely adopted in researches.

For the same set of papers (**SciPileLarge**, 10K papers), we organize them according to their assigned labels in OpenAlex. Since both hierarchies contain four levels, no additional structural adaptation is required. We apply the same LLM evaluator used in our main experiments to traverse each hierarchy and attempt to locate papers based on the taxonomy structure.

Paper Title	**OpenAlex**	**SCYCHIC**
Subfield-Level (L2) Comparison		
Differentially Fed Dual-Band Base Station Antenna (5G)	Aerospace Engineering	Next-Generation Wireless Communications
Abnormalities in Diffusional Kurtosis Metrics Related to Head Impact	Epidemiology	Neurological Disorders and Brain Injury
New Sea Spray Generation Function for Spume Droplets	Earth-Surface Processes	Oceanic Mixing and Air-Sea Flux
Vortex Flow and Cavitation in Diesel Engines	Electrical Engineering	Multiphase Flow and Combustion
Topic-Level (L3) Comparison		
Visual Transformers and CNNs for Disease Classification	COVID-19 Diagnosis	Deep Learning for Medical Image Analysis
A Cyclone-Centered Perspective on Sea Ice Motion	Arctic Ice Dynamics	Atmospheric Dynamics and Polar Cyclones
Performance of Chip-Scale Optical Frequency Comb Generators	Fiber Laser Tech.	Integrated Photonics for Communications

Table 8: Side-by-side comparison of OpenAlex and SCYCHIC classifications at corresponding hierarchy levels.

C Extracting Paper Contributions

As we discuss in §3.2, below are prompts and examples for extracting different contributions (*problem*, *solution*, *result* and *topics*) from papers' titles and abstracts. we utilize the GPT-4o model (`gpt-4o-2024-08-06`) to generate all contribution extractions along with detailed rationales explaining the extraction decisions.

C.1 Prompt for Extracting *Problem/Solution/Result* Contributions

We use the prompt below to extract contributions from the paper's title and abstract. After finishing the extraction, the three contributions will be saved into the original `json` file. Please see §3.2 for more information.

Contributions Extraction from Paper

Consider the following following paper:

Title: {title}

Abstract: {abstract}

Extract the relevant content of the above abstract into the following JSON structure.
For certain fields that the information is not found in the abstract, leave them empty (empty string).

```
{
    "problem": {
        "overarching problem domain": "".
        "challenges/difficulties": "",
        "research question/goal": "",
        "novelty of the problem": "",
        "knowns or prior work": "",
    },
    "solution": {
        "overarching solution domain": "".
        "solution approach": "",
        "novelty of the solution": "",
        "knowns or prior work": "",
    },
    "results": {
        "findings/results": "",
        "potential impact of the results": "",
    }
}
```

Figure 3: Prompt for extracting *Problem/Solution/Result* contributions

C.2 Prompt for Extracting *Topic* Contributions and Rationales

This section has the prompt of generating topics and rationales from papers given their titles and abstracts. The prompt provides the model with a system role instruction that describes the task, title, and abstract, and also an example to get the specified output format.

Topics and Rationales Generation

You are an experienced scientist who is going to read and review research papers.

Paper Title: {title}
Paper Abstract: {abstract}

Read the above given Title and Abstract for a research paper and
Generate topics that are represented in the given Title and Abstract.
Example output format:

```json
{
 "topics": [
  {
   "topic": "Entity Taxonomy Creation",
   "rationale": "The research focuses on generating a comprehensive entity taxonomy using LLMs."
  },
  {
   "topic": "Iterative Prompting Techniques",
   "rationale": "Highlights the use of iterative prompting to refine entity classifications."
  },
  {
   "topic": "GPT-4 and GPT-4 Turbo",
   "rationale": "Explores the capabilities of these advanced LLMs in taxonomy development."
  },
  {
   "topic": "Information Extraction",
   "rationale": "Demonstrates applications like relation and event argument extraction."
  },
  {
   "topic": "Computational Linguistics",
   "rationale": "Emphasizes contributions to AI-related and linguistic computational tasks."
  }
 ]
}
```

Figure 4: Prompt of *Topic* and *Rationale* Generation

C.3 Examples for *Problem/Solution/Result/Topic* contributions extracted from papers

Below we show examples of paper titles and abstracts, and different contributions (*Problem/Solution/Result/Topic*) we extract by language model.

Problem/Solution/Result/Topic *contributions from scientific papers*

Title: Sixfold excitations in electrides

Abstract: Due to the lack of full rotational symmetry in condensed matter physics, solids exhibit new excitations beyond Dirac and Weyl fermions, of which the sixfold excitations have attracted considerable interest owing to the presence of maximum degeneracy in bosonic systems. Here, we propose that a single linear dispersive sixfold excitation can be found in the electride Li_{12} Mg_3 Si_4 and its derivatives. The sixfold excitation is formed by the floating bands of elementary band representation A@12a originating from the excess electrons centered at the vacancies (i.e., the 12a Wyckoff sites). There exists a unique topological bulk-surface-edge correspondence for the spinless sixfold excitation, resulting in trivial surface "Fermi arcs" but topological hinge arcs. All gapped k_z slices belong to a two-dimensional higher-order topological insulating phase, which is protected by a combined symmetry T S_{4z} and characterized by a quantized fractional corner charge Q_{corner} = 3|e|/4. Consequently, the hinge arcs are obtained in the hinge spectra of the S_{4z}-symmetric rod structure. The state with a single sixfold excitation, stabilized by both nonsymmorphic crystalline symmetries and time-reversal symmetry, is located at the phase boundary and can be driven into various topologically distinct phases by explicit breaking of symmetries, making these electrides promising platforms for the systematic studies of different topological phases.

Contribution - Problem Statement

```
{
 "overarching_problem_domain":
    "Condensed matter physics",
 "challenges/difficulties":
    "Lack of full rotational symmetry
    in solids leading to new excitatio
    beyond Dirac and Weyl fermions",
 "research_question/goal":
    "Investigate sixfold excitations
    in electrides"
}
```

Contribution - Solution Statement

```
{
 "overarching_solution_domain":
    "Electrides and topological phases
 "solution_approach":
    "Propose that a single linear
    dispersive sixfold excitation can
    be found in the electride
    Li12Mg3Si4 and its derivatives",
 "novelty_of_the_solution":
    "Unique topological bulk-surface-e
    correspondence for the spinless
    sixfold excitation"
}
```

Contribution - Result Statement

```
{
 "findings/results":
    "The sixfold excitation is formed by
    floating bands of elementary band
    representation A@12a. All gapped
    kz slices belong to two-dimensional
    higher-order topological insulating
    phase, characterized by a quantized
    fractional corner charge Qcorner = 3|e|/4.
    Hinge arcs are obtained in the hinge
    spectra of the S4z-symmetric rod
    structure.",
 "potential_impact_of_the_results":
    "Electrides are promising platforms
    for systematic studies of different
    topological phases."
}
```

Contribution - Topic: *'Electrides', 'Electrides in Condensed Matter Physics', 'Higher-Order Topological Insulators', 'Non-symmorphic Symmetries', 'Sixfold Excitation in Solids', 'Sixfold Excitations', 'Symmetry Breaking in Topological Materials', 'Topological Bulk-Surface-Edge Correspondence', 'Topological Phase Transitions', 'Topological Phases in Condensed Matter Physics', 'Topological Properties'*

Title: The Tin Pest Problem as a Test of Density Functionals Using High-Throughput Calculations

Abstract: At ambient pressure tin transforms from its ground-state semi-metal α-Sn (diamond structure) phase to the compact metallic β-Sn phase at 13 • C (286K). There may be a further transition to the simple hexagonal γ-Sn above 450K. These relatively low transition temperatures are due to the small energy differences between the structures, $\approx$ 20 meV/atom between α-and β-Sn. This makes tin an exceptionally sensitive test of the accuracy of density functionals and computational methods. Here we use the high-throughput Automatic-FLOW (AFLOW) method to study the energetics of tin in multiple structures using a variety of density functionals. We look at the successes and deficiencies of each functional. As no functional is completely satisfactory, we look Hubbard U corrections and show that the Coulomb interaction can be chosen to predict the correct phase transition temperature. We also discuss the necessity of testing high-throughput calculations for convergence for systems with small energy differences.

Contribution - Problem Statement

```
{
 "overarching_problem_domain":
    "Density functionals and computati
    methods for phase transitions in
    materials.",
 "challenges/difficulties":
    "Small energy differences between
    phases of tin make it a sensitive
    test for the accuracy of density
    functionals.",
 "research_question/goal":
    "To study the energetics of tin in
    multiple structures using a variet
    of density functionals and assess
    their accuracy."
}
```

Contribution - Solution Statement

```
{
 "overarching_solution_domain":
    "High-throughput computational
    methods and density functional
    theory.",
 "solution_approach":
    "Using the high-throughput
    Automatic-FLOW (AFLOW) method
    to study tin's energetics with
    various density functionals.",
 "novelty_of_the_solution":
    "Application of Hubbard U
    corrections to improve predictions
    of phase transition temperatures."
}
```

Contribution - Result Statement

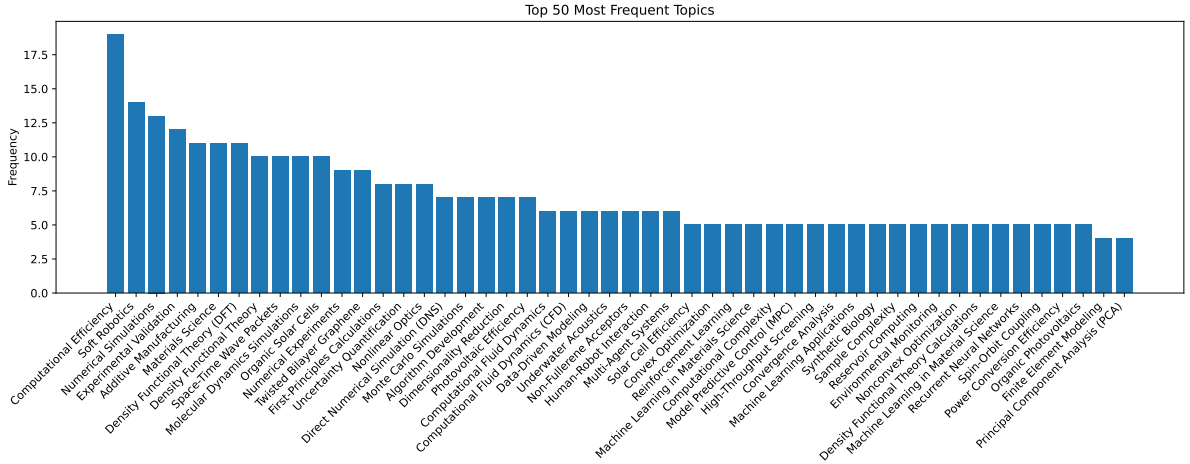
```
{
 "findings/results":
    "No functional is all satisfactory,
    but Hubbard U corrections can be chosen
    to predict the correct phase
    transition temperature.",
 "potential_impact_of_the_results":
    "Improved accuracy in predicting
    phase transitions in materials
    with small energy differences."
}
```

Contribution - Topic: *'Convergence Testing in Computational Simulations', 'Density Functional Theory (DFT) Accuracy', 'High-Throughput Computational Methods', 'Hubbard U Corrections', 'Tin Phase Transitions'*

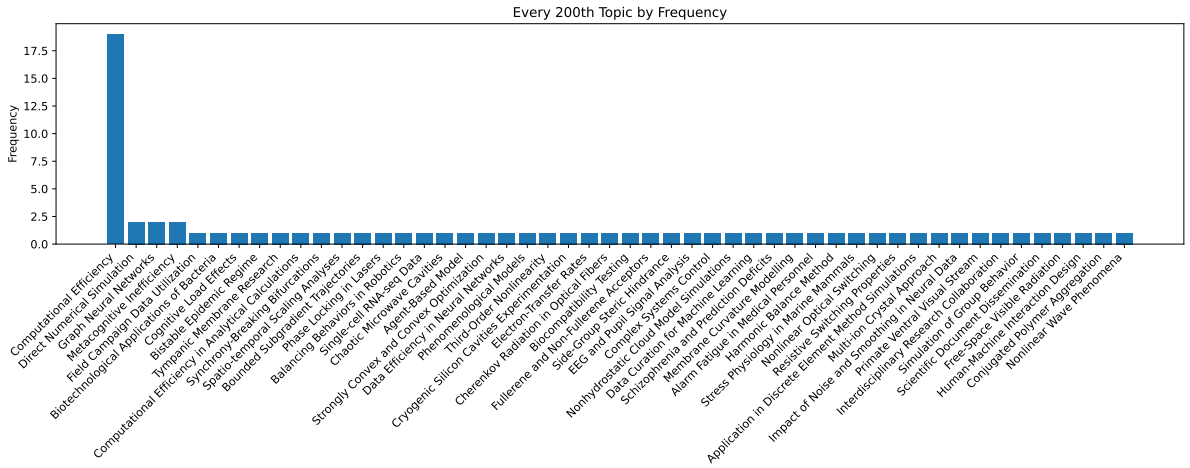
Table 9: Examples of extracted *problem/solution/result/topic* contributions from scientific paper abstracts.

C.4 Distribution of Extracted Topics

This section shows the distribution of various topics extracted from the papers based on frequency. This gives us an idea of what kind of topics were extracted.



(a) Top-50 topics by frequency in decreasing order



(b) Sampled topics (every 200)

Figure 5: Distribution of topics extracted from **SciPile**: (a) Top-50 topics, (b) Every 200 topics. Refer §3.2 for more information.

D Compiling a Seed Hierarchy

As we discuss in §4.3, we make a few adjustments to the seed hierarchy that we obtain from Wikipedia. Specifically:

1. We added "Theoretical Computer Science" and "Information Theory" as separate branches under "Formal Sciences" due to their unique characteristics;
2. We moved "Astronomy" under "Physical Science";
3. "Astronomy", "Geology" and "Oceanography" are listed under "Earth Science" but since these topics are not specific to early, we move them up one layer so that they're directly under "Physical Science"; The Wikipedia article groups Geology, Oceanography, and Meteorology under ;
4. We added "History" as a branch under "Social Sciences";
5. We included "Cell Biology" and "Genetics" under "Biological Sciences" as they were relevant and their inclusion would only help in better hierarchy creation.

These modifications aim to refine the hierarchy, ensuring it accurately reflects the distinct areas within each scientific domain. The resulting hierarchy is included in Fig.6.

```
{
    "Science":{
        "Formal Sciences":{
            "Logic":{},
            "Mathematics":{},
            "Statistics":{},
            "Computer Science":{},
            "Information Theory":{},
            "Systems Theory":{},
            "Decision Theory":{}
        },
        "Natural Sciences":{
            "Physical Science":{
                "Physics":{
                    "Classical Mechanics":{},
                    "Thermodynamics and statistical mechanics":{},
                    "Electromagnetism and photonics":{},
                    "Relativity":{},
                    "Quantum Mechanics":{},
                    "Atomic and molecular physics":{},
                    "Condensed matter physics":{},
                    "Optics and acoustics":{},
                    "High energy particle physics":{},
                    "Nuclear physics":{},
                    "Cosmology":{},
                    "Interdisciplinary Physics":{}
                },
                "Chemistry":{
                    "Physical Chemistry":{},
                    "Organic Chemistry":{},
                    "Inorganic Chemistry":{},
                    "Analytical Chemistry":{},
                    "Biological Chemistry":{},
                    "Theoretical Chemistry":{},
                    "Interdisciplinary Chemistry":{}
                },
                "Earth Science":{},
                "Astronomy":{},
                "Geology":{},
                "Oceanography":{},
                "Meteorology":{}
            },
            "Biological Sciences":{
                "Biochemistry":{},
                "Cell Biology":{},
                "Genetics":{},
                "Ecology":{},
                "Microbiology":{},
                "Botany":{},
                "Zoology":{}
            }
        },
        "Social Sciences":{
            "Anthropology":{},
            "Economics":{},
            "Political Science":{},
            "Sociology":{},
            "Psychology":{},
            "Geography":{},
            "History":{}
        }
    }
}
```

Figure 6: The seed hierarchy used by our FLMSCI baselines. See §D for details.

E FLMSCI: LLM-based Baselines

This section includes the pipeline and prompts used for FLMSCI (parallel) and FLMSCI(incremental) from §4.3.

E.1 Pipeline for FLMSCI (parallel)

This section demonstrates the pipeline used for FLMSCI (par) right from extracting topics and rationales to obtaining a final taxonomy with papers. (Refer to §4.3 for more information).

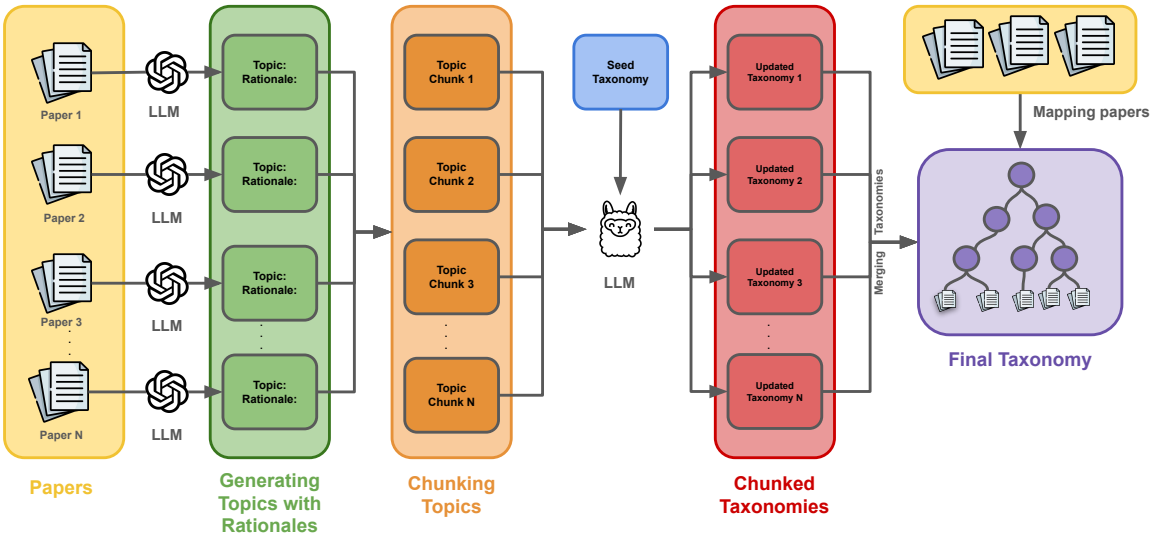


Figure 7: Pipeline for of FLMSCI (parallel).

E.2 Prompt for FLMSCI (parallel)

This prompt guides a large language model (LLM) to expand an existing scientific taxonomy - the seed taxonomy (Refer to D) by adding a set of new topics in a logical and consistent manner. With a clear list of instructions it ensures accurate placement and also preserves the original structure. This prompt was used with `Llama-3.3-70B-Instruct`. (Refer to §4.3 for more information.)

FLMSCI (parallel) Prompt

You are a scientific domain expert. You have an initial "seed taxonomy" of scientific concepts and a list of new topics to integrate into that taxonomy. Please carefully analyze these new topics and update the seed taxonomy so that:

1. You must retain the current structure of the seed taxonomy and respect all existing categories.

2. Place each and every topic from the list given below.

2. You are free to add new branches or subcategories only where necessary to fit the new topics in a consistent, hierarchical ("is-a") manner.

3. Each topic from the list must appear exactly once. Do not introduce any new topics beyond those in the list.

4. Ensure each new topic is placed under the correct parent concept based on its semantic meaning or specialization level.

5. Return your updated taxonomy as valid JSON, containing both the original seed hierarchy and the newly incorporated topics.

Below is your seed taxonomy (in JSON). Make sure to preserve its structure as much as possible:

{seed_taxonomy}

Here is the list of new topics that must be integrated:

{topics_chunk}

Focus on logical placement of each term to maintain an accurate scientific hierarchy.

Figure 8: Prompt of FLMSCI (incremental) pipeline

E.3 Demonstration of actions for FLMSCI (incremental)

This section demonstrates the various actions (`add sibling`, `make parent`, `go down` and `discard`) that are available for the LLM to take at various levels of taxonomy building. Refer to §4.3 for more information.

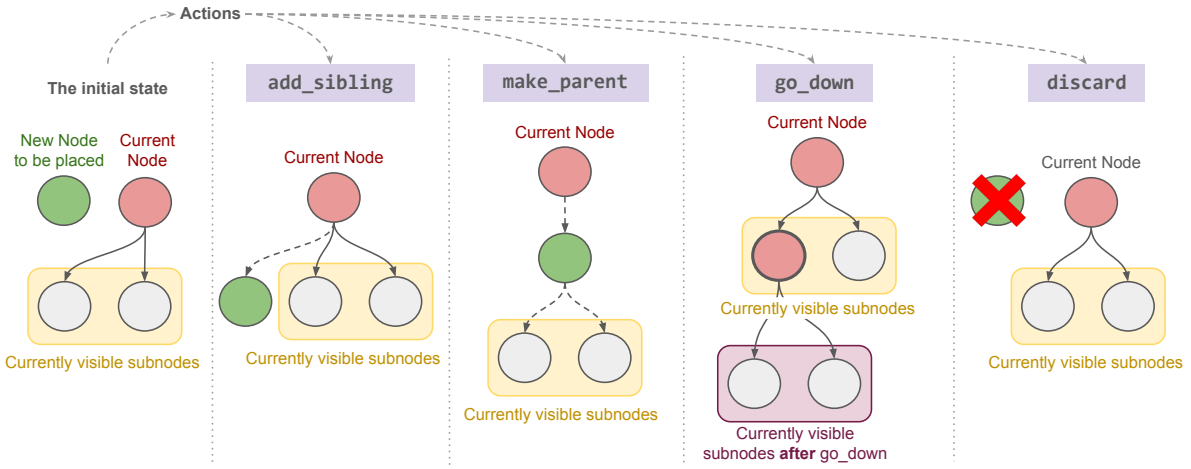


Figure 9: Actions for FLMSCI (incremental)

E.4 Prompt for FLMSCI (incremental)

This prompt is used to place new scientific topics into an existing seed taxonomy (Refer §D) incrementally. The model evaluates multiple possible actions based on the available action options. The prompt clearly instructs its priorities explicitly to give a hint to the model. The example usage and example output format help to get the response in a valid format. This prompt was used for `Llama-3.3-70B-Instruct`.

```
SUBNODE_DESCRIPTIONS = {
  "Formal Sciences": "Focuses on abstract systems and formal methodologies grounded in logic, mathematics, and symbolic reasoning. Provides theoretical frameworks (e.g., statistics, computer science, systems theory) used to model and solve problems across empirical disciplines and technology.",
  "Natural Sciences": "Investigates the physical universe and living organisms through empirical observation, experimentation, and theoretical analysis. Includes physical sciences (e.g., physics, chemistry, astronomy) and biological sciences (e.g., genetics, ecology) to uncover fundamental laws and processes governing nature.",
  "Social Sciences": "Studies human behavior, societies, and institutions using qualitative and quantitative methods. Encompasses disciplines like psychology, economics, and political science to analyze cultural, economic, and social interactions within historical and geographic contexts."
}
```

Figure 10: Descriptive statement used for contextualizing layer-1 items in the seed hierarchy, used in FLMSCI (incremental). See §4.3 for broader context.

```
You are building a scientific topics based hierarchy.

Path traced until now: {current_path}
Subnode options available at this level:
subnodes = [{subnodes}]
New topic: "{new_topic}"

Evaluate all possible actions listed below equally before choosing the most appropriate one.
Choose the action that best preserves a logical hierarchy, semantic clarity, and appropriate
    abstraction level.

**Priority Guidance**:
1. FIRST consider "go_down" if ANY existing subnode could reasonably contain the new topic as a
    specialization
2. THEN consider "make_parent" if multiple existing subnodes could be grouped under a new category
3. ONLY use "add_sibling" if the topic is FUNDAMENTALLY distinct from all existing subnodes at this
    level
4. "discard" should be used for low-quality or redundant topics

**Critical Rules**:
- A node about "Applications of X" should ALWAYS go under X, not as a sibling
- Specific methods/tools belong under their parent field (e.g., "PCR" under "Molecular Biology")
- Avoid creating flat structures

Possible actions:
1) "go_down" - Use this if the topic: {new_topic} is a *more specific* subtype of one of the "
    subnodes" and belongs *within* it.
2) "add_sibling" - Use this if the topic: {new_topic} is on the same level of abstraction as the
    existing "subnodes". It should be added *alongside* them as a direct child of `{current_path
    [-1]}`.
3) "discard" - Use this if the topic: {new_topic} is irrelevant, redundant, or already captured under
     another topic.
4) "make_parent" - Use this when the new topic: {new_topic} or any one of the "subnodes" is broader
    or more abstract than one or more of the subnodes. In that case, make the new topic a direct
    child of `{current_path[-1]}` and move the relevant subnodes under it. Return them in `"
    child_nodes": [...]`.

### Example of desired usage for each action:
1) "go_down"
   - "node": must be the name of one of the existing "subnodes"
   - "explanation": an optional text with reasoning
   - "child_nodes", "parent_node": not used.

2) "add_sibling"
   - "node": {new_topic}
   - "parent_node": {current_path[-1]}
   - "explanation": optional
   - "child_nodes": not used.

3) "discard"
   - "node": {new_topic}
   - "explanation": optional
   - "parent_node", "child_nodes": not used

4) "make_parent"
   - "node": {new_topic} or one of the "subnodes" whichever is more appropriate
   - "child_nodes": array of the subnodes that must be moved under the new node
   - "explanation": optional
   - "parent_node": not used

Your output must be valid JSON only:
{{
  "action": "go_down"|"add_sibling"|"make_parent"|"discard",
  "node": "string",
  "parent_node": "string or null",  // only used if action = add_sibling
  "child_nodes": ["string", ...],   // only used if action = make_parent
  "explanation": "string (optional)"
}}
No extra text.
```

Figure 11: The detailed prompt used in the execution of our FLMSCI (incremental) baseline. See §4.3 for broader context.

F Further Details on Collection of Science Papers

This section provides more context on our piles of papers in our experiments from §5.1. **SciPileLarge** is an extension of **SciPile**. For each paper in **SciPile**, we extract five relevant keywords using an LLM (see Fig.12) and query the Semantic Scholar API[2] with these keywords to retrieve additional relevant papers.

We apply three filtering criteria to ensure quality: (a) **Citation Count:** A paper must have a minimum number of citations to be considered reliable. The minimum citation count is calculated using the formula: $(2 + 3 \times (2025 - \texttt{publish\_year})$. (b) **Abstract Length:** A paper must have an abstract with at least 250 tokens, as measured by the tokenizer of `Llama-3.1-8B-Instruct`. (c) **Publication Venue:** A paper must be published in a recognized journal or conference. For each keyword, we select up to five papers that meet all criteria. This approach maintains the disciplinary distribution of our seed dataset **SciPile**while expanding our corpus to $10K$ papers.

Keyword Extraction for Dataset Expansion

Title: {title}

Abstract: {abstract}

Generate exactly 5 relevant keyword phrases for this research paper. Each keyword phrase should be no more than 6 words long.

Return only a JSON array containing these 5 keywords. No explanations or other text.

Figure 12: Prompt of Keyword Extraction for Dataset Expansion

G Hyperparameters of SCYCHIC

Here shows the models and hyperparameters we use for the experiments mentioned in §5.3. We utilize the GPT-4o model (`gpt-4o-2024-08-06`) to generate all contribution extractions along with detailed rationales explaining the extraction decisions. For `summarizer`, we use `Llama-3.3-70B-Instruct` (Grattafiori et al., 2024) for its superiority of following instructions among open-source models, and use `gte-Qwen2-7B-instruct` as our `embedder`. For clustering algorithm, we apply k-means clustering. The number of clusters for each layer is (6, 40, 276) when conducting experiments on **SciPile** ($2K$ papers), and (6, 40, 276, 1250) when on **SciPileLarge** ($10k$ papers).

[2] https://www.semanticscholar.org/product/api

H Additional Experiments of SCYCHIC

H.1 Evaluation Results on SciPile

Method	Accuracy (%)		LLM Cost		Hierarchy Structure		
	Strict-Acc ↑	L1-Acc ↑	Avg. # of Input Tokens ↓	# of Calls ↓	Depth	Avg. Branching Factor	Max. Branching Factor
Contributions type: Problem Statement							
SCYCHIC	**51.1** ± 3.8	**81.7** ± 2.6	2624				20
↳Top-down	49.0 ± 3.7	80.3 ± 2.7	2953	322	3	7.1	18
↳Bottom-up	45.9 ± 5.0	69.3 ± 8.1	2177				16
Contributions type: Solution Statement							
SCYCHIC	**48.8** ± 6.1	**82.3** ± 1.1	2343				16
↳Top-down	45.9 ± 5.5	79.2 ± 3.4	2521	322	3	7.1	19
↳Bottom-up	36.7 ± 2.6	67.0 ± 4.3	1990				14
Contributions type: Results Statement							
SCYCHIC	46.4 ± 5.2	76.4 ± 6.9	2654				16
↳Top-down	**47.3** ± 3.1	**80.5** ± 4.4	2718	322	3	7.1	16
↳Bottom-up	40.0 ± 10.7	64.0 ± 8.9	2210				13

Table 10: Evaluation results of SCYCHIC and the corresponding baselines on the 2K (**SciPile**) dataset.

H.2 Detailed Evaluation Results on *Topic* Contributions

Here we show the complete evaluation results mentioned in §5.2. SCYCHIC, FLMSCI (**par**allel) and FLMSCI(**inc**remental) are using *Topic* as contribution type.

Method	Accuracy (%)		LLM Cost		Hierarchy Structure				
	Strict-Acc ↑	L1-Acc ↑	Avg. # of Input Tokens ↓	# of Calls ↓	Max Depth	Avg Depth	Avg Bran. Factor	Max Bran. Factor	# of Items
Contributions type: Topic									
SCYCHIC	**14.9** ± 2.7	**65.7** ± 4.4	5017	322	3	3	40.9	128	11k
↳Top-down	14.5 ± 4.7	62.5 ± 7.4	6440	322	3	3	40.9	104	11k
↳Bottom-up	13.9 ± 5.3	54.4 ± 12.7	3988	322	3	3	40.9	119	11k
↳FLMSCI (par)	4.0 ± 2.8	32.0 ± 6.3	8896	226	9	6.2	13.9	734	**9.9K**
↳FLMSCI (inc)	**18.0** ± 5.3	**91.0** ± 4.0	4040	**61K**	14	7.7	3.6	704	**10.4K**

Table 11: Evaluation results of SCYCHIC, FLMSCI (**par**allel) and FLMSCI(**inc**remental) when using *Topic* as contribution type. "Bran." stands for "Branching". All methods show poor Strict-Acc ($\leq 18.0\%$), highlighting the challenging nature of the task. On one hand, FLMSCI (inc) achieves the highest accuracy, showing the feasibility of building hierarchies by pure LLM-based methods. However, it incurs substantial computational costs, about 200× higher than other methods. In contrast, SCYCHIC offers a balanced performance profile with competitive accuracy (14.9% Strict-Acc, 65.7% L1-Acc) while maintaining significantly lower computational requirements.

H.3 Comparison of Different Embedding models

For the `embedder` mentioned in §4.1. We evaluate three embedding models—Qwen's `gte-Qwen2-7B-instruct` (Li et al., 2023), OpenAI's `text-embedding-3-large`, and `text-embedding-ada-002`. The first two performe similarly, whereas `text-embedding-ada-002` produce markedly weaker results. Given the comparable overall performance between the two leading models, we selecte `gte-Qwen2-7B-instruct` for our main experiments due to its strong balanced performance across both metrics, superior Sctric-Acc results, and practical advantages as an open-weight model that offers greater accessibility and cost-effectiveness for reproducible research.

H.4 Experiments of Prompt Engineering

We investigate the effect of different prompts on the final quality of hierarchy. In the main text, for the `summarizer` mentioned in §4.1, we use the detailed version prompt which is carefully curated. For

Models→	text-embedding-3-large		gte-Qwen2-7B-instruct		text-embedding-ada-002	
Metrics→	L1-Acc	Sctric-Acc	L1-Acc	Sctric-Acc	L1-Acc	Sctric-Acc
PROBLEM	**86.7** ± 4.6	46.7 ± 0.9	81.7 ± 2.6	**51.1** ± 3.8	76.0 ± 4.4	41.7 ± 5.2
SOLUTION	80.3 ± 3.4	36.7 ± 1.7	**82.3** ± 1.1	**48.8** ± 6.1	63.5 ± 2.0	31.0 ± 3.2
RESULTS	**84.7** ± 5.7	44.0 ± 0.8	76.4 ± 6.9	**46.4** ± 5.2	74.6 ± 3.4	41.0 ± 8.7

Table 12: Performance comparison across three embedding models and contribution types. `gte-Qwen2-7B-instruct` demonstrates superior Sctric-Acc performance across all categories, while `text-embedding-3-large` excels in L1-Acc for *problem* and *results*. `text-embedding-ada-002` shows consistently weaker performance across both metrics.

comparison, we also conduct the experiments with a much simpler prompt.

Detailed (Curated) Prompt

You are a scientific research expert specializing in identifying and analyzing research problems and challenges. Your task is to analyze a collection of research papers or research clusters and provide a focused analysis of the research problems they address.
The input could be either a collection of individual papers or research cluster summaries. Based on the content, you need to:

1. Identify the core research problems and challenges being addressed
2. Determine the overarching problem domain that connects these research efforts
3. Analyze the specific difficulties, gaps, or limitations that motivate this research
4. Understand the research questions or goals that drive these investigations
5. Generate an appropriate cluster name that captures the essence of the problem space
6. Provide a comprehensive problem-focused analysis

Here is the content to analyze: {}
Remember to:

- Focus specifically on problems, challenges, and research gaps rather than solutions
- Be specific about the technical difficulties and limitations being addressed
- Identify both theoretical and practical challenges
- Consider interdisciplinary connections in problem formulation
- Maintain scientific accuracy and use precise terminology
- Generate only one JSON format output that must follow the structure exactly

Please format your response as a JSON object with the following structure:

```
{
    "Cluster Name": "A clear and specific title focusing
    on the problem domain (No less than 5 words)",

    "Problem": {
        "overarching problem domain": "The broad scientific
        domain where these problems exist",
        "challenges/difficulties": "Specific technical,
        theoretical, or practical challenges that these
        papers address",
        "research question/goal": "The fundamental
        research questions or objectives that motivate
        this work"
    }
}
```

Simple Prompt

You are a scientific research expert specializing in identifying and analyzing research problems and challenges.
Analyze the input %s and output one JSON object:

```
{
    "Cluster Name": "A clear and specific title (No less
    than 5 words)",
    "Problem": {
    "overarching problem domain": "",
    "challenges/difficulties": "",
    "research question/goal": ""
  }
}
```

Instructions
Extract key themes and concepts.
Identify the common thread that links the items.
Craft a clear, specific title ($\geq$ 5 words) for Cluster Name.
Return only the JSON—nothing else.

Table 13: Comparison of Detailed (Curated) and Simple Prompts

The results show that across all contributions, the curated prompt offers significantly better quality hierarchies.

Prompt type ↓	Embedder→	`text-embedding-3-large`		`gte-Qwen2-7B-instruct`	
	Metrics→	L1-Acc	Sctric-Acc	L1-Acc	Sctric-Acc
Simplified	*problem*	75.0 ± 4.6	33.7 ± 3.7	61.0 ± 0.8	24.7 ± 1.7
Detailed		**86.7** ± 4.6	**46.7** ± 0.9	**81.7** ± 2.6	**51.1** ± 3.8
Simplified	*solution*	65.3 ± 3.4	32.7 ± 2.6	59.0 ± 2.8	21.7 ± 2.9
Detailed		**80.3** ± 3.4	**36.7** ± 1.7	**82.3** ± 1.1	**48.8** ± 6.1
Simplified	*results*	77.7 ± 4.1	38.0 ± 4.6	66.7 ± 3.3	27.7 ± 2.5
Detailed		**84.7** ± 5.7	**44.0** ± 0.8	**76.4** ± 6.9	**46.4** ± 5.2

Table 14: Performance comparison between simplified and detailed prompts across different embedding models and contribution types. Detailed prompts consistently outperform simplified prompts across all scenarios, with improvements ranging from 7.0 to 23.3 % for L1-Acc and 3.0 to 26.4 % for Sctric-Acc. The `gte-Qwen2-7B-instruct` model shows the largest performance gains, with L1-Acc improvements of 20.7, 23.3, and 9.7 % for *problem*, *solution*, and *results* respectively.

I Visualization and Examples of Inter-Cluster Citations

Below is the figure of comparing inter- and intra-cluster citation counts mentioned in §5.5.

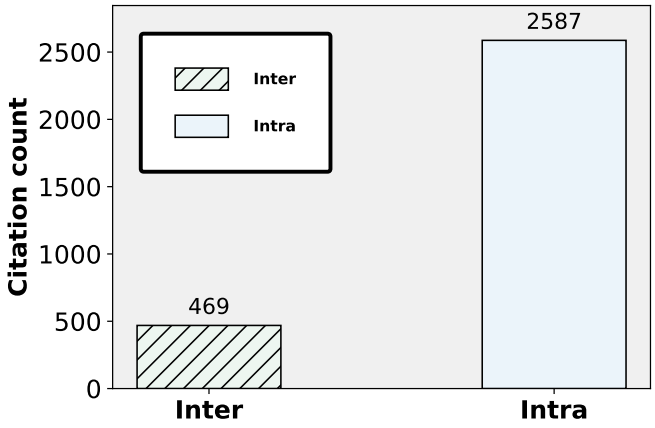


Figure 13: Comparison of inter(green)- and intra(blue)-cluster citation counts. 84.7% citations are between papers in the same top layer cluster, and the rest inter-cluster citations are mostly theory-to-application works, which proves the reliability of SCYCHIC.

Below are examples of citations between papers in different top-layer clusters. These examples show that many inter-cluster citations represent theory-to-application connections, while the last row illustrates cross-disciplinary citations between research fields. Importantly, all papers involved are correctly categorized—the inter-cluster citations reflect legitimate relationships rather than classification errors.

Citing Paper		Cited Paper
Rationale: AI research grounding in foundational cognitive science theory		
Minding Language Models' (Lack of) Theory of Mind: A Plug-and-Play Multi-Character Belief Tracker	→	**Does the chimpanzee have a theory of mind?**
Challenges and Limitations in ML and AI		*Neuroscience, Cognitive Psychology, and Neurotechnology*
Rationale: Theory-to-application for THz photonics		
Terahertz topological photonic integrated circuits for 6G and beyond	→	**Topological photonics**
Advanced Materials Challenges		*Quantum Systems and Materials Science*
Rationale: Hardware implementation citing quantum network theory		
Cavity electro-optics in thin-film lithium niobate	→	**Quantum internet: A vision for the road ahead**
Advanced Materials Challenges		*Quantum Systems and Materials Science*
Rationale: Manufacturing citing characterization techniques		
Creating Quantum Emitters in Hexagonal Boron Nitride	→	**Nanoscale Imaging and Control of hBN Single Photon Emitters**
Advanced Materials Challenges		*Quantum Systems and Materials Science*
Rationale: Cross-disciplinary bridge between biology and quantum physics		
Magnetic field effects in biology from radical pair mechanism	→	**Quantum biology revisited**
Neuroscience, Cognitive Psychology, and Neurotechnology		*Quantum Systems and Materials Science*

Table 15: Examples of cross-cluster citations. Each row shows the citing paper, the cited paper, their cluster names, and the citation rationale.

J Demonstration of Hierarchy

Below is a snippet of our final hierarchy result as mentioned in §5.6.

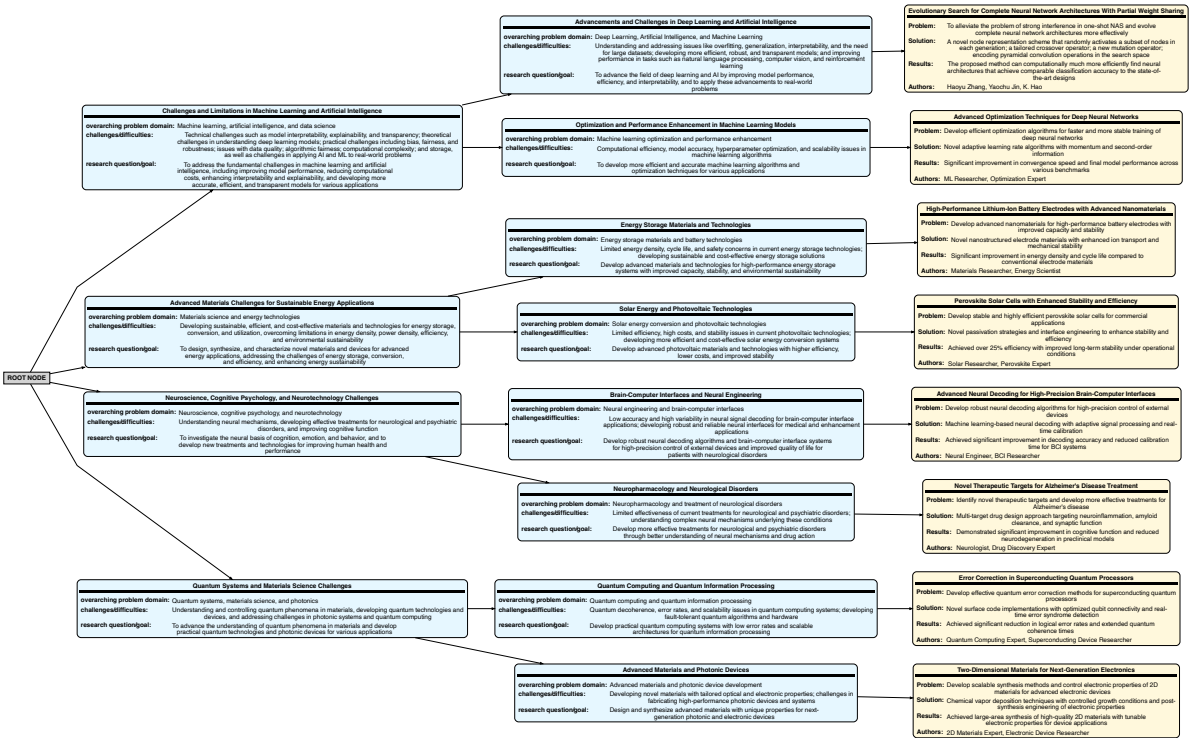


Figure 14: Above is a small example of a final hierarchy generated by SCYCHIC on the **SciPileLarge** dataset. The original hierarchy has 4 levels, use papers' *problem* contribution. Due to space constraints, this snippet shows only two levels of clusters above the individual papers.