# TAXOALIGN: Scholarly Taxonomy Generation Using Language Models

**Avishek Lahiri*** **Yufang Hou†** **Debarshi Kumar Sanyal***
*Indian Association for the Cultivation of Science, Kolkata, India
†IT:U Interdisciplinary Transformation University Austria, Linz, Austria
avisheklahiri2014@gmail.com,
yufang.hou@it-u.at, debarshi.sanyal@iacs.res.in

## Abstract

Taxonomies play a crucial role in helping researchers structure and navigate knowledge in a hierarchical manner. They also form an important part in the creation of comprehensive literature surveys. The existing approaches to automatic survey generation do not compare the structure of the generated surveys with those written by human experts. To address this gap, we present our own method for automated taxonomy creation that can bridge the gap between human-generated and automatically-created taxonomies. For this purpose, we create the CS-TAXOBENCH benchmark which consists of 460 taxonomies that have been extracted from human-written survey papers. We also include an additional test set of 80 taxonomies curated from conference survey papers. We propose TAXOALIGN, a three-phase topic-based instruction-guided method for scholarly taxonomy generation. Additionally, we propose a stringent automated evaluation framework that measures the structural alignment and semantic coherence of automatically generated taxonomies in comparison to those created by human experts. We evaluate our method and various baselines on CS-TAXOBENCH, using both automated evaluation metrics and human evaluation studies. The results show that TAXOALIGN consistently surpasses the baselines on nearly all metrics. The code and data can be found at https://github.com/AvishekLahiri/TaxoAlign.

## 1 Introduction

In scientific research, a taxonomy is constructed around a well-defined topic to integrate relevant research findings under a unified framework, thereby facilitating deeper understanding among researchers and industry practitioners alike. In the general domain, taxonomies have been proven to be useful tools (Wang et al., 2017), which exhibit the capability of enhancing the performance of various

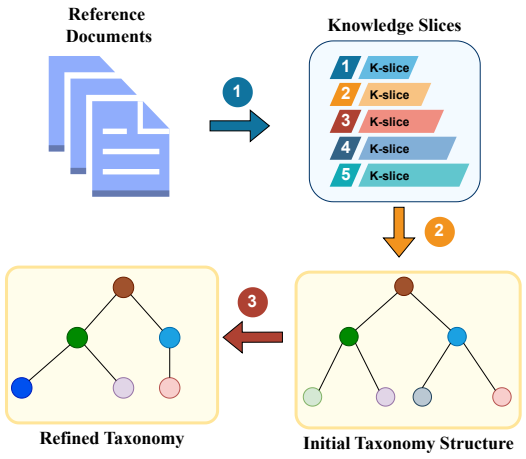


Figure 1: Schematic representation of TAXOALIGN: (1) Knowledge Slice Creation (2) Taxonomy Verbalization (3) Taxonomy Refinement

Natural Language Processing (NLP) and Information Retrieval (IR) tasks such as question answering (Harabagiu et al., 2003; Yang et al., 2017), textual entailment (Geffet and Dagan, 2005), personalized recommendation (Zhang et al., 2014), query understanding (Hua et al., 2017), information extraction (Hou et al., 2019; Şahinuç et al., 2024) and knowledge graph construction (Hou et al., 2021; Mondal et al., 2021). Taxonomies have also found a place in real-world deployment applications such as biomedical systems (Köhler et al., 2013), information management (Nickerson et al., 2013) and e-commerce (Aanen et al., 2015; Mao et al., 2020).

In this paper, we conceptualize the task of *automated scholarly taxonomy generation*. When given the taxonomy topic and a set of related reference papers, the task for the model is to reason over this large set of reference documents and generate a taxonomy tree that is both concise and provides maximal coverage of the reference documents. Automating this process reduces the time, effort, and energy



*Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30203–30223
November 4-9, 2025 

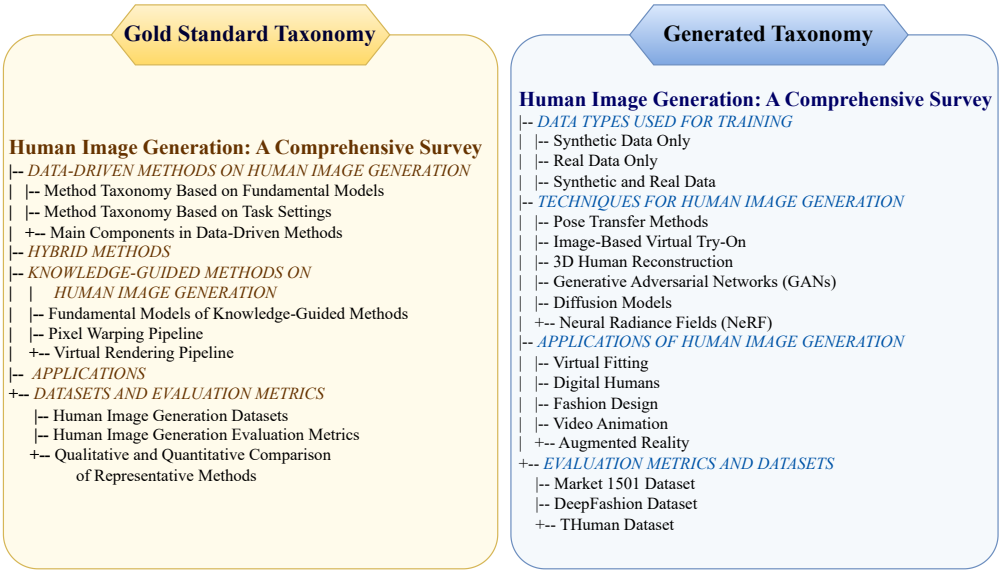


Figure 2: A comparison of a gold standard taxonomy tree and a generated taxonomy tree using TAXOALIGN for the topic "Human Image Generation: A Comprehensive Survey". The generated taxonomy shown here uses Mistral-7B-Instruct-v0.3 for creation of knowledge slices and Llama-3.1-Tülu-3-8B for the taxonomy verbalization component and GPT-4o-mini for refining the generated taxonomy.

researchers spend organizing research within a specific topic.

While Large Language Models (LLMs) are increasingly applied across a wide range of tasks, they face notable challenges in understanding and performing domain-specific tasks (Li et al., 2024; Cai et al., 2025). Additionally, their reasoning over long contexts is limited by constraints in context window size (Liu et al., 2024). In our initial pilot study, we observed that prompting LLMs could generate some topic-related sub-topics, but they were not remotely aligned to the human-written taxonomy. For example, "*3D Models and Mapping*", "*Generative Adversarial Network*" and "*Data and Annotation*" were the first-level nodes/sub-topics that were generated by Tülu3 (Lambert et al., 2025) for the topic "*Human Image Generation: A Comprehensive Survey*" when supplied with all the summaries of the cited documents. In general, the generated nodes were distantly relevant to the topic but were not at close to the human-written ones as shown in Figure 2 (left).

Recent prior work has focused extensively on the end-to-end automated creation of survey papers (Wang et al., 2024; Liang et al., 2025; Yan et al., 2025; Kang and Xiong, 2025). In contrast, scholarly taxonomy generation remains a relatively unexplored area, with no well-articulated open-source data resources currently available for this task. Moreover, prior work does not compare the structure of generated surveys with those authored by human experts. There is also a lack of an evaluation framework capable of assessing both the structural similarity and semantic coherence between automatically generated and human-written taxonomy trees.

To address this gap, we curate and release CS-TAXOBENCH, a comprehensive benchmark that is designed for the task of scholarly taxonomy generation (Section 3). Our benchmark consists of 460 human-written taxonomies (accompanied by their corresponding reference papers) that have been extracted from survey articles published in Computer Science journals in 2020 – 2024. We also curate an additional test set made up of 80 taxonomies extracted from conference survey papers.

We further develop TAXOALIGN, our own intuitive LLM-based pipeline for generating taxonomies (Section 4). TAXOALIGN consists of three parts: *Knowledge Slice Creation*, *Taxonomy Verbalization* and *Taxonomy Refinement*. We compare our method with a range of baselines to show the effectiveness of our method. Figure 1 demonstrates our proposed framework, while Figure 2 (right) shows an example of a taxonomy generated using TAXOALIGN.

Finally, we present an automated evaluation framework for the comparison of taxonomy-tree

structures (Section 5). For this purpose, we develop two metrics of our own – the average degree score metric to judge structural similarity and the level-order traversal comparison metric to judge semantic similarity. In addition, we adopt soft recall and entity recall metrics, originally proposed for evaluating outline generation, to assess the quality of the generated taxonomies (Fränti and Mariescu-Istodor, 2023; Shao et al., 2024; Kang and Xiong, 2025). We also employ LLM-as-a-judge for qualitative evaluation. TAXOALIGN outperforms all baselines across nearly all metrics, including human evaluation. To facilitate future research, we make our code and dataset publicly available at https://github.com/AvishekLahiri/TaxoAlign.

## 2 Related Work

**Taxonomy Construction.** Taxonomy learning has been attempted in NLP through the decades by captalizing on the semantic relations in text (Hearst, 1992; Pantel and Pennacchiotti, 2006; Suchanek et al., 2006; Ponzetto and Strube, 2011; Rios-Alvarado et al., 2013; Dietz et al., 2012; Liu et al., 2012; Diederich and Balke, 2007; Wang et al., 2010; Kang et al., 2016; Kozareva and Hovy, 2010; Velardi et al., 2013). Recent approaches such as HiGTL (Hu et al., 2025) and the method of Martel and Zouaq (2021) introduce graph- and clustering-based techniques for taxonomy learning. Most of these methods use pattern-based or clustering-based methods, whereas TAXOALIGN leverages the power of LLMs to construct taxonomies in the scientific domain.

**Scientific Survey Generation and Knowledge Synthesis.** Recently, there has been some interest among researchers to generate surveys from a corpus of research papers. AutoSurvey (Wang et al., 2024), SURVEYX (Liang et al., 2025), ResearchArena (Kang and Xiong, 2025), Qwen-long (Lai et al., 2024) and SURVEYFORGE (Yan et al., 2025) are some of the prominent techniques proposed for this task. In literature-based knowledge synthesis, LLMs have been used to generate scientific leaderboards (Şahinuç et al., 2024; Timmer et al., 2025), literature review tables (Newman et al., 2024), or to synthesize biomedical evidence in the format of forest plots (Pronesti et al., 2025a,b). Our work focuses on scientific survey taxonomy generation and contributes to the broader agenda of AI for Science (Eger et al., 2025).

## 3 CS-TAXOBENCH

### 3.1 Overview

In graph theory, a tree is defined as an undirected connected graph with no cycles. A taxonomy tree $T$ is a tree in which the root represents the taxonomy topic and the child nodes represent sub-topics and grandchild nodes represent more fine-grained topics. Survey papers typically propose a taxonomy which is expanded upon in the sections and sub-sections of the paper. Therefore, in most cases, the structure of the paper closely mirrors the nodes and connections in the taxonomy tree. We leverage this pattern to extract scholarly taxonomies, using our taxonomy extraction module to first derive outlines from survey papers and then analyze them to construct the final taxonomy.

### 3.2 Desiderata

We list out the desiderata we used for selecting taxonomies for inclusion in our benchmark dataset. Our overall goal was to ensure that the selected taxonomies are of high-quality with a logical flow and each node is grounded in a set of reference papers. We decide on the following desiderata for curating our dataset: (1) each taxonomy should be based on a specific research topic, and the taxonomy should provide optimal coverage of the given topic; (2) the taxonomies should be human-made, i.e., they should not be generated artificially; (3) the taxonomy trees should be multi-layered, i.e., they should have at least two levels.

### 3.3 Automated Taxonomy Extraction

**Data Source Selection**: Survey papers form a rich resource for taxonomies in scientific literature. Therefore, we select survey papers as our primary source of data. To mitigate the risk of data contamination while ensuring that the papers in consideration are of high quality, we select survey papers from "*ACM Computing Surveys*" (ACM CSUR), which is a highly reputed journal in the field of Computer Science research with an Impact Factor of 23.8.[1] This is one of the top venues in Computer Science that publishes survey papers relating to the areas of computing research and practice.

**Research Paper Selection**: We select a time frame of five years between 2020 to 2024 for the purpose of the creation of our dataset. A total of 1,165 papers were accepted in ACM CSUR during this period, of which 325 are open-access and 285

[1] https://dl.acm.org/journal/csur


Figure 3: An overview of the pipeline for the curation of our dataset.

have copies available on arXiv.[2] Due to licensing restrictions, we include only the open-access and arXiv articles in our study.

**Filtering**: We use Docling (Team, 2024) to extract text from the research paper PDFs. Since arXiv papers are not always in a consistent format, we remove those with noisy layouts or Docling parsing errors. After filtering, 499 papers remain.

**Reference Paper Matching**: We retrieve the abstracts of the reference papers by parsing data from Semantic Scholar[3], which hosts approximately 214 million research documents. If fewer than 50% of a survey paper's references are available on Semantic Scholar, we exclude that paper from the final version of our proposed datasets. Following this criterion, 39 out of the original 499 papers are excluded, leaving us with a final set of 460 papers. The average percentage of available citations in the final set of papers is shown in Table 1.

**Taxonomy Extraction**: We extract the headings, subheadings, and sub-subheadings from the retrieved text of the survey paper to construct the taxonomy. The title of the survey paper is treated as the overall taxonomy topic. We discard all headings that contain terms such as "Introduction", "related work", "problem formulation", "summary", "conclusion", "result", "future", "discussion", "background" and "overview" as they do not contribute to the core taxonomy. We also remove those nodes for which we cannot extract reference papers from Semantic Scholar.

| Statistics | Value |
|---|---|
| No. of taxonomies in train set | 400 |
| No. of taxonomies in test set | 60 |
| Total no. of cited papers | 79,027 |
| Cited papers present in S2 | 60,373 |
| % of available cited papers | 76.40% |

Table 1: Statistics of our proposed benchmark. Note that here S2 refers to Semantic Scholar (Lo et al., 2020).

[2] https://arxiv.org/

[3] https://www.semanticscholar.org/me/research

### 3.4 Dataset Statistics

Our entire benchmark contains 460 taxonomies along with the reference papers that are used to build each taxonomy. The details about the statistics of our benchmark are present in Table 1. On average, there are around 131 reference papers for each taxonomy in our benchmark.

### 3.5 Manual Annotation

To evaluate the taxonomy extraction method, we manually annotate a set of 10 taxonomy trees from as many survey papers. We use a Python package, TreeLib[4], to annotate these taxonomies from their respective survey papers. The papers were selected such that there were explicitly-defined taxonomies in them. To compare the annotated and generated trees, we compare the paths from each node to the root node. Each path is treated as an individual element. The precision, recall and F1 between the annotated and extracted taxonomies were found to be 83.92%, 94.35% and 88.83% respectively, thereby demonstrating a high degree of correlation. The only errors originated due to some general section headers like "Two sea changes in Natural Language Processing" (Liu et al., 2023), which were not present in the annotated trees.

### 3.6 Additional Test Set: Survey Papers in Conferences

We create an additional new test set from conference survey papers for testing on a different distribution of survey papers. For this purpose, we chose survey papers published in 2024 in IJCAI (which has a dedicated survey track) and all the ACL* conferences (ACL, NAACL, EMNLP and EACL). There are a total of 86 survey papers in these conferences in 2024. We carry out our proposed filtering strategies, which narrows down the total number of papers to 80. Therefore, we have 80 taxonomy trees with an average of about 71 reference papers for each tree. This set is solely used as a test set.

[4] https://treelib.readthedocs.io/en/latest/

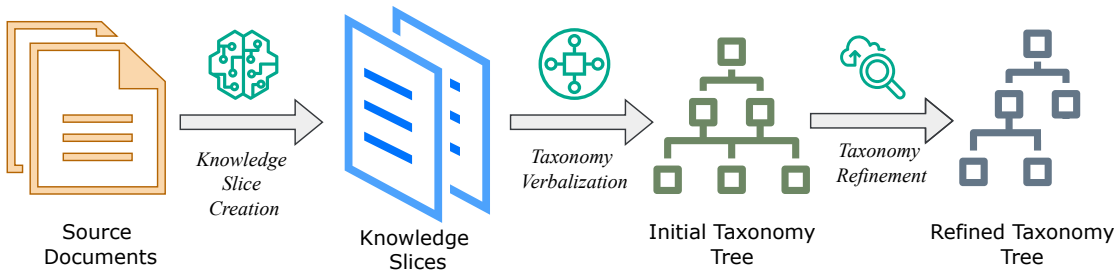


Figure 4: An overview of the proposed TAXOALIGN pipeline.

## 4 TAXOALIGN

### 4.1 Task Formulation

Given a corpus of documents $D$ and a topic $t$, the task is to automatically construct a hierarchical taxonomy tree $T$ whose nodes represent relevant topics and subtopics derived from $D$. The goal for $T$ is to comprehensively and meaningfully categorize the information contained within the entire corpus that relates to topic $t$.

To solve this task, we propose a method TAXOALIGN, comprising three components: Knowledge Slice Creation, Taxonomy Verbalization and Taxonomy Refinement. Figure 4 shows the three-stage pipeline of our proposed method.

### 4.2 Knowledge Slice Creation

In this step, we use a LLM to identify segments of text within each research article that are highly relevant to the taxonomy topic. We refer to these segments as knowledge slices. A knowledge slice is thus a highlighted portion of the document that strongly relates to the taxonomy topic.

This stage helps in extracting information that guides the model in later steps about the topics and subtopics present in the cited papers related to the taxonomy. More importantly, the number of cited papers in a survey paper is quite extensive, thereby making it quite impossible to fit all the papers in the model's input context window. This step ensures that all the cited papers can be accommodated within the context length of recent LLMs.

### 4.3 Taxonomy Verbalization

We opt for the instruction tuning of an LLM with the most pertinent taxonomy topic-related information from the reference papers that has been extracted in the previous stage.

The main objective is to teach the model to generate meaningful and concise taxonomies which are grounded in the given information, and most importantly, teach the model to learn the structure of the taxonomy trees. Finetuning helps in preserving the structure of the taxonomy in a major way, which is a feature that is lacking in the direct prompting-based methods.

### 4.4 Taxonomy Refinement

The verbalization phase is followed by a refinement stage, which evaluates and refines the connections between the parent and the child nodes. It checks whether each node is grounded in the document knowledge slices. If the tree contains too few nodes, it expands the node set to achieve a greater coverage of the documents using their corresponding knowledge slices. This refinement strategy is executed by prompting an LLM with stronger reasoning capabilities than those used in the previous stages.

## 5 Evaluation

We present both the new evaluation approaches we have developed and the existing methods used in this domain. The main challenge in comparing a generated tree with the gold tree lies in aligning the two structures. The two trees should be structurally similar as well as semantically aligned. However, alignment is challenging because the trees may differ significantly in their hierarchical structures or exhibit low lexical overlap, as has been encountered in similar problems such as table schema alignment evaluation (Newman et al., 2024).

### 5.1 Average Degree Score

The first condition for two trees to be considered similar is that their structures should be similar. We design this metric to judge the structural similarity of the generated tree and the gold standard tree.

In a graph, the average degree is calculated as the average number of edges connected to a node in

the graph. For any tree with $N$ nodes, the number of edges is $N-1$, which gives the average degree of $2(N-1)/N$. Therefore, to judge the structural similarity between the gold standard tree $T$ and the predicted tree $T'$, we find the average degree score $\Delta$, which is the ratio between the average degree of $T'$ and the average degree of $T$.

$$\Delta = \frac{\sum_{i=1}^{m} d(T'_i)}{\sum_{i=1}^{n} d(T_i)} \quad (1)$$

where, $d(t)$ represents the degree of a node $t$, and $m$ and $n$ are the number of nodes in the trees $T$ and $T'$ respectively. In the ideal case, the value of $\Delta$ should be 1. If the generated tree $T'$ is more branched out than the original tree $T$, then the value of $\Delta$ is greater than 1, while if $T'$ is less branched out than it should be, then the value of $\Delta$ is less than 1. We report the final score as the mean of all the average degree scores in the test corpus.

### 5.2 Level-order Traversal Comparison

The hierarchical structure of a tree makes it difficult to implement standard text generation metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), or BERTScore (Zhang et al., 2019). Therefore, we propose a metric to compare the gold standard tree and the generated tree using level-order traversal. More specifically, we traverse the tree in such a way that all the nodes present in the same level are traversed completely before traversing the next level. After converting the entire tree into a single list through level-order traversal, we calculate the the corpus-level BLEU-2, ROUGE-L and BERTScore.

### 5.3 Node Soft Recall and Node Entity Recall

We use the metrics *Node Soft Recall* and *Node Entity Recall*, following the evaluation protocol in prior work (Fränti and Mariescu-Istodor, 2023; Shao et al., 2024; Kang and Xiong, 2025). These metrics compare the generated and the ground-truth taxonomy trees using semantic similarity and lexical overlap between them, respectively. Node Soft Recall (NSR) is dependent on soft cardinality of a tree (Jimenez et al., 2010), which is given by,

$$c(T) = \sum_{i=1}^{n} \frac{1}{n \sum_{j=1}^{n} \mathrm{Sim}(T_i, T_j)} \quad (2)$$

where $\mathrm{Sim}(T_i, T_j)$ is the cosine similarity between the SENTENCE-BERT (Reimers and Gurevych, 2019) embeddings of the taxonomy trees $T_i$ and $T_j$. The Node Soft Recall between two trees $T$ and $T'$ is defined as,

$$\mathrm{NSR}(T, T') = \frac{c(T) + c(T') - c(T \cup T')}{c(T')} \quad (3)$$

Since in most baselines, there is a large mismatch between the number of nodes of the generated taxonomy tree and the original tree, we tweak the original heading soft recall in Shao et al. (2024) by inserting a normalizing factor in Eq. 2, i.e., we divide by the number of nodes to offset the effect of a large node count in the tree.

Node Entity Recall (NER) between the gold standard tree $T$ and the generated tree $T'$ is defined as the percentage of entities that are present both in $T$ and $T'$. Formally, it can be expressed as,

$$\mathrm{NER}(T, T') = \frac{|\mathrm{Ent}(T) \cap \mathrm{Ent}(T')|}{|\mathrm{Ent}(T)|} \quad (4)$$

where $|\mathrm{Ent}(T)|$ represents the number of entities in $T$. In our case, we track Noun Phrases (NP) for better coverage. We use the chunking model from FLAIR (Akbik et al., 2019) for this purpose.

### 5.4 LLM-as-a-Judge

We prompt a stronger LLM, GPT-4.1, with the gold-standard taxonomy tree and the generated taxonomy tree, and ask it to evaluate the generated tree on a scale of 1 to 5 based on the structural similarity and the semantic similarity of the two trees. The LLM judges whether the generated tree aligns with the gold tree and whether they are coherent. The prompt is described in Appendix B.

## 6 Baselines

In this section, we present the baselines against which we evaluate our method TAXOALIGN.

**AutoSurvey** (Wang et al., 2024): The outline generation part of this method randomly divides the reference papers into several groups, which results in the creation of multiple outlines. The language model then amalgamates these outlines to construct a single comprehensive outline. For a fair comparison with our method, we provide the reference papers, in contrast to the original work. We run only the outline generation step of AutoSurvey, instead of generating the whole article.

**STORM** (Shao et al., 2024): The pre-writing stage of this method involves researching the given

topic through simulated conversations. A draft outline is generated initially by prompting the LLM with the topic only. To generate the final outline, the LLM is prompted with the topic, simulated conversations as well as the draft outline. Like in AutoSurvey, here also we provide the pipeline with the reference papers.

**Topic only**: In this baseline, we simply prompt a LLM with the taxonomy topic and ask it to generate a corresponding tree that best fits the topic. The primary goal is to evaluate how effectively the model can generate such a tree using only its parametric knowledge.

**Topic + Keyphrases**: We use a LLM to extract the top keyphrases from each of the reference (cited) papers that form the basis of the taxonomy. This provides the top phrases of each paper, enabling us to fit the essential content of all references within the limited context window of LLMs. We then prompt another LLM to generate the taxonomy tree based on the set of keyphrases.

**TAXOALIGN w/o Taxonomy Verbalization w/o Taxonomy Refinement**: We use the knowledge slices used in our proposed method. We simply prompt the model to generate the taxonomy tree based on these slices. This allows us to isolate and assess the effect of the latter stages of our own method, specifically those that occur after knowledge-slice creation.

## 7 Experimental Setup

### 7.1 Base Models

We use the open-domain **Mistral-7B-Instruct-v0.3** (Jiang et al., 2023) and **Meta-Llama-3-8B-Instruct** (Grattafiori et al., 2024) for extracting the knowledge slices. In the taxonomy verbalization phase, we finetune **Llama-3.1-Tülu-3-8B** (Lambert et al., 2025) and **SciLitLLM1.5-7B** (Li et al., 2024) respectively. For the refinement stage, we use a closed-domain model **GPT-4o-mini**. For the prompting stage in the TAXOALIGN w/o Taxonomy Verbalization w/o Taxonomy Refinement baseline, we also test using three large open-source reasoning models (Qwen's **QwQ-32B** (Team, 2025b), DeepSeek-AI's **DeepSeek-R1-Distill-Qwen-32B** (DeepSeek-AI, 2025) and NovaSky-AI's **Sky-T1-32B**(Team, 2025a)) using the previously generated knowledge slices. Prompts and model details are present in Appendix A. We choose the Tülu and the SciLitLLM models for instruction tuning in TAXOALIGN as well as prompting because they include extensive scientific research-related data in their pre-training or continual pre-training corpus.

### 7.2 Hyperparameter Choices

In the taxonomy verbalization stage, we instruction-tune LLMs using QLoRA (Dettmers et al., 2023). QLoRA uses 4-bit NormalFloat, Double Quantization and Paged Optimizers on the LoRA fine-tuning approach (Hu et al., 2022). Each language model is instruction-tuned for 800 steps with an input context window of $16,384$ and a output context window of $1,024$. The learning rate is $2e-4$ and the training batch size is 1. For instruction-tuning, we use simple intuitive prompts based on training data from CS-TAXOBENCH with Alpaca-like [5]instruction format (Taori et al., 2023). The instruction format is given in the Appendix B.1. We instruct LLMs in our experiments to generate taxonomies with a maximum depth of three. For the taxonomy verbalization part in our method or in any of the baselines, we set $1,024$ as the maximum number of new tokens to be generated by the model. All experiments are done on a single A100.

## 8 Results and Analysis

We evaluated our method and the baseline methods using the proposed metrics. We summarize the results of our experiments in Table 2. Additional results with more models are in Table 5 of Appendix C. We find that the TAXOALIGN w/o Taxonomy Verbalization w/o Taxonomy Refinement baseline performs second best to our method on most metrics. This indicates that the knowledge slices are an important tool for this task. This baseline has high average degree score compared to our method that reveals that our Taxonomy Verbalization and Refinement stages could effectively reduce the gap between the generated and gold taxonomy trees while enhancing the quality of node labels. In our experiments, we find that the BLEU-2, ROUGE-L and BERTScore values are much less than what we typically encounter in tasks like machine translation or question answering. This suggests substantial scope for improvement in this task, as a huge gap remains between the human-created and machine-generated results.

### 8.1 Structural Similarity

Our method consistently achieves an average degree score ($\Delta$) close to 1, while the $\Delta$ value ob-

[5] https://huggingface.co/datasets/tatsu-lab/alpaca

| Method | Model | Δ | Level-order Traversal | | | NSR | NER | LLM judge |
|---|---|---|---|---|---|---|---|---|
| | | | BLEU-2 | ROUGE-L | BERTScore | | | |
| AutoSurvey | Prompt: GPT-4o-mini | 4.4659 | 0.0016 | 0.1784 | 0.8256 | 1.0903 | 0.1982 | 2.4333 |
| STORM | Prompt: GPT-4o-mini | 6.151 | 0.0012 | 0.1349 | 0.8166 | 1.0727 | 0.1539 | 2.2000 |
| Topic only | Prompt: Tülu | **1.4274** | 0.0052 | 0.2359 | 0.8376 | 1.4187 | 0.1373 | 2.0833 |
| Topic + Keyphrases | Keyphrase: LLaMa; Prompt: Tülu | 4.4517 | 0.0018 | 0.1584 | 0.8134 | 1.1103 | 0.1491 | 2.4167 |
| | Keyphrase: Mistral; Prompt: Tülu | 4.91 | 0.0014 | 0.1432 | 0.8100 | 1.0996 | 0.1640 | 2.4167 |
| TAXOALIGN w/o Tax. Verbaliz. w/o Tax. Refine. | K-Slice: LLaMa; Prompt: Tülu | 5.486 | 0.0037 | 0.159 | 0.8123 | 0.9571 | 0.2074 | 2.4833 |
| | K-Slice: Mistral; Prompt: Tülu | 6.1125 | 0.0029 | 0.1465 | 0.8087 | 1.0791 | **0.2197** | 2.4333 |
| | K-Slice: LLaMa; Prompt: Sky-T1-32B | 6.4486 | 0.0020 | 0.1761 | 0.8170 | 1.0804 | 0.2135 | 2.3966 |
| | K-Slice: Mistral; Prompt: Sky-T1-32B | 7.1965 | 0.0022 | 0.1933 | 0.8221 | 1.0948 | 0.2103 | 2.4211 |
| TAXOALIGN | K-Slice: LLaMa; T- Verbal.: Tülu; T-Refine.: GPT-4o-mini | 1.6687 | **0.0132** | **0.2975** | 0.8501 | 1.3244 | 0.1986 | 2.4167 |
| | K-Slice: Mistral; T- Verbal.: Tülu; T-Refine.: GPT-4o-mini | 1.668 | 0.0051 | 0.2974 | **0.8517** | **1.3635** | 0.1872 | **2.5000** |

Table 2: Results of our method, TAXOALIGN, compared with AutoSurvey, STORM, Topic-only, Topic+Keyphrases and TAXOALIGN w/o Taxonomy Verbalization w/o Taxonomy Refinement, on the original test set.

| Method | Model | Δ | Level-order Traversal | | | NSR | NER | LLM judge |
|---|---|---|---|---|---|---|---|---|
| | | | BLEU-2 | ROUGE-L | BERTScore | | | |
| TAXOALIGN w/o Tax. Verbaliz. w/o Tax. Refine. | K-Slice: LLaMa; Prompt: Tulu | 6.361 | 0.0019 | 0.1643 | 0.8182 | 1.0716 | 0.2683 | 2.275 |
| | K-Slice: Mistral; Prompt: Tulu | 7.2083 | 0.0034 | 0.1598 | 0.8159 | 1.0737 | 0.2653 | 2.2125 |
| TAXOALIGN | K-Slice: LLaMa; T-Verbal.: Tulu; T-Refine.: gpt-4o-mini | **2.1924** | 0.0058 | **0.3091** | **0.8542** | 1.2129 | 0.2566 | 2.2875 |
| | K-Slice: Mistral; T-Verbal.: Tulu; T-Refine.: gpt-4o-mini | 2.3617 | 0.013 | 0.3004 | 0.8522 | 1.2072 | **0.2716** | **2.35** |

Table 3: Results of TAXOALIGN compared with TAXOALIGN w/o Taxonomy Verbalization w/o Taxonomy Refinement on the additional test set of conference papers.

tained by other baseline methods are much higher. This indicates that our generated tree is much closer to the human-written taxonomy tree in terms of structure. We observe that the $\Delta$ value is closest to 1 for the Topic-only baseline. This is mainly because when provided with only the topic, the language model generates a small tree due to lack of parametric knowledge, as has been established by the consistent low scores for this baseline on the rest of the metrics. Other baseline methods tend to produce overly large trees ($\Delta > 2.9$) with an excessive number of nodes and branches, increasing the likelihood of hallucinations and structural divergence from the gold-standard tree. We show examples of generated taxonomies using AutoSurvey and STORM in comparison with TAXOALIGN in Figures 5 and 6 respectively in Appendix D.

### 8.2 Semantic Similarity

In terms of the level-order traversal, we observe that our method produces comprehensively better BLEU-2, ROUGE-L and BERTScore in all the cases when compared with the baselines. We observe that all baselines have similarly low scores for level-order traversal, indicating that the generated nodes exhibit low lexical overlap with the gold data. In comparison, our method produces more coherent labels and nodes in the taxonomy tree.

In terms of Node Soft Recall (NSR), our method performs better than the other baselines, showing the similarity between the generated and gold node

labels. The TAXOALIGN w/o Taxonomy Verbaliz. w/o Taxonomy Refine. baseline performs better in some cases in terms of the Node Entity Recall metric, which is mainly because this baseline generates large trees, as has been demonstrated by the $\Delta$ value, and larger trees contribute to greater match in the Noun Phrase chunks. Using LLM-as-a-judge, TAXOALIGN also outperforms the baselines, reaffirming the metric-based results and showing that it generates taxonomies closer to the human-written ones in both structure and intent.

### 8.3 Testing with Conference Survey Papers

We tested our three-phase pipeline (that had been trained on 400-instance training data) on the new test set introduced in Section 3.6, the results for which are given in Table 3. Extended results with more models are present in Table 6 of Appendix C. We observe that our method comfortably outperforms the baseline on this test set too, which is consistent with the results reported in Table 2.

### 8.4 Error Analysis

We present an error analysis based on a manual evaluation of instances from the test set in Table 1 using our proposed TAXOALIGN pipeline. For illustration, we provide a example for each of the three stages of the pipeline in Figure 7 of Appendix E. Below, we summarize the common errors observed at each stage of the pipeline.

**Knowledge Slices + Prompting:** Direct generation from knowledge slices creates more verbose taxonomies that contain irrelevant information leading to the nodes not being very specific or not pertaining to the topic directly. Another major factor is the presence of repeated numbers of the same nodes or sub-trees in the taxonomy.

**Knowledge Slices + Taxonomy Verbalization:** Structurally, the taxonomies are closer to the gold standard taxonomies, but there are some factual errors that persist. We observe that the generated taxonomies suffer from the problem of hallucinated node labels or are too short.

**Knowledge Slices + Taxonomy Verbalization + Taxonomy Refinement:** The generated trees are more aligned to the gold standard trees in terms of structure and semantic coherence. Still, the generated trees suffer from a low number of layer-wise exact matches. The generated trees are certainly more interpretable than the previous stages and the overall tree also presents a coherent structure.

| Method | Structure | Content |
|---|---|---|
| TAXOALIGN | 3.17 | 2.62 |
| AutoSurvey | 2.17 | 2.25 |

Table 4: Mean ratings from human evaluation of the structure and content similarity for TAXOALIGN and AutoSurvey.

## 9 Human Evaluation

We use human evaluation to complement the automated framework. Three annotators with domain knowledge were asked to rate the surveys generated by TAXOALIGN and AutoSurvey. The annotators are instructed to assess based on (1) the structural commonalities between the gold and generated taxonomy trees and then (2) the semantic coherence of the generated tree with respect to the gold tree. The evaluation is done using a 5-point Likert scale on 20 randomly sampled data instances from the test set of CS-TAXOBENCH. The inter-annotator agreement is calculated as 0.61 and 0.73 (Krippendorff's $\alpha$). The results are shown in Table 4. The mean ratings show TAXOALIGN outperforms AUTOSURVEY in both structural and content quality.

To verify the consistency between our LLM-as-a-judge evaluation and the human evaluation, we first average the scores assigned by human annotators for each taxonomy tree. We then compare these with the LLM-generated scores using Spearman's rank correlation coefficient. Thereby, we obtain a Spearman's rho value of 0.527 which indicates a strong positive correlation. These results suggest that our LLM-as-a-judge evaluation method aligns well with human preferences, providing a reliable proxy for human judgment.

## 10 Conclusion

Automating scholarly taxonomy generation can help researchers and practitioners efficiently navigate the vast body of scholarly literature. To facilitate this, we present CS-TAXOBENCH, a benchmark comprising 460 taxonomies from journey surveys and 80 from conference surveys, along with TAXOALIGN, a method that uses instruction tuning and refinement on topic-related information extracted from papers. We introduce two new metrics and show that TAXOALIGN outperforms the baselines on most evaluation measures.

## Limitations

We construct CS-TAXOBENCH from a single journal within a defined time frame to ensure consistency in taxonomy quality. However, additional open-access journals and conference venues could also be explored for future curation.

We do not focus on the retrieval of the reference papers from a corpus of papers. While this is an important task for end-to-end taxonomy construction given only the taxonomy topic, we focus more on creating and evaluating taxonomies when provided with a set of reference documents.

Although we improve the structure and semantic coherence between the human-written and the generated taxonomies using TAXOALIGN, there is a a lot of scope for improvement in this field. Therefore, this is a encouraging field of work in which the community can work in the coming days.

## References


Steven S. Aanen, Damir Vandic, and Flavius Frasincar. 2015. Automated product taxonomy mapping in an e-commerce environment. *Expert Syst. Appl.*, 42(3):1298–1313.

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota. Association for Computational Linguistics.

Hengxing Cai, Xiaochen Cai, Junhan Chang, Sihang Li, Lin Yao, Wang Changxin, Zhifeng Gao, Hongshuai Wang, Li Yongge, Mujie Lin, Shuwen Yang, Jiankun Wang, Mingjun Xu, Jin Huang, Xi Fang, Jiaxi Zhuang, Yuqi Yin, Yaqi Li, Changhong Chen, Zheng Cheng, Zifeng Zhao, Linfeng Zhang, and Guolin Ke. 2025. SciAssess: Benchmarking LLM proficiency in scientific literature analysis. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2335–2357, Albuquerque, New Mexico. Association for Computational Linguistics.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

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

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

Steffen Eger, Yong Cao, Jennifer D'Souza, Andreas Geiger, Christian Greisinger, Stephanie Gross, Yufang Hou, Brigitte Krenn, Anne Lauscher, Yizhi Li, Chenghua Lin, Nafise Sadat Moosavi, Wei Zhao, and Tristan Miller. 2025. Transforming science with large language models: A survey on ai-assisted scientific discovery, experimentation, content generation, and evaluation. *Preprint*, arXiv:2502.05151.

Pasi Fränti and Radu Mariescu-Istodor. 2023. Soft precision and recall. *Pattern Recognition Letters*, 167:115–121.

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

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, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer,

Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai

Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

SANDA M. Harabagiu, STEVEN J. MAIORANO, and MARIUS A. PAŞCA. 2003. Open-domain textual question answering techniques. *Natural Language Engineering*, 9(3):231–267.

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

Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, and Debasis Ganguly. 2019. Identification of tasks, datasets, evaluation metrics, and numeric scores for scientific leaderboards construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5203–5213, Florence, Italy. Association for Computational Linguistics.

Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, and Debasis Ganguly. 2021. TDMSci: A specialized corpus for scientific literature entity tagging of tasks datasets and metrics. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 707–714, Online. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

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

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

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Sergio Jimenez, Fabio Gonzalez, and Alexander Gelbukh. 2010. Text comparison using soft cardinality. In *String Processing and Information Retrieval*, pages 297–302, Berlin, Heidelberg. Springer Berlin Heidelberg.

Hao Kang and Chenyan Xiong. 2025. Researcharena: Benchmarking large language models' ability to collect and organize information as research agents. *Preprint*, arXiv:2406.10291.

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

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

Sebastian Köhler, Sandra C. Doelken, Christopher J. Mungall, Sebastian Bauer, Helen V. Firth, Isabelle Bailleul-Forestier, Graeme C. M. Black, Danielle L. Brown, Michael Brudno, Jennifer Campbell, David R. FitzPatrick, Janan T. Eppig, Andrew P. Jackson, Kathleen Freson, Marta Girdea, Ingo Helbig, Jane A. Hurst, Johanna Jähn, Laird G. Jackson, Anne M. Kelly, David H. Ledbetter, Sahar Mansour, Christa L. Martin, Celia Moss, Andrew Mumford, Willem H. Ouwehand, Soo-Mi Park, Erin Rooney Riggs, Richard H. Scott, Sanjay Sisodiya, Steven Van Vooren, Ronald J. Wapner, Andrew O. M. Wilkie, Caroline F. Wright, Anneke T. Vulto-van Silfhout, Nicole de Leeuw, Bert B. A. de Vries, Nicole L. Washingthon, Cynthia L. Smith, Monte Westerfield, Paul Schofield, Barbara J. Ruef, Georgios V. Gkoutos, Melissa Haendel, Damian Smedley, Suzanna E. Lewis, and Peter N. Robinson. 2013. The human phenotype ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Research*, 42(D1):D966–D974.

Yuxuan Lai, Yupeng Wu, Yidan Wang, Wenpeng Hu, and Chen Zheng. 2024. Instruct large language models to generate scientific literature survey step by step. *Preprint*, arXiv:2408.07884.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. 2025. Tulu 3: Pushing frontiers in open language model post-training. *Preprint*, arXiv:2411.15124.

Sihang Li, Jin Huang, Jiaxi Zhuang, Yaorui Shi, Xiaochen Cai, Mingjun Xu, Xiang Wang, Linfeng Zhang, Guolin Ke, and Hengxing Cai. 2024. Scilitllm: How to adapt llms for scientific literature understanding. *Preprint*, arXiv:2408.15545.

Xun Liang, Jiawei Yang, Yezhaohui Wang, Chen Tang, Zifan Zheng, Shichao Song, Zehao Lin, Yebin Yang, Simin Niu, Hanyu Wang, Bo Tang, Feiyu Xiong, Keming Mao, and Zhiyu li. 2025. Surveyx: Academic survey automation via large language models. *Preprint*, arXiv:2502.14776.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

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

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Yuning Mao, Tong Zhao, Andrey Kan, Chenwei Zhang, Xin Luna Dong, Christos Faloutsos, and Jiawei Han. 2020. Octet: Online catalog taxonomy enrichment with self-supervision. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 2247–2257, New York, NY, USA. Association for Computing Machinery.

Félix Martel and Amal Zouaq. 2021. Taxonomy extraction using knowledge graph embeddings and hierarchical clustering. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, SAC '21, page 836–844, New York, NY, USA. Association for Computing Machinery.

Ishani Mondal, Yufang Hou, and Charles Jochim. 2021. End-to-end construction of NLP knowledge graph. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1885–1895, Online. Association for Computational Linguistics.

Benjamin Newman, Yoonjoo Lee, Aakanksha Naik, Pao Siangliulue, Raymond Fok, Juho Kim, Daniel S Weld, Joseph Chee Chang, and Kyle Lo. 2024. ArxivDIGESTables: Synthesizing scientific literature into tables using language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9612–9631, Miami, Florida, USA. Association for Computational Linguistics.

R. Nickerson, U. Varshney, and J. Muntermann. 2013. A method for taxonomy development and its application in information systems. *European Journal of Information Systems*, 22(3):336–359.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

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

Massimiliano Pronesti, Joao H Bettencourt-Silva, Paul Flanagan, Alessandra Pascale, Oisín Redmond, Anya Belz, and Yufang Hou. 2025a. Query-driven document-level scientific evidence extraction from biomedical studies. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28034–28051, Vienna, Austria. Association for Computational Linguistics.

Massimiliano Pronesti, Michela Lorandi, Paul Flanagan, Oisin Redmon, Anya Belz, and Yufang Hou. 2025b. Enhancing study-level inference from clinical trial papers via rl-based numeric reasoning. *Preprint*, arXiv:2505.22928.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

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

Furkan Şahinuç, Thy Thy Tran, Yulia Grishina, Yufang Hou, Bei Chen, and Iryna Gurevych. 2024. Efficient performance tracking: Leveraging large language models for automated construction of scientific leaderboards. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language*

*Processing*, pages 7963–7977, Miami, Florida, USA. Association for Computational Linguistics.

Yijia Shao, Yucheng Jiang, Theodore Kanell, Peter Xu, Omar Khattab, and Monica Lam. 2024. Assisting in writing Wikipedia-like articles from scratch with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6252–6278, Mexico City, Mexico. Association for Computational Linguistics.

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

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Deep Search Team. 2024. Docling technical report. Technical report.

NovaSky Team. 2025a. Sky-t1: Train your own o1 preview model within $450. https://novasky-ai.github.io/posts/sky-t1. Accessed: 2025-01-09.

Qwen Team. 2025b. Qwq-32b: Embracing the power of reinforcement learning.

Roelien C Timmer, Yufang Hou, and Stephen Wan. 2025. A position paper on the automatic generation of machine learning leaderboards. *Preprint*, arXiv:2505.17465.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

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

David Wadden, Kejian Shi, Jacob Morrison, Aakanksha Naik, Shruti Singh, Nitzan Barzilay, Kyle Lo, Tom Hope, Luca Soldaini, Shannon Zejiang Shen, Doug Downey, Hannaneh Hajishirzi, and Arman Cohan. 2024. Sciriff: A resource to enhance language model instruction-following over scientific literature. *Preprint*, arXiv:2406.07835.

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

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

Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Min Zhang, Qingsong Wen, Wei Ye, Shikun Zhang, and Yue Zhang. 2024. Autosurvey: Large language models can automatically write surveys. In *Advances in Neural Information Processing Systems*, volume 37, pages 115119–115145. Curran Associates, Inc.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023. How far can camels go? exploring the state of instruction tuning on open resources. *Preprint*, arXiv:2306.04751.

Xiangchao Yan, Shiyang Feng, Jiakang Yuan, Renqiu Xia, Bin Wang, Bo Zhang, and Lei Bai. 2025. Surveyforge: On the outline heuristics, memory-driven generation, and multi-dimensional evaluation for automated survey writing. *Preprint*, arXiv:2503.04629.

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.

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

# A Our Method – Prompts and Models

## A.1 Models

- **Mistral-7B-Instruct-v0.3** (Jiang et al., 2023): The Mistral group of models leverages grouped-query attention (GQA) for faster inference, coupled with sliding window attention (SWA) to effectively handle sequences of arbitrary length with a reduced inference cost.

Compared to version 0.2, this model can process and respond more effectively to diverse tasks and instructions, owing to its expanded vocabulary of 32,768 tokens and support for the v3 tokenizer. The model can carry out operations that call for outside data since it supports function calling.

- **Meta-Llama-3-8B-Instruct** (Grattafiori et al., 2024): Llama is a family of pre-trained foundational language models that have been open-sourced by Meta in recent times. The Meta-Llama-3-8B-Instruct is trained on a mix of publicly available online data with a knowledge cutoff of March, 2023. The tuned versions of Llama3 use Supervised Fine-Tuning (SFT) and Reinforcement Learning with Human Feedback (RLHF) to align with human preferences.

- **Llama-3.1-Tülu-3-8B** (Lambert et al., 2025): Tülu (Wang et al., 2023) is a set of models that are instruction-tuned on LLaMA (Touvron et al., 2023) using a mixture of publicly available, synthetic and human-created datasets. Building upon the Llama 3.1 basic models, Tülu-3 (Lambert et al., 2025) models are trained using Direct Preference Optimization (DPO), Supervised Fine-Tuning (SFT), and a technique called Reinforcement Learning with Verifiable Rewards (RLVR).

- **SciLitLLM1.5-7B** (Li et al., 2024): It is a very recently released LLM designed for the task of scientific literature understanding that has been trained using both Continual Pre-Training (CPT) and Supervised Fine-Tuning (SFT).This strategy is used on Qwen2.5 to obtain SciLitLLM. The CPT stage uses 73,000 textbooks and 625,000 academic papers, while the SFT stage uses SciLitIns, SciRIFF (Wadden et al., 2024) and Infinity-Instruct[6]. We use the SciLitLLM 7B[7] for our experimental purposes.

- **QwQ-32B** (Team, 2025b): QwQ is designed for complex problem-solving and logical reasoning tasks and is based on Qwen2.5. The model is text-only and focuses on tasks like multi-step reasoning, complex decision-making, and research assistance.

- DeepSeek-AI's **DeepSeek-R1-Distill-Qwen-32B** (DeepSeek-AI, 2025): DeepSeek-R1-Distill-Qwen-32B is an open-source, distilled large language model (LLM) based on the Qwen2.5 32B architecture, utilizing the knowledge from the DeepSeek-R1 reasoning model. It is optimized for language understanding, reasoning, and text generation tasks and is known for outperforming other open-source models, including OpenAI's o1-mini, on various industry benchmarks.

- **Sky-T1-32B** (Team, 2025a): This model has been developed by the NovaSky team at UC Berkeley. It excels in mathematical and coding reasoning, outperforming some advanced closed-source models and other open-source alternatives on various benchmarks. The model was created by fine-tuning the Qwen 2.5 32B instruct model with a high-quality, 17,000-item dataset.

### A.2 Knowledge Slice-Prompt

You will receive a document and a topic. Your task is to identify the knowledge-slices within the document that are very relevant to the given topic. A knowledge-slice is a piece of information representing the highlights of the document related to the given topic i.e. each knowledge-slice should be such that it both represents an important point in the document, but at the same time, the knowledge-slice should pertain closely to the given topic. Also, the knowledge-slice should not represent any additional information that is not present in the document.
[Document]
document-text
[Topic]
taxonomy-topic
Please ONLY return the relevant knowledge-slices in the form of a list enclosed within square brackets. Your response should be in the following format:
[Knowledge-Slices]
[Knowledge-Slice 1, Knowledge-Slice 2,..., Knowledge-Slice n]
[Your response]

[6] https://huggingface.co/datasets/BAAI/Infinity-Instruct
[7] https://huggingface.co/Uni-SMART/SciLitLLM

### A.3 Taxonomy Verbalization-Prompt

A taxonomy is a tree-structured semantic hierarchy that establishes a classification of the existing literature under a common topic. You will receive a taxonomy topic along with a collection of documents. Your task is to create a taxonomy tree using the given topic and based on the highlights of the documents i.e. create new child nodes by identifying generalizable sub-level topics from the document highlights that can act as child nodes to the taxonomy topic, which acts as the root node . The taxonomy tree should be created such that it looks as if all the given documents are a part of the taxonomy. There may be several levels in the tree i.e. each node may contain child nodes, but the total depth of the tree should not exceed three. The topics in all the levels of the tree except the last level must not be too specific so that it can accommodate future sub-topics i.e. child nodes.
- The nodes at the last level of the hierarchy i.e. the leaf nodes should reflect a single topic instead of a combination of topics.
- Each node label is a small and concise phrase.
[Response Format Instructions]
- The output tree is to be formatted as shown in the example such that the root node is the taxonomy topic and each child node is connected to its parent.
[Example Output]
example-output
[Taxonomy Topic]
taxonomy-topic
[Documents]
Doc-1
Doc-2
Doc-3
Please ONLY return the taxonomy tree in the output format as shown in the example above.
[Your response]

### A.4 Taxonomy Refinement- Prompt

A taxonomy is a tree-structured semantic hierarchy that establishes a classification of the existing literature under a common topic. You will receive a taxonomy tree along with a collection of documents. The root node of the taxonomy tree is the overall taxonomy topic. Your task is to refine the taxonomy tree such that there is a clear connection between the parent node and the subsequent child nodes. Each node must be a well-defined topic that is grounded in the input document highlights. Do not alter the root node of the tree i.e. the taxonomy topic. Your task is to alter the other nodes only if deemed necessary i.e. only if a better viable replacement is found. Please try to adhere to the structure of the given taxonomy tree as much as possible. Only if the given taxonomy tree is restricted to less than five nodes, then generate the taxonomy tree on your own. Strictly adhere to the format of the tree shown here.
[Example Output]
example-output
[Taxonomy Topic]
taxonomy-topic
[Documents]
Doc-1
Doc-2
Doc-3
Please ONLY return the edited taxonomy tree in the output format as shown in the example above.
[Your response]

## B LLM-as-a-Judge Prompt

A taxonomy is a tree-structured semantic hierarchy that establishes a classification of the existing literature under a common topic. You are given a gold standard taxonomy tree and a generated taxonomy tree and your task is to respond with an appropriate score after comparing the two. Two taxonomy trees are said to be structurally similar if the number of nodes and branches are similar in number. If one tree has too many or too less nodes and branches than the gold tree, then they

are said to be structurally dissimilar. Two taxonomy trees are said to be semantically similar if their nodes have values with close meanings or are matching entirely. Please respond with only the score based on the following criteria:
Score 1: The generated taxonomy has no similarity at all with the gold standard taxonomy i.e. the structure and the intent of the generated taxonomy is totally different from that of the gold standard taxonomy.
Score 2: The generated taxonomy have only a few nodes that has a semantic match with the nodes in the gold standard taxonomy and the structure of the generated taxonomy is a little similar to that of the gold standard taxonomy. The structure of the generated tree is very less similar to the gold standard tree but the intent of both taxonomies is similar.
Score 3: The generated taxonomy has a reasonable similarity to the generated taxonomy in terms of structural similarity and semantic similarity. The structure of both trees are similar but some nodes are different in the two taxonomies.
Score 4: The generated taxonomy has good logical consistency with that of the gold standard taxonomy in terms of semantic matching of the nodes between the two with the structure of the generated taxonomy is very similar to that of the gold standard taxonomy. The two taxonomies only differ for a small number of instances.
Score 5: The generated taxonomy is fully similar in terms of semantic matching and structure to the gold standard taxonomy.
Gold Standard Taxonomy:
gold-taxonomy
Generated Taxonomy:
generated-taxonomy
[Your Response]

### B.1 Instruction Format for Finetuning

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.
### Instruction:
[Instruction prompt (present in Appendix A.2)]
### Input:
[Knowledge slices]
### Response:
[Gold Standard Taxonomy Tree]

## C Extended Results

We show additional results using a expanded set of models on the original test set and the additional conference paper test in Tables 5 and 6 respectively.

## D Output Example Comparison

We see in Figure 5 and 6 that the taxonomy trees generated using TAXOALIGN are much less verbose than the corresponding taxonomy trees generated using AutoSurvey or STORM.

## E Error Analysis

We show an example of the results obtained in the three stages of our TAXOALIGN pipeline in Figure 7. The stages are Knowledge Slices + Prompting, Knowledge Slices + Taxonomy Verbalization and Knowledge Slices + Taxonomy Verbalization + Taxonomy Refinement.

| Method | Model | Δ | Level-order Traversal | | | NSR | NER | LLM judge |
|---|---|---|---|---|---|---|---|---|
| | | | BLEU-2 | ROUGE-L | BERTScore | | | |
| AutoSurvey | Prompt: GPT-4o-mini | 4.4659 | 0.0016 | 0.1784 | 0.8256 | 1.0903 | 0.1982 | 2.4333 |
| STORM | Prompt: GPT-4o-mini | 6.151 | 0.0012 | 0.1349 | 0.8166 | 1.0727 | 0.1539 | 2.2000 |
| Topic only | Prompt: Tülu | **1.4274** | 0.0052 | 0.2359 | 0.8376 | 1.4187 | 0.1373 | 2.0833 |
| Topic + Keyphrases | Keyphrase: LLaMa; Prompt: Tülu | 4.4517 | 0.0018 | 0.1584 | 0.8134 | 1.1103 | 0.1491 | 2.4167 |
| | Keyphrase: LLaMa; Prompt: SciLitLLM | 8.0766 | 0.0022 | 0.192 | 0.8168 | 1.2170 | 0.1578 | 1.6833 |
| | Keyphrase: Mistral; Prompt: Tülu | 4.91 | 0.0014 | 0.1432 | 0.8100 | 1.0996 | 0.1640 | 2.4167 |
| | Keyphrase: Mistral; Prompt: SciLitLLM | 6.6771 | 0.0029 | 0.1676 | 0.8084 | 1.2522 | 0.1670 | 1.6500 |
| TAXOALIGN w/o Taxonomy Verbalization w/o Taxonomy Refinement | K-Slice: LLaMa; Prompt: Tülu | 5.486 | 0.0037 | 0.159 | 0.8123 | 0.9571 | 0.2074 | 2.4833 |
| | K-Slice: LLaMa; Prompt: SciLitLLM | 2.9139 | 0.0058 | 0.1964 | 0.823 | 1.2968 | 0.1619 | 2.1000 |
| | K-Slice: Mistral; Prompt: Tülu | 6.1125 | 0.0029 | 0.1465 | 0.8087 | 1.0791 | **0.2197** | 2.4333 |
| | K-Slice: Mistral; Prompt: SciLitLLM | 3.3845 | 0.0033 | 0.2122 | 0.8206 | 1.3194 | 0.1504 | 2.0167 |
| | K-Slice: LLaMa; Prompt: QwQ-32B | 5.4111 | 0.0019 | 0.1545 | 0.8042 | 1.0791 | 0.1958 | 2.4500 |
| | K-Slice: Mistral; Prompt: QwQ-32B | 5.8538 | 0.0019 | 0.1503 | 0.8078 | 1.0944 | 0.2066 | 2.4071 |
| | K-Slice: LLaMa; Prompt: DeepSeek-R1-Dist.-Qwen-32B | 6.7846 | 0.0016 | 0.1428 | 0.8037 | 1.0514 | 0.2087 | 2.2807 |
| | K-Slice: Mistral; Prompt: DeepSeek-R1-Dist.-Qwen-32B | 7.2543 | 0.0018 | 0.1489 | 0.8092 | 0.8434 | **0.2255** | 2.2143 |
| | K-Slice: LLaMa; Prompt: Sky-T1-32B | 6.4486 | 0.0020 | 0.1761 | 0.8170 | 1.0804 | 0.2135 | 2.3966 |
| | K-Slice: Mistral; Prompt: Sky-T1-32B | 7.1965 | 0.0022 | 0.1933 | 0.8221 | 1.0948 | 0.2103 | 2.4211 |
| TAXOALIGN | K-Slice: LLaMa; T- Verbal.: Tülu; T-Refine.: GPT-4o-mini | 1.6687 | **0.0132** | **0.2975** | 0.8501 | 1.3244 | 0.1986 | 2.4167 |
| | K-Slice: LLaMa; T- Verbal.: SciLitLLM; T-Refine.: GPT-4o-mini | 1.7358 | 0.0081 | 0.29 | 0.8484 | 1.2956 | 0.1875 | 2.4833 |
| | K-Slice: Mistral; T- Verbal.: Tülu; T-Refine.: GPT-4o-mini | 1.668 | 0.0051 | 0.2974 | **0.8517** | **1.3635** | 0.1872 | **2.5000** |
| | K-Slice: Mistral; T- Verbal.: SciLitLLM; T-Refine.: GPT-4o-mini | 2.1709 | 0.0053 | 0.284 | 0.8484 | 1.265 | 0.1966 | 2.4833 |

Table 5: Results of our method compared with baselines like AutoSurvey, Topic-only, Topic+Keyphrases andTAXOALIGN w/o Taxonomy Verbalization w/o Taxonomy Refinement.

| Method | Model | Δ | Level-order Traversal | | | NSR | NER | LLM judge |
|---|---|---|---|---|---|---|---|---|
| | | | BLEU-2 | ROUGE-L | BERTScore | | | |
| TAXOALIGN w/o Tax. Verbaliz. w/o Tax. Refine. | K-Slice: LLaMa; Prompt: Tulu | 6.361 | 0.0019 | 0.1643 | 0.8182 | 1.0716 | 0.2683 | 2.275 |
| | K-Slice: LLaMa; Prompt: SciLitLLM | 6.5221 | 0.0026 | 0.1957 | 0.8183 | 1.2095 | 0.2267 | 1.9375 |
| | K-Slice: Mistral; Prompt: Tulu | 7.2083 | 0.0034 | 0.1598 | 0.8159 | 1.0737 | 0.2653 | 2.2125 |
| | K-Slice: Mistral; Prompt: SciLitLLM | 5.9387 | 0.0032 | 0.2039 | 0.8211 | **1.3243** | 0.2176 | 1.875 |
| TAXOALIGN | K-Slice: LLaMa; T-Verbaliz.: Tulu; T-Refine.: gpt-4o-mini | **2.1924** | 0.0058 | **0.3091** | **0.8542** | 1.2129 | 0.2566 | 2.2875 |
| | K-Slice: LLaMa; T-Verbaliz.: SciLitLLM; T-Refine.: gpt-4o-mini | 2.5551 | **0.0127** | 0.3034 | 0.851 | 1.1927 | 0.2614 | 2.2625 |
| | K-Slice: Mistral; T-Verbaliz.: Tulu; T-Refine.: gpt-4o-mini | 2.3617 | 0.013 | 0.3004 | 0.8522 | 1.2072 | **0.2716** | **2.35** |
| | K-Slice: Mistral; T-Verbaliz.: SciLitLLM; T-Refine.: gpt-4o-mini | 3.1779 | 0.0042 | 0.2845 | 0.8465 | 1.1806 | 0.267 | 2.3125 |

Table 6: Results of TAXOALIGN compared with TAXOALIGN w/o Taxonomy Verbalization w/o Taxonomy Refinement.


Figure 5: A comparison of a gold standard taxonomy tree and a generated taxonomy tree using AutoSurvey.

**Gold Standard Taxonomy**

**Human Image Generation: A Comprehensive Survey**

|-- *DATA-DRIVEN METHODS ON HUMAN IMAGE GENERATION*
| |-- Method Taxonomy Based on Fundamental Models
| |-- Method Taxonomy Based on Task Settings
| +-- Main Components in Data-Driven Methods
|-- *HYBRID METHODS*
|-- *KNOWLEDGE-GUIDED METHODS ON*
| | *HUMAN IMAGE GENERATION*
| |-- Fundamental Models of Knowledge-Guided Methods
| |-- Pixel Warping Pipeline
| +-- Virtual Rendering Pipeline
|-- *APPLICATIONS*
+-- *DATASETS AND EVALUATION METRICS*
|-- Human Image Generation Datasets
|-- Human Image Generation Evaluation Metrics
+-- Qualitative and Quantitative Comparison of Representative Methods

**STORM**

**Human Image Generation: A Comprehensive Survey**

|-- Introduction
| |-- Definition of Human Image Generation
| |-- Importance and Applications
| +-- Overview of Current Technologies
|-- Historical Background
| |-- Early Techniques
| |-- Traditional Artistic Methods
| +-- Early Photography
| |-- Evolution of Technology
| |-- Digital Revolution
| +-- Rise of Computational Techniques
| +-- Key Milestones
| |-- Notable Projects and Innovations
| +-- Influential Figures in the Field
|-- Methods of Human Image Generation
| |-- Traditional Methods
| |-- Manual Drawing and Painting
| +-- Photography and Digital Manipulation
| |-- Computational Methods
| |-- Procedural Generation
| |-- Generative Adversarial Networks (GANs)
| |-- Neural Networks
| +-- Variational Autoencoders (VAEs)
| +-- Hybrid Approaches
| +-- Combining Traditional and Computational Techniques
|-- Applications
| |-- Entertainment and Media
| |-- Video Games
| |-- Movies and Animation
| +-- Virtual Influencers and Digital Celebrities
| |-- Fashion and Advertising
| |-- Digital Fashion Design
| +-- Personalized Marketing
| |-- Medical and Scientific Visualization
| |-- Anatomical Visualization
| +-- Simulation in Medical Training
| |-- Art and Creative Expression
| |-- Digital Art Communities
| +-- Interactive Art Installations
| +-- Other Emerging Applications
| |-- Education and Training
| +-- Social Media and Content Creation
|-- Challenges and Limitations
| |-- Technical Challenges
| |-- Quality and Realism
| +-- Computational Resources
| |-- Ethical Concerns
| |-- Copyright and Ownership
| +-- Representation and Bias
| +-- Legal Issues
| |-- Privacy Considerations
| +-- Regulation of AI-generated Content
|-- Future Trends
| |-- Advances in AI and Machine Learning
| |-- Improved Algorithms and Techniques
| +-- Accessibility of Image Generation Tools
| |-- Integration with Augmented and Virtual Reality
| |-- Enhanced User Experiences
| +-- New Platforms for Creativity
| +-- Potential Impact on Society
| |-- Implications for Employment and Industry
| +-- Cultural Shifts in Art and Media
+-- Conclusion
|-- Final Thoughts on the Future of Human Image Generation
|-- Academic Journals
|-- Books
|-- Online Resources
|-- Related Organizations
+-- Notable Conferences and Workshops

Figure 6: A comparison of a gold standard taxonomy tree and a generated taxonomy tree using STORM.

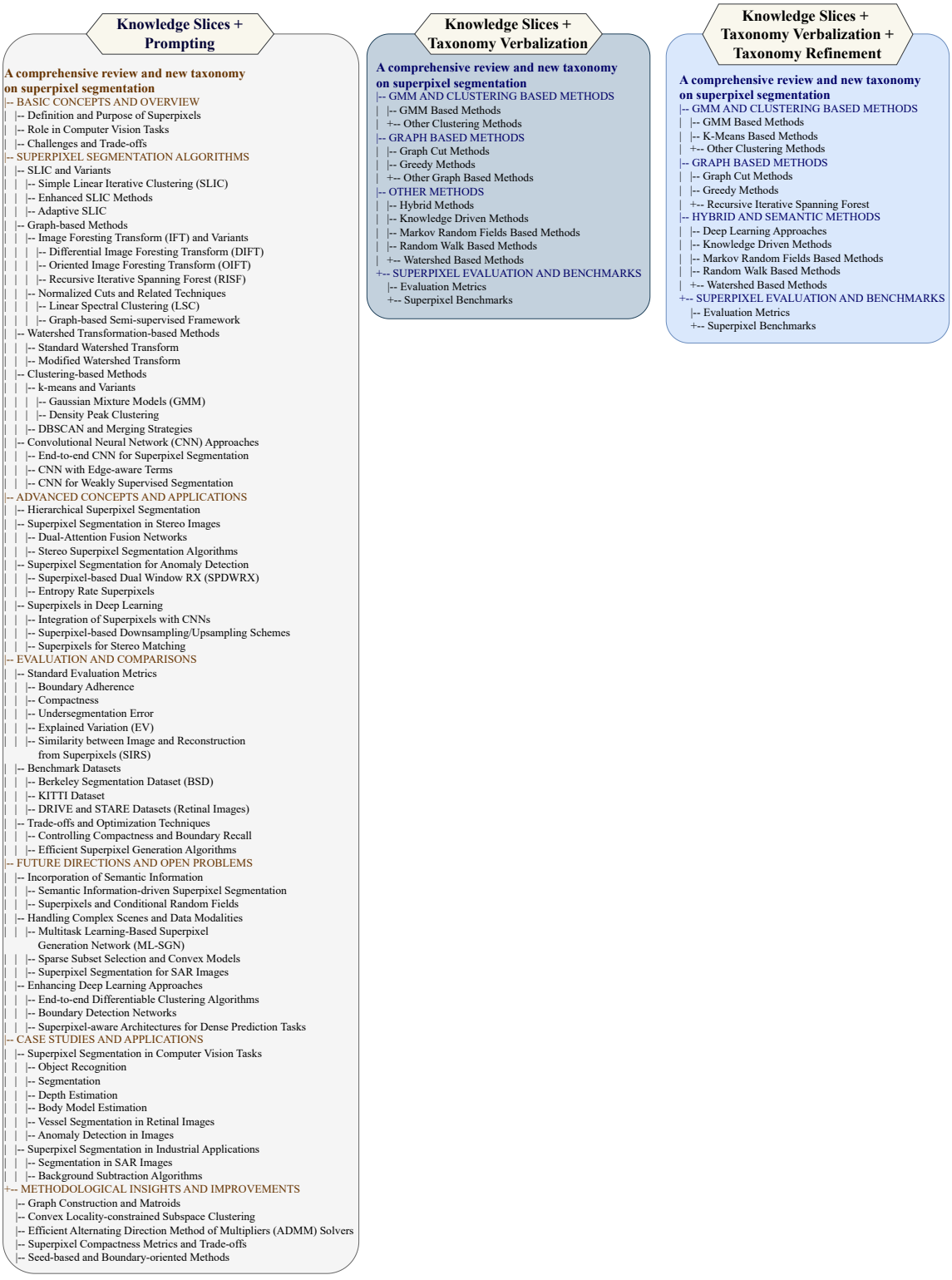


Figure 7: A comparison of the results at the end of each stage of the TAXOALIGN pipeline for the topic "A comprehensive review and new taxonomy on super-pixel segmentation". The generated taxonomy shown here uses Mistral-7B-Instruct-v0.3 for creation of knowledge slices and Llama-3.1-Tülu-3-8B for the taxonomy verbalization component and GPT-4o-mini for refining the generated taxonomy.