

Large Language Models for Creation, Enrichment and Evaluation of Taxonomic Graphs

Semantic Web
Vol. 17(1) 1–30
© The Author(s) 2025
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/22104968251404186
journals.sagepub.com/home/swj



Viktor Moskvoretskii^{1*}, Irina Nikishina^{2*}, Ekaterina Neminova³, Alina Lobanova³, Alexander Panchenko^{4,5} and Chris Biemann²

Abstract

Taxonomies play a crucial role in organizing knowledge for various natural language processing tasks. Recent advancements in large language models (LLMs) have opened new avenues for automating taxonomy-related tasks with greater accuracy. In this paper, we explore the potential of contemporary LLMs in learning, evaluating and predicting taxonomic relations across multiple lexical semantic tasks. We propose novel method for creation of taxonomy-based instruction datasets. With the use of this dataset based on WordNet we build TaxoLLaMA, a unified model fine-tuned designed to handle a wide range of taxonomy-related tasks such as taxonomy construction and enrichment, hypernym discovery, and lexical entailment. The experimental results demonstrate that our model achieves state-of-the-art performance on 11 out of 16 tasks and ranked second on 4 other tasks. We also explore LLM ability for constructed taxonomies graph refinement and present comprehensive ablation study and thorough error analysis supported by both manual and automated techniques.

Keywords

taxonomy, WordNet, lexical semantics, semantic graph, taxonomy construction, lexical entailment

Received: November 18, 2023; accepted: November 18, 2025

Editor: Guest Editors KG Construction 2024

Solicited reviewers: John McCrae, Associate Professor at the Data Science Institute, Insight Centre for Data Analytics and ADAPT centre at the University of Galway, Ireland; Pablo Calleja, Professor of Computer Science, Boadilla del Monte, Madrid, Spain; Four anonymous reviews

1 Introduction

One of the core ideas of Semantic Web is to extend the current Web by enabling machines to understand and respond to complex human requests based on the meaning of various data objects, rather than shallow representations, such as keywords. Central to this vision is the structuring of data in a way that allows for meaningful interconnections between different data points, such as documents, images, or concepts. Taxonomies, which classify and organize concepts into

¹École Polytechnique Fédérale de Lausanne, Switzerland

²Universität Hamburg, Germany

³HSE University, Russia

⁴Skolkovo Institute of Science and Technology, Russia

⁵Artificial Intelligence Research Institute, Russia

*Equal contribution. Corresponding author. Work was done while at Skoltech.

Corresponding Author:

Viktor Moskvoretskii, École Polytechnique Fédérale de Lausanne, Switzerland.

Email: viktor.moskvoretskii@epfl.ch



hierarchical structures, are essential to this process as they provide the backbone for organizing information in a way that is both accessible and meaningful. By categorizing data into well-defined classes and relationships, taxonomies facilitate the creation of ontologies, which are more complex frameworks that define the relationships between concepts in the Semantic Web. Such ontologies are supposed to enable more accurate data retrieval, allowing for richer, more nuanced interactions with web content. In essence, taxonomies serve as the building blocks of the Semantic Web, providing the necessary structure for data. Thus integration of taxonomies into the Semantic Web framework could enhance ability to handle complex queries, making it a more powerful tool for knowledge discovery and data management.

More formally, taxonomy is a directed acyclic graph that organizes concepts through various relationships, with each node representing a specific concept connected to others via IS-A relations. Prime examples of such a taxonomy is Princeton WordNet for the English language (Miller, 1998)¹ (further will be referred as the WordNet) or Open English WordNet (McCrae et al., 2019)² which is a project further developing of the WordNet in an open source collaborative manner.³ For an easy interoperability with the Semantic Web the latter resource is shared in various formats, including RDF (Turtle) format.⁴ WordNets for other languages are also available and many of them being maintained under the Global WordNet Association.⁵ These semantic graphs connect terms into a hierarchical structure via hyponymy/hypernymy relations, while also featuring other types of relations, such as synonymy, antonymy, metonymy, troponymy, etc. WordNet not only includes nodes but also provides definitions, multiple lemmas, and unique sense numbers to distinguish between different meanings within the same synset.

The use of taxonomies is well-justified in various NLP tasks, including entity linking (Corro et al., 2015), named entity recognition (Toral & Muñoz, 2006), and several others (Lenz & Bergmann, 2023; Wang et al., 2023). Yet, despite the widespread adoption of large language models (LLMs), taxonomies continue to be constructed and curated primarily through the manual efforts of such as language experts, for example, linguists or lexicographers, but also enthusiasts and crowdworkers. Earlier neural approaches to natural language processing have struggled to automate taxonomy construction effectively, but this limitation may not apply to the latest generation of LLMs. While some research has shown that Transformer models underperform in this area, these studies were conducted using much less powerful language models than those available today (Hanna & Mareček, 2021; Nikishina et al., 2022a; Radford et al., 2019).

Recent studies of LLMs highlight their impressive capacity to internally store vast amounts of knowledge (Kauf et al., 2023; Sun et al., 2024; Tang et al., 2023). Additionally, as these models scaled up, they demonstrated emerging in-context learning abilities, enabling rapid adaptation to new tasks (Dong et al., 2024). These observations suggest that LLMs could be also effectively leveraged for lexical semantic tasks, such as taxonomy construction. However, despite some previous attempts to apply LLMs in this domain, research remains limited. The few studies that have explored LLMs for lexical semantics have primarily focused on hyponymy and hypernymy relationships, with little attention given to other types of graph relations (Chernomorchenko et al., 2024; Nikishina et al., 2023, 2022b).

Moreover, these studies have generally been limited to hypernym discovery, neglecting the broader range of tasks that taxonomies can support. For instance, research on taxonomy enrichment often uses LLMs only to extract representations that are then fed into Graph Neural Networks (GNNs) (Scarselli et al., 2008) or other simpler graph embedding models, such as node2vec (Grover & Leskovec, 2016), rather than directly employing LLMs for the full range of tasks (Jiang et al., 2022).

In this paper, we aim to fill the gap in existing research by exploring how modern foundation models can learn and apply taxonomy graph relations across multiple lexical semantic tasks. Specifically, we focus on using a single LLM to tackle four distinct tasks simultaneously: taxonomy construction, hypernym discovery, taxonomy enrichment, and lexical entailment. We hypothesize that contemporary LLMs, when pretrained exclusively on the Princeton WordNet, can effectively learn taxonomy relations by leveraging their inherent language knowledge and align it with the established human-labeled structure.

To sum up, the contribution of the paper is as follows:

1. We *investigate* the ability of LLMs to generate taxonomic structures and predict entities at within various levels of its tree structure.
2. We introduce a novel dataset creation *method* that encompasses a variety of taxonomy-related subtasks, including hypernym prediction, hyponym prediction, insertion between two existing nodes, and synset mixing, expanding beyond previous setups that focused solely on hypernym prediction.
3. Using the developed method, we create *datasets* based on WordNet for training a taxonomy-based LLM enriched with word definitions using Wikidata⁶ and ChatGPT⁷.
4. Using the dataset, we train TaxoLLaMA, a unified *model* tailored to handle a wide range of tasks requiring taxonomic knowledge. The model achieves state-of-the-art results in a battery of common lexical-semantic problems.

5. We conduct a comprehensive *error analysis* across all tasks using both manual and automated methods, including the evaluation of error patterns and model performance with the assistance of ChatGPT.
6. Finally, we *demonstrate* the capability of LLMs to refine existing taxonomies by incorporating multiple relationships they have learned.

We make produced data, code and models in this study publicly available.⁸

This work is an extended version of research originally presented in two conference papers (Moskvoretskii et al., 2024a, 2024b). The novelty of the present journal article compared to these prior publications lies in the following additional contributions:

1. We deliver *experiments using additional LLMs* in zero- and few-shot settings: Phi-3-mini⁹, Qwen2.5¹⁰, and LLaMA3.1-8b¹¹.
2. We *update the TaxoLLaMA model* by fine-tuning LLaMA-3.1¹² and update the results for all datasets (while originally it was trained on LLaMa-2¹³).
3. We perform an *ablation study* on the consistency and performance for the TaxoLLaMA by different numbers of generations.
4. We report results of taxonomy construction on an *additional evaluation metric* Fowlkes-Mallows (F&M) Index commonly used in other taxonomy induction studies.
5. Finally, we perform *additional experiments* which investigate:
 - how LLMs can resolve graph cycles using extracted relations to improve quality of taxonomy construction;
 - how LLMs can leverage multiple relations to refine an already constructed graph;
 - how bidirectional relations can be used to refine constructed taxonomies.

2 Related Work

In this section, we provide a brief overview of previous approaches to the lexical semantics tasks that are the focus of our experiments. We explore the development of graph and taxonomy construction methods and discuss the challenges where taxonomic knowledge has shown to be particularly advantageous.

2.1 Taxonomies and Large Language Models

To the best of our knowledge, most existing papers do not consider generative transformers for taxonomy learning, while research mostly had focused on encoder-based rather than GPT-style models for taxonomy learning. Notable examples include using pre-trained BERT encoder to estimate hypernymy (Chen et al., 2021; Davies et al., 2023; Hanna & Mareček, 2021). Most studies involving LLMs in taxonomy construction have explored the use of models like LM-Scorer (Jain & Espinosa Anke, 2022), which employs BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) among masked LMs, and GPT-2 (Radford et al., 2019) among causal LMs. These studies typically employ zero-shot sentence probing or experiment with prompts for taxonomy learning. However, their results have not surpassed the state-of-the-art GNN models for tasks like TexEval-2. Notably, there is a lack of research comparing these methods to more recent open-source models such as LLaMA-2 (Touvron et al., 2023a) and Mistral (Jiang et al., 2023) for taxonomy-related tasks, that is the part of the current paper.

2.2 Hypernym Discovery

The task of hypernymy discovery involves generating a list of hypernyms for a given hyponym, as illustrated in Figure 1(a). A recent contribution in this area is a taxonomy-adapted, fine-tuned T5 model introduced by Nikishina et al. (2023). Prior to this, several approaches have been explored. The 300-sparsans method (Berend et al., 2018) improves upon the traditional word2vec technique. The Hybrid model (Held & Habash, 2019) combines the k-Nearest Neighbor method with Hearst patterns. CRIM (Bernier-Colborne & Barrière, 2018), recognized as the best performer in the SemEval competition, uses a Multilayer Perceptron (MLP) structure with a contrastive loss function. Lastly, the Recurrent Mapping Model (RMM) (Bai et al., 2021) employs an MLP with residual connections and a contrastive-like loss function.

2.3 Taxonomy Enrichment

The task of taxonomy enrichment involves determining the most suitable position for a missing node within a taxonomy, addressed in SemEval-2016 Task 14 (Jurgens & Pilehvar, 2016). Over the past few years, various architectures have

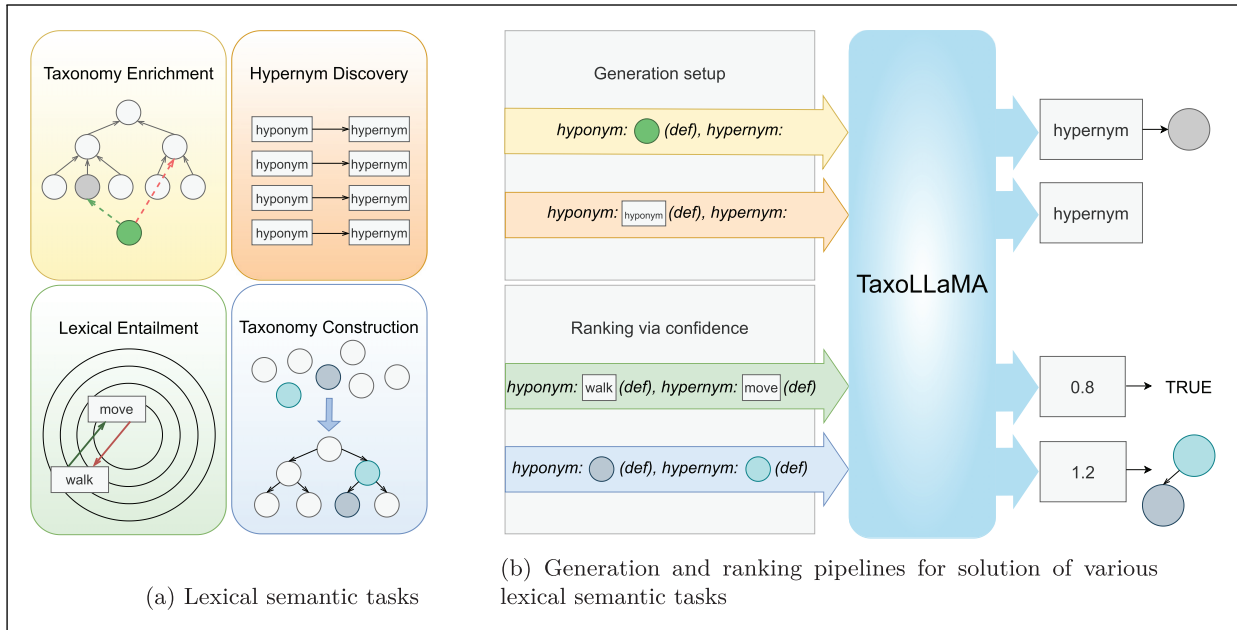


Figure 1. Examples with input and output for each task are highlighted by color. Rectangle “hypernym” denotes a word generated by the model; circle means a node from the graph. Confidence score determines the existence of a relationship between the two nodes provided in the input. (a) Lexical semantic tasks and (b) Generation and ranking pipelines for solution of various lexical semantic tasks.

been developed to tackle this task. TMN (Zhang et al., 2021) uses multiple scoring mechanisms to identify ⟨hyponym, hypernym⟩ pairs for a given query concept. TaxoEnrich (Jiang et al., 2022) utilizes two LSTM networks (Staudemeyer & Morris, 2019) to encode information about both ancestors and descendants. Additionally, TaxoExpan (Shen et al., 2020) employs a Graph Neural Network (GNN) (Scarselli et al., 2008) to predict whether the query concept is a child of an anchor concept.

2.4 Taxonomy Construction

The task of taxonomy construction is focused on building a domain taxonomy starting from a raw list of terms. Previously, this task was solved with the use of GNN, such as Graph2Taxo (Shang et al., 2020) or employing zero-shot language model for scoring pairs or mask token probability, such as LMScorer and RestrictMLM (Jain & Espinosa Anke, 2022). However, some approaches differ with focus on Hearst patterns boosted with Poincare embeddings for refinement.

2.5 Lexical Entailment

The task of lexical entailment involves classifying the semantic connections between word pairs. For instance, if we consider the term “tiger” (a hyponym), it inherently suggests the broader category “big cat” (a hypernym).

Recent research in lexical entailment includes various innovative models. SeVeN (Espinosa-Anke & Schockaert, 2018) encodes relationships between words, while Pair2Vec (Joshi et al., 2019) and a modified GloVe approach from Jameel et al. (2018) utilize word co-occurrence vectors along with Pointwise Mutual Information to understand semantic connections. The LEAR model (Vulić & Mrkšić, 2018), on the other hand, fine-tunes Euclidean space to better reflect hyponymy–hypernymy relationships. Graph-based approaches, the “Global” Entailment Graph (GBL) (Hosseini et al., 2018) employs a GNN focusing on local learning, while its evolution, the “Contextual” Entailment Graph (CTX) (Hosseini et al., 2021), enhances this by integrating contextual link prediction. The CTX model was later improved with an entailment smoothing technique proposed by McKenna et al. (2023), which currently holds SOTA for this task.

3 Methodology

In this section, we describe the process of building an instruction-tuning dataset specifically designed for taxonomy learning using LLMs and further fine-tuning.

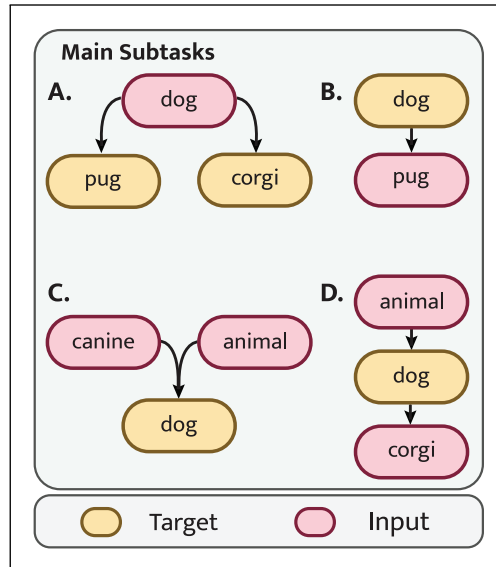


Figure 2. Examples of IS-A relation structures: (A) hyponym prediction, (B) hypernym prediction, (C) synset mixing, and (D) synset insertion.

3.1 Dataset Construction Method

The dataset creation process is largely based on the English Princeton WordNet 3.0, chosen for its structured and well-maintained organization. Our focus is mainly on the noun subgraph, not only because it represents the most frequent category in WordNet, but also because recent research (Lazaridou et al., 2021) has identified it as a challenging class for language models to master.

We begin dataset creation by utilizing a directed acyclic graph (DAG) derived from WordNet, which is structured around “IS-A” relationships. Next, we randomly select edges or subsets from this graph, dividing them into different subsets while taking into account all possible tree operations. A comprehensive explanation of the dataset construction algorithm can be found in Section 3.1.1.

We posit that a diverse dataset encompassing various taxonomy relations offers two key advantages:

- A diverse dataset enhances the model’s ability to generalize, enabling it to understand broader relationships between words across a wide range of subtasks.
- A diverse dataset also enables the model to develop and apply various strategies for effective taxonomy construction.

To account for the widest possible range of tree operations within the graph, we gather four distinct subsets, with a particular emphasis on hyponym and hypernym prediction. The tasks include the following scenarios (as illustrated in Figure 2):

- **Hyponym prediction** (task A): Predicting a list of hyponyms associated with a given synset from the taxonomy.
- **Hypernym prediction** (task B): Identifying the hypernym based on the provided input word.
- **Synset mixing** (task C): Predicting a single hyponym by combining information from two different synsets.
- **Insertion** (task D): Determining a word when given both its hypernym and hyponym.

We ensure that our test and training datasets are completely distinct, with no overlap between them. Specifically, none of the test nodes is included in any of the subtask scenarios. The statistics for each subset are detailed in Table 1.

Table 1. Statistics of Taxonomy Subtask Samples Used for Training and Evaluation.

Category	TaxoLLaMA _{multi}		TaxoLLaMA _{hyp}		TaxoLLaMA _{bench}	
	Train	Test	Train	Test	Train	Test
Hypernym prediction	1 338	364	44 772	0	36 775	0
Hyponym prediction	16 789	828	0	0	0	0
Synset mixing	1 461	47	0	0	0	0
Insertion	648	35	0	0	0	0
Total	20 236	1 274	44 772	0	36 775	0

Each column shows the number of samples per subtask type in different training configurations. **TaxoLLaMA_{multi}** includes all four relation types (hyponym, hypernym, insertion, synset mixing). **TaxoLLaMA_{hyp}** is trained only on hypernym prediction. **TaxoLLaMA_{bench}** shares the same training configuration as TaxoLLaMA but is constructed to avoid any overlap with test nodes in our application tasks, making it suitable for evaluating generalization to unseen taxonomic structures.

3.1.1 Formal Description. More formally we represent the method in Algorithm 1. The input of this algorithm are subsets A, B, C, D mentioned above corresponding to subtasks. These sets are derived from the graph, which are represented as a collection of the following mini-sets as following:

Algorithm 1. Method for dataset construction.

Input: Sets A, B, C, D sampled from a graph

Output: Sets for *Train* and *Test*

```

1: Train := Empty Array
2: Test := Empty Array
3: Collect sets  $A, B, C, D$ .
4: while  $(A \cup B \cup C \cup D) \neq \emptyset$  do
5:    $cur\_set \sim \mathbb{P}_{data}$ 
6:    $cur\_sample = cur\_set.pop()$ 
7:   if  $cur\_sample \overset{t}{\bar{\cap}} \overline{Train} = \emptyset$  then
8:      $to\_test \sim \mathbb{P}_{test}$ 
9:     if  $to\_test == 1$  then
10:      Test.append( $cur\_sample$ )
11:     else
12:      Train.append( $cur\_sample$ )
13:     end if
14:   else
15:     Train.append( $cur\_sample$ )
16:   end if
17: end while
18: return (Train, Test)

```

$$A_i = \{h, \{c_j\}_{j=1}^{deg^+h}\} \in A,$$

$$B_i = \{h, c\} \in B,$$

$$C_i = \{h_1, h_2, c\} \in C,$$

$$D_i = \{g, h, c\} \in D.$$

Here, c denotes hyponyms, h hypernyms, and g hyperhypernyms (hypernyms of hypernyms (two levels above in the taxonomy)). A, B, C , and D correspond to the subtasks in Figure 2.

To facilitate comprehensive set intersections, we introduce the concept of “deep intersection,” denoted as $\bar{\cap}$. This operation goes beyond the intersection of individual elements in two sets by considering the intersection between the

elements within the subsets of each set. It is mathematically expressed as:

$$S_1 \bar{\cap} S_2 = \bigcup_{ij} (S_{1i} \cup S_{2j}).$$

In the next phase, our goal is to generate random training and testing sets while aiming to balance the categories as much as possible, although some relations are less frequently represented. We ensure that the training set primarily consists of hyponym and hypernym predictions, while other types of samples are evenly distributed. This task is challenging due to the potential for significant overlap among different cases and the sequence in which samples are collected. To manage this complexity, we introduce a distribution over subtasks, denoted as \mathbb{P}_{data} . This allows us to manually adjust the probability of sampling each subtask, giving us greater control over the composition of the dataset.

To regulate the likelihood of samples being allocated to the test set, a Bernoulli distribution was considered, denoted as \mathbb{P}_{test} , with a parameter p , where p defines the probability of assigning a sample to the test set, while $q = 1 - p$ corresponds to assigning it to the training set. To ensure that each type of relation is effectively learned, we manually selected the optimal probability values to balance the category distribution in the dataset:

For \mathbb{P}_{data} : $P(A) = 0.51$, $P(B) = 0.39$, $P(C) = 0.05$, and $P(D) = 0.05$.

For \mathbb{P}_{test} : $p = 0.05$ and $q = 0.95$.

During data collection, we utilize the “pop()” operation, which removes and returns the last element from a set.

To manage the complexities associated with dominant word categories, we perform a topological sort on the graph. We then ensure that no vertex in our sets has a level lower than a specified parameter, referred to as “level.” This condition is expressed as: $\forall i, S \quad \forall v \in S_i : TopSort(v) \geq level$. For our collected data, we set $level = 3$.

We also designate a “target” vertex for each element within the subtasks. This enables us to monitor the inclusion of this specific target vertex in the test set, ensuring the integrity of our evaluation process. The definitions of these “target” vertices vary depending on the subtask and can be outlined as follows:

- $A'_i = \{c_j\}$: The focus is on tracking all hyponyms. If the hyponyms haven’t been encountered in the training set, the target cannot be determined. However, encountering the hypernym in the test set is acceptable since it is provided in the prompt.
- $B'_i = c$: If the hyponym hasn’t been seen, it indicates that this pair has not been encountered. Otherwise, the hyponym would have been added to the tracking. Therefore, if the hyponym is unseen, it implies the corresponding edge has not been observed.
- $C'_i = c$: Similarly, if the hyponym hasn’t been seen, it indicates that the target hasn’t been observed.
- $D'_i = h, c$: This scenario involves tracking two edges: $g - h$ and $h - c$, analogous to cases A and B . This restriction ensures that both edges are handled appropriately.

4 Training LLMs for Taxonomy Tasks

In this section, we investigate LLM’s ability to learn taxonomic relations.

4.1 Models

Using our collected dataset, we train a series of models under the TaxoLLaMA family:

- **TaxoLLaMA_{multi}**: LLaMA 2 7B (4-bit) model fine-tuned with LoRA on all types of taxonomy relations.
- **TaxoLLaMA_{hyp}**: LLaMA 2 7B (4-bit) model trained with LoRA exclusively on hypernymy relations.
- **TaxoLLaMA_{bench}**: A variant of TaxoLLaMA_{hyp} trained on a disjoint subset of the taxonomy, ensuring that none of the nodes in the training set appear in any test set, used for evaluating generalization.
- **TaxoLLaMA3.1_{multi}**: LLaMA 3.1 8B Instruct model fine-tuned on the full set of taxonomy relations.
- **TaxoLLaMA3.1_{hyp}**: LLaMA 3.1 8B Instruct model trained only on hypernymy relations.

In addition to our fine-tuned models, we evaluate several modern instruction-tuned LLMs in both zero-shot and few-shot settings: Qwen2.5-7B-Instruct¹⁴, Qwen2.5-1.5B-Instruct¹⁵, Phi-3-Mini-4K-Instruct¹⁶, GPT-2¹⁷, and Mistral-7B¹⁸. All ablation studies are conducted using TaxoLLaMA_{multi} as it offers a representative configuration while remaining computationally feasible.

4.2 Prompting

Our inputs include a system prompt that looks as follows:

1. [INST] <<SYS>> You are a helpful assistant. List all the possible words divided with a comma. Your answer should not include anything except the words divided by a comma <</SYS>>

Then we introduce a technical-style input prompt and the expected output format:

1. hypernym: dog.n.01 — hyponyms: [/INST]
2. pug, corgi,

In addition to the data collected through the described algorithm, it is crucial to disambiguate sense of the input node. We employ 3 ways of doing it: incorporating number ID from wordnet, lemmas and definitions in following prompt versions:

1. hypernym: dog.n.01 — hyponyms: [/INST]
2. hypernym: dog (dog, domestic dog, Canis familiaris) — hyponyms: [/INST]
3. hypernym: dog (a member of the genus Canis that has been domesticated by man since prehistoric times) — hyponyms: [/INST]

Since definitions might not be available for certain subtasks during inference—such as lexical entailment or taxonomy enrichment – we also generate definitions using ChatGPT for test sets that lack pre-existing explanations or source them from Wikidata.

For generating definitions, we used the web interface of ChatGPT 3.5 of February 2024 and the “gpt-3.5-turbo” model from the same period. The prompts used for these requests, along with the statistics of the generated definitions, are detailed in the Appendix A, specifically in Examples 9-10 and Table 16. This step is crucial, as experiments have shown that the absence of definitions can significantly reduce the model’s performance (Moskvoretskii et al., 2024b).

4.3 Evaluation

We evaluate the performance of our models using the Mean Reciprocal Rank (MRR), a metric that indicates the rank position of the first correct answer. We chose MRR over other ranking metrics, such as Precision@k or MAP because they might impose overly stringent criteria, which do not reflect abilities of understanding taxonomy. To assess the models, we create a list of potential candidates, each separated by a comma, and then match these candidates with the target words.

4.4 Fine-Tuning Details

To optimize several models, we applied a 4-bit quantization technique. Subsequently, we fine-tuned them using LoRA (Hu et al., 2022) for one training epoch with a batch size of 64. We used the AdamW optimizer with a learning rate of 3×10^{-4} , coupled with a cosine annealing scheduler. For any additional fine-tuning, the models were trained with a reduced batch size of 2.

4.5 Applications

We further hypothesize that such trained models will be effective in solving taxonomy-related out-of-domain tasks. We propose two ways of adapting to tasks:

Generative approach directly employs the generation abilities: starting with an input node, the model generates a list of potential completions. This approach is utilized for example in the hypernym discovery task and taxonomy enrichment task.

Ranking approach involves assessing the hypernymy relation through perplexity calculations, where a lower perplexity score indicates a stronger relationship. Specifically, for a candidate hypernym h and a given hyponym c , we compute the perplexity $\text{PPL}(h | c)$ based on the model’s generative likelihood. Perplexity is defined as:

$$\text{PPL}(x) = \exp \left(-\frac{1}{n} \sum_{i=1}^n \log P(x_i | x_{<i>i</i>}) \right),$$

Table 2. Mean Reciprocal Rank (MRR) scores across four taxonomy subtask types: Hyponym prediction, Hypernym prediction, Insertion, and Synset Mixing, along with the overall mean score.

Model	Hyponym	Hypernym	Insertion	Synset Mixing	Mean
zero-shot					
GPT-2	0.006	0.033	0.018	0.027	0.021
Phi-3-mini-4k-instruct	0.144	0.205	0.040	0.065	0.113
Llama3.1-8b-instruct	0.117	0.232	0.066	0.011	0.107
Qwen2.5-7B-Instruct	0.150	0.336	0.133	0.072	0.173
Qwen2.5-1.5B-Instruct	0.103	0.324	0.056	0.051	0.134
few-shot					
Phi-3-mini-4k-instruct	0.135	0.132	0.036	0.027	0.082
Llama3.1-8b-instruct	0.174	0.340	0.045	0.011	0.143
Qwen2.5-7B-Instruct	0.249	0.422	0.243	0.122	0.259
Qwen2.5-1.5B-Instruct	0.085	0.371	0.049	0.056	0.140
fine-tuning					
TaxoLLaMA _{multi} Numbers	0.099	0.267	0.262	0.239	0.162
TaxoLLaMA _{multi} Lemmas	0.127	0.293	0.329	0.218	0.188
TaxoLLaMA _{multi} Definitions	0.123	<u>0.494</u>	0.436	0.234	0.247
Mistral-7B Definitions	0.085	<u>0.498</u>	0.436	0.160	0.218
TaxoLLaMA3.1 _{hyp}	0.329	0.517	0.171	<u>0.250</u>	0.317
TaxoLLaMA3.1 _{multi}	<u>0.288</u>	0.269	0.221	0.280	<u>0.265</u>

The best-performing model on each task is shown in **bold**, and the second-best is underlined. Fine-tuned variants of our proposed models – TaxoLLaMA and TaxoLLaMA3.1 – demonstrate substantial improvements.

Table 3. Evaluation of Precision and Recall metrics for the Hyponym subtask using the TaxoLLaMA_{multi} Definitions model.

Subtask	MAP	P@1	P@5	P@15	R@1	R@5	R@10
Hyponym	0.092	0.110	0.098	0.087	0.076	0.103	0.103

where $x = h | c$ is the model input (e.g., a prompt including the hyponym followed by the candidate hypernym), and $P(x_i | x_{<i>$ is the predicted token probability. The lower the perplexity, the more confident the model is in the hypernym prediction. This ranking-based approach is applied, for example, in the taxonomy construction and lexical entailment tasks.

4.6 Results

The results of tested models are summarized in Table 2. In the zero-shot setting, all base models perform poorly, with MRRs generally below 0.2. Notably, Qwen2.5 models outperform GPT-2 and Phi-3-mini, especially on hypernym detection, which shows that it might be easier to predict hypernyms. However, scores remain low overall, highlighting the difficulty of taxonomic inference without supervision.

Table 3 presents precision and recall metrics for the hyponym prediction task using the TaxoLLaMA_{multi} Definitions model. This analysis aims to explain the low Hyponym scores observed in Table 2.

The results show that Precision is higher than Recall at the earliest ranks, indicating that the model’s top predictions are generally accurate. However, as the rank increases, recall improves and eventually surpasses precision, suggesting that the model retrieves more relevant hyponyms but with lower accuracy at deeper levels. This pattern reflects a trade-off between precision and recall, which contributes to the overall modest MAP scores for the hyponym subtask.

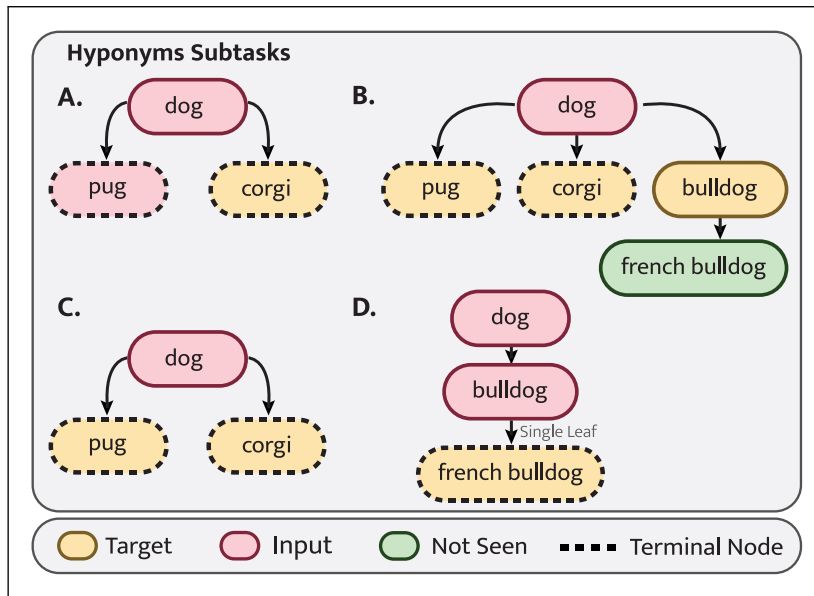
The few-shot setting leads to modest gains, particularly for Qwen2.5-7B, which achieves 0.259 average MRR and the highest insertion score. However, Phi-3-mini and LLaMA3.1-8B show only marginal improvements or even regressions, suggesting limited benefit from prompting for these tasks.

In contrast, the fine-tuned models significantly outperform all others. Our proposed latest models—TaxoLLaMA3.1 and TaxoLLaMA3.1-ALL achieve the top two average scores. TaxoLLaMA3.1 ranks first overall, with the highest scores on hyponym (0.329), hypernym (0.517), and strong performance on synset mixing. Interestingly, while TaxoLLaMA3.1-ALL trails slightly in average score, it leads on synset mixing (0.280), indicating its robustness across structural variations.

The evaluation reveals a clear contrast between TaxoLLaMA and TaxoLLaMA3.1. The former performs better on the Insertion subtask but struggles with Hyponym prediction, while the latter shows improved results on Hyponyms but a

Table 4. Evaluation of Precision and Recall metrics for the Hyponym and Insertion subtasks using the TaxoLLaMA3.1 model.

Subtask	MAP	P@1	P@5	P@15	R@1	R@5	R@10
Hyponym	0.243	0.247	0.251	0.255	0.076	0.244	0.256
Insertion	0.327	0.057	0.329	0.371	0.057	0.329	0.371

**Figure 3.** Examples of hyponym subtasks: Leaves Divided (A), Internal Nodes (B), Only Leaves (C), Single Leaves (D).

decline in Insertion performance. As shown in Table 4, TaxoLLaMA3.1 achieves higher Precision and Recall on Insertion, but a lower MRR. This suggests that although it retrieves more correct answers overall, it identifies correct Hyponyms earlier in the ranking, indicating stronger early precision on that subtask.

Among the disambiguation techniques, definitions performs best, particularly on hypernym and insertion tasks. The Mistral-7B backbone achieves similar high performance on hypernym and insertion but underperforms on other tasks.

Interestingly, the top-performing model, TaxoLLaMA3.1, underperforms on the insertion task compared to earlier TaxoLLaMA variants. This may suggest that modern models tend to prioritize easier-to-learn taxonomic patterns during optimization, potentially overlooking more structurally complex relations like insertion.

We believe that the size of the model is the main contributing factor, rather than pre-training data amount. Despite using lemmas or definitions for disambiguation, the score does not change drastically for worst cases, showing that disambiguation is not the key problem. Moreover, the underperformance may be linked to the sequential nature of LM loss in instruction tuning. With multiple correct answers, it poses a problem to properly apply loss, as different orders of correct nodes would imply completely different loss values. The problem usually arises with hyponym prediction.

4.7 Subtypes of Hyponyms

To provide better understanding of underlying processes, we splitted the hyponymy cases into more detailed. The narrower case, reflected at Figure 3 are as follows:

- **Leaves Divided** (3A): all hyponyms required to be terminal nodes, with half of them passed as input, other half is considered a target.
- **Internal Nodes** (3B): Hyponyms are required to have at least one internal node.
- **Only Leaves** (3C): all target hyponyms are terminal nodes.
- **Single Leaves** (3D): hyponyms are required to be terminal nodes, and they are the only hyponyms for the node.

The results in Table 5 show that terminal nodes are predicted better as internal. We believe that this results stems from ambiguity of internal nodes, as we noted through manual examination of them. The main issue with predicting internal

Table 5. MRR Scores for the LLaMA-2 model with Different Hyponyms Prediction Subtasks; Column Names Correspond to Figure 3.

Model	3A	3B	3C	3D
#Samples	117	115	110	486
Numbers	0.152	0.113	0.220	0.068
Lemmas	0.179	0.154	0.220	0.100
Definition	0.175	0.163	0.268	0.081

Table 6. Difference in MRR Scores Between the *easy* and *hard* Subsets for Each Taxonomy Learning Subtask, Computed as: $\Delta\text{MRR} = \text{MRR}_{\text{easy}} - \text{MRR}_{\text{hard}}$. Positive Values (Highlighted in Green) Indicate that the Model Performed Better on the *easy* Subset; Negative Values (Highlighted in Red) Indicate Better Performance on the *hard* Subset. Results are shown for Three Input Variants: WordNet Numbers, Lemmas, and Definitions.

	Hyponym	Internal Nodes	Leaves Divided	Only Leaves	Single Leaves	Insertion	Hypernym	Synset Mixing	Mean
Numbers	-0.036	0.001	-0.046	-0.240	0.006	0.079	-0.055	0.089	-0.005
Lemmas	-0.036	-0.019	-0.004	-0.124	-0.016	-0.028	-0.053	0.120	-0.001
Definition	-0.042	-0.044	0.029	-0.194	0.000	0.025	0.006	0.097	0.054

nodes (3B), is prediction of more distant nodes (with $\text{hop} \geq 2$) instead of the direct hyponyms. Additionally, the scenario in (3A) shows lower performance compared to predicting all possible hyponyms (3C). This suggests that the key issue is not simply ambiguity, as it would be resolved with cohyponyms, but rather the model’s difficulty in generating the appropriate hyponyms. The model’s predictive scope seems constrained by the candidates provided in the input. The scenario involving a single leaf hyponym (3D) proves to be particularly challenging to predict, even when hypernyms are provided as input. This difficulty might be due to the complexity and relative rarity of such instances in natural language, making them harder for the model to learn and generate accurately.

4.8 Common Words Versus Terminology

To better understand the consistently low average results, we closely examined the model outputs and found that the complexity of the dataset could be a significant factor. Some synsets within the WordNet taxonomy may be overly specialized, which poses a challenge for the model when predicting hyponyms or hypernyms. To investigate this possibility, we categorized our dataset into two distinct groups: commonly known words (classified as the “*easy*” category) and more specialized terms, jargon, or rare words (classified as the “*hard*” category). This categorization was carried out with the help of three computational linguistics experts, who annotated the test set. They were asked to classify a sample as “*hard*” if it contained at least one word that could be considered a term, jargon, or rare, and as “*easy*” if it did not. The level of agreement among the annotators, as measured by Krippendorff’s alpha, was 0.67, which is high enough to consider the annotations valid and reliable.

We revisit the performance metrics for both the “*easy*” and “*hard*” subsets and summarized the results in Table 6. Interestingly, models generally performed better on the “*hard*” nodes, especially when it came to predicting hyponyms. However, when using our best model that incorporates word definitions, the “*easy*” instances yielded higher scores, particularly in cases that did not involve hyponym predictions. This trend, though, is not consistent across all prompt types; in some cases, “*hard*” instances were more accurately predicted, even when dealing with hypernyms or internal nodes.

We believe the results of the ablation study suggest that the model tends to predict less common words more accurately. This could be because the candidate pool for these terms is smaller, allowing the model to focus more directly on the correct answers. Additionally, the model likely encounters these rare words less frequently and typically within consistent, specific contexts, which might enhance its predictive accuracy for such terms.

5 Taxonomy Construction

In this section we describe the application of TaxoLLaMA to taxonomy construction task.

We test the TaxoLLaMA versions on the downstream task: SemEval 2016 Task 13. We use the Eurovoc taxonomies (“*Science*”, “*Environment*”) and Wordnet “*Food*” from SemEval-2016 (Bordea et al., 2016). These datasets are commonly used as a benchmark for testing models’ abilities of taxonomy construction. As mentioned in Section 3.1.1, the test set was deliberately excluded during the TaxoLLaMA training.

To create the taxonomy, we use an uncertainty-based ranking approach. This technique involves assessing the hypernymy relation through perplexity calculations, where a lower perplexity score indicates a stronger relationship. We

Table 7. F1 Scores for the Taxonomy Construction Task on Three SemEval-2016 Domain-Specific Datasets: Science, Environment, and Food.

	TexEval-2 best	TAXI+	Graph2Taxo pure	Graph2Taxo best	LMScorer	RestrictMLM	TaxoLLaMA _{hyp}	TaxoLLaMA _{bench}	TaxoLLaMA _{multi}
Science	31.3	41.4	39.0	47.0	31.8	37.9	<u>44.55</u>	<u>42.36</u>	<u>44.12</u>
Environment	30.0	30.9	37.0	<u>40.0</u>	26.4	23.0	45.13	44.82	42.03
Food	<u>36.01</u>	34.1	–	–	24.9	24.9	51.71	51.18	42.35

Our models’ TaxoLLaMA_{hyp}, TaxoLLaMA_{bench}, and TaxoLLaMA_{multi} outperform previous approaches on the Environment and Food domains and achieve competitive results on the Science domain. **Bold** indicates the best result, and underlined values mark the second-best. Notably, our models are trained solely on WordNet and do not rely on domain-specific taxonomies, yet still generalize well across unseen taxonomic structures.

Table 8. F1 Scores in Comparison to Fowlkes & Mallows Index for Taxonomy Construction Task.

		TaxoLLaMA _{hyp}	TaxoLLaMA _{bench}	TaxoLLaMA _{all}
Science	F1 score	44.55	42.36	<u>44.18</u>
	F & M	<u>46.07</u>	44.47	48.73
Environment	F1 score	45.13	<u>44.82</u>	42.03
	F & M	46.65	<u>45.59</u>	42.43
Food	F1 score	51.71	<u>51.18</u>	42.35
	F & M	51.5	51.01	–

calculate the perplexity for every possible edge and retain only those below an optimal threshold, determined via brute-force search over a predefined grid. We have not used definitions, as it is infeasible to generate them in this setting. We also apply self-refinement based on hypernymy perplexity to resolve self-loops and delete multiple parental edges, which is further discussed in Section 5.2 and Section 5.4.

5.1 Results

In Table 7, we showcase the F1-scores for the Science, Environment and Food datasets. We evaluate our three models version against earlier methods.

Table 8 presents F1 scores alongside the Fowlkes & Mallows (F&M) index for evaluating the performance of our taxonomy construction approach across the three domains described above. While the F1 score captures the harmonic mean of precision and recall, it does not fully reflect structural alignment in hierarchical tasks. The F&M index, which measures pairwise clustering similarity, is included to provide a complementary perspective on how well the predicted taxonomies preserve the underlying hierarchical structure.

Our results indicate that our method outperforms all existing models on the Environment and Food domains and ranks second on the Science domain. The top-performing approach for the “Science” dataset, Graph2Taxo (Shang et al., 2020), achieves its best score through a GNN-based cross-domain transfer framework, specifically during their ablation study. Interestingly, the framework’s default setup does not produce the highest scores (refer to Shang et al. (2020) (pure) in Table 7). It is also clear that zero-shot LM performed the worst on average, underscoring the need for specific fine-tuning and stronger models (Jain & Espinosa Anke, 2022).

5.2 Multiple Parental Nodes

Typically, having multiple parent nodes in taxonomies and ontologies is rare, usually with no more than three parents. We analyzed how our LLM constructs the graph across various thresholds, with the results presented in Table 9. The findings show that assigning multiple parents is common when using non-optimal thresholds, and while less frequent, it still occurs with optimal thresholds.

We addressed the issue of multiple parent nodes using several techniques:

- **Delete All:** Remove all parental edges from nodes that have multiple parents. This approach operates under the assumption that if the LLM is uncertain about a parental relationship, it’s better to omit such connections altogether.
- **Perplexity:** Retain only the edge with the lowest perplexity value. This method ensures that only the most probable parental connection, as determined by the LLM, is maintained.

Table 9. Distribution of Parent Counts Across Graph Types and Subsets.

Subset	Graph Type	2 Parents	3 Parents	4 Parents	5 Parents	6 Parents	7 Parents	8 Parents	9 Parents	10+ Parents
Environment	Noisy Graph	6	10	13	11	8	13	12	17	141
	Optimal Graph	35	18	2	1	–	–	–	–	–
Science	Noisy Graph	6	5	6	6	13	11	10	10	38
	Optimal Graph	5	–	–	–	–	–	–	–	–

Table 10. Differences in F1 Scores for Different Methods for Self-Refinement of LLM in Comparison with the Perplexity Method (best). Baseline Refers to the Taxonomy Construction Without any Refinement Strategy.

Task	No refinement (baseline)	Delete All	Perplexity (best)	Synset Mixing Two	Synset Mixing One	Compose	Perplexity (Cycles)	Hyponymy
Environment	−0.020	−0.016	0.000	−0.087	−0.009	0.000	−0.010	−0.017
Science	−0.020	−0.007	0.000	−0.022	−0.013	−0.007	−0.016	−0.018

- **Synset Mixing:** Keep two parental nodes if the LLM’s synset mixing perplexity falls below a specified threshold. This technique leverages the LLM’s ability to combine synsets and assesses the likelihood of such combinations resulting in the target node. We explore two variations:
 - * **Synset Mixing One:** Applies the same threshold used during edge construction.
 - * **Synset Mixing Two:** Uses a different, specifically chosen threshold for evaluation.
- **Compose:** Combines the Perplexity and Synset Mixing methods. If two parental nodes do not meet the Synset Mixing threshold criteria, the Perplexity method is applied as a fallback.

The results in Table 10 indicate that most of the self-refinement methods for handling multiple parents improve the quality of the graph in comparison to the baseline, but are still worse than the best (Perplexity) method. The simple perplexity rule proves to be the most effective. We believe this is due to the LLM’s stronger ability to encode hypernym relations, while its synset mixing capability is less developed, likely due to limited data during pretraining.

5.3 Hypernym–Hyponym Validation

In this section, we explore how an LLM can validate edges for both hypernymy and hyponymy relations. After constructing the graph using hypernymy, we investigate the impact of removing edges that fall above the hyponymy threshold on the overall quality. The results presented in Figure 4 demonstrate that using hyponymy for graph refinement can be beneficial, though it requires careful calibration. However, it is not as effective as the refinement techniques used for resolving multiple parental nodes.

5.4 Cycles Resolution

Cycles are typically rare in taxonomies, and self-loops should not exist at all, as they contradict the fundamental structure of taxonomies. To address self-loops and larger cycles, we primarily use the perplexity rule, similar to the approach described in Section 5.2, by removing the edge with the highest perplexity.

We also considered eliminating cycles involving three or more nodes by leveraging the LLM’s ability to evaluate the insertion of a node followed by the deletion of the least probable connection. However, cycles with three or more nodes are rare when using optimal thresholds and are not included in our analysis, as they consistently result in lower scores compared to the optimal threshold.

The results in Table 10 show an overall improvement with this procedure, particularly in the scientific domain, where closely related concepts are more likely to form loops.

5.5 Taxonomy Construction Strategies

5.5.1 Hypernymy Vs Hyponymy. This experiment presented in Table 11 show that predicting hypernyms performs significantly better than predicting hyponyms, which is coherent with the scores for the respective subtasks during the fine-tuning step.

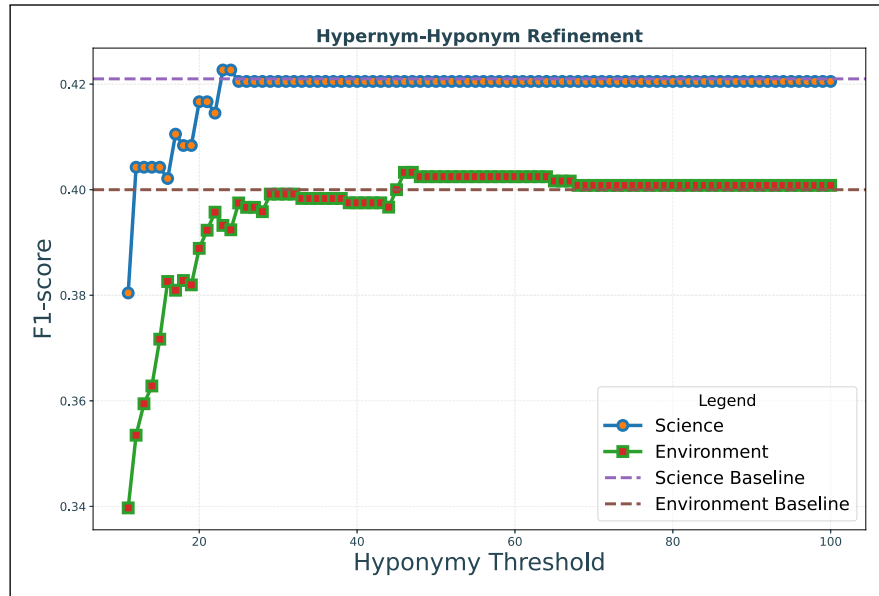


Figure 4. The graph for hypernym–hyponym validation for Science and Environment. X axis shows the threshold for hyponymy, Y axis shows the resulting score. Dashed lines indicate initial score without hyponymy validation.

Table 11. Results for the Downstream TexEval-2 Task Comparing Different Fine-Tuned Models, Methods for Graph Construction, and Templates for Model Inputs. Hyper Approach Stands for Hypernym Prediction and Hypo for Hyponym Prediction.

Approach	Method	Template	Sci	Env
TaxoLLaMA _{multi}	brute-force	hyper	0.419	0.409
		hypo	0.192	0.115
with lemma	dfs	hyper	0.340	0.213
		hypo	0.137	0.142
TaxoLLaMA _{multi}	brute-force	hyper	0.426	0.380
		hypo	0.188	0.116
with empty lemma	dfs	hyper	0.338	0.213
		hypo	0.127	0.129
TaxoLLaMA _{multi}	brute-force	hyper	0.416	0.411
		hypo	0.185	0.116
with numbers	dfs	hyper	0.186	0.186
		hypo	0.125	0.138

5.5.2 Construction Methods. We explored two techniques for building a taxonomic graph. For both of them, we traverse a predefined grid and select the best threshold based on evaluation metrics. However, the search spaces differ: the *brute-force* method evaluates all possible edges globally at each threshold level, while the *DFS-style* approach builds the graph incrementally. It begins at the root and recursively explores child nodes, appending new edges only if their confidence exceeds the current threshold. This mimics a depth-first traversal and aims to enforce local coherence during construction.

Results in Table 11 show that brute-force outperformed the DFS-style approach. That could happen due to error accumulation during graph traversal. Incorrect decision on the first couple levels significantly limits our possible edge space.

5.5.3 Prompt. Prompt was ablated with adding lemmas, empty lemma or specific WordNet number with corresponding models. For prompting with lemmas (as we have no additional lemmas unlike in WordNet), we tried two approaches (duplicate lemma in listing; provide no lemma at all):

1. “hypernym: cat (cat) — hyponyms:”
2. “hypernym: cat () — hyponyms:”

Table 12. MRR Performance on Hypernym Discovery. * Refers to the Systems that Rely on the Provided Dataset only, without LLM Pretraining or Additional Data being Used. Zero-shot is Trained on the WordNet Data only, Without Fine-Tuning on the Target Dataset.

	1A: English	2A: Medical	2B: Music	1B: Italian	1C: Spanish
CRIM* (Bernier-Colborne & Barrière, 2018)	36.10	54.64	60.93	–	–
Hybrid* (Held & Habash, 2019)	34.07	64.47	77.24	–	–
RMM* (Bai et al., 2021)	39.07	54.89	74.75	–	–
T5 Nikishina et al. (2023)	45.22	44.73	53.35	24.04	27.50
300-sparsans* (Berend et al., 2018)	–	–	–	25.14	37.56
TaxoLLaMA _{hyp} zero-shot	38.05	43.09	42.7	1.95	2.21
TaxoLLaMA _{bench} zero-shot	37.66	42.2	44.36	1.47	2.08
TaxoLLaMA _{hyp} fine-tuned	54.39	77.32	80.6	51.58	57.44
TaxoLLaMA _{bench} fine-tuned	51.59	73.82	78.63	50.95	58.61

Results in Table 11 show that the best result is obtained with either empty lemma or technical numbers. We believe that model could be distracted when the lemma is repeated, therefore scores are lower. It is unexpected that model with WordNet number has shown outperformance for Environment and Strong result for Science, possibly due to more straightforward task.

6 Hypernym Discovery

We evaluate TaxoLLaMA on the hypernym discovery task from SemEval-2018 (Camacho-Collados et al., 2018) using a generative approach. This task includes an English test set for general hypernyms, as well as two domain-specific sets for “Music” and “Medical.” Additionally, there are general test sets available for Italian and Spanish. The performance is assessed using the Mean Reciprocal Rank (MRR) metric. We employ a zero-shot approach, where the model is tested without fine-tuning on the training datasets. Notably, the test set is distinct from WordNet and may require multiple hops to reach hypernyms, making it suitable for both general and narrow domains.

6.1 Results

The results for the English language, presented in Table 12, show that both the fine-tuned TaxoLLaMA and TaxoLLaMA_{bench} models significantly surpass previous SOTA results. Although the zero-shot performance of our models is somewhat lower than their fine-tuned counterparts, they still achieve outcomes comparable to earlier results in general English tasks and remain competitive in domain-specific tasks, despite the fact that previous methods were all fine-tuned.

6.1.1 Multilingual Performance. In the case of Italian and Spanish, the fine-tuned model exceeds previous SOTA results. This success might be attributed to the model’s inherent multilingual capabilities, given that LLaMA-2 was initially designed to be multilingual, even though fine-tuning was conducted solely on English pairs. However, the zero-shot performance reveals challenges in generating accurate hypernyms for languages other than English. It is important to note that Italian and Spanish data were not part of the instruction tuning dataset.

6.1.2 Zero-shot Performance. To better understand the underperformance in zero-shot scenarios, we analyzed the impact of fine-tuning across different domains and languages, as depicted in Figure 5(a). The analysis shows that, apart from task 2B, the model surpasses previous SOTA results with as few as 50 samples for fine-tuning. Furthermore, the varying scores emphasize the model’s sensitivity to the quality and characteristics of the training data.

6.1.3 Few-shot Performance. We further investigated the few-shot learning approach for Italian and Spanish to evaluate the model’s adaptability in an in-context learning setting, as depicted in Figure 5(b). The model surpassed previous SOTA benchmarks for Italian, showing a near-logarithmic improvement with 30 and 50 shots, but did not perform as well for Spanish. We attribute this suboptimal few-shot performance to the 4-bit quantization and the relatively small model size. Smaller models generally underperform on various benchmarks compared to their larger counterparts, as demonstrated by the example of LLaMA-2 (Touvron et al., 2023b). Moreover, smaller or quantized models have limited capacity compared to larger models, a finding supported by earlier research (Egiazarian et al., 2024; Frantar et al., 2022; Lin et al., 2024; Wang et al., 2022). As it has been already seen (Lin et al., 2024), the benefits of few-shot learning are less pronounced in quantized models compared to full-precision models.

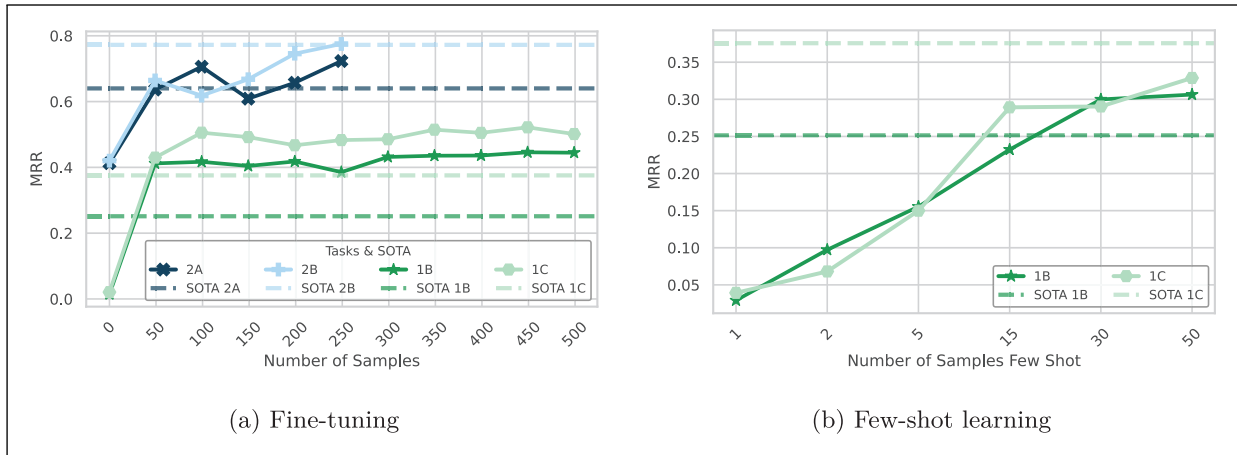


Figure 5. Experiments for domain and language adaptation on the hypernym discovery datasets. (a) Fine-tuning and (b) Few-shot learning

7 Taxonomy Enrichment

In this section we evaluate TaxoLLaMA on the taxonomy enrichment task. Following the methodology of previous studies (Jiang et al., 2022; Zhang et al., 2021), the task is considered as ranking graph nodes based on their probability of being the correct hypernym. The aim is to position the correct hypernyms at the top of the ranking, ensuring the node is accurately placed within the taxonomy. In our approach, we utilize the generative method, as shown in Figure 1(b).

The taxonomy enrichment benchmark includes datasets such as WordNet Noun, WordNet Verb, MAG-PSY, and MAG-CS (Jiang et al., 2022; Shen et al., 2020). To maintain consistency with the TaxoExpan test set (Shen et al., 2020), we selected 1,000 nodes from each dataset. In line with (Jiang et al., 2022), we utilize scaled MRR (Ying et al., 2018) as the key evaluation metric. This metric is derived by multiplying MRR by 10 and then averaging it across all correct hypernyms associated with each node.

To improve disambiguation, we created definitions for MAG datasets that lacked predefined explanations, either by generating them with ChatGPT or retrieving them from Wikidata. We utilized the ChatGPT 3.5 web interface and the “gpt-3.5-turbo” model, both from February 2024, for generating these definitions. The prompts used and the statistics related to the generated definitions are provided in Appendix A, specifically in Examples 9-10 and Table 16. This step is essential, as missing definitions can lead to a decrease in model performance, as highlighted in Moskvoretskii et al. (2024b).

7.1 Results

The results in Table 13 indicate that our model outperforms all previous approaches on the WordNet Noun and WordNet Verb datasets. However, it falls short of the current SOTA method on the more specialized MAG-CS and MAG-PSY taxonomies, even with fine-tuning. Interestingly, TaxoLLaMA_{bench}, despite having access to less data, unexpectedly delivered better performance on the MAG datasets. To gain further insight into the reasons for the overall underperformance, we conducted an in-depth error analysis, which is discussed in Section 9.1.

8 Lexical Entailment

In this section, we show the application of TaxoLLaMA to the lexical entailment task. For our evaluation, we rely on the Hyperlex benchmark (Vulić et al., 2017) alongside the ANT entailment subset (Guillou & de Vroe, 2023), which serves as a detailed refinement of the Levy/Holt dataset (Holt, 2019).

8.1 Ant

This dataset features sentence pairs that differ by a single argument within their syntactic structure (e.g., “The audience *applauded* the comedian” vs. “The audience *observed* the comedian,” as shown in Table 2 of Guillou and de Vroe (2023)). Each pair is classified into one of several relationships: antonymy, synonymy, directional

Table 13. Scaled MRR Across Tasks for Taxonomy Enrichment. Here, “n/a” Stands for “not Applicable”, as TaxoLLaMA has Already Seen WordNet Data and its Performance Cannot be Considered as Zero-Shot. Zero-shot is Trained on the WordNet Data only, Without Fine-Tuning on the Target Dataset.

	MAG-CS	MAG-PSY	Noun	Verb
TaxoExpan (Shen et al., 2020)	19.3	44.1	39.0	32.5
GenTaxo (Zeng et al., 2021)	23.9	46.4	28.6	42.8
TMN (Zhang et al., 2021)	24.3	53.1	36.7	35.4
TaxoEnrich (Jiang et al., 2022)	57.8	58.3	44.2	45.2
TaxoLLaMA _{hyp} zero-shot	7.4	7.3	n/a	n/a
TaxoLLaMA _{bench} zero-shot	8.5	6.6	n/a	n/a
TaxoLLaMA _{hyp} fine-tuned	24.9	29.8	48.0	52.4
TaxoLLaMA _{bench} fine-tuned	<u>30.2</u>	31.4	45.9	51.9

entailment, or non-directional entailment (the reverse of directional entailment). For sentences that exhibit an entailment relationship, we treat the differing elements as hypernym–hyponym pairs.

The ranking method here is enriched with confidence scores. The confidence score is the ratio between forward and reversed perplexity. The forward perplexity is the regular one, and the reversed is obtained by first reversing hypernym and hyponym roles.

Based on these confidence scores entailment relations are assessed as the ratio of the hypernym to hyponym ranking scores, with normalization by the L2 norm to estimate the probability of entailment. For example, we compute the perplexity score of “move” as a hypernym of “walk” ($PPL_{m \rightarrow w}$) and the reverse ($PPL_{w \rightarrow m}$). The ratio $\frac{PPL_{m \rightarrow w}}{PPL_{w \rightarrow m}}$ between these scores then reflects the model’s confidence in the entailment relationship.

Additionally, we developed TaxoLLaMA_{verb} specifically for this subtask. This model was pre-trained exclusively on verbs from WordNet, with the aim of better capturing the taxonomy structure of verbs.

8.2 HyperLex

This dataset is designed to assess entailment for both verbs and nouns, using a scale from 0 to 10. A score of 0 signifies no entailment, whereas a score of 10 represents strong entailment. The objective is to maximize correlation with the gold-standard scores. For this dataset, we apply the ranking approach directly, without any additional processing and usage of confidence scores.

Earlier approaches typically generate embeddings and then train a basic SVM on the Hyperlex training set. Fine-tuned models, such as RoBERTa, require significant computational resources and are specifically adapted to the Hyperlex dataset. In contrast, our zero-shot model utilizes perplexities directly as predictions, eliminating the need for any additional training. As a result, direct comparisons may not fully account for the distinct methodologies and resource demands, highlighting the importance of evaluating each method within its own specific context.

8.3 Results

8.3.0.1 Results on the ANT Dataset. The results presented in Table 14(a) compare our models with previous SOTA performances on the ANT dataset. A significant observation is the clear disparity in performance between TaxoLLaMA, trained on both nouns and verbs, and TaxoLLaMA_{verb}, specialized exclusively in verbs.

TaxoLLaMA_{verb} outperforms TaxoLLaMA in the lexical entailment task, indicating potential challenges in processing nouns and verbs together, which may hinder effective verb learning. This could be related to the constraints of quantization and LORA adapter tuning. Interestingly, this issue appears to be specific to the entailment task, as it does not arise in other tasks like taxonomy enrichment, which also involves a verb dataset. The discrepancy might be due to the metrics used, which require precise normalized perplexity rankings.

Table 14(a) reveals that TaxoLLaMA_{verb} attains SOTA performance in Average Precision and ranks second in normalized AUC. However, it is important to note that the comparison with previous SOTA results is somewhat imbalanced, as the top-performing models leveraged additional Entailment Smoothing (McKenna et al., 2023) to enhance their performance. This technique has not yet been applied to our models, suggesting a potential avenue for future improvements.

8.3.0.2 Results on the HyperLex Dataset. Table 14(b) highlights the effectiveness of our model, outperforming the previous SOTA in a zero-shot scenario for the “Lexical” subset and securing second place for the “Random” subset. Interestingly, while most models tend to perform better on the random subset, our approach deviates from this trend,

Table 14. Results of Experiment for the Lexical Entailment Tasks on the ANT (a) and HyperLex (b) Datasets.

(a) Performance on the lexical entailment ANT dataset. *Zero-shot* is trained on the WordNet data without fine-tuning on the target dataset.

	AUC _N	AP
GBL (Hosseini et al., 2018)	3.79	58.36
CTX (Hosseini et al., 2021)	15.44	65.66
GBL-P _{K=4} (McKenna et al., 2023)	13.91	64.71
CTX-P _{K=4} (McKenna et al., 2023)	25.86	67.47
TaxoLLaMA zero-shot	0.89	51.61
TaxoLLaMA _{bench} zero-shot	2.82	54.24
TaxoLLaMA _{verb} zero-shot	<u>19.28</u>	69.51

(b) Spearman correlation for lexical and random test subsets of Hyperlex dataset. *Zero-shot* is trained on the WordNet data without fine-tuning on the target dataset.

Setting	Model	Lexical	Random
fine-tuned	RoBERTa best (Pitarch et al., 2023)	79.4	82.8
	RoBERTa mean (Pitarch et al., 2023)	65.8	63.8
	LEAR (Vulić & Mrkšić, 2018)	54.4	69.2
zero-shot	Relative (Camacho-Collados et al., 2019)	54.3	58.4
	Pair2Vec (Joshi et al., 2019)	33.4	54.3
	GRV SI (Jameel et al., 2018)	48.3	55.4
	SeVeN (Espinosa-Anke & Schockaert, 2018)	46.9	62.7
	FastText	43.9	54.3
	TaxoLLaMA	70.2	<u>59.3</u>

indicating that the larger training size of the random subset may provide greater advantages to other methods. Despite the simplicity of our zero-shot method, it still delivers impressive results. Future research could investigate incorporating this score as a meta-feature in task-specific models, or refining our entire model for better alignment.

9 Error Analysis

In this section, we examine the errors produced by the TaxoLLaMA model, delve into the underlying causes of these inaccuracies, and propose strategies for improving the performance of LLMs when applied to taxonomies.

9.1 Hypernym Discovery and Taxonomy Enrichment

Since we use the same generative approach for both hypernym discovery and taxonomy enrichment, we conduct a combined error analysis. This process is divided into four steps: (i) conducting a manual review to pinpoint the most frequent errors; (ii) performing an automatic error analysis using ChatGPT; (iii) comparing and consolidating the common errors identified; and (iv) classifying these errors with the help of ChatGPT.

We begin by selecting approximately 200 random samples from both the hypernym discovery and taxonomy enrichment datasets and provide explanations for the model’s failure to generate the correct hypernym. Through this process, we identify four categories of errors: (i) predicted hypernyms are excessively broad; (ii) Incorrect or irrelevant definition; (iii) the model fails to produce relevant candidates within the same semantic domain; (iv) miscellaneous cases that do not fit into the other categories.

We further provide in Appendix the prompt to request that ChatGPT generate potential error types 11, the resulting output 12, and the Table 17 summarizing the error types identified across multiple runs. Afterward, we combine the error types identified both manually and automatically into the following categories:

- **Overly Broad Predictions:** The model frequently generates predictions that are broader than the intended hypernym.
- **Overly Narrow Predictions:** Some predictions are too specific and do not capture the generality of the true hypernym.
- **Inaccurate Predictions:** The model sometimes predicts terms that are semantically similar to the correct hypernym but fails to match the exact wording required.
- **Conceptual Ambiguity:** The model struggles with input words or concepts that have ambiguous meanings, resulting in incorrect predictions.

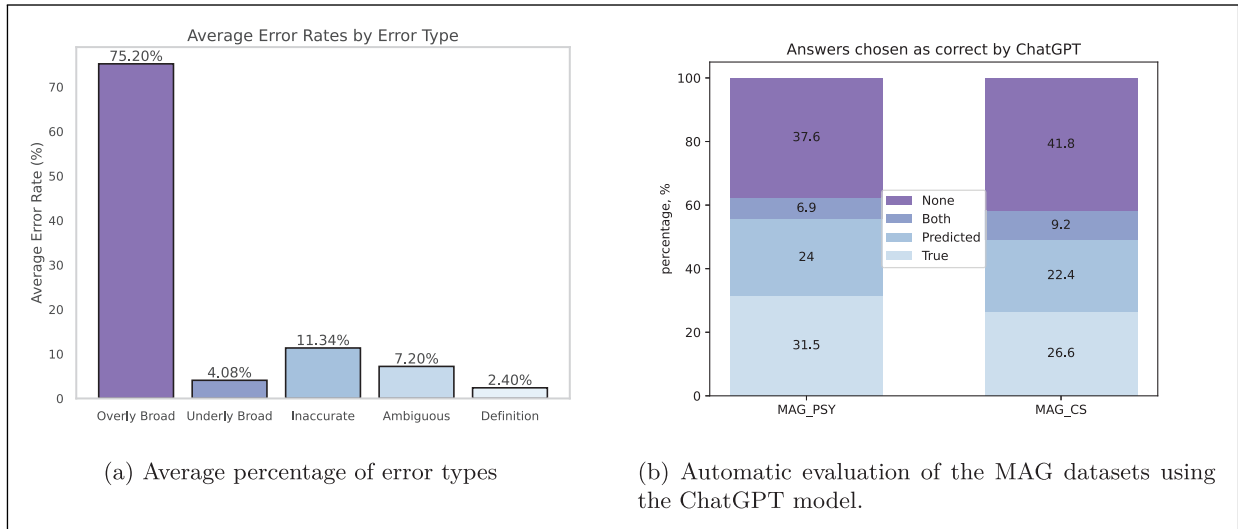


Figure 6. (a) Average percentage of error types across hypernym discovery and taxonomy enrichment datasets. (b) Automatic evaluation of the MAG datasets using the ChatGPT model. The label “True” represents the number of instances where ChatGPT favored the gold-standard answers over those generated by TaxoLLaMA; “Predicted” indicates cases where ChatGPT preferred the output from TaxoLLaMA. Additionally, ChatGPT could select “Both” if it found both answers equally acceptable or “None” if neither answer was preferred. (a) Average percentage of error types and (b) Automatic evaluation of the MAG datasets using the ChatGPT model.

- **Incorrect Definitions:** Errors occur when the model is misled by inaccurate or incorrect definitions retrieved from external sources.

To classify incorrectly predicted instances, we used the prompt provided in Appendix A, as shown in Example 13. The outcomes for each task and dataset are detailed in Table 18 and Figure 6(a) in Appendix B, which illustrate the average error distribution. Additionally, Table 19 includes an example corresponding to each type of error. The most prevalent problem, affecting 75% of the cases, is the prediction of overly broad concepts. This issue is likely due to the model’s adaptation to domain-specific datasets that are more expansive than WordNet, such as those in the “Music” and “Medical” domains.

In the case of Italian and Spanish, substantial inaccuracies were primarily due to the grammatical complexities inherent in these languages, compounded by dataset limitations, linguistic nuances, and insufficient pre-training data. Likewise, the MAG datasets encountered challenges related to specificity and ambiguity, which resulted in TaxoLLaMA underperforming compared to WordNet-based datasets, as highlighted in Table 13.

A manual review of the MAG taxonomies reveals misclassifications, such as “olfactory toxicity in fish” being incorrectly categorized as a hyponym of “neuroscience.” To further evaluate the accuracy of the predicted hypernyms, we leveraged ChatGPT, drawing inspiration from recent research (Rafailov et al., 2023). We provided ChatGPT with the input queries, predicted nodes, and ground truth nodes, asking for a preference. As shown in Figure 6(b), ChatGPT often preferred neither of the options, with ground truth hypernyms being favored only slightly more often than the predicted ones. An example of the input query used is detailed in Appendix A, Example 14.

Our evaluation of the overlap between the MAG datasets and WordNet data reveals that they have little in common. Specifically, only 5% of the nodes in the MAG graph are also found in the WordNet graph. The overlap is even less in terms of edges, with only 2% in the CS domain and 4% in the PSY domain matching WordNet connections. Additionally, 92% of the identified connections lack any corresponding path within the WordNet structure. Among the connections that do overlap, we discovered that 28% in CS and 10% in PSY mistakenly identify nodes as their own hypernyms. These disparities highlight why TaxoLLaMA performs less effectively on MAG datasets, as they differ significantly from the WordNet-based data used during training.

In our final analysis, we visualized the embeddings, which highlighted a clear divergence between the predicted outcomes and the actual ground truth within the MAG subsets—a divergence that was not observed in the WordNet data. Detailed findings from this visualization are discussed in Appendix C.

Table 15. Statistics of Original Graph and the Constructed Graph with Highest FI Score. The Lower Part of the Table Corresponds to Constructed Graph Statistics.

Metric	Science	Environment	Food
	Original		
# Nodes	125	261	1486
# Edges	124	261	1533
	Constructed		
# Nodes	78	216	1132
# Edges	71	507	1372
# Nodes Missing	48	45	354
# Weak Components	8	5	51
# Nodes w/o original hypernym	4	5	39
# Nodes w/o path to original hypernym	29	70	308
# Nodes w/ path to original hypernym	44	140	784
Mean Distance to original hypernym	1.02	1.15	1.06

Table 16. Statistics on Definitions Generated with ChatGPT for Different Tasks.

Dataset	Total	Generated with ChatGPT	From Wikidata
MAG PSY	23,156	12,823	10,333
MAG CS	29,484	5,714	23,770
ANT	5,933	5,933	–
HyperLex	2,307	2,307	–

9.2 Taxonomy Construction

Our detailed assessment of the predicted graphs across different domain datasets, based on the data in Table 15, reveals consistent trends. In most cases, the gold standard graphs exhibit a higher number of edges, except for the environment domain. Interestingly, the model tends to miss entire clusters of nodes rather than isolated ones: around 30% of the nodes in the TaxoLLaMA graph are disconnected from their true parents, indicating they belong to separate components.

Although some paths generated by the model are highly accurate, its overall performance is inconsistent—either perfectly on target or completely off course. Frequently, paths with high perplexity are mistakenly discarded, suggesting the model struggles particularly with concepts that are neither highly specific nor overly broad but fall somewhere in the middle of the taxonomy.

This issue is exacerbated by the use of perplexity as a relative metric, where some edges are excluded because they exceed the defined perplexity threshold. However, adjusting the threshold to be more lenient can lead to the creation of incorrect edges. This challenge highlights the need to explore alternative methods, such as employing LLMs as embedding tools, to improve the model’s performance.

9.3 Lexical Entailment

Our review of the ANT dataset revealed that it comprises nearly 3,000 test samples but only 589 distinct verbs. This suggests that errors associated with a single verb could potentially be repeated multiple times throughout the dataset. However, when we looked at the overlap with WordNet, we found that only 7 of these verb forms matched.

After lemmatization, the number of unique verbs increases to 338, yet around 42% still cannot be found in WordNet. Moreover, for the verbs that do exist in WordNet, no corresponding paths were identified, which may have negatively impacted the model’s performance in this task.

Hyperlex offers more favorable statistics, with nearly 50% of the words being unique and 88% included in WordNet. However, only 27% of the word pairs are represented in the taxonomy, and 99% of these pairs are missing a connecting path.

Perplexity-related errors tend to have high values when dealing with polysemous pairs, such as “spade is a type of card,” and low values for synonyms or paraphrases, which indicates semantic closeness without implying a hypernymy relationship. This suggests that the model struggles with lexical diversity and ambiguity, highlighting the necessity of robust disambiguation capabilities in entailment tasks. Additional details are provided in Appendix D.

Table 17. 12 Error Types Made by TaxoLLaMA for Hypernym Prediction Detected by ChatGPT.

Error Type	Description
Overly Broad Predictions	The model often generates predictions that encompass a broader concept than the true hypernym.
Underly Broad Predictions	Some predictions are too narrow and fail to capture the broader concept represented by the true hypernym
Inclusion of Unrelated Terms	In some cases, the model includes terms in its predictions that are not directly related to the input word or true hypernym.
Repetition of Terms	The model occasionally repeats terms in its predictions, which might indicate redundancy or lack of diversity in its output.
Inadequate Coverage of Concepts	Some input words and true hypernyms receive predictions that lack comprehensive coverage of related concepts
Semantic Shift	The model might exhibit errors related to semantic shift, where the predicted terms are semantically related to the input word but do not accurately reflect the intended meaning or context.
Conceptual Ambiguity	The model may struggle with ambiguous input words or concepts, leading to predictions that lack clarity or specificity.
Domain-Specific Knowledge	Errors may arise due to a lack of domain-specific knowledge or understanding of specialized terminology.
Cultural or Contextual Bias	The model's predictions may be influenced by cultural or contextual biases inherent in the training data. This could lead to inaccuracies, especially when dealing with topics or concepts that vary across cultures or contexts.
Incomplete Understanding of Relationships	The model may struggle to understand complex relationships between concepts, leading to inaccurate predictions.
Word Sense Disambiguation	Errors may occur due to difficulties in disambiguating between different senses of a word.
Knowledge Gap	The model's predictions may reflect gaps in its knowledge or understanding of certain concepts, resulting in inaccurate or incomplete responses.

Table 18. Errors Type Distribution Across Subset Datasets for Hypernym Prediction: Hypernym Discovery and Taxonomy Enrichment.

	1A: English	2A: Medical	2B: Music	1B: Italian	1C: Spanish	MAG-CS	MAG-PSY	Noun	Verb
Error 1	72.49%	93.75%	100.0%	54.69%	49.08%	66.48%	85.43%	81.45%	73.39%
Error 2	2.61%	0.00%	0.0%	10.03%	10.62%	5.40%	1.40%	4.10%	2.58%
Error 3	9.44%	1.25%	0.0%	27.83%	12.09%	13.21%	11.20%	6.05%	21.03%
Error 4	15.46%	5.00%	0.0%	7.44%	18.21%	10.94%	0.70%	6.64%	0.43%
Error 5	0.00%	0.00%	0.00%	0.00%	0.00%	3.98%	1.26%	1.76%	2.58%

10 Conclusion

In this article, we comprehensively explored the use of LLMs for learning taxonomic relations, evaluating their effectiveness, and applying them to various downstream tasks. To facilitate taxonomy learning, we developed a dataset collection method using WordNet. Our fine-tuned models achieved state-of-the-art performance across several lexical semantic tasks, including taxonomy construction, hypernym discovery, taxonomy enrichment, and lexical entailment. Specifically, our models secured the top performance in 11 out of 16 tasks and ranked second in 4 others, demonstrating that LLMs are well-suited for solving taxonomy-related challenges.

Additionally, we conducted an extensive ablation study on our model, focusing on the learning of hyponymy by categorizing it into subtypes and levels of difficulty. Our findings shown that hyponymy is generally more challenging to learn than hypernymy, particularly for concepts located in the middle of the graph. Furthermore, our results suggest that some taxonomy relations are easier to learn for specialized terminology rather than for common concepts. The study also highlighted the potential of LLMs to refine existing taxonomies by utilizing multiple learned taxonomic relations to assess the accuracy of edges, which significantly improved overall performance. For taxonomy construction task, our experiments showed that hypernymy plays a crucial role, and that basic, straightforward brute-force methods currently yield the best results.

Table 19. Examples for each Error Type Made by TaxoLLaMA for Hypernym Prediction Detected by ChatGPT.

Error Type	Input	True	Predicted
Overly Broad Predictions	weed	vascular plant	plant, plant life, vegetable, vegetation, vascular plant, phanerogam, green plant, green thing, plant of the class, plant of the division
Underly Broad Predictions	audio	auditory communication	signal, transmission, transmission signal, communication, communication signal, message, message signal, information, information signal, data, data signal, notification, notification signal, signaling
Inaccurate Predictions	unconventionality	unorthodoxy	unorthodoxiness, unorthodoxy, unconformity, heterodoxy, dissidence, disobedience, disloyalty, disaffection, dissent, disagreement
Conceptual Ambiguity	giant red paintbrush	indian paintbrush	paintbrushrush, paint brush, larkspur, delphinium, liliaceous plant, lilioid monocot, lilioid
Incorrect definitions	advance directive status	documentation	advance directive, directive, decision support system, clinical decision support system, health information technology

Definition: A do-not-resuscitate order (DNR), also known as Do Not Attempt Resuscitation (DNAR), Do Not Attempt Cardiopulmonary Resuscitation (DNACPR)

Lastly, we carried out an in-depth analysis of model errors, revealing inconsistencies between WordNet and other taxonomies, and underscoring the need to revisit and possibly revise MAG taxonomies due to numerous misaligned relations.

Acknowledgements

The work of Irina Nikishina and Chris Biemann was supported by the DFG through the project “ACQuA: Answering Comparative Questions with Arguments” (grants BI 1544/7- 1 and HA 5851/2- 1) as part of the priority program “RATIO: Robust Argumentation Machines” (SPP 1999). The work of Alexander Panchenko was supported by the RSF project - 25-71-30008 “Laboratory for reliable, adaptive, and trustworthy Artificial Intelligence.”

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Notes

1. <https://wordnet.princeton.edu>
2. <https://en-word.net>
3. <https://github.com/globalwordnet/english-wordnet>
4. <https://en-word.net/static/english-wordnet-2024.ttl.gz>
5. <https://globalwordnet.github.io>
6. <http://wikidata.org>
7. <https://chat.openai.com>
8. <https://github.com/uhh-It/taxonomic-graphs-swj>
9. <https://huggingface.co/microsoft/Phi-3-mini-4k-instruct>
10. <https://huggingface.co/Qwen/Qwen2.5-7B-Instruct>
11. <https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>
12. <https://ai.meta.com/blog/meta-llama-3-1>
13. <https://huggingface.co/meta-llama/Llama-2-7b>
14. <https://huggingface.co/Qwen/Qwen2.5-7B-Instruct>
15. <https://huggingface.co/Qwen/Qwen2.5-1.5B-Instruct>
16. <https://huggingface.co/microsoft/Phi-3-mini-4k-instruct>
17. <https://huggingface.co/openai-community/gpt2>

18. <https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3>

References

- Bai, Y., Zhang, R., Kong, F., Chen, J., & Mao, Y. (2021). Hypernym discovery via a recurrent mapping model. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (pp. 2912–2921). Online: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.findings-acl.257>
- Berend, G., Makrai, M., & Földiák, P. (2018). 300-sparsans at SemEval-2018 task 9: Hypernymy as interaction of sparse attributes. In M. Apidianaki, S. M. Mohammad, J. May, E. Shutova, S. Bethard, & M. Carpuat (Eds.), *Proceedings of the 12th International workshop on semantic evaluation* (pp. 928–934). New Orleans, LA: Association for Computational Linguistics. <https://doi.org/10.18653/v1/S18-1152>
- Bernier-Colborne, G., & Barrière, C. (2018). CRIM at SemEval-2018 task 9: A hybrid approach to hypernym discovery. In M. Apidianaki, S. M. Mohammad, J. May, E. Shutova, S. Bethard, & M. Carpuat (Eds.), *Proceedings of the 12th international workshop on semantic evaluation* (pp. 725–731). New Orleans, LA: Association for Computational Linguistics. <https://doi.org/10.18653/v1/S18-1116>
- Bordea, G., Lefever, E., & Buitelaar, P. (2016). SemEval-2016 task 13: Taxonomy extraction evaluation (TExEval-2). In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)* (pp. 1081–1091). San Diego, CA: Association for Computational Linguistics. <https://doi.org/10.18653/v1/S16-1168>
- Camacho-Collados, J., Delli Bovi, C., Espinosa-Anke, L., Oramas, S., Pasini, T., Santus, E., Shwartz, V., Navigli, R., & Saggion, H. (2018). SemEval-2018 task 9: Hypernym discovery. In *Proceedings of The 12th international workshop on semantic evaluation* (pp. 712–724). New Orleans, LA: Association for Computational Linguistics. <https://doi.org/10.18653/v1/S18-1115>
- Camacho-Collados, J., Espinosa-Anke, L., Jameel, S., & Schockaert, S. (2019). A latent variable model for learning distributional relation vectors. In *Proceedings of the twenty-eighth international joint conference on artificial intelligence (IJCAI-19)* (pp. 4911–4917). International joint conferences on artificial intelligence organization. <https://doi.org/10.24963/ijcai.2019/682>
- Chen, C., Lin, K., & Klein, D. (2021). Constructing taxonomies from pretrained language models. In *Proceedings of the 2021 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies* (pp. 4687–4700). Online: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.naacl-main.373>
- Chernomorchenko, P., Panchenko, A., & Nikishina, I. (2024). Leveraging taxonomic information from large language models for hyponymy prediction. In *Analysis of images, social networks and texts—11th International Conference, AIST 2023, LNCS*, (Vol. 14486). Springer. https://link.springer.com/chapter/10.1007/978-3-031-54534-4_4
- Corro, L. D., Abujabal, A., Gemulla, R., & Weikum, G. (2015). FINET: Context-aware fine-grained named entity typing. In L. Márquez, C. Callison-Burch, J. Su, D. Pighin, & Y. Marton (Eds.), *Proceedings of the 2015 conference on empirical methods in natural language processing (EMNLP 2015), Lisbon, Portugal, September 17–21, 2015* (pp. 868–878). The Association for Computational Linguistics. <https://doi.org/10.18653/v1/d15-1103>
- Davies, A., Jiang, J., & Zhai, C. (2023). Competence-based analysis of language models. *CoRR* abs/2303.00333. <https://doi.org/10.48550/arXiv.2303.00333>
- Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies, NAACL-HLT 2019, Minneapolis, MN, June 2–7, 2019, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. <https://doi.org/10.18653/v1/n19-1423>
- Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., Xia, H., Xu, J., Wu, Z., Chang, B., Sun, X., Li, L., & Sui, Z. (2024). A survey on in-context learning. In Y. Al-Onaizan, M. Bansal, & Y. N. Chen (Eds.), *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 1107–1128). Miami, FL: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.emnlp-main.64>
- Egiazarian, V., Panferov, A., Kuznedelev, D., Frantar, E., Babenko, A., & Alistarh, D. (2024). Extreme compression of large language models via additive quantization. In *Proceedings of the 41st international conference on machine learning, ICML '24*. JMLR.org.
- Espinosa-Anke, L., & Schockaert, S. (2018). SeVeN: Augmenting word embeddings with unsupervised relation vectors. In E. M. Bender, L. Derczynski, & P. Isabelle (Eds.), *Proceedings of the 27th international conference on computational linguistics* (pp. 2653–2665). Santa Fe, NM: Association for Computational Linguistics. <https://aclanthology.org/C18-1225>
- Frantar, E., Ashkboos, S., Hoeffler, T., & Alistarh, D. (2022). GPTQ: Accurate post-training quantization for generative pre-trained transformers. *CoRR* abs/2210.17323. <https://doi.org/10.48550/arXiv.2210.17323>
- Grover, A., & Leskovec, J. (2016). node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 855–864).
- Guillou, L., & de Vroe, S. B. (2023). ANT dataset. <https://github.com/lguillou/ant>.

- Hanna, M., & Mareček, D. (2021). Analyzing BERT's knowledge of hypernymy via prompting. In *Proceedings of the Fourth blackboxnlp workshop on analyzing and interpreting neural networks for NLP* (pp. 275–282). Punta Cana, Dominican Republic: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.blackboxnlp-1.20>
- Held, W., & Habash, N. (2019). The effectiveness of simple hybrid systems for hypernym discovery. In A. Korhonen, D. Traum, & L. Márquez (Eds.), *Proceedings of the 57th annual meeting of the Association for Computational Linguistics* (pp. 3362–3367). Florence, Italy: Association for Computational Linguistics. <https://doi.org/10.18653/v1/P19-1327>
- Holt, X. (2019). Probabilistic models of relational implication. *arXiv preprint arXiv:1907.12048*.
- Hosseini, M. J., Chambers, N., Reddy, S., Holt, X. R., Cohen, S. B., Johnson, M., & Steedman, M. (2018). Learning typed entailment graphs with global soft constraints. *Transactions of the Association for Computational Linguistics*, 6, 703–717. https://doi.org/10.1162/tacl_a_00250
- Hosseini, M. J., Cohen, S. B., Johnson, M., & Steedman, M. (2021). Open-domain contextual link prediction and its complementarity with entailment graphs. In M. F. Moens, X. Huang, L. Specia, & S. W. Yih (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 2790–2802). Punta Cana, Dominican Republic: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.findings-emnlp.238>
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). Lora: Low-rank adaptation of large language models. In *The Tenth international conference on learning representations (ICLR 2022), Virtual Event, April 25–29, 2022*. OpenReview.net. <https://openreview.net/forum?id=nZeVKeeFYf9>.
- Jain, D., & Espinosa Anke, L. (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* (pp. 151–156). Seattle, WA: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.starsem-1.13>
- Jameel, S., Bouraoui, Z., & Schockaert, S. (2018). Unsupervised learning of distributional relation vectors. In I. Gurevych & Y. Miyao (Eds.), *Proceedings of the 56th annual meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 23–33). Melbourne, Australia: Association for Computational Linguistics. <https://doi.org/10.18653/v1/P18-1003>
- Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de Las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M. A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W. E. (2023). Mistral 7b. *ArXiv abs/2310.06825*. <https://api.semanticscholar.org/CorpusID:263830494>.
- Jiang, M., Song, X., Zhang, J., & Han, J. (2022). TaxoEnrich: Self-supervised taxonomy completion via structure-semantic representations. In *Proceedings of the ACM web conference 2022, WWW '22* (pp. 925–934). New York, NY: Association for Computing Machinery. <https://doi.org/10.1145/3485447.3511935>
- Joshi, M., Choi, E., Levy, O., Weld, D., & Zettlemoyer, L. (2019). Pair2vec: Compositional word-pair embeddings for cross-sentence inference. In *Proceedings of the 2019 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies, Volume 1 (Long and Short Papers)* (pp. 3597–3608). Minneapolis, MN: Association for Computational Linguistics. <https://www.aclweb.org/anthology/N19-1362>.
- Jurgens, D., & Pilehvar, M. T. (2016). SemEval-2016 task 14: Semantic taxonomy enrichment. In S. Bethard, M. Carpuat, D. Cer, D. Jurgens, P. Nakov, & T. Zesch (Eds.), *Proceedings of the 10th International workshop on semantic evaluation (SemEval-2016)* (pp. 1092–1102). San Diego, CA: Association for Computational Linguistics. <https://doi.org/10.18653/v1/S16-1169>
- Kauf, C., Ivanova, A. A., Rambelli, G., Chersoni, E., She, J. S., Chowdhury, Z., Fedorenko, E., & Lenci, A. (2023). Event knowledge in large language models: The gap between the impossible and the unlikely. *Cognitive Science*, 47(11), e13386. <https://doi.org/10.1111/cogs.13386>
- Lazaridou, A., Kuncoro, A., Gribovskaya, E., Agrawal, D., Liska, A., Terzi, T., Gimenez, M., de Masson d'Autume, C., Ruder, S., Yogatama, D., Cao, K., Kociský, T., Young, S., & Blunsom, P. (2021). Pitfalls of static language modelling. *CoRR abs/2102.01951*. <https://arxiv.org/abs/2102.01951>
- Lenz, M., & Bergmann, R. (2023). Case-based adaptation of argument graphs with wordnet and large language models. In S. Massie, & S. Chakraborti (Eds.), *Case-based reasoning research and development* (pp. 263–278). Cham: Springer Nature Switzerland. ISBN 978-3-031-40177-0.
- Lin, J., Tang, J., Tang, H., Yang, S., Chen, W. M., Wang, W. C., Xiao, G., Dang, X., Gan, C., & Han, S. (2024). Awq: Activation-aware weight quantization for on-device llm compression and acceleration. In P. Gibbons, G. Pekhimenko, & C. D. Sa (Eds.), *Proceedings of machine learning and systems* (Vol. 6, pp. 87–100). https://proceedings.mlsys.org/paper_files/paper/2024/file/42a452cbafa9dd64e9ba4aa95cc1ef21-Paper-Conference.pdf
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *CoRR abs/1907.11692*. <http://arxiv.org/abs/1907.11692>.
- McCrae, J. P., Rademaker, A., Bond, F., Rudnicka, E., & Fellbaum, C. (2019). English WordNet 2019—An open-source WordNet for English. In P. Vossen, & C. Fellbaum (Eds.), *Proceedings of the 10th Global Wordnet Conference* (pp. 245–252). Wrocław, Poland: Global Wordnet Association. <https://aclanthology.org/2019.gwc-1.31/>

- McKenna, N., Li, T., Johnson, M., & Steedman, M. (2023). Smoothing entailment graphs with language models. In J. C. Park, Y. Arase, B. Hu, W. Lu, D. Wijaya, A. Purwarianti, & A. A. Krisnadhi (Eds.), *Proceedings of the 13th international joint conference on natural language processing and the 3rd conference of the Asia-Pacific chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 551–563). Nusa Dua, Bali: Association for Computational Linguistics. <https://aclanthology.org/2023.ijcnlp-main.37>
- Miller, G. A. (1998). *WordNet: An Electronic Lexical Database*. MIT Press.
- Moskvoretskii, V., Neminova, E., Lobanova, A., Panchenko, A., & Nikishina, I. (2024a). TaxoLLaMA: WordNet-based model for solving multiple lexical semantic tasks. In L. W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd Annual meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2331–2350). Bangkok, Thailand: Association for Computational Linguistics. <https://aclanthology.org/2024.acl-long.127>
- Moskvoretskii, V., Panchenko, A., & Nikishina, I. (2024b). Are large language models good at lexical semantics? a case of taxonomy learning. In N. Calzolari, M. Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)* (pp. 1498–1510). Torino, Italy: ELRA and ICCL. <https://aclanthology.org/2024.lrec-main.133>
- Nikishina, I., Chernomorchenko, P., Demidova, A., Panchenko, A., & Biemann, C. (2023). Predicting terms in IS-a relations with pre-trained transformers. In J. C. Park, Y. Arase, B. Hu, W. Lu, D. Wijaya, A. Purwarianti, & A. A. Krisnadhi (Eds.), *Findings of the Association for Computational Linguistics: IJCNLP-AAACL 2023 (Findings)* (pp. 134–148). Nusa Dua, Bali: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.findings-ijcnlp.12>
- Nikishina, I., Tikhomirov, M., Logacheva, V., Nazarov, Y., Panchenko, A., & Loukachevitch, N. (2022a). Taxonomy enrichment with text and graph vector representations. *Semantic Web*, 13(3), 441–475. <https://doi.org/10.3233/SW-212955>
- Nikishina, I., Vakhitova, A., Tutubalina, E., & Panchenko, A. (2022b). Cross-modal contextualized hidden state projection method for expanding of taxonomic graphs. In D. Ustalov, Y. Gao, A. Panchenko, M. Valentino, M. Thayaparan, T. H. Nguyen, G. Penn, A. Ramesh, & A. Jana (Eds.), *Proceedings of TextGraphs-16: Graph-based methods for natural language processing* (pp. 11–24). Gyeongju, Republic of Korea: Association for Computational Linguistics. <https://aclanthology.org/2022.textgraphs-1.2>
- Pitarch, L., Bernad, J., Dranca, L., Bobed Lisboa, C., & Gracia, J. (2023). No clues good clues: out of context lexical relation classification. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st Annual meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 5607–5625). Toronto, Canada: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.acl-long.308>
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *Open Access Library Journal*, 1(8), 9.
- Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in neural information processing systems* (Vol. 36, pp. 53728–53741). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2023/file/a85b405ed65c6477a4fe8302b5e06ce7-Paper-Conference.pdf
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 3982–3992). Hong Kong, China: Association for Computational Linguistics. <https://doi.org/10.18653/v1/D19-1410>
- Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2008). The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1), 61–80. <https://doi.org/10.1109/TNN.2008.2005605>
- Shang, C., Dash, S., Chowdhury, M. F. M., Mihindukulasooriya, N., & Gliozzo, A. (2020). Taxonomy construction of unseen domains via graph-based cross-domain knowledge transfer. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics* (pp. 2198–2208). Online: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.199>
- Shen, J., Shen, Z., Xiong, C., Wang, C., Wang, K., & Han, J. (2020). Taxoexpand: Self-supervised taxonomy expansion with position-enhanced graph neural network. In *Proceedings of The Web Conference 2020, WWW '20* (pp. 486–497). New York, NY: Association for Computing Machinery. ISBN 9781450370233. <https://doi.org/10.1145/3366423.3380132>
- Staudemeyer, R. C., & Morris, E. R. (2019). Understanding lstm—A tutorial into long short-term memory recurrent neural networks. *arXiv preprint arXiv:1909.09586* <https://doi.org/10.48550/arXiv.1909.09586>
- Sun, K., Xu, Y., Zha, H., Liu, Y., & Dong, X. L. (2024). Head-to-tail: How knowledgeable are large language models (LLMs)? A.K.A. will LLMs replace knowledge graphs? In K. Duh, H. Gomez, & S. Bethard (Eds.), *Proceedings of the 2024 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies (Volume 1: Long Papers)* (pp. 311–325). Mexico City, Mexico: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.naacl-long.18>
- Tang, R., Zhang, X., Lin, J., & Ture, F. (2023). What do llamas really think? Revealing preference biases in language model representations. *arXiv preprint arXiv:2311.18812* <https://doi.org/10.48550/arXiv.2311.18812>

- Toral, A., & Muñoz, R. (2006). A proposal to automatically build and maintain gazetteers for named entity recognition by using Wikipedia. In *Proceedings of the Workshop on NEW TEXT Wikis and blogs and other dynamic text sources*. <https://aclanthology.org/W06-2809>.
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Canton-Ferrer, C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., & Fuller, B., ... Scialom, T. (2023a). Llama 2: Open foundation and fine-tuned chat models. *CoRR* abs/2307.09288. <https://doi.org/10.48550/arXiv.2307.09288>
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Canton-Ferrer, C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., & Fuller, B., ... Scialom, T. (2023b). LLaMA 2: Open foundation and fine-tuned chat models. *CoRR* abs/2307.09288. <https://doi.org/10.48550/arXiv.2307.09288>
- Vulić, I., Gerz, D., Kiela, D., Hill, F., & Korhonen, A. (2017). HyperLex: A large-scale evaluation of graded lexical entailment. *Computational Linguistics*, 43(4), 781–835. https://doi.org/10.1162/COLI_a_00301
- Vulić, I., & Mrkšić, N. (2018). Specialising word vectors for lexical entailment. In M. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2018 conference of the north american chapter of the Association for Computational Linguistics: Human language technologies, Volume 1 (Long Papers)* (pp. 1134–1145). New Orleans, LA: Association for Computational Linguistics. <https://doi.org/10.18653/v1/N18-1103>
- Wang, X., Li, Y., Wang, H., & Lv, M. (2023). Mkbqa: Question answering over knowledge graph based on semantic analysis and priority marking method. *Applied Sciences*, 13(10), 6104. <https://doi.org/10.3390/app13106104>
- Wang, Y., Liu, C., Chen, K., Wang, X., & Zhao, D. (2022). SMASH: Improving SMALL language models' few-SHOT ability with prompt-based distillation. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022* (pp. 6608–6619). Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.findings-emnlp.492>
- Ying, R., He, R., Chen, K., Eksombatchai, P., Hamilton, W. L., & Leskovec, J. (2018). Graph convolutional neural networks for web-scale recommender systems. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery; data Mining, KDD '18*. ACM. <http://dx.doi.org/10.1145/3219819.3219890>
- Zeng, Q., Lin, J., Yu, W., Cleland-Huang, J., & Jiang, M. (2021). Enhancing taxonomy completion with concept generation via fusing relational representations. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining, KDD '21* (pp. 2104–2113). New York, NY: Association for Computing Machinery. ISBN 9781450383325. <https://doi.org/10.1145/3447548.3467308>
- Zhang, J., Song, X., Zeng, Y., Chen, J., Shen, J., Mao, Y., & Li, L. (2021). Taxonomy completion via triplet matching network. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 35, pp. 4662–4670). <https://doi.org/10.1609/aaai.v35i5.16596>

Appendix A. ChatGPT for Definition Generation and Automatic Error Analysis

We employed two distinct prompts, referenced as 9 and 10, with ChatGPT to create definitions for datasets that originally lacked them. Specifically, the MAG PSY and MAG CS datasets used in taxonomy enrichment, along with the ANT and HyperLex datasets for lexical entailment, did not have predefined definitions. To address this, we designed custom prompts for each dataset type. For hypernym prediction, the prompts are geared towards generating a definition for a single word, whereas for lexical entailment, definitions for two words are generated simultaneously to assist in disambiguation. The resulting definition statistics are provided in Table 16.

(9)

Write a definition for the word/phrase in one sentence.

Example:

Word: caddle

Definition: act as a caddie and carry clubs for a player

Word: eszopiclone 3 mg

Definition:

(10)

Write a definition for Word 1 and Word 2. Each definition should be in one sentence. If a word is ambiguous, use the other word to disambiguate it.

Example:

Word 1: depression

Word 2: melancholy

Definition 1: a mental state characterized by a pessimistic sense of inadequacy and a despondent lack of activity

Definition 2: a constitutional tendency to be gloomy and depressed

Word 1: conflict

Word 2: disagreement

Example 11 shows how input was provided to ChatGPT for the purpose of automatically detecting error types in hypernym prediction. The corresponding output generated by the model is presented in Example 12. Additionally, Example 13 includes a prompt example used for automated error classification.

(11)

Here is the data in four columns separated with "\t":

topic: psychology input word: social construction of schizophrenia true: psychiatry, predictions: psychosis, social, psychological, and philosophical aspects, social psychol
psych, schizophrenia, antipsychotic, psychotic, antipsychotic

topic: psychology input word: use heroin true: methadone maintenance, predictions: heroin
opium tincture, heroin, opium, tincture, heroin, opium, and morphine, man, heroin, drug
addiction

topic: psychology input word: coordinated management of meaning true: social
psychology, predictions: communication disorders, language and literacy, and social studi
language and literacy, metaphor, metonymy, and politics, metonymy, metaphor

There is a topic, input word, correct hypernym (true), and model prediction
(predicted). Your task is to compare true answers and model predictions and summarize what
are the main error types of the model.

(12)

To summarize the main error types of the model's predictions compared to the true hypernym:

1. Overly Broad Predictions: The model often generates predictions that encompass a broad concept than the true hypernym.

For example:

"medication effects" was predicted with terms like "drug effects", "drug interactions", and "medication", which are broader than the true hypernym "clinical psychology" and "psychiatry."

2. Underly Broad Predictions: Conversely, some predictions are too narrow and fail to capture the broader concept represented by the true hypernym.

For example:

"schizoid personality" was predicted with terms like "personality disorders" and "personality theory," which are narrower than the true hypernym "psychosis."

3. Inclusion of Unrelated Terms: In some cases, the model includes terms in its prediction that are not directly related to the input word or true hypernym.

For example:

"pirisudanol" was predicted with terms like "dopamine receptor d1" and "dopamine receptor d2", which are unrelated to the true hypernym "psychiatry".

4. Repetition of Terms: The model occasionally repeats terms in its predictions, which might indicate redundancy or lack of diversity in its output.

For example:

"pirisudanol" was predicted with repeated terms like "piracetam" and "hydergine."

5. Inadequate Coverage of Concepts: Some input words and true hypernyms receive predictions that lack comprehensive coverage of related concepts.

For example:

"result category" was predicted with terms like "item response theory" and "qualitative

research," which might not fully encapsulate the scope of the true hypernym "mathematics education."

(13)

You are a helpful linguist. Your task is to do error classification for a dataset for hypernym prediction in a general domain. Here are possible classes:

1. Overly Broad Predictions: The model often generates predictions encompassing a broader concept than the true hypernym.
2. Underly Broad Predictions: Conversely, some predictions are too narrow and fail to capture the broader concept represented by the true hypernym.
3. Inaccurate Predictions: The model may predict words that are very semantically close to the true hypernym, but struggles with fitting into the exact wording
4. Conceptual Ambiguity: The model may struggle with ambiguous (polysemantic/multivalued) input words or concepts, leading to incorrect predictions.
5. Incorrect definitions: The model gets confused with the incorrect/ inaccurate definition retrieved from external sources

You will be given an input word/phrase, true hypernym, and candidate hypernyms. Please, return a Python dict of error classes

{1: 1, 2: 5, 3: 1, \ldots , 100:3}) for all instances below:

id: 1, input word: parathyroid_hormone, true hypernym: hormone, predicted: hormonal agent, hormon, hematopoietic growth factor, growth factor of the blood, growth regulator, growth substance, growth

\cdots id: 100, input word: proofreader, true hypernym: printer, predicted: reader, audience, audience member, spectator, viewer, listener, listener-in, hearer, recipient, witness, watcher, observer

The prompt shown in Example 14 was used with ChatGPT to automatically evaluate TaxoLLaMA results, as manual analysis revealed that the gold standard answers in the MAG PSY and MAG CS datasets might not always be reliable. Consequently, ChatGPT was tasked with selecting between the dataset's gold standard answer and the model's predicted candidate.

(14)

Here are the words in the psychological domain. Your task is to choose hypernym which is more relevant given two options.

Answer 1 / 2 / both / none

Example: social construction of schizophrenia option 1: psychosis option 2: psychiatry

Answer: 2

abdominal air sac option 1: air sac option 2: trachea

Answer:

Appendix B. Error Type Analysis

This section outlines the distribution of error types across different datasets for hypernym prediction, as detailed in Table 18. Furthermore, Table 19 provides an example of each error type, as classified by ChatGPT.

Appendix C. Distribution Visualization for Taxonomy Enrichment

In this section, we explore the distribution of ground truth and model predictions within the SentenceBert embedding space (Reimers & Gurevych, 2019). To ensure the results were not tied to a specific initialization, we performed two model runs with different seeds. We then projected the predicted candidates and ground truth hypernyms into the SentenceBert embedding space. For better visualization, we first reduced the embeddings to 50 dimensions using Principal Component Analysis (PCA), followed by t-SNE to condense the data into a two-dimensional space.

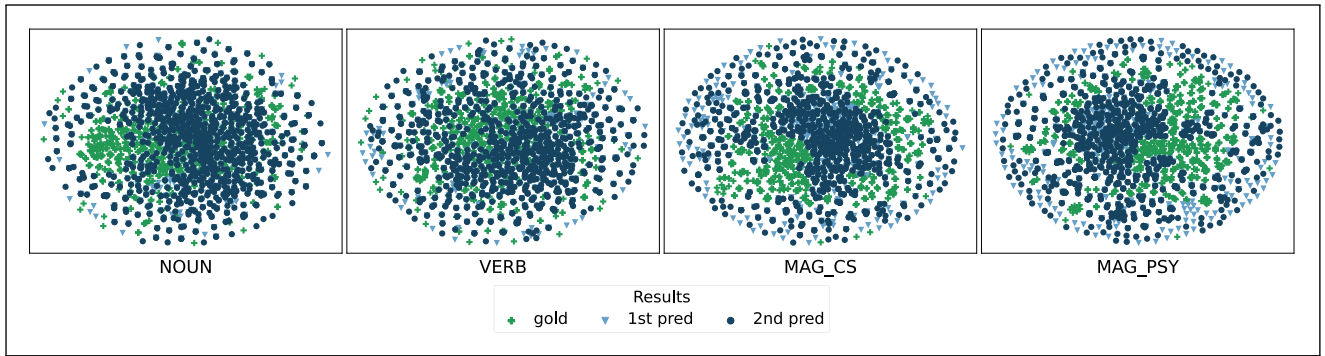


Figure 7. t-SNE plot of distributions of ground truth nodes and predicted nodes for taxonomy enrichment tasks. Each point represents a node, embedded with SentenceBert. Color represents ground truth or model predictions (we ran 2 predictions with different seeds)

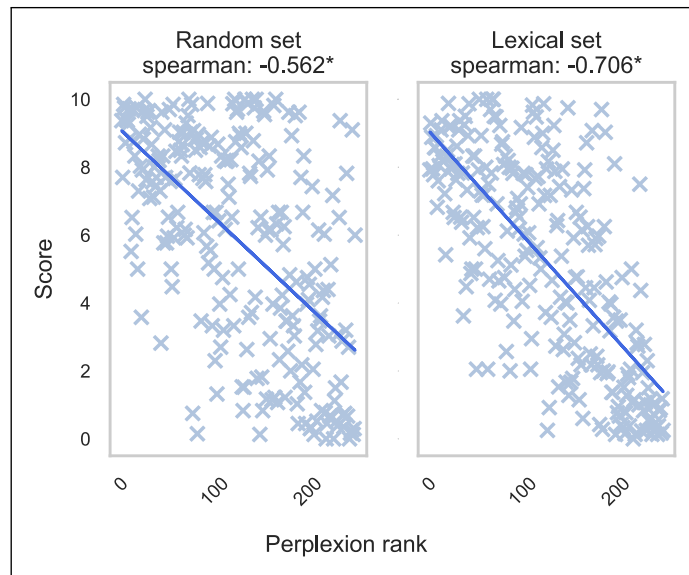


Figure 8. Correlation plot of the perplexion ranks with the annotator's score on Hyperlex test sets. The line over the dots is a trend found with linear regression. * shows that correlation has a p-value lower than $1e^{-4}$.

Figure 7 uncovers differences in how WordNet and the MAG subsets (MAG_CS and MAG_PSY) are represented in the embedding space. For WordNet, there is considerable overlap between the model's predictions and the gold standard, with only a few exceptions, likely linked to lower-ranked candidates. In contrast, the MAG subsets form two distinct clusters that barely overlap, indicating a notable divergence between the predicted and true hypernyms. Moreover, the MAG subsets contain more outliers, suggesting that the model may have missed the correct hypernym sense entirely in several instances. These observations could be influenced by the SentenceBert model's limitations, especially when dealing with concepts that are not well-represented in the training data.

Appendix D. Hyperlex Correlation Analysis

We also evaluated correlations using traditional methods for both test sets, as shown in Figure 8. A clear pattern emerges when a linear regression line is added to the data points, though this pattern is heavily influenced by outliers, particularly in the Random set. This finding is consistent with those from taxonomy construction, where the model also faces difficulties in accurately processing middle nodes or pairs with moderate entailment strength.

When evaluating gold scores in the range of 2 to 8, the Random set shows no discernible trend, underscoring the model's inconsistency in this area. The Lexical set, on the other hand, exhibits a slightly more defined trend within the

same range. Nonetheless, in both sets, pairs with either strong or minimal entailment are more reliably categorized. This differentiation significantly enhances the overall correlation, contributing to an encouraging correlation score.

Appendix E. Hyperparameters

Our investigation revealed the model’s pronounced sensitivity to both learning rate and scheduler settings. In the initial experiments, the successful application of a high learning rate was largely attributed to the LORA adapter, which subtly adjusts weights without causing major disruptions. However, when engaging in full model fine-tuning, we encountered significant instability, with the model oscillating between overfitting and underfitting, underscoring the need for refined hyperparameter tuning. Additionally, implementing 4-bit quantization requires careful calibration of the learning rate, as this compression method significantly alters the weight distribution, making it necessary to employ strategies that effectively restore the model’s knowledge.

During fine-tuning, we chose a smaller batch size to better align the model with datasets that often have limited samples. However, increasing the learning rate and batch size did not enhance performance, likely due to the model having fewer steps to adapt to domain-specific features. This was not the case during WordNet pre-training, where different trends were observed.

In contrast to certain instruction tuning strategies, our method does not calculate loss based on the instruction itself but rather focuses exclusively on the target tokens.

Our experiments were carried out on Nvidia A100 or Quadro RTX 8000 GPUs. Pre-training for both TaxoLLaMA and TaxoLLaMA_{bench} took around 6 GPU hours, while TaxoLLaMA_{verb} required less than 1 hour. Fine-tuning the MAG subsets took 5 GPU hours due to the extended definitions, whereas other datasets were fine-tuned in under an hour.