LLM-Hype: A Targeted Evaluation Framework for Hypernym-Hyponym Identification in Large Language Models

Qiu Ji
Nanjing University of Posts and Telecommunications
Nanjing, China
qiuji@njupt.edu.cn

Pengfei Shen
Nanjing University of Posts and Telecommunications
Nanjing, China
1224097324@njupt.edu.cn

Haolei Zhu
Nanjing University of Posts and Telecommunications
Nanjing, China
1224097340@njupt.edu.cn

Guilin Qi
Southeast University
Nanjing, China
gqi@seu.edu.cn

Yang Sheng
Nanjing University of Posts and Telecommunications
Nanjing, China
1222097628@njupt.edu.cn

Lianlong Wu
University of Oxford
Oxford, United Kingdom
lianlong.wu@cs.ox.ac.uk

Kang Xu
Nanjing University of Posts and Telecommunications
Nanjing, China
kxu@njupt.edu.cn

Yuan Meng
Southeast University
Nanjing, China
yuan_meng@seu.edu.cn

ABSTRACT

Understanding taxonomic relationships, such as hypernymy and hyponymy, is a fundamental aspect of conceptual reasoning. While large language models (LLMs) have shown impressive performance across a variety of NLP tasks, their ability to recognize and infer hierarchical relationships remains underexplored. In this work, we introduce **LLM-Hype**, a targeted evaluation framework designed to systematically assess LLMs' capability in hypernym-hyponym reasoning. Our framework constructs taxonomy graphs from diverse resources that encode hierarchical semantics, we generate carefully labeled test cases containing both positive and negative examples. To probe model reasoning under varied conditions, we design four complementary prompting strategies: (1) direct prompts assessing internalized knowledge, (2) definition-based prompts augmenting queries with natural language definitions, (3) structure-based prompts leveraging structural cues from taxonomies, and (4) hybrid prompts combining definitional and structural information. We conduct a comprehensive evaluation across five topic-specific datasets and a total of 12 representative LLMs. Experimental results reveal that definition-based prompts consistently yield the highest accuracy, underscoring the effectiveness of explicit semantic context. In contrast, structure-based prompts do not provide consistent benefits and may degrade performance in some cases. Among the evaluated models, GPT-4o and Gemini-2.5 demonstrate robust and stable performance, while GLM-4 exhibits divergent behavior, suggesting underlying differences in reasoning strategies or training data. Overall, LLM-Hype offers a robust framework for analyzing conceptual hierarchy understanding in LLMs and provides valuable insights into their reasoning capabilities and limitations.

VLDB Workshop Reference Format:

Qiu Ji, Pengfei Shen, Haolei Zhu, Guilin Qi, Yang Sheng, Lianlong Wu, Kang Xu, and Yuan Meng. LLM-Hype: A Targeted Evaluation Framework for Hypernym-Hyponym Identification in Large Language Models. VLDB 2025 Workshop: LLM+Graph.

1 INTRODUCTION

Hierarchical semantic relationships, particularly hypernym-hyponym ("is-a") relations, are fundamental to linguistic semantics and play a pivotal role in various natural language processing (NLP) applications, such as word sense disambiguation, information retrieval, text classification, and knowledge graph and ontologies construction [16, 20, 29]. These is-a relationships enable systems to infer implicit knowledge by leveraging structured taxonomies or ontologies. For instance, recognizing that a "dalmatian" is a type of "dog" allows NLP systems to generalize and reason across different levels of abstraction, improving both precision and recall in downstream tasks.

Traditional methods for identifying hypernym-hyponym relations can be categorized into pattern-based approaches, distributional methods, and hybrid techniques. Pattern-based methods rely on predefined linguistic patterns (e.g., Hearst patterns) to detect these relations [21, 40]. While they often achieve high precision, their coverage is inherently limited, as they may not capture novel or uncommon expressions and typically require labor-intensive

Corresponding author: gqi@seu.edu.cn.

Proceedings of the VLDB Endowment. ISSN 2150-8097.

pattern engineering. Distributional methods, grounded in the distributional hypothesis, infer taxonomic relations by analyzing contextual co-occurrence patterns in large corpora [26, 47]. Although effective within the scope of the training domain, they often struggle to generalize to out-of-domain settings and may lack sensitivity to fine-grained semantic distinctions. Hybrid approaches seek to integrate the precision of pattern-based methods with the broader coverage of distributional models [14, 48]. However, their performance remains highly dependent on the quality and diversity of the underlying data sources, and they continue to face challenges such as data sparsity, noise, and representation bias.

Recent advances in large language models (LLMs), such as GPT-4 [34] and LLaMA [43], have opened up new possibilities for conceptual and relational reasoning. Unlike traditional methods that depend on explicit patterns or distributional signals, LLMs are trained on massive corpora and learn to encode semantic knowledge implicitly through next-token prediction objectives [8]. This enables them to generalize beyond surface-level cues and potentially infer hypernym-hyponym relations even in the absence of explicit syntactic indicators. Moreover, LLMs have demonstrated impressive performance on a wide range of zero-shot and few-shot tasks, suggesting a latent capacity for taxonomic reasoning [27, 46]. However, despite their empirical success, it remains unclear how reliably and systematically LLMs can identify and reason over hierarchical relations—a question that calls for targeted and rigorous evaluation.

In this paper, we present LLM-Hype, a targeted evaluation framework for systematically assessing the ability of LLMs to infer hypernym-hyponym relationships. Our framework builds taxonomy graphs from a diverse set of resources that encode hierarchical semantics via concept definitions, instance-level data, and schema-level structures. From these graphs, we generate carefully labeled test cases containing both positive and negative examples. To probe model reasoning under varied conditions, we design four complementary prompting strategies: (1) direct prompts, which assess the model's internalized knowledge by asking it to judge a relationship with no additional context; (2) definition-based prompts, which augment the query with natural language definitions of the involved concepts to test reasoning over explicit semantic content; (3) structure-based prompts, which provide structural cues from the taxonomy to evaluate the model's capacity for structured reasoning; and (4) hybrid prompts, which integrate both definitional and structural information, offering a comprehensive input that challenges the model to synthesize heterogeneous signals. Each test case is paired with all prompt types and submitted to the LLMs under evaluation.

We conduct a systematic evaluation of LLM performance across diverse datasets, prompting strategies, and model families, using the number of correct and incorrect identifications, along with overall accuracy, as evaluation metrics. Specifically, we use two data sources: a traditional hypernym-hyponym dataset (SE16) and a widely used knowledge graph (DBpedia). From these sources, we construct five topic-specific datasets, each containing 20 positive and 20 negative examples randomly sampled from a larger pool of test cases. In total, 12 representative LLMs are included in the comparison. Our experimental results reveal several key trends. Definition-based prompts generally yield the highest accuracy, suggesting that providing semantic context helps models make more informed relational judgments. In contrast, prompts incorporating structural information do not consistently lead to performance gains and may even hinder performance in certain cases. In terms of model comparison, some LLMs—such as GPT-4o and Gemini-2.5—demonstrate consistently strong and stable performance, while others, like GLM-4, show distinct behavior compared to the majority of models, indicating potential differences in reasoning strategies or training biases.

Our key contributions are as follows:

- **Taxonomy-Aware Evaluation Framework (LLM-Hype):** We propose LLM-Hype, a comprehensive evaluation framework for systematically assessing the hypernym-hyponym inference ability of LLMs. This framework leverages diverse hierarchical resources (e.g. concept definitions and knowledge graphs) to construct taxonomy graphs, which are used to generate labeled test cases with both positive and negative examples. This method advances over ad-hoc benchmarks by explicitly modeling taxonomy graphs, ensuring comprehensive coverage of semantic relationships while controlling for data bias–a critical innovation for reliable LLM evaluation in relational reasoning.
- **Systematic Prompting Strategies for Hierarchical Reasoning:** We design and evaluate four distinct prompting strategies—direct, definition-based, structure-based, and hybrid prompts—to probe different aspects of model reasoning. These strategies provide varied levels of semantic context, from simple queries to integrated definitional and structural information. Unlike traditional single-prompt approaches, these strategies systematically probe different reasoning facets–internal knowledge, semantic understanding, structural logic, and integrated reasoning–enabling a fine-grained analysis of model capabilities and limitations in taxonomic tasks.
- **Systematic Evaluation Across Datasets and Model Families:** We conduct a robust evaluation across five topic-specific datasets constructed from two data sources (SE16 and DBpedia), and 12 representative LLMs. This enables a thorough comparison of model performance under different prompting conditions, highlighting trends and model-specific behaviors.
- **Key Insights on Prompting and Model Performance:** Our experiments reveal that definition-based prompts generally enhance accuracy, while structural prompts do not always lead to improvements. Additionally, model comparison reveals notable differences in reasoning strategies, with some LLMs (e.g., GPT-4o and Gemini-2.5) showing strong performance stability, while others (e.g., GLM-4) display more divergent behavior, suggesting variations in training and reasoning.

The paper is organized as follows: Section 2 provides background knowledge on ontologies and large language models. Section 3 details our evaluation approach. Section 4 presents the experimental results, while Section 5 reviews related work. Finally, Section 6 concludes the paper and outlines directions for future research.

2 BACKGROUND KNOWLEDGE

This section provides an overview of key concepts related to taxonomy graph, ontologies, and LLMs.

2.1 Taxonomy Graph

A taxonomy graph is a structured representation of hierarchical relationships among concepts, typically organized through is-a (hypernym-hyponym) links. In such graphs, nodes represent concepts or entities, and directed edges indicate subsumption relationships, where one concept is a subclass or a more specific instance of another. This structure forms a backbone for organizing knowledge in a way that supports inheritance, abstraction, and category-based reasoning [19].

Unlike general knowledge graphs that may contain a wide variety of relation types (e.g., part-of, located-in, causes), taxonomy graphs focus specifically on semantic hierarchies. They are commonly derived from structured ontologies (e.g., WordNet [32]), instance-level knowledge bases (e.g., DBpedia [4]), or curated schema-level taxonomies. These graphs are foundational in many tasks, such as concept classification, semantic search, question answering, and textual entailment [33].

In the context of evaluating LLMs, taxonomy graphs provide a grounded and interpretable structure for probing the model's ability to infer hypernym-hyponym relationships. Reasoning accurately over these hierarchical structures requires models to recognize abstract category inclusion, generalization, and transitive subsumption—all key aspects of symbolic and conceptual understanding [7].

2.2 Ontologies

An ontology is a structured representation of knowledge that consists of classes, object properties, data properties, and individuals. In this framework, individuals represent concrete instances in a specific domain, classes denote sets of such instances, object properties describe relationships between individuals, and data properties associate individuals with literal values such as numbers, strings, or dates. For example, the property hasWeight is a data property linking a person to a numeric value, while hasSpouse is an object property representing a relation between two individuals. Ontologies are commonly expressed using the Web Ontology Language (OWL) [22], a widely adopted standard in the Semantic Web.

A central component of an ontology is the taxonomy graph, which encodes hierarchical relationships among concepts via subclass (or is-a) links [23]. These structures are derived from TBox axioms such as $C \sqsubseteq D$, indicating that concept C is a subclass (hyponym) of D (hypernym). The resulting taxonomy forms a directed acyclic graph that supports concept subsumption, inheritance, and logical reasoning.

Beyond taxonomic hierarchies, ontologies also incorporate rich structural semantics. These include domain and range constraints on properties, disjointness axioms, property chains, and cardinality restrictions. Such schema-level structures provide additional context about how concepts and relationships are organized and interact. This structural information is particularly useful for enabling inference, detecting inconsistencies, and supporting knowledge-based reasoning in applications such as semantic search, information integration, and question answering.

In this work, we utilize both taxonomic hierarchies and structural relations from ontologies and knowledge graphs to construct evaluation scenarios for assessing LLMs' capability in conceptual reasoning, with a focus on hypernym-hyponym inference.

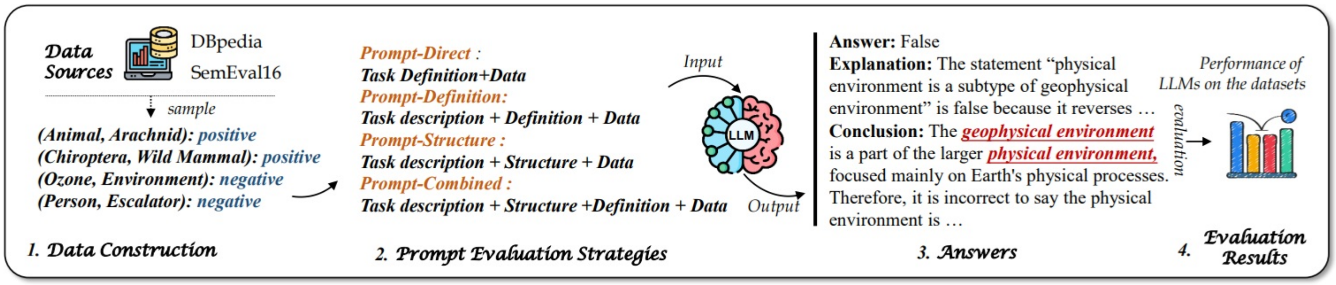
2.3 Large Language Models

Large Language Models (LLMs) are a class of deep learning-based NLP models designed to generate, comprehend, and manipulate natural language texts at scale. Their core innovation lies in leveraging massive datasets and powerful architectures—primarily the Transformer — to capture semantic, syntactic, and contextual nuances of language. Trained with self-supervised objectives, LLMs use mechanisms such as autoregression or self-attention to model language effectively. Over recent years, LLMs have demonstrated remarkable capabilities across a wide range of tasks, including text generation [37], machine translation [28], dialogue systems [2], and code generation [24], making them a central pillar in the advancement of modern artificial intelligence.

The development of LLMs builds upon a series of foundational milestones in NLP. The era of word embeddings, exemplified by models such as Word2Vec [31] and GloVe [38], marked the transition from symbolic or statistical methods to distributed semantic representations. However, these embeddings were static, failing to capture context-dependent meanings. Contextualized models like ELMo [39] addressed this limitation by introducing dynamic word representations using bidirectional LSTMs. The introduction of the Transformer architecture by Vaswani et al. [44] revolutionized the field by enabling parallelizable and scalable self-attention mechanisms. The pretrain-finetune paradigm, popularized by BERT [13], enabled general-purpose language models through large-scale unsupervised pretraining followed by supervised finetuning. In parallel, autoregressive models such as GPT [8] emphasized generative capabilities and cross-task generalization. More recent models like GPT-4 [35], Flamingo [1], and ChatGPT integrate multimodal inputs and leverage Reinforcement Learning from Human Feedback (RLHF) [36] to align model behavior with human preferences, improving both response quality and safety.

To conduct a representative and comprehensive evaluation of current LLMs, we selected a range of state-of-the-art models by leading research institutions and technology companies (see Table 1). The release dates listed in the table are primarily sourced from Wikipedia[1] or official websites. These models include OpenAI's GPT-3.5, GPT-4 [34], and GPT-4o, which reflect major milestones in scaling and instruction tuning; Google's Gemini 2.5 [18], a powerful multimodal model designed for advanced reasoning and integration across input types; Meta's LLaMA-3.3 [30], an influential open-weight model widely used in academic research; and several competitive models from the Chinese AI ecosystem, such as GLM-4 (Zhipu AI) [50], Qwen-2.5 (Alibaba) [3], Baichuan-4 Turbo (Baichuan) [5], Doubao-1.5 (ByteDance) [9], and Hunyuan (Tencent) [42], all of which demonstrate strong performance in multilingual and instruction-following tasks. We also included DeepSeek-R1

[1] https://en.wikipedia.org/wiki/


Data Sources
DBpedia
SemEval16
sample
(Animal, Arachnid): positive
(Chiroptera, Wild Mammal): positive
(Ozone, Environment): negative
(Person, Escalator): negative
1. Data Construction
Prompt-Direct :
Task Definition+Data
Prompt-Definition:
Task description + Definition + Data
Prompt-Structure :
Task description + Structure + Data
Prompt-Combined :
Task description + Structure +Definition + Data
2. Prompt Evaluation Strategies
Input
LLM
Output
3. Answers
evaluation
Performance of LLMs on the datasets
4. Evaluation Results

Answer: False
Explanation: The statement "physical environment is a subtype of geophysical environment" is false because it reverses …
Conclusion: The geophysical environment is a part of the larger physical environment, focused mainly on Earth's physical processes. Therefore, it is incorrect to say the physical environment is …

Figure 1: The framework to evaluate the ability of a LLM to identify hypernym-hyponym relationships.

Model name	Version	Release date	Affiliated institution
GPT-3.5	3.5	2022.11	OpenAI
GPT-4	4	2023.03	OpenAI
GLM-4	4	2024.01	Zhipu AI
GPT-4o	4o	2024.05	OpenAI
Baichuan-4	4-Turbo	2024.05	Baichuan Intelligent Technology
Qwen-2.5	2.5-72b-instruct	2024.09	Alibaba
DeepSeek-R1	R1	2024.11	DeepSeek
DeepSeek-V3	V3	2024.12	DeepSeek
LLaMA-3.3	3.3-70b-instruct	2024.12	Meta
Doubao-1.5	1.5-pro-32k	2025.01	ByteDance
Hunyuan	Standard	2025.02	Tencent
Gemini-2.5	2.5-pro-exp	2025.03	Google

Table 1: Overview of selected LLMs for evaluation.

and DeepSeek-V3 [12], open-weight models designed for both code and natural language tasks. The selection covers a diverse set of model architectures, affiliations, and release timelines, aiming to capture the breadth of current LLM development across different regions and application domains. This ensures that the comparative evaluation reflects both the technical advancements and the global landscape of LLM deployment.

3 APPROACH

In this section, we first present an overview of our proposed approach for evaluating the reasoning depth of LLMs in detecting logical conflicts within DL ontologies. We then describe in detail the two core tasks of our approach: dataset construction and prompt design.

3.1 Evaluation Framework

The overall process of our approach is illustrated in Figure 1. We present *LLM-Hype*, a targeted evaluation framework designed to assess the ability of LLMs to identify hypernym-hyponym relationships.

The framework consists of three key components: dataset construction, prompt design, and evaluation metric specification. It starts with a collection of data sources, including traditional hypernym-hyponym datasets, knowledge graphs, and ontologies. These sources offer rich semantic structures that form the foundation of the evaluation benchmark. From these resources, we construct concept hierarchies and derive labeled pairs — positive examples representing valid hypernym-hyponym relations, and negative examples such as sibling or semantically unrelated concepts.

To systematically probe the reasoning abilities of LLMs, we design a variety of prompts to leverage different types of information, including explicit relational cues, textual definitions of concepts, and structural signals derived from hierarchical concept graphs. By varying the content and context presented in these prompts, we aim to explore how LLMs interpret and utilize semantic information when identifying hypernym-hyponym relationships. This strategy allows us to assess not only overall performance, but also the models' sensitivity to different forms of semantic representation.

For evaluation, we use the number of correctly / incorrectly identified samples, and the accuracy in binary classification as the primary metric—measuring whether an LLM correctly identifies hypernymy. We conduct comparative experiments across multiple LLM families and prompt types, providing a multidimensional perspective on performance. This framework not only benchmarks current capabilities but also highlights the strengths and limitations of LLMs in structured semantic reasoning tasks.

3.2 Dataset Construction

The dataset construction process begins with the creation of a concept hierarchy graph, built upon existing data sources. Based on this hierarchy, both positive and negative examples are generated for evaluation.

To construct the concept hierarchy, we consider two main sources: traditional hypernym-hyponym extraction datasets and ontology-based datasets rich in structural information. Traditional datasets typically provide pairs of concepts with annotated hypernym-hyponym relations along with textual definitions for each concept. In contrast, ontology-based datasets not only declare subsumption relations but also offer additional semantic information, such as formal definitions, and a variety of relationships among concepts, properties, and instances. For each extracted hypernym-hyponym pair, we first initialize an empty directed graph. Each node in the graph corresponds to either a hypernym or a hyponym. A directed edge is added from a hyponym node to a hypernym node if such a relation exists between them. For clarity, we define the following variables and functions:

- *ls*: The set of all leaf nodes, i.e., nodes with no outgoing edges.
- *rs*: The set of all root nodes, i.e., nodes with no incoming edges.
- *paths*(q, r): The set of all paths from a leaf node q to a reachable root node r.
- *pathNodes*(q, p): For a given path $p \in paths(q, r)$, this denotes the set of all intermediate nodes on the path, excluding q. Each node in *pathNodes*(q, p) is considered a hypernym of q.
- *anchors*(q): The union of all intermediate nodes from all valid paths from q to any reachable root node, i.e.,

$$anchors(q) = \bigcup_{root \in rs} \bigcup_{p \in paths(q, root)} pathNodes(q, p).$$

When constructing positive examples, we compute anchors for each leaf node q, and define the corresponding hypernym-hyponym pairs as:

$$nodesPairs(q) = \{(q, sup) \mid sup \in anchor,\ anchor \in anchors(q)\}$$

That is, each pair (q, sup) represents a hyponym-hypernym relation, where q is the hyponym and *sup* is one of its hypernyms derived from the hierarchy. We denote the complete set of positive examples as *allNodesPairs*, defined as:

$$allNodesPairs = \bigcup_{leaf \in ls} nodesPairs(leaf)$$

The set *allNodesPairs* constitutes candidate positive examples to be used in the evaluation process.

To construct negative examples, we utilize both hierarchical and structural information in the following three ways: (1) We identify sibling nodes—i.e., nodes that share the same immediate parent in the hierarchy—based on the commonly adopted assumption in ontology construction that sibling concepts are mutually exclusive within the same classification scheme. For each pair of sibling nodes, we generate a negative example. (2) We construct negative examples by inverting the semantic relations used in the construction of positive examples. Specifically, we refer to the templates of positive prompts and reverse their semantic direction to generate negative counterparts. (3) We extract additional negative examples using disjointness relations. For any node n, if a concept c is explicitly declared to be disjoint with n and c is formally defined in the dataset, then both (n, c) and (c, n) are treated as negative examples. Furthermore, for each ancestor node a of n, we also construct the negative pairs (a, c) and (c, a).

It should be noted that: (1) Based on the methods for constructing positive and negative examples, a large number of candidate examples may be generated. For specific experiments, a fixed number of examples can be selected using an appropriate sampling strategy. (2) Certain negative examples may lead to incorrect constructions due to special cases. For instance, suppose a classification schema defines both Student and MasterStudent as subclasses of Person. In such a case, creating a negative example like (MasterStudent, Student) would be invalid. To ensure correctness, all automatically selected negative examples are manually verified. If any incorrect negative example is identified, it is replaced with another selected example from the remaining pool of negative examples.

3.3 Prompt Design

In this section, we design four distinct types of prompts to systematically evaluate the capability of LLMs in identifying hypernym-hyponym relationships. These prompts are constructed based on different levels and sources of contextual information, aiming to disentangle the contribution of internal knowledge, definitional semantics, and structural signals. The four types are as follows:

- **Direct Prompts (Prompt-Direct)** rely solely on explicitly querying whether a hypernym-hyponym relation holds between a given node pair. These prompts exclude any definitional or structural context, requiring the LLM to rely entirely on its internal knowledge. The model is then asked to verify the correctness of this claim.
- **Definition-based Prompts (Prompt-Definition)** extend the direct prompt by providing natural language definitions for both concepts in the given node pair. These definitions can be obtained from the source dataset, generated by the LLM, or retrieved from external lexical resources. The goal is to supply semantic content that may help the model better interpret the conceptual relationship before making a judgment.
- **Structure-based Prompts (Prompt-Structure)** incorporate structural information associated with the given concepts. Depending on the underlying dataset, this may include concept-level hierarchy (e.g., ancestors, descendants, siblings), instance-level connections (e.g., shared instances), or schema-level axioms (e.g., disjointness or domain/range constraints). When both instance-level and schema-level information are available, the prompt includes both, encouraging the model to reason using the structured context.
- **Combined Prompts (Prompt-Combined)** fuse both definitional and structural information. Specifically, the prompt presents definitions of the given concepts along with their structural context derived from the concept graph. This comprehensive design provides the richest informational grounding, intended to assess the model's ability to synthesize heterogeneous signals in making relational judgments.

In the following, we design specific prompt templates for a node pair (A, B). In each template, `[MASK]` is filled with either true or false. The interpretation of `[p]` depends on the direction of the taxonomy initialization. If the taxonomy is top-down, `[p]` represents a superclass relation (e.g., "the superclass", "the parent class", "a supertype"); in a bottom-up taxonomy, `[p]` denotes a subclass relation (e.g., "a subclass", "the child class", "a subtype", "a kind").

First, we design a prompt template for **Prompt-Direct**, following the format proposed in [17]:

> *"Identify whether the following statement is true or false: (A |B) is [p] of (B |A). This statement is [MASK]. You need to give not only true or false, but also an explanation of why you made this choice."*

Such prompts serve as a baseline for assessing the language model's ability to recognize taxonomic relationships without relying on any external context.

We further design a second type of prompt template that incorporates the definitions of the concepts involved:

> *"Suppose [qm] is defined as [def(qm)], and [an] is defined as [def(an)]. Identify whether the following statement is true or false: (A |B) is [p] of (B |A). This statement is [MASK]. You need to give not only true or false, but also an explanation of why you made this choice."*

In this template, `[def(A)]` and `[def(B)]` denote the textual definitions of the hyponym and hypernym, respectively.

We also propose a third type of prompt template that leverages the structural relations associated with the candidate concept:

> *"Suppose [A] has the following relations: NL1, NL2, ... Suppose [B] has the following relations: NL1', NL2', ... Based on the above relational information, identify whether the following statement is true or false: ([A] | [B]) is [p] of ([B] | [A]). This statement is [MASK]. You must provide not only true or false, but also an explanation for your decision."*

Here, `NL1, NL2, ...` and `NL1', NL2', ...` refer to natural language expressions of structural relations associated with `[A]` and `[B]`, respectively — such as their parent classes, subclasses, instances, or related properties. The model is instructed to use these relational cues to assess the hierarchical relationship between `[A]` and `[B]`.

Finally, we design a fourth type of prompt template that integrates both structural relations and concept definitions to support more comprehensive reasoning:

> *"Suppose [A] has the following relations: NL1, NL2, ... Suppose [B] has the following relations: NL1', NL2', ... Suppose [A] is defined as [def(A)], and [B] is defined as [def(B)]. Based on the above definitions and relational information, identify whether the following statement is true or false: ([A] | [B]) is [p] of ([B] | [A]). This statement is [MASK]. You must provide not only true or false, but also an explanation for your decision."*

By combining structural and definitional information, this prompt encourages the model to engage in multi-source reasoning.

4 EXPERIMENTS

In this section, we first describe the dataset constructed using the method presented in Section 3.2. We then evaluate the impact of different prompting strategies on model performance. Based on the hard cases identified in this initial evaluation, we further compare the performance of various LLMs.

4.1 Datasets

To construct a concept hierarchy graph, we selected two well-established sources of hypernym-hyponym data: the SemEval-2016 dataset and the DBpedia knowledge graph. The SemEval-2016 dataset originates from Task 13 of the SemEval-2016 competition: Taxonomy Extraction Evaluation [6]. This task focuses on developing NLP systems capable of automatically identifying and organizing scientific concepts and terms into taxonomies that accurately represent domain-specific knowledge structures. From this dataset, we selected three classification schemes: **SE-Environment**, **SE-Food**, and **SE-Science**. DBpedia, by contrast, is an open-source structured knowledge base extracted from Wikipedia, which transforms unstructured textual content into structured, queryable data. Given the extensive volume of information available in DBpedia, we used SPARQL queries—the standard language for querying RDF data [25]—to extract both annotation data and structural relationships associated with two high-level: classes Animal and Person. The resulting datasets are referred to as **DBpedia-Animal** and **DBpedia-Person**, respectively. In total, five data sources were selected, and then the corresponding concept hierarchy graphs were constructed.

Based on the constructed graphs, both positive and negative examples were generated. Specifically, over 10,000 positive and 10,000 negative examples were produced for SE-Environment; over 500 positive and 3,000 negative examples for SE-Food; and over 6,000 positive and 6,000 negative examples for SE-Science. Given the large volume of examples generated from these datasets, we randomly sampled 20 positive and 20 negative examples from each dataset for use in the experiment. For the DBpedia-based datasets, we directly constructed 20 positive and 20 negative examples for each of DBpedia-Animal and DBpedia-Person.

4.2 Effect of Different Prompting Strategies

To evaluate the impact of different prompting strategies, we select three representative models: DeepSeek-R1, GLM-4, and GPT-4. First, we compare the number of incorrect identifications each model makes on the constructed dataset. Then, we analyze the models' performance across different types of test samples. Finally, we provide a summary of the overall performance under each prompting strategy.

First of all, we compare the number of false identifications made by DeepSeek-R1, GLM-4, and GPT-4 across five datasets and four prompt types (see Figure 2). In the figure, bars are omitted when a model correctly identifies all test examples for a given dataset. From the figure, we can obtain the following observations. **(1) Comparing DBpedia and SE16 datasets:** The examples in the two DBpedia-based datasets are generally much simpler than those in the SE16 datasets. All three models are mostly able to correctly determine the hypernym-hyponym relationships for DBpedia samples. Notably, with the Prompt-Combined setting, all three models are capable of making entirely correct judgments on these datasets. In contrast, the SE16 datasets are considerably more challenging. No prompt type yields perfect performance across SE16, and each dataset results in at least 5 incorrect judgments by each model. The SE-Environment dataset is especially difficult, with each model making more than 10 errors. This is primarily because the positive and negative examples in the DBpedia datasets involve more common and widely known concept pairs, such as Person–Astronaut or Athlete, whereas the SE16 datasets include many uncommon, domain-specific, or technical terms—such as defoliation, seismic monitoring, pragmatics, and morphology—making semantic relationship identification more difficult. **(2) Comparing the four prompt types:** Overall, Prompt-Definition performs the best. For DeepSeek-R1 and GPT-4, the two structurally enriched prompts

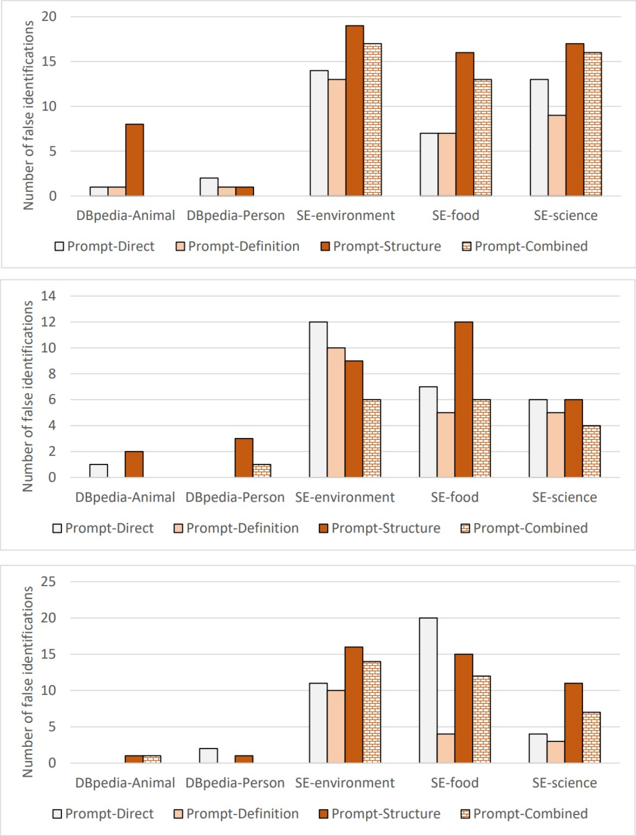


Figure 2: Number of false identifications across different datasets and prompts using DeepSeek-R1, GLM-4 and GPT-4, respectively.

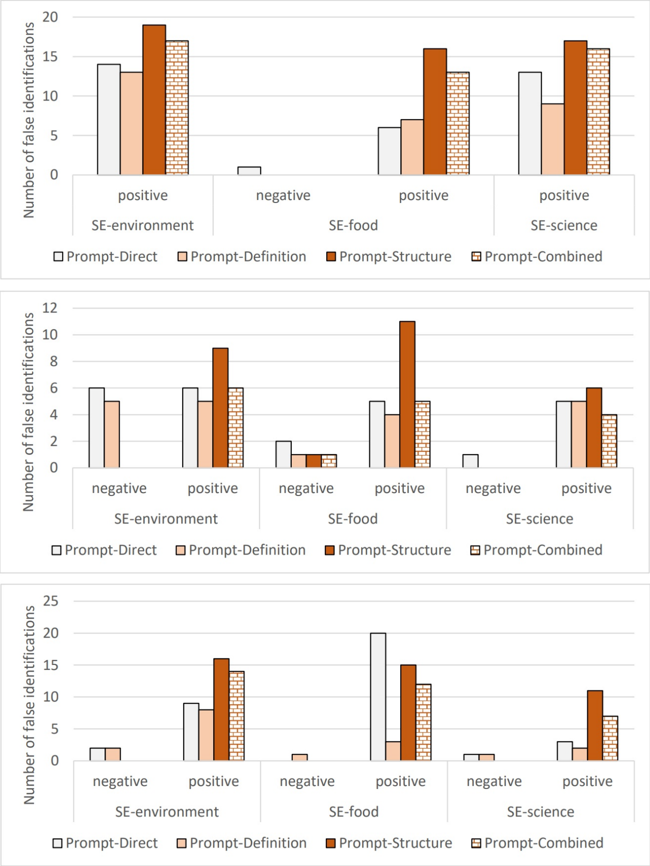


Figure 3: False identifications across example types and prompt types on SE16 datasets using DeepSeek-R1, GLM-4, and GPT-4.

do not provide any improvements in hypernym recognition. However, for GLM-4 on the SE-Environment dataset, these structured prompts lead to clear performance gains. For instance, the simplest prompt, Prompt-Direct, results in 12 incorrect identifications, whereas the two structured prompts reduce the number of errors to 6 and 8, respectively.

Second, we compare the number of false identifications across example types and prompt types on SE16 datasets using DeepSeek-R1, GLM-4, and GPT-4 (see Figure 3). As in Figure 2, bars are omitted when all positive or negative examples are correctly identified by a model. From the figure, it is evident that negative examples are generally easier to be identified correctly than positive ones. All three models are able to correctly classify nearly all negative examples across the SE16 datasets, except for a few errors made by GLM-4 on the SE-environment dataset. Moreover, for negative examples, the two prompt types that incorporate structural information nearly achieve 100% accuracy. This can be largely attributed to the conceptual complexity of the SE datasets: determining that two concepts do not have a hypernym-hyponym relationship is often much easier than determining that they do, especially when domain-specific or uncommon terms are involved.

Finally, we compare identification accuracy across all datasets and the four prompt types (see Figure 4). Several trends emerge from the results: GLM-4 consistently outperforms the other models, achieving high accuracy across all prompt types, with the Prompt-Combined configuration yielding the best overall performance. GPT-4 also demonstrates strong results, particularly with Prompt-Definition, which delivers the highest accuracy among its prompt variants. DeepSeek-R1 performs moderately well, with Prompt-Definition again being the most effective, though its accuracy drops noticeably under the Prompt-Structure condition. In contrast, GPT-3.5 exhibits the weakest performance, especially when using Prompt-Structure, indicating limited ability to interpret structured prompts. Overall, Prompt-Definition tends to deliver robust performance across models, while Prompt-Structure proves to be the least effective—particularly for GPT-3.5 and DeepSeek-R1. These findings suggest that clear definitional information is more beneficial than structured or combined prompts for identifying hypernym-hyponym relationships. In other words, structural information should be incorporated selectively when supporting such relational identification.

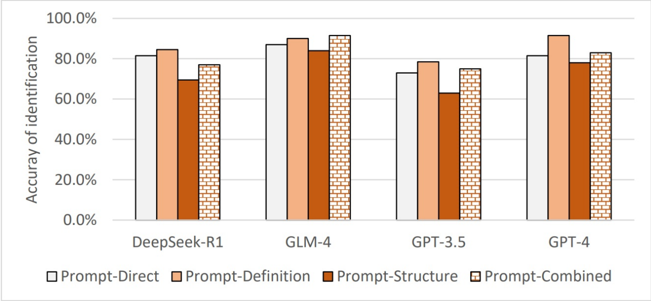


Figure 4: Accuracy of identification across all datasets and prompts.

4.3 Performance Comparison of Different LLMs

To further evaluate the performance of more LLMs, we selected 40 challenging cases from the constructed positive and negative examples based on the experimental results presented in the previous section. A challenging case refers to a positive or negative example that at least one LLM failed to correctly identify. The specific selection details are shown in Table 2. For the comparison, we selected 12 LLMs as introduced in Section 2.3. To better highlight the differences among models, we limited the evaluation to the Prompt-Direct and Prompt-Combined settings.

Dataset	Negative examples	Positive Examples
DBpedia-Animal	0	9
DBpedia-Person	0	10
SE-environment	6	5
SE-food	3	2
SE-science	2	3

Table 2: Number of selected challenging cases.

Figure 5 presents the number of false identifications made by 12 LLMs on the 40 challenging cases under two prompting strategies: Prompt-Direct and Prompt-Combined. From the figure, we can draw the following observations:

- Prompt-Combined generally results in fewer errors on negative examples across most models, indicating that combining prompts enhances the ability of LLMs to correctly reject invalid hypernym-hyponym relations. For example, DeepSeek-V3 shows a significant reduction in errors, from 9 under Prompt-Direct to just 1 under Prompt-Combined.
- In contrast, Prompt-Combined does not consistently improve performance on positive examples. With the exception of DeepSeek-R1, GLM-4, GPT-3.5, and Qwen-2.5, most models made more errors under Prompt-Combined than under Prompt-Direct. For instance, DeepSeek-V3 produced only 2 errors under Prompt-Direct but 11 under Prompt-Combined.
- Unlike most other models, GLM-4 performed worst on negative examples but best on positive ones. It produced the fewest errors on positive examples—two and one under Prompt-Direct and Prompt-Combined, respectively—yet made the highest number of errors on negative examples.
- Among all the models, Gemini-2.5, GPT-4, GPT-4o, and Doubao-1.5 exhibited relatively stable performance, producing a moderate and balanced number of errors across both positive and negative cases.

LLMs	Number of false idenfications	Number of correct idenfications	Accuracy
Baichuan-4	17	63	78.75%
DeepSeek-R1	15	65	81.25%
DeepSeek-V3	16	64	80.00%
GPT-4	12	68	**85.00**%
GPT-4o	10	70	**87.50**%
LLaMA-3.3	13	67	83.75%
Doubao-1.5	13	67	83.75%
GPT-3.5	32	48	60.00%
GLM-4	13	67	83.75%
Gemini-2.5	11	69	**86.25**%
Hunyuan	13	67	83.75%
Qwen-2.5	20	60	75.00%

Table 3: Performance of LLMs over the challenging cases without distinguishing between positive and negative examples or different prompting strategies.

To evaluate the overall performance of the selected LLMs, we present aggregated experimental results—without differentiating between positive and negative examples or varying prompt strategies—based on the number of correct and incorrect identifications, as well as overall accuracy (see Table 3). Among all evaluated models, GPT-4o achieved the highest accuracy (87.50%), closely followed by Gemini-2.5 (86.25%) and GPT-4 (85.00%), demonstrating strong and consistent performance in recognizing hypernym-hyponym relations. Models such as LLaMA-3.3, Doubao-1.5, GLM-4, and Hunyuan also performed well, each achieving an accuracy of 83.75%, reflecting their robustness in this task. In contrast, GPT-3.5 exhibited the weakest performance, with a significantly lower accuracy of 60.00%, suggesting its limitations in handling fine-grained semantic structures. Overall, newer and more advanced models consistently outperform earlier or less capable ones, underscoring the importance of architectural improvements and training data scale in relation extraction tasks.

5 RELATED WORK

Understanding hypernym-hyponym relations is essential for building taxonomies, structuring knowledge, and supporting conceptual reasoning. While this topic has a long history in traditional NLP, recent work has started to explore whether LLMs can internalize and reason about such hierarchical relationships.

[15] examined LLMs' ability to capture conceptual hierarchies through concept classification tasks. Their results suggest that LLMs exhibit a partial understanding of hierarchical relations, particularly when the class structure is implicitly embedded in the

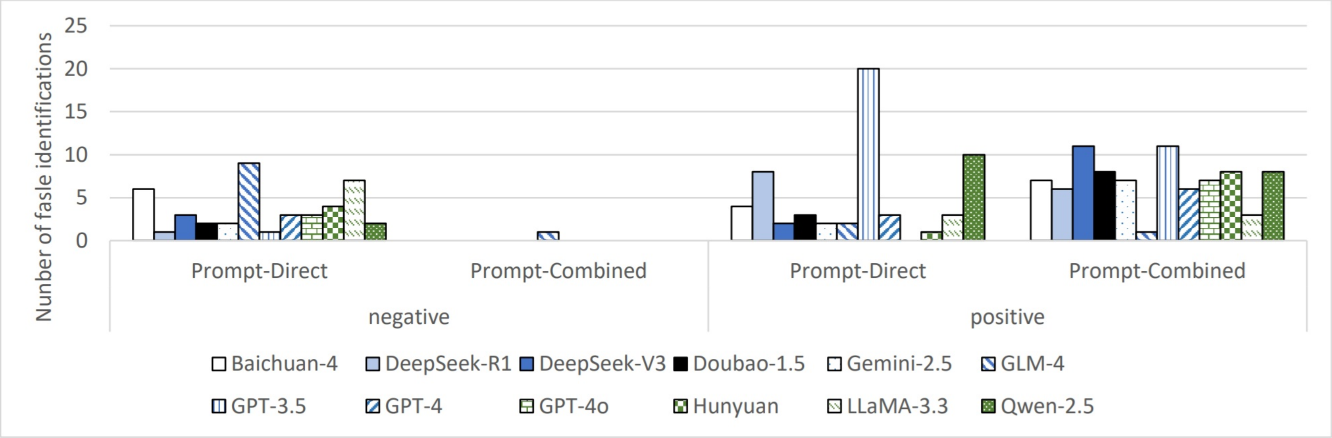


Figure 5: False identifications made by large language models across various example and prompt types.

training data. However, their evaluation was limited to recognizing coarse-grained category labels rather than explicitly reasoning about hypernym-hyponym links.

[11] investigated the role of input representation in guiding LLMs' interpretation of structured knowledge, including hierarchical relations. By comparing linearized triples to fluent text, they showed that input formatting can significantly influence the model's ability to capture taxonomic structure—highlighting the importance of prompt design in eliciting hypernymic reasoning.

[10] compared prompting and fine-tuning strategies for LLM-based taxonomy construction. Their findings indicate that prompting can effectively induce hierarchical structures, especially in low-resource settings. However, their focus was on generating taxonomies rather than evaluating whether models actually understand the underlying hypernymy.

[49] proposed Chain-of-Layer, an iterative prompting framework for building taxonomies with LLMs. By constructing hierarchies layer by layer and using ensemble filtering to reduce hallucination, they demonstrated improved generation quality. Yet, the evaluation focused on output structure, not on the reasoning process behind hypernym-hyponym decisions.

[41] introduced TaxoGlimpse, a benchmark testing LLMs across 10 taxonomies. They observed that while models perform well in general domains, their performance degrades on fine-grained or domain-specific concepts, raising questions about their robustness in identifying specific hypernymic links.

Wang et al. [45] evaluated LLMs' reasoning ability over description logic ontologies, with a focus on formal constructs such as concept inclusion, disjointness, and role restrictions. Their study included tests on whether models can infer hypernym-hyponym and disjoint relations under logical semantics. While their work highlights the potential of LLMs in structured reasoning tasks, it is grounded in synthetic DL-based benchmarks and emphasizes symbolic compositionality rather than naturalistic taxonomic understanding. In contrast, our work assesses hypernym-hyponym recognition in a realistic, non-formalized setting using data derived from encyclopedic and lexical taxonomies.

In contrast to these prior works, our approach does not aim to induce or classify taxonomies, but rather to systematically evaluate whether LLMs can correctly identify and reason about hypernym-hyponym relationships. We construct labeled test cases from real-world hierarchical datasets and design multiple prompt strategies that probe models from linguistic, definitional, and structural perspectives. Our evaluation isolates the reasoning component and highlights model strengths and weaknesses in understanding conceptual hierarchies—offering a targeted diagnostic framework rather than a generative or classificatory one.

6 CONCLUSIONS AND FUTURE WORK

In this paper, we introduced LLM-Hype, a targeted evaluation framework for systematically assessing the ability of LLMs to infer hypernym-hyponym relationships. By constructing taxonomy graphs from diverse data sources and generating carefully labeled test cases, our framework enables controlled and fine-grained evaluation of conceptual hierarchy understanding in LLMs. We further proposed four complementary prompting strategies to probe different dimensions of model reasoning. Through extensive experiments across multiple datasets and 12 representative LLMs, we found that definition-based prompts consistently led to the highest accuracy, highlighting the importance of explicit semantic context. In contrast, structure-based prompts offered mixed results, and in some cases, even degraded performance. Moreover, we observed notable differences in model behavior, with some LLMs like GPT-4o and Gemini-2.5 showing robust performance, while others such as GLM-4 exhibited divergent reasoning patterns. These findings provide valuable insights into the strengths and limitations of current LLMs in taxonomic reasoning.

In future work, we plan to extend LLM-Hype in several directions. First, we aim to enrich the taxonomy graphs with multilingual and domain-specific resources to evaluate model generalization across languages and subject areas. Second, we intend to incorporate more fine-grained reasoning categories—such as transitivity, asymmetry, and negation—to better characterize the nature of model errors. Third, we will explore automated prompt optimization and dynamic prompting strategies to further enhance the diagnostic power of the framework. Finally, integrating human-in-the-loop evaluations may provide additional interpretability and complement automated metrics.

ACKNOWLEDGMENTS

This work was supported by the National Social Science Foundation Key Program of China (Number 23&ZD222). Additional funding was provided by the Natural Science Foundation of China (Number: 62202240).

REFERENCES

[1] Jean-Baptiste Alayrac, Jeff Donahue, and et. al. Pauline Luc. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (Eds.).

[2] Atheer Algherairy and Moataz Ahmed. 2025. Prompting large language models for user simulation in task-oriented dialogue systems. *Comput. Speech Lang.* 89 (2025), 101697. https://doi.org/10.1016/J.CSL.2024.101697

[3] Alibaba Cloud. 2025. Qwen 2.5: A Party of Foundation Models! https://www.alibabacloud.com/blog/qwen2-5-a-party-of-foundation-models_601782. Accessed: 2025-05-06.

[4] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. 2007. DBpedia: A Nucleus for a Web of Open Data. In *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007 (Lecture Notes in Computer Science)*, Vol. 4825. Springer, 722–735.

[5] Baichuan Intelligence. 2024. Baichuan 4: Comprehensive AI Model Release. https://xueqiu.com/9919963656/291029762. Accessed: 2025-05-06.

[6] Georgeta Bordea, Els Lefever, and Paul Buitelaar. 2016. SemEval-2016 Task 13: Taxonomy Extraction Evaluation (TExEval-2). In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*, Steven Bethard, Daniel M. Cer, Marine Carpuat, David Jurgens, Preslav Nakov, and Torsten Zesch (Eds.). The Association for Computer Linguistics, 1081–1091.

[7] Zied Bouraoui, Jose Camacho-Collados, and Steven Schockaert. 2020. Inducing taxonomies from word embeddings: A survey of methods and evaluation strategies. *Computational Linguistics* 46, 4 (2020), 835–876.

[8] Tom B. Brown, Benjamin Mann, Nick Ryder, and et al. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[9] ByteDance. 2025. Doubao 1.5 Pro: Performance Surpassing GPT-4o and Claude3.5Sonnet. https://www.aibase.com/news/14931. Accessed: 2025-05-06.

[10] Boqi Chen, Fandi Yi, and Dániel Varró. 2023. Prompting or Fine-tuning? A Comparative Study of Large Language Models for Taxonomy Construction. In *ACM/IEEE International Conference on Model Driven Engineering Languages and Systems, MODELS 2023 Companion, Västerås, Sweden, October 1-6, 2023.* IEEE, 588–596. https://doi.org/10.1109/MODELS-C59198.2023.00097

[11] Xinbang Dai, Yuncheng Hua, Tongtong Wu, Yang Sheng, Qiu Ji, and Guilin Qi. 2025. Large language models can better understand knowledge graphs than we thought. *Knowl. Based Syst.* 312 (2025), 113060. https://doi.org/10.1016/J.KNOSYS.2025.113060

[12] DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. https://arxiv.org/abs/2501.12948. Accessed: 2025-05-06.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186.

[14] Chongren Feng, Jiwei Qin, and Yuhang Zhang. 2024. EP-BoxE: A method for hypernym discovery based on extended patterns and box embeddings. *J. Intell. Fuzzy Syst.* 46, 3 (2024), 5801–5810. https://doi.org/10.3233/JIFS-235181

[15] Chao Feng, Xinyu Zhang, and Zichu Fei. 2023. Knowledge Solver: Teaching LLMs to Search for Domain Knowledge from Knowledge Graphs. *CoRR* abs/2309.03118 (2023). https://doi.org/10.48550/ARXIV.2309.03118 arXiv:2309.03118

[16] Sohom Ghosh, Ankush Chopra, and Sudip Kumar Naskar. 2023. Learning to Rank Hypernyms of Financial Terms Using Semantic Textual Similarity. *SN Comput. Sci.* 4, 5 (2023), 610. https://doi.org/10.1007/S42979-023-02134-Z

[17] Hamed Babaei Giglou, Janice D'Souza, and Sören Auer. 2023. LLMs4OL: Large Language Models for Ontology Learning. In *Proceedings of the International Semantic Web Conference (ISWC)*. Springer Nature Switzerland, Cham, 408–427.

[18] Google. 2025. Gemini 2.5: Our newest Gemini model with thinking. https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/. Accessed: 2025-05-06.

[19] Nicola Guarino. 1998. Formal ontology and information systems. *Proceedings of FOIS* 98, 1998 (1998), 81–97.

[20] Seyed Mohammad Hossein Hashemi, Mostafa Karimi Manesh, and Mehrnoush Shamsfard. 2024. SKH-NLP at LLMs4OL 2024 Task B: Taxonomy Discovery in Ontologies Using BERT and LLaMA 3. In *LLMs4OL 2024: The 1st Large Language Models for Ontology Learning Challenge at the 23rd ISWC,* Co-located with the 23rd International Semantic Web Conference (ISWC 2024)*, Baltimore, Maryland, USA, November 11-15, 2024 (Open Conference Proceedings)*, Hamed Babaei Giglou, Jennifer D'Souza, and Sören Auer (Eds.), Vol. 4. TIB Open Publishing, 103–111. https://doi.org/10.52825/OCP.V4I.2483

[21] Marti A. Hearst. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. In *14th International Conference on Computational Linguistics, COLING 1992, Nantes, France, August 23-28, 1992.* 539–545. https://aclanthology.org/C92-2082/

[22] Pascal Hitzler, Markus Krötzsch, Bijan Parsia, Peter F. Patel-Schneider, and Sebastian Rudolph. 2009. OWL 2 Web Ontology Language Primer. W3C Recommendation. Available at: https://www.w3.org/TR/owl2-primer/.

[23] Pascal Hitzler, Markus Krötzsch, and Sebastian Rudolph. 2010. *Foundations of Semantic Web Technologies*. Chapman and Hall/CRC Press. http://www.semantic-web-book.org/

[24] Rasha Ahmad Husein, Hala Aburajouh, and Cagatay Catal. 2025. Large language models for code completion: A systematic literature review. *Comput. Stand. Interfaces* 92 (2025), 103917. https://doi.org/10.1016/J.CSI.2024.103917

[25] Xun Jian, Yue Wang, Xiayu Lei, Libin Zheng, and Lei Chen. 2020. SPARQL Rewriting: Towards Desired Results. In *Proceedings of the 2020 International Conference on Management of Data, SIGMOD Conference 2020, online conference [Portland, OR, USA], June 14-19, 2020.* ACM, 1979–1993.

[26] Song Jiang, Qiyue Yao, Qifan Wang, and Yizhou Sun. 2024. A Single Vector Is Not Enough: Taxonomy Expansion via Box Embeddings (Extended Abstract). In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024.* ijcai.org, 8421–8426. https://www.ijcai.org/proceedings/2024/934

[27] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large Language Models are Zero-Shot Reasoners. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (Eds.). http://papers.nips.cc/paper_files/paper/2022/hash/8bb0d291acd4acf06ef112099c16f326-Abstract-Conference.html

[28] Philip Lippmann, Konrad Skublicki, Joshua B. Tanner, Shonosuke Ishiwatari, and Jie Yang. 2025. Context-Informed Machine Translation of Manga using Multimodal Large Language Models. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (Eds.). Association for Computational Linguistics, 3444–3464.

[29] Ningchen Ma, Dong Wang, Hongyun Bao, Lei He, and Suncong Zheng. 2023. KEPL: Knowledge Enhanced Prompt Learning for Chinese Hypernym-Hyponym Extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, 5858–5867. https://doi.org/10.18653/V1/2023.EMNLP-MAIN.358

[30] Meta Platforms, Inc. 2024. Llama 3.3: 70B Multilingual Instruction-Tuned Model. https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct. Accessed: 2025-05-06.

[31] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.).

[32] George A. Miller. 1995. WordNet: A Lexical Database for English. *Commun. ACM* 38, 11 (1995), 39–41. https://doi.org/10.1145/219717.219748

[33] Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a Very Large Multilingual Semantic Network. In *ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden*, Jan Hajic, Sandra Carberry, and Stephen Clark (Eds.). The Association for Computer Linguistics, 216–225. https://aclanthology.org/P10-1023/

[34] OpenAI. 2023. GPT-4 Technical Report. *CoRR* abs/2303.08774 (2023). https://doi.org/10.48550/ARXIV.2303.08774 arXiv:2303.08774

[35] OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774* (2023).

[36] Long Ouyang, Jeffrey Wu, and et. al. Xu Jiang. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (Eds.).

[37] Michael-Andrei Panaitescu-Liess, Zora Che, Bang An, Yuancheng Xu, Pankayaraj Pathmanathan, Souradip Chakraborty, Sicheng Zhu, Tom Goldstein, and Furong Huang. 2025. Can Watermarking Large Language Models Prevent Copyrighted Text Generation and Hide Training Data?. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, Toby Walsh, Julie Shah, and Zico Kolter (Eds.). AAAI Press, 25002–25009.

[38] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, Alessandro Moschitti, Bo Pang, and Walter Daelemans (Eds.). ACL, 1532–1543. https://doi.org/10.3115/V1/D14-1162

[39] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, Marilyn A. Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, 2227–2237.

[40] Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2004. Learning Syntactic Patterns for Automatic Hypernym Discovery. In *Advances in Neural Information Processing Systems 17 [Neural Information Processing Systems, NIPS 2004, December 13-18, 2004, Vancouver, British Columbia, Canada]*. 1297–1304. https://proceedings.neurips.cc/paper/2004/hash/358aee4cc897452c00244351e4d91f69-Abstract.html

[41] Yushi Sun, Xin Hao, Kai Sun, Yifan Xu, Xiao Yang, Xin Luna Dong, Nan Tang, and Lei Chen. 2024. Are Large Language Models a Good Replacement of Taxonomies? *Proc. VLDB Endow.* 17, 11 (2024), 2919–2932. https://doi.org/10.14778/3681954.3681973

[42] Tencent. 2025. Tencent's New Hunyuan Turbo S AI Model Outpaces DeepSeek R1. https://opentools.ai/news/tencents-new-hunyuan-turbo-s-ai-model-outpaces-deepseek-r1. Accessed: 2025-05-06.

[43] Hugo Touvron, Thibaut Lavril, Gautier Izacard, and et al. 2023. LLaMA: Open and Efficient Foundation Language Models. *CoRR* abs/2302.13971 (2023). https://doi.org/10.48550/ARXIV.2302.13971 arXiv:2302.13971

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5998–6008.

[45] Keyu Wang, Guilin Qi, Jiaqi Li, and Songlin Zhai. 2024. Can Large Language Models Understand DL-Lite Ontologies? An Empirical Study. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, 2503–2519. https://aclanthology.org/2024.findings-emnlp.141

[46] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (Eds.). http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html

[47] Yishi Xu, Dongsheng Wang, Bo Chen, Ruiying Lu, Zhibin Duan, and Mingyuan Zhou. 2022. HyperMiner: Topic Taxonomy Mining with Hyperbolic Embedding. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (Eds.). http://papers.nips.cc/paper_files/paper/2022/hash/cd004fa45fc1fe5c0218b7801d98d036-Abstract-Conference.html

[48] Geonil Yun, Yongjae Lee, A-Seong Moon, and Jaesung Lee. 2023. Hypert: hypernymy-aware BERT with Hearst pattern exploitation for hypernym discovery. *J. Big Data* 10, 1 (2023), 141. https://doi.org/10.1186/S40537-023-00818-0

[49] Qingkai Zeng, Yuyang Bai, Zhaoxuan Tan, Shangbin Feng, Zhenwen Liang, Zhihan Zhang, and Meng Jiang. 2024. Chain-of-Layer: Iteratively Prompting Large Language Models for Taxonomy Induction from Limited Examples. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM 2024, Boise, ID, USA, October 21-25, 2024*, Edoardo Serra and Francesca Spezzano (Eds.). ACM, 3093–3102. https://doi.org/10.1145/3627673.3679608

[50] Zhipu AI. 2024. GLM-4: Next-Generation Foundation Model Released at Zhipu DevDay. https://zhipuai.cn/devday. Accessed: 2025-05-06.