Are Large Language Models a Good Replacement of Taxonomies? [Experiment, Analysis & Benchmark]

Yushi Sun
HKUST
ysunbp@cse.ust.hk

Hao Xin
HKUST
hxinaa@cse.ust.hk

Kai Sun
Meta Reality Labs
sunkaicn@meta.com

Yifan Ethan Xu
Meta Reality Labs
ethanxu@meta.com

Xiao Yang
Meta Reality Labs
xiaoyangfb@meta.com

Xin Luna Dong
Meta Reality Labs
lunadong@meta.com

Nan Tang
HKUST (GZ) / HKUST
nantang@hkust-gz.edu.cn

Lei Chen
HKUST(GZ) / HKUST
leichen@cse.ust.hk

ABSTRACT

Large language models (LLMs) demonstrate an impressive ability to internalize knowledge and answer natural language questions. Although previous studies validate that LLMs perform well on general knowledge while presenting poor performance on long-tail nuanced knowledge, the community is still doubtful about whether the traditional knowledge graphs should be replaced by LLMs. In this paper, we ask *if the schema of knowledge graph (i.e., taxonomy) is made obsolete by LLMs*. Intuitively, LLMs should perform well on common taxonomies and at taxonomy levels that are common to people. Unfortunately, there lacks a comprehensive benchmark that evaluates the LLMs over a wide range of taxonomies from common to specialized domains and at levels from root to leaf so that we can draw a confident conclusion. To narrow the research gap, we constructed a novel taxonomy hierarchical structure discovery benchmark named TaxoGlimpse[1] to evaluate the performance of LLMs over taxonomies. TaxoGlimpse covers ten representative taxonomies from common to specialized domains with in-depth experiments of different levels of entities in this taxonomy from root to leaf. Our comprehensive experiments of eighteen state-of-the-art LLMs under three prompting settings validate that LLMs can still not well capture the knowledge of specialized taxonomies and leaf-level entities.

1 INTRODUCTION

Recently, we have witnessed the rapid advancements of large language models (LLMs) such as GPTs [21] and Llamas [72]. These LLMs have demonstrated impressive abilities in a wide range of applications such as question answering [71], information retrieval [78], news summarization [76], entity relation extraction [73], among many others, disrupting and redefining the development of these areas. However, several previous studies also pointed out that LLMs are significantly less knowledgeable in domain-specific long-tail knowledge details [25, 70], sparking a growing debate about whether traditional knowledge graphs will be replaced by LLMs in real applications [30, 70, 71].

As the schema of knowledge graphs, taxonomies provide a structured way to organize and categorize knowledge, which is indeed a kind of "knowledge about knowledge" (or meta-knowledge), serving as an important asset in different applications such as knowledge/information management [68], data integration [61], knowledge extraction [47], and domain-specific recommendation [42]. The general form of taxonomies involves a hierarchical structure that organizes entities and concepts into categories based on their characteristics or relationships. Typically, taxonomies follow a tree-like structure, where each category is represented as a node, and the relationships between categories are depicted as hypernymy (Is-A) links. The top-level category represents the broadest classification, while the lower-level categories become more specific.

Traditionally, taxonomies are used to assist entity searching (e.g., "best health tracker" as a shopping query), category display, and knowledge reasoning, which rely on the basic "Is-A" relation constructed between parent and child entities. Recently, LLMs have shown the ability to learn world knowledge, which opens up an opportunity to also store the "Is-A" taxonomy structure in LLMs' parameters. This raises an important question: Are traditional hierarchical structures in taxonomies made obsolete by LLMs [24, 59, 70]?

Example 1. [**Taxonomies.**] *Figure 1 shows taxonomy snippets from common (at the top) to specialized (at the bottom) domains ranked by the popularity of taxonomies as illustrated in Figure 2. From left to right, we present the entities from root to leaf levels.*

[**From Common to Specialized Domains.**] *To get an understanding of LLMs' knowledgeability in determining the "Is-A" relations in taxonomies, we prompt GPT-4 [21] model with: " Is <child entity> a type of <parent entity>?" and record the overall annotation accuracy of GPT-4 [21] on each taxonomy. Specifically, GPT-4 achieves* 85.7% *accuracy on Google taxonomy, while achieving* 70.8% *and* 62.6% *accuracies on ACM-CCS and Glottolog taxonomies respectively, which is consistent with the intuition that GPT-4 performs better on common domains while exhibiting poorer performance on specialized domains.*

[**From Root to Leaf Levels.**] *We further query each level of the exemplar chain of Glottolog taxonomy. The queries are again provided in a child-to-parent manner: e.g., Is Sinitic language a type of Sino-Tibetan language? We observe that GPT-4 gave incorrect answers at the Hailu to Hakka-Chinese and the Hakka-Chinese to Middle-Modern-Sinitic levels, while correctly answering the rest, which means it tends to be more knowledgeable near the root levels while becoming less reliable near the leaf levels of Glottolog.* □

Example 1 shows that LLMs' knowledgeability in taxonomies varies based on multiple factors such as the popularity of the taxonomies or the depth at which a question is posed. Motivated by

[1]All datasets collection were done by HKUST. The source code, data, and/or other artifacts have been made available at https://github.com/ysunbp/TaxoGlimpse.

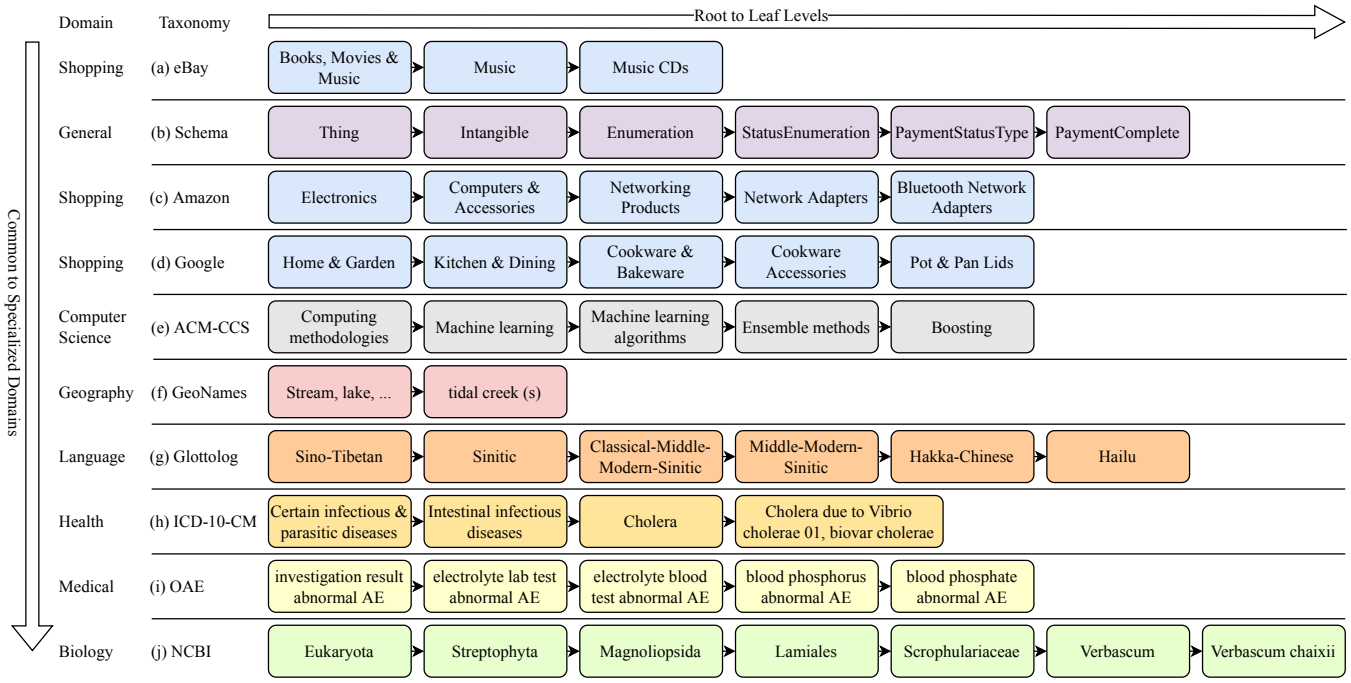


Figure 1: Exemplar chain of entities snippets of ten taxonomies. From top to bottom, we list the taxonomy snippets from common domains to specialized domains. From left to right, we present entities from the root to leaf levels.

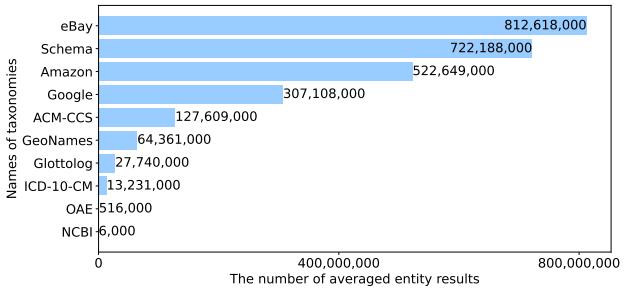


Figure 2: The popularity of different taxonomies.

the observations, we conduct a systematic study to address a crucial question: **Can LLMs Effectively Replace Taxonomies?**

The importance of the study is three-fold: (1) **Industrial users** can understand if constructing and maintaining traditional taxonomies is worth investing in; (2) **LLM developers** can learn about the pros and cons of their models in taxonomies and improve accordingly to help users better perform taxonomy-related tasks with LLMs; and (3) **Database researchers** can innovate on the novel forms of taxonomy structures, and explore interesting and meaningful research problems/application domains that may boost the reasoning of LLMs.

Challenges. Addressing the question faces three challenges:

(C1) The Absence of a Comprehensive Benchmark. To our knowledge, there is no comprehensive benchmark that can effectively answer the key research question. Such a benchmark should encompass a wide variety of taxonomies taking into account diverse characteristics such as popularity, domains, and complexity.

(C2) Formulating an Evaluation Strategy. Taxonomies possess a unique hierarchical structure, which presents a challenge when designing probing strategies to thoroughly evaluate the knowledgeability of LLMs from root to leaf levels.

(C3) Diversity of LLMs. The field of LLMs is rapidly evolving, with a multitude of base models available and an even greater number of configurations such as model sizes and prompting methods. It is critical to establish a systematic approach for a thorough study.

Contributions. We make the following notable contributions.

(1) TaxoGlimpse: A New Benchmark. In response to challenges (C1) and (C2), we constructed a benchmark, namely TaxoGlimpse, which covers taxonomies from eight different domains ranging from common to specialized domains. The taxonomies selected for each domain are representative, and have a range of different numbers of entities, levels, and trees. Moreover, we designed the questions on each level of different taxonomies, enabling users to have an in-depth glimpse into the knowledge of LLMs from root to leaf of the taxonomies.

(2) A Comprehensive Evaluation. To resolve challenge (C3), we evaluate eighteen state-of-the-art LLMs, including GPTs, Claude-3, Llama-2 (7B, 13B, 70B), Llama-3 (8B, 70B), Mistral (7B), Mixtral (8*7B), and so on, to have decent coverage of state-of-the-art LLMs and their variants. In addition, we consider common prompt engineering settings, such as zero-shot, few-shot, and Chain-of-Thoughts, to systematically evaluate the knowledge of LLMs.

(3) Key Research Questions. By designing the new benchmark and conducting comprehensive experiments, we answer the following questions.

(i) *How reliable are LLMs for determining hierarchical structures in different taxonomies?* LLMs perform well on common taxonomies (e.g., Shopping); however, their performance downgrades significantly on specialized taxonomies (e.g., Biology).

(ii) *Do LLMs perform equally well among different levels of taxonomies?* LLMs roughly achieve progressively worse performance from root to leaf in most taxonomies.

(iii) *Do normal methods that improve LLMs increase the reliability?* The increase in sizes and the adoption of domain-agnostic fine-tuning of LLMs may not lead to an increase in performance. The domain-specific instruction tuning leads to stable and significant performance improvements.

(iv) *Do different prompting settings influence the performance?* The performance changes of best LLMs brought by few-shot and Chain-of-Thoughts prompting settings are minimal.

(4) Future Opportunities. For practitioners, we recommend a) continuing manual taxonomy construction for specialized domains such as Language, and near the root levels of taxonomies for the purposes of displaying; b) leveraging LLMs for taxonomy-based search and reasoning in common domains such as shopping, to save manual work in constructing lower level of taxonomies. For researchers, we suggest exploring the development of taxonomies in an LLM-tree-structure-combined form, where entities reside implicitly as LLM embeddings or exist explicitly in link forms.

Moreover, we publish our code and datasets on GitHub [20] to attract more research in this direction.

2 BENCHMARK CONSTRUCTION AND QUESTION DESIGN

2.1 Benchmark Construction

We selected taxonomies from eight domains to cover a wide range of taxonomic knowledge. When selecting the taxonomies, we considered the following criteria:

(1) The taxonomies are publicly accessible;

(2) The taxonomies cover different domains with different popularity to ensure a comprehensive view of LLMs' performance from common to specialized domains;

(3) The characteristics (number of entities, number of levels, number of trees) of taxonomies are diverse to ensure representativity; and

(4) The taxonomies are representative in each domain and widely used by their respective communities.

With these criteria in mind, we selected taxonomies from eight domains: Shopping (Google Product Category [4], Amazon Product Category [3], and eBay Categories [15]), General (Schema.org [36]), Computer Science Research (ACM Computing Classification System [2]), Geography (GeoNames [16]), Language (Glottolog [9]), Health (ICD-10-CM [11]), Medical (OAE [40]), and Biology (The NCBI Taxonomy Database [12]).

Figure 2 depicts the popularity of the selected taxonomies, measured by the average number of results returned by google.com by searching (exact match) the names of 100 randomly sampled concepts from each taxonomy. Specifically, eBay, Schema.org, Amazon, and Google taxonomies are the representatives of the common taxonomies that cover common entities known to ordinary people. ACM-CCS, GeoNames, Glottolog, ICD-10-CM, OAE, and NCBI are the specialized taxonomies that are domain-specific and likely to be accessed by domain experts. Apart from the popularity of taxonomies, we also considered the characteristics of different taxonomies. As shown in Table 1, the characteristics we selected give the audience the basic information about the scale of the taxonomies (number of entities), the depth of the taxonomies (number of levels), the width of the taxonomies (number of trees), the rough shape of the taxonomies (number of nodes and classes in each level). The selected taxonomies have a range of cover of different numbers of entities (from 500 to 2M), number of levels (from 2 to 7), and number of tree roots (from 3 to 245), representing well the diverse distribution of the morphology of taxonomies in different application domains and scenarios. We now discuss the data collection details on each domain.

Shopping Taxonomies. We selected Google Product Category [4], Amazon Product Category [3], and eBay Categories [15], which are the representative taxonomies in the shopping domains: Google Product Category is used by Google Shopping, which is the most widely used for product price comparison in the United States according to [6]. Amazon Product Category is from Amazon.com, which is the most visited e-commerce shopping website in the United States [8]. The eBay Categories is from ebay.com, which is another popular online shopping platform.

Despite that all these taxonomies target the shopping domain, they have significant differences in size and organization of categorization. As shown in Table 1, the Amazon Product Category is larger in the number of entities and the size of top-level classifications. As a result, the Amazon Product Category provides a finer-grained classification of products. By evaluating LLMs on the three shopping taxonomies, we can gain a comprehensive view of LLMs' performance in the shopping domain.

We collected the Google Product Category (US version) from the official link provided by Google [4] and crawled the Amazon Product Category and eBay Categories from [3] and [15], respectively. In order to gain a holistic view of LLMs performance in different levels of the taxonomies, we pre-processed and divided the entities into five levels for the Google and Amazon taxonomies: root level, level 1, level 2, level 3, and level 4 or lower for the Google and Amazon taxonomies and three levels for the eBay taxonomy. We present the detailed statistics of the three taxonomies in Table 1 and exemplar snippets of the three taxonomies in Figure 1(a),(c),(d).

General Taxonomies. We adopted the Schema.org taxonomy [36] as a representative for the general domain, which is a community effort to develop schemas for the structured data from the internet. Schema.org is a general domain taxonomy covering a wide range of concepts on the internet and serves as the basis for other general-purpose knowledge bases such as YAGO [69].

As shown in Table 1, the Schema.org taxonomy contains six levels with a total number of 1346 entities, covering coarse concepts such as Thing to fine-grained concepts such as PaymentComplete. We used the newest release v26.0 of Schema.org from the official link [19].

Table 1: Statistics of taxonomies.

Domain	Taxonomy	# of entities	# of levels	# of trees	# of nodes and classes in each level
Shopping	eBay	595	3	13	13-110-472
	Amazon	43814	5	41	41-507-3910-13579-25777
	Google	5595	5	21	21-192-1349-2203-1830
General	Schema	1346	6	3	3-17-215-403-436-272
Computer Science	ACM-CCS	2113	5	13	13-84-543-1087-386
Geography	GeoNames	689	2	9	9-680
Language	Glottolog	11969	6	245	245-712-1048-1205-1366-7393
Health	ICD-10-CM	4523	4	22	22-155-963-3383
Medical	OAE	9547	5	181	181-1854-3817-2587-1108
Biology	NCBI	2190125	7	53	53-309-514-1859-10215-107615-2069560

Computer Science Research Taxonomies. For the computer science research domain, we selected the ACM Computing Classification System (ACM CCS) [2], which is the standard classification system for papers in the computer science field. The CCS concept taxonomy is widely used by researchers to accurately categorize their work so that other researchers can easily overview the main topics and quickly refer to related papers.

As shown in Table 1, we considered five levels in the ACM CCS concept taxonomy. We provide an example for the entities in ACM CCS in Figure 1(e). We adopted the ACM CCS concept taxonomy version 2012 through ACM's website [1].

Geography Taxonomies. We selected the GeoNames taxonomy [16] for the geography domain. The GeoNames taxonomy is representative of this domain covering a two-level classification of the common geographical concepts.

As shown in Table 1, the GeoNames taxonomy contains two levels with 689 concepts. We downloaded the GeoNames taxonomy from the official data release website [16].

Language Taxonomies. We chose Glottolog taxonomy [37, 38, 60] to represent the language domain, which is widely used by linguists. Glottolog offers an extensive inventory of languages, language families, and dialects found across the globe, that linguists need to be able to identify [28].

As shown in Table 1, we considered six levels in the Glottolog taxonomy with a total number of 11,969 languoid entities. The six levels of Glottolog cover a taxonomic structure from language family (e.g., Sino-Tibetan in Figure 1(g)) to a specific language (e.g., Hakka-Chinese) or dialect (e.g., Hailu). We adopted the release v4.8 of Glottolog from [9].

Health Taxonomies. We selected the ICD-10-CM taxonomy [11] for evaluation. ICD-10-CM is a representative candidate for the health domain designed by the Centers for Disease Control and Prevention of the US.

As shown in Table 1, ICD-10-CM taxonomy contains four levels. The root to level 3 concepts correspond to the rough classification of diseases based on body system or condition, the detailed classification, common disease group, and disease entities with different causes. Level 3 entities can be considered as the instances in the ICD-10-CM taxonomy. We present a snippet for ICD-10-CM in Figure 1(h) for better understanding. The ICD-10-CM taxonomy is accessed through the simple_icd_10_CM 2.0.1 package [5].

Medical Taxonomies. We selected the OAE taxonomy (Ontology of Adverse Events) [40], which is a taxonomy specialized for adverse events. The OAE taxonomy has been developed to standardize the annotation of adverse events, integrate various adverse event data, and support computer-assisted reasoning.

As shown in Table 1, we considered five levels with a total number of 9547 entities in the OAE taxonomy from the coarse- to fine-grained classifications of the adverse events. We adopted the newest release v1.2.47 of OAE from [7].

Biology Taxonomies. We selected the NCBI Taxonomy Database [64, 66] as a representative in the biology domain. The NCBI taxonomy serves as the primary repository for standard nomenclature and classification within the International Nucleotide Sequence Database Collaboration (INSDC) and encompasses several prominent databases, including GenBank, ENA (EMBL), and DDBJ [35].

Following the instructions provided by [66], we considered seven levels in the taxonomy, aligning with the biological taxonomy order: 1) superkingdom/kingdom/high-level clade, 2) phylum, 3) class, 4) order, 5) family, 6) genus, and 7) species. An example for the seven levels is presented in Figure 1(j). We downloaded the version (Sep 2023) of the NCBI taxonomy at the time when our experiments started from the official website [12].

2.2 Question Design

We first discuss the question templates we designed for each taxonomy, followed by the question generation process for each respective question type.

Question Templates. In order to understand LLMs' ability to discover hierarchical relationships in taxonomies and to take into account the characteristics of different taxonomies, we designed the simple-formed True/False templates and Multiple-Choice Question (MCQ) templates for each domain in Tables 2 and 3.

We observed similar results when using slight paraphrasing of the templates (the slight paraphrasing for the True/False questions replaces the words "a type of" with "a kind of" and "a sort of"; while for the MCQ questions, we replace the word "appropriate" with "suitable" and "proper"), so will report results on these templates only and present the full experimental results on all the template variants in our GitHub repository [20].

Tables 2 and 3 present the detailed templates we used for evaluating the LLMs on True/False and MCQ question types respectively.

Question Generation. The questions were generated concerning the levels of child entities. For each taxonomy, we randomly sample entities from each level of the taxonomy except the root level. The sample sizes were determined based on the number of entities in

Table 2: Question templates (True/False).

Domains	Question Templates
Shopping	Are <child-type> products a type of <parent-type> products? answer with (Yes/No/I don't know)
General	Is <child-type> entity type a type of <parent-type> entity type? answer with (Yes/No/I don't know)
Computer Science	Is <child-type> computer science research concept a type of <parent-type> computer science research concept? answer with (Yes/No/I don't know)
Geography	Is <child-type> geographical concept a type of <parent-type> geographical concept? answer with (Yes/No/I don't know)
Language	Is <child-type> language a type of <parent-type> language? answer with (Yes/No/I don't know)
Health / Biology	Is <child-type> a type of <parent-type>? answer with (Yes/No/I don't know)
Medical	Is <child-type> Adverse Events concept a type of <parent-type> Adverse Events concept? answer with (Yes/No/I don't know)

Table 3: Question templates (MCQ).

Domains	Question Templates
Shopping	What is the most appropriate supertype of <child-type> product? A) B) C) D)
General	What is the most appropriate supertype of <child-type> entity type? A) B) C) D)
Computer Science	What is the most appropriate supertype of <child-type> research concept? A) B) C) D)
Geography	What is the most appropriate supertype of <child-type> geographical concept? A) B) C) D)
Language	What is the most appropriate supertype of <child-type> language? A) B) C) D)
Health / Biology	What is the most appropriate supertype of <child-type>? A) B) C) D)
Medical	What is the most appropriate supertype of <child-type> Adverse Events concept? A) B) C) D)

each level, with a confidence level of 95% and a margin of error of 5% as suggested by [13]. As shown in Table 2, besides <child-type>, we also need to obtain <parent-type> to form valid True/False questions. As for the MCQs in Table 3, we need to obtain four options to form each valid question. For ease of demonstration of relationship inside a taxonomy, we considered the following notations: e_n denotes an entity in level n, ($n = 0, 1, 2, ...$); E_n denotes the set of all entities in level n; $e_n.p$ denotes the entity e_n's intermediate parent entity (direct parent entity); $e_n.s$ denotes the set of sibling entities of e_n. To comprehensively understand LLMs' performance on taxonomic data, we consider the following generation modes:

- **positive:** Directly get the intermediate parent entity $e_n.p$ of the sampled child entity e_n.
- **negative-easy:** Randomly sample a negative parent entity from the set $E_{n-1} - \{e_n.p\}$.
- **negative-hard:** Randomly sample a negative parent entity from the set $(e_n.p).s$ (uncles of the child entity).
- **MCQ:** Randomly sample three negative options from the set $(e_n.p).s$, and preserve the parent entity $e_n.p$ as the ground truth option.

The reason why we generated negative-hard and negative-easy questions is to provide hard negatives and easy negatives. Intuitively, the negative-hard questions tend to be more difficult since the negative samples are siblings of the ground truth parent entity (i.e., uncles), which means these entities are more similar to the ground truth in comparison with randomly sampled negative samples, serving as the hard negatives for the LLMs. The evaluation was conducted in three sets of data for each level of taxonomies: positive + negative-easy, positive + negative-hard, and MCQ, which were denoted as easy, hard, and MCQ datasets respectively. Detailed statistics of the easy, hard, and MCQ datasets at each level of different taxonomies are presented in Table 4.

3 EXPERIMENTAL SETTINGS

In this section, we introduce the LLMs considered in our experiments, the implementation details we adopted, and the metrics.

3.1 Large Language Models

We now introduce the LLM series considered in our experiments. In order to comprehensively evaluate the performance of state-of-the-art LLMs, we selected nine popular LLM series with eighteen models to conduct the experiments.

- **GPTs** [21]: The Generative Pre-trained Transformers series, are advanced language models by OpenAI. We selected GPT-3.5 and GPT-4 as two representatives for evaluation. The models are close-sourced and accessed through API only.
- **Claude-3** [14]: Claude-3 is the newest release of the Claude family models by Anthropic, which is close-sourced and claims to set new benchmarks for multiple cognitive tasks. We experimented with the best variant Claude-3-Opus.
- **Llama-2s** [72]: The Llama-2 series is a set of open-sourced large language models released by Meta. We adopted Llama-2 7B, 13B, and 70B models with chat settings, which are suitable for the question-answering application scenario.
- **Llama-3s** [22]: The Llama-3 series is a novel set of open-sourced large language models released by Meta in April 2024. We adopted Llama-3 8B and 70B models with instruct settings.
- **Flan-T5s** [31]: The Flan-t5s is an encoder-decoder LLM series developed by Google. We selected the best models from the series: Flan-t5-3B and Flan-t5-11B.
- **Falcons** [23]: Developed by TIIUAE, the Falcon series is claimed to achieve comparable performance with Llama-2s in question answering tasks [23]. We chose Falcon-Instruct models with 7B and 40B parameters for our experiments, which are optimized for chat format.
- **Vicunas** [29]: The Vicuna series [29] are the large language models developed based on the weights through domain-agnostic instruction fine-tuning. We include these models to investigate if domain-agnostic fine-tuning improves performance. We adopted Vicunas 7B, 13B, and 33B.
- **Mistrals** [44, 45]: Designed by Mistral AI, the Mistral and Mixtral models claimed to outperform the Llama-2 13B on several reasoning benchmarks. We adopted the Mistral-7B-Instruct and Mixtral-8*7B-Instruct models.
- **LLMs4OL** [24]: LLMs4OL is the state-of-the-art approach that utilizes instruction tuning [32] based on the Flan-T5-3B model to perform ontology learning tasks. Different

Table 4: Statistics of datasets.

		eBay	Amazon	Google	Schema	ACM-CCS	GeoNames	Glottolog	ICD-10-CM	OAE	NCBI
Level 1-root	Easy	176	438	258	34	138	492	500	222	638	344
	Hard	176	438	258	34	138	492	500	222	638	344
	MCQ	88	219	129	17	69	246	250	111	319	172
Level 2-1	Easy	430	700	600	276	450	n/a	564	550	700	440
	Hard	430	700	597	276	450	n/a	564	550	700	439
	MCQ	215	350	300	138	225	n/a	282	275	350	220
Level 3-2	Easy	n/a	748	656	394	568	n/a	584	690	670	638
	Hard	n/a	748	653	394	567	n/a	584	690	670	636
	MCQ	n/a	374	328	197	284	n/a	192	345	335	319
Level 4-3	Easy	n/a	758	636	410	386	n/a	600	n/a	572	742
	Hard	n/a	758	626	410	370	n/a	600	n/a	572	741
	MCQ	n/a	379	318	205	193	n/a	300	n/a	286	371
Level 5-4	Easy	n/a	n/a	n/a	320	n/a	n/a	732	n/a	n/a	766
	Hard	n/a	n/a	n/a	320	n/a	n/a	732	n/a	n/a	766
	MCQ	n/a	n/a	n/a	160	n/a	n/a	366	n/a	n/a	383
Level 6-5	Easy	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	770
	Hard	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	770
	MCQ	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	385
Total	Easy	606	2644	2150	1434	1542	492	2980	1462	2580	3700
	Hard	606	2644	2134	1434	1525	492	2980	1462	2580	3696
	MCQ	303	1322	1075	717	771	246	1490	731	1290	1850

from other LLMs, the LLMs4OL is the only domain-specific finetuned approach selected for comparison.

3.2 Implementation Details

We interacted with GPTs through Azure OpenAI API and the OpenAI official API. The employed GPT-3.5 version is 2023-05-15 and the GPT-4 version is 2023-11-06-preview. For other LLMs, we used 8 GeForce RTX 3090 GPUs and 4 NVIDIA A100 GPUs for the deployment. For the Vicuna series, we adopted the latest models Vicuna-7B-v1.5, Vicuna-13B-v1.5, and Vicuna-33B-v1.3. The original implementation of the LLMs4OL model is limited in general, geography, and medical domains. We adapted and fine-tuned the model to other taxonomy domains by preserving the official implementation details [18] suggested by LLMsOL as much as possible. To obtain stable outputs from LLMs, we employed the most deterministic hyper-parameter settings (e.g., temperature=0). All LLMs were experimented with the same question set for all the experiments.

3.3 Metrics

To cater to the needs for evaluating the quality of the answers provided by LLMs, similar to previous works [43, 70], we selected the following metrics for evaluation: **accuracy** (A), and **miss rate** (M), which measures the number of questions that LLMs give correct answers and " I don't know" over the number of all questions respectively. We consider an LLM as a good model if it achieves high accuracy with a low miss rate.

4 EXPERIMENTAL RESULTS

In this section, we analyze the experimental results of LLMs following multiple different research questions focusing on the overall performance, performance with respect to levels of taxonomies, the relationship between performance and model sizes, domain-agnostic fine-tuning, and domain-specific fine-tuning and influence from the prompting settings. Additionally, we include an additional instance typing experiment to further evaluate LLMs on a taxonomy-related task.

4.1 How reliable are LLMs for discovering hierarchical structures in different taxonomies?

We present the performance of LLMs on Hard, Easy, and MCQ datasets in Tables 5, 6, and 7.

Accuracy. We observe an overall decreasing trend in accuracy for LLMs from the common to specialized taxonomies on the three datasets, demonstrating a drop in LLM's reliability when we go from common to specialized taxonomies. Exceptions are the OAE and ICD-10-CM taxonomies, where the LLMs achieve good performance. LLMs perform well on the OAE taxonomy might be due to the high similarity in terms of names between the parent and child concepts as shown in Figure 1. While the ICD-10-CM data might be covered by the training data of the LLMs since it is also widely used in many common and non-medical domains (e.g., insurance billing processes) [10]. On the NCBI, Glottolog and GeoNames hard datasets, the accuracy of the best LLM is only around 70%. We attribute these phenomena to the fact that the domain knowledge of common taxonomies tends to be covered by the pre-training data of LLMs, while the knowledge of specialized taxonomies such as NCBI, Glottolog, and GeoNames is scarce on the internet and thus is less likely to be included in the pre-training data. Accurately determining hierarchical structures on specialized domains still requires support from the traditional taxonomy learning approaches.

Miss Rate. When analyzing the miss rates of different LLMs, we observe that Flan-T5-3B, Flan-T5-11B, and LLMs4OL have zero miss rates, in other words, they always provide their best guesses, while Llama-2-7B and Falcon-40B tend to be conservative: always provide "I don't know" as responses. We further observe rises in miss rates of

Table 5: Overall results on hard datasets.

		eBay	Amazon	Google	Schema	ACM-CCS	GeoNames	Glottolog	ICD-10-CM	OAE	NCBI
GPT-3.5	*A*	0.891	0.724	0.814	0.591	0.617	0.598	0.510	0.838	0.767	0.495
	M	0.021	0.138	0.042	0.324	0.150	0.057	0.298	0.063	0.144	0.301
GPT-4	*A*	0.921	0.806	0.857	0.734	0.708	0.652	0.626	0.917	0.822	0.653
	M	0.003	0.051	0.011	0.193	0.017	0.002	0.154	0.001	0.035	0.132
Claude-3	*A*	0.901	0.668	0.781	0.315	0.624	0.679	0.244	0.871	0.766	0.449
	M	0.033	0.231	0.090	0.663	0.111	0.138	0.714	0.041	0.161	0.456
Llama-2-7B	*A*	0.201	0.052	0.092	0.000	0.032	0.006	0.001	0.114	0.004	0.000
	M	0.789	0.946	0.903	1.000	0.963	0.994	0.999	0.871	0.996	1.000
Llama-2-13B	*A*	0.898	0.766	0.822	0.712	0.658	0.543	0.192	0.811	0.757	0.457
	M	0.000	0.002	0.003	0.010	0.011	0.006	0.681	0.027	0.078	0.252
Llama-2-70B	*A*	0.899	0.806	0.836	0.616	0.687	0.553	0.305	0.826	0.747	0.535
	M	0.000	0.000	0.000	0.000	0.003	0.000	0.467	0.023	0.017	0.130
Llama-3-8B	*A*	0.880	0.789	0.774	0.788	0.664	0.663	0.608	0.878	0.851	0.691
	M	0.000	0.000	0.000	0.000	0.000	0.000	0.078	0.000	0.000	0.011
Llama-3-70B	*A*	0.904	0.770	0.824	0.419	0.705	0.693	0.388	0.881	0.800	0.551
	M	0.000	0.050	0.011	0.531	0.048	0.073	0.474	0.003	0.088	0.231
Flan-T5-3B	*A*	0.899	0.781	0.835	0.743	0.672	0.539	0.584	0.767	0.838	0.593
	M	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Flan-T5-11B	*A*	0.919	0.793	0.864	0.786	0.698	0.520	0.589	0.842	0.856	0.633
	M	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Falcon-7B	*A*	0.597	0.547	0.556	0.501	0.550	0.537	0.503	0.636	0.497	0.587
	M	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Falcon-40B	*A*	0.434	0.253	0.348	0.013	0.043	0.108	0.021	0.454	0.007	0.013
	M	0.515	0.711	0.591	0.987	0.950	0.858	0.975	0.489	0.991	0.986
Vicuna-7B	*A*	0.827	0.725	0.728	0.699	0.599	0.705	0.637	0.757	0.813	0.609
	M	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002
Vicuna-13B	*A*	0.690	0.625	0.601	0.581	0.527	0.492	0.301	0.666	0.424	0.350
	M	0.007	0.037	0.022	0.093	0.023	0.114	0.557	0.077	0.450	0.460
Vicuna-33B	*A*	0.759	0.713	0.682	0.728	0.591	0.728	0.496	0.772	0.820	0.522
	M	0.000	0.000	0.000	0.020	0.000	0.000	0.277	0.001	0.002	0.187
Mistral	*A*	0.583	0.474	0.532	0.214	0.433	0.240	0.146	0.478	0.405	0.176
	M	0.262	0.361	0.230	0.750	0.308	0.691	0.818	0.410	0.528	0.772
Mixtral	*A*	0.805	0.739	0.738	0.707	0.618	0.604	0.394	0.840	0.789	0.482
	M	0.000	0.000	0.000	0.017	0.111	0.041	0.450	0.018	0.052	0.313
LLMs4OL	*A*	0.904	0.849	0.860	0.912	0.753	0.677	0.711	0.891	0.906	0.725
	M	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

GPT-3.5, GPT-4, Vicuna-13B, and Vicuna-33B on the Glottolog and NCBI taxonomies, which are the difficult specialized taxonomies that most LLMs perform poorly on. This is desirable since these models have learned to be cautious in taxonomies where they do not have sufficient domain knowledge.

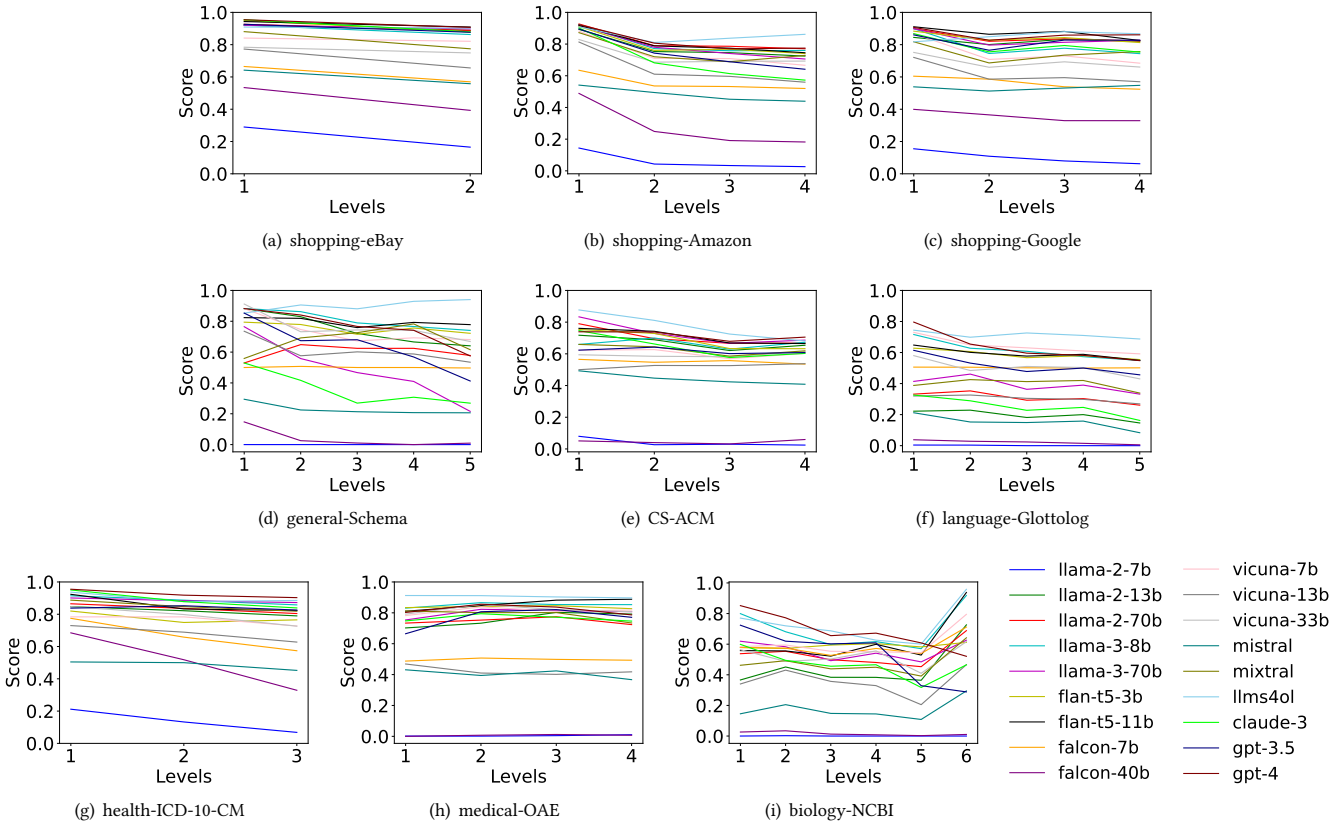
Different question types. Comparing the experimental results between the Easy, Hard, and MCQ datasets, we observe that providing MCQ options significantly reduces the miss rates of the LLMs. For instance, the average miss rates of the Llama-3-70B model reduce from 0.151 on the Hard datasets to 0.005 on the MCQ datasets. The average accuracy of Llama-3-70B in turn rises from 0.694 to 0.791.

Finding 1: The state-of-the-art LLMs are reliable in more common domains such as Shopping and General; while lacking sufficient domain knowledge in more specialized domains such as Computer Science Research, Biology, Language, and Geography.

4.2 Do LLMs perform equally well among different levels of taxonomies?

To answer this question, we conducted experiments on each level (level n to level n-1) of the taxonomies, because of the page limit, we only presented the accuracy results of hard datasets in Figure 3. Since the GeoNames taxonomy only has two concept levels, resulting in only one set of experiments (level 1 to root), we omit the demonstration of its results in the figure.

Accuracy. For common shopping taxonomies, as shown in Figures 3(a), 3(b), and 3(c), despite fluctuation, the accuracy of all LLMs tend to decrease as we go from the shallow levels (root) to deep levels (leaf). Most LLMs can achieve around 80% accuracy in all levels in these taxonomies. A similar root-to-leaf performance decline trend can also be observed on taxonomies Schema.org, ACM-CCS, Glottolog, and ICD-10-CM as shown in Figures 3(d), 3(e), 3(f), and 3(g). On the general domain taxonomy Schema.org, LLMs4OL, the best LLM achieves over 90% accuracy across different levels, indicating its mastery of general domain knowledge. We surprisingly



(a) shopping-eBay (b) shopping-Amazon (c) shopping-Google

(d) general-Schema (e) CS-ACM (f) language-Glottolog

(g) health-ICD-10-CM (h) medical-OAE (i) biology-NCBI

Figure 3: Accuracies for different levels of questions in hard datasets of different taxonomies under the zero-shot setting.

observe that the root-to-leaf performance decline trend does not apply to the NCBI taxonomy as shown in Figure 3(i): Most LLMs experience sudden performance uplifts at the last level. We notice that the last level in the NCBI taxonomy corresponds to the set of species-to-genus questions. We dived into the NCBI taxonomy database and discovered that this might be due to the data property of biological taxonomy: The names between species and corresponding genus tend to be similar in forms. For example, the ancestor chain of the species *Verbascum chaixii* is *Eukaryota-Strepyophyta-Magnoliopsida-Scrophulariaceae-Verbascum-Verbascum chaixii*, where the names between species and genus are very similar in forms. As a result, LLMs may take advantage of the similarity in forms between species and genus and thus achieve good performance at the last level. Despite the performance uplifts at the last level, the performances of the state-of-the-art LLMs at the middle levels (e.g., level 3, level 4, and level 5) are still very poor: slightly better than random guessing. A performance uplift trend from root to leaf levels can be observed in the OAE taxonomy, which also has similar name forms between the parent and child concepts near the leaf levels.

Finding 2: LLMs tend to present a root-to-leaf performance decline trend in most of the taxonomies. Additional support to improve LLMs' performance on leaf-level entities remains a promising direction for future ontology learning research.

4.3 Do normal methods that improve LLMs increase the accuracy?

We consider three normal methods that improve LLM reliability discussed by other works [24, 72, 77] to see if they work on taxonomies: improving the model size, providing domain-agnostic instruction fine-tuning, conducting domain-specific instruction fine-tuning.

Larger Model Sizes. Tables 5, 6, and 7 show LLMs with different sizes and their corresponding performances in different taxonomies. Since the GPTs and Claude-3 do not release information on model sizes, we only analyze the open-sourced models in this section. Specifically, for the Llama-2 series and Flan-T5 series, we notice that Llama-2-70B outperforms Llama-2-13B and Llama-2-7B and Flan-T5-11B outperforms Flan-T5-3B in most taxonomies, which indicates that increasing the sizes of LLMs can improve the models' performance for Llama-2 and Flan-T5. However, for the Vicunas and Flacons on the easy and hard datasets: Vicuna-7B outperforms Vicuna-13B in all the taxonomies and achieves better performance than Vicuna-33B in half of the taxonomies; Falcon-7B significantly outperforms Falcon-40B in all taxonomies. Besides, we observe that on the easy and hard datasets, the miss rates of Falcon-40B are significantly higher than those of Falcon-7B, which means Falcon-40B tends to be more conservative in answering hierarchical structure discovery questions and thus generates more "I don't know" answers. This observation coincides with the observation regarding Falcon-40B presented in a previous study [70]. We attribute this phenomenon to the fact that once the LLM is sufficiently large, the differences in pre-training data and strategies play a vital role in determining the performance of answering hierarchical structure discovery questions in taxonomies.

Domain-Agnostic Fine-Tuning. We further compare the Llama-2 series and the Vicuna series as shown in Tables 5, 6, and 7 to

consider the effect of domain-agnostic fine-tuning. As discussed in [29, 77], the Vicuna models are fine-tuned Llama-2 models based on the dialog data in the ShareGPT dataset and thus should produce answers with higher quality. Naturally, an interesting question is whether such domain-agnostic fine-tuning can improve the performance of LLMs in answering taxonomy structure questions. However, we observe that although Vicuna-7B significantly improves the performance of Llama-2-7B, Vicuna-13B is outperformed by its original model Llama-2-13B on the easy and hard datasets. On the MCQ dataset, Vicuna-13B improves the performance of Llama-2-13B on some taxonomies only. The reason might be that the domain-agnostic dialog data fine-tuning may have positive effects on the model to better understand the question format (Vicuna-7B and Llama-2-7B), while it might not improve the performance significantly and stably or even bring negative effects if the model can already well understand the question format because of the miss-match of knowledge coverage between the domain-agnostic fine-tuning data and the domain-specific taxonomy data (Vicuna-13B and Llama-2-13B).

Domain-specific Fine-tuning. Considering the domain-specific fine-tuning, we observe that the instruction-tuned LLMs4OL largely outperforms its backbone model Flan-T5-3B. Specifically, the averaged accuracy over all taxonomies of LLMs4OL boosts the averaged accuracy of Flan-T5-3B by 12.9%, 12.9%, and 17.0% on the easy, hard, and MCQ datasets, which showcases the significant benefit of performing domain-specific instruction fine-tuning.

Finding 3: Normal methods, including using larger model sizes and domain-agnostic fine-tuning, may not lead to an increase in performance. The domain-specific fine-tuning leads to a stable and significant performance uplift. Indeed, the answer quality for the hierarchical structure discovery questions in taxonomies is related to the domain knowledge coverage of the pretraining data. Introducing domain-specific fine-tuning can increase the domain knowledge coverage of the LLMs.

4.4 Do different prompting settings influence the performance?

Similar to the prompting settings adopted by previous work [43], we introduced two additional prompting settings to further evaluate the performance of LLMs: Few-shot learning, and Chain-of-Thoughts (CoT). As shown in [72] and [49], few-shot and CoT prompting techniques can improve LLMs' performance. Therefore, we want to include these prompting settings to see if they can improve LLMs' performance on taxonomies. For the few-shot setting, following [43], we conducted five-shot experiments. To avoid introducing bias in the examples, we sample positive and negative pairs with equal probability. In addition, to investigate if improving the reasoning ability of LLMs enhances the performance [75], we conducted chain-of-thoughts (CoT) experiments following [49] by providing an extra prompt "Let's think step by step." at the end of the questions to guide LLMs through more reasoning steps. Please refer to Figure 5 for an example of the few-shot and CoT settings of our experiments.

We present the radar charts of the performance of representative LLMs in hard datasets of different taxonomies under zero-shot, few-shot, and CoT prompting settings in Figure 4.

Few-Shot Prompting. Few-shot prompting can improve the performance of some LLMs in answering hierarchical structure discovery questions in different taxonomies, yet cannot significantly improve the performance of the top-performing LLMs. We observe that compared to zero-shot prompting, few-shot prompting reduces the miss rates of LLMs: the miss rates of Llama-2-7B reduce significantly (Figure 4(d)), and its corresponding accuracy in turn increases (Figure 4(c)). Llama-2-7B benefits from the few-shot prompting: changing from scoring less than or close to 20% accuracy in all taxonomies to achieving comparable performance to Flan-T5-3B on some taxonomies. We believe the miss rates reduce because few-shot prompting provides concrete question-answering pairs so that the models are more confident in imitating the prompt examples provided to generate their best guesses. However, the performance uplift of the LLMs with low miss rates is not significant. For instance, the changes to the performance of Flan-T5-11B in most taxonomies are not significant (Figure 4(b)), which implies that the effect brought by few-shot prompting mainly lies in reducing the miss rates of the LLMs instead of improving models' answering accuracy.

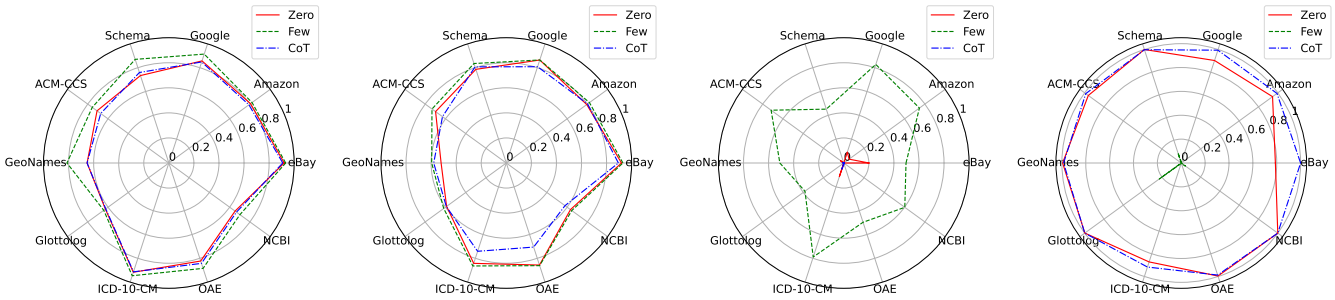
Chain-of-Thoughts (CoT). The CoT prompting harms the performance of some of the LLMs, but the influence brought to the top-performing LLMs is minimal. Comparing the miss rates of zero-shot and CoT prompting, we observe that by introducing CoT prompting, the miss rates of Llama-2-7B rise (Figure 4(d)). We attribute this phenomenon to the fact that the hierarchical structure discovery questions in taxonomies are simple-formed questions, that do not require complex reasoning, and thus CoT may not be helpful for this type of question-answering task, which coincides with the observation in a recent study that CoT is helpful for complex reasoning process [75]. Despite the phenomenon that some LLMs are influenced by CoT prompting, we find that the performance of GPT-4 is stable under CoT prompting: remains unchanged or drops by only a small extent (Figure 4(a)).

Finding 4: The performance changes brought by few-shot and CoT are minimal to the best LLMs such as GPT-4. The main effect of these prompting settings is to influence the miss rates of LLMs, instead of directly improving the accuracy.

4.5 Instance Typing

Other than determining the hierarchical structures of taxonomies, which is the core task we focus on in this paper, we want to further investigate the reliability of LLMs on a taxonomy-related task: instance typing (i.e., the task of determining the types of instance entities under the leaf entities in the taxonomies.) to spark more discussion and thoughts.

We first define the instances as follows: The instances in Google and Amazon taxonomies are defined as product names under each leaf entity: we additionally crawled the product names of each leaf entity from Google shopping [17] and Browsenodes [4] websites as the instances. As for the ICD-10-CM taxonomy, we define the entities as diseases with different causes, which can be obtained as the fourth-level entities in the taxonomy. The species entities in the NCBI taxonomy are considered as instances. Moreover, we treat the leaf entity language as the instances in the Glottolog taxonomy. The leaf entity adverse events are considered as the



(a) GPT-4, accuracy (b) Flan-T5-11B, accuracy (c) Llama-2-7B, accuracy (d) Llama-2-7B, miss rate

Figure 4: Radar charts for representative LLMs under different prompting settings in hard datasets.

```
# Few-shot
-----------------------------------------------------------------------------------------------------------
Example: Is <child-type> a type of <parent-type>? answer with (Yes/No/I don't know)
Yes.
Example: Is <child-type> a type of <uncle-type>? answer with (Yes/No/I don't know)
No.
Example: Is <child-type> a type of <parent-type>? answer with (Yes/No/I don't know)
Yes.
Example: Is <child-type> a type of <uncle-type>? answer with (Yes/No/I don't know)
No.
Example: Is <child-type> a type of <parent-type>? answer with (Yes/No/I don't know)
Yes.

Is <child-type> a type of <parent-type>? answer with (Yes/No/I don't know)

#############################################################################

# Chain-of-Thoughts
-----------------------------------------------------------------------------------------------------------
Is <child-type> a type of <parent-type>? answer with (Yes/No/I don't know) Let's think
step by step.
```

Figure 5: Few-shot and Chain-of-Thoughts examples.

instances for the OAE taxonomy. GeoNames and eBay taxonomies do not provide valid entities. Besides, for eBay Categories, we fail to find a proper way to crawl its product information as entities. The cases of Schema.org and ACM-CCS are complex: Schema.org and ACM-CCS do not have well-defined instances under their leaf entity concepts; besides, there is no appropriate data source for us to crawl proper entities for these taxonomies. As a result, we skip the instance typing experiment for these four taxonomies.

We adopted the same question templates as shown in Table 2. Following a similar True/False question generation manner described in Section 2.2, we generated the instance typing pairs in each level. For example, given an instance i, which is under an entity e_k in level k of taxonomy, we preserve the following instance typing pairs: $(i \rightarrow e_k)$, $(i \rightarrow e_k.p)$, ... $(i \rightarrow e_k.r)$, mark as the instance typing pairs in level k, $k-1$, ..., 0, where $e_k.p$ and $e_k.r$ are the intermediate parent and root entities of the entity e_k. Similar to Section 2.2, we generate negative hard samples and negative easy samples. We record the performances of LLMs under the zero-shot prompting setting and present the results on hard datasets in Figure 6 due to the similar trends between easy and hard datasets.

In general, we have the following observations: 1) Similar to the main experiments discussed in Section 4.1, the performance of LLMs presents a common to specialized decline except for the ICD-10-CM and OAE taxonomies due to a similar analysis presented in Section 4.1. 2) The overall instance typing performance drops as we go from root to leaf levels, except for the OAE and NCBI taxonomies whose concept names are highly overlapping near the leaf levels.

Finding 5: These validate our hypothesis that LLMs are reliable in performing tasks in more common taxonomies, which means instead of manually constructing and maintaining deep and intricate taxonomies in these common domains, we can rely on LLMs to complete most of the ontology learning work. However, in specialized taxonomies, such as NCBI and Glottolog, the better practice is still relying on the traditional tree-like structures.

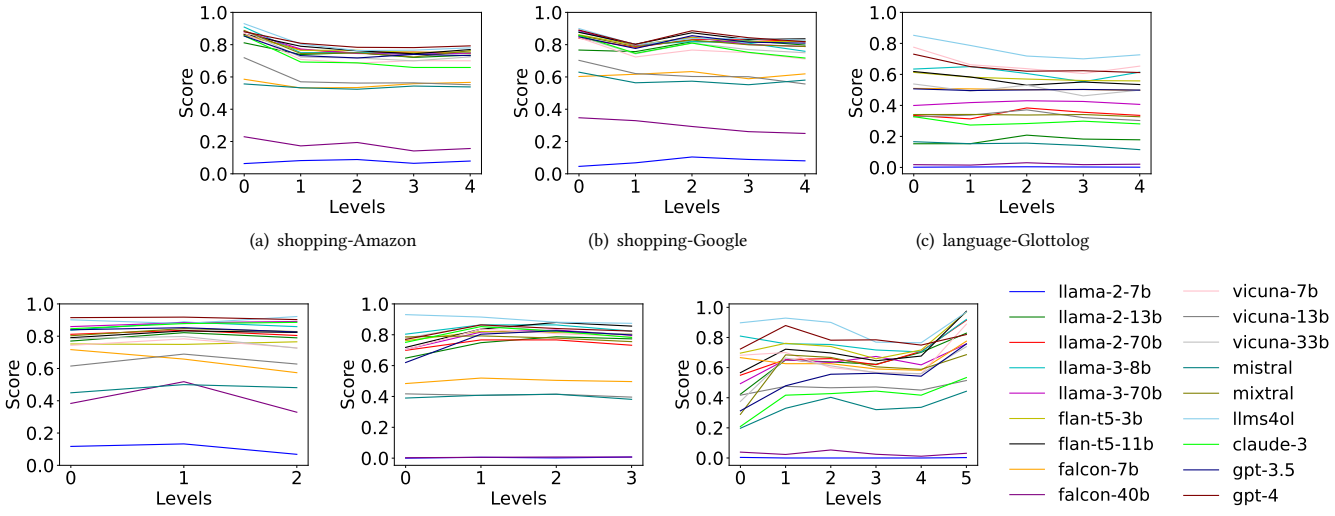
5 DISCUSSION

5.1 The Future of Taxonomy and LLMs

The experimental analysis showcases that despite the integration of taxonomy knowledge within the parameters of LLMs, the coverage of their knowledge in specialized domains and deeper parts of taxonomies is still limited. Specifically, the state-of-the-art LLMs demonstrate their mastery of taxonomic knowledge in common domains such as shopping and general, however, their performances in more specialized domains such as computer science research, biology, geography, and language are unsatisfactory. Besides, the root-to-leaf performance decline almost happens on every LLM in most taxonomies we experimented with, which indicates insufficient coverage of knowledge in deeper levels of taxonomies.

Nevertheless, LLMs have ushered in a paradigm shift in ontology learning, prompting a fundamental reconsideration of the most effective representation of taxonomies. Our vision for the next-generation taxonomy is to combine LLMs with the traditional tree-structure taxonomy to form a novel taxonomy form, where the hierarchical knowledge implicitly resides inside LLMs' weights or explicitly presents as the traditional "Is-A" parent-child structure. The harmonious synthesis of these two modalities, harnessing both the cutting-edge advancement of LLMs' knowledge and the reliability of traditional tree structure, presents a captivating and fertile research domain, which we shall explore in greater depth.

Common Taxonomies. Common taxonomies that are likely to be covered by the abundant training data of LLMs, such as shopping and general, should be encoded inside the LLMs. Despite that in some use cases such as relation display and visualization, there might still be places for the traditional taxonomic structure near root levels to exist, the majority of the use cases (such as entity searching and knowledge reasoning) in common taxonomies can be well handled by LLMs. We provide a concrete example to study



(a) shopping-Amazon (b) shopping-Google (c) language-Glottolog

(d) health-ICD-10-CM (e) medical-OAE (f) biology-NCBI

Figure 6: Instance typing experiments with respect to different levels of questions in hard datasets of different taxonomies.

the possibility of taxonomy replacement on Amazon Product Category in Section 5.3. LLMs already demonstrate high reliability as shown in the hierarchical structure discovery and instance typing experiments for these taxonomies. The manually constructed and maintained taxonomies in these domains may not be needed shortly. The few errors (less than 25%) that LLMs still have in these taxonomies are likely to get addressed by performing fine-tuning based on existing studies [41, 74].

More Specialized Taxonomies. The more specialized taxonomies that domain experts generally use, such as language, computer science research, biology, and geography, are likely to remain in their current tree-structure forms or change to LLM-tree-structure-combined forms. Since the state-of-the-art LLMs are still not ready to provide reliable responses for these more specialized taxonomies, especially near the leaf levels, where the performance of LLMs gets significantly poor or fluctuates. We also discovered that although LLMs can perform well near the root levels of these more specialized taxonomies, their performance near the leaf levels can be significantly worse (Glottolog and ACM-CCS) or unstable (NCBI). Therefore, we recommend that industrial practitioners continue with the current tree-structure taxonomies in specialized domains to ensure reliability; while the research communities should start exploring the possibility of LLM-tree-structure-combined taxonomy forms: the entities near the roots are transformed into LLMs' weights, while the entities near the leaves should remain in the traditional tree-structure form to achieve both high accuracy and minimal maintaining and constructing cost for ontology learning.

5.2 Limitations

As discussed in Section 5.1, LLMs without instruction fine-tuning struggle to achieve satisfactory performance at low levels of the specialized taxonomies, which is indeed a limitation of LLMs for taxonomy replacement. In this sense, a possible solution is to replace the traditional taxonomies on some of the levels where the LLMs can achieve high and stable performance. Please refer to Section 5.3 for an example of taxonomy replacement. Besides, the domain-specific instruction fine-tuning [32] (LLMs4OL) improves the performance of the original LLM (Flan-T5-3B) at low levels of the specialized taxonomies as shown in Figures 3(e)-3(i), which makes it a possible alternative to resolve the limitation.

Despite the fact that domain-specific instruction fine-tuning requires high-quality labeled data and induces high training costs, we believe these issues could be resolved by introducing domain adaptation techniques as discussed by some pilot works [34, 39, 48]. However, the effectiveness of these techniques on taxonomies is not yet validated and we think this should be a promising future direction to explore.

5.3 Case Study

To provide a concrete example of the integration of traditional taxonomy structure and LLMs, we conducted a case study on the performance and feasibility of the integrated solution with the Amazon Product Category. We replaced the nodes in level-4 or lower of the Amazon Product Category with the Llama-2-70B model, while preserving the nodes in root to level-3 for relation display and visualization purposes. Specifically, suppose there is a level-3 concept named "Stationery" and has descendants "Pen" and "Pencil" and there is a customer who searches for pencil products. Then traditionally, if he/she relies on the traditional taxonomy structure, the query would match the level-4 "Pencil" node and then get the product list under the "Pencil" category. After we remove the level-4 concepts, his/her query would instead first ask about the parent concept of the query concept "Pencil" with an accuracy of over 70% as shown in Figure 3(b). Then the query would find and match the level-3 concept "Stationery", and ask Llama-2-70B to return all the pencil products from the set of stationery products.

The performance of such replacement is evaluated as follows: given a level 4 or lower concept e_k, with a list of products under e_k: l_{e_k}. We denote the list of products of the siblings of e_k as $l_{e_k.s}$. We record the precision and recall of the product list $\hat{l_{e_k}}$ returned by Llama-2-70B when given the full stationery product list $\{l_{e_k} \cup l_{e_k.s}\}$. We sampled the leaf concepts with a confidence level of 95% and a margin of error of 5% similar to the **Question Generation** in Section 2.2 to form the experimental dataset.

By performing the replacement, we save 25777/43814 = 59% of the construction and maintenance cost of the taxonomy, as shown in

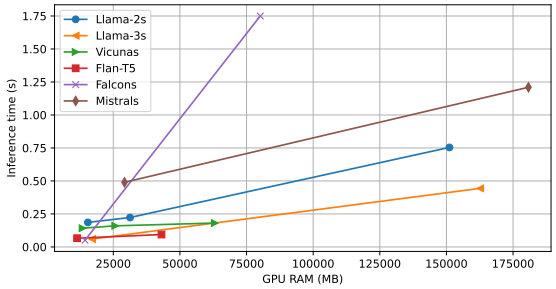


Figure 7: Scalability of different model series.

Table 1. The precision and recall of the integrated solution are 0.713 and 0.792 respectively. As such, by replacing the level 4 or lower concepts with Llama-2-70B, we saved 59% taxonomy construction and maintenance cost, while achieving an overall precision and recall of over 70%.

Note that we may replace more layers to achieve lower taxonomy construction and maintenance costs and introduce more advanced fine-tuning techniques for the LLMs or adopt ranking techniques [53, 56] to achieve better precision and recall. The case study serves as a pilot study of the feasibility of the integration of taxonomy and LLMs optimizing and refining the integration solution could be a promising research topic for the community to work on.

5.4 Scalability

We analyze the scalability of the LLM series by recording the GPU RAM and the average time costs induced by each LLM model in the corresponding LLM series during the inference of zero-shot taxonomy questions. The experimental results are presented in Figure 7. Specifically, we considered the scalability of six open-sourced LLM series: Llama-2s, Llama-3s, Vicunas, Flan-T5s, Falcons, and Mistrals. We observe that Flan-T5s, Vicunas, and Llama-3s present relatively good scalability, as the model size grows, the inference time does not increase significantly, which is especially important for their adoption in real-world taxonomy-related applications.

6 RELATED WORK

Benchmarks and experimental analysis. Many QA benchmarks were developed to evaluate the ability of language models [26, 33, 46, 51, 67]. However, these benchmarks do not focus on the taxonomy data and common to specialized domain knowledge. Two recent works further the study of long-tail domain knowledge by constructing new QA benchmarks [50, 58]. Sun et al. [70] introduced Head-to-tail, a novel benchmark that systematically analyzes the factuality of LLMs over KG entities from common to long-tail. Luo et al. [57] proposed an automatic question generation method to generate factual questions from common to specialized domains. These works mainly analyze LLMs' performance on KG instead of taxonomy. Although a pilot work LLMs4OL [24] explored the possibility of utilizing LLMs to perform ontology learning, the domain coverage of their evaluation is rather limited: mainly on the common and bio-medical domains, which cannot comprehensively reflect LLMs' knowledge in various taxonomies from common to specialized domains. Besides, LLMs4OL failed to provide an in-depth analysis of LLMs' performance in different levels of the taxonomies, which is indeed important for the audience to explore whether LLMs can replace taxonomies. As discussed in [70], LLMs are less knowledgeable for the long-tail, nuanced knowledge in KGs. Intuitively, we believe a similar phenomenon should present in different levels of taxonomies, i.e., LLMs are less knowledgeable at the leaf levels of taxonomies and thus should be an issue considered if the users plan to use LLMs in replacement of the traditional taxonomies. However, no existing study on taxonomies considered this direction. Compared with existing works, TaxoGlimpse is the first benchmark that systematically covers the taxonomies instead of KG from common to specialized domains with in-depth root-to-leaf analysis, which is less touched by previous studies.

LLM prompting and its settings. As the major way to probe the knowledge of LLMs towards taxonomies, we would like to briefly introduce the LLM prompting methods with different prompting settings. The prompts adopted by us to evaluate the LLMs' performance over taxonomies are named prefix prompts [52, 54, 55], which are suitable for the general question-answering scenario. As discussed by [55], the methods to design prompts can be classified into manual template engineering and automated template learning. In our paper, we chose the manual template engineering approach, which is to manually design intuitive templates based on experience. The reason why we did not adopt the advanced automatic template generation approaches is that the primary focus of our work is to provide an initial analysis of LLMs' performance on taxonomies and we believe that manually crafted templates can achieve the goal of reflecting LLMs' knowledge as done by [62, 70].

The popular prompting settings include zero-shot, few-shot, Chain-of-Thoughts, etc. Zero-shot is straightforward: the LLMs take the input question and return the corresponding answer directly [63]. Instead of directly querying the LLMs with the question, under few-shot settings, users additionally provide several examples of questions and answers, and then query the LLM with the desired question and receive the response [27, 65]. Chain-of-Thoughts (CoT) prompting [49, 75] guides the LLMs to break down a complex reasoning question into several intermediate steps. By solving the questions step-by-step, the LLMs return more reasonable answers, especially for questions that require complex reasoning.

7 CONCLUSION

In this paper, we introduced TaxoGlimpse, a novel taxonomy hierarchical structure benchmark that comprehensively evaluates the performance of LLMs over different taxonomies from common to specialized domains, from root to leaf levels. We systematically evaluated the performances of eighteen state-of-the-art LLMs under three popular prompting settings: zero-shot, few-shot, and Chain-of-Thoughts at different levels of ten representative taxonomies. Four highly concerned research questions were proposed and resolved and we provided valuable insights into future research opportunities for industrial users, LLM developers, and database researchers. Our comprehensive evaluation shows that even the best-performing LLM presents unsatisfactory performances at specialized taxonomies and for entities near the leaf levels. In response, we suggest future research directions to combine the LLMs with traditional taxonomies so as to create novel neural-symbolic taxonomies that have the best of both worlds.

REFERENCES

[1] 2012. ACM CCS Concept 2012 link. Retrieved Jan 10, 2024 from https://dl.acm.org/pb-assets/dl_ccs/acm_ccs2012-1626988337597.xml

[2] 2012. ACM Computing Classification System. Retrieved Jan 10, 2024 from https://dl.acm.org/ccs

[3] 2019. Amazon's Product Category. Retrieved Jan 10, 2024 from https://www.browsenodes.com/

[4] 2021. Google Product Category. Retrieved Jan 10, 2024 from https://www.google.com/basepages/producttype/taxonomy.en-US.txt

[5] 2021. ICD-10-CM package. Retrieved Jan 10, 2024 from https://pypi.org/project/simple-icd-10/

[6] 2022. Google Shopping Statistics. Retrieved Jan 10, 2024 from https://www.statista.com/statistics/1341380/most-well-known-price-comparison-portals-in-the-united-states/

[7] 2022. OAE-website. Retrieved Apr 17, 2024 from https://bioportal.bioontology.org/ontologies/OAE

[8] 2023. Amazon's Product Category statistics. Retrieved Jan 10, 2024 from https://www.statista.com/forecasts/997230/most-popular-online-shops-in-the-us

[9] 2023. Glottolog-4.8. Retrieved Jan 10, 2024 from https://glottolog.org/meta/downloads

[10] 2023. ICD-10-CM for public. Retrieved Jan 10, 2024 from https://www.verywellhealth.com/icd-10-codes-5271405

[11] 2023. ICD-10-CM taxonomy information. Retrieved Jan 10, 2024 from https://www.cdc.gov/nchs/icd/icd-10-cm.htm

[12] 2023. NCBI data download. Retrieved Jan 10, 2024 from https://ftp.ncbi.nlm.nih.gov/pub/taxonomy/

[13] 2023. Qualtrics. Retrieved Jan 10, 2024 from https://www.qualtrics.com/au/experience-management/research/determine-sample-size/?rid=ip&prevsite=en&newsite=au&geo=HK&geomatch=au

[14] 2024. Claude-v3-documentation. Retrieved Apr 18, 2024 from https://www.anthropic.com/news/claude-3-family

[15] 2024. eBay. Retrieved Apr 17, 2024 from https://www.ebay.com/n/all-categories

[16] 2024. geonames-website. Retrieved Apr 17, 2024 from https://www.geonames.org/export/codes.html

[17] 2024. Google Shopping Website. Retrieved Jan 10, 2024 from https://shopping.google.com/

[18] 2024. LLMs4OL-code. Retrieved Apr 17, 2024 from https://github.com/HamedBabaei/LLMs4OL/tree/main/TaskB

[19] 2024. schema-website. Retrieved Apr 17, 2024 from https://github.com/schemaorg/schemaorg/blob/main/data/releases/26.0/schemaorg-current-https-types.csv

[20] 2024. TaxoGlimpse experimental results. Retrieved Apr 17, 2024 from https://github.com/ysunbp/TaxoGlimpse/tree/main/exp-results

[21] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774* (2023).

[22] AI@Meta. 2024. Llama 3 Model Card. (2024). https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md

[23] Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867* (2023).

[24] Hamed Babaei Giglou, Jennifer D'Souza, and Sören Auer. 2023. LLMs4OL: Large language models for ontology learning. In *International Semantic Web Conference*. Springer, 408–427.

[25] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023* (2023).

[26] Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 conference on empirical methods in natural language processing*. 1533–1544.

[27] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[28] Andrew Caines, Christian Bentz, Dimitrios Alikaniotis, Fridah Katushemererwe, and Paula Buttery. 2016. The Glottolog data explorer: Mapping the world's languages. *Proceedings of VisLR II: Visualization as Added Value in the Development, Use and Evaluation of Language Resources* (2016), 38–53.

[29] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 Apr 2023)* (2023).

[30] Yejin Choi. 2023. Common Sense: The Dark Matter of Language and Intelligence. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*. 2–2.

[31] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416* (2022).

[32] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research* 25, 70 (2024), 1–53.

[33] Mohnish Dubey, Debayan Banerjee, Abdelrahman Abdelkawi, and Jens Lehmann. 2019. Lc-quad 2.0: A large dataset for complex question answering over wikidata and dbpedia. In *The Semantic Web–ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part II 18*. Springer, 69–78.

[34] Yuqi Fang, Pew-Thian Yap, Weili Lin, Hongtu Zhu, and Mingxia Liu. 2024. Source-free unsupervised domain adaptation: A survey. *Neural Networks* (2024), 106230.

[35] Scott Federhen. 2012. The NCBI taxonomy database. *Nucleic acids research* 40, D1 (2012), D136–D143.

[36] Ramanathan V Guha, Dan Brickley, and Steve Macbeth. 2016. Schema. org: evolution of structured data on the web. *Commun. ACM* 59, 2 (2016), 44–51.

[37] Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2023. Glottolog 4.8. (2023).

[38] Harald Hammarström and Robert Forkel. 2022. Glottocodes: Identifiers Linking Families, Languages and Dialects to Comprehensive Reference Information. *Semantic Web Journal* 13, 6 (2022), 917–924. https://content.iospress.com/articles/semantic-web/sw212843

[39] Pengrui Han, Rafal Kocielnik, Adhithya Saravanan, Roy Jiang, Or Sharir, and Anima Anandkumar. 2024. ChatGPT Based Data Augmentation for Improved Parameter-Efficient Debiasing of LLMs. *arXiv preprint arXiv:2402.11764* (2024).

[40] Yongqun He, Sirarat Sarntivijai, Yu Lin, Zuoshuang Xiang, Abra Guo, Shelley Zhang, Desikan Jagannathan, Luca Toldo, Cui Tao, and Barry Smith. 2014. OAE: the ontology of adverse events. *Journal of biomedical semantics* 5 (2014), 1–13.

[41] Linmei Hu, Zeyi Liu, Ziwang Zhao, Lei Hou, Liqiang Nie, and Juanzi Li. 2023. A survey of knowledge enhanced pre-trained language models. *IEEE Transactions on Knowledge and Data Engineering* (2023).

[42] Jin Huang, Zhaochun Ren, Wayne Xin Zhao, Gaole He, Ji-Rong Wen, and Daxiang Dong. 2019. Taxonomy-aware multi-hop reasoning networks for sequential recommendation. In *Proceedings of the twelfth ACM international conference on web search and data mining*. 573–581.

[43] Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, et al. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *arXiv preprint arXiv:2305.08322* (2023).

[44] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825* (2023).

[45] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088* (2024).

[46] Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551* (2017).

[47] Giannis Karamanolakis, Jun Ma, and Xin Luna Dong. 2020. TXtract: Taxonomy-Aware Knowledge Extraction for Thousands of Product Categories. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 8489–8502.

[48] Rafal Kocielnik, Sara Kangaslahti, Shrimai Prabhumoye, Meena Hari, Michael Alvarez, and Anima Anandkumar. 2023. Can you label less by using out-of-domain data? Active & transfer learning with few-shot instructions. In *Transfer Learning for Natural Language Processing Workshop*. PMLR, 22–32.

[49] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems* 35 (2022), 22199–22213.

[50] Rohan Kumar, Youngmin Kim, Sunitha Ravi, Haitian Sun, Christos Faloutsos, Ruslan Salakhutdinov, and Minji Yoon. 2023. Automatic Question-Answer Generation for Long-Tail Knowledge. (2023).

[51] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics* 7 (2019), 453–466.

[52] Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691* (2021).

[53] Hang Li. 2022. *Learning to rank for information retrieval and natural language processing*. Springer Nature.

[54] Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190* (2021).

[55] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Comput. Surv.* 55, 9, Article 195 (jan 2023), 35 pages. https://doi.org/10.1145/3560815

[56] Tie-Yan Liu et al. 2009. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval* 3, 3 (2009), 225–331.

[57] Linhao Luo, Trang Vu, Dinh Phung, and Reza Haf. 2023. Systematic Assessment of Factual Knowledge in Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 13272–13286.

[58] Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 9802–9822. https://doi.org/10.18653/v1/2023.acl-long.546

[59] Fabian Neuhaus. 2023. Ontologies in the era of large language models–a perspective. *Applied Ontology* 18, 4 (2023), 399–407.

[60] Sebastian Nordhoff and Harald Hammarström. 2011. Glottolog/Langdoc: Defining dialects, languages, and language families as collections of resources. In *First International Workshop on Linked Science 2011-In conjunction with the International Semantic Web Conference (ISWC 2011)*.

[61] Panagiotis Papadimitriou, Panayiotis Tsaparas, Ariel Fuxman, and Lise Getoor. 2012. TACI: Taxonomy-aware catalog integration. *IEEE Transactions on knowledge and data engineering* 25, 7 (2012), 1643–1655.

[62] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066* (2019).

[63] Bernardino Romera-Paredes and Philip Torr. 2015. An embarrassingly simple approach to zero-shot learning. In *International conference on machine learning*. PMLR, 2152–2161.

[64] Eric W Sayers, Mark Cavanaugh, Karen Clark, James Ostell, Kim D Pruitt, and Ilene Karsch-Mizrachi. 2019. GenBank. *Nucleic acids research* 47, Database issue (2019), D94.

[65] Timo Schick and Hinrich Schütze. 2020. It's not just size that matters: Small language models are also few-shot learners. *arXiv preprint arXiv:2009.07118* (2020).

[66] Conrad L Schoch, Stacy Ciufo, Mikhail Domrachev, Carol L Hotton, Sivakumar Kannan, Rogneda Khovanskaya, Detlef Leipe, Richard Mcveigh, Kathleen O'Neill, Barbara Robbertse, et al. 2020. NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database* 2020 (2020), baaa062.

[67] Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. 2021. Simple Entity-Centric Questions Challenge Dense Retrievers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 6138–6148. https://doi.org/10.18653/v1/2021.emnlp-main.496

[68] Yasemin Sen. 2019. Knowledge as a valuable asset of organizations: Taxonomy, management and implications. In *Management science: Foundations and innovations*. Springer, 29–48.

[69] Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*. 697–706.

[70] Kai Sun, Yifan Ethan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. 2023. Head-to-tail: How knowledgeable are large language models (llm)? AKA will llms replace knowledge graphs? *arXiv preprint arXiv:2308.10168* (2023).

[71] Yiming Tan, Dehai Min, Yu Li, Wenbo Li, Nan Hu, Yongrui Chen, and Guilin Qi. 2023. Evaluation of ChatGPT as a question answering system for answering complex questions. (2023).

[72] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

[73] Somin Wadhwa, Silvio Amir, and Byron C Wallace. 2023. Revisiting relation extraction in the era of large language models. *arXiv preprint arXiv:2305.05003* (2023).

[74] Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Guihong Cao, Daxin Jiang, Ming Zhou, et al. 2020. K-adapter: Infusing knowledge into pre-trained models with adapters. *arXiv preprint arXiv:2002.01808* (2020).

[75] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.

[76] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2023. Benchmarking large language models for news summarization. *arXiv preprint arXiv:2301.13848* (2023).

[77] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685* (2023).

[78] Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji-Rong Wen. 2023. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107* (2023).

8 SUPPLEMENTARY MATERIALS

This section contains the supplementary materials for our paper. In Section 8.1, we present the experimental results on Easy and MCQ datasets.

8.1 Experimental results

We present the experimental results on Easy and MCQ datasets in Tables 6, and 7.

Table 6: Overall results on easy datasets.

		eBay	Amazon	Google	Schema	ACM-CCS	GeoNames	Glottolog	ICD-10-CM	OAE	NCBI
GPT-3.5	*A*	0.921	0.775	0.920	0.593	0.711	0.598	0.563	0.851	0.778	0.529
	M	0.026	0.148	0.045	0.334	0.169	0.057	0.310	0.099	0.167	0.354
GPT-4	*A*	0.946	0.879	0.962	0.773	0.860	0.648	0.710	0.964	0.866	0.789
	M	0.002	0.044	0.008	0.183	0.014	0.002	0.141	0.001	0.032	0.089
Claude-3	*A*	0.932	0.758	0.910	0.331	0.784	0.679	0.256	0.953	0.869	0.486
	M	0.018	0.199	0.068	0.664	0.121	0.138	0.736	0.021	0.095	0.494
Llama-2-7B	*A*	0.196	0.053	0.090	0.000	0.032	0.006	0.001	0.115	0.004	0.000
	M	0.804	0.946	0.908	1.000	0.967	0.994	0.999	0.880	0.996	1.000
Llama-2-13B	*A*	0.926	0.798	0.886	0.727	0.789	0.543	0.149	0.815	0.714	0.411
	M	0.000	0.012	0.005	0.020	0.024	0.006	0.733	0.077	0.145	0.310
Llama-2-70B	*A*	0.931	0.865	0.945	0.616	0.817	0.553	0.292	0.884	0.761	0.536
	M	0.000	0.001	0.000	0.001	0.004	0.000	0.487	0.040	0.022	0.138
Llama-3-8B	*A*	0.931	0.871	0.940	0.819	0.829	0.663	0.665	0.915	0.893	0.785
	M	0.000	0.000	0.000	0.000	0.000	0.000	0.086	0.020	0.000	0.015
Llama-3-70B	*A*	0.939	0.797	0.927	0.376	0.787	0.693	0.354	0.899	0.804	0.514
	M	0.005	0.103	0.039	0.602	0.104	0.073	0.591	0.055	0.139	0.399
Flan-T5-3B	*A*	0.926	0.826	0.934	0.751	0.732	0.539	0.588	0.811	0.851	0.605
	M	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Flan-T5-11B	*A*	0.944	0.834	0.944	0.804	0.737	0.520	0.595	0.871	0.875	0.643
	M	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Falcon-7B	*A*	0.607	0.553	0.574	0.504	0.577	0.533	0.504	0.677	0.495	0.618
	M	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Falcon-40B	*A*	0.434	0.255	0.357	0.013	0.042	0.106	0.021	0.449	0.005	0.012
	M	0.541	0.732	0.636	0.987	0.957	0.860	0.977	0.536	0.994	0.987
Vicuna-7B	*A*	0.919	0.814	0.915	0.723	0.748	0.705	0.671	0.877	0.878	0.676
	M	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002
Vicuna-13B	*A*	0.812	0.708	0.741	0.580	0.575	0.492	0.297	0.668	0.409	0.347
	M	0.035	0.115	0.084	0.133	0.059	0.114	0.620	0.234	0.526	0.530
Vicuna-33B	*A*	0.871	0.839	0.857	0.792	0.737	0.728	0.536	0.887	0.887	0.601
	M	0.000	0.000	0.000	0.010	0.000	0.000	0.272	0.016	0.001	0.216
Mistral	*A*	0.571	0.460	0.509	0.211	0.428	0.240	0.146	0.467	0.405	0.176
	M	0.347	0.480	0.433	0.774	0.471	0.691	0.842	0.491	0.565	0.811
Mixtral	*A*	0.898	0.829	0.894	0.745	0.656	0.604	0.369	0.873	0.809	0.451
	M	0.005	0.002	0.002	0.024	0.202	0.041	0.526	0.048	0.066	0.465
LLMs4OL	*A*	0.929	0.890	0.908	0.932	0.844	0.677	0.739	0.935	0.933	0.748
	M	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table 7: Overall results on MCQ datasets.

		eBay	Amazon	Google	Schema	ACM-CCS	GeoNames	Glottolog	ICD-10-CM	OAE	NCBI
GPT-3.5	*A*	0.931	0.782	0.894	0.842	0.709	0.504	0.534	0.915	0.824	0.531
	M	0.013	0.067	0.025	0.040	0.044	0.053	0.254	0.023	0.038	0.209
GPT-4	*A*	0.947	0.849	0.932	0.907	0.790	0.695	0.683	0.964	0.900	0.701
	M	0.000	0.027	0.001	0.000	0.001	0.000	0.003	0.000	0.004	0.009
Claude-3	*A*	0.947	0.830	0.922	0.894	0.741	0.638	0.497	0.962	0.885	0.577
	M	0.000	0.043	0.003	0.020	0.022	0.004	0.401	0.004	0.032	0.286
Llama-2-7B	*A*	0.488	0.399	0.393	0.304	0.259	0.305	0.313	0.406	0.283	0.271
	M	0.000	0.005	0.002	0.008	0.003	0.004	0.089	0.004	0.035	0.002
Llama-2-13B	*A*	0.680	0.551	0.547	0.448	0.450	0.313	0.305	0.695	0.429	0.368
	M	0.003	0.026	0.037	0.070	0.006	0.016	0.072	0.015	0.088	0.001
Llama-2-70B	*A*	0.881	0.704	0.794	0.604	0.556	0.350	0.352	0.776	0.613	0.359
	M	0.000	0.004	0.000	0.004	0.001	0.004	0.015	0.003	0.002	0.000
Llama-3-8B	*A*	0.865	0.725	0.817	0.689	0.610	0.431	0.449	0.896	0.813	0.449
	M	0.000	0.017	0.006	0.014	0.006	0.000	0.101	0.001	0.001	0.003
Llama-3-70B	*A*	0.941	0.790	0.898	0.805	0.729	0.598	0.634	0.956	0.905	0.650
	M	0.000	0.021	0.001	0.010	0.001	0.000	0.011	0.000	0.002	0.003
Flan-T5-3B	*A*	0.898	0.756	0.878	0.799	0.664	0.455	0.506	0.836	0.777	0.524
	M	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Flan-T5-11B	*A*	0.904	0.805	0.902	0.841	0.696	0.634	0.583	0.877	0.817	0.550
	M	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Falcon-7B	*A*	0.261	0.259	0.275	0.233	0.240	0.256	0.260	0.254	0.262	0.255
	M	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.003	0.000	0.000
Falcon-40B	*A*	0.578	0.465	0.525	0.368	0.389	0.305	0.036	0.711	0.168	0.117
	M	0.168	0.332	0.269	0.494	0.198	0.215	0.960	0.059	0.593	0.804
Vicuna-7B	*A*	0.617	0.493	0.527	0.425	0.384	0.313	0.409	0.528	0.473	0.392
	M	0.000	0.001	0.000	0.000	0.000	0.008	0.017	0.011	0.166	0.005
Vicuna-13B	*A*	0.838	0.636	0.747	0.523	0.543	0.317	0.205	0.833	0.691	0.362
	M	0.013	0.100	0.061	0.137	0.079	0.146	0.664	0.045	0.196	0.375
Vicuna-33B	*A*	0.327	0.307	0.338	0.266	0.261	0.264	0.253	0.401	0.285	0.252
	M	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.009	0.000
Mistral	*A*	0.828	0.666	0.713	0.692	0.582	0.415	0.436	0.765	0.768	0.519
	M	0.000	0.014	0.006	0.006	0.003	0.000	0.179	0.008	0.004	0.044
Mixtral	*A*	0.924	0.768	0.876	0.775	0.707	0.537	0.611	0.923	0.797	0.634
	M	0.000	0.040	0.008	0.038	0.008	0.020	0.089	0.005	0.003	0.067
LLMs4OL	*A*	0.954	0.851	0.924	0.921	0.796	0.683	0.660	0.921	0.941	0.650
	M	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000