Are Large Language Models a Good Replacement of Taxonomies?

Yushi Sun[1], Hao Xin[1], Kai Sun[3], Yifan Ethan Xu[3], Xiao Yang[3], Xin Luna Dong[3], Nan Tang[1,2], Lei Chen[1,2]

[1]The Hong Kong University of Science and Technology,
[2]The Hong Kong University of Science and Technology (Guangzhou),
[3]Meta Reality Labs

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

香港科技大學(廣州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

Meta

Outline

- Background and Motivation
- Benchmark: TaxoGlimpse
- Experiment
- Discussion
- Summary

1. Background and Motivation

- Taxonomies provide a structured way to organize and categorize knowledge, which is indeed a kind of ``knowledge about knowledge" (meta-knowledge).
- Typically, nodes in taxonomies follow a tree-like structure and the relationships between nodes are depicted as hypernymy (Is-A) links.

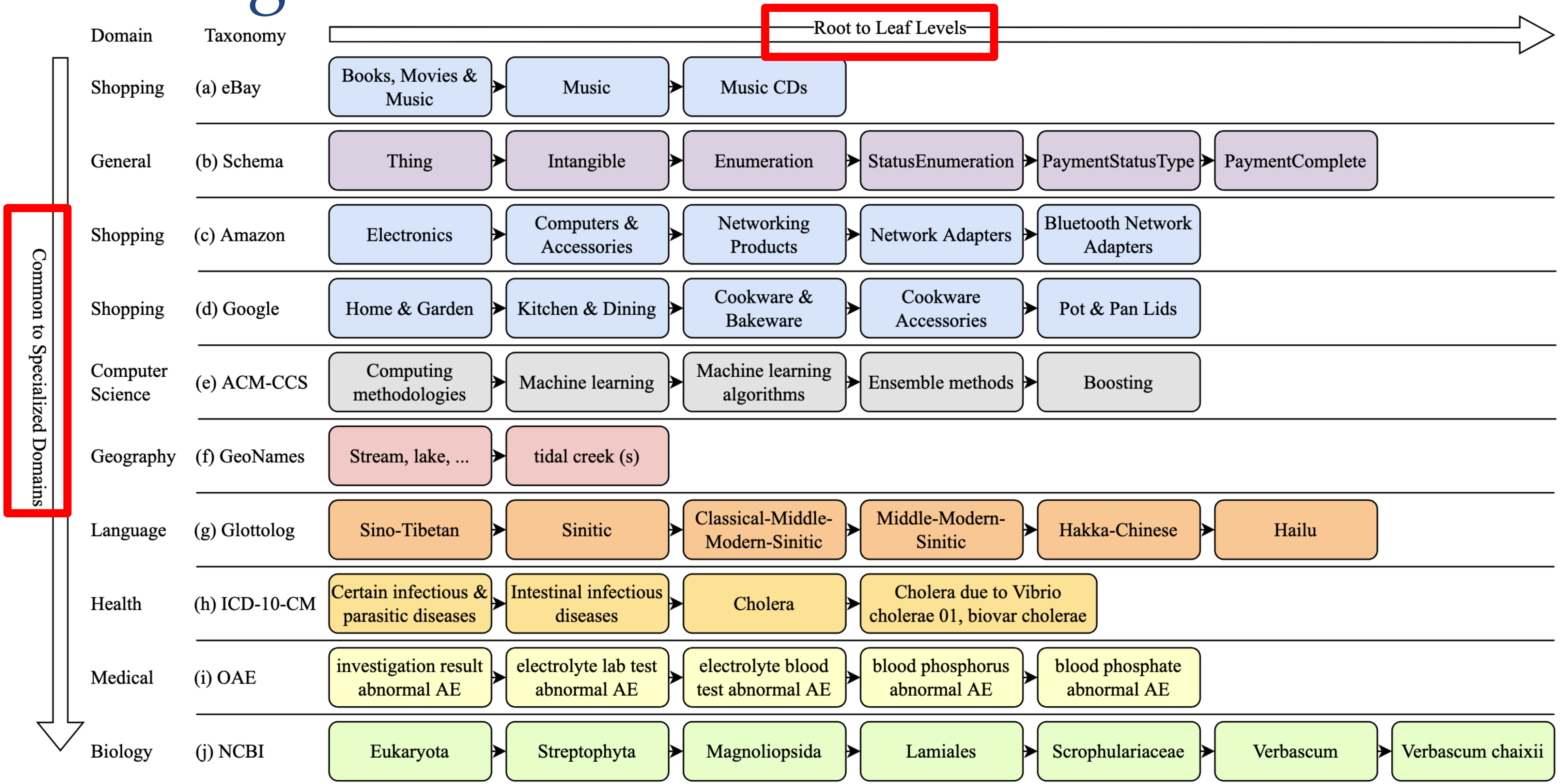
[1] Andreas, “Taxonomy: Tracing Its Greek Roots to Modern Biological Classification - U speak Greek,” *U speak Greek*, Dec. 25, 2023. https://uspeakgreek.com/science/biology/taxonomy-tracing-its-greek-roots-to-modern-biological-classification/ (accessed Aug. 18, 2024).

1. Background and Motivation

- Recently, we have witnessed the rapid advancements of large language models (LLMs) such as GPTs and Llamas. These LLMs have demonstrated impressive abilities in internalizing knowledge [2].
- Can LLMs internalize taxonomy structures?
- Are traditional taxonomies made obsolete by LLMs?

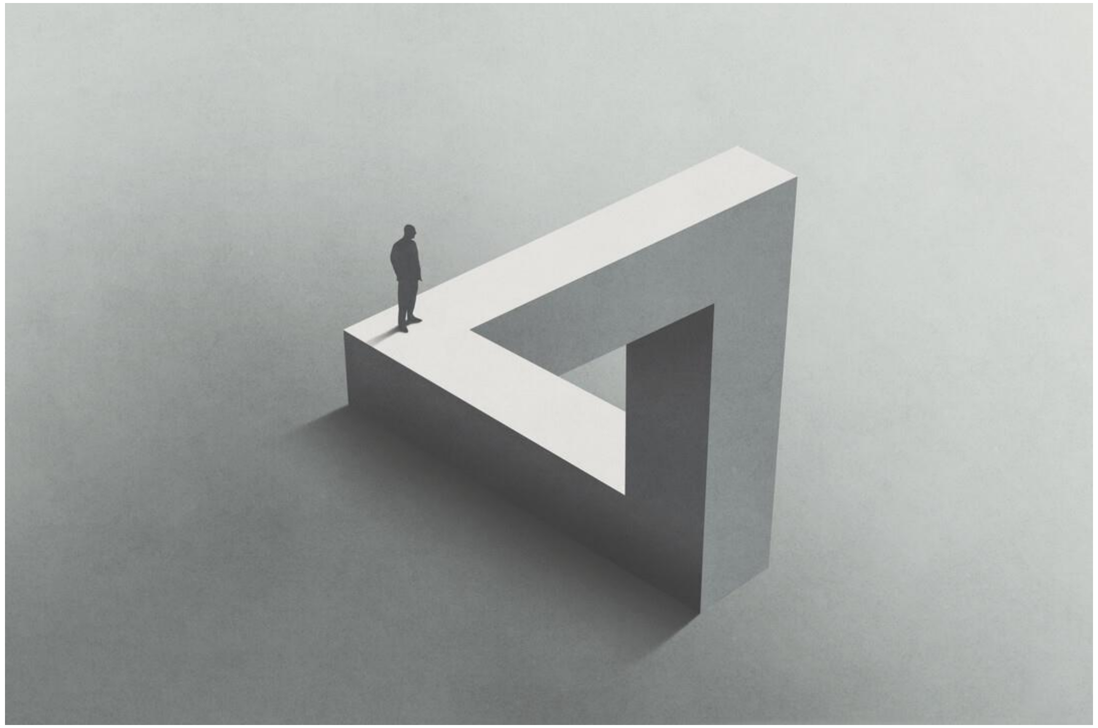
[2] K. Sun, Y. Xu, Hanwen Zha, Y. Liu, and Xin Luna Dong, "Head-to-Tail: How Knowledgeable are Large Language Models (LLMs)? A.K.A. Will LLMs Replace Knowledge Graphs?," Jan. 2024, doi: https://doi.org/10.18653/v1/2024.naacl-long.18.

1. Background and Motivation



1. Background and Motivation

- 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 meaningful research problems/application domains that boost the reasoning of LLMs.

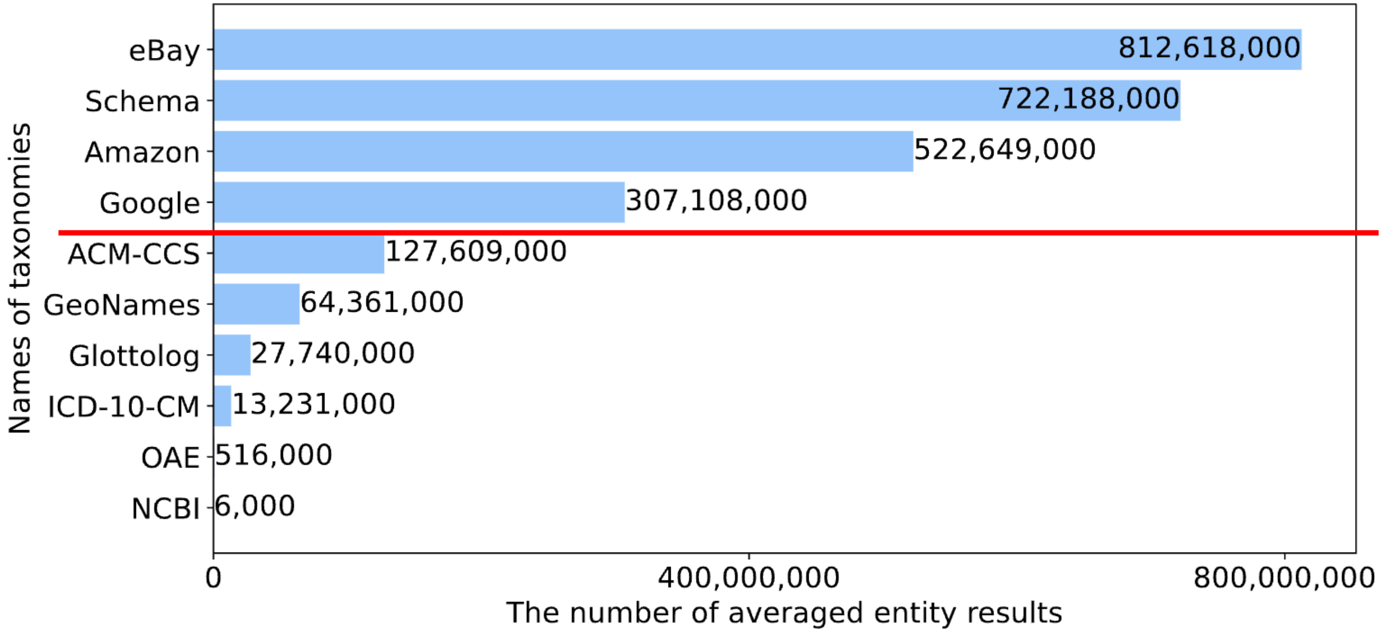


Outline

- Background and Motivation
- Benchmark: TaxoGlimpse
- Experiment
- Discussion
- Summary

2. Benchmark

- Taxonomies: 10 taxonomies on 8 domains:
- Common taxonomies:
  - Shopping domain: eBay, Amazon, Google
  - General domain: Schema.org
- Specialized taxonomies:
  - CS domain: ACM-CCS
  - Geography domain: GeoNames
  - Language domain: Glottolog
  - Health domain: ICD-10-CM
  - Medical domain: OAE
  - Biology domain: NCBI



2. Benchmark

- Design of questions: adopt simple True/False question

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)

2. Benchmark

- Generation of question set

	eBay	Amazon	Google	Schema	ACM-CCS	GeoNames	Glottolog	ICD-10-CM	OAE	NCBI
Level 1-root	176	438	258	34	138	492	500	222	638	344
Level 2-1	430	700	597	276	450	n/a	564	550	700	439
Level 3-2	n/a	748	653	394	567	n/a	584	690	670	636
Level 4-3	n/a	758	626	410	370	n/a	600	n/a	572	741
Level 5-4	n/a	n/a	n/a	320	n/a	n/a	732	n/a	n/a	766
Level 6-5	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	770
Total	606	2644	2134	1434	1525	492	2980	1462	2580	3696

2. Benchmark

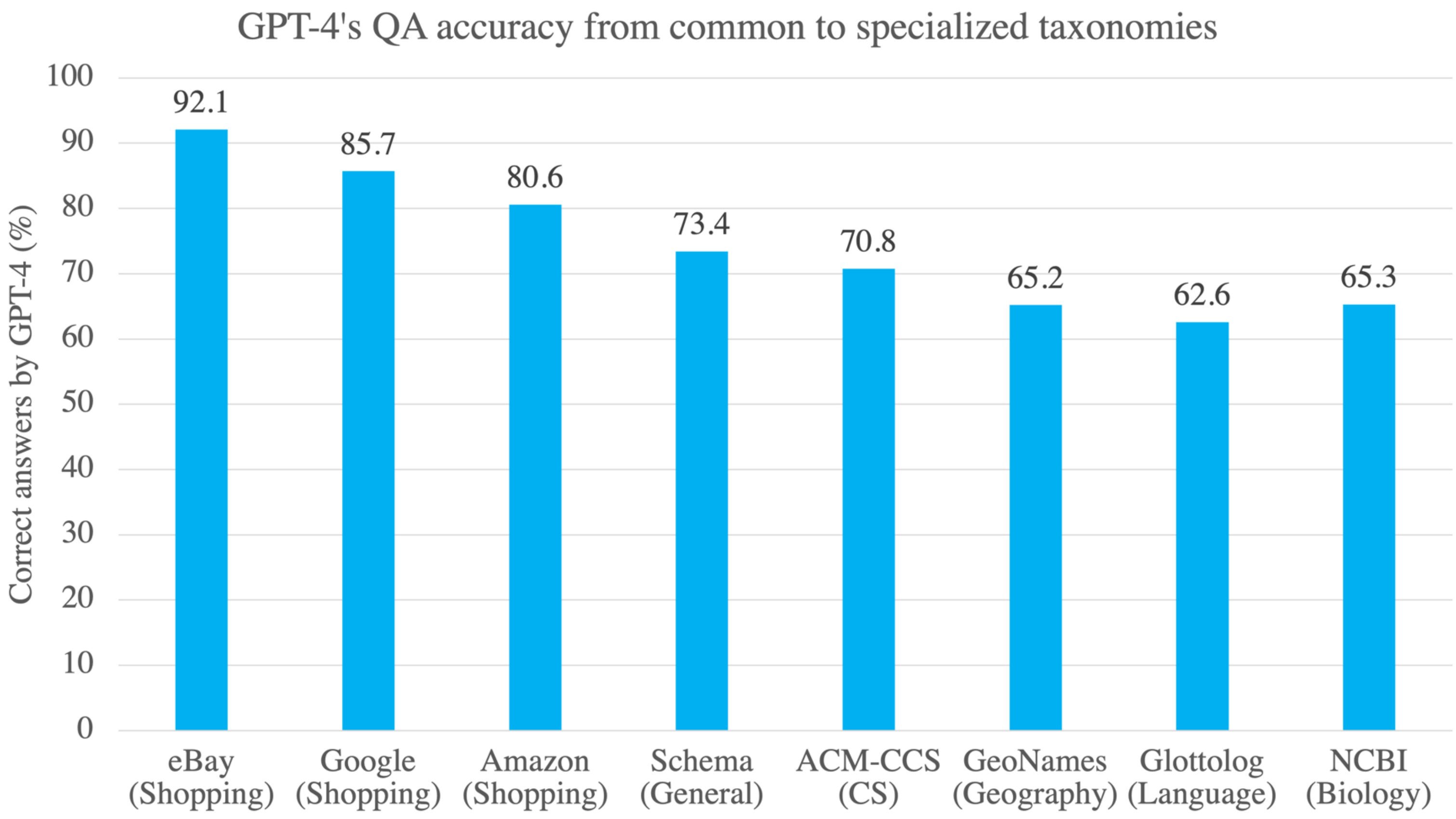
- LLMs considered:
  - Open-source:
    - Llama-2s: 7B, 13B, 70B
    - Llama-3s: 8B, 70B
    - Flan-T5s: 3B, 11B
    - Falcons: 7B, 40B
    - Vicunas: 7B, 13B, 33B
    - Mistrals: 7B, 8*7B
- Closed-source:
  - GPTs: GPT 3.5, GPT 4
  - Claude-3-Opus
- Fine-tuned:
  - LLMs4OL

Outline

- Background and Motivation
- Benchmark: TaxoGlimpse
- **Experiment**
- Discussion
- Summary

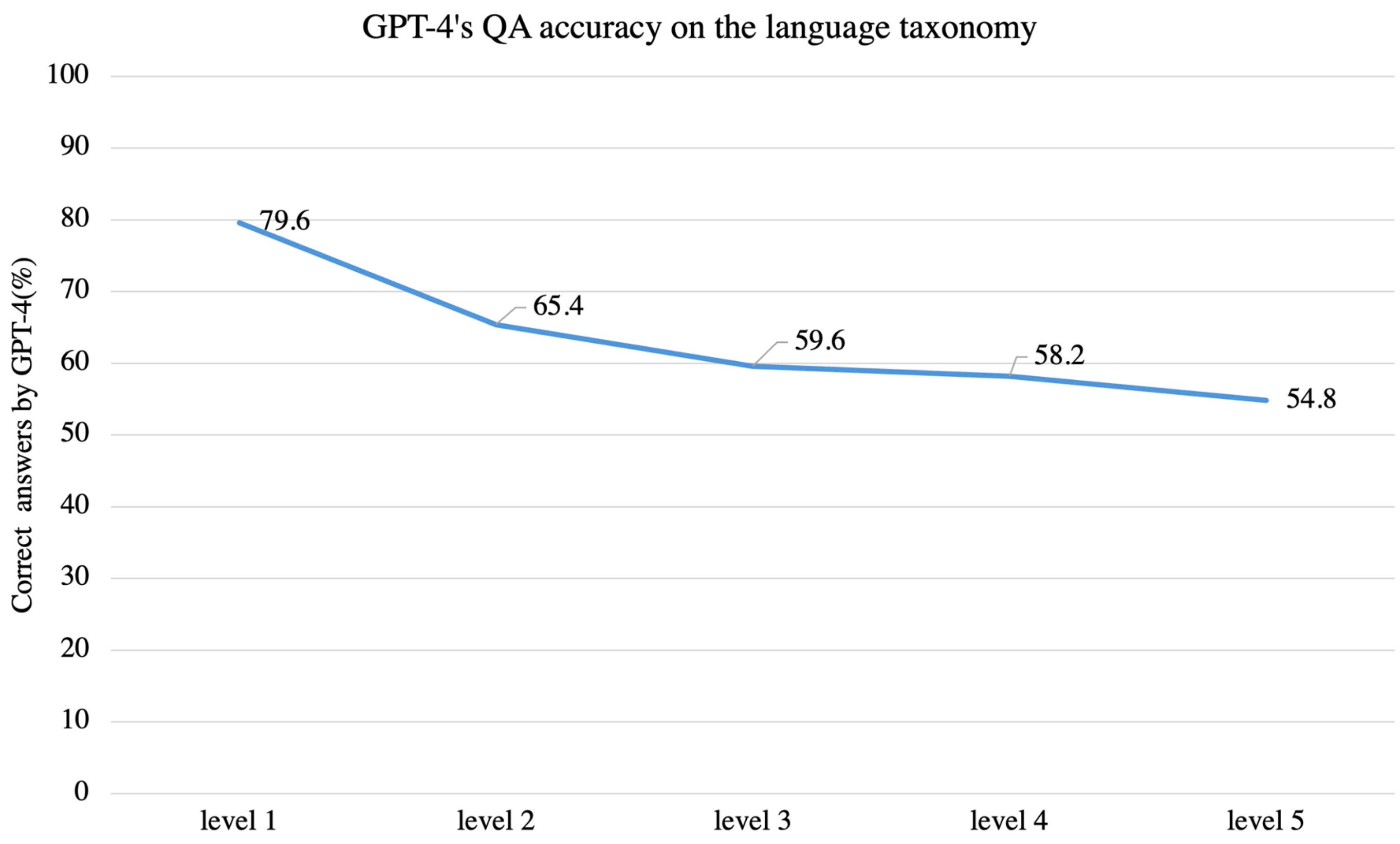
3. Experiment

- RQ1: How reliable are LLMs for discovering hierarchical structures in different taxonomies?
- The best LLMs perform well on common taxonomies (e.g., eBay, with over 90% accuracy); however, the performance downgrades on specialized taxonomies to around 60%.



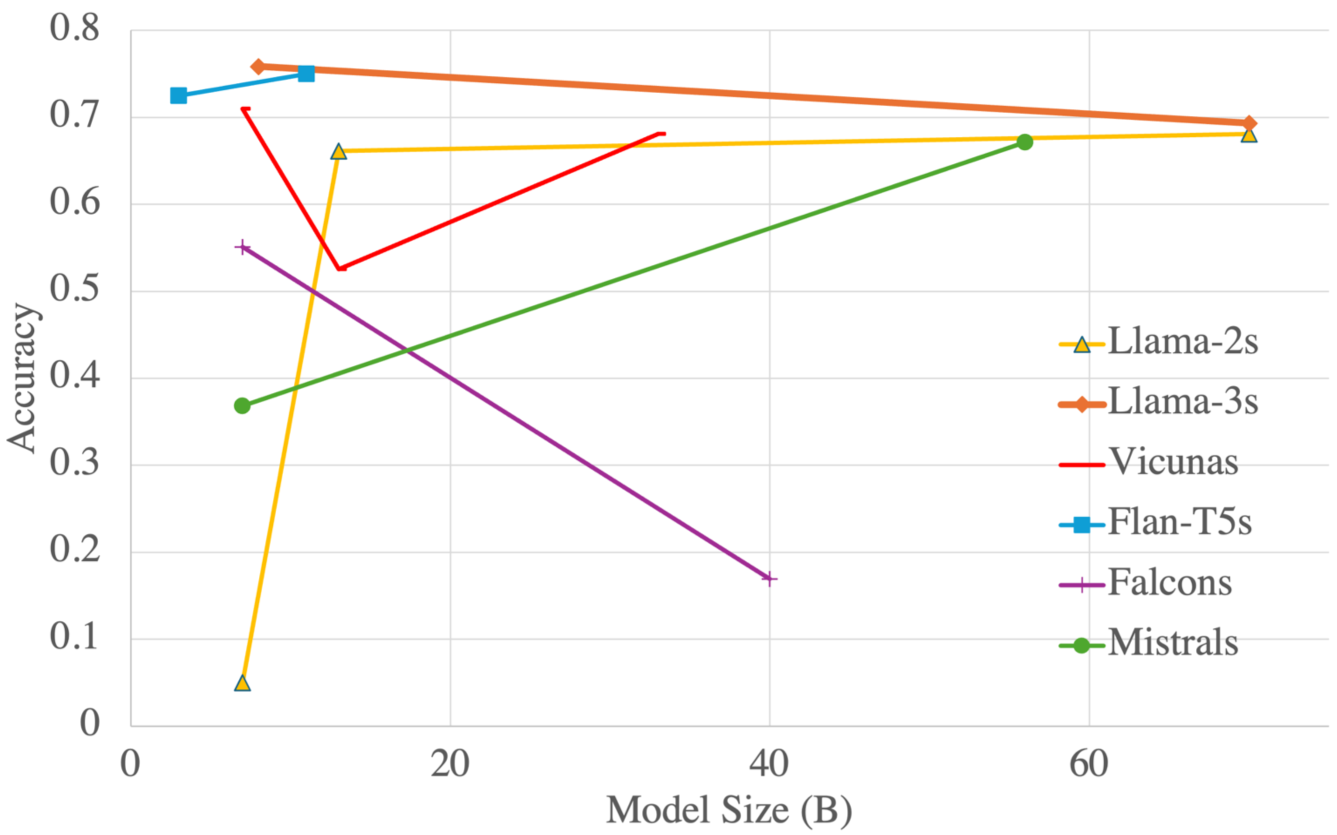
3. Experiment

- RQ2: Do LLMs perform equally well among different levels of taxonomies?
- LLMs roughly achieve progressively worse performance from root to leaf in most taxonomies (e.g., drops by relatively over 30% on Language taxonomy).



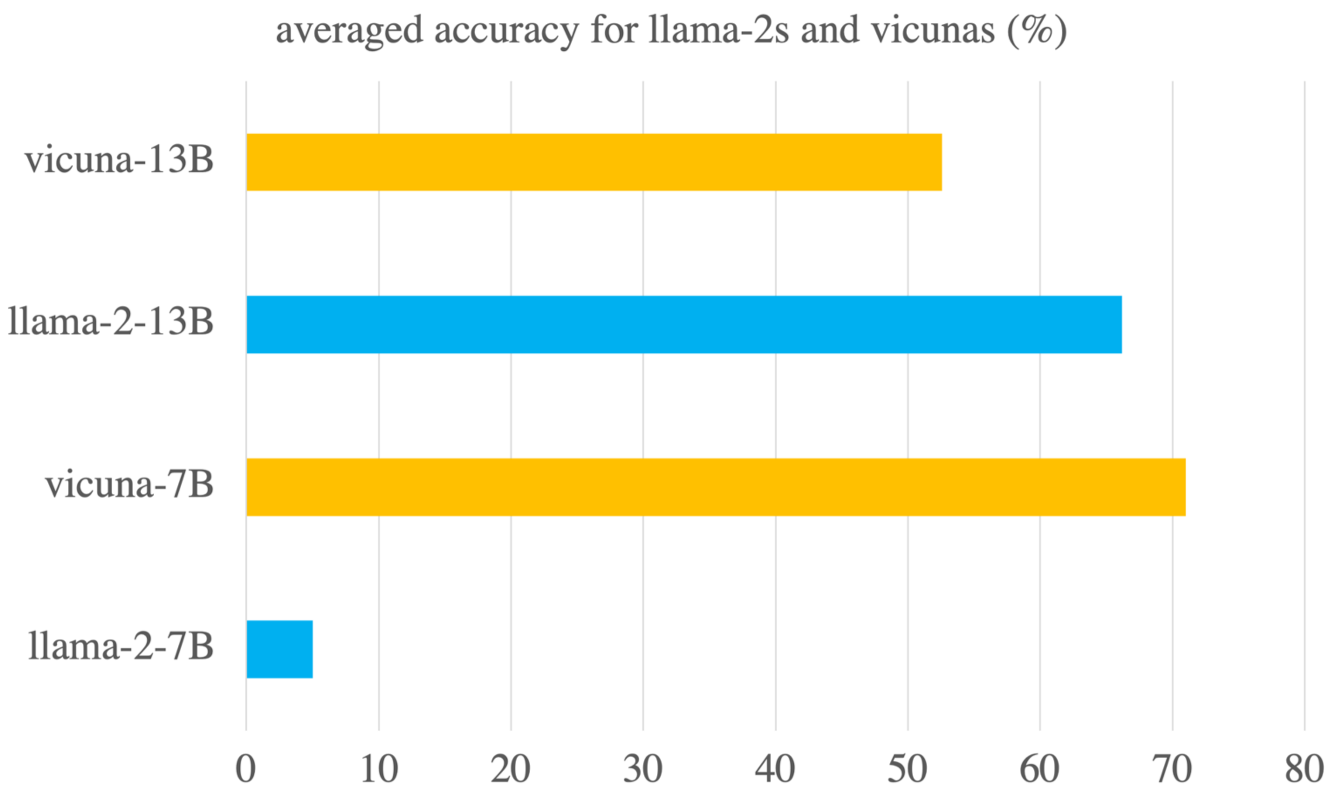
3. Experiment

- RQ3: Do normal methods that improve LLMs increase the accuracy?
  - RD3.1: Can we improve LLMs' performance by increasing the sizes of the LLMs used?
  - The increase in sizes of LLMs may not lead to an increase in performance.



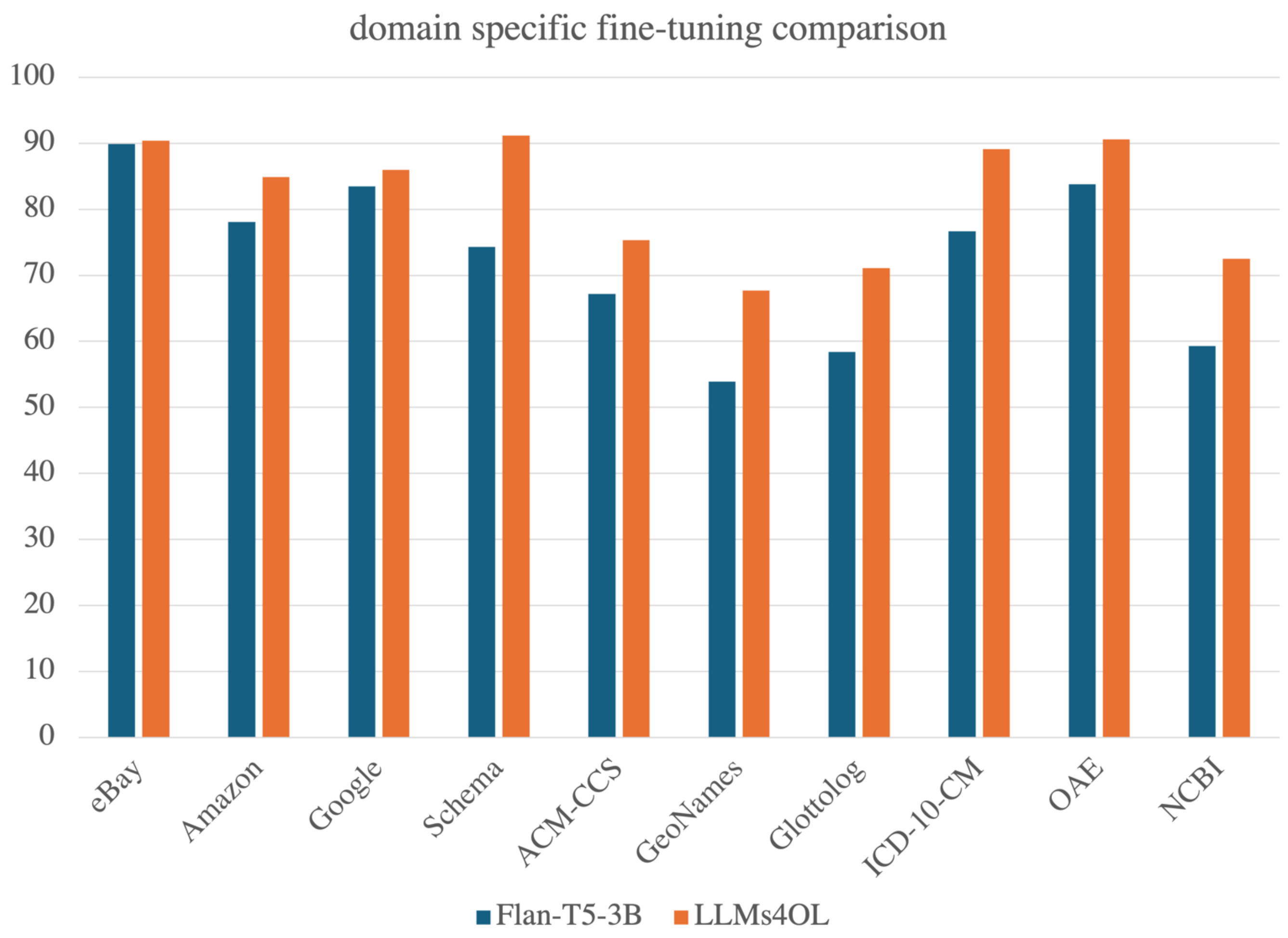
3. Experiment

- RQ3: Do normal methods that improve LLMs increase the accuracy?
  - RD3.2: Can we improve LLMs' performance by adopting domain-agnostic fine-tuning?
  - The adoption of domain-agnostic fine-tuning of LLMs may not lead to an increase in performance.



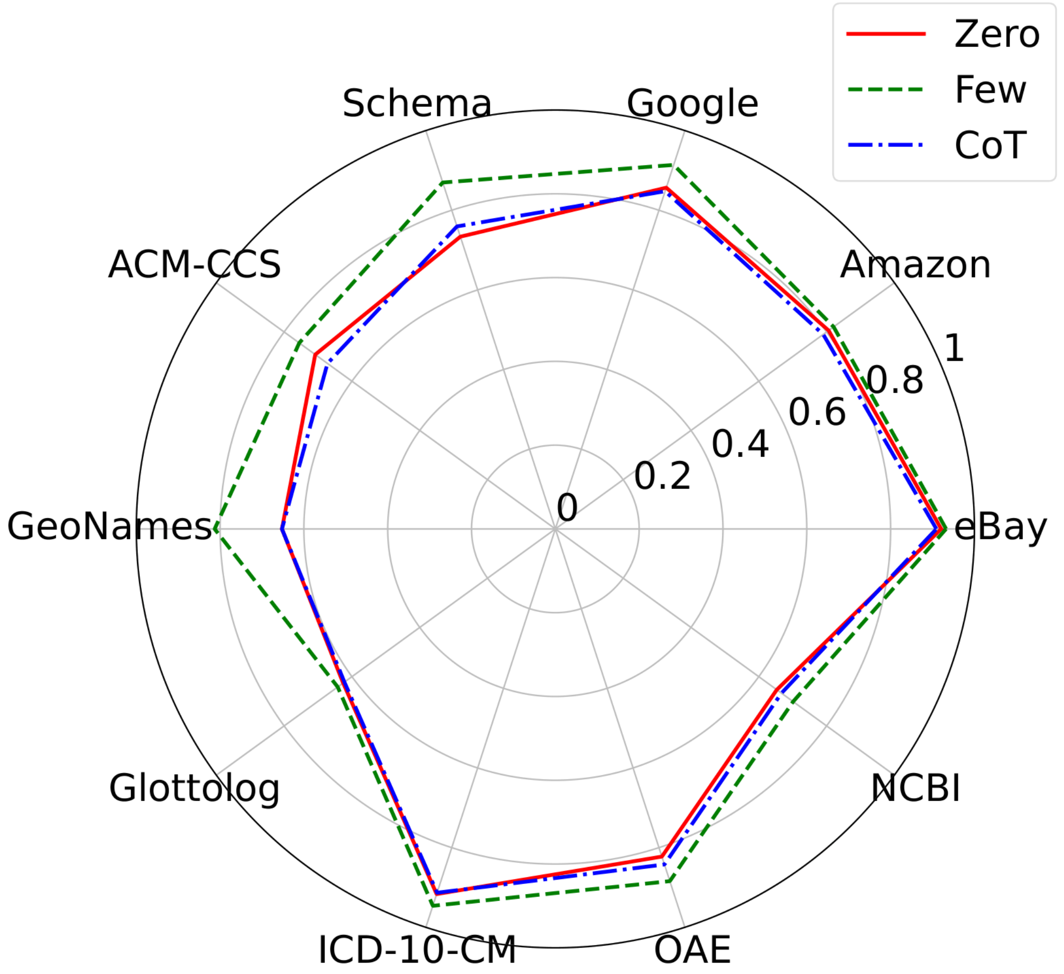
3. Experiment

- RQ3: Do normal methods that improve LLMs increase the accuracy?
  - RD3.3: Can we improve LLMs' performance by adopting domain-specific instruction tuning?
  - The adoption of domain-specific instruction tuning leads to stable and significant improvements.



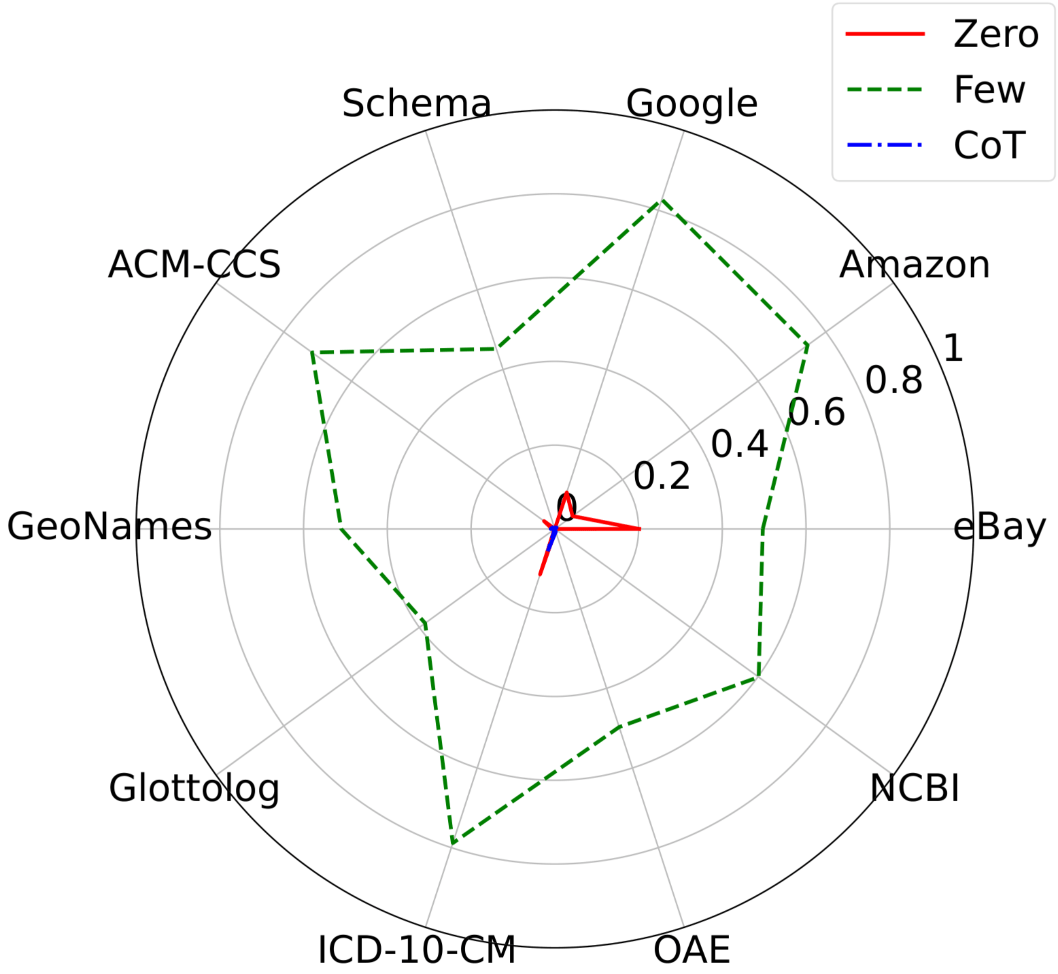
3. Experiment

- RQ4: 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. The main effect of prompting settings is to influence the miss rates instead of the accuracy of LLMs.



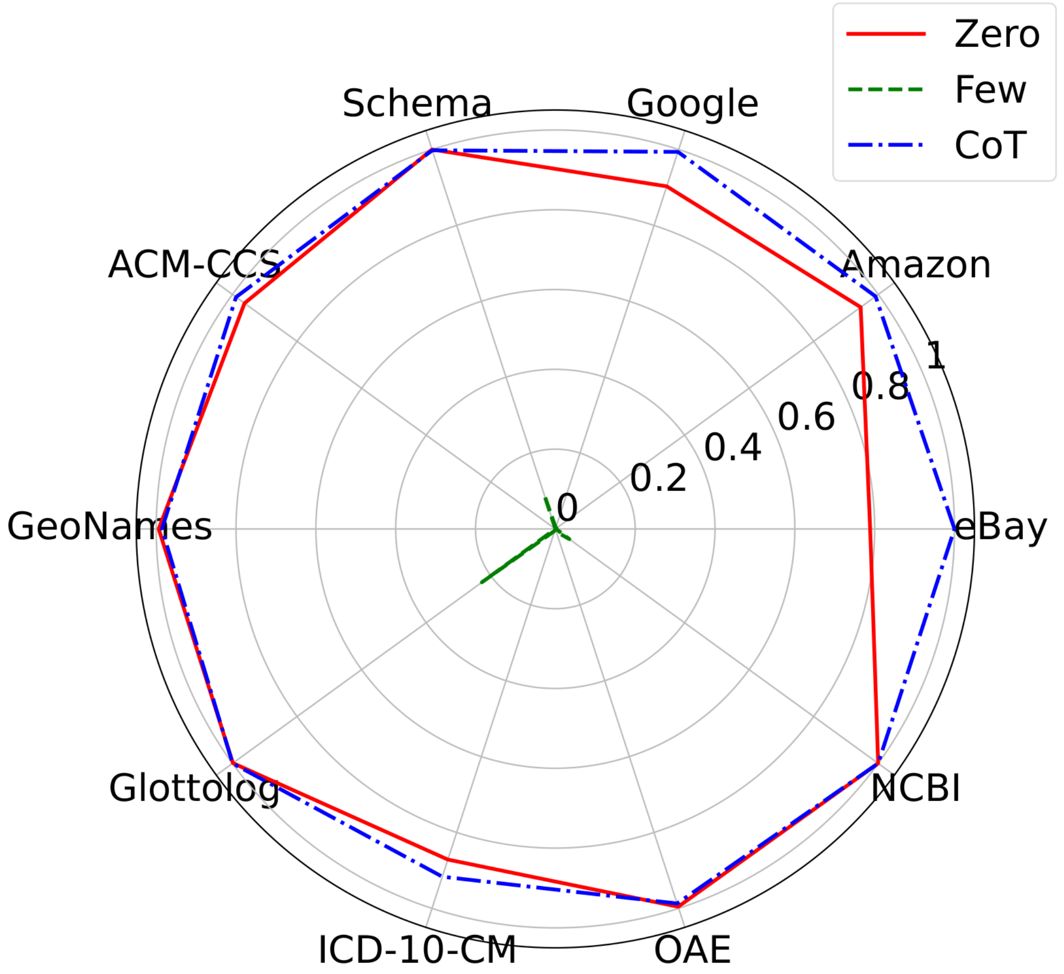
3. Experiment

- RQ4: 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. The main effect of prompting settings is to influence the miss rates instead of the accuracy of LLMs.



3. Experiment

- RQ4: 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. The main effect of prompting settings is to influence the miss rates instead of the accuracy of LLMs.



Outline

- Background and Motivation
- Benchmark: TaxoGlimpse
- Experiment
- Discussion
- Summary

4. Discussion

- The future of taxonomies:
  - Common taxonomies: Such as shopping, should be encoded inside the LLMs (a case study provided in our paper).
    - In some use cases such as relation display and visualization, the traditional taxonomic structure near root levels may still be needed. The majority of the use cases (such as entity searching and knowledge reasoning) in common taxonomies can be well handled by LLMs.
  - Specialized taxonomies: Such as language, 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.

Outline

5. Summary

- 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.
- Four highly concerned research questions were proposed and resolved and we provided valuable insights into future research.
- Our comprehensive evaluation shows that LLMs present 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 to create novel neural-symbolic taxonomies that have the best of both worlds.

Thank you for your listening!

The full paper of TaxoGlimpse:

My personal website: