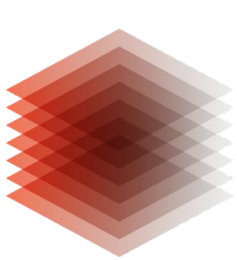
LEIBNIZ-INFORMATIONSZENTRUM
TECHNIK UND NATURWISSENSCHAFTEN
UNIVERSITÄTSBIBLIOTHEK



TIB

# LLMs4OL: Large Language Models for Ontology Learning

**Hamed Babaei Giglou**

**6 March 2024**
**Ontology Summit 2024**

Leibniz Association

# Ontology Learning

- **Ontology Learning (OL):** involves automatically identifying terms, types, relations, and potentially axioms from textual information to construct an ontology, with ultimate goal of knowledge acquisition and representation.

- A quick review of traditional OL approaches in the field:
  - lexico-syntactic pattern mining
  - clustering
  - statistical approach

- Aim of existing approaches:
  - Learn new concepts/relations or enrich ontology with new concepts
  - Elicit or prune or extend a taxonomy
  - Build taxonomy

D'Souza, J. (2023). Ontology Learning from Text. Open Research Knowledge Graph. https://doi.org/10.48366/R586328, ORKG: https://orkg.org/comparison/R586328/

# Paradigm Shift in OL




**Recent advancement in NLP through LLMs offer a paradigm shift in OL methods**

# Goal

**Within LLMs4OL we hypothesize:**

- **Can LLMs effectively apply their language pattern capturing capability to OL, which involves automatically extracting and structuring knowledge from natural language text?**

**Experimental Considerations:**

- Do the characteristics of LLMs justify ontology learning?
- Which LLMs to test for the LLMs4OL task hypothesis?

# Primitive Ontology

## Maedche et al.:

1. A set of strings that describe terminological lexical entries ***L*** for conceptual **types**.

2. A set of conceptual types ***T***.

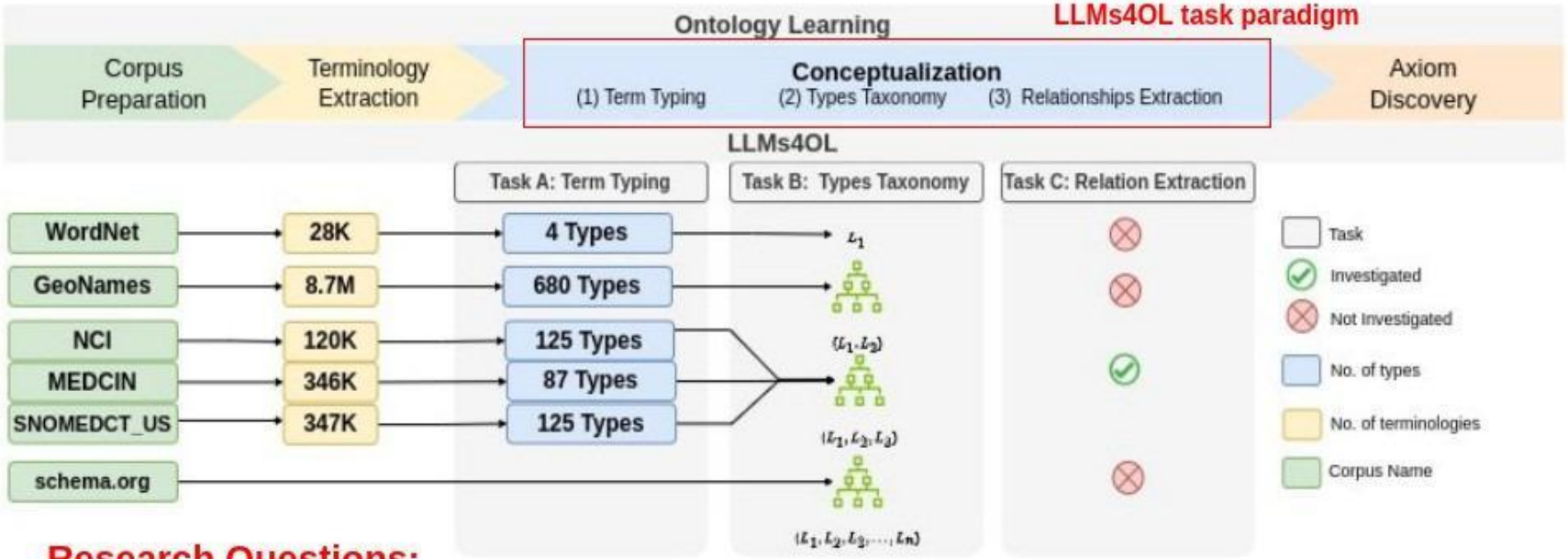
3. A **taxonomy of types** in a hierarchy **H_T.**

4. A set of **non-taxonomic relations *R*** described by their domain and range restrictions arranged in a heterarchy of relations **H_R.**

5. A set of **axioms *A*** that describe additional constraints on the ontology and make implicit facts explicit.

Maedche, A., Staab, S.: Ontology learning for the semantic web. IEEE Intell. Syst. 16(2), 72–79 (2001)

# LLMs4OL task Paradigm

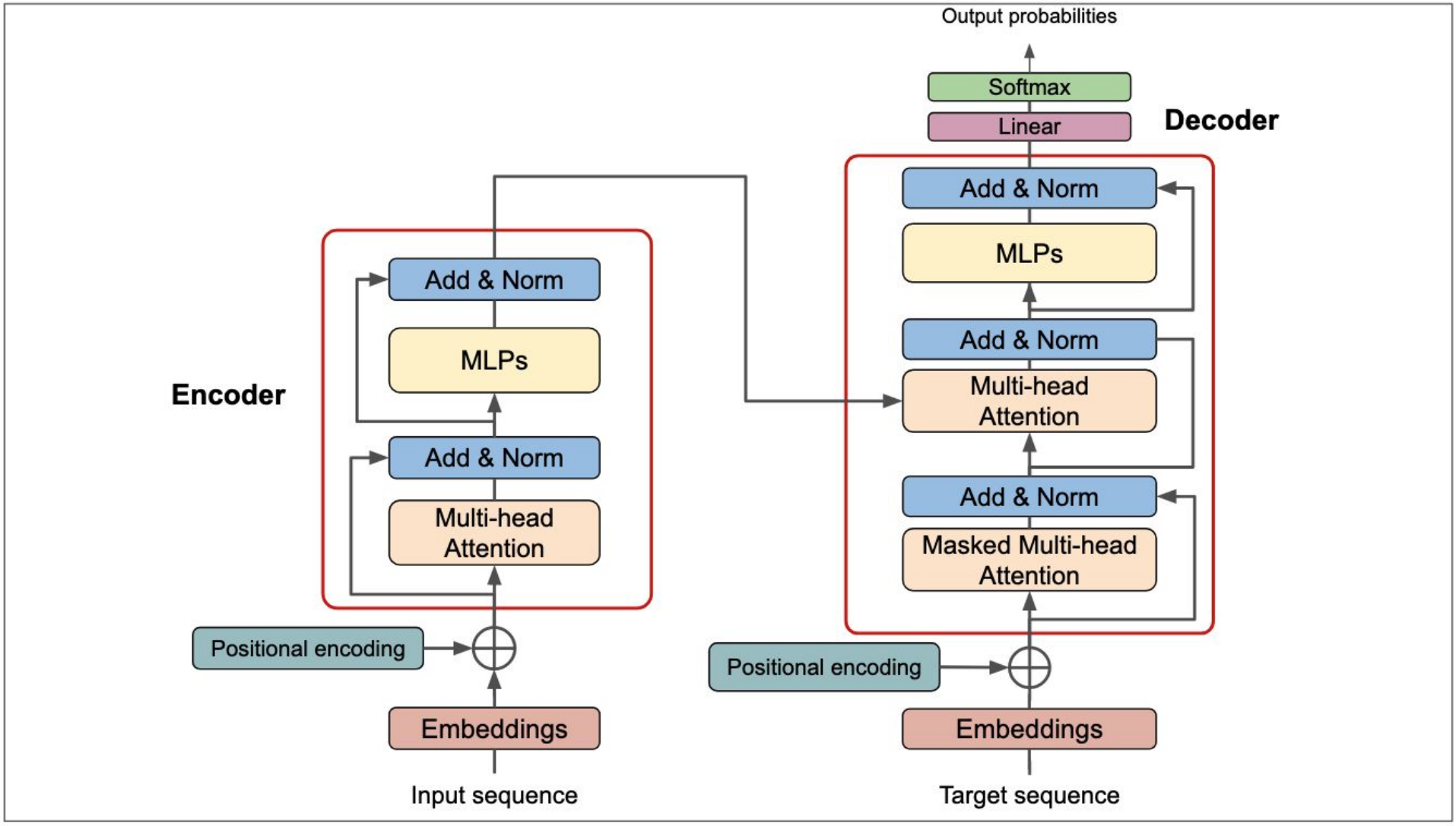



**Research Questions:**

- TASK A: **Term Typing** – How effective are LLMs for automated type discovery to construct an ontology? – RQ1
- TASK B: **Type Taxonomy Discovery** – How effective are LLMs to recognize a type taxonomy i.e. the "is-a" hierarchy between types? – RQ2
- TASK C: **Type Non-Taxonomic Relation Extraction** – How effective are LLMs to discover non-taxonomic relations between types? – RQ3

# Large Language Models

## Encoder, Decoder, Encoder-Decoder



# Chosen LLMs – 11 Models

- **Encoder-only**
  - BERT-Large (340M)
  - PubMedBERT (340M)
- **Decoder-only**
  - BLOOM-1B7 (1.7B)
  - BLOOM-3B (3B)
  - LLaMA (7B)
  - GPT-3 (175B)
  - GPT-3.5 (174B) – ChatGPT
  - GPT-4 (>1T)
- **Encoder-Decoder**
  - BART-Large (400M)
  - Flan-T5-Large (780M)
  - Flan-T5-XL (3B)

# Prompt Engineering

**We aim to use:**

- Standard prompting (Wei et al.) as it represents a fundamental approach for instructing LLMs
  - Within this prompting technique we will consider LLMs as knowledge base (Petroni et al.), to see how they can be reusable for OL.

**We aim to NOT use complex prompting, Why?**

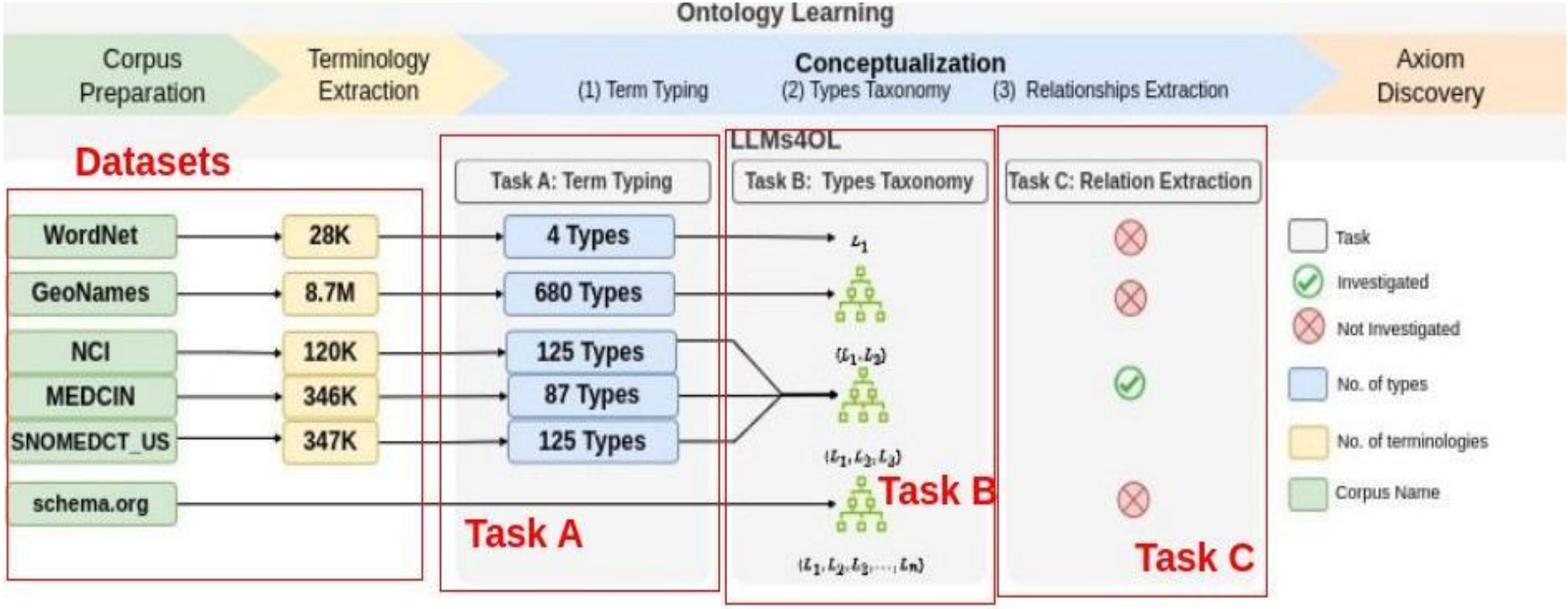
- Our current focus is on the initial exploration of the task to identify the areas where we need further improvement.
- Understand how much we have accomplished so far before delving into more complex techniques like CoT, ToT, and decomposed prompting.

**So:**

Once we have a clearer picture of the model's capabilities and limitations, we can then consider other prompt engineering approaches by formulating OL as a stepwise reasoning task.

Petroni, F., et al.: Language models as knowledge bases? In: EMNLP-IJCNLP. ACL (2019).
Wei, J., et al.: Finetuned language models are zero-shot learners. In: International Conference on Learning Representations (2022). https://openreview.net/forum?id=gEZrGCozdqRReturn.

# Empirical Validation Scenario



# Ontological Sources – Task A Datasets




- 5 sources, 3 domains: Lexicosemantics (WordNet) – Geography (GeoNames) – Biomedicine (UMLS subontologies – NCI, MEDICIN, SNOMEDCT_US)
- Dataset stats:

| | WordNet | GeoNames | NCI | MEDICIN | SNOMEDCT_US |
|---|---|---|---|---|---|
| *Train Set Size* | 40,559 | 8,078,865 | 96,177 | 277,028 | 278,374 |
| *Test Set Size* | 9,470 | 702,510 | 24,045 | 69,258 | 69,594 |
| *Types* | 4 | 680 | 125 | 87 | 125 |

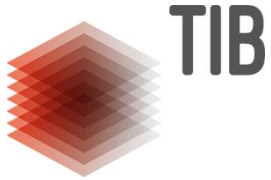
TASK A: Term Typing – Generalized types are discovered for a lexical term.

$f^{A}_{cloze-prompt}(L) := [S?].\ [L]\ [P_{domain}]\ is\ a\ [MASK].$

$f^{A}_{prefix-prompt}(L) = [instruction] + f^{A}_{cloze-prompt}(L)$

1. the company manufactures plastic chairs. manufacture POS is a [MASK]. → verb
2. Perform a sentence completion on the following sentence:
   Sentence: Roc Mele is a place in Andorra. Roc Mele is a → mountain
3. primary adrenal insufficiency in medicine is a ? disease / syndrome

# Ontological Sources – Task B Datasets (1/2)



- 3 sources, 3 domains: Geography (GeoNames) – Biomedicine (UMLS)–Web content types (schema.org)
- Dataset Construction Method:
  1. $\forall(a \in T_n, b \in T_{n+1}) \longmapsto aRb \wedge b\neg Ra$
  2. $\forall(a \in T_n, b \in T_{n+1}, c \in T_{n+2}); aRb \wedge bRc \longmapsto aRc$
  3. $\forall(a \in T_n, b \in T_{n+1}, c \in T_{n+2}); c\neg Rb \wedge b\neg Ra \longmapsto c\neg Ra$
- Dataset stats:

| | **GeoNames** | **UMLS** | **schema.org** |
|---|---|---|---|
| Types | 689 | 127 | 797 |
| Levels | 2 | 3 | 6 |
| Positive/Negative Samples | 680/680 | 254/254 | 2,670/2,670 |
| *Train/Test split* | 272/1,088 | 101/407 | 1,086/4,727 |

# Ontological Sources – Task B Datasets (2/2)

TASK B: Type Taxonomy Discovery – Taxonomic hierarchy between pairs of types are discovered.

$f^{B}_{c-prompt}(a, b) := [a|b]$ *is* $[P_{hierarchy}]$ *of* $[b|a]$. *This statement is* $[MASK]$.

$f^{B}_{prefix-prompt}(a, b) = [instruction] + f^{B}_{cloze-prompt}(a, b)$

1. Reservation is the superclass of Bus Reservation. This statement is a {"mask"}. → True
2. Reservation is a subclass of Bus Reservation. This statement is a {"mask"} . → False
3. Fungus is the parent class of Infectious Agent Class. This statement is a {"mask"} . → False
4. Landmarks or Historical Buildings is a child class of Place. This statement is a {"mask"} . → True

# Ontological Sources – Task C Datasets

- 1 source, 1 domain: Biomedicine (UMLS)
- UMLS reports 53 non-taxonomic relations across its 127 term types.
- Dataset stats:

| Task | UMLS |
|---|---|
| Non-Taxonomic Relations | 53 |
| Positive/Negative Samples | 5,641/1,896 |
| *Train/Test Split* | 1,507/6,030 |

- TASK C: Type Non-Taxonomic Relation Extraction – Non-taxonomic semantic heterarchical relations between types are discovered.

$f^{C}_{cloze-prompt}(h, r, t) := [h]\ is\ [r]\ [t].\ This\ statement\ is\ [MASK].$

$f^{C}_{prefix-prompt}(h, r, t) = [instruction] + f^{C}_{cloze-prompt}(h, r, t)$

1. Virus is interacts with Human. This statement is a {"mask"} . → True

2. Identify whether the following statement is true or false:
   Statement: Behavior is manifestation of Organism Function. → True

# Zero-Shot Testing – Summary of Results

RQ1 – RQ2 – RQ3

| Task | Dataset | BERT-Large | PubMedBERT | BART-Large | Flan-T5-Large | Flan-T5-XL | BLOOM-1b7 | BLOOM-3b | GPT-3 | GPT-3.5 | LLaMA-7B | GPT-4 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | WordNet | 27.9 | - | 2.2 | 31.3 | 52.2 | 79.2 | 79.1 | 37.9 | **91.7** | 81.4 | 90.1 |
| | GeoNames | 38.3 | - | 23.2 | 13.2 | 33.8 | 28.5 | 28.8 | 22.4 | 35.0 | 29.5 | **43.3** |
| | NCI | 11.1 | 5.9 | 9.9 | 9.0 | 9.8 | 12.4 | 15.6 | 12.7 | 14.7 | 7.7 | **16.1** |
| | SNOMEDCT | 21.1 | 28.5 | 19.8 | 24.3 | 31.6 | 37.0 | **37.7** | 24.4 | 25.0 | 13.8 | 27.8 |
| | MEDCIN | 8.7 | 15.6 | 12.7 | 13.0 | 18.5 | 28.8 | **29.8** | 25.7 | 23.9 | 4.9 | 23.7 |
| B | GeoNames | 54.5 | - | 55.4 | 59.6 | 52.4 | 36.7 | 48.3 | 53.2 | **67.8** | 33.5 | 55.4 |
| | UMLS | 48.2 | 33.7 | 49.9 | 55.3 | 64.3 | 38.3 | 37.5 | 51.6 | 70.4 | 32.3 | **78.1** |
| | schema.org | 44.1 | - | 52.9 | 54.8 | 42.7 | 48.6 | 51.3 | 51.0 | **74.4** | 33.8 | 74.3 |
| C | UMLS | 40.1 | 42.7 | 42.4 | 46.0 | **49.5** | 43.1 | 42.7 | 38.8 | 37.5 | 20.3 | 41.3 |

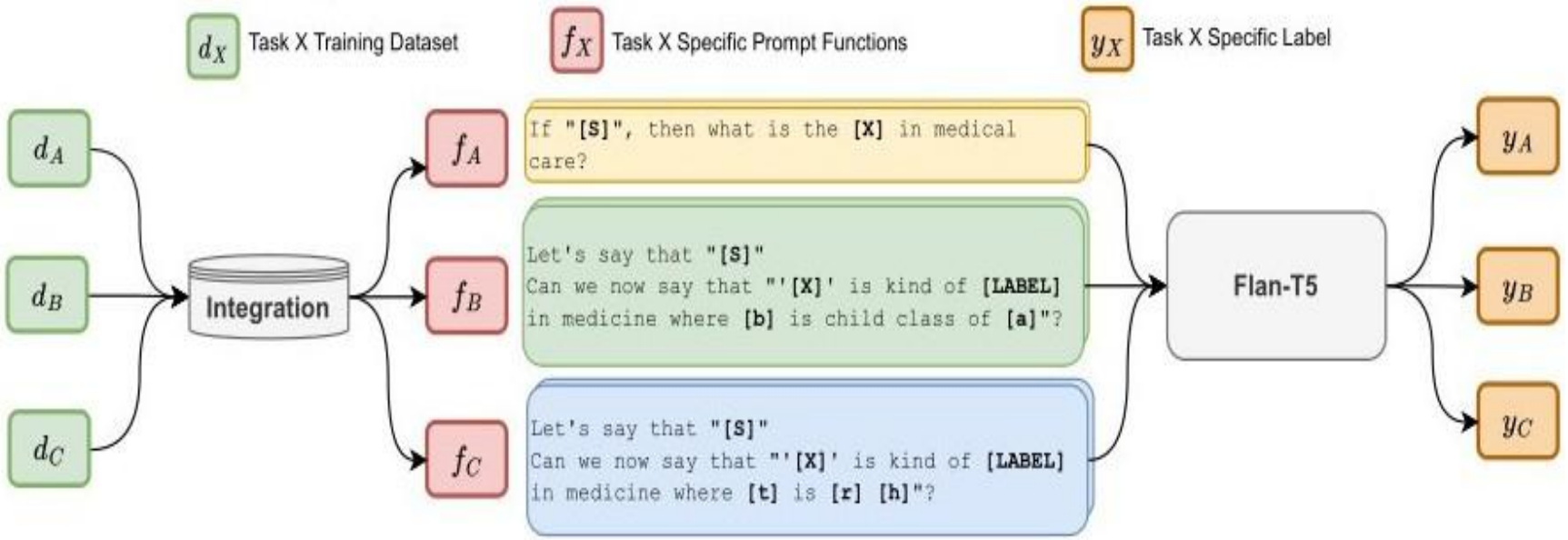
1. Task A - WordNet obtains high performance (simple type space)
2. Task A - GeoName obtains promising result (more fine-grained type space)
3. Task B - A higher number of parameters are successful
4. Task C - More challenging task fails to cope with LLMs
5. PubMedBERT fails to obtain better performance on biomedical datasets

# Zero-Shot Testing – Findings

- With an increase in the number of parameters, we can expect an improvement in OL
- LLMs are showing good capabilities toward OL.
- Zero-shot test results indicate that while LLMs seems to be promising for OL, they would need task-specific finetuning to be a practically viable solution.

# Finetuning LLM – Flan-T5

- Adapted instruction tuning workflow as described in Flan-T5
- Instructions are instantiated from a small selection of samples using tasks train set for each knowledge source

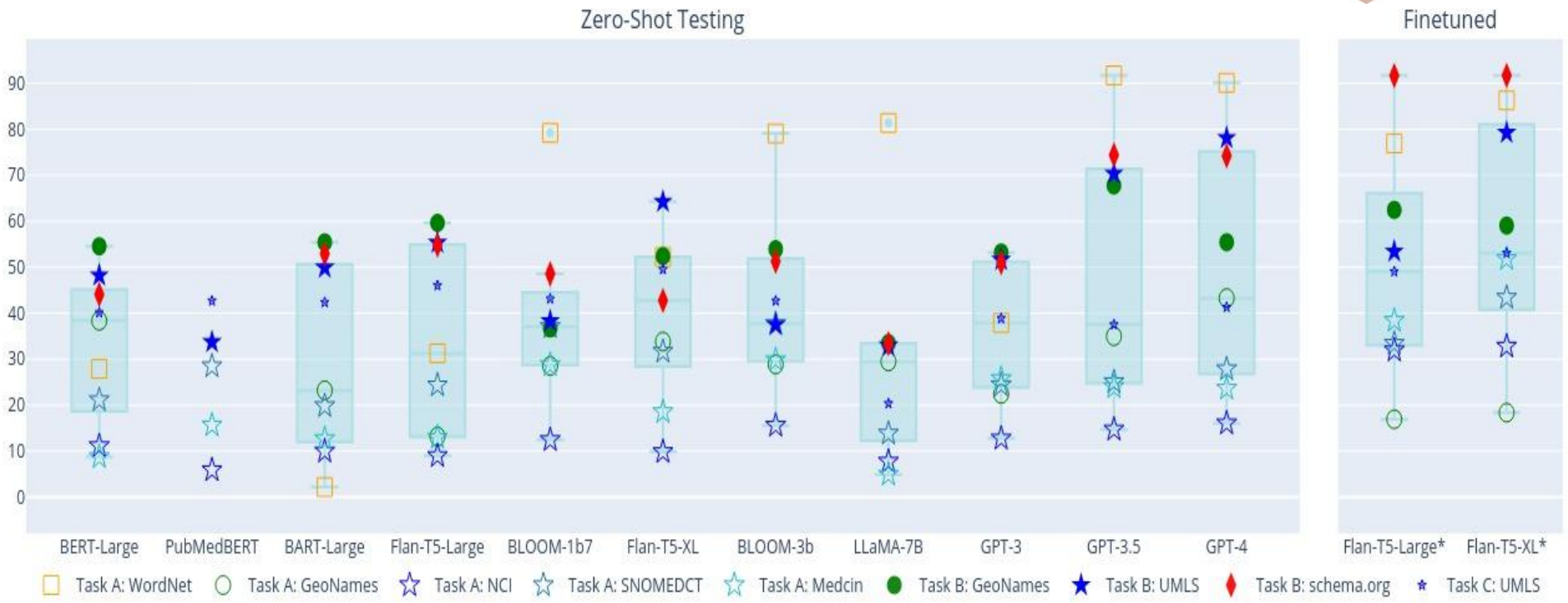


# Finetuning LLM – Results

| Task | Dataset | Zero-Shot Test | Flan-T5-Large | Flan-T5-XL |
|---|---|---|---|---|
| | *WordNet* | **91.7 (GPT-3.5)** | 76.9 | 86.3 |
| | *GeoNames* | **43.3 (GPT-4)** | 16.9 | 18.4 |
| *A* | *NCI* | 16.1 (GPT-4) | 31.9 | **32.8** |
| | *SNOMEDCT_US* | 37.7 (BLOOM-3b) | 33.4 | **43.4** |
| | *MEDCIN* | 29.8 (BLOOM-3b) | 38.4 | **51.8** |
| | *GeoNames* | **67.8 (GPT-3.5)** | 62.5 | 59.1 |
| *B* | *UMLS* | 78.1 (GPT-4) | 53.4 | **79.3** |
| | *Schema.Org* | 74.4 (GPT-3.5) | 91.7 | **91.7** |
| *C* | *UMLS* | 49.5 (Flan-T5-XL) | 49.1 | **53.1** |

- 25% improvement on average from zero-shot to finetuning on Flan-T5
- To obtain better performance we don't need to have larger LLM!

# Further Analysis




## Limitations

- Finetuning – low number of training shots in finetuning but it still doing a good job here!
- LLaMA-7B overall low performance – Absence of few-shot testing
- NCI low performance in most of the evaluations – Heuristics-based generated answer set mapping to the gold standard.

## Conclusion

- We explored LLMs potential for OL through our introduced conceptual framework, LLMs4OL.
- Extensive experiments on 11 LLMs across three OL tasks demonstrate the paradigm's proof of concept.
- The obtained empirical results show that foundational LLMs are not sufficiently suitable for ontology construction that entails a high degree of reasoning skills and domain expertise.
- When LLMs effectively fine-tuned they just might work as suitable assistants, alleviating the knowledge acquisition bottleneck, for ontology construction.
- A codebase with detailed results is shared: https://github.com/HamedBabaei/LLMs4OL

## Future:

- Still, we need to explore more recent LLMs.
- Incorporate more ontologies in this study.
- Build a benchmark dataset that considers more domains.
- Optimize three LLMs4OL tasks.

# News!

**ISWC-2024 Challenge:** There is lots of work but a limited number of interested people in exploring the fitness of LLMs for OL. In response to this and attracting more people:

- We are organizing **[ISWC 2024 Challenge] LLMs4OL: The Large Language Models for Ontology Learning 2024 Challenge**
  - With aim of attracting more novel approaches and works for OL.
  - Tasks: Term Typing, Taxonomy Discovery, and Non-Taxonomic Relation Extraction.
    - Few-shot testing phase – train/test split which information such as types and taxonomy will be shared for test and participant will be asked to extract the types, taxonomy and non-taxonomic relations
    - Zero-shot testing phase – there will be no train set.
  - Considering more ontologies:
    - DBpedia
    - Gene Ontology
    - Food Ontology
    - Basic Formal Ontology
    - Ontologies from Humanities and Social Sciences

Hamed Babaei Giglou, Jennifer D'Souza, and Sören Auer (2024). LLMs4OL: The Large Language Models for Ontology learning 2024 Challenge. ISWC 2024 Challenge




# Thank You!