Zero-shot Link Prediction Prompts

LP Prompt.

```
We have two {domain} related entities: A: {entity_1} and B: {entity_2}.

Do you think learning {entity_1} will help in understanding {entity_2}?

Hints:
1. Answer YES or NO only.
2. This is a directional relation, which means if the answer is "YES", (B, A) is
   false, but (A, B) is true.
3. Your answer will be used to create a knowledge graph.

{Additional Information}
```

LP Prompt With Chain-of-Thought.

```
We have two {domain} related entities: A: {entity_1} and B: {entity_2}.

Assess if learning {entity_1} is a prerequisite for understanding {entity_2}.

Employ the Chain of Thought to detail your reasoning before giving a final answer.

# Identify the Domain and entities: Clearly define A and B within their domain.
  Understand the specific content and scope of each entity.

# Analyze the Directional Relationship: Determine if knowledge of entity A is
  essential before one can fully grasp entity B. This involves considering if A
  provides foundational knowledge or skills required for understanding B.

# Evaluate Dependency: Assess whether B is dependent on A in such a way that
  without understanding A, one cannot understand B.

# Draw a Conclusion: Based on your analysis, decide if understanding A is a
  necessary prerequisite for understanding B.

# Provide a Clear Answer: After detailed reasoning, conclude with a distinct answer
  : <result>YES</result> if understanding A is a prerequisite for understanding B,
  or <result>NO</result> if it is not.
```

Extraction Prompt.

```
### Instruction:
You are a domain expert in natural language processing, and now you are building a
knowledge graph in this domain.

Given a context (### Content), and a query entity (### entity), do the following:

1. Extract the query entity and in-domain entities from the context, which should
   be fine-grained: could be introduced by a lecture slide page, or a whole
   lecture, or possibly to have a Wikipedia page.

2. Determine the relations between the query entity and the extracted entities, in
   a triplet format: (<head entity>, <relation>, <tail entity>). The relation
   should be functional, aiding learners in understanding the knowledge. The query
   entity can be the head entity or tail entity.

   We define 7 types of the relations:

   a) Compare: Represents a relation between two or more entities where a
      comparison is being made. For example, "A is larger than B" or "X is more
      efficient than Y."

   b) Part-of: Denotes a relation where one entity is a constituent or component of
      another. For instance, "Wheel is a part of a Car."

   c) Conjunction: Indicates a logical or semantic relation where two or more
      entities are connected to form a group or composite idea. For example, "Salt
      and Pepper."

   d) Evaluate-for: Represents an evaluative relation where one entity is assessed
      in the context of another. For example, "A tool is evaluated for its
      effectiveness."

   e) Is-a-Prerequisite-of: This dual-purpose relation implies that one entity is
      either a characteristic of another or a required precursor for another. For
      instance, "The ability to code is a prerequisite of software development."

   f) Used-for: Denotes a functional relation where one entity is utilized in
      accomplishing or facilitating the other. For example, "A hammer is used for
      driving nails."

   g) Hyponym-Of: Establishes a hierarchical relation where one entity is a more
      specific version or subtype of another. For instance, "A Sedan is a hyponym
      of a Car."

3. Please note some relations are strictly directional. For example, "A tool is
   evaluated for B" indicates (A, Evaluate-for, B), NOT (B, Evaluate-for, A).
   Among the seven relation types, only "a) Compare" and "c) Conjunction" are not
   direction-sensitive.

4. You can also extract triplets from the extracted entities, and the query entity
   may not be necessary in the triplets.

5. Your answer should ONLY contain a list of triplets, each triplet is in this
   format: (entity, relation, entity). For example: "(entity, relation, entity)
   (entity, relation, entity)." No numbering and other explanations are needed.

6. If ### Content is empty, output None.
```

Fusion Prompt.

```
### Instruction: You are a knowledge graph builder.
    Now please fuse two sub-knowledge graphs about the entity "{entity}".

Graph 1: {LLM-KG}   Graph 2: {E-G}

Rules for Fusing the Graphs:
1. Union the entities and edges.

2. If two entities are similar, or refer to the same entity, merge them into one
   entity, keeping he one that is meaningful or specific. For example, "lstm"
   versus "long short-term memory",  please keep "long short-term memory".

3. Only one relation is allowed between two entities. If there is a conflict, read
   the "### Background" to help you keep the correct relation. knowledge to keep the
   correct one. For example, (ROUGE, Evaluate-for, question answering model) and
   (ROUGE,Used-for , question answering model) are considered to be conflicts.

4. Once step 3 is done, consider every possible entity pair not covered in step 2.
   For example, take an entity from Graph 1, and match it from Graph 2. Then,
   please refer to "### Background" to summarize new triplets.

Hint: the relation types and their definition. You can use it to do Step 3.
We define 7 types of the relations:

   a) Compare: Represents a relation between two or more entities where a
      comparison is being made. For example, "A is larger than B" or "X is more
      efficient than Y."

   b) Part-of: Denotes a relation where one entity is a constituent or component of
      another. For instance, "Wheel is a part of a Car."

   c) Conjunction: Indicates a logical or semantic relation where two or more
      entities are connected to form a group or composite idea. For example, "Salt
      and Pepper."

   d) Evaluate-for: Represents an evaluative relation where one entity is assessed
      in the context of another. For example, "A tool is evaluated for its
      effectiveness."

   e) Is-a-Prerequisite-of: This dual-purpose relation implies that one entity is
      either a characteristic of another or a required precursor for another. For
      instance, "The ability to code is a prerequisite of software development."

   f) Used-for: Denotes a functional relation where one entity is utilized in
      accomplishing or facilitating the other. For example, "A hammer is used for
      driving nails."

   g) Hyponym-Of: Establishes a hierarchical relation where one entity is a more
      specific version or subtype of another. For instance, "A Sedan is a hyponym
      of a Car."

### Background:
{background}

### Output Instruction:
    Output the new merged data by listing the triplets. Your answer should ONLY contain
    triplets in this format: (entity, relation, entity). No other explanations or numbering
    are needed. Only triplets, no intermediate results.
```

*Link Prediction with **Doc.***

```
We have two {domain} related entities: A: {entity_1} and B: {entity_2}.

Do you think learning {entity_1} will help in understanding {entity_2}?

Hints:
1. Answer YES or NO only.
2. This is a directional relation, which means if the answer is "YES", (B, A) is
   false, but (A, B) is true.
3. Your answer will be used to create a knowledge graph.

And here are related contents to help:
{related documents concatenation}
```

*Link Prediction with **Con.***

```
We have two {domain} related entities: A: {entity_1} and B: {entity_2}.

Do you think learning {entity_1} will help in understanding {entity_2}?

Hints:
1. Answer YES or NO only.
2. This is a directional relation, which means if the answer is "YES", (B, A) is
   false, but (A, B) is true.
3. Your answer will be used to create a knowledge graph.

And here are related contents to help:

We know that {entity_1} is a prerequisite of the following entities:
{1-hop successors of entity_1 from training data};

The following entities are the prerequisites of {entity_1}:
{1-hop predecessors of entity_1 from training data}.

We know that {entity_2} is a prerequisite of the following entities:
{1-hop successors of entity_2 from training data};

The following entities are the prerequisites of {entity_2}:
{1-hop predecessors of entity_2 from training data}.
```

*Link Prediction with **Wiki.***

```
We have two {domain} related entities: A: {entity_1} and B: {entity_2}.

Do you think learning {entity_1} will help in understanding {entity_2}?

Hints:
1. Answer YES or NO only.
2. This is a directional relation, which means if the answer is "YES", (B, A) is
   false, but (A, B) is true.
3. Your answer will be used to create a knowledge graph.

And here are related contents to help:
{Wikipedia introductory paragraph of {entity_1}}
{Wikipedia introductory paragraph of {entity_2}}
```

GraphRAG's Prompt Tuning for Entity/Relationship Extraction.

```
-Goal-
Given a text document that is potentially relevant to this activity, first identify all the
    entities needed from the text in order to capture the information and ideas in the text.
    Next, introduce each relation concept by defining the relation, and then report all
    relationships among the identified entities according to the predefined relations. These
    predefined relations and seed entities include:

-Relation Concepts and Definitions-:
a) Compare: Represents a relation between two or more entities where a comparison is being
    made. For example, "A is larger than B" or "X is more efficient than Y."
b) Part-of: Denotes a relation where one entity is a constituent or component of another. For
     instance, "Wheel is a part of a Car."
c) Conjunction: Indicates a logical or semantic relation where two or more entities are
    connected to form a group or composite idea. For example, "Salt and Pepper."
d) Evaluate-for: Represents an evaluative relation where one entity is assessed in the
    context of another. For example, "A tool is evaluated for its effectiveness."
e) Is-a-Prerequisite-of: This dual-purpose relation implies that one entity is either a
    characteristic of another or a required precursor for another. For instance, "The ability
     to code is a prerequisite of software development."
f) Used-for: Denotes a functional relation where one entity is utilized in accomplishing or
    facilitating the other. For example, "A hammer is used for driving nails."
g) Hyponym-of: Establishes a hierarchical relation where one entity is a more specific
    version or subtype of another. For instance, "A Sedan is a hyponym of a Car."

-Steps-
1. Identify all entities: For each identified entity, extract the following information:
- entity_name: Name of the entity,
Format each entity as ("entity"{tuple_delimiter}<entity_name>)

2. Identify all relations: From the entities identified in step 1, determine the relation
    between each pair of entities based on the predefined relation concepts (Compare, Part-of
    , Conjunction, Evaluate-for, Is-a-Prerequisite-of, Used-for, and Hyponym-of). For each
    pair of related entities:
- source_entity: Name of the source entity as identified in step 1
- target_entity: Name of the target entity as identified in step 1
- relationship_type: Select the appropriate relation from the predefined relations
- relationship_strength: a numeric score indicating strength of the relationship between the
    source entity and target entity

Format each relationship as ("relationship"{tuple_delimiter}<source_entity>{tuple_delimiter}<
    target_entity>{tuple_delimiter}<relationship_type>{tuple_delimiter}<relationship_strength
    >)
Return output: Provide the list of all entities and relationships identified in steps 1 and
    2. Use {record_delimiter} as the list delimiter. When finished, output {
    completion_delimiter}.


######################
-Real Data-:
######################
text: {input_text}
######################
output:
```

Experimental Setup

In our experimental setup, we employed Hugging Face's `LLaMA-2-70b-chat-hf`[4] and `LLaMA-3-70b-chat-hf`[5] for LLaMA2 and LLaMA3 on a cluster equipped with 8 NVIDIA A100 GPUs. For GPT-3.5 and GPT-4, we used OpenAI's `gpt-3.5-turbo`, `gpt-4-1106-preview`, and `gpt-4o` APIs, respectively, each configured with a temperature setting of zero. The RAG models are implemented using Embedchain [35]. To solve TutorQA tasks, we implemented our pipeline using LangChain[6]. The total budget spent on this project, including the cost of the GPT API service, is approximately 500 USD.

Additional Corpora Description

TutorialBank We obtained the most recent version of TutorialBank from the authors, which consists of 15,583 manually curated resources. This collection includes papers, blog posts, textbook chapters, and other online resources. Each resource is accompanied by metadata and a publicly accessible URL. We downloaded the resources from these URLs and performed free text extraction. Given the varied data formats such as PDF, PPTX, and HTML, we encountered some challenges during text extraction. To ensure text quality, we filtered out sentences shorter than 25 words. Ultimately, this process yielded 559,217 sentences suitable for RAG and finetuning experiments.
NLP-Papers We downloaded conference papers from EMNLP, ACL, and NAACL spanning the years 2021 to 2023. Following this, we utilized Grobid (https://github.com/kermitt2/grobid) for text extraction, resulting in a collection of 4,787 documents with clean text.

Ablation Study

Prompting Strategies In Tab. 7, we explore the impact of different prompting strategies for entity graph recovery, comparing CoT and zero-shot prompts across both NLP and CV domains. The results indicate the introduction of CoT is not improving. We further find that CoT Prompting more frequently results in negative predictions. This finding serves as a drawback for our study, as it somewhat suppresses the performance of our system. This observation highlights the need to balance the impact of CoT on the rigor and complexity of predictions, especially in the context of graph recovery.

Model	NLP		CV	
	Acc	F1	Acc	F1
GPT-4 zs	0.7639	0.7946	0.7391	0.7629
GPT-4 CoT	0.7342	0.6537	0.6122	0.4159

Table 7: Comparison of zero-shot and CoT prompts with GPT-4: Results on NLP and CV.

Finetuning We further explore the impact of finetuning on additional datasets, with results detailed in Table 8. Specifically, we utilize LLaMA2-70b [36], finetuning it on two previously mentioned datasets: TutorialBank and NLP-Papers. Both the zero-shot LLaMA and the finetuned models are employed to generate answers. As these answers are binary (`YES` or `NO`), we can calculate both the accuracy and F1 score for evaluation. However, the results indicate that finetuning does not yield positive outcomes. This can be attributed to two potential factors: 1) the poor quality of data, and 2) limited effectiveness in aiding the graph recovery task. We leave this part as the future work.

Dataset	**Acc**	**F1**
LLaMA2-70b	**0.6058**	**0.6937**
TutorialBank	0.4739	0.0764
NLP Papers	0.5435	0.6363

Table 8: Comparison of the effect of finetuning: Results on NLP domain.

Ablation Study: RAG Data for Link Prediction

We explore the potential of external data in enhancing entity graph recovery. This is achieved by expanding the `{Additional Information}` part in the **LP Prompt**. We utilize LLaMa as the **Base** model, focusing on the NLP domain. We introduce three distinct settings: **Doc.**: In-domain lecture slides data as free-text; **Con.**: Adding one-hop neighboring entities from the training set as additional information related to the query entities. **Wiki.**: Incorporating the introductory paragraph of the Wikipedia page of each query entity. As illustrated in Fig 6, our findings indicate that incorporating LectureBankCD documents (Doc.) significantly diminishes performance. This decline can be attributed

[4]https://huggingface.co/meta-LLaMA
[5]https://huggingface.co/meta-LLaMA/Meta-LLaMA-3-70B
[6]https://www.langchain.com/

to the introduction of noise and excessively lengthy content, which proves challenging for the LLM to process effectively. Conversely, the inclusion of neighboring entities (Con.) markedly enhances the base model's performance. However, it relies on training data, rendering it incompatible with our primary focus on the zero-shot setting. Incorporating Wikipedia content also yields improvements and outperforms the use of LectureBankCD, likely due to higher text quality.

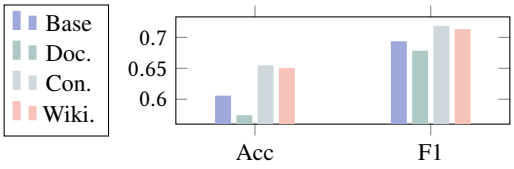

Figure 6: Link Prediction Ablation Study: Comparison of models with external data.

TutorQA

Benchmark Details

We show the data analysis in Tab. 9.

Task	Question Token			entity Count			Number
	Max	Min	Mean	Max	Min	Mean	
T1	77	61	68.00	-	-	-	250
T2	27	22	23.48	7	1	1.79	250
T3	40	34	36.66	8	2	3.36	250
T4	88	76	83.00	-	-	-	250
T5	21	18	19.26	8	1	4.76	100
T6	54	42	48.62	-	-	-	100

Table 9: TutorQA data statistics comparison: The answers in T1 are only "True" or "False", and the answers in T4 are relations, while the answers in T6 are free text with open-ended answers.

GraphRAG Results

We extend the results in Tab. 6 by adding GraphRAG as a baseline, the full version of the evaluation is shown in Tab. 10. Based on the established indexing pipelines in knowledge graph construction, we utilize GraphRAG's query engine with the local search method to directly ask the questions in TutorQA. Notably, the performance of GraphRAG appears less satisfactory, which may be due to an evaluation approach that is not well-suited for GraphRAG's results. For example, in Task 5, GraphRAG produces concepts with very broad or specific terms with a bad granularity, such as *predict sentiment, emotion cause pair extraction, emotional support conversation*. This observation holds across other tasks, where achieving higher scores requires a more granular concept list. This indicates the critical importance of Step 1, which involves generating a well-defined seed concept, in the Graphusion pipeline.

Task 2 and Task 3: case study

Knowledge Graph Construction Analysis

Average Rating We compare expert ratings on the Graphusion KGC results produced by four models: LLaMA, GPT-3.5, GPT-4, and GPT-4o. Fig. 7 and 8 display the average ratings for entity quality and relation quality, respectively, grouped by relation type. Most types achieve an average rating of around 3 (full score) in entity quality, indicating that the extracted triplets contain good in-domain entities. In contrast, the ratings for relation quality are slightly lower. GPT-4 and GPT-4o perform better in relation prediction.

Relation Type Distribution We then compare the Graphusion results for each relation type across the four selected base LLMs, as shown in Fig. 9. All models tend to predict `Prerequisite_of` and `Used_For` relations. The results from LLaMA show relatively even distributions across relation types, whereas the results from the GPT family do not.

Word cloud Visualization Finally, in Fig. 10, we present a word cloud visualization of the entities extracted by Graphusion, comparing the four base LLMs. High-frequency entities include `word embedding`, `model`, `neural network`, `language model`, and others.

Setting	T1	T2	T3	T4	T5
GPT4o zs	69.20	64.42	66.61	44.00	11.45
GPT4o RAG	64.40	65.06	69.31	40.80	10.02
GraphRAG	60.40	64.19	67.45	42.00	8.96
Ours	**92.00**	**80.29**	**77.85**	**50.00**	**15.65**

(a) Evaluation on Tasks 1-5. T1, T4: accuracy; T2, T3: similarity score; T5: hit rate.

Model	Relevancy	Coverage	Convincity	Factuality
GPT4o zs	4.75	4.84	4.38	4.63
GPT4o RAG	4.73	4.71	4.58	4.71
GraphRAG	3.94	4.08	4.13	4.45
Ours	**4.85**	**4.91**	**4.72**	**4.77**

(b) Expert evaluation on Task 6.

Table 10: Results for TutorQA evaluations across various tasks.

***Question* [Task2]**	*In the field of Natural Language Processing, I want to learn about **multilingual model**. What entities should I learn first?*
GPT-4o	**Tokenization, Embeddings, Transfer Learning, Cross-lingual Transfer, Zero-shot Learning, Multilingual Corpora, Language Modeling, Fine-tuning, Evaluation Metrics, Pretrained Models**
Ours	**language models, machine translation, cross-lingual embeddings, transfer learning, tokenization, fine-tuning**
***Question* [Task3]**	*In the field of Natural Language Processing, I know about **natural language processing intro**, now I want to learn about **t-sne**. What entity path should I follow?*
GPT-4o	**natural language processing, dimensionality reduction, t-SNE, perplexity, high-dimensional data, data visualization, machine learning**
Ours	**natural language processing intro, vector representations, t-sne**

Table 11: Case study on TutorQA Task 2 and Task 3: GPT-4o, and GPT-4o-Graphusion.

Question	*Given the following edges constituting an entity subgraph, please identify and select the possible type of relationship between **natural language generation** and **natural language understanding**.*
GPT-4o	**Is-a-Prerequisite-of**
Ours	**Conjunction**

Table 12: Case study on TutorQA Task 4: GPT-4o, and GPT-4o-Graphusion.

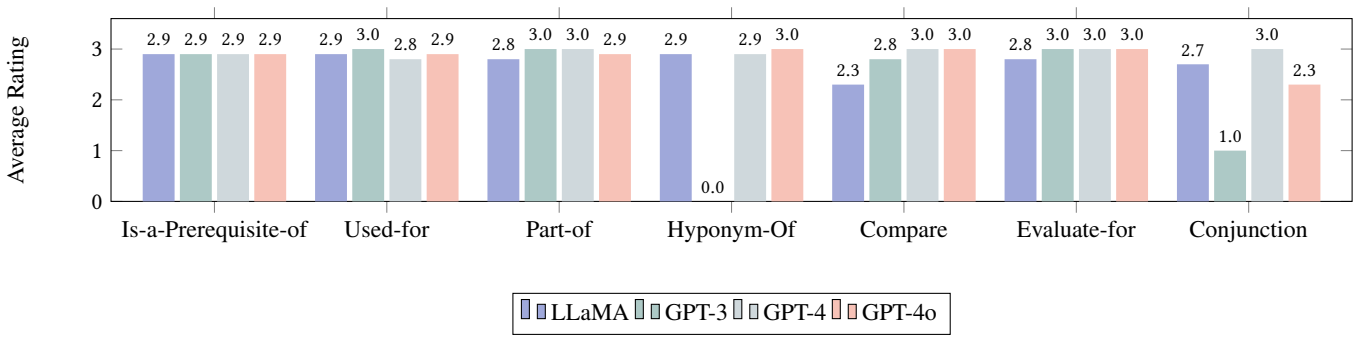

Figure 7: Entity quality rating by human evaluation, grouped by relation type.

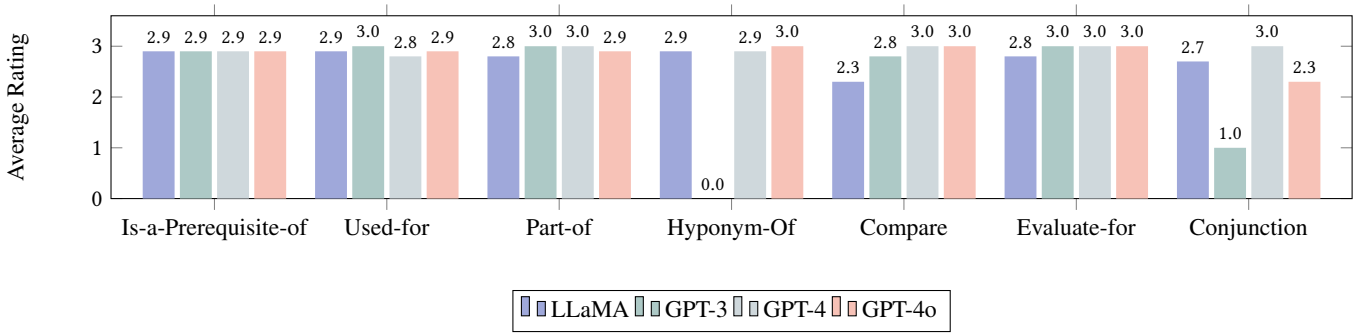

Figure 8: Relation quality rating by human evaluation, grouped by relation type.

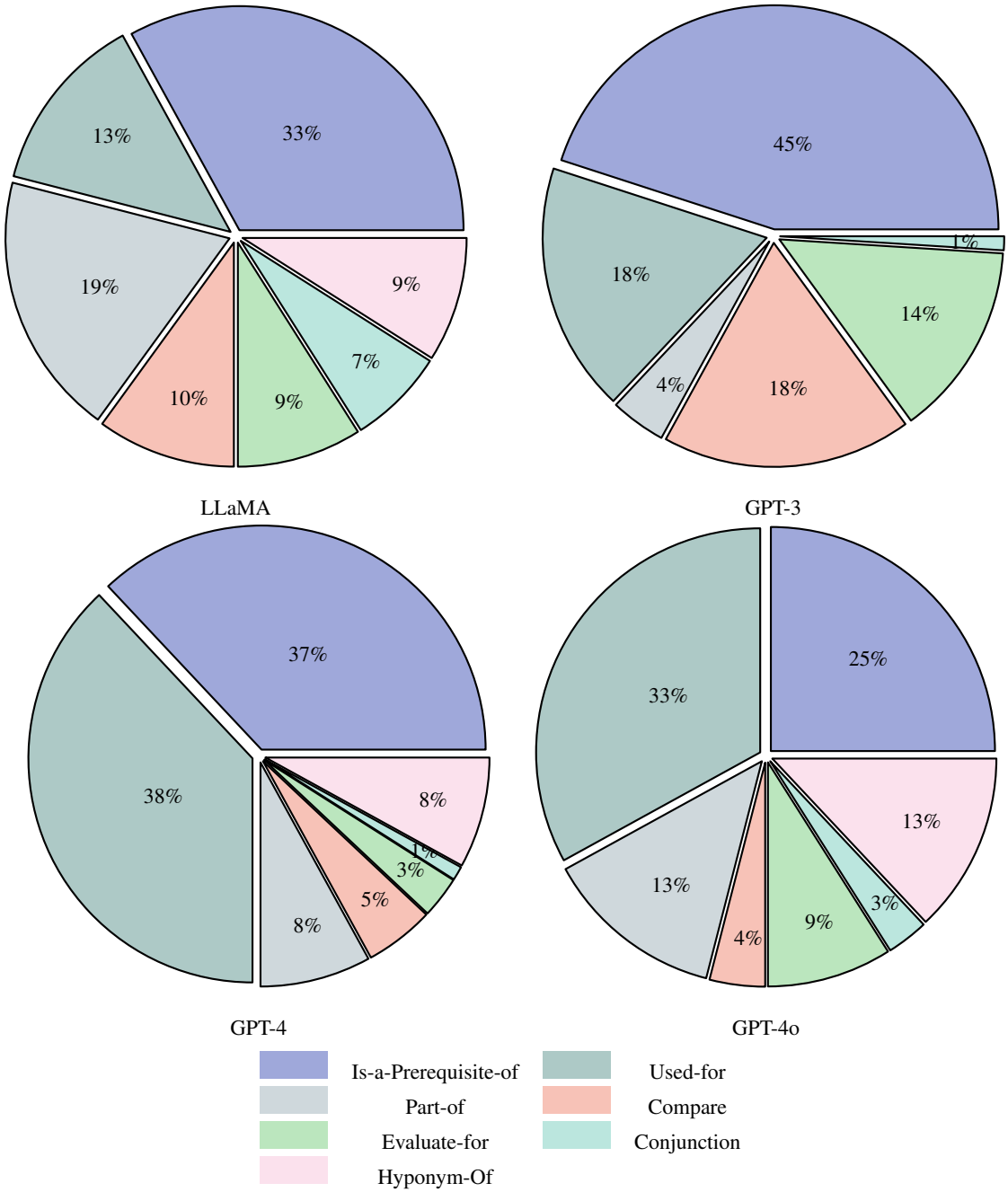

Figure 9: Relation type distribution.

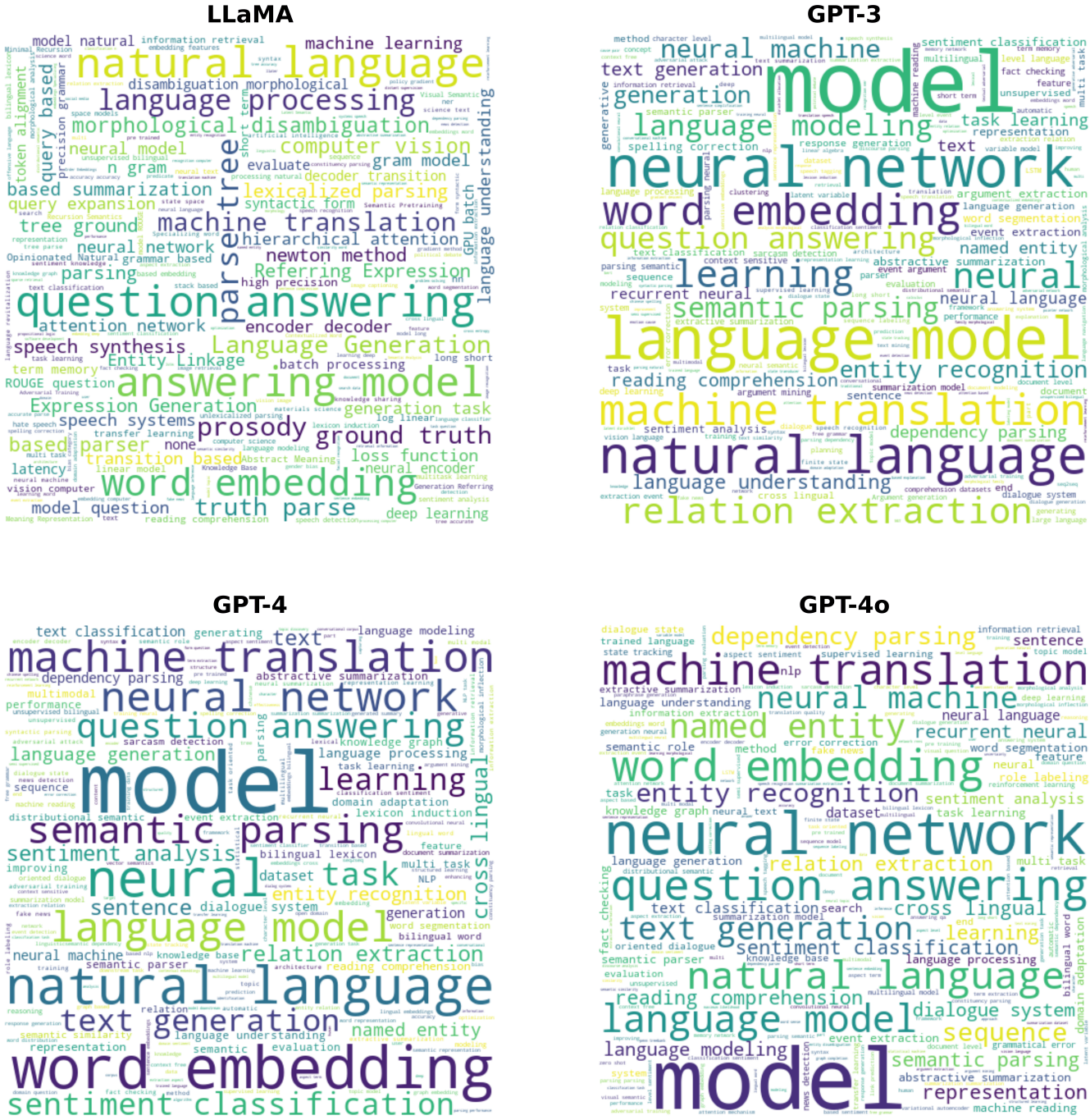

Figure 10: Word cloud visualization for extracted entities.