# Supplementary Material for "LLMs4OL: Large Language Models for Ontology Learning"

Hamed Babaei Giglou[0000−0003−3758−1454], Jennifer D'Souza[0000−0002−6616−9509], and Sören Auer[0000−0002−0698−2864]

TIB Leibniz Information Centre for Science and Technology, Hannover, Germany
{hamed.babaei,jennifer.dsouza,auer}@tib.eu

## Table of Contents



## 1 Zero-Shot Testing Templates

### 1.1 Task A

**Prompt Templates for WordNet Dataset.** Templates for WordNet in zero-shot testing for Task A are presented in Table 1.

**Prompt Templates for GeoNames Dataset.** Templates for GeoNames in zero-shot testing for Task A are presented in Table 2. As a sentence $S$, we used $[L]$ *is a place in* $[COUNTRY]$. template.

**Prompt Templates for UMLS Dataset.** Templates for UMLS sources (NCI, MEDCIN, and SNOMEDCT_US) in zero-shot testing for Task A are presented in Table 3.

### 1.2 Task B

Templates for GeoNames, UMLS, and Schema.Org in zero-shot testing for Task B is for LLMs are presented in Table 4.

**Table 1.** The WordNet zero-shot testing prompt templates for task A. $L$ represents lexical entries, $S$ represents sentence containing $L$. In the BERT/BART LLMs, for BART the $[MASK]$ is being replaced by $< mask >$ ".

| **LLMs** | *Prompt Templates* |
|---|---|
| *BERT/BART* | [S]. [L] POS is a $[MASK]$ . |
| | [S]. [L] part of speech is a $[MASK]$ . |
| | [S]. '[L]' POS is a $[MASK]$ . |
| | [S]. '[L]' part of speech is a $[MASK]$ . |
| | [L] POS is a $[MASK]$ . |
| | [L] part of speech is a $[MASK]$ . |
| | '[L]' POS is a $[MASK]$ . |
| | '[L]' part of speech is a $[MASK]$ . |
| *Flan-T5* | [S]. [L] POS is a ? |
| | [S]. [L] part of speech is a ? |
| | [S]. '[L]' POS is a ? |
| | [S]. '[L]' part of speech is a ? |
| | [L] POS is a ? |
| | [L] part of speech is a ? |
| | '[L]' POS is a ? |
| | '[L]' part of speech is a ? |
| *BLOOM/GPT-3* | Perform a sentence completion on the following sentence: \n Sentence: [S]. [L] POS is a |
| | Perform a sentence completion on the following sentence: \n Sentence: [S]. [L] part of speech is a |
| | Perform a sentence completion on the following sentence: \n Sentence: [S]. '[L]' POS is a |
| | Perform a sentence completion on the following sentence: \n Sentence: [S]. '[L]' part of speech is a |
| | Perform a sentence completion on the following sentence: \n Sentence: [L] POS is a |
| | Perform a sentence completion on the following sentence: \n Sentence: [L] part of speech is a |
| | Perform a sentence completion on the following sentence: \n Sentence: '[L]' POS is a |
| | Perform a sentence completion on the following sentence: \n Sentence: '[L]' part of speech is a |
| *LLaMA* | Perform a sentence completion on the following sentence: [S]. [L] POS is a ___. \n The answer is |
| | Perform a sentence completion on the following sentence: [S]. [L] part of speech is a ___. \n The answer is |
| | Perform a sentence completion on the following sentence: [S]. '[L]' POS is a ___. \n The answer is |
| | Perform a sentence completion on the following sentence: [S]. '[L]' part of speech is a ___. \n The answer is |
| | Perform a sentence completion on the following sentence: [L] POS is a ___. \n The answer is |
| | Perform a sentence completion on the following sentence: [L] part of speech is a ___. \n The answer is |
| | Perform a sentence completion on the following sentence: '[L]' POS is a ___. \n The answer is |
| | Perform a sentence completion on the following sentence: '[L]' part of speech is a ___. \n The answer is |

**Table 2.** The GeoNames zero-shot testing prompt templates for task A. $L$ represents lexical entries, $S$ represents sentence containing $L$. In the BERT/BART LLMs, for BART the $[MASK]$ is being replaced by $< mask >$ ".

| **LLMs** | **Prompt Templates** |
|---|---|
| *BERT/BART* | [S]. [L] is a $[MASK]$ .<br>[S]. [L] geographically is a $[MASK]$ .<br>[S]. '[L]' is a $[MASK]$ .<br>[S]. '[L]' geographically is a $[MASK]$ .<br>[L] is a $[MASK]$ .<br>[L] geographically is a $[MASK]$ .<br>'[L]' is a $[MASK]$ .<br>'[L]' geographically is a $[MASK]$ . |
| *Flan-T5* | [S]. [L] is a ?<br>[S]. [L] geographically is a ?<br>[S]. '[L]' is a ?<br>[S]. '[L]' geographically is a ?<br>[L] is a ?<br>[L] geographically is a ?<br>'[L]' is a ?<br>'[L]' geographically is a ? |
| *BLOOM/GPT-3* | Perform a sentence completion on the following sentence: \n Sentence: [S]. [L] is a |
| | Perform a sentence completion on the following sentence: \n Sentence: [S]. [L] geographically is a |
| | Perform a sentence completion on the following sentence: \n Sentence: [S]. '[L]' is a |
| | Perform a sentence completion on the following sentence: \n Sentence: [S]. '[L]' geographically is a |
| | Perform a sentence completion on the following sentence: \n Sentence: [L] is a |
| | Perform a sentence completion on the following sentence: \n Sentence: [L] geographically is a |
| | Perform a sentence completion on the following sentence: \n Sentence: '[L]' is a |
| | Perform a sentence completion on the following sentence: \n Sentence: '[L]' geographically is a |
| *LLaMA* | Perform a sentence completion on the following sentence: [S]. [L] is a ___. \n The answer is |
| | Perform a sentence completion on the following sentence: [S]. [L] geographically is a ___. \n The answer is |
| | Perform a sentence completion on the following sentence: [S]. '[L]' is a ___. \n The answer is |
| | Perform a sentence completion on the following sentence: \n [S]. '[L]' geographically is a ___. \n The answer is |
| | Perform a sentence completion on the following sentence: \n [L] is a ___. \n The answer is |
| | Perform a sentence completion on the following sentence: \n [L] geographically is a ___. \n The answer is |
| | Perform a sentence completion on the following sentence: \n '[L]' is a ___. \n The answer is |
| | Perform a sentence completion on the following sentence: \n '[L]' geographically is a ___. \n The answer is |

**Table 3.** The UMLS zero-shot testing prompt templates for task A. $L$ represents lexical entries, $S$ represents sentence containing $L$. In the BERT/BART LLMs, for BART the $[MASK]$ is being replaced by $< mask >$ ".

| **LLMs** | **Prompt Templates** |
|---|---|
| *BERT/BART* | [S]. [L] in medicine is a $[MASK]$ . |
| | [S]. [L] in biomedicine is a $[MASK]$ . |
| | [S]. '[L]' in medicine is a $[MASK]$ . |
| | [S]. '[L]' in biomedicine is a $[MASK]$ . |
| | [L] in medicine is a $[MASK]$ . |
| | [L] in biomedicine is a $[MASK]$ . |
| | '[L]' is a $[MASK]$ . |
| | '[L]' in biomedicine is a $[MASK]$ . |
| *Flan-T5* | [S]. [L] in medicine is a ? |
| | [S]. [L] in biomedicine is a ? |
| | [S]. '[L]' in medicine is a ? |
| | [S]. '[L]' in biomedicine is a ? |
| | [L] in medicine is a ? |
| | [L] in biomedicine is a ? |
| | '[L]' in medicine is a ? |
| | '[L]' in biomedicine is a ? |
| *BLOOM/GPT-3* | Perform a sentence completion on the following sentence: \n Sentence: [S]. [L] in medicine is a |
| | Perform a sentence completion on the following sentence: \n Sentence: [S]. [L] in biomedicine is a |
| | Perform a sentence completion on the following sentence: \n Sentence: [S]. '[L]' in medicine is a |
| | Perform a sentence completion on the following sentence: \n Sentence: [S]. '[L]' in biomedicine is a |
| | Perform a sentence completion on the following sentence: \n Sentence: [L] in medicine is a |
| | Perform a sentence completion on the following sentence: \n Sentence: [L] in biomedicine is a |
| | Perform a sentence completion on the following sentence: \n Sentence: '[L]' in medicine is a |
| | Perform a sentence completion on the following sentence: \n Sentence: '[L]' in biomedicine is a |
| *LLaMA* | Perform a sentence completion on the following sentence: Sentence: [S]. [L] in medicine is a ___. \n The answer is |
| | Perform a sentence completion on the following sentence: [S]. [L] in biomedicine is a ___. \n The answer is |
| | Perform a sentence completion on the following sentence: [S]. '[L]' in medicine is a ___. \n The answer is |
| | Perform a sentence completion on the following sentence: [S]. '[L]' in biomedicine is a ___. \n The answer is |
| | Perform a sentence completion on the following sentence: [L] in medicine is a ___. \n The answer is |
| | Perform a sentence completion on the following sentence: [L] in biomedicine is a ___. \n The answer is |
| | Perform a sentence completion on the following sentence: '[L]' in medicine is a ___. \n The answer is |
| | Perform a sentence completion on the following sentence: '[L]' in biomedicine is a ___. \n The answer is |

**Table 4.** The GeoNames, UMLS, and Schema.Org zero-shot testing prompt templates for task B. In the type pairs $(a, b)$ or $(b, a)$, where $a$ is parent and $b$ is child.

| LLMs | Prompt Templates |
|---|---|
| *BERT* | [a] is the superclass of [b]. This statement is a $[MASK]$ .<br>[b] is a subclass of [a]. This statement is a $[MASK]$ .<br>[a] is the parent class of [b]. This statement is a $[MASK]$ .<br>[b] is a child class of [a]. This statement is a $[MASK]$.<br>[a] is a supertype of [b]. This statement is a $[MASK]$ .<br>[b] is a subtype of [a]. This statement is a $[MASK]$ .<br>[a] is an ancestor class of [b]. This statement is a $[MASK]$ .<br>[b] is a descendant class of [a]. This statement is a $[MASK]$ . |
| *BART* | [a] is the superclass of [b]. This statement is a $< mask >$ .<br>[b] is a subclass of [a]. This statement is a $< mask >$ .<br>[a] is the parent class of [b]. This statement is a $< mask >$ .<br>[b] is a child class of [a]. This statement is a $< mask >$.<br>[a] is a supertype of [b]. This statement is a $< mask >$ .<br>[b] is a subtype of [a]. This statement is a $< mask >$ .<br>[a] is an ancestor class of [b]. This statement is a $< mask >$ .<br>[b] is a descendant class of [a]. This statement is a $< mask >$ . |
| *Flan-T5* | [a] is the superclass of [b]. This statement is a<br>[b] is a subclass of [a]. This statement is a<br>[a] is the parent class of [b]. This statement is a<br>[b] is a child class of [a]. This statement is a<br>[a] is a supertype of [b]. This statement is a<br>[b] is a subtype of [a]. This statement is a<br>[a] is an ancestor class of [b]. This statement is a<br>[b] is a descendant class of [a]. This statement is a |
| *BLOOM/GPT-3* | Identify whether the following statement is true or false: \n Statement: [a] is the superclass of [b]. \n This statement is a |
| | Identify whether the following statement is true or false: \n Statement: [b] is a subclass of [a]. \n This statement is a |
| | Identify whether the following statement is true or false: \n Statement: [a] is the parent class of [b]. \n This statement is a |
| | Identify whether the following statement is true or false: \n Statement: [b] is a child class of [a]. \n This statement is a |
| | Identify whether the following statement is true or false: \n Statement: [a] is a supertype of [b]. \n This statement is a |
| | Identify whether the following statement is true or false: \n Statement: [b] is a subtype of [a]. \nThis statement is a |
| | Identify whether the following statement is true or false: \n Statement: [a] is an ancestor class of [b]. \n This statement is a |
| | Identify whether the following statement is true or false: \n Statement: [b] is a descendant class of [a]. \n This statement is a |

## 2 Flan-T5 Training Setups and Hyperparameters

We finetune Flan-T5 LM on three tasks and evaluate them on all three tasks using zero-shot testing. It involved employing different sources, i.e. WordNet (task A), GeoNames (task A and B), UMLS (the NCI source representing medical sources in task A, B, and C), and schema.org (task B). Considering task A as an 8-shot instance for training we combined samples task B and C training with the condition that only samples that are in the task A 8-shot instances are considered for inclusion. Next, using task-specific prompt templates, Flan-T5 inputs are generated for finetuning. Following, using designed prompt templates Flan-T5 is fine-tuned.

We utilized a consistent training strategy for all datasets and models, except for a few hyperparameters: batch size and finetuning steps. All the models were trained using AdamW optimizer with a learning rate of 1e-5. For the Flan-T5-Large model, a batch size of 8 is used during training, while for the Flan-T5-XL model, a batch size of 4 is employed on all datasets. The WordNet and Schema.Org datasets were finetuned for 5 training epochs on both models, similarly, UMLS was finetuned using 10 epochs, while GeoNames was finetuned on Flan-T5-Large for 10 epochs and on Flan-T5-XL for 6 epochs.

## 3 Detailed Results

The Table 5 and Table 6 represent prompt template results across all the templates and LLMs for term typing. While the Table 7 represents prompt template results across all the templates and LLMs for taxonomy discovery.

**Table 5.** The detailed results of zero-shot testing and finetuning across seven LLMs reported for ontology learning Task A, term typing in MAP@1. The results are in percentages. The $*$ denotes finetuning model results. Results for WordNet and GeoName datasets.

| **Dataset** | **LLMs** | **Prompt Templates** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $t_1$ | $t_2$ | $t_3$ | $t_4$ | $t_5$ | $t_6$ | $t_7$ | $t_8$ |
| WordNet | BERT-Large | 2.19 | 9.36 | 9.18 | 19.41 | 4.72 | 19.34 | 9.93 | 27.85 |
| | BART-Large | 0.01 | 0.28 | 0.22 | 2.16 | 0.01 | 0.03 | 0.0 | 0.19 |
| | Flan-T5-Large | 0.17 | 19.70 | 5.54 | 31.26 | 0.0 | 3.03 | 5.70 | 26.80 |
| | BLOOM-1b7 | 66.83 | 71.53 | 79.20 | 76.84 | 40.08 | 61.96 | 68.39 | 70.03 |
| | Flan-T5-XL | 2.81 | 40.26 | 17.83 | 52.21 | 0.01 | 7.75 | 18.47 | 18.85 |
| | BLOOM-3b | 63.33 | 75.29 | 79.08 | 77.06 | 37.40 | 65.32 | 68.99 | 71.62 |
| | LLaMA-7B | 37.26 | 74.61 | 70.16 | 75.97 | 24.28 | 76.61 | 67.41 | 81.38 |
| | GPT-3 | 15.32 | 26.55 | 37.86 | 27.57 | 8.47 | 27.13 | 27.51 | 24.65 |
| | GPT-3.5 | 24.27 | 80.81 | 89.46 | 91.72 | 0.81 | 60.76 | 49.38 | 82.41 |
| | GPT-4 | - | - | - | 90.11 | - | - | - | - |
| | *Flan-T5-Large* * | 73.32 | 76.74 | 54.57 | 76.90 | 10.83 | 61.36 | 54.29 | 69.32 |
| | *Flan-T5-XL* * | 84.51 | 84.77 | 77.27 | 86.28 | 50.23 | 76.46 | 72.38 | 80.51 |
| GeoNames | BERT-Large | 38.34 | 29.79 | 30.86 | 35.32 | 23.61 | 25.66 | 11.32 | 30.44 |
| | BART-Large | 8.47 | 0.57 | 2.23 | 0.98 | 21.48 | 20.51 | 7.83 | 23.21 |
| | Flan-T5-Large | 11.55 | 3.57 | 13.16 | 4.68 | 9.45 | 6.05 | 8.17 | 7.38 |
| | BLOOM-1b7 | 2.71 | 2.54 | 2.89 | 3.20 | 28.51 | 18.38 | 25.86 | 19.80 |
| | Flan-T5-XL | 33.81 | 15.71 | 19.77 | 20.78 | 15.36 | 12.41 | 18.43 | 15.82 |
| | BLOOM-3b | 3.76 | 4.70 | 2.64 | 3.43 | 28.84 | 18.08 | 25.64 | 20.71 |
| | LLaMA-7B | 29.49 | 14.16 | 25.54 | 15.95 | 13.91 | 9.44 | 17.79 | 16.79 |
| | GPT-3 | 22.42 | 8.72 | - | 7.50 | - | - | - | - |
| | GPT-3.5 | 35.00 | - | - | - | - | - | - | - |
| | GPT-4 | 43.28 | - | - | - | - | - | - | - |
| | *Flan-T5-Large* * | 15.08 | 15.17 | 14.93 | 15.12 | 15.77 | 16.28 | 15.93 | 16.91 |
| | *Flan-T5-XL* * | 18.35 | 18.12 | 18.12 | 17.91 | 17.26 | 17.32 | 17.45 | 17.64 |

**Table 6.** The detailed results of zero-shot testing and finetuning across seven LLMs reported for ontology learning Task A, term typing in MAP@1. The results are in percentages. The $*$ denotes finetuning model results. Results for NCI, SNOMEDCT_US, and MEDCIN datasets.

| **Dataset** | **LLMs** | **Prompt Templates** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $t_1$ | $t_2$ | $t_3$ | $t_4$ | $t_5$ | $t_6$ | $t_7$ | $t_8$ |
| NCI | BERT-Large | 9.94 | 9.76 | 2.61 | 2.90 | 11.09 | 10.96 | 1.12 | 1.36 |
| | PubMedBERT | 5.87 | 5.36 | 4.52 | 2.79 | 3.36 | 1.61 | 1.33 | 0.65 |
| | BART-Large | 7.09 | 7.87 | 5.14 | 6.32 | 9.10 | 9.94 | 7.24 | 8.26 |
| | Flan-T5-Large | 4.59 | 5.06 | 7.53 | 8.96 | 3.06 | 4.25 | 5.48 | 5.84 |
| | BLOOM-1b7 | 12.03 | 12.10 | 11.22 | 12.43 | 10.95 | 10.45 | 11.13 | 11.49 |
| | Flan-T5-XL | 4.44 | 5.65 | 7.41 | 9.83 | 2.12 | 3.29 | 3.87 | 6.28 |
| | BLOOM-3b | 13.77 | 14.35 | 12.94 | 14.41 | 14.26 | 14.06 | 14.92 | 15.56 |
| | LLaMA-7B | 3.78 | 4.05 | 3.24 | 4.77 | 3.67 | 3.92 | 5.25 | 7.71 |
| | GPT-3 | 9.30 | 9.17 | 11.03 | 12.74 | 9.37 | 8.75 | 9.14 | 9.11 |
| | GPT-3.5 | 11.04 | 9.52 | 14.70 | 14.22 | 8.56 | 8.13 | 12.68 | 11.24 |
| | GPT-4 | - | - | 16.05 | - | - | - | - | - |
| | *Flan-T5-Large* * | 30.60 | 31.59 | 31.32 | 31.92 | 29.11 | 29.28 | 31.29 | 30.79 |
| | *Flan-T5-XL* * | 31.51 | 30.99 | 32.78 | 32.05 | 30.01 | 29.70 | 31.76 | 31.35 |
| SNOMEDCT | BERT-Large | 19.83 | 8.02 | 1.06 | 0.12 | 21.10 | 12.76 | 0.45 | 0.04 |
| | PubMedBERT | 28.48 | 22.47 | 13.91 | 5.70 | 7.96 | 3.58 | 2.29 | 1.51 |
| | BART-Large | 19.16 | 19.81 | 4.16 | 4.04 | 17.54 | 17.89 | 10.06 | 9.43 |
| | Flan-T5-Large | 19.26 | 19.89 | 21.04 | 24.32 | 8.07 | 8.90 | 11.54 | 12.92 |
| | BLOOM-1b7 | 32.43 | 37.02 | 13.78 | 19.97 | 29.48 | 30.40 | 31.24 | 33.86 |
| | Flan-T5-XL | 25.21 | 26.23 | 30.09 | 31.65 | 7.21 | 8.22 | 15.58 | 17.22 |
| | BLOOM-3b | 34.26 | 37.69 | 27.18 | 27.87 | 31.06 | 32.21 | 33.29 | 35.47 |
| | LLaMA-7B | 7.56 | 6.75 | 7.89 | 8.06 | 10.74 | 10.80 | 13.15 | 13.81 |
| | GPT-3 | 21.06 | 20.33 | 22.73 | 24.36 | 19.20 | 18.99 | 20.20 | 20.09 |
| | GPT-3.5 | 21.81 | 17.99 | 25.02 | 24.50 | 18.24 | 15.71 | 22.71 | 19.87 |
| | GPT-4 | - | - | 22.36 | 27.83 | - | - | - | - |
| | *Flan-T5-Large* * | 32.27 | 31.99 | 31.56 | 31.36 | 32.00 | 31.50 | 33.39 | 33.05 |
| | *Flan-T5-XL* * | 43.39 | 42.03 | 42.76 | 41.75 | 40.89 | 40.31 | 42.60 | 42.48 |
| MEDCIN | BERT-Large | 7.33 | 1.25 | 0.14 | 0.05 | 8.71 | 1.19 | 0.08 | 0.01 |
| | PubMedBERT | 15.62 | 9.71 | 5.20 | 1.58 | 5.68 | 2.32 | 1.27 | 0.61 |
| | BART-Large | 11.67 | 12.65 | 2.27 | 2.31 | 9.40 | 9.22 | 5.47 | 4.82 |
| | Flan-T5-Large | 9.30 | 8.08 | 10.97 | 12.96 | 2.89 | 3.59 | 6.71 | 6.78 |
| | BLOOM-1b7 | 27.58 | 28.67 | 2.70 | 4.97 | 26.38 | 28.76 | 26.89 | 26.69 |
| | Flan-T5-XL | 15.24 | 15.89 | 18.04 | 18.51 | 4.47 | 5.44 | 11.14 | 11.09 |
| | BLOOM-3b | 23.05 | 28.31 | 14.39 | 10.82 | 22.58 | 24.23 | 27.30 | 29.81 |
| | LLaMA-7B | 3.40 | 2.80 | 3.37 | 3.73 | 4.90 | 4.47 | 3.17 | 3.80 |
| | GPT-3 | 22.40 | 22.56 | 25.72 | 24.91 | 19.75 | 17.80 | 19.92 | 18.57 |
| | GPT-3.5 | 22.51 | 22.06 | 23.92 | 23.58 | 20.46 | 19.84 | 22.37 | 20.23 |
| | GPT-4 | - | - | 21.25 | 23.61 | - | - | - | - |
| | *Flan-T5-Large* * | 38.37 | 36.37 | 37.43 | 35.86 | 31.26 | 30.00 | 33.11 | 31.91 |
| | *Flan-T5-XL* * | 51.80 | 50.90 | 51.80 | 51.16 | 47.88 | 45.38 | 49.86 | 49.09 |

**Table 7.** The detailed results of zero-shot testing and finetuning across seven LLMs reported for ontology learning Task B, type taxonomy discovery in F1-score. The results are in percentages. The $*$ denotes finetuning model results.

| **Dataset** | **LLMs** | **Prompt Templates** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $t_1$ | $t_2$ | $t_3$ | $t_4$ | $t_5$ | $t_6$ | $t_7$ | $t_8$ |
| GeoNames | BERT-Large | 41.00 | 51.69 | 40.55 | 48.70 | 37.16 | 41.07 | 41.70 | 54.54 |
| | BART-Large | 38.11 | 41.03 | 40.55 | 52.50 | 39.09 | 45.80 | 36.67 | 55.40 |
| | Flan-T5-Large | 59.63 | 48.24 | 54.08 | 48.24 | 44.40 | 51.30 | 36.40 | 38.44 |
| | BLOOM-1b7 | 33.16 | 31.04 | 33.16 | 32.83 | 33.77 | 33.53 | 36.67 | 32.92 |
| | Flan-T5-XL | 49.37 | 44.05 | 45.09 | 52.41 | 43.92 | 46.34 | 49.98 | 44.29 |
| | BLOOM-3b | 35.85 | 39.12 | 53.92 | 30.22 | 35.62 | 33.60 | 48.26 | 37.73 |
| | LLaMA-7B | 33.49 | 33.49 | 33.49 | 33.49 | 33.49 | 33.49 | 33.49 | 33.49 |
| | GPT-3 | 43.43 | 51.74 | 42.70 | 53.20 | 46.04 | 52.56 | 45.49 | 52.62 |
| | GPT-3.5 | 59.40 | 47.79 | 67.78 | 41.95 | 48.02 | 51.72 | 45.25 | 43.86 |
| | GPT-4 | 38.56 | 52.46 | 34.00 | 38.89 | 44.06 | 55.43 | 33.78 | 36.23 |
| | *Flan-T5-Large*$^*$ | 42.53 | 59.40 | 40.29 | 62.46 | 46.03 | 57.41 | 42.49 | 62.04 |
| | *Flan-T5-XL*$^*$ | 48.41 | 34.80 | 55.23 | 46.96 | 57.48 | 36.29 | 59.05 | 49.26 |
| UMLS | BERT-Large | 48.21 | 38.84 | 41.46 | 40.41 | 45.88 | 40.91 | 41.04 | 42.92 |
| | PubMedBERT | 33.71 | 33.71 | 33.71 | 33.71 | 33.71 | 33.71 | 33.71 | 33.71 |
| | BART-Large | 36.02 | 48.21 | 41.42 | 49.90 | 39.37 | 47.47 | 42.39 | 45.46 |
| | Flan-T5-Large | 47.55 | 51.22 | 55.32 | 40.94 | 49.45 | 50.87 | 44.23 | 42.90 |
| | BLOOM-1b7 | 33.71 | 36.18 | 33.71 | 38.26 | 33.71 | 35.89 | 33.27 | 33.60 |
| | Flan-T5-XL | 64.25 | 46.53 | 51.00 | 41.54 | 60.07 | 42.83 | 51.25 | 41.18 |
| | BLOOM-3b | 33.16 | 37.23 | 34.82 | 35.77 | 33.16 | 35.89 | 33.05 | 37.48 |
| | LLaMA-7B | 32.94 | 32.94 | 32.94 | 32.94 | 32.94 | 32.94 | 32.94 | 32.94 |
| | GPT-3 | 51.58 | 49.41 | 49.86 | 42.90 | 50.57 | 46.07 | 45.36 | 46.72 |
| | GPT-3.5 | 61.38 | 70.38 | 63.91 | 66.82 | 63.14 | 67.27 | 56.64 | 64.41 |
| | GPT-4 | 41.19 | 76.99 | 42.55 | 63.88 | 50.28 | 78.11 | 36.59 | 60.72 |
| | *Flan-T5-Large*$^*$ | 37.17 | 48.66 | 36.07 | 42.12 | 48.39 | 46.65 | 53.42 | 35.97 |
| | *Flan-T5-XL*$^*$ | 63.69 | 50.04 | 36.91 | 41.34 | 78.12 | 50.12 | 79.25 | 39.27 |
| schema.org | BERT-Large | 43.85 | 41.17 | 44.06 | 43.20 | 43.70 | 40.05 | 42.15 | 43.72 |
| | BART-Large | 34.62 | 38.69 | 39.28 | 52.90 | 38.20 | 41.17 | 43.26 | 42.74 |
| | Flan-T5-Large | 46.98 | 49.92 | 46.11 | 54.78 | 40.27 | 54.47 | 42.06 | 47.93 |
| | BLOOM-1b7 | 33.39 | 47.83 | 33.39 | 39.77 | 38.92 | 48.56 | 44.35 | 39.57 |
| | Flan-T5-XL | 42.70 | 33.45 | 33.59 | 42.76 | 36.69 | 34.04 | 33.75 | 36.45 |
| | BLOOM-3b | 41.64 | 47.16 | 47.98 | 45.25 | 39.73 | 40.75 | 51.28 | 48.73 |
| | LLaMA-7B | 33.37 | 33.37 | 33.37 | 33.37 | 33.37 | 33.37 | 33.37 | 33.37 |
| | GPT-3 | 49.64 | 49.28 | 50.97 | 48.03 | 47.19 | 48.63 | 48.87 | 49.48 |
| | GPT-3.5 | 56.84 | 74.38 | 58.52 | 70.16 | 53.35 | 72.35 | 54.16 | 71.03 |
| | GPT-4 | 58.47 | 72.82 | 65.83 | 63.30 | 50.56 | 74.24 | 57.45 | 63.69 |
| | *Flan-T5-Large*$^*$ | 35.35 | 85.43 | 29.82 | 89.24 | 41.30 | 91.68 | 42.46 | 56.39 |
| | *Flan-T5-XL*$^*$ | 91.06 | 57.46 | 74.68 | 65.32 | 91.54 | 50.63 | 91.70 | 33.33 |