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

Su	pplementary Material for "LLMs4OL: Large Language Models for	
Or	ntology Learning"]
1	Zero-Shot Testing Templates 1.1 Task A	1
2	Flan-T5 Training Setups and Hyperparameters	6
3	Detailed Results	6
1	Zero-Shot Testing Templates	
1.1	1 Task A	
Pr	compt Templates for WordNet Dataset. Templates for WordNet in z	ero.

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 > ".

į.	
LLMs	Prompt Templates
	[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 $[MASK]$.
BERT/BART	[L] POS is a $[MASK]$.
	[L] part of speech is a $[MASK]$.
	'[L]' POS is a $[MASK]$.
	'[L]' part of speech is a $[MASK]$.
	[S]. [L] POS is a ?
	[S]. [L] part of speech is a ?
	[S]. '[L]' POS is a ?
	[S]. '[L]' part of speech is a?
Flan- $T5$	[L] POS is a ?
	[L] part of speech is a ?
	'[L]' POS is a ?
	'[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: [S]. '[L]' POS is a
	Perform a sentence completion on the following sentence: \n
	Sentence: [S]. '[L]' part of speech is a
BLOOM/GPT-3	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
	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
LLaMA	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
	[S]. $[L]$ is a $[MASK]$.
	[S]. [L] geographically is a $[MASK]$.
	[S]. $[L]$ is a $[MASK]$.
DEDT /D A DT	[S]. $'[L]'$ geographically is a $[MASK]$.
BERT/BART	[L] is a $[MASK]$.
	[L] geographically is a $[MASK]$.
	'[L]' is a $[MASK]$.
	'[L]' geographically is a $[MASK]$.
	[S]. [L] is a?
	[S]. [L] geographically is a ?
	[S]. '[L]' is a ?
Elan TE	[S]. '[L]' geographically is a ?
Flan-T5	[L] is a ?
	[L] geographically is a ?
	'[L]' is a ?
	'[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: [S]. '[L]' is a
	Perform a sentence completion on the following sentence: \n
BLOOM/GPT-3	Sentence: [S]. '[L]' geographically is a
DLOOM/GI 1-3	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
	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].
LLaMA	'[L]' geographically is a \n The answer is
BBWIIII	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

4 Babaei Giglou et al.

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
	[S]. [L] in medicine is a $[MASK]$.
	[S]. [L] in biomedicine is a $[MASK]$.
	[S]. '[L]' in medicine is a $[MASK]$.
DEDT/DADT	[S]. '[L]' in biomedicine is a $[MASK]$.
BERT/BART	[L] in medicine is a $[MASK]$.
	[L] in biomedicine is a $[MASK]$.
	'[L]' is a $[MASK]$.
	'[L]' in biomedicine is a $[MASK]$.
	[S]. [L] in medicine is a?
	[S]. [L] in biomedicine is a ?
	[S]. '[L]' in medicine is a ?
חו חוי	[S]. '[L]' in biomedicine is a?
Flan- $T5$	[L] in medicine is a ?
	[L] in biomedicine is a ?
	'[L]' in medicine is a ?
	'[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: [S]. '[L]' in medicine is a
	Perform a sentence completion on the following sentence: \n
DIOOM/ODT 6	Sontoneo: [S] '[I]' in biomodicino is a
BLOOM/GPT-3	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
	Perform a sentence completion on the following sentence: Sen-
	tence: [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 \dots \n The answer is
	Perform a sentence completion on the following sentence: [S].
LLaMA	'[L]' in biomedicine is a \n The answer is
LLUMA	Perform a sentence completion on the following sentence: [L] in
	medicine is a \dots \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]'

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
	[a] is the superclass of [b]. This statement is a $[MASK]$.
	[b] is a subclass of [a]. This statement is a $[MASK]$.
	[a] is the parent class of [b]. This statement is a $[MASK]$.
BERT	[b] is a child class of [a]. This statement is a $[MASK]$.
DERI	[a] is a supertype of [b]. This statement is a $[MASK]$.
	[b] is a subtype of [a]. This statement is a $[MASK]$.
	[a] is an ancestor class of [b]. This statement is a $[MASK]$.
	[b] is a descendant class of [a]. This statement is a $[MASK]$.
	[a] is the superclass of [b]. This statement is a $\langle mask \rangle$.
	[b] is a subclass of [a]. This statement is a $< mask >$.
	[a] is the parent class of [b]. This statement is a $< mask >$.
BART	[b] is a child class of [a]. This statement is a $< mask >$.
DARI	[a] is a supertype of [b]. This statement is a $\langle mask \rangle$.
	[b] is a subtype of [a]. This statement is a $\langle mask \rangle$.
	[a] is an ancestor class of [b]. This statement is a $\langle mask \rangle$.
	[b] is a descendant class of [a]. This statement is a $< mask >$.
	[a] is the superclass of [b]. This statement is a
	[b] is a subclass of [a]. This statement is a
	[a] is the parent class of [b]. This statement is a
חו שי	[b] is a child class of [a]. This statement is a
Flan-T5	[a] is a supertype of [b]. This statement is a
	[b] is a subtype of [a]. This statement is a
	[a] is an ancestor class of [b]. This statement is a
	[b] is a descendant class of [a]. This statement is a
	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
DI COM / CDT A	Statement: [b] is a child class of [a]. \n This statement is a
BLOOM/GPT-3	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.

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

Dataset	LLMs	Prompt Templates							
Dataset	LLIVIS	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8
	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
WordNet	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 st	84.51	84.77	77.27	86.28	50.23	76.46	72.38	80.51
	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
GeoNames	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 st	18.35	18.12	18.12	17.91	17.26	17.32	17.45	17.64

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

$ \begin{array}{c ccccccccccccccccccccccccccccccccccc$	2 1.36 3 0.65 4 8.26 8 5.84 13 11.49 7 6.28 92 15.56
PubMedBERT 5.87 5.36 4.52 2.79 3.36 1.61 1.6 BART-Large 7.09 7.87 5.14 6.32 9.10 9.94 7.5 Flan-T5-Large 4.59 5.06 7.53 8.96 3.06 4.25 5.6	3 0.65 4 8.26 8 5.84 13 11.49 7 6.28 92 15.56
BART-Large 7.09 7.87 5.14 6.32 9.10 9.94 7.5 Flan-T5-Large 4.59 5.06 7.53 8.96 3.06 4.25 5.4	4 8.26 8 5.84 13 11.49 7 6.28 92 15.56
Flan-T5-Large 4.59 5.06 7.53 8.96 3.06 4.25 5.4	8 5.84 13 11.49 7 6.28 92 15.56
	13 11.49 7 6.28 92 15.56
	7 6.28 92 15.56
NCI BLOOM-1b7 12.03 12.10 11.22 12.43 10.95 10.45 11.	92 15.56
Flan-T5-XL 4.44 5.65 7.41 9.83 2.12 3.29 3.3	
BLOOM-3b 13.77 14.35 12.94 14.41 14.26 14.06 14.	
LLaMA-7B 3.78 4.05 3.24 4.77 3.67 3.92 5.5	5 7.71
GPT-3 9.30 9.17 11.03 12.74 9.37 8.75 9.	4 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
BERT-Large 19.83 8.02 1.06 0.12 21.10 12.76 0.	5 0.04
PubMedBERT 28.48 22.47 13.91 5.70 7.96 3.58 2.5	9 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
SNOMEDCT 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.	20.09
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	33.05
Flan-T5-XL * 43.39 42.03 42.76 41.75 40.89 40.31 42.	60 42.48
BERT-Large 7.33 1.25 0.14 0.05 8.71 1.19 0.0	8 0.01
PubMedBERT 15.62 9.71 5.20 1.58 5.68 2.32 1.5	7 0.61
BART-Large 11.67 12.65 2.27 2.31 9.40 9.22 5.4	7 4.82
Flan-T5-Large 9.30 8.08 10.97 12.96 2.89 3.59 6.	1 6.78
MEDCIN 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.	7 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.

BERT-Large 41.00 51.00 40.55 48.70 37.16 41.07 41.70 54.54 BERT-Large 59.63 48.24 54.00 31.00 36.67 55.40 BLOOM-1b7 33.16 31.04 33.16 32.83 33.77 33.53 36.67 32.92 GeoNames Flan-T5-XL 49.37 44.05 45.09 52.41 43.92 46.34 49.98 42.90 BLOOM-3b 33.49 33.49 33.49 33.49 33.40 33.49 33.49 GPT-3 43.43 51.74 42.70 53.20 46.04 52.56 45.20 52.60 GPT-3 43.43 51.74 42.70 53.20 46.04 52.56 45.20 52.60 GPT-4 38.56 52.61 34.00 38.89 44.06 55.43 33.78 36.23 Flan-T5-Large 48.41 34.80 52.32 46.94 54.20 52.60 Flan-T5-Large 48.41 34.80 52.23 46.94 55.43 33.78 36.23 Flan-T5-Large 48.21 38.84 41.46 40.41 45.88 40.91 42.90 BERT-Large 48.21 38.84 41.46 40.41 45.88 40.91 42.90 Talan-T5-Large 48.21 38.84 41.42 49.90 39.37 47.47 42.39 42.40 Flan-T5-Large 47.55 51.22 55.32 40.94 49.45 50.87 42.30 42.40 BLOOM-1b7 33.71 36.18 33.71	Dataset	LLMs	Prompt Templates							
BART-Large 38.11 41.03 40.55 52.50 39.09 45.80 36.67 55.40 GeoNames Flan-T5-Large 59.63 48.24 54.08 48.24 44.40 51.30 36.40 38.44 GeoNames Flan-T5-XL 49.37 44.05 45.09 52.41 43.92 46.34 49.98 42.92 GeONames Flan-T5-XL 49.37 44.05 45.09 52.42 33.49 33	Dataset	LLIVIS	$\overline{t_1}$		t_3	t_4	t_5	t_6	t_7	t_8
GeoNames Flan-T5-Large 59.63 48.24 54.08 48.24 44.40 51.30 36.40 38.44 GeoNames Flan-T5-XL 49.37 44.05 45.09 52.41 43.29 46.24 49.93 44.29 35.62 33.53 36.67 32.92 BLOOM-3b 38.49 33.49 32.42 25.25 43.60 26.26 <td></td> <td>BERT-Large</td> <td>41.00</td> <td>51.69</td> <td>40.55</td> <td>48.70</td> <td>37.16</td> <td>41.07</td> <td>41.70</td> <td>54.54</td>		BERT-Large	41.00	51.69	40.55	48.70	37.16	41.07	41.70	54.54
BLOOM-1b7 33.16 31.04 33.16 32.83 33.77 33.53 36.67 32.92 GeoNames 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 GPT-3 43.43 51.74 42.70 53.20 46.04 52.56 45.29 52.62 GPT-3 58.40 51.77 47.79 67.78 41.95 48.02 51.72 45.25 43.60 GPT-4 38.56 52.46 34.00 38.89 44.06 55.43 33.78 36.23 Flan-T5-Large 48.21 38.80 55.23 46.96 57.48 36.29 59.05 49.26 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 <t< td=""><td></td><td>BART-Large</td><td>38.11</td><td>41.03</td><td>40.55</td><td>52.50</td><td>39.09</td><td>45.80</td><td>36.67</td><td>55.40</td></t<>		BART-Large	38.11	41.03	40.55	52.50	39.09	45.80	36.67	55.40
GeoNames Flan-T5-XL 49.37 44.05 45.09 52.41 43.92 46.34 49.98 44.29 LLaMA-7B 35.85 39.12 53.92 30.22 35.62 33.49 33.62 33.62 34.60 42.14 49.45 56.87 42.14 49.92		Flan-T5-Large	59.63	48.24	54.08	48.24	44.40	51.30	36.40	38.44
BLOOM-3b 35.85 39.12 53.92 30.22 35.62 33.69 33.49 34.69 52.62 62.64 46.04 52.64 48.21 48.21 34.00 38.89 44.06 55.43 36.29 59.05 49.26 49.26 49.26 49.26 49.24 42.04 41.04 45.88 40.91 41.04 42.92 49.24 42.92 49.24 42.93 45.64 42.92 42.94 49.45 50.87 44.23 42.92 42.94 49.45 50.87 44.23 42.92		BLOOM-1b7	33.16	31.04	33.16	32.83	33.77	33.53	36.67	32.92
LLaMA-7B 33.49 33.49 33.49 33.49 33.49 33.49 33.49 33.49 33.49 33.49 33.49 53.49 52.62 45.62 52.62 45.63 45.62 45.62 45.63 45.62 45.62 45.63 45.62 45.62 45.63 45.62	GeoNames	Flan-T5-XL	49.37	44.05	45.09	52.41	43.92	46.34	49.98	44.29
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 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 33.71 34.74 42.90 JMLS BLOOM-1b7 33.71 36.18 33.71 38.26 33.71 35.89 33.27 33.60 JMLS BLOOM-3b 33.16 37.23 34.82 35.77 33.16 35.9 33.25 35.60 JMLS BLOOM-3b 33.16 37.2		BLOOM-3b	35.85	39.12	53.92	30.22	35.62	33.60	48.26	37.73
GPT-3.5 59.40 47.79 67.78 41.95 48.02 51.72 45.25 43.66 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 BERT-Large 48.21 38.84 41.46 40.41 45.88 40.91 41.04 42.92 PubMedBERT 33.71 35.89 33.27 35.66 BLOOM-1b7 33.76 37.23 31.82 35.77 33.16 35.99 37.48 LLaMA-7B 32.94		LLaMA-7B	33.49	33.49	33.49	33.49	33.49	33.49	33.49	33.49
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 BERT-Large 48.21 38.84 41.46 40.41 45.88 40.91 41.04 42.92 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 UMLS BLOOM-1b7 33.71 36.18 33.71 38.26 33.71 35.89 33.25 35.75 44.93 45.46 46.07 42.83 51.25 41.18 42.90 44.94 42.91 50.77 42.33 33.05 37.48 42.92 43.294 32.94 32.94 32.94 <td></td> <td>GPT-3</td> <td>43.43</td> <td>51.74</td> <td>42.70</td> <td>53.20</td> <td>46.04</td> <td>52.56</td> <td>45.49</td> <td>52.62</td>		GPT-3	43.43	51.74	42.70	53.20	46.04	52.56	45.49	52.62
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 BERT-Large 48.21 38.84 41.46 40.41 45.88 40.91 41.04 42.92 PubMedBERT 33.71 35.89 33.27 35.60 UMLS 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		GPT-3.5	59.40	47.79	67.78	41.95	48.02	51.72	45.25	43.86
Flan-T5-XL* 48.41 34.80 55.23 46.96 57.48 36.29 59.05 49.26 BERT-Large 48.21 38.84 41.46 40.41 45.88 40.91 41.04 42.92 PubMedBERT 33.71 33.60 44.29 50.87 44.23 42.90 UMLS 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 37.93 32.94		GPT-4	38.56	52.46	34.00	38.89	44.06	55.43	33.78	36.23
BERT-Large 48.21 38.84 41.46 40.41 45.88 40.91 41.04 42.92 PubMedBERT 33.71 47.47 42.39 45.46 UMLS 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 32.94 32.94 32.94 32.94 32.94 32.94		$Flan$ - $T5$ - $Large^*$	42.53	59.40	40.29	62.46	46.03	57.41	42.49	62.04
PubMedBERT 33.71 33.71 33.71 33.71 33.71 33.71 33.71 33.71 33.71 33.71 33.71 33.71 33.71 33.71 33.71 33.71 33.71 33.71 33.71 34.66 Flan-T5-Large 47.55 51.22 55.32 40.94 49.45 50.87 44.23 42.90 UMLS 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		$Flan$ - $T5$ - XL^*	48.41	34.80	55.23	46.96	57.48	36.29	59.05	49.26
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 UMLS 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.		BERT-Large	48.21	38.84	41.46	40.41	45.88	40.91	41.04	42.92
UMLS Flan-T5-Large 47.55 51.22 55.32 40.94 49.45 50.87 44.23 42.90 UMLS 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		${\bf PubMedBERT}$	33.71	33.71	33.71	33.71	33.71	33.71	33.71	33.71
UMLS 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 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-Large 43.85 41.17 44.06 43.20 43.70 40.05 42.15 43.72 BART-Large 46.92 38.69 39.28 52.90 38.20 41.17 <		BART-Large	36.02	48.21	41.42	49.90	39.37	47.47	42.39	45.46
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 46.72 56.64 64.41 64.11 66.82 63.14 67.27 56.64 64.41		Flan-T5-Large	47.55	51.22	55.32	40.94	49.45	50.87	44.23	42.90
BLOOM-3b 33.16 37.23 34.82 35.77 33.16 35.89 33.05 37.48 LLaMA-7B 32.94 46.72 56.64 64.71 46.74 67.72 56.64 64.41 66.72 78.11 36.59 60.72 78.27 78.12 35.97 79.25 39.27 78.25 39.27 78.25 39.27 39.27 39.27 48.21 49.26 42.74 42.60 47.93 48.27 48.52 48.52 48.52	UMLS	BLOOM-1b7	33.71	36.18	33.71	38.26	33.71	35.89	33.27	33.60
LLaMA-7B 32.94 46.72 46.72 46.72 46.72 46.72 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 36.69 50.04 36.91 41.34 78.12 50.12 79.25 39.27 BART-Large 43.85 41.17 44.06 43.20 43.70 40.05 42.15 43.72 BLOOM-1b7 33.39 47.83 33.39 39.77 38.92 48.56 44.35 39.57 schema.org Flan-		Flan-T5-XL	64.25	46.53	51.00	41.54	60.07	42.83	51.25	41.18
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 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 schema.org Flan-T5-XL 42.70 33.45 33.59 39.77 38.92 48.56 44.35 39.57 schema.org 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 </td <td></td> <td>BLOOM-3b</td> <td>33.16</td> <td>37.23</td> <td>34.82</td> <td>35.77</td> <td>33.16</td> <td>35.89</td> <td>33.05</td> <td>37.48</td>		BLOOM-3b	33.16	37.23	34.82	35.77	33.16	35.89	33.05	37.48
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 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 schema.org 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 <td></td> <td>LLaMA-7B</td> <td>32.94</td> <td>32.94</td> <td>32.94</td> <td>32.94</td> <td>32.94</td> <td>32.94</td> <td>32.94</td> <td>32.94</td>		LLaMA-7B	32.94	32.94	32.94	32.94	32.94	32.94	32.94	32.94
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 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 schema.org 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 33.37 33.37 33.37 33.37 33.37 <td></td> <td>GPT-3</td> <td>51.58</td> <td>49.41</td> <td>49.86</td> <td>42.90</td> <td>50.57</td> <td>46.07</td> <td>45.36</td> <td>46.72</td>		GPT-3	51.58	49.41	49.86	42.90	50.57	46.07	45.36	46.72
$\begin{array}{c ccccccccccccccccccccccccccccccccccc$		GPT-3.5	61.38	70.38	63.91	66.82	63.14	67.27	56.64	64.41
Flan-T5-XL* 63.69 50.04 36.91 41.34 78.12 50.12 79.25 39.27 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 schema.org 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 33.37 33.37 48.63 48.87 49.48 GPT-3 56.84		GPT-4	41.19	76.99	42.55	63.88	50.28	78.11	36.59	60.72
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 schema.org 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 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		$Flan$ - $T5$ - $Large^*$	37.17	48.66	36.07	42.12	48.39	46.65	53.42	35.97
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 schema.org 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 33.37 33.37 33.37 49.48 GPT-3 49.64 49.28 50.97 48.03 47.19 48.63 48.87 49.48 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 <td></td> <td>$Flan$-$T5$-XL^*</td> <td>63.69</td> <td>50.04</td> <td>36.91</td> <td>41.34</td> <td>78.12</td> <td>50.12</td> <td>79.25</td> <td>39.27</td>		$Flan$ - $T5$ - XL^*	63.69	50.04	36.91	41.34	78.12	50.12	79.25	39.27
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 schema.org 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 33.37 49.48 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		BERT-Large	43.85	41.17	44.06	43.20	43.70	40.05	42.15	43.72
BLOOM-1b7 33.39 47.83 33.39 39.77 38.92 48.56 44.35 39.57 schema.org 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 33.37 33.37 49.48 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		BART-Large	34.62	38.69	39.28	52.90	38.20	41.17	43.26	42.74
schema.org 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 33.37 33.37 33.37 34.79 49.48 49.48 49.48 49.48 49.48 49.48 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-Large	46.98	49.92	46.11	54.78	40.27	54.47	42.06	47.93
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		BLOOM-1b7	33.39	47.83	33.39	39.77	38.92	48.56	44.35	39.57
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	schema.org	Flan-T5-XL	42.70	33.45	33.59	42.76	36.69	34.04	33.75	36.45
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		BLOOM-3b	41.64	47.16	47.98	45.25	39.73	40.75	51.28	48.73
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		LLaMA-7B	33.37	33.37	33.37	33.37	33.37	33.37	33.37	33.37
$ \begin{array}{llllllllllllllllllllllllllllllllllll$		GPT-3	49.64	49.28	50.97	48.03	47.19	48.63	48.87	49.48
$Flan-T5-Large^*$ 35.35 85.43 29.82 89.24 41.30 91.68 42.46 56.39		GPT-3.5	56.84			70.16	53.35	72.35	54.16	71.03
8		GPT-4	58.47	72.82	65.83	63.30	50.56	74.24	57.45	63.69
Flan-T5-XL* 91.06 57.46 74.68 65.32 91.54 50.63 91.70 33.33			35.35				41.30		42.46	56.39
		$Flan-T5-XL^*$	91.06	57.46	74.68	65.32	91.54	50.63	91.70	33.33