# A RelEntLess Benchmark for Modelling Graded Relations between Named Entities

**Anonymous ACL submission**


## Abstract

Relations such as “is influenced by”, “is known for” or ”is a competitor of” are inherently graded: we can rank entity pairs based on how well they satisfy these relations, but it is hard to draw a line between those pairs that satisfy them and those that do not. Such graded relations play a central role in many applications, yet they are typically not covered by existing Knowledge Graphs. In this paper, we consider the possibility of using Large Language Models (LLMs) to fill this gap. To this end, we introduce a new benchmark, in which entity pairs have to be ranked according to how much they satisfy a given graded relation. The task is formulated as a few-shot ranking problem, where models only have access to a description of the relation and five prototypical instances. We use the proposed benchmark to evaluate state-of-the-art relation embedding strategies as well as several publicly available LLMs and closed conversational models such as GPT-4. We find that smaller language models struggle to outperform a naive baseline. Overall, the best results are obtained with the 11B parameter Flan-T5 model and the 13B parameter OPT model, where further increasing the model size does not seem to be beneficial. For all models, a clear gap with human performance remains.


## 1 Introduction

Language Models (LMs) capture an abundance of factual and commonsense knowledge about the world (Petroni et al., 2019; Roberts et al., 2020; Heinzerling and Inui, 2021; West et al., 2022; Hao et al., 2022; Cohen et al., 2023). Given two entities, Large Language Models (LLMs) can straightforwardly be used to obtain a description of how these entities are related, although with some caveats for less popular entities (Mallen et al., 2022). However, relations are often a matter of degree (Rosch, 1975; Turney, 2006; Vulić et al., 2017). For instance, suppose we are interested in modelling whether one entity has been *influenced by* another one. While we could argue that most contemporary pop music has been influenced by the Beatles, clearly there are some bands that have been influenced more directly than others. Graded relations such as *influenced by*, *competitor of* or *similar to* are typically not found in traditional Knowledge Graphs (KGs), while they can nonetheless be of central importance to applications. For instance, in the context of financial NLP, we may need to know which companies are leaders and which are followers in a given field, who is competing with whom, and what strategic alliances exist. As another example, music recommendation systems often suggest artists based on the user’s listening history, but these suggestions would be more helpful if the system could identify artists that have influenced or were influenced by artists the user already likes, as opposed to merely identifying similar artists. Studying how such relations can be modelled is thus clearly an important but under-explored research problem.

The subjective nature of graded relations makes it difficult to include them in traditional KGs. Moreover, for many of these relations, it would simply not be feasible to list all the (graded) instances in a comprehensive way. Taking inspiration from existing work on extracting KGs from LLMs, we therefore ask the following question: *are current LLMs capable of modelling graded relations between named entities in a meaningful way?* The task of modelling graded relations offers a number of unique challenges for LLMs. First, since this is essentially a ranking task, designing suitable prompts is not straightforward. Second, the task requires making very fine-grained distinctions. For instance, while we can say that *Microsoft is known for Windows* and *Apple is known for MacOS*, the former statement represents a more prototypical instance of the *known for* relation, as Apple is perhaps best known for its hardware products (e.g. iPhone). It is currently unclear to what extent LLMs are able to capture such subtle differences. Finally, modelling graded relations requires comparing entities of different types. For instance, the *known for* relation has instances such as (*Microsoft*,*Windows*), (*the Beatles*, *Hey Jude*) and even (*France*,*wine*). Comparing instances of such a diverse nature poses a particular challenge, as such comparisons are almost never expressed in text.

In this paper, we introduce RELENTLESS[1], a new dataset aimed at furthering the study of graded relations between named entities. Our dataset covers five common graded relations: competitor/rival of, friend/ally of, influenced by, known for, and similar to. We evaluate the ability of LLMs to rank entity pairs according to how much they satisfy these relations, given a description of the relation and five prototypical examples. Analysing the performance of several recent LLMs (Chung et al., 2022; Iyer et al., 2022), including GPT-4 (OpenAI, 2023), we find the best models to achieve a Spearman rank correlation of around 0.6. This shows that recent LLMs capture fine-grained relational knowledge to a meaningful extent, while at the same time still leaving a significant gap with human performance. For the open-source LLMs, we find that while the largest models achieve strong results, smaller models fail to outperform a naive baseline based on fastText vectors (Bojanowski et al., 2017). GPT-3 performs well, albeit slightly below the best variants of Flan-T5 and OPT. Finally, we found ChatGPT and GPT-4 hard to use for this task, since the OpenAI API[2] does not allow computing perplexity scores. As a result, we were not able to outperform GPT-3 with these models.

## 2 Related Work

**Benchmarks for Graded Relations** RELENTLESS was inspired by the SemEval 2012 Task 2 dataset on modelling relational similarity (Jurgens et al., 2012), which we will refer to as *RelSim*. RelSim covers 79 fine-grained relations, which are organised into 10 categories, such as *part-whole* (e.g. car:engine), *attribute* (e.g. beggar:poor) and *cause-purpose* (enigma:puzzlement). For each of the fine-grained relations, a ranking of concept pairs is provided, which reflects how prototypical these pairs are as instances of the relation. However, RelSim only considers concepts, whereas our focus is on named entities. To the best of our knowledge, the problem of modelling relational similarity between named entities has not yet been considered.

HyperLex (Vulić et al., 2017) is focused on modelling hypernymy as a graded relation. It involves ranking concept pairs according to how prototypical they are of the hypernymy relation. As for RelSim, named entities were explicitly excluded from this dataset. More broadly, word similarity benchmarks also follow the format of ranking concept pairs according to the degree to which a graded relation is satisfied, i.e. similarity.

Benchmarks with analogy questions (Turney et al., 2003; Ushio et al., 2021b; Chen et al., 2022) also relate to the problem of modelling graded relations. These benchmarks typically follow a multiple-choice format, where one word pair is given (e.g. eye:seeing), and the system has to predict which among a given set of candidate answer pairs is most analogous to the query pair (e.g. ear:hearing). Most existing benchmarks again focus on concepts. Moreover, where named entities are involved, the task degenerates to predicting whether two entity pairs have the same relation, i.e. the problem of measuring degrees of relatedness is not considered for named entities.

**Language Models as Knowledge Bases** The idea of using language models as knowledge bases was popularised by Petroni et al. (2019), and has gained considerable further traction with the advent of LLMs. For instance, several authors have proposed strategies for extracting knowledge graphs from LLMs (West et al., 2022; Hao et al., 2022; Cohen et al., 2023). While the idea of modelling graded relations has not been considered, Hao et al. (2022) focused on relations that are not covered by traditional knowledge graphs, such as "is capable of but not good at". Similarly, our motivation for studying graded relations between named entities is also to complement what is captured by KGs.

## 3 Dataset

We consider the five relations which are shown in Table 1. These relations were chosen because of their graded character and because they can apply to a broad range of entities. We created a dataset with annotated entity pairs for each of the relations in three phases. We recruited a diverse annotation team in terms of age, gender, ethnicity and nationality; however, all annotators come from an academic setting: four undergraduate students, one PhD stu-

[1] The name RELENTLESS refers to Relations between Entities, where Less refers to the idea of ordering. The dataset will be made available upon the acceptance of the paper.

[2] https://openai.com/blog/openai-api

| Relation Type | Val | Test | Prototypical examples | Middle rank examples |
|---|---|---|---|---|
| competitor/rival of | 20 | 84 | Dell : HP, Sprite : 7 Up, Israel : Palestine, Liverpool FC : Manchester United, Microsoft Teams : Slack | Macallan : Suntory, Marvel Comics : D.C. Comics, Borussia Dortmund : PSG, UK : France, Doctor Who : Game of Thrones |
| friend/ally of | 20 | 88 | Australia : New Zealand, Aznar : Bush, Extinction Rebellion : Greta Thunberg, Elsa : Anna, CIA : MI6 | Kylo Ren : Rey, UK : Commonwealth, Darth Vader : Emperor Palpatine, The Beatles : Queen, Mark Drakeford : Rishi Sunak |
| influenced by | 20 | 90 | Europe : European Union, Plato : Socrates, Ethereum : Bitcoin, Messi : Maradona, Impressionism : Edouard Manet | Mike Tyson : Muhammad Ali, US : NASA, Acer : Asus, Vincent van Gogh : Bipolar disorder, Conservative Party : Labour Party |
| known for | 20 | 105 | Russell Crowe : Gladiator, Cadbury : chocolate, Paris : Eiffel Tower, Leonardo Da Vinci : Mona Lisa, Apple : iPhone | New Zealand : sheep, Le Corbusier : purism art, Sean Connery : Finding Forrester, Qualcomm : smartphones, Nikola Tesla : robotics |
| similar to | 20 | 89 | Coca-Cola : Pepsi, Ligue 1 : Bundesliga, Australia : New Zealand, The Avengers : The Justice League, Tesco : Sainsburys | NATO : United Nations, Iraq : Iran, cement : concrete, Cornwall : Brittany, Adele : Ed Sheeran |

Table 1: Overview of the considered relations, showing the numbers of entity pairs in the validation and test sets, the five prototypical training examples, and five examples from the middle of the ranking of the entity pairs in the validation set.

| | |
|---|---|
| 5: | This is clearly a positive example, and I would expect everyone to agree with this view. |
| 4: | I consider this to be a positive example, but I would not be surprised if some knowledgeable people consider this word pair to be borderline. |
| 3: | I consider this to be a borderline case: I find it hard to decide whether this is a positive or a negative example. |
| 2: | I consider this to be a negative example, but I would not be surprised if some knowledgeable people consider this word pair to be borderline. |
| 1: | This is clearly a negative example, and I would expect everyone to agree with this view. |

Table 2: Rating scale for the 2nd annotation phase.

dent and two faculty members. The students were recruited through an internal student employment service and were offered a remuneration of around $20 per hour. The total annotation effort was about 160 hours. The annotation process was split into three phases.

**First phase** In the first phase, the annotators were asked to provide 15 entity pairs for each of the five relations. Specifically, the aim was to provide 5 prototypical examples (i.e. entity pairs that clearly satisfy the relationship), 5 borderline positive pairs, which only satisfy the relationship to some extent, and 5 borderline negative pairs, which do not satisfy the intended relationship but are nonetheless related in a similar way. After removing duplicates, this resulted in an average of 114 entity pairs for each relation, and 573 pairs in total. We augmented these entity pairs with a number of randomly chosen entity pairs. The entities for these random pairs were selected from the 50,000 most popular Wikidata entities, in terms of the number of page views of the associated Wikipedia article.

**Second phase** In the second phase, each annotator scored all the entity pairs that were provided in phase 1, using the 5-point scale shown in Table 2. For this phase, annotators were encouraged to consult web sources (e.g. search engines such as Google) for a limited time in order to familiarize themselves with the considered entities, if needed. This was the most time-consuming annotation phase, taking almost 10 hours on average per annotator to complete.

**Third phase** The third and final phase was aimed at resolving disagreements between the annotations from the second phase. Specifically, for each entity pair where there was a difference of 3 points between the highest and the lowest score, the annotator(s) with a diverging view were asked to check their previous annotation, and to either update their score or to provide a justification. A total of 255 unique entity pairs were checked in this way (310 scores were checked in total). We subsequently verified the justifications that were provided. In 13 cases, the justifications suggested that the other annotators might have missed a salient point. For these cases, the annotators with the opposite view were asked to re-check their previous annotation.

| | A | B | C | D | E | F | G | Others |
|---|---|---|---|---|---|---|---|---|
| A | 100 | 62 | 81 | 71 | 75 | 75 | 75 | 84 |
| B | 62 | 100 | 61 | 57 | 62 | 57 | 60 | 66 |
| C | 81 | 61 | 100 | 73 | 72 | 74 | 75 | 84 |
| D | 71 | 57 | 73 | 100 | 67 | 67 | 70 | 77 |
| E | 75 | 62 | 72 | 67 | 100 | 70 | 72 | 77 |
| F | 75 | 57 | 74 | 67 | 70 | 100 | 69 | 76 |
| G | 75 | 60 | 75 | 70 | 72 | 69 | 100 | 79 |
| AVG | 77 | 66 | 77 | 72 | 74 | 73 | 74 | 77 |

Table 3: Spearman correlation (%) between each pair of annotators (A,...,G), and between each annotator and the average score provided by the other six averaged over all the five relation types after the 3rd and final quality enhancement annotation round.

The final ranking for each relation was obtained by averaging the scores of the 7 annotators.

Table 3 summarises the agreement between the annotators in terms of Spearman's rank correlation.[3] The table shows the correlation between the individual annotators, as well as the correlation between each annotator and the average of the scores from the six other annotators. The reconciliation step improved the average agreement over all the annotators from 70 to 77.[4]

We split the annotated entity pairs as follows. First, we selected a small training set consisting of five prototypical pairs for each relation. This training set could be used, for instance, for few-shot prompting strategies. The entity pairs were selected (i) to be among the top-ranked entity pairs and (ii) to be sufficiently diverse (i.e. including entities of different types). Next, for each relation, we randomly selected 20 of the remaining entity pairs to be used as a validation set.[5] The remaining entity pairs constitute the test set. Table 1 shows the prototypical entity pairs that were selected for each relation, as well as five examples of entity pairs from the validation set. The latter were selected from the middle of the ranking, typically with an average score of 3 to 4. We use the Spearman rank correlation between the predicted ranking and the ground truth ranking as the evaluation metric.[6]

[3] In Appendix A, we include the breakdown of the annotator agreement scores per relation type.

[4] Details about the agreement before the reconciliation step can be found in the appendix.

[5] This validation set was not used in our main experiments, but it was considered in the few-shot analysis (see subsection 6.2). However, we release the full validation set so it can be used for further testing and experimentation without the risk of overfitting on the test set

[6] The final annotated dataset, along with the guidelines provided to annotators in each phase, are available in the supplementary material.

## 4 Baselines

**Human Performance** As a proxy for human performance, we report the average Spearman rank correlation between each annotator and the average of the other annotators, referred to as *Human Upperbound*. Please note that this upperbound is computed based on the test set, and thus slightly differs from the average agreement in Table 3. Furthermore, note that we only estimate human performance to provide a reference for interpreting the results. Doing this accurately is challenging. For instance, we can already see large differences in agreement across the different annotators, suggesting that the best annotators would perform much better than what is suggested by the given upperbound. Conversely, one may also argue that because of the reconciliation step in the third phrase, we are overestimating human performance.

### 4.1 Embedding Models

**Word Embedding.** First, we consider the fastText (Bojanowski et al., 2017) embeddings that were trained on Common Crawl with subword information[7]. Inspired by the tradition of modelling word analogies using vector differences (Mikolov et al., 2013), we represent each entity pair by subtracting the fastText embedding of the first entity from the embedding of the second entity. We refer to the resulting vector as the fastText relation embedding. For a given relation, we score an entity pair by taking the maximum cosine similarity between its fastText relation embedding and the embedding of the five prototypical examples.[8] We use the maximum, rather than e.g. the average, due to the diverse nature of these prototypical examples. We refer this approach as fastText$_{\text{pair}}$.

As a naive baseline, we also consider a variant in which an entity pair is scored by taking the cosine similarity between the word embeddings of the two entities. Note that this baseline ignores both the description of the relation and the prototypical examples. It is based on the idea that prototypical pairs often involve closely related entities. We refer to this approach as fastText$_{\text{word}}$.

**RelBERT.** RelBERT (Ushio et al., 2021a) is a RoBERTa model that was fine-tuned to encode word pairs such that analogous word pairs are represented by similar vectors. We use RelBERT models

[7] https://fasttext.cc/

[8] Empirically, we confirmed that indeed using the maximum leads to better results overall.

that were initialised from $\text{RoBERTa}_{\text{BASE}}$[9] and from $\text{RoBERTa}_{\text{LARGE}}$[10]. For a given relation, we score each entity pair as the maximum cosine similarity between its RelBERT encoding and the RelBERT encoding of the five prototypical examples.

### 4.2 Language Models

To score entity pairs using LMs, we create a prompt from the description of the relation and the five prototypical examples. The score of the entity pair then corresponds to the perplexity of the prompt. We consider two prompt templates: a binary question answering (QA) template similar to the instructions provided to Flan-T5 for the task (Longpre et al., 2023), and a targeted list completion template (LC). Writing the five prototypical examples as $[A_i, B_i]_{i=1...5}$ and the target entity pair as $[C, D]$, the QA template has the following form:

> Answer the question by yes or no. We know that $[A_1, B_1], \ldots, [A_5, B_5]$ are examples of <desc>. Are $[C, D]$ <desc> as well?
> Yes

The LC template has the following form:

> Complete the following list with examples of <desc>
> $[A_1, B_1]$
> $\vdots$
> $[A_5, B_5]$
> $[C, D]$

In both templates, <desc> is the description of the relation, as follows:

- *Rival:* entities that are competitors or rivals
- *Ally:* entities that are friends or allies
- *Inf:* what has influenced different entities
- *Know:* what entities are known for
- *Sim:* entities that are similar

We use the following LMs: OPT (Zhang et al., 2022), OPT-IML (Iyer et al., 2022), T5 (Raffel et al., 2020), Flan-T5 (Chung et al., 2022), and Flan-UL2 (Tay et al., 2023), where the model weights are obtained via HuggingFace (Wolf et al., 2020)[11]. We also use GPT-3 (Brown et al., 2020), which is a private model and subject to be changed every six months; we use `davinci`, which is the most powerful GPT-3 model available via the OpenAI API [12][13]. We compute the perplexity over the whole input text for OPT, OPT-IML and GPT-3, while we use the last line of the input text (i.e., "Yes" for the QA template and $[C, D]$ for the LC template) to compute the perplexity on the decoder for T5, Flan-T5, and Flan-UL2.

We test two conversational LMs: ChatGPT (or `gpt-3.5-turbo`) and GPT-4 (`gpt-4`). These models are only available through the OpenAI API. Unfortunately, for these models, the API does not allow us to obtain the log-likelihood of each token. Therefore, we instead use a prompt which asks to sort the list of entity pairs directly[14].

## 5 Results

Table 4 summarises the results. The best result is achieved by $\text{Flan-T5}_{\text{XXL}}$ with the QA template, which scores 62.0%. In general, the performance of this model remains far below the performance upper bound suggested by the inter-annotator agreement (77%). Surprisingly, however, for the *rival of* relation, the human upper bound is outperformed by Flan-UL2. In contrast, the *friend/ally of* relation appears to be particularly challenging. Among the LM methods, the LC template generally leads to the best results, but not for Flan-T5 and Flan-UL2. This is not entirely surprising given that Flan models have been fine-tuned using instructions similar to the QA template (see subsection 4.2). Beyond the encoder-decoder LMs, $\text{OPT}_{\text{13B}}$ and $\text{GPT-3}_{\text{davinci}}$ perform the best, even outperforming the instruction fine-tuned OPTs (OPT-IML and $\text{OPT-IML}_{\text{MAX}}$). $\text{GPT-3}_{\text{davinci}}$ is the best model in the *influenced by* and *known for* relations. Although $\text{Flan-T5}_{\text{XXL}}$ and Flan-UL2 perform best on average, they perform poorly on the *influenced by* relation, underperforming $\text{GPT-3}_{\text{davinci}}$ and $\text{OPT}_{\text{13B}}$ by a wide margin. Among the embedding based models, fastText generally performs poorly. The performance of $\text{RelBERT}_{\text{LARGE}}$ is remarkably strong, considering that this is a small concept-based relation model that was not trained on relations between named en-

[9] https://huggingface.co/relbert/relbert-roberta-base

[10] https://huggingface.co/relbert/relbert-roberta-large

[11] A complete list of the models on huggingface we used can be found in Appendix B.

[12] https://openai.com

[13] All the OpenAI models are from the checkpoint that was live during May 2023.

[14] A complete prompt can be found in Appendix C

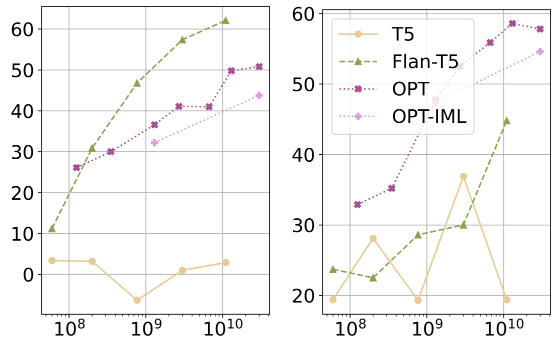
| | | | Model | Inst-FT | Model Size | Rival | Ally | Inf | Know | Sim | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Human Upperbound* | | | | | | 75.9 | 78.0 | 70.5 | 82.0 | 80.2 | 77.3 |
| Embedding | | | fastText$_{\text{word}}$ | | - | 25.0 | 10.0 | 7.0 | 24.0 | 20.0 | 17.0 |
| Embedding | | | fastText$_{\text{pair}}$ | | - | 28.0 | 12.0 | 3.0 | 20.0 | 21.0 | 17.0 |
| Embedding | | | RelBERT$_{\text{BASE}}$ | | 110M | 58.0 | 15.0 | 30.0 | 24.0 | 28.0 | 31.0 |
| Embedding | | | RelBERT$_{\text{LARGE}}$ | | 335M | 64.0 | 20.0 | 20.0 | 44.0 | 53.0 | 40.0 |
| LM | *LC template* | T5 | T5$_{\text{SMALL}}$ | | 60M | 20.0 | 33.0 | 24.0 | 11.0 | 10.0 | 19.0 |
| LM | *LC template* | T5 | T5$_{\text{BASE}}$ | | 220M | 35.0 | 35.0 | 38.0 | 20.0 | 13.0 | 28.0 |
| LM | *LC template* | T5 | T5$_{\text{LARGE}}$ | | 770M | 29.0 | 8.0 | 26.0 | 11.0 | 22.0 | 19.0 |
| LM | *LC template* | T5 | T5$_{\text{XL}}$ | | 3B | 47.0 | 28.0 | 50.0 | 33.0 | 26.0 | 37.0 |
| LM | *LC template* | T5 | T5$_{\text{XXL}}$ | | 11B | 33.0 | 8.0 | 24.0 | 18.0 | 15.0 | 19.0 |
| LM | *LC template* | T5 | Flan-T5$_{\text{SMALL}}$ | ✓ | 60M | 38.0 | 33.0 | 24.0 | 16.0 | 7.0 | 24.0 |
| LM | *LC template* | T5 | Flan-T5$_{\text{BASE}}$ | ✓ | 220M | 36.0 | 31.0 | 28.0 | 17.0 | -0.0 | 22.0 |
| LM | *LC template* | T5 | Flan-T5$_{\text{LARGE}}$ | ✓ | 770M | 41.0 | 19.0 | 36.0 | 24.0 | 22.0 | 29.0 |
| LM | *LC template* | T5 | Flan-T5$_{\text{XL}}$ | ✓ | 3B | 40.0 | 17.0 | 35.0 | 27.0 | 31.0 | 30.0 |
| LM | *LC template* | T5 | Flan-T5$_{\text{XXL}}$ | ✓ | 11B | 61.0 | 32.0 | 47.0 | 44.0 | 40.0 | 45.0 |
| LM | *LC template* | T5 | Flan-UL2 | ✓ | 20B | 60.0 | 28.0 | 49.0 | 53.0 | 37.0 | 45.0 |
| LM | *LC template* | OPT | OPT$_{\text{125M}}$ | | 125M | 41.0 | 37.0 | 51.0 | 23.0 | 13.0 | 33.0 |
| LM | *LC template* | OPT | OPT$_{\text{350M}}$ | | 300M | 41.0 | 33.0 | 47.0 | 36.0 | 18.0 | 35.0 |
| LM | *LC template* | OPT | OPT$_{\text{1.3B}}$ | | 1.3B | 58.0 | 39.0 | 54.0 | 45.0 | 42.0 | 48.0 |
| LM | *LC template* | OPT | OPT$_{\text{13B}}$ | | 13B | 72.0 | 41.0 | 55.0 | 70.0 | 55.0 | 59.0 |
| LM | *LC template* | OPT | OPT$_{\text{30B}}$ | | 30B | 71.0 | 39.0 | 57.0 | 69.0 | 53.0 | 58.0 |
| LM | *LC template* | OPT | OPT-IML$_{\text{30B}}$ | ✓ | 30B | 65.0 | 36.0 | 55.0 | 70.0 | 47.0 | 55.0 |
| LM | *LC template* | OPT | OPT-IML$_{\text{MAX-30B}}$ | ✓ | 30B | 62.0 | 36.0 | 57.0 | 67.0 | 46.0 | 53.0 |
| LM | *LC template* | GPT | GPT-3$_{\text{davinci}}$* | | - | 72.0 | 39.0 | **64.0** | **73.0** | 47.0 | 59.0 |
| LM | *QA template* | T5 | T5$_{\text{SMALL}}$ | | 60M | 10.0 | -13.0 | 17.0 | -6.0 | 8.0 | 3.0 |
| LM | *QA template* | T5 | T5$_{\text{BASE}}$ | | 220M | 15.0 | -7.0 | 6.0 | -12.0 | 14.0 | 3.0 |
| LM | *QA template* | T5 | T5$_{\text{LARGE}}$ | | 770M | -3.0 | 4.0 | -12.0 | -19.0 | -1.0 | -6.0 |
| LM | *QA template* | T5 | T5$_{\text{XL}}$ | | 3B | -2.0 | 12.0 | -8.0 | 17.0 | -14.0 | 1.0 |
| LM | *QA template* | T5 | T5$_{\text{XXL}}$ | | 11B | 7.0 | 1.0 | -1.0 | 11.0 | -4.0 | 3.0 |
| LM | *QA template* | T5 | Flan-T5$_{\text{SMALL}}$ | ✓ | 60M | 31.0 | -0.0 | 21.0 | -3.0 | 8.0 | 11.0 |
| LM | *QA template* | T5 | Flan-T5$_{\text{BASE}}$ | ✓ | 220M | 41.0 | 28.0 | 46.0 | 17.0 | 22.0 | 31.0 |
| LM | *QA template* | T5 | Flan-T5$_{\text{LARGE}}$ | ✓ | 770M | 67.0 | 39.0 | 24.0 | 49.0 | 56.0 | 47.0 |
| LM | *QA template* | T5 | Flan-T5$_{\text{XL}}$ | ✓ | 3B | 75.0 | 44.0 | 44.0 | 61.0 | 63.0 | 57.0 |
| LM | *QA template* | T5 | Flan-T5$_{\text{XXL}}$ | ✓ | 11B | 74.0 | **56.0** | 44.0 | 70.0 | 66.0 | **62.0** |
| LM | *QA template* | T5 | Flan-UL2 | ✓ | 20B | **79.0** | 51.0 | 47.0 | 67.0 | 57.0 | 60.0 |
| LM | *QA template* | OPT | OPT$_{\text{125M}}$ | | 125M | 35.0 | 31.0 | 46.0 | 10.0 | 9.0 | 26.0 |
| LM | *QA template* | OPT | OPT$_{\text{350M}}$ | | 350M | 38.0 | 35.0 | 37.0 | 21.0 | 19.0 | 30.0 |
| LM | *QA template* | OPT | OPT$_{\text{1.3B}}$ | | 1.3B | 44.0 | 33.0 | 46.0 | 29.0 | 31.0 | 37.0 |
| LM | *QA template* | OPT | OPT$_{\text{13B}}$ | | 13B | 63.0 | 39.0 | 43.0 | 61.0 | 43.0 | 50.0 |
| LM | *QA template* | OPT | OPT$_{\text{30B}}$ | | 30B | 61.0 | 38.0 | 48.0 | 62.0 | 45.0 | 51.0 |
| LM | *QA template* | OPT | OPT-IML$_{\text{30B}}$ | ✓ | 30B | 57.0 | 37.0 | 36.0 | 53.0 | 35.0 | 44.0 |
| LM | *QA template* | OPT | OPT-IML$_{\text{MAX-30B}}$ | ✓ | 30B | 58.0 | 36.0 | 39.0 | 43.0 | 42.0 | 43.0 |
| LM | *QA template* | GPT | GPT-3$_{\text{davinci}}$* | | - | 67.0 | 35.0 | 50.0 | 61.0 | 35.0 | 50.0 |
| Conv. LM | | | ChatGPT* | | - | -0.9 | 32.5 | 17.5 | 15.5 | 14.7 | 17.9 |
| Conv. LM | | | GPT-4* | | - | 62.5 | 55.8 | 35.9 | 60.8 | **69.3** | 56.9 |

Table 4: Spearman's rank correlation (%) on the test set. The LMs are grouped by the template (QA or LC), the model family, and instruction-fine-tuned or not. The best correlation in each relation type is highlighted by bold characters. Model size is measured as the number of parameters. Models marked with * are not openly available.

tities. As far as the OpenAI conversational models are concerned, we can see that GPT-4 achieves the best result on the *similar to* relation. The poor performance of ChatGPT suggests that the considered list ranking prompt may be hard to understand for this model, or that the task of ranking around 100 pairs may be too complicated. We also observed that ChatGPT tends to omit more pairs from its output than GPT-4 (see Appendix D).

## 6 Analysis

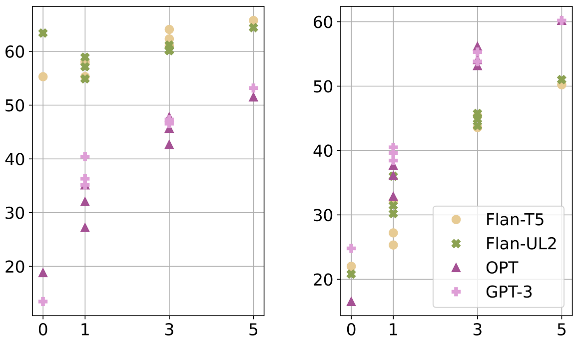
We now aim to gain a better understanding of the behaviour of LMs. First, we analyse the effect of model size (subsection 6.1). Then, we experiment




(a) QA template

(b) LC template

Figure 1: Average Spearman's rank correlation results among the five relation types along with the model size.




(a) QA template

(b) LC template

Figure 2: Spearman's rank correlation averaged over the five relation types with different number of the prototypical examples. For 1-shot and 3-shot examples, we report the each correlation of the three individual runs.

with different zero-shot and few-shot learning setups (subsection 6.2), and we present a qualitative analysis of the predictions (subsection 6.3). For the latter two analyses, we focus on the best performing models for each LM family from the main experiment, using their optimal prompts: Flan-UL2, Flan-T5$_{\text{XXL}}$, OPT$_{\text{13B}}$, and GPT-3$_{\text{davinci}}$.[15]

### 6.1 Model Size

In this section, we analyse the effect of model size. Figure 1 visualises the performance of the different model families in function of model size. For Flan-T5, OPT, and OPT-IML we can see a strong correlation between performance and size. Nevertheless, the result of the largest OPT models suggests that a plateau in performance may have been reached at 13B. Moreover, for T5 we do not see an improvement in performance for larger models[16].

### 6.2 Zero-shot/Few-shot Learning

In the main experiments, for each relation, models had access to a description as well as five prototypical examples. To analyse the impact of these five examples, we now describe experiments in which only the description is provided (i.e. zero-shot) or where only 1 or 3 examples are given (few-shot). For the few-shot setting, we use the same QA and LC templates as in the main experiment. For the 3-shot experiments, we randomly choose 3 of the 5 examples, and similar for the 1-shot experiments. Since this introduces some randomness, we report results for three different samples[17].

Figure 2a shows the results for the QA template. We can see that all models improve when more prototypical examples are provided, with the zero-shot performance of Flan-UL2 being an outlier. Remarkably, Flan-UL2 achieves 62.5% accuracy in the zero-shot setting, which is competitive with the 5-shot results in Table 4. Flan-T5$_{\text{XXL}}$ also achieves a zero-shot result of 54.5%, which is better than most of the models in the main (5-shot) experiments. In the zero-shot setting, OPT$_{\text{13B}}$ performs better than GPT-3$_{\text{davinci}}$, but GPT-3$_{\text{davinci}}$ quickly improves as more examples are provided, clearly outperforming OPT$_{\text{13B}}$ in the 5-shot setting. Figure 2b shows the results for the LC template. We again see that providing more examples benefits all models. Unlike for the QA template, however, Flan-T5$_{\text{XXL}}$ performs poorly in the zero-shot setting. Moreover, OPT$_{\text{13B}}$ now sees the largest improvement between the zero-shot and 5-shot settings.

### 6.3 Qualitative Analysis

To better understand the predictions of the models, we analyse the most flagrant mistakes. Specifically, we focus on those entity pairs whose predicted rank is in the top 30%, while being in the bottom 30% of the gold ranking, and vice versa. Table 5 and Table 6 show the entity pairs from the test set for which this was the case. For this analysis, we look at the models with their optimal templates: i.e., Flan-T5 and Flan-UL2 with the QA template, and the other models with the LC template.

When looking at the instances that mistakenly end up in the top 30%, we see entities which are closely related (e.g. "Coca-Cola : Pepsi") while not actually satisfying the intended relation. We can see several cases where entities with similar

[15]Note that we omit Flan-UL2 from the model size analysis as there is only a single Flan-UL2 model.

[16]In Appendix E we include a more detailed breakdown of the results of this model size experiment by relation type.

[17]The prompt used in the zero-shot/few-shot learning can be found in Appendix F

| Model | Relation | Incorrectly predicted to be in the top 30% |
|---|---|---|
| Flan-T5$_{\text{XXL}}$ | Ally | Armenia : Azerbaijan, Liam Gallagher : Noel Gallagher, Russia : Georgia |
| | Inf | Harry Potter : Wizard of Oz, heavy metal : punk music, Luke Bryan : Hank Williams, James Brown : Michael Jackson |
| | Sim | sphinx : sphynx, New York : York, cannoli : canneloni |
| Flan-UL2 | Rival | Serena Williams : Andy Murray |
| | Ally | Liam Gallagher : Noel Gallagher, Google : Samsung |
| | Inf | Harry Potter : Wizard of Oz, heavy metal : punk music, James Brown : Michael Jackson |
| | Know | Belgium : wine |
| | Sim | sphinx : sphynx, cannoli : canneloni |
| OPT$_{\text{13B}}$ | Rival | Serena Williams : Andy Murray |
| | Ally | Joseph Stalin : Josip Broz Tito, Armenia : Azerbaijan, Sophia Loren : Marlon Brando |
| | Inf | Joe Biden : Donald Trump, Harry Potter : Wizard of Oz, Singaporean food : Malaysian food |
| | Know | Coca-Cola : Pepsi, Steve Jobs : AirPods |
| GPT-3$_{\text{davinci}}$ | Rival | Serena Williams : Andy Murray |
| | Ally | Joseph Stalin : Josip Broz Tito, Armenia : Azerbaijan, Liam Gallagher : Noel Gallagher |
| | Inf | Harry Potter : Wizard of Oz |
| | Know | Coca-Cola : Pepsi |
| | Sim | Nicolae Ceaușescu : Javier Hernández |

Table 5: Test examples of incorrect predictions made by the three best models in the top 30%.

| Model | Relation | Incorrectly predicted to be in the bottom 30% |
|---|---|---|
| Flan-T5$_{\text{XXL}}$ | Rival | Isaac Newton : Gottfried Leibniz |
| | Ally | China : North Korea, Ron Weasley : Neville Longbottom, Windows : Xbox |
| | Inf | Prince Harry : Monarchy, trending music : TikTok, Coca-Cola : Pepsi, Apple Music : Spotify, Pepsi : Coca-Cola, Hoover : Dyson |
| | Know | Corsica : Napoleon Bonaparte, France : cheese |
| | Sim | Suits : Law&Order, Shark : Bush |
| Flan-UL2 | Ally | Tata Motors : Jaguar, China : North Korea, HSBC : BlackRock, Coca-Cola : McDonald's, Huawei : China |
| | Inf | Prince Harry : Monarchy, trending music : TikTok, Wales : Westminster, Theresa May : David Cameron |
| | Know | Europe : The Final Countdown, Corsica : Napoleon Bonaparte, OpenAI : ChatGPT |
| | Sim | Minnesota : Wisconsin, Shark : Bush, Glastonbury : Roskilde |
| OPT$_{\text{13B}}$ | Ally | FTX : Alameda Research, Red Bull : GoPro, HSBC : BlackRock, Microsoft : LinkedIn, Windows : Xbox |
| | Inf | Prince Harry : Monarchy, trending music : TikTok, Wales : Westminster |
| | Know | OpenAI : ChatGPT, UK : rain |
| | Sim | pill : tablet, Great Britian : British Empire, fusilli : rotini, Shark : Bush |
| GPT-3$_{\text{davinci}}$ | Rival | Netflix : Disney Plus |
| | Ally | FTX : Alameda Research, Rishi Sunak : Joe Biden, Microsoft : LinkedIn, Windows : Xbox |
| | Inf | Prince Harry : Monarchy, trending music : TikTok, Stephen King : Arthur Machen |
| | Know | OpenAI:ChatGPT |
| | Sim | Homebase : IKEA, fusilli : rotini, Shark : Bush, Primark : Shein |

Table 6: Test examples of incorrect predictions made by the three best models in the bottom 30%.

names are mistakenly predicted to be similar (e.g. sphinx : sphynx, New York : York, cannoli : canneloni). Several models also mistakenly predict "Serena Williams : Andy Murray" as an instance of the rival-of relation, presumably because the model has learned that players from the same sport are often rivals. When looking at the examples from the bottom 30%, we can see entities which only recently became prominent (e.g. FTX and Alameda Research), highlighting the limitation of using language models that have not been trained on the most recent data. The "Corsica : Napoleon Bonaparte", "Prince Harry : Monarchy" and "trending music : TikTok" examples illustrate how the models can struggle with cases involving entities of different semantic types.

## 7 Conclusions

In this paper, we have proposed the task of modelling graded relations between named entities, with a new dataset. The task consists in ranking entity pairs according to how much they satisfy a given graded relation, where models only have access to the description of the relation and five prototypical instances per relation. To assess the difficulty of the task, we analysed a large number of baselines, including public LLMs of up to 30B parameters, state-of-the-art relation embedding models, and closed LLMs such as GPT-4. We found significant performance differences between the largest LMs and their smaller siblings, which highlights the progress achieved in NLP in the last few years by scaling up LMs. However, even the largest models trail human performance by around 15 percentage points.

## Limitations

Our dataset is aimed at testing the ability of LMs to understand graded relations between named entities. In particular, the size of the dataset makes it unsuitable for training models (beyond the few-shot setting). Furthermore, our dataset is limited to five relation types. We believe these relations to be among the most prominent graded relations between named entities. Nonetheless, there are clearly various other relations that could be considered, especially in domain-specific settings. While the annotation process involved comprehensive quality control mechanisms, the dataset may have inherited some of the biases of the annotators. The annotators were diverse in terms of gender, nationality and cultural background, but all came from the the same academic setting. Since the annotation is inherently subjective, this may be reflected in the final dataset. Finally, the task may have a temporal component in which some relationships may change over time. Our annotations represents the views of the annotators at a particular moment in time. In future, the dataset could be extended, to provide different temporal snapshots, which would allow an evaluation of ability of LMs to model temporal context.

## Ethics Statement

Our data has been created and labelled by human annotators. As such, we have ensured that proper training was provided, and that annotators were paid fairly through our institutional student job provider. We also acknowledge the potential biases of our dataset, and the potentially sensitive nature of examples related to political or religious content. To mitigate this issue, we have relied on a diverse set of annotators, and we have provided guidelines about avoiding sensitive content.

## References


Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the Annual Conference on Neural Information Processing Systems*.

Jiangjie Chen, Rui Xu, Ziquan Fu, Wei Shi, Zhongqiao Li, Xinbo Zhang, Changzhi Sun, Lei Li, Yanghua Xiao, and Hao Zhou. 2022. E-KAR: A benchmark for rationalizing natural language analogical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3941–3955, Dublin, Ireland. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Roi Cohen, Mor Geva, Jonathan Berant, and Amir Globerson. 2023. Crawling the internal knowledge-base of language models. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1856–1869, Dubrovnik, Croatia. Association for Computational Linguistics.

Shibo Hao, Bowen Tan, Kaiwen Tang, Hengzhe Zhang, Eric P Xing, and Zhiting Hu. 2022. Bertnet: Harvesting knowledge graphs from pretrained language models. *arXiv preprint arXiv:2206.14268*.

Benjamin Heinzerling and Kentaro Inui. 2021. Language models as knowledge bases: On entity representations, storage capacity, and paraphrased queries. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1772–1791, Online. Association for Computational Linguistics.

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Dániel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, et al. 2022. Opt-iml: Scaling language model instruction meta learning through the lens of generalization.

David Jurgens, Saif Mohammad, Peter Turney, and Keith Holyoak. 2012. SemEval-2012 task 2: Measuring degrees of relational similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 356–364, Montréal, Canada. Association for Computational Linguistics.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Hannaneh Hajishirzi, and Daniel Khashabi. 2022. When not to trust language models: Investigating effectiveness and limitations of parametric and non-parametric memories. *arXiv preprint arXiv:2212.10511*.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *arXiv*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

Eleanor Rosch. 1975. Cognitive representations of semantic categories. *Journal of experimental psychology: General*, 104(3):192.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Siamak Shakeri, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. 2023. Ul2: Unifying language learning paradigms.

Peter D. Turney. 2006. Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416.

Peter D. Turney, Michael L. Littman, Jeffrey Bigham, and Victor Shnayder. 2003. Combining independent modules in lexical multiple-choice problems. In *Recent Advances in Natural Language Processing III*, pages 101–110.

Asahi Ushio, Jose Camacho-Collados, and Steven Schockaert. 2021a. Distilling relation embeddings from pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9044–9062, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Asahi Ushio, Luis Espinosa Anke, Steven Schockaert, and Jose Camacho-Collados. 2021b. BERT is to NLP what AlexNet is to CV: Can pre-trained language models identify analogies? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3609–3624, Online. Association for Computational Linguistics.

Ivan Vulić, Daniela Gerz, Douwe Kiela, Felix Hill, and Anna Korhonen. 2017. HyperLex: A large-scale evaluation of graded lexical entailment. *Computational Linguistics*, 43(4):781–835.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic knowledge distillation: from general language models to commonsense models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4602–4625, Seattle, United States. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models.

## A Annotator Agreement for each Relation

Table 7 show the Spearman correlation for each relation type as well as the average over all the relation types before the 3rd and final quality enhancement annotation round. Table 8, Table 9, Table 10, Table 11, and Table 12 show the Spearman

| | A | B | C | D | E | F | G | Others |
|---|---|---|---|---|---|---|---|---|
| A | 100 | 53 | 77 | 63 | 64 | 68 | 67 | 80 |
| B | 53 | 100 | 52 | 43 | 47 | 46 | 48 | 56 |
| C | 77 | 52 | 100 | 63 | 58 | 67 | 68 | 79 |
| D | 63 | 43 | 63 | 100 | 48 | 54 | 59 | 66 |
| E | 64 | 47 | 58 | 48 | 100 | 57 | 59 | 65 |
| F | 68 | 46 | 67 | 54 | 57 | 100 | 62 | 70 |
| G | 67 | 48 | 68 | 59 | 59 | 62 | 100 | 73 |
| AVG | 70 | 55 | 69 | 61 | 62 | 65 | 66 | 70 |

Table 7: Spearman correlation (%) between each pair of annotators (A,...,G), and between each annotator and the average score provided by the other six averaged over all the five relation types **before the 3rd and final quality enhancement annotation round.**

| | A | B | C | D | E | F | G | Others |
|---|---|---|---|---|---|---|---|---|
| A | 100 | 55 | 79 | 69 | 74 | 78 | 79 | 86 |
| B | 55 | 100 | 46 | 35 | 58 | 57 | 50 | 54 |
| C | 79 | 46 | 100 | 75 | 67 | 73 | 75 | 80 |
| D | 69 | 35 | 75 | 100 | 52 | 66 | 68 | 74 |
| E | 74 | 58 | 67 | 52 | 100 | 69 | 67 | 74 |
| F | 78 | 57 | 73 | 66 | 69 | 100 | 65 | 79 |
| G | 79 | 50 | 75 | 68 | 67 | 65 | 100 | 79 |
| AVG | 76 | 57 | 74 | 66 | 70 | 73 | 72 | 75 |

Table 8: Spearman correlation (%) on the *competitor/rival of* relation between each pair of annotators (A,...,G), and between each annotator and the average score provided by the other six **after the 3rd and final quality enhancement annotation round.**

| | A | B | C | D | E | F | G | Others |
|---|---|---|---|---|---|---|---|---|
| A | 100 | 73 | 85 | 69 | 74 | 78 | 73 | 87 |
| B | 73 | 100 | 74 | 52 | 64 | 72 | 65 | 75 |
| C | 85 | 74 | 100 | 68 | 72 | 77 | 74 | 87 |
| D | 69 | 52 | 68 | 100 | 63 | 59 | 65 | 69 |
| E | 74 | 64 | 72 | 63 | 100 | 67 | 70 | 76 |
| F | 78 | 72 | 77 | 59 | 67 | 100 | 75 | 80 |
| G | 73 | 65 | 74 | 65 | 70 | 75 | 100 | 78 |
| Avg | 79 | 71 | 78 | 68 | 73 | 76 | 75 | 79 |

Table 9: Spearman correlation (%) on the *friend/ally of* relation between each pair of annotators (A,...,G), and between each annotator and the average score provided by the other six **after the 3rd and final quality enhancement annotation round.**

correlation for each relation type after the 3rd and final quality enhancement annotation round.

## B Models on HuggingFace

Table 13 shows the model alias on the HuggingFace of the LMs we used in our experiment.

## C Conversational Model Baselines

Writing the list of target word pairs as $[C_i, D_i]_{i=1\dots n}$, our prompt has the following

| | A | B | C | D | E | F | G | Others |
|---|---|---|---|---|---|---|---|---|
| A | 100 | 50 | 76 | 68 | 69 | 59 | 71 | 76 |
| B | 50 | 100 | 55 | 63 | 49 | 32 | 54 | 55 |
| C | 76 | 55 | 100 | 74 | 70 | 69 | 76 | 84 |
| D | 68 | 63 | 74 | 100 | 65 | 52 | 70 | 76 |
| E | 69 | 49 | 70 | 65 | 100 | 65 | 71 | 71 |
| F | 59 | 32 | 69 | 52 | 65 | 100 | 62 | 61 |
| G | 71 | 54 | 76 | 70 | 71 | 62 | 100 | 78 |
| AVG | 70 | 58 | 74 | 70 | 70 | 63 | 72 | 71 |

Table 10: Spearman correlation (%) on the *influenced by* relation between each pair of annotators (A,...,G), and between each annotator and the average score provided by the other six **after the 3rd and final quality enhancement annotation round.**

| | A | B | C | D | E | F | G | Others |
|---|---|---|---|---|---|---|---|---|
| A | 100 | 74 | 84 | 78 | 80 | 80 | 77 | 88 |
| B | 74 | 100 | 71 | 70 | 73 | 65 | 70 | 76 |
| C | 84 | 71 | 100 | 77 | 77 | 75 | 80 | 88 |
| D | 78 | 70 | 77 | 100 | 76 | 82 | 75 | 83 |
| E | 80 | 73 | 77 | 76 | 100 | 71 | 76 | 81 |
| F | 80 | 65 | 75 | 82 | 71 | 100 | 71 | 80 |
| G | 77 | 70 | 80 | 75 | 76 | 71 | 100 | 82 |
| AVG | 82 | 75 | 81 | 80 | 79 | 78 | 78 | 83 |

Table 11: Spearman correlation (%) on the *known for* relation between each pair of annotators (A,...,G), and between each annotator and the average score provided by the other six **after the 3rd and final quality enhancement annotation round.**

| | A | B | C | D | E | F | G | Others |
|---|---|---|---|---|---|---|---|---|
| A | 100 | 58 | 82 | 74 | 79 | 78 | 73 | 82 |
| B | 58 | 100 | 61 | 64 | 64 | 59 | 61 | 68 |
| C | 82 | 61 | 100 | 74 | 75 | 74 | 70 | 79 |
| D | 74 | 64 | 74 | 100 | 77 | 77 | 73 | 83 |
| E | 79 | 64 | 75 | 77 | 100 | 75 | 78 | 84 |
| F | 78 | 59 | 74 | 77 | 75 | 100 | 74 | 79 |
| G | 73 | 61 | 70 | 73 | 78 | 74 | 100 | 78 |
| AVG | 78 | 67 | 76 | 77 | 78 | 77 | 75 | 79 |

Table 12: Spearman correlation (%) on the *similar to* relation between each pair of annotators (A,...,G), and between each annotator and the average score provided by the other six **after the 3rd and final quality enhancement annotation round.**

form:

> Consider the following reference list of <desc>:
> $[A_1, B_1]$
> :
> $[A_5, B_5]$
> Now sort the entity pairs from the following list based on the extent to which they also represent <desc> in descending order. Do not include the pairs from the

| Model | Name on HuggingFace |
|---|---|
| RelBERT$_{\text{BASE}}$ | `relbert/relbert-roberta-base` |
| RelBERT$_{\text{LARGE}}$ | `relbert/relbert-roberta-large` |
| OPT$_{\text{125M}}$ | `facebook/opt-125m` |
| OPT$_{\text{350M}}$ | `facebook/opt-350m` |
| OPT$_{\text{1.3B}}$ | `facebook/opt-1.3b` |
| OPT$_{\text{2.7B}}$ | `facebook/opt-2.7b` |
| OPT$_{\text{6.7B}}$ | `facebook/opt-6.7b` |
| OPT$_{\text{13B}}$ | `facebook/opt-13b` |
| OPT$_{\text{30B}}$ | `facebook/opt-30b` |
| OPT$_{\text{66B}}$ | `facebook/opt-66b` |
| OPT-IML$_{\text{1.3B}}$ | `facebook/opt-iml-1.3b` |
| OPT-IML$_{\text{30B}}$ | `facebook/opt-iml-30b` |
| OPT-IML$_{\text{MAX-1.3B}}$ | `facebook/opt-iml-max-1.3b` |
| OPT-IML$_{\text{MAX-30B}}$ | `facebook/opt-iml-max-30b` |
| T5$_{\text{SMALL}}$ | `t5-small` |
| T5$_{\text{BASE}}$ | `t5-base` |
| T5$_{\text{LARGE}}$ | `t5-large` |
| T5$_{\text{XL}}$ | `t5-3b` |
| T5$_{\text{XXL}}$ | `t5-11b` |
| Flan-T5$_{\text{SMALL}}$ | `google/flan-t5-small` |
| Flan-T5$_{\text{BASE}}$ | `google/flan-t5-base` |
| Flan-T5$_{\text{LARGE}}$ | `google/flan-t5-large` |
| Flan-T5$_{\text{XL}}$ | `google/flan-t5-xl` |
| Flan-T5$_{\text{XXL}}$ | `google/flan-t5-xxl` |
| Flan-UL2$_{\text{20B}}$ | `google/flan-ul2` |

Table 13: The language models used in the paper and their corresponding alias on HuggingFace model hub.

| | ChatGPT | GPT-4 |
|---|---|---|
| Rival | -0.9 (0.0%) | 62.5 (100.0%) |
| Ally | 42.5 (56.8%) | 55.8 (100.0%) |
| Inf | 17.5 (91.1%) | 35.9 (94.4%) |
| Know | 15.5 (86.7%) | 60.8 (100.0%) |
| Sim | 14.7 (80.9%) | 69.3 (98.9%) |
| AVG | 17.9 (63.1%) | 56.9 (98.7%) |

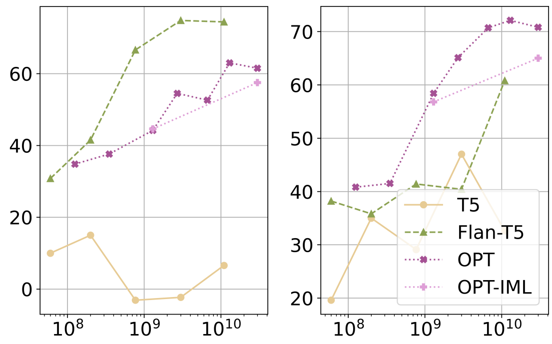
Table 14: Spearman's rank correlation (%) on the test set for conversational LMs with the percentage of word pairs included in the output.

reference list. The output should contain
all the entity pairs from the following list
and no duplicates:
$[C_1, D_1]$
$\vdots$
$[C_n, D_n]$

These conversational models often omit entity pairs from the output, especially those with lower similarity to the reference pairs. To deal with this, we simply concatenate those removed pairs to the bottom of the sorted output list.

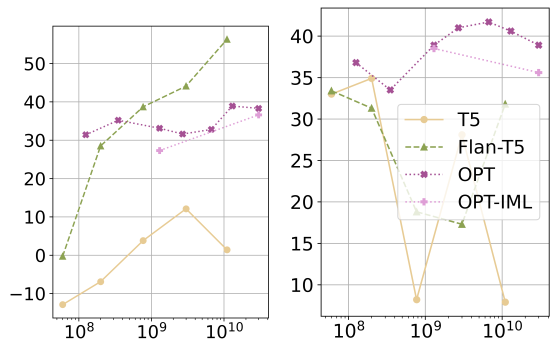
## D Conversational LMs

Table 14 shows the results and percentage of retrieved pairs of the conversational LMs.



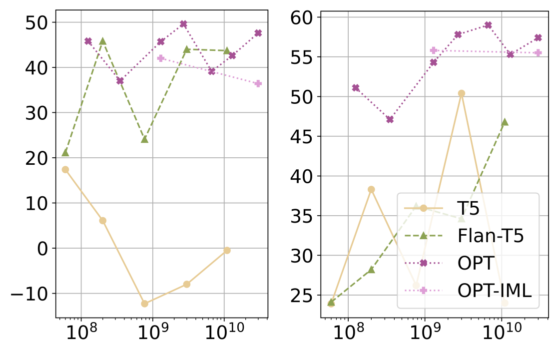

(a) QA template (b) LC template

Figure 3: Spearman's rank correlation for the *competitor/rival of* relation type along with the model size.




(a) QA template (b) LC template

Figure 4: Spearman's rank correlation for the *friend/ally of* relation type along with the model size.

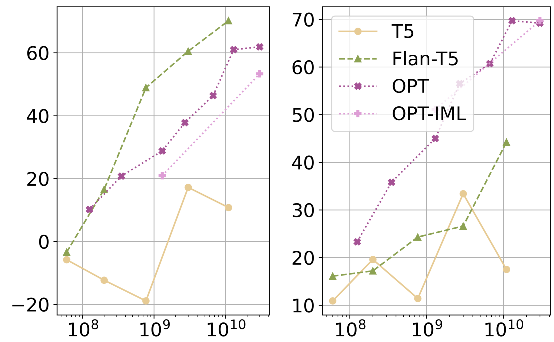



(a) QA template (b) LC template

Figure 5: Spearman's rank correlation for the *influenced by* relation type along with the model size.

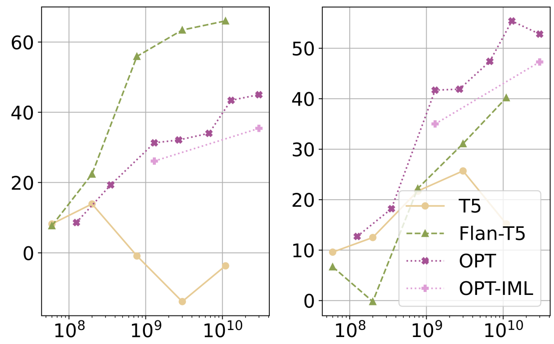
## E Additional Results

Figure 3, Figure 4, Figure 5, Figure 6, and Figure 7 show the performance improvement along with the model size for individual relation types. Figure 8, Figure 9, Figure 10, Figure 11, and Figure 12 show the zero-shot and few-shot evaluation result for individual relation types.



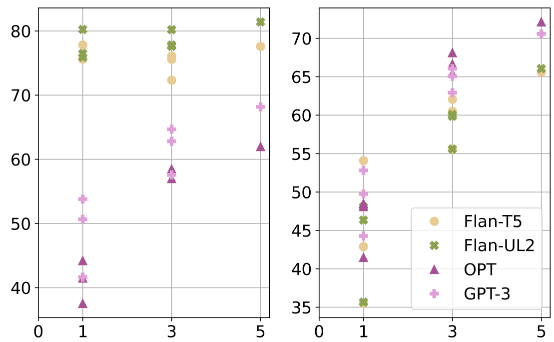

(a) QA template (b) LC template

Figure 6: Spearman's rank correlation for the *known for* relation type along with the model size.



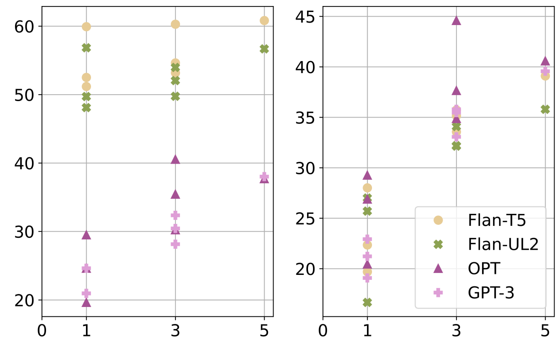

(a) QA template (b) LC template

Figure 7: Spearman's rank correlation for the *similar to* relation type along with the model size.



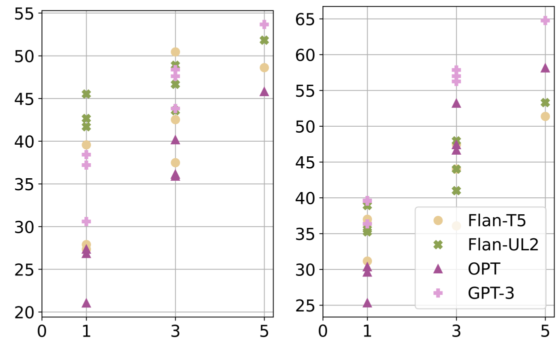

(a) QA template (b) LC template

Figure 8: Spearman's rank correlation for *competitor/rival of* relation with different number of the prototypical examples.



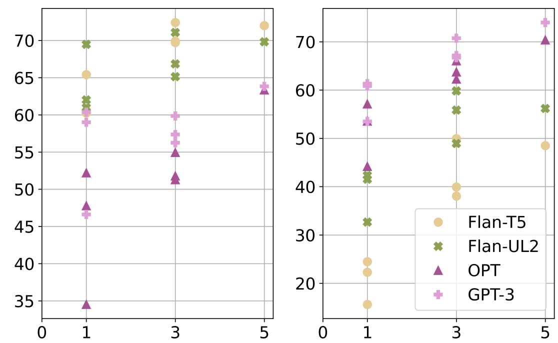

(a) QA template (b) LC template

Figure 9: Spearman's rank correlation for *friend/ally of* relation with different number of the prototypical examples.




(a) QA template (b) LC template

Figure 10: Spearman's rank correlation for *influenced by* relation with different number of the prototypical examples.




(a) QA template (b) LC template

Figure 11: Spearman's rank correlation for *known for* relation with different number of the prototypical examples.

# F Prompt for Zero-shot/Few-shot Learning

The QA template for zero-shot/few-shot learning are:

> Answer the question by yes or no. Are $[C, D]$ <desc>?
> Yes

while the zero-shot LC template has the following form:

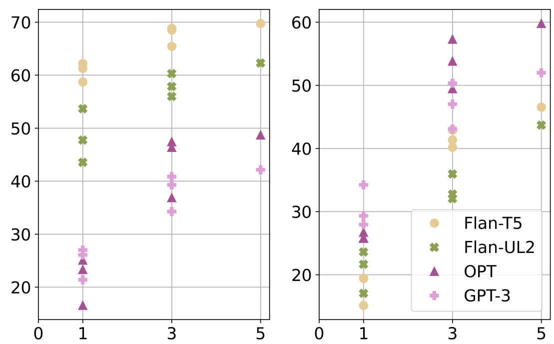
> Complete the following list with examples of <desc>?
> $[C, D]$

# G Full Results

Table 15 shows the result for all the LMs we considered in the paper.

| | | | Model | Inst-FT | Model Size | Rival | Ally | Inf | Know | Sim | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Human Upperbound* | | | | | | 75.9 | 78.0 | 70.5 | 82.0 | 80.2 | 77.3 |
| Embedding | | | fastText$_{\text{word}}$ | | - | 25.0 | 10.0 | 7.0 | 24.0 | 20.0 | 17.0 |
| | | | fastText$_{\text{pair}}$ | | - | 28.0 | 12.0 | 3.0 | 20.0 | 21.0 | 17.0 |
| | | | RelBERT$_{\text{BASE}}$ | | 110M | 58.0 | 15.0 | 30.0 | 24.0 | 28.0 | 31.0 |
| | | | RelBERT$_{\text{LARGE}}$ | | 335M | 64.0 | 20.0 | 20.0 | 44.0 | 53.0 | 40.0 |
| LM | *LC template* | T5 | T5$_{\text{SMALL}}$ | | 60M | 20.0 | 33.0 | 24.0 | 11.0 | 10.0 | 19.0 |
| | | | T5$_{\text{BASE}}$ | | 220M | 35.0 | 35.0 | 38.0 | 20.0 | 13.0 | 28.0 |
| | | | T5$_{\text{LARGE}}$ | | 770M | 29.0 | 8.0 | 26.0 | 11.0 | 22.0 | 19.0 |
| | | | T5$_{\text{XL}}$ | | 3B | 47.0 | 28.0 | 50.0 | 33.0 | 26.0 | 37.0 |
| | | | T5$_{\text{XXL}}$ | | 11B | 33.0 | 8.0 | 24.0 | 18.0 | 15.0 | 19.0 |
| | | | Flan-T5$_{\text{SMALL}}$ | ✓ | 60M | 38.0 | 33.0 | 24.0 | 16.0 | 7.0 | 24.0 |
| | | | Flan-T5$_{\text{BASE}}$ | ✓ | 220M | 36.0 | 31.0 | 28.0 | 17.0 | -0.0 | 22.0 |
| | | | Flan-T5$_{\text{LARGE}}$ | ✓ | 770M | 41.0 | 19.0 | 36.0 | 24.0 | 22.0 | 29.0 |
| | | | Flan-T5$_{\text{XL}}$ | ✓ | 3B | 40.0 | 17.0 | 35.0 | 27.0 | 31.0 | 30.0 |
| | | | Flan-T5$_{\text{XXL}}$ | ✓ | 11B | 61.0 | 32.0 | 47.0 | 44.0 | 40.0 | 45.0 |
| | | | Flan-UL2 | ✓ | 20B | 60.0 | 28.0 | 49.0 | 53.0 | 37.0 | 45.0 |
| | | OPT | OPT$_{\text{125M}}$ | | 125M | 41.0 | 37.0 | 51.0 | 23.0 | 13.0 | 33.0 |
| | | | OPT$_{\text{350M}}$ | | 300M | 41.0 | 33.0 | 47.0 | 36.0 | 18.0 | 35.0 |
| | | | OPT$_{\text{1.3B}}$ | | 1.3B | 58.0 | 39.0 | 54.0 | 45.0 | 42.0 | 48.0 |
| | | | OPT$_{\text{2.7B}}$ | | 2.7B | 65.0 | 41.0 | 58.0 | 56.0 | 42.0 | 52.0 |
| | | | OPT$_{\text{6.7B}}$ | | 6.7B | 71.0 | 42.0 | 59.0 | 61.0 | 47.0 | 56.0 |
| | | | OPT$_{\text{13B}}$ | | 13B | 72.0 | 41.0 | 55.0 | 70.0 | 55.0 | 59.0 |
| | | | OPT$_{\text{30B}}$ | | 30B | 71.0 | 39.0 | 57.0 | 69.0 | 53.0 | 58.0 |
| | | | OPT-IML$_{\text{1.3B}}$ | ✓ | 1.3B | 57.0 | 39.0 | 56.0 | 51.0 | 35.0 | 47.0 |
| | | | OPT-IML$_{\text{30B}}$ | ✓ | 30B | 65.0 | 36.0 | 55.0 | 70.0 | 47.0 | 55.0 |
| | | | OPT-IML$_{\text{MAX-1.3B}}$ | ✓ | 1.3B | 55.0 | 37.0 | 57.0 | 49.0 | 33.0 | 46.0 |
| | | | OPT-IML$_{\text{MAX-30B}}$ | ✓ | 30B | 62.0 | 36.0 | 57.0 | 67.0 | 46.0 | 53.0 |
| | | GPT | GPT-3$_{\text{davinci}}$* | | - | 72.0 | 39.0 | **64.0** | **73.0** | 47.0 | 59.0 |
| | *QA template* | T5 | T5$_{\text{SMALL}}$ | | 60M | 10.0 | -13.0 | 17.0 | -6.0 | 8.0 | 3.0 |
| | | | T5$_{\text{BASE}}$ | | 220M | 15.0 | -7.0 | 6.0 | -12.0 | 14.0 | 3.0 |
| | | | T5$_{\text{LARGE}}$ | | 770M | -3.0 | 4.0 | -12.0 | -19.0 | -1.0 | -6.0 |
| | | | T5$_{\text{XL}}$ | | 3B | -2.0 | 12.0 | -8.0 | 17.0 | -14.0 | 1.0 |
| | | | T5$_{\text{XXL}}$ | | 11B | 7.0 | 1.0 | -1.0 | 11.0 | -4.0 | 3.0 |
| | | | Flan-T5$_{\text{SMALL}}$ | ✓ | 60M | 31.0 | -0.0 | 21.0 | -3.0 | 8.0 | 11.0 |
| | | | Flan-T5$_{\text{BASE}}$ | ✓ | 220M | 41.0 | 28.0 | 46.0 | 17.0 | 22.0 | 31.0 |
| | | | Flan-T5$_{\text{LARGE}}$ | ✓ | 770M | 67.0 | 39.0 | 24.0 | 49.0 | 56.0 | 47.0 |
| | | | Flan-T5$_{\text{XL}}$ | ✓ | 3B | 75.0 | 44.0 | 44.0 | 61.0 | 63.0 | 57.0 |
| | | | Flan-T5$_{\text{XXL}}$ | ✓ | 11B | 74.0 | **56.0** | 44.0 | 70.0 | 66.0 | **62.0** |
| | | | Flan-UL2 | ✓ | 20B | **79.0** | 51.0 | 47.0 | 67.0 | 57.0 | 60.0 |
| | | OPT | OPT$_{\text{125M}}$ | | 125M | 35.0 | 31.0 | 46.0 | 10.0 | 9.0 | 26.0 |
| | | | OPT$_{\text{350M}}$ | | 350M | 38.0 | 35.0 | 37.0 | 21.0 | 19.0 | 30.0 |
| | | | OPT$_{\text{1.3B}}$ | | 1.3B | 44.0 | 33.0 | 46.0 | 29.0 | 31.0 | 37.0 |
| | | | OPT$_{\text{2.7B}}$ | | 2.7B | 54.0 | 32.0 | 50.0 | 38.0 | 32.0 | 41.0 |
| | | | OPT$_{\text{6.7B}}$ | | 6.7B | 53.0 | 33.0 | 39.0 | 46.0 | 34.0 | 41.0 |
| | | | OPT$_{\text{13B}}$ | | 13B | 63.0 | 39.0 | 43.0 | 61.0 | 43.0 | 50.0 |
| | | | OPT$_{\text{30B}}$ | | 30B | 61.0 | 38.0 | 48.0 | 62.0 | 45.0 | 51.0 |
| | | | OPT-IML$_{\text{1.3B}}$ | ✓ | 1.3B | 45.0 | 27.0 | 42.0 | 21.0 | 26.0 | 32.0 |
| | | | OPT-IML$_{\text{30B}}$ | ✓ | 30B | 57.0 | 37.0 | 36.0 | 53.0 | 35.0 | 44.0 |
| | | | OPT-IML$_{\text{MAX-1.3B}}$ | ✓ | 1.3B | 42.0 | 25.0 | 38.0 | 16.0 | 29.0 | 30.0 |
| | | | OPT-IML$_{\text{MAX-30B}}$ | ✓ | 30B | 58.0 | 36.0 | 39.0 | 43.0 | 42.0 | 43.0 |
| | | GPT | GPT-3$_{\text{davinci}}$* | | - | 67.0 | 35.0 | 50.0 | 61.0 | 35.0 | 50.0 |
| Conv. LM | | | ChatGPT* | | - | -0.9 | 32.5 | 17.5 | 15.5 | 14.7 | 17.9 |
| | | | GPT-4* | | - | 62.5 | 55.8 | 35.9 | 60.8 | **69.3** | 56.9 |

Table 15: Spearman's rank correlation (%) on the test set. The LMs are grouped by the template (QA or LC), the model family, and instruction-fine-tuned or not. The best correlation in each relation type is highlighted by bold characters. Model size is measured as the number of parameters. Models marked with * are not openly available.




(a) QA template (b) LC template

Figure 12: Spearman's rank correlation for *similar to* relation with different number of the prototypical examples.