Testing the Ability of Language Models to Interpret Figurative Language

Anonymous ACL submission

Abstract

Figurative and metaphorical language are commonplace in discourse, and figurative expressions play an important role in communication and cognition. However, figurative language has been a relatively under-studied area in NLP, and it remains an open question to what extent modern language models can interpret nonliteral phrases. To address this question, we introduce Fig-QA, a Winograd-style nonliteral language understanding task consisting of correctly interpreting paired figurative phrases with divergent meanings. We evaluate the performance of several state-of-the-art language models on this task, and find that although language models achieve performance significantly over chance, they still fall short of human performance, particularly in zero- or few-shot settings. This suggests that further work is needed to improve the nonliteral reasoning capabilities of language models.[1]

1 Introduction

All our words are but crumbs that fall down from the feast of the mind (Gibran, 1926). When humans read such a metaphorical phrase, how do they interpret it? Conceptual metaphors structure our everyday language and are used to map everyday physical experiences and emotions onto abstract concepts (Lakoff and Johnson, 1981). They allow us to communicate complex ideas, to emphasize emotions, and to make humorous statements (Fussell and Moss, 2008). However, despite relating words in a way that differs from their accepted definition, these phrases are readily interpreted by human listeners, and are common in discourse (Shutova, 2011), occurring on average every three sentences (Mio and Katz, 1996; Fussell and Moss, 2008)

The ability to interpret figurative language has been viewed as a bottleneck in natural language understanding, but it has not been studied as widely as literal language (Shutova, 2011; Tong et al., 2021). Figurative language often relies on shared commonsense or cultural knowledge, and in some cases may be difficult to solve using language statistics. This presents a challenge to language models (LMs), as strong LMs trained only on text may not be able to make sense of the physical world, nor the social or cultural knowledge that language is grounded in (Bender and Koller, 2020; Bisk et al., 2020).

Most previous work on figurative language focuses on metaphor detection, where a model is trained to *identify* the existence of metaphors in text (Tsvetkov et al., 2014; Stowe and Palmer, 2018; Leong et al., 2020), with datasets consisting mostly of conventionalized metaphors and idioms in wide use. However, identifying these common metaphors that already appear often in language may be an easy task for LMs, and not fully test their ability to interpret figurative language. The little work that exists on metaphor interpretation frames it as a task linking metaphorical phrases to literal rewordings, either through paraphrase detection (Bizzoni and Lappin, 2018) or paraphrase generation (Shutova, 2010; Su et al., 2017; Mao et al., 2018) (details in § 7) While interesting, this work does not take into account the fact that metaphors are rich with different implications that may vary depending on the context.

In this work, we ask whether or not LMs can correctly *make inferences regarding creative, relatively novel metaphors* generated by humans. This task is harder for two reasons: (1) *inference* is harder than *identification* or *paraphrasing*, as it requires understanding the underlying semantics, and (2) the metaphors in our dataset are novel creations, and many may not appear even once in the LMs' training data. We propose a minimal task inspired by the Winograd schema (Levesque et al., 2012), where LMs are tasked with choosing the entailed phrase from two opposite metaphorical phrases. An example of a paired sentence is "Her

[1] Code is available at https://anonymous.4open.science/r/metaphor-qa-5083/ under the MIT license.

commitment is as sturdy as (plywood/oak)". The correct answer would be either "She was (committed/uncommitted)". This can also be seen as an entailment task, where input x is the premise, and the output y is the hypothesis.[2]

We crowdsource a benchmark **Fig-QA**, consisting of 10,256 such metaphors and implications (§ 2), which can be used to evaluate the nonliteral reasoning abilities of LMs or for more broad studies of figurative language in general (we provide preliminary analyses in § 3). Through extensive experiments over strong pre-trained LMs (§ 4), we find that although they can be fine-tuned to do reasonably well, their few-shot performance falls significantly short of human performance (§ 5). An in-depth analysis (§ 6) uncovers several insights: (1) LMs do not make use of the metaphorical context well, instead relying on the probability of interpretations alone, (2) the task of associating a metaphor with an interpretation is more difficult than the reverse, (3) even strong models such as GPT-3 make inexplicable errors that are not well-aligned with human ones, indicating that further work is needed to properly model nonliteral language.

2 Dataset Creation and Validation

2.1 Crowdsourcing Task

We crowdsourced data from workers on Amazon Mechanical Turk (details in Appendix A). Workers were asked to generate paired metaphors with different meanings, as well as literal implications of the two metaphors in context. We instructed workers to try to generate rare or creative metaphors, namely “metaphors that would not appear often in text on the internet, books, social media, or news sites, but that can still be easily understood by people.” Workers were given examples of valid pairs that fit the format, and examples of invalid ones to discourage errors. Some examples of generated pairs are displayed in Table 1.

In order to help workers, we employ the *randomness as genesis* and *narrow limits of change* principles of Cognitive Load Theory (Sweller, 2006). To add soft constraints, we generate 3 different random words to be shown to each batch of workers. However, workers were not required to use these words, as we wanted to encourage maximal diversity. In order to ensure that the random words were metaphorically rich, we selected them based on metaphorical frames in Lakoff and Johnson (1981).

2.2 Data Validation

The dataset was manually validated by three authors of this paper. Each author covered roughly one-third, evenly split between training, validation, and test. Examples were excluded if they (a) did not make sense given the figurative expression, (b) had grammar or spelling errors that rendered them unintelligible, or (c) did not follow the format of the task. Examples of excluded samples are included in Appendix B. We collected 13,324 sentences and interpretations from the crowdsourcing task, and 10,256 sentences remained after filtering.

2.3 Final Dataset

The release version of our dataset contains the named data splits in Table 2. The medium train, dev, and test splits were generated from partitioning the first stage of data collected. The large train split additionally contains all the new examples from the second collection stage, and the small train split is a small random sample.

3 Figurative Language Typologies

In this sample, we perform an analysis of the collected data to demonstrate its trends and categorize examples for further error analysis. Specifically, we examine (a) subjects, objects, and relations, and (b) types of common-sense knowledge needed to interpret the metaphor.

3.1 Figurative Language Structure

We note that most metaphors and similes can be characterized by three components, (S, R, O), where S is a subject, R is a relation, and O is an object. For instance, "Her commitment is as sturdy as plywood" can be written (Her commitment, sturdy, plywood). Interpretation involves inferring an attribute of the subject by extracting a relational attribute from the object (Fauconnier and Turner, 2003). In a simile, R is explicit, whereas it is usually implicit in a metaphor. The most common subjects, relations, and objects in the medium train dataset are shown in Figure 1. These were obtained by first segmenting the phrases with syntactic patterns constructed from observation, followed by lemmatization and removal of punctuation and determiners "the", "an", "a" and "that". There are 441

[2] The opposing meanings help to avoid ambiguity in the correct answer, make the task intuitive for human annotators, and help prevent annotation artifacts that have plagued other NLI datasets (Gururangan et al., 2018).

Paired sentences	Possible answers
The pilot flew like a ballet dancer The pilot flew like a modern dancer	The pilot flew in a (**restrained way** | creative way) The pilot flew in a (restrained way | **creative way**)
The meteor was as bright as New York City The meteor was as bright as coal	The meteor was (**very bright** | not bright at all) The meteor was (very bright | **not bright at all**)
The atom is like a solar system The atom is like a cloud	Electrons (**orbit the nucleus** | are in probability densities) Electrons (orbit the nucleus | **are in probability densities**)
He hustles like rent was due three days ago He hustles like he's a billionaire's son.	He (**hustles hardcore**. | doesn't hustle at all.) He (hustles hardcore | **doesn't hustle at all**)
Life is as easy as kindergarten for a high school senior Life is as easy as kindergarten for a newborn	Life is (**basic** | beyond comprehension) Life is (basic | **beyond comprehension**)

Table 1: Example sentences from the dataset

Train			Dev	Test
S	M	L		
200	1,458	8,016	1,094	1,146

Table 2: Examples in each data split

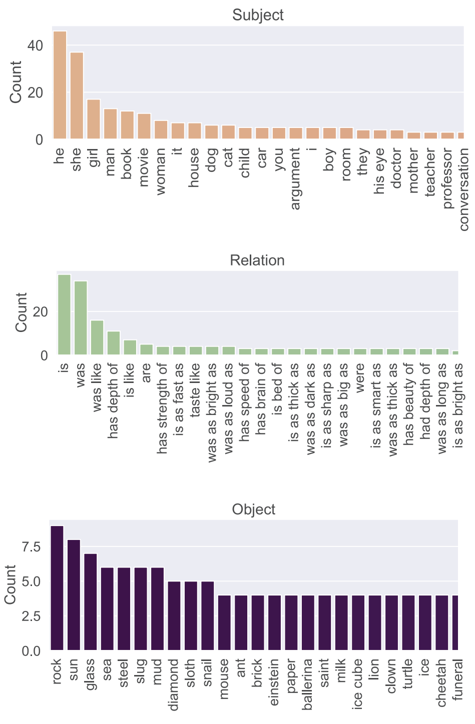

Figure 1: Visualization of 25 most frequent subjects, relations, and objects in the medium train set.

unique subjects, 646 unique relations, and 1,198 unique objects in the medium training set.

3.2 Common-sense Knowledge Types

Next, we examined the test set to determine the types of commonsense knowledge needed to interpret metaphors. Through thematic analysis, we devised 4 categories based on commonsense knowledge, which are not mutually exclusive: common-sense object knowledge, visual metaphors, common-sense social understanding, and cultural knowledge. The same 3 paper authors annotated the test set for these categories, with annotators responsible for separate categories.

Common-sense object knowledge consisted of metaphors that made reference to properties of common objects and animals, such as volume, height or mass of objects, or properties of materials. 68.35% of the test-set was found to require common-sense object knowledge.

Visual metaphors were a subset of common-sense object metaphors, primarily relying on the visual modality, including attributes such as brightness or colour. Some visual metaphors also sketched a vivid visual scene. These examples comprised 14.73% of the test set.

Common-sense social understanding examples required knowing about how humans would react in different circumstances, or required knowing about human emotions. These examples comprised 27.55% of the test set.

Cultural metaphors required knowing cultural traditions, works of art/artefacts, or religion. Due to crowdworkers being recruited from the US, these are centered around US culture. These examples comprised 16.56% of the test set.

4 Baseline Models and Evaluation

4.1 Auto-regressive Language Models

Auto-regressive LMs generate a probability distribution of the next token given all preceding tokens. As such, we can directly compute the probability of a sentence by multiplying the conditional probability of each token at every time step.

$$\tilde{P}(w_1...w_N) = p(w_1) \prod_{i=2}^{N} p(w_i|w_1...w_{i-1})$$

Type of knowledge required	Paired sentences
Common-sense (objects)	The new mattress is just as comfortable as sleeping on a (cloud/rocks outside)
Visual	The professor’s argument had the clarity of a (crystal glass/marine fog)
Common-sense (social)	She is as embarrassed as a kid that (forgot homework/got an A)
Cultural	The construction was as disastrous as the (1981 musical Cats/The 2019 film based on the musical Cats)

Table 3: Metaphor types based on types of knowledge required (not mutually exclusive)

The ability to directly extract probabilities enables the *zero-shot* reasoning of these LMs. For a pair of metaphorical expressions x_1 and x_2 with two corresponding interpretations y_1 and y_2, we feed in the concatenation of the metaphor and the interpretation to the pretrained model without finetuning. We define “forward” and “backward” probabilities assigned to interpretations and figurative language expressions, respectively. For the **forward probability**, for figurative phrase x_i and correct answer y_i, we take

$$P(y_i|x_i) = \frac{P(x_i, y_i)}{P(x_i, y_i) + P(x_i, y_j)}$$

since there are only two answer options. From this, we can calculate accuracy when we taking the indicator of $P(y_i|x_i) > 0.5$. Similarly for the **backward probability** (predicting phrase based on answer), we take

$$P(x_i|y_i) = \frac{P(x_i, y_i)}{P(x_i, y_i) + P(x_j, y_i)}$$

with analogous backward accuracy.[3]

We examine three state-of-the-art large transformer-based LMs of this category: **GPT-2** (with 117M parameters, trained on 40GB of text), **GPT-neo** (with 1.3B parameters, trained on 800GB of text) and **GPT-3** (4 variants between 350M and 175B parameters, trained on 45TB on text) (Radford et al., 2019; Black et al., 2021; Brown et al., 2020). We also examine the performance of these models after finetuning on the training data. GPT-2 and GPT-neo were trained with a batch size of 8, with early stopping on the medium dataset with a patience of 1 epoch, and a minimal hyperparameter search was done with learning rates 1e-5 to 5e-5. GPT-3 was trained with the default parameters of the GPT-3 finetuning API.

4.2 Masked Language Models

We also evaluate the performance of masked LMs on this task. Unlike auto-regressive LMs, masked LMs cannot directly output the probability of a sentence, so it is not possible to directly test the zero-shot performance of these models. Instead, we test the transfer performance by first finetuning them in two ways: on WinoGrande, which is also a binary choice task based on common-sense reasoning, and on several NLI datasets, including SNLI, MNLI, FEVER-NLI and ANLI (Nie et al., 2020; Sakaguchi et al., 2020). The input to the model trained on WINOGRANDE is formatted as `[CLS][metaphor][SEP][answer1][SEP][answer2]`, and we use an extra linear layer on the `[CLS]` token embedding to perform the classification. In addition to the transfer performance, we also use contrastive finetuning by feeding in each metaphor along with both answer choices, and training the model with our dataset to classify which answer is correct. For the NLI model, we examine accuracy using all three labels the model was originally trained with (entailment, neutral, and contradiction), as well as using a forced binary choice paradigm in which the logits for the contradiction label are subtracted from the logits for the entailment label, and the higher "entailment score" is the ending the model predicts. We treat these two conditions as the analog of “zero-shot" for these models.

We examine two masked LMs that are commonly used as baselines on many NLP tasks: **BERT** (Devlin et al., 2019), a transformer-based LM jointly trained on the masked LM and next sentence prediction objectives, and **RoBERTa** (Liu et al., 2019), an improved variant of BERT which consistently outperforms BERT across most tasks. We use the large variant of both models (350M parameters). BERT and RoBERTa were finetuned on the medium dataset for 8 epochs with batch size 8, following the setting in (Sakaguchi et al., 2020). A hyperparameter search was done with learning rates 5e-6 to 2e-5. Both BERT and RoBERTa were used for the Winogrande experiments, while only

[3] In actuality, we use the length-normalized probability that a model assigns to a sentence as a heuristic for the total probability, to minimize the effect that the length of a sentence has on the decision (though this is not the probability of the sequence in a strict sense): $P(w_1...w_N) = \exp(-\frac{1}{N}\log\tilde{P}(w_1...w_N))$. Initial experimentation showed marginal differences in accuracy when using these two methods, so we used normalized probabilities by default.

Model	Zero-shot	Tuned (L)	Tuned (XL)
GPT-2	53.93	54.80	62.65
GPT-neo 1.3B	56.89	69.98	72.00
GPT-3 Ada	59.08	69.17	73.56
GPT-3 Babbage	62.91	73.97	77.31
GPT-3 Curie	65.35	**79.04**	**81.94**
GPT-3 Davinci	**68.41**	-	-
BERT	58.14	83.16	85.69
RoBERTa	**66.18**[4]	**89.22**	**90.32**
Human	94.42	-	-
Human (confident)	**95.39**	-	-

Table 4: Zero-shot and finetuned test accuracies (%), finetuned is averaged across 5 seeds.

RoBERTa was used for the NLI experiment.

4.3 Human Performance

To estimate the expected human performance on this task, we ran a benchmark on the test set with 10 human volunteers who were not authors of the paper. The human annotators were not shown any training examples, so this would be equivalent to the zero-shot setting for models. Participants ranged from 20 to 29 years old, and there were 5 male and 5 female participants. 5 each were native- and non-native English speakers respectively. Participants were mainly graduate student volunteers.

We shuffled the test set and split it into 10 partitions of ≈115 examples for each annotator. Due to differences in vocabulary or cultural background, we instructed participants to mark examples where they weren't confident, such as those that contained words or cultural references they didn't understand.

5 Results

5.1 Inference Results

The first question is **whether strong LMs can interpret metaphors at all when presented with two opposing meanings, in zero-shot or supervised settings**. These results are presented in Table 4. The results for masked language models are higher than those for autoregressive language models, and fine-tuning significantly improves performance for all models.

Zero-shot Performance For the zero-shot setting, we examine the test accuracy based on zero-shot forward probabilities for the GPT models, and the pseudo "zero-shot" transfer performance for BERT and RoBERTa using models pretrained on the WinoGrande task (Sakaguchi et al., 2020). As shown, the GPT-3 models outperform the GPT-2 and GPT-neo models. Among the GPT-3 models, there is a clear correlation between model size and performance, with the largest model (GPT-3 Davinci) achieving the highest zero-shot test accuracy. BERT and RoBERTa achieve accuracies within the range of GPT-3 models. While our models mostly perform much better than chance in the zero-shot setting, there is still a large gap of 26 percentage points between our best model and human level performance.

Fine-tuned Performance For the fine-tuned setting, all listed models are fine-tuned on the small dataset split. GPT models were trained with language modeling loss, whereas BERT and RoBERTa are trained with contrastive loss. We did not evaluate fine-tuning of GPT-3 Davinci due to budget. Overall, fine-tuning improved accuracy significantly for all models, with GPT-3 models uniformly improving by about 13 percentage points, and BERT/RoBERTa improving by about 25 points. Our best model after fine-tuning is RoBERTa, which reaches within 5% of our human performance.

Prompting We also experiment with prompting methods. Firstly, we use a simple *suffix* prompting method, where we simply append the phrase "that is to say" between the metaphor and the interpretation, which we hypothesized may "explain" to the LM that the previous statement is figurative. We also evaluate the effectiveness of the *examples* method, by appending k random correct metaphor/interpretation pairs before the actual pair we are testing. The results of these experiments can be seen in Figure 2. We found that the suffix method provided a small (1-2%) improvement over the baseline, while the example method was generally ineffective.

Backward accuracies Note the accuracies reported in this section are for the forward direction, and the backward direction is reported in Appendix C. Backward accuracies are lower, with GPT-3 Curie for example having a 7% reduction in accuracy in the zero-shot case. This suggests that selecting a metaphorical expression to match a literal phrase is more challenging than the reverse for LMs.

[4] This is the accuracy score when transferring from Winogrande. Pretrained NLI results were 50.47 when using original labels (entailment/contradiction/neutral), and 66.32 when forcing a binary decision.

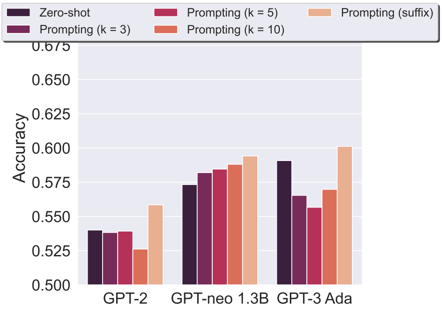

Figure 2: Comparison of prompting methods with autoregressive models

5.2 Generation Results

Next, we examine **if models can generate sensible interpretations for metaphors**. Given the difficulty of evaluating text generation, compounded by the difficulty of figurative language, we opted for manual evaluation of one tenth of the test dataset using generations of the strongest auto-regressive model: GPT-3 Davinci ($\approx$175B parameters).

The metaphor was given as input to the model, and 4 completions were generated for each metaphor, with a maximum length of 100 tokens. Suffix prompting was also used because of the lack of context, with "That is to say, " appended to each metaphor. Only the first sentence of the output was evaluated. The temperature parameter was determined through grid search through values [0.2, 0.4, 0.6, 0.8, 1] on a small separate set of metaphors. A human annotator inspected the generated completions and found that a temperature of 0.4 produced the most correct results.

Three paper authors labelled completions generated by GPT-3 Davinci as either correct, incorrect, or literal. In some cases, there were valid interpretations that were not the same as the answer given by crowdworkers, which were also marked correct. If the model simply restated the metaphor with no interpretation, the completion was marked as literal. Because some metaphors are ambiguous when presented without context, those examples were not counted. The inter-rater reliability was moderate due to differing standards for correctness (Krippendorff's α = 0.5567). The majority vote was taken between annotators' judgments.

GPT-3 Davinci's accuracy, counting literalized metaphors as incorrect, was 50.8%. Not counting literalized metaphors, accuracy was 63.9%. In 37.7% of cases, GPT-3 generated contradictory completions among the 4 completions. There was at least one correct completion for 78.1% of the metaphors, but only 19.3% of metaphors had all completions correct. Examples of annotated generations can be found in Appendix F.

6 Performance and Error Analysis

With these results in mind, we examine **what kinds of errors models make, and what factors make the task difficult.**. This is covered in § 6. We find that autoregressive models rely on the probability of each answer by itself to predict the answer, and that this holds true for all models, before and after training. We find that models have difficulty in interpreting "sarcastic" metaphors, and sometimes inexplicably interpret very simple metaphors wrong. We also examine error typology according to the commonsense typology of § 3.2 and find that models improve significantly on object, visual and social commonsense when trained, but not on cultural commonsense.

6.1 Reliance on Probability of Answers

We find that models often rely solely on the probability of answers y_1 and y_2 to make their predictions, regardless of the context. This led models to make the same prediction for the paired sentences in many cases. Figure 3 and Table 5 show that this trend improves with fine-tuning, and that GPT-3 is best able to disentangle the probability of y_i and the probability of $P(y_i|x_i)$, but all three models show a heavy tendency to predict based on the relative probability of an answer alone.

We hypothesize that this may be one reason why BERT and RoBERTa achieve the best finetuned performance; they use a contrastive finetuning strategy providing both the correct and incorrect options as input to the model. On the other hand, the GPT models were finetuned with only the correct option, making the comparison unfair. One way to finetune GPT models contrastively is to include both options into a cleverly engineered prompt, but we leave this as a direction for future work.

6.2 Other Factors Influencing Correctness

We also examined the influence of several other factors on correctness. The point-biserial correlation between length of the context phrase and the binary correctness value was -0.1544 with a p-value of 1.50×10^{-7}, indicating that longer phrases are harder to interpret correctly. The point-biserial correlation between answer probability and binary cor-

Model	r	p
	Untrained	
GPT-2	0.8128	6.700×10^{-136}
GPT-neo	0.7891	6.075×10^{-123}
GPT-3	0.7392	4.329×10^{-100}
	Trained	
GPT-2	0.6765	6.700×10^{-78}
GPT-neo	0.6689	1.456×10^{-75}
GPT-3	0.4157	2.598×10^{-25}

Table 5: Spearman r-values and p-values between $P(y_i|x_i)$ and $P(y_i)$

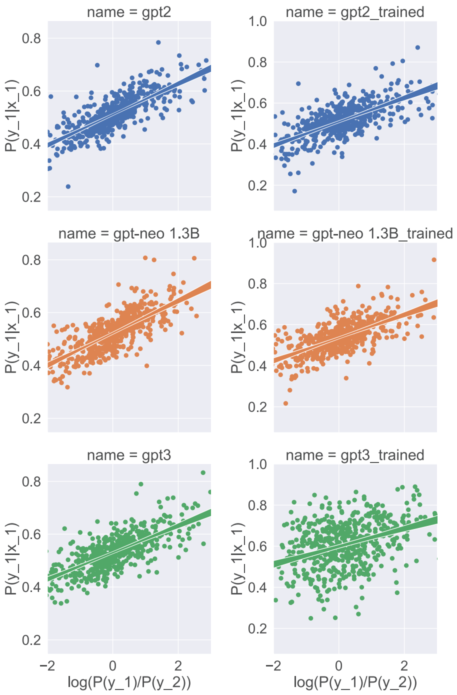

Figure 3: Models over-rely on probability of the answer to do their predictions. y-axis is probability of the first interpretation (answer) given metaphor while x-axis is log odds of the first interpretation.

rectness was 0.1840, with a p-value of 3.50×10^{-10}, indicating that examples where the answer was already more probable were more likely to be answered correctly, in line with our findings that models tended to predict the answer that was already more plausible alone.

Furthermore, we conducted an analysis on subjects, objects, and relations as defined in § 3.1. We examined accuracy by part of speech patterns in each part of the metaphor, as well as by wordnet hypernyms present in each part of the metaphor. This is detailed in Appendix D and Appendix E (Fellbaum, 1998). We used NLTK for POS tagging (Loper and Bird, 2002).

6.3 Qualitative Analysis of Error Trends

Common Sense Knowledge We first examine the error tendencies by the type of common sense knowledge described in § 3.2. Table 6 summarizes accuracies for these types of commonsense questions compared to humans.

Model	Obj	Vis	Soc	Cul
		Untrained		
GPT-2	52.17	52.07	55.38	**58.42**
GPT-neo	56.38	55.62	56.01	**62.10**
GPT-3 Curie	75.00	71.00	72.47	**78.42**
		Trained		
GPT-2	53.57	51.48	**57.91**	57.37
GPT-neo	70.15	**72.78**	68.67	70.00
GPT-3 Curie	**87.50**	84.62	83.86	83.16
BERT	87.37	**92.31**	84.18	77.37
RoBERTa	91.20	**94.08**	89.56	83.68
Human	95.41	**96.45**	93.99	90.00

Table 6: The performance of models across different commonsense categories, in terms of accuracy on examples annotated with that category (%). The strongest category of each model is highlighted.

We find that both humans and trained models tend to find object commonsense and visual commonsense metaphors easier to interpret. We find that as models improve, most of the performance gain comes from the object, visual and social commonsense categories. Interestingly, the untrained models do quite well on cultural examples, but do not improve much on the culture category when trained. This makes sense, as the cultural examples tend to be quite disparate, so training would not help as much with other examples.

Sarcastic Metaphors For both humans and LMs, many of the errors are "sarcastic" metaphors, such as saying "the girl was as bubbly as still water" to mean "the girl was bland", rather than "the girl was vivacious". These sentences can be difficult if the model or human focuses on simple word association (bubbly with vivacious) without reading the entire sentence to understand the sarcasm.

Inexplicable Errors We examined the errors made by GPT-3 Curie (trained) and found that there was little overlap with mistakes made by humans. Of the 64 human errors, 13 were also errors made by GPT-3. GPT-3 made many more "obvious" errors, such as predicting "The ball is a big red sun" to mean "the ball is small" rather than "the ball is big and red" This is in contrast to the sentences in which humans made errors, which often contained rare vocabulary or unfamiliar cultural references.

7 Related work

7.1 Figurative Language Identification

Most existing work focuses on identifying figurative language at the word level. The VU Amsterdam Metaphor Corpus (VUA) is the largest available corpus of metaphorical language, annotated by humans (Steen et al., 2010). Two shared tasks on metaphor identification have been run (Leong et al., 2018, 2020). Both have utilized the VUA corpus, and the latter also introduced the TOEFL corpus, sampled from essays written by non-native English speakers (Leong et al., 2020; Beigman Klebanov et al., 2018). Most participants in the shared tasks used neural models, notably BERT, RoBERTa, and Bi-LSTMs (Leong et al., 2020; Bizzoni and Ghanimifard, 2018; Gao et al., 2018; Pramanick et al., 2018). These models are generally improved when augmented with semantic data, such as concreteness, and multimodal information.

Despite the utility of these tasks and datasets, they have drawbacks. Most of the metaphor use is conventional, so this task does not capture novel metaphors well. The word-level annotation also does not lend itself well to capturing extended conceptual metaphors. Finally, metaphor interpretation may be a more difficult, although less studied, task.

7.2 Figurative Language Interpretation

Recent studies mostly focus on metaphor paraphrases, either through identification (Bizzoni and Lappin, 2018) or generation (Shutova, 2010; Su et al., 2017; Mao et al., 2018). However, there has not been as much work done on interpretation as on detection, and framing metaphor interpretation as a paraphrase task may not capture the emergent meaning of metaphors, such as the intended emotion, or the interaction of subject, relation and object in the metaphor (Tong et al., 2021; Mohammad et al., 2016).

Other work has focused on interpreting figurative language in narratives in context, based on plausible continuations of figurative language such as idioms and similes from stories (Chakrabarty et al., 2021) or dialogues (Jhamtani et al., 2021). This represents a promising direction, and our work focuses on expanding our understanding of LMs' ability to interpret non-conventionalized metaphors.

7.3 Human Language Processing

Humans typically do not have any more difficulty processing metaphorical statements in context compared to literal statements (Fussell and Moss, 2008; Glucksberg, 2003). This may be because certain words serve as a *dual reference*, which is to say they refer simultaneously to a physical referent and an abstract superordinate category (Glucksberg, 2003). For instance, "shark" may refer to literal sharks, as well as anything that is considered vicious, leading to utterances such as "that lawyer is a shark".

Metaphorical language processing has also been studied in second-language learners, in the case of idioms. In most cases, the meaning of an unfamiliar idiom is inferred from the context or from word association (Cooper, 1999; Carston and Wearing, 2011; Wolff and Gentner, 2000).

As LMs excel at word-association based tasks, this is an encouraging finding. However, there is still a gap between LM and human performance even in our task, in which one answer is obviously wrong when the input is correctly understood.

8 Conclusion

We present a Winograd-like benchmark task to test the ability of LMs to reason about figurative language, based on large-scale collection of creative metaphors written by humans. We find a large gap between LM zero-shot and human performance on this dataset, but show that models can be fine-tuned to perform well on this particular task.

We hope that this work will encourage further study of nonliteral reasoning in LMs, especially in few-shot settings. Given that metaphorical reasoning may play a role in problem-solving and linguistic creativity, the development of models, training methods, or datasets that enable metaphorical reasoning may improve models' abilities to reason creatively and draw analogies between situations that may appear to be different on the surface. One avenue we hope to investigate is multimodal metaphors, as this dataset currently includes only text-based metaphors. Nonliteral expressions also remain understudied cross-linguistically, but further work on identifying and interpreting metaphors in other languages may also improve the abilities of multilingual models.

9 Ethical Considerations

9.1 Potential Risks

Figurative language has the potential to be used in a harmful way, especially against minority and historically disadvantaged groups. Such language is often emotionally charged or used to insult others, so we took care to remove any examples that were potentially offensive, especially toward protected groups. We acknowledge that this was based on our own judgment, and generically insulting language (for instance, a metaphor that implies that someone is ugly) was not removed because it was not insulting toward any particular individual.

All examples from this dataset are also in English, as it is the language that all authors speak, and this was a preliminary dataset, being the first of its type. However, figurative language is not just important in English, and we leave investigation of figurative language in other languages as future work.

9.2 Terms of Use of Artefacts Used

Additional datasets we used were the Winogrande dataset, SNLI, MNLI, FEVER-NLI and ANLI. Winogrande is licensed under the Apache 2.0 license, which allows modification and distribution, fitting our use case. SNLI is licensed under a Creative Commons Attribution ShareAlike 4.0 International license, which allows us to share and adapt the work as long as we give attribution. The majority of MNLI is licensed under OANC, which allows free use. The fiction section of this dataset consists mostly of works in the public domain, but several stories are licensed: *Seven Swords* is available under a Creative Commons Share-Alike 3.0 Unported License, while *Living History* and *Password Incorrect* are available under Creative Commons Attribution 3.0 Unported Licenses. These licenses allow sharing and adaptation with attribution. FEVER-NLI is licensed under an MIT license, which also allows modification, distribution, and reuse. ANLI is licensed under Creative Commons Attribution-NonCommercial 4.0 International, which also allows sharing and reuse as long as we give attribution.

Models used were GPT-2, GPT-neo, GPT-3, BERT and RoBERTa. GPT-2 and GPT-neo are licensed under an MIT license, which does not place any restrictions on its use. BERT is licensed under an Apache License 2.0, which allows modification and distribution. RoBERTa is licensed under a GNU General Public License v2.0. This fits our use case as we are only running and studying the model. GPT-3 is licensed by Microsoft, and we used the public API to receive output.

9.3 Computational Infrastructure and Computing Budget

To run our computational experiments, we had access to a compute cluster, but minimal compute is needed to run the experiments in this paper. We generally did not use more than 2 GPUs at a time. The only models that required GPU parallelism were the GPT-neo models. An estimated 20 GPU hours are required.

Our computing budget was roughly 100USD, of which 50USD came from a free coupon for cloud services. We also used roughly 20USD on credits for the GPT-3 API.

References

Beata Beigman Klebanov, Chee Wee (Ben) Leong, and Michael Flor. 2018. A corpus of non-native written English annotated for metaphor. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 86–91, New Orleans, Louisiana. Association for Computational Linguistics.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online. Association for Computational Linguistics.

Yuri Bizzoni and Mehdi Ghanimifard. 2018. Bigrams and BiLSTMs two neural networks for sequential metaphor detection. In *Proceedings of the Workshop on Figurative Language Processing*, pages 91–101, New Orleans, Louisiana. Association for Computational Linguistics.

Yuri Bizzoni and Shalom Lappin. 2018. Predicting human metaphor paraphrase judgments with deep neural networks. In *Proceedings of the Workshop on Figurative Language Processing*, pages 45–55, New Orleans, Louisiana. Association for Computational Linguistics.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow.

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.

Robyn Carston and Catherine Wearing. 2011. Metaphor, hyperbole and simile: A pragmatic approach. *Language and Cognition*, 3(2):283–312.

Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. 2021. It's not rocket science : Interpreting figurative language in narratives. *ArXiv*, abs/2109.00087.

Thomas C. Cooper. 1999. Processing of idioms by l2 learners of english. *TESOL Quarterly*, 33:233–262.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Gilles Fauconnier and Mark Turner. 2003. Conceptual blending, form and meaning.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Susan Fussell and Mallie Moss. 2008. Figurative language in emotional communication.

Ge Gao, Eunsol Choi, Yejin Choi, and Luke Zettlemoyer. 2018. Neural metaphor detection in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 607–613, Brussels, Belgium. Association for Computational Linguistics.

Kahlil Gibran. 1926. *Sand and Foam; a book of aphorisms*. A.A Knopf.

Sam Glucksberg. 2003. The psycholinguistics of metaphor. *Trends in cognitive sciences*, 7:92–96.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Taylor Berg-Kirkpatrick. 2021. Investigating robustness of dialog models to popular figurative language constructs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7476–7485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

G. Lakoff and M. Johnson. 1981. *Metaphors we Live By*. University of Chicago Press.

Chee Wee (Ben) Leong, Beata Beigman Klebanov, Chris Hamill, Egon Stemle, Rutuja Ubale, and Xianyang Chen. 2020. A report on the 2020 VUA and TOEFL metaphor detection shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 18–29, Online. Association for Computational Linguistics.

Chee Wee (Ben) Leong, Beata Beigman Klebanov, and Ekaterina Shutova. 2018. A report on the 2018 VUA metaphor detection shared task. In *Proceedings of the Workshop on Figurative Language Processing*, pages 56–66, New Orleans, Louisiana. Association for Computational Linguistics.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, page 552–561. AAAI Press.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. *CoRR*, cs.CL/0205028.

Rui Mao, Chenghua Lin, and Frank Guerin. 2018. Word embedding and WordNet based metaphor identification and interpretation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1222–1231, Melbourne, Australia. Association for Computational Linguistics.

J. S Mio and A. N Katz. 1996. *Metaphor: Implications and Applications*. Psychology Press.

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 23–33, Berlin, Germany. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Malay Pramanick, Ashim Gupta, and Pabitra Mitra. 2018. An LSTM-CRF based approach to token-level metaphor detection. In *Proceedings of the Workshop on Figurative Language Processing*, pages 67–75, New Orleans, Louisiana. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. In *AAAI*.

Ekaterina Shutova. 2010. Automatic metaphor interpretation as a paraphrasing task. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1029–1037, Los Angeles, California. Association for Computational Linguistics.

Ekaterina Shutova. 2011. Computational approaches to figurative language.

Gerard J Steen, Alleta G Dorst, Berenike Herrmann, Anna A Kaal, Tina Krennmayr, and Trynetje Pasma. 2010. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. John Benjamins.

Kevin Stowe and Martha Palmer. 2018. Leveraging syntactic constructions for metaphor identification. In *Proceedings of the Workshop on Figurative Language Processing*, pages 17–26, New Orleans, Louisiana. Association for Computational Linguistics.

Chang Su, Shuman Huang, and Yijiang Chen. 2017. Automatic detection and interpretation of nominal metaphor based on the theory of meaning. *Neurocomputing*, 219:300–311.

John Sweller. 2006. Discussion of ’emerging topics in cognitive load research: Using learner and information characteristics in the design of powerful learning environments’. *Applied Cognitive Psychology - APPL COGNITIVE PSYCHOL*, 20:353–357.

Xiaoyu Tong, Ekaterina Shutova, and Martha Lewis. 2021. Recent advances in neural metaphor processing: A linguistic, cognitive and social perspective. In *NAACL 2021*.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 248–258, Baltimore, Maryland. Association for Computational Linguistics.

P. Wolff and Dedre Gentner. 2000. Evidence for role-neutral initial processing of metaphors. *Journal of experimental psychology. Learning, memory, and cognition*, 26 2:529–41.

A Crowdsourcing Details

We crowdsource metaphorical expressions and their interpretations through Amazon Mechanical Turk. Workers were recruited from the United States and were limited to those who had a $> 98\%$ approval rating on the platform, and who had also completed more than 1000 Human Intelligence Tasks (HITs). Data collection was split into two stages: in the first stage, 1458 train examples, and all the dev and test examples were collected. In the second stage, the remaining 6558 training examples were collected. We identified some workers who created especially good examples in the first stage, and recruited them back for more examples in the second stage. Workers were paid $0.33 for each pair of sentences and were asked to generate 3 pairs at a time. An author of this paper wrote an initial pilot set of sentences, and timed themselves while writing some sentences. They found that each pair took around 1 minute to write, though this varied (less creative examples took less time, while more creative examples took more time). This extrapolates to an hourly rate of 19.80 USD, which is above the minimum wage in all US states, where workers were located.

Our HIT task was structured as follows: At the top of the page, the workers are shown the following instructions: "Your task is to generate three pairs of sentences with opposite or very different meanings, both of which contain rare/creative metaphors, which means metaphors that would not appear often in text on the internet, books, social media or news sites, but that can still be easily understood by people. For each metaphor, you should also provide a literal (non-metaphorical) sentence with the same meaning." Then, we display one example of a valid sentence pair. There is a button that opens a modal with more detailed instructions and some more valid/invalid examples for reference. Below that, we display three random words, which workers are encouraged to use in their sentences if they get stuck. Finally, we display three sets of 5 text fields for workers to fill in: one for the start phrase, two for each metaphorical phrase, and two for each literal interpretation. As the user types in each start phrase, we prepend a copy of their phrase before the corresponding metaphor fields in the UI using some embedded JavaScript, which we found helped reduce confusion and resulted in less improperly formatted responses.

We launched many batches of these HITs until we had collected the desired quantity of data. Then, we converted the form responses into sentence pairs and validated each pair by hand before adding it to our dataset.

B Invalid Examples

Figurative language examples collected from crowdworkers were excluded if they (a) did not make sense given the meaning and the metaphorical expression, (b) had grammar or spelling errors that rendered them unintelligible, or (c) did not follow the format specified by the task template.

Examples are given below:

1. Do not make sense given the meaning and the metaphorical expression

Paired sentences	Possible answers
He was resourceful like toilet paper He was resourceful like a mess	He was very resourceful. He wasn't resourceful at all
The night was as long as a spool of thread The night was as long as a winding road	The night is long The night dragged on
the concert of the lession is a main and a major the concert of the lession features	we concert everyone we concert our loved one

Table 7: Examples that were rejected due to being nonsensical.

2. Significant grammar or spelling errors

Paired sentences	Possible answers
fallten data are very much trusted fallten data are very valuable	fallten are nice flatten are safe
CAR IS BIRD FEATHEAR CAR IS COTTON	CAR SITE IS ROUGH CAR SITE IS HARD
Inflation is as natural as Minnesota rainfall in June Inflation is as natural as Minnesota snowfall in June	Inflation is perfectly natural Patient is in a natural result of other things

Table 8: Examples that were rejected due to having significant spelling or grammar errors.

3. Do not follow format

Paired sentences	Possible answers
This attack is as weak as a feather This attack is as weak as a breeze	The attack is useless The attack doesn't work
My car motor is dusty like old cave My car motor is dusty like abandon building	Car motor is very rusty car motor is very dusty
the writer is stuck between a rock And another hard place the writer is stuck between a rock And a pebble	He is just stuck doesnt have a choice The writer can get over the pebble

Table 9: Examples that were rejected due to not following the specified format.

Efforts were made to ensure that the final dataset contains no offensive content or personally identifiable information. WorkerID and other potentailly personally identifying information were not included.

C Backward accuracies

Model	Zero-shot	Fine-tuned (L)
GPT-2	52.18	52.00
GPT-neo 1.3B	54.36	63.44
GPT-3 Curie	**58.46**	**74.83**

Table 10: Zero-shot and finetuned backward autoregressive model accuracies on the test set

D Accuracy breakdown by Part-of-Speech

D.1 Subject

Part of speech	Accuracy	Frequency
NN	0.8569	538
PRP	0.8526	156
PRP$ NN	0.9	110
NN NN	0.8889	63
DT NN	0.8182	44
NN NN NN	0.9375	32
JJ NN	0.9167	12

Table 11: Accuracy breakdown and frequency of parts of speech in metaphor subjects. Only part-of-speech patterns with greater than 10 occurrences are shown.

D.2 Relation

Part of speech	Accuracy	Frequency
VBZ NN IN	0.8421	152
VBD RB JJ IN	0.8904	146
VBZ RB JJ IN	0.8889	99
VBZ	0.8352	91
VBD NN IN	0.8806	67
VBD	0.9180	61
VBN IN	0.9545	22
NN IN	0.8636	22
VBD JJ IN	0.9048	21
NNS IN	0.8889	18
VBD IN	0.8462	13
VBZ IN	1.0	13
VBD RB VBN IN	0.8182	11

Table 12: Accuracy breakdown and frequency of parts of speech in metaphor relations. Only part-of-speech patterns with greater than 10 occurrences are shown.

D.3 Object

Part of speech	Accuracy	Frequency
NN	0.8788	429
NN NN	0.8992	129
JJ NN	0.8352	91
NN IN NN	0.8372	43
JJ NN NN	0.8710	31
NN NN NN	0.9130	23
VBG NN	0.9545	22
NN IN JJ NN	0.6154	13
PRP$ NN	1.0	11
JJ	0.6364	11
NN IN NN NN	0.8182	11

Table 13: Accuracy breakdown and frequency of parts of speech in metaphor objects. Only part-of-speech patterns with greater than 10 occurrences are shown.

E Accuracy breakdown by hypernyms

E.1 Subject

Synset	Accuracy	Frequency
adult.n.01	0.8736	182
male.n.02	0.8684	152
woman.n.01	0.7391	46
female.n.02	0.9130	46
show.n.03	0.875	24
product.n.02	0.8636	22
motor_vehicle.n.01	0.9048	21
activity.n.01	0.8421	19
emotion.n.01	0.6667	18
publication.n.01	0.8333	18
feline.n.01	0.9375	16
being.n.01	0.7143	14
performer.n.01	0.8333	12
canine.n.02		12
body_covering.n.01	0.8333	12
vessel.n.03	0.8333	12
sound.n.01	1.0	12
domestic_animal.n.01	0.9167	12
person.n.01	0.8	10
scheme.n.01	0.9	10
contestant.n.01	1.0	10

Table 14: Accuracy breakdown and frequency of WordNet hypernyms in metaphor subjects. Only hypernyms with 10 or greater occurrences are shown.

E.2 Object

Synset	Accuracy	Frequency
time_period.n.01	0.85	20
natural_object.n.01	0.8947	19
person.n.01	0.8824	17
large_integer.n.01	0.9286	14
adult.n.01	1.0	14
solid.n.01	0.9167	13
male.n.02	1.0	13
child.n.02	0.8333	12
body_of_water.n.01	0.75	12
body_covering.n.01	0.8333	12
digit.n.01	0.9167	12
region.n.01	0.8182	11
beverage.n.01	0.8182	11
juvenile.n.01	0.8182	11
container.n.01	0.9	10
rodent.n.01	0.9	10
feline.n.01	0.7	10
building.n.01	0.8	10
time_unit.n.01	0.9	10
travel.v.01	0.7	10

Table 15: Accuracy breakdown and frequency of WordNet hypernyms in metaphor objects. Only hypernyms with 10 or greater occurrences are shown.

F Generation examples

Generation examples can be found in Table 16.

Startphrase	Completion	Label
His temper was a marshmallow. That is to say,	he was very easily angered, and once he was, he was very hard to cool down	Incorrect
He’s got the swimming ability of a pack of dolphins. That is to say,	he’s got none	Incorrect
The villain is as beautiful as the Joker. That is to say,	he’s not	Correct
The child has the energy of a sloth. That is to say,	he is lazy	Correct
The girl moved as fast as a turtle. That is to say,	she moved at a turtle’s pace	Literal
The sadness of her death was a flea.That is to say,	It was a flea that was a sadness	Literal

Table 16: Examples of completions generated by GPT-3 Davinci.