Metaphor Understanding Challenge Dataset for LLMs

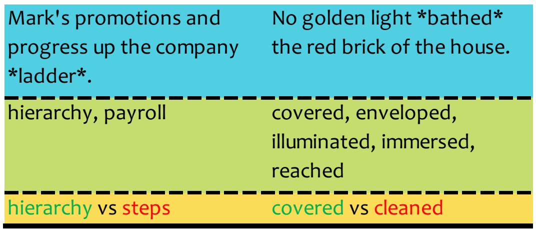
Xiaoyu Tong[♡], **Rochelle Choenni**[♡], **Martha Lewis**[◇,♣] and **Ekaterina Shutova**[♡]
[♡]ILLC, University of Amsterdam, the Netherlands
[◇]School of Engineering Mathematics and Technology, University of Bristol, UK
[♣]Santa Fe Institute, Santa Fe, NM, USA
{x.tong,r.m.v.k.choenni,e.shutova}@uva.nl
martha.lewis@bristol.ac.uk

Abstract

Metaphors in natural language are a reflection of fundamental cognitive processes such as analogical reasoning and categorisation, and are deeply rooted in everyday communication. Metaphor understanding is therefore an essential task for large language models (LLMs). We release the **M**etaphor **Un**derstanding **Ch**allenge Dataset (MUNCH), designed to evaluate the metaphor understanding capabilities of LLMs. The dataset provides over 10k paraphrases for sentences containing metaphor use, as well as 1.5k instances containing inapt paraphrases. The inapt paraphrases were carefully selected to serve as control to determine whether the model indeed performs full metaphor interpretation or rather resorts to lexical similarity. All apt and inapt paraphrases were manually annotated. The metaphorical sentences cover natural metaphor uses across 4 genres (academic, news, fiction, and conversation), and they exhibit different levels of novelty. Experiments with LLaMA and GPT-3.5 demonstrate that MUNCH presents a challenging task for LLMs. The dataset is freely accessible at https://github.com/xiaoyuisrain/metaphor-understanding-challenge.

1 Introduction

Large language models (LLMs), such as BERT (Devlin et al., 2019), GPT-3 (Brown et al., 2020) and LLaMA (Touvron et al., 2023), have become a common paradigm in natural language processing (NLP). Several benchmarks have been proposed to investigate the capabilities of LLMs (Srivastava, 2022; Liang et al., 2022; Hendrycks et al., 2021); and comprehensive analyses have been conducted, evaluating their performance on a range of NLU tasks (Zhong et al., 2023; Qin et al., 2023; Kocoń et al., 2023; Ye et al., 2023; Bang et al., 2023). The community has extensively examined LLM performance on question answering, summarisation, sentiment analysis, natural language inference; a few studies have also shed light on LLMs' analogical reasoning capabilities (Czinczoll et al., 2022; Webb et al., 2023). However, the ability of LLMs to comprehend metaphor—a fundamental linguistic and cognitive tool—is still poorly understood.



Mark's promotions and progress up the company *ladder*.	No golden light *bathed* the red brick of the house.
hierarchy, payroll	covered, enveloped, illuminated, immersed, reached
hierarchy vs steps	covered vs cleaned

Figure 1: MUNCH dataset samples. Each metaphor sample has a ∗highlighted∗ word that is metaphorically used, and is accompanied by up to 5 crowdsourced paraphrases: Substituting the highlighted word with one of the provided words should result in an apt paraphrase. For a selection of metaphor samples, we also provide expert annotation: a pair of correct and incorrect substitution words.

Metaphors are linguistic expressions based on conceptual mappings between a target and a source domain (Lakoff and Johnson, 1980). The verb phrase *to <u>stir</u> excitement*, for example, is based on the conceptual metaphor FEELING IS LIQUID, with FEELING (excitement) being the target domain and LIQUID (something that can be stirred) the source domain. The metaphor compares FEELING with LIQUID, introducing vividness into the description of an otherwise intangible emotional impact. Such cross-domain mappings are sets of systematic ontological correspondences, mapping concepts and their relational structure across distinct domains. Performing this mapping is an essential part of reasoning involved in the interpretation of metaphorical language (Lakoff, 2014; Grady et al., 1999; Gentner and Markman, 1997).

Humans use metaphors so naturally and frequently that they largely fly under our radar. In

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3517–3536
August 11-16, 2024 ©2024 Association for Computational Linguistics

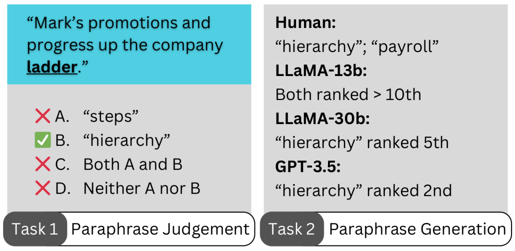


Figure 2: Two tasks for MUNCH: Given a sentence containing a metaphorically used word, a model is prompted to 1) select correct paraphrases from two given candidates (Paraphrase Judgement), and 2) paraphrase the sentence by replacing the highlighted metaphorically used word (Paraphrase Generation).

one of the largest metaphor corpora annotated by linguists, the VU Amsterdam Metaphor Corpus (VUA; Steen et al., 2010b), every 8th word is metaphorical, as averaged over four different genres, including academic and conversation. LLMs, therefore, require the ability to comprehend metaphor in order to have a full command of language. As such, metaphor understanding is an essential task for evaluating the capabilities of LLMs.

Several corpora have been created that contain metaphor annotations at either word or sentence level. These include the VUA corpus (Steen et al., 2010b), the LCC metaphor datasets (Mohler et al., 2016) and the metaphor-emotion dataset of Mohammad et al. (2016), among others. These datasets have been widely used to develop and evaluate automated metaphor identification systems (see Tong et al. (2021) for a survey), but they do not contain information of how the annotated metaphors are interpreted. On the other hand, several works developed datasets with a focus on interpretation, typically casting the problem as a paraphrasing task (Shutova, 2010; Bizzoni and Lappin, 2018; Joseph et al., 2023). Yet, those datasets were often small in scale (containing 200–1000 instances) and were not designed to test the reasoning process by which a metaphor is interpreted, which remains an open question.

This paper presents a novel Metaphor Understanding Challenge Dataset for LLMs (MUNCH). It provides over 10k paraphrases for metaphorical sentences and 1.5k triples of a metaphorical sentence and two candidate paraphrases, which could be apt or inapt (for dataset examples see Figure 1; for statistics see Appendix A). The metaphorical sentences were extracted from VUA texts, spanning four genres (academic, news, fiction, and conversation) and featuring metaphors at different levels of novelty. Each metaphorical sentence contains a content word that is marked as metaphorically used. A candidate paraphrase replaces the metaphorical word with another word, so that the resulting sentence is the same as the reference sentence except for that one word, therefore representing a lexical substitution task. An apt paraphrase shows correct contextual interpretation of the metaphor while an inapt paraphrase uses a word that is related to a literal, source domain sense of the metaphorical word (see the examples of correct and incorrect substitution words in Figure 1). Such a setup of the task is inspired by the conceptual metaphor theory (Lakoff and Johnson, 1980) and allows us to investigate whether the model performs full metaphor interpretation by cross-domain mapping or rather resorts to more shallow lexical similarity. In order to investigate this in a more controlled fashion, we opted for a lexical substitution task. Specifically, we test whether the model consistently chooses the correct target domain paraphrase (therefore, fully interpreting the metaphor) or rather bases its decisions on lexical similarity and chooses the inapt paraphrase that is similar in meaning to the literal use of the metaphorical word.

We set up a fill-in-the-blank task to crowdsource apt paraphrases, and manually selected the best paraphrases using expert knowledge. We also manually created inapt paraphrases from WordNet synsets, so that the apt and inapt paraphrases reflect the target and source domains of the metaphors respectively. Specifically, the inapt paraphrases are synonyms or hypernyms associated with the word's literal use (the source domain).

Using this dataset, we tested the metaphor understanding capabilities of LLaMA-13B, 30B, and GPT-3.5 zero-shot in two tasks: paraphrase judgement and paraphrase generation, as illustrated in Figure 2. Our results show that both tasks are challenging for the models. In particular, the models are prone to confuse the target and source domains of the metaphors, as they often fail to distinguish the inapt paraphrases from the apt paraphrases or reference sentences. Our experiments also reveal that LLMs' metaphor understanding capabilities are associated with genre, metaphor novelty, and POS of the metaphorical word. The MUNCH tasks thus allow us to gain insight into how LLMs process metaphors as well as how this remarkable ability can be improved in the future.

2 Related work

Steen et al. (2010b) created VUA, which marks out **metaphor-related words (MRWs)** in a 4-million-word subset of the British National Corpus. MRWs are lexical units implicative of underlying cross-domain mappings; they can be directly or indirectly used, depending on their contextual meaning. Consider the sentence "A small five-year-old perched like a mosquito on the beginners' pony". The noun *mosquito* is a **direct metaphor**, as it literally means mosquito (the source domain) in this context. The verb *perched* is an **indirect metaphor**, because it has a more basic usage, as in "A pair of glasses were perched on the bridge of his nose".

VUA has been widely used in studies on automated metaphor detection (Leong et al., 2018, 2020; Choi et al., 2021; Zhang and Liu, 2022; Li et al., 2023). However, the corpus does not specify the conceptual metaphors indicated by the MRWs or provide annotation for interpreting the metaphors. The corpus is not directly applicable to automated metaphor interpretation.

Shutova (2010) defined automated metaphor interpretation as a paraphrasing task: Given a metaphorical expression where a word is marked as metaphorically used, the model should replace this word with another word to render a literal paraphrase of the expression. For example, the verb phrase *stir excitement*, where *stir* is used metaphorically, should be paraphrased as *provoke excitement*.

Bizzoni and Lappin (2018) created a Metaphor Paraphrase Evaluation Corpus (MPEC), which provides correct and incorrect paraphrases for ∼200 short sentences containing metaphor use; paraphrases could greatly differ from the reference sentences. Joseph et al. (2023) created the NewsMet dataset, which consists of 1k verbal metaphors in news headlines as well as their literal equivalents; incorrect paraphrases are not provided.

Several recent studies approached metaphor understanding as an inference task. The IMPLI dataset (Stowe et al., 2022) includes entailed and non-entailed sentences for ∼900 metaphorical sentences. The FLUTE dataset (Chakrabarty et al., 2022b) provides entailment and contradiction pairs for 1500 metaphorical sentences (including 750 similes). Fine-tuned transformer-based models reached > 0.8 accuracies in these 2 studies in predicting the class of a given sentence pair.

Recent studies also employed multiple-choice and generative tasks to assess LLMs' ability to reason with metaphorical language. The MiQA benchmark (Comșa et al., 2022) uses such tasks to test whether models can distinguish metaphorical and literal uses of the same words; 150 conventional metaphors are involved. The Fig-QA task (Liu et al., 2022) includes 10k similes (a type of direct metaphor) and requires models to distinguish a pair of metaphors of opposite meanings. Chakrabarty et al. (2022a) examined LLMs' figurative language understanding by asking them to generate text after encountering an idiom or simile.

The MUNCH dataset provides 3k samples of indirect metaphors, 10k correct paraphrases, and 1.5k incorrect paraphrases. It is therefore one of the largest datasets for paraphrasing of indirect metaphors. The candidate paraphrases are also systematically different from the ones in previous datasets, as we tailored the dataset for testing LLMs' understanding of metaphors as cross-domain mappings and correctly capturing the underlying relational structures. We summarise differences between MUNCH and previous datasets and provide more details for the latter in Appendix A.

3 Data collection: metaphor samples

The metaphor samples in our dataset were selected from the publicly available metaphor corpus VUA. Each metaphor sample is a sentence containing a highlighted MRW, the metaphorical word to be interpreted; a paraphrase uses a single word to substitute the metaphorical word. We use two criteria for selecting metaphorical sentences: novelty of the metaphor and possibility of single-word substitution. We explicate our selection process below.

The novelty criterion. We employed novelty scores from Do Dinh et al. (2018) to increase the proportion of novel metaphors in our dataset. Scores range from -1 to 1. VUA contains a large proportion of conventional metaphors: The metaphorical use of the word can be found in a dictionary of contemporary language use (Steen et al., 2010a). As LLMs might have encountered enough data for such conventional metaphor uses during pre-training, the understanding of such metaphors should be relatively easy. To render a more challenging dataset, we excluded MRWs with novelty scores below -0.3. Metaphors with a novelty score higher than -0.3 could still be conventional: The crowd workers who provided the novelty annotations in Do Dinh et al. (2018) relied on their intuition instead of a dictionary like Steen et al. (2010a).

And metaphorical uses included in dictionaries may still be considered novel by lay people. We chose -0.3 as the threshold in order to collect a large and diverse dataset as a starting point.

The single-word criterion. To ensure that the metaphorical sentences can be paraphrased via single-word substitution, we excluded MRWs that are marked as direct metaphors, as well as a portion of indirect metaphors. Directly used MRWs usually occur in a sequence, such as "I knew the pathway like the back of my hand". They are thus not suitable for single-word substitution. Also, the direct metaphor *back of my hand* refers literally to the back of the speaker's hand—its contextual meaning is directly associated with the source domain. This is contrary to our task setup, where apt paraphrases (contextual meaning) should be associated with target domains. We therefore opted to focus on indirect metaphors in this study.

Within the category of indirect metaphors, we filtered out new-formations, consecutive MRWs, and proper names. New-formations are words that do not have an entry in dictionary, so VUA annotated the parts that do have corresponding entries. For example, in the phrase *a rose-tinted vision of the world*, the word *rose-tinted* was a new-formation; so *rose* and *tinted* are marked as separate MRWs in VUA and received their separate novelty scores. We filtered these out because a single metaphorical word should have a single novelty score (*rose-tinted* has two), yet it is hard to paraphrase *rose* or *tinted* instead of *rose-tinted* altogether.

Likewise, we excluded cases where multiple content words marked as indirect metaphors occur consecutively, such as *take place*, *long road home*, *great leap forward*. These often involve fixed expressions or phrases that either should be replaced as a whole or should not be marked as consecutive indirect metaphors. We also excluded metaphorical words that are part of a proper name, which, like fixed expressions, need to be treated as a whole. For example, the proper name *Nord Stream* would lose its meaning if one changed the metaphorical word *stream* into another word.

4 Annotation of apt paraphrases

Crowdsourcing task. We constructed a fill-in-the-blank task to crowdsource (apt) paraphrases for the metaphorical sentences. Each task included 30 sentences to be paraphrased, so that the task can be finished within 30 minutes. Under each sentence, the workers were presented with a copy of the sentence where the metaphorical word is replaced with a blank; they were asked to fill the blank with a single word so that the new sentence is a semantically and grammatically apt paraphrase of the reference sentence. If they were not able to paraphrase the sentence, they were asked to explain why it was difficult. Examples of good and bad answers were provided in the instructions (see Appendix B).

The workers were recruited via Prolific[1]. We set prescreening criteria to only include adult (age > 18) native English speakers who were living in an English-speaking country and did not have any language-related disorders. The workers were asked to confirm within the task that they met these criteria. After giving consent to participate and reading the instructions, they were also required to correctly paraphrase a trial sentence in order to access the task. More details (worker's consent, the trial sentence) are given in Appendix B.

We released 99 tasks in total and collected 5 data points for each of the 2970 reference sentences. We received single-word substitutions for 2953 sentences (the other 17 are presented and explained in Appendix B), and 61% of them got repeated answers—multiple workers submit the same paraphrase despite the question being open-ended. This confirms the reliability of our task.

Expert validation. For a selection of the reference sentences for which we later annotated inapt paraphrases (Section 5), we further validated the crowdsourced paraphrases to determine the best paraphrase for creating the triples (one metaphor sample, two candidate paraphrases).

We used both majority vote and expert knowledge to find one best paraphrase for each reference sentence. For each sentence, we first sorted the collected single-word substitutions from the most to the least popular (in terms of how many participants proposed that substitution). The apt paraphrase that was proposed by the highest number of participants was verified by the authors and selected as the best paraphrase for that reference sentence.

When multiple apt paraphrases have the same number of votes, we chose the one that is clearly within the target domain—that is, the paraphrase clearly shows that the metaphorical word is interpreted in its contextual sense. For instance, we received 5 different single-word substitutions for the metaphorical word *attack* in the sentence "...he

[1] https://www.prolific.co/

has become involved in a row over his attack on the "Pharisees" of British society". These are *remarks*, *views*, *offense*, *incursion*, and *disagreement*, each proposed by a single participant. All of them can be considered apt paraphrases. We chose *remarks* because it clearly shows the metaphorical word *attack* is interpreted in the ARGUMENT domain. The meaning of *offense* and *disagreement* are more abstract and could involve other conceptual domains; the paraphrases that replace *attack* with *views* and *incursion* respectively are still metaphorical, as *view* can be associated with VISION and *incursion*, like the metaphorical word itself, is still in the domain of BATTLE. These four are thus less preferable with regard to the purpose of our dataset.

While we managed to find one best paraphrase for most reference sentences, there are 45 for which we selected two paraphrases as the best, as the two received the same votes and are equally apt. There are also 11 sentences for which no paraphrase was selected. These are cases where the given context is insufficient for determining the contextual meaning of the metaphorical word.

5 Annotation of inapt paraphrases

Tong (2021) shows that incorrect paraphrases based on the basic sense of the metaphorical word are the least distinguishable from correct ones (i.e., paraphrases based on the contextual sense) with respect to aptness. To render truly challenging inapt paraphrases for our task, we therefore created inapt paraphrases exclusively from basic senses.

We employed WordNet for identifying basic senses and obtaining sense-specific synonyms, following the annotation guidelines presented in Appendix C. For each metaphorical word, we first locate the WordNet synsets that correspond to its more basic meaning (relative to its contextual meaning in the reference sentence). Then we go through the synonyms (or hypernyms when no synonyms are provided) under the basic-sense synsets and select those that are clearly associated with the metaphor's source domain and would render a grammatical (but inapt) paraphrase.

	ACPROSE	NEWS	FICTION	CONVRSN	TOTAL
	1061	922	593	377	2953
N	50%	38%	35%	25%	40%
V	35%	42%	39%	51%	40%
A	15%	20%	26%	24%	20%

Table 1: Number of metaphor samples per genre (academic, news, fiction, conversation), and the percentage of sentences where the metaphorical word is a noun (N), a verb (V), or either an adjective or an adverb (A).

We went through all 2970 sentences released for the crowdsourcing task and found inapt paraphrases for 991 of them. After removing items lacking apt paraphrases (either because no single-word substitutions were crowdsourced or because none of the collected ones are of sufficient quality), we created 1492 triples for 728 metaphorical sentences, including 1072 triples with an apt and an inapt paraphrase, 375 triples with two inapt paraphrases, and 45 with two apt paraphrases.

Inter-annotator agreement. From the 991 sentences for which inapt paraphrases were identified, we randomly selected 200 to be annotated by a second annotator. The annotator was a PhD candidate in linguistics specialising in metaphor research. We explained the annotation process to the second annotator through a meeting and the guidelines in Appendix C. The Gwet's gamma coefficient for the agreement between the two expert annotators is 0.84.

6 Data analysis

The dataset contains approximately the same number of samples from academic and news genres, and fewer samples from fiction and conversation, as shown in Table 1. These metaphor samples cover metaphorical use of content words in all four parts of speech. Noun and verb MRWs are of a higher proportion compared to adjectives and adverbs. In news and fiction, these two categories have similar percentages. The academic genre contains more noun MRWs than verbs whereas in conversation the situation is reversed: Half of the metaphorical words are verbs, while the percentage of nouns is similar to that of adjectives and adverbs.

As we excluded MRWs of novelty scores lower than -0.3, the metaphor samples exhibit a wider range of novelty scores above 0 than below 0 (see Appendix D). Meanwhile, a large proportion of the metaphor samples could be considered only slightly novel or conventional (novelty scores between -0.3 and 0.3). Metaphor samples of the highest novelty scores can be from any of the four genres. Despite their different proportion in the entire dataset (Table 1), all four genres include metaphor samples across all levels of perceived novelty.

We also examined the cosine similarity between the metaphorical words and apt and inapt substitu-

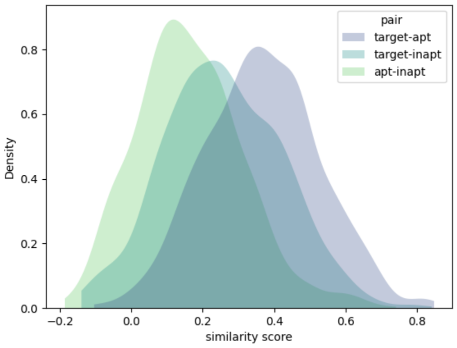


Figure 3: Distribution of the cosine similarity between target-apt, target-inapt, and apt-inapt pairs.

tions. Since the inapt paraphrases were based on the more basic meaning of the MRWs (section 5), we expected inapt substitution words to be more similar to the metaphorical words than apt substitution words. We computed the cosine similarity scores using `glove-wiki-gigaword-300` embeddings (Pennington et al., 2014), accessed through the `gensim` Python library. Figure 3 shows the distribution of cosine similarity scores for 1006 triples, excluding the ones containing out-of-vocabulary words. Surprisingly, the target-apt pairs tend to have higher cosine similarity scores than the target-inapt pairs. The plot suggests that the 3 pairs are distinguishable in terms of cosine similarity scores, with target-apt pairs being the most similar, and apt-inapt the least similar. This might be associated with the fact that our metaphorical sentences were sampled from VUA, which, being representative of metaphor use in natural discourse, includes a large percentage of conventional metaphors. Nonetheless, the majority of the cosine similarity scores are above 0, and the 3 pairs still share a wide range of similarity scores. The distribution plot is therefore also suggestive of the reliability of our dataset, as well as its potential challenge for LLMs.

7 Model evaluation

We evaluated LLaMA-13B, LLaMA-30B, and GPT-3.5 (`text-davinci-003`) on two tasks: (Task 1) **paraphrase judgement**, which requires a model to select correct paraphrases for a given reference sentence from given candidates; and (Task 2) **paraphrase generation**, which asks a model to generate correct paraphrases for a given reference sentence. The paraphrase judgement task used the 1492 triples that include inapt paraphrases; the generation task used all 2953 metaphorical sentences. Details regarding computational budget is given in Appendix E.

```
Select words that can replace the highlighted word in the given
sentence without altering the sentence's meaning.
Sentence: ... *extending* the Government's borrowing power ...
Option A: increasing
Option B: exserting
Option C: Both Option A and Option B
Option D: Neither Option A nor Option B
Correct answer: Option
```

Figure 4: Example prompt for the *Word-judgement* task (the *Implicit* condition). The given sentence is shortened for illustration.

7.1 Paraphrase judgement

We evaluate the LLMs in a prompting setup. We test the models' ability to interpret metaphor under different conditions. In the first scenario, we prompt the model by providing the reference sentence with the metaphorical word highlighted and two candidate replacement words for it (*Word-judgement*). In the second scenario, each of the candidate replacement words is embedded in the sentence (*Sentence-judgement*). In both cases the model needs to solve a multiple choice task. Besides providing the apt and inapt paraphrases (Options A and B) as answer options, we also complement them with Option C, that both candidates are correct, or Option D, that neither are correct. See Figure 2 for an example. We expect *Word-judgement* to be more challenging, as the model would need an additional inference step compared to sentence judgement, to replace the metaphorical word with the two given options and (implicitly) form the intended paraphrases.

For both *Word-judgement* and *Sentence-judgement* setups, we further investigate whether it makes a difference if the model is explicitly "told" that the task is to paraphrase a metaphor or not. This results in three further conditions: *Implicit* (not mentioning metaphor in the prompt), *Metaphor-Sent* (revealing that the reference sentence contains a metaphor), and *Metaphor-Word* (revealing that the specific highlighted word in the sentence is metaphorically used). The *Implicit* condition corresponds best to the real-life application of LLMs, where the model needs to be able to interpret metaphors without being instructed that metaphors are there.

We tested LLaMA-13B and 30B, and GPT-3.5 in each of the 6 conditions, using 3 prompts for each condition (the prompts are listed in Appendix E).

	LLaMA-13B	LLaMA-30B	GPT-3.5
Word-judge			
Implicit	.28 (.18)	.21 (.10)	.23 (.10)
M-Sent	.30 (.16)	.19 (.09)	.20 (.10)
M-Word	.33 (.18)	.21 (.08)	.20 (.08)
Sent-judge			
Implicit	.13 (.06)	.14 (.03)	.17 (.07)
M-Sent	.12 (.07)	.17 (.03)	.16 (.06)
M-Word	.10 (.08)	.27 (.05)	.21 (.02)

Table 2: Mean (SD) accuracies across 3 prompts for each paraphrase judgement condition.

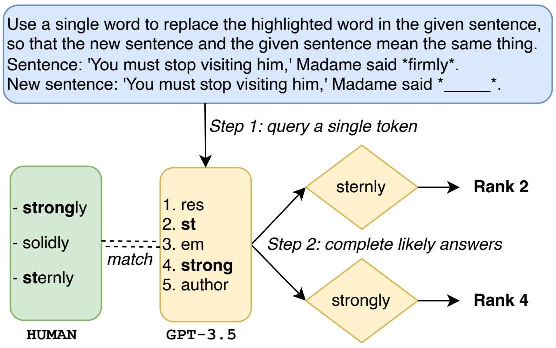

Use a single word to replace the highlighted word in the given sentence, so that the new sentence and the given sentence mean the same thing.
Sentence: 'You must stop visiting him,' Madame said *firmly*.
New sentence: 'You must stop visiting him,' Madame said *______*.

Figure 5: Procedure of the paraphrase generation task, using GPT-3.5 prompt and outputs as example. We first ask the model to generate a single token to get a glimpse of its top 5 answers. For each token that matches the beginning of a human answer, we let the model complete it to see whether it is a complete match.

Table 2 shows the mean accuracy and standard deviation for each model in each condition. The random baseline achieves an accuracy of 0.25, as there is always only one correct option out of the given four. The performance of all three models was below the random baseline in most cases, except for LLaMA-13B in the *Word-judgement* tasks and LLaMA-30B in the *Metaphor-Word* condition of the *Sentence-judgement* task. Meanwhile, the accuracy of LLaMA-13B varied a great deal across different prompts in the *Word-judgement* tasks.

The *Sentence-judgement* task seems to be more challenging than *Word-judgement* for the models. For LLaMA-30B and GPT-3.5, the task was particularly difficult when they were not instructed to focus on the metaphorical word, and were not informed that the word is metaphorically used (the *Metaphor-Word* condition). For LLaMA-13B, all 3 *Sentence-judgement* conditions are similarly difficult. However, its higher accuracies in the *Word-judgement* tasks also indicate the benefit of instructing the model to focus on the metaphorical word.

	MRR	Recall@5	Recall@10
LLaMA-13B	.33 (.02)	.22 (.02)	.33 (.02)
LLaMA-30B	.47 (.03)	.28 (.02)	.40 (.03)
GPT-3.5	.54 (.02)	.32 (.01)	-

Table 3: Mean (SD) performance across 3 prompts in the paraphrase generation task. Recall@10 does not apply to GPT-3.5 as the OpenAI API only allows access to the top-5 answers.

7.2 Paraphrase generation

The purpose of this task is to compare model and human performance in paraphrasing metaphorically used words. The prompts were thus designed to be semantically close to the instructions in our crowdsourcing task (Section 4). The model answers were generated in two steps (see Figure 5). We first let the models generate a single token—this allowed us to access the models' ranking of all tokens in their vocabulary. Of these, we selected the ones that match human annotations and let the models complete them into words. The completions were then compared with human annotations to determine the rank of each expected answer.

We tested the models on 3 prompts (see Appendix E) and their mean performance in terms of mean reciprocal rank (MRR), recall at top 5 paraphrases and recall at top 10 paraphrases is shown in Table 3. GPT-3.5 performed best and LLaMA-30B came second. The models' performance was also more stable across different prompts as compared to the paraphrase judgement task. Nonetheless, all three models clearly preferred different answers as compared to human annotators.

8 Discussion

Paraphrase judgement. We looked into the type of errors the models made in paraphrase judgement. The number of each combination of expected and predicted answers for each model is in Appendix F. We found that LLaMA-30B and GPT-3.5 could ignore the semantic differences between a given sentence and an inapt paraphrase, as they tend to predict both candidates as correct when presented with one or more inapt paraphrases. LLaMA-13B, on the other hand, tends to assume that the two given candidates always contain one apt and one inapt paraphrase. Nonetheless, it did not seem capable of distinguishing the two, as it made a similar number of Option A and Option B predictions.

Paraphrase generation. We examined the top-ranked answers of the models and found 4 categories that the ‘incorrect’ or unexpected answers could fall into. **1) Nonsensical:** For the sentence “...for this number line I would say...”, GPT-3.5 gives *thus* as the best substitution word, ignoring the meaning of *line*, whereas LLaMA-13B repeats *number*. **2) Lack of contextual understanding:** In “...he touched both sides of the coin...”, GPT-3.5 replaces the word *sides* with *facets*, suggesting that it neglects details of the meaning of the sentence (that a coin only has 2 sides). **3) Ungrammatical:** On the other hand, the model may have understood the metaphor, but fails to convert its understanding into a suitable substitution word. In “...they all shared the emphasis on ‘her’...”, LLaMA-30B suggests *concurred* as the best answer, implying that the meaning of *shared* is understood, but that grammatical agreement has been sacrificed. **4) Preference:** Finally, the disagreement between the models and human annotators may simply be a matter of preference. For “For a man whom Rebecca West ... called ‘repulsive’ and ‘treacherous’...”, crowd workers provide 4 possible answers: *revolting*, *disgusting*, *awful*, and *grotesque*. Both LLaMA-30B and GPT-3.5 give *odious* as the best answer. Here, both the human annotators and the models understand and can paraphrase the sentence, and it is hard to say whose answer is best.

Factors associated with model performance. We also examined the association between model performance and 3 factors: genre, metaphor novelty, and the POS of the metaphorical word. The details are available in Appendix F. We found metaphors of higher novelty scores to be more difficult for LLaMA-30B in paraphrase judgement, and for GPT-3.5 in paraphrase generation. The association between genre or POS and model performance tends to differ per model and task. The fiction genre, for example, is the easiest for the LLaMA models in paraphrase generation; yet it is the most difficult for GPT-3.5 in the generation task and for LLaMA-30B in the judgement task. Similarly, noun metaphors are the easiest for LLaMA-30B in the generation task and for GPT-3.5 in the judgement task. Meanwhile for GPT-3.5 in the generation task, adverb metaphors become the easiest.

To sum up, the results of the two paraphrase tasks indicate that the LLMs are unable to (fully) understand some of the metaphors in our dataset. The paraphrase judgement task further reveals that the models have difficulty distinguishing the metaphors’ source domains (implied by the inapt paraphrases) and target domains (implied by the reference sentences and apt paraphrases). This further suggests that the models are unlikely to perform reasoning across semantic domains; when they succeed in understanding the metaphor, they may still reason in ways that are different from humans. This means that for downstream NLP tasks such as opinion mining, bias detection, humour detection, and intent recognition, the LLMs could overlook the entailment of a metaphor. In machine translation as well as summarisation of highly figurative or poetic texts, the problems may manifest as incorrect or peculiar explanation of metaphors.

A direction for improvement is to mark out metaphor uses in texts and direct the model’s attention to them: In the paraphrase judgement task, the models reach higher accuracies when the metaphorical word is marked out (in the *Word-judgement* task or in a *Metaphor-Word* condition). However, since the models generally performed poorly in the experiments, the LLMs may need to be fine-tuned in order to better understand metaphors. When fine-tuning, one can consider increasing the proportion of certain metaphor types in training data, as genre, metaphor novelty, and POS of the metaphorical word are all associated with model performance. Future studies could first employ MUNCH to detect the weak points of an LLM and then curate training data accordingly.

9 Conclusion

We release a dataset of manually created apt and inapt paraphrases for metaphorical sentences and present two metaphor understanding tasks, which we demonstrate to be challenging for current LLMs. The errors the models make in the paraphrase generation task indicate various levels of misunderstanding of the metaphors. In the paraphrase judgement task, the models’ accuracy was lower than the random baseline in the majority of the cases; a closer look at their errors reveals that the models had difficulty in detecting the inaptness of the inapt paraphrases. The experiments also show that the models performed better when being instructed to focus on the metaphorical word, and that genre and the POS and novelty of the metaphorical word are all potential factors that affect model performance.

10 Limitations

We designed the metaphor understanding tasks to be representative of a lexical substitution task: The metaphorical word is the only difference between a reference sentence and a candidate paraphrase. This setup makes it possible to examine whether LLMs indeed perform metaphor interpretation or resort to lexical similarity when they encounter metaphorically used words. At the same time, however, it also means that the models' understanding of multi-word metaphors and direct metaphors (e.g., similes) could not be tested using our dataset.

We tested the lastest and state-of-the-art LLMs available at the time that our study was ongoing, but newer LLMs such as GPT-4 and Llama 2 emerged shortly after the completion of our study. We suggest running a data contamination test before evaluating newer LLMs using our dataset.

This study reveals that LLMs have difficulty distinguishing the target and source domains of linguistic metaphors. A more extensive analysis is desirable to uncover more differences between LLMs and humans in terms of metaphor interpretation.

11 Ethics statement

We abide by the ACL Code of Ethics. The metaphor resources used in this study are publicly available. The crowdsourcing task was approved by an ethics committee. The crowd workers received fair payment (9 GBP per hour), and no personal information was collected or stored in our database.

The metaphor samples in our dataset come from excerpts of natural discourse. They may therefore involve bias, taboo, violence, or other aspects of everyday language use that could be considered negative (we also pointed this out to the crowd workers before they gave their consent to participate, as presented in Appendix B). Nonetheless, these are integral parts of language use, and should be properly understood by NLP systems, which is precisely what this paper aims at.

Acknowledgements

The crowdsourcing task (for collecting apt paraphrases) was funded by the Faculty of Engineering at the University of Bristol, through the faculty's Pump-Priming Fund. We thank Dr. Jie Fu for co-annotating the inapt paraphrases. We also thank all anonymous reviewers who provided constructive feedback for previous versions of this paper.

References

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. 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.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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 *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. 2022a. It's not rocket science: Interpreting figurative language in narratives. *Transactions of the Association for Computational Linguistics*, 10:589–606.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022b. FLUTE: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. MelBERT: Metaphor detection via contextualized late interaction using metaphorical identification theories. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1763–1773, Online. Association for Computational Linguistics.

Iulia Comșa, Julian Eisenschlos, and Srini Narayanan. 2022. MiQA: A benchmark for inference on metaphorical questions. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 373–381, Online only. Association for Computational Linguistics.

Tamara Czinczoll, Helen Yannakoudakis, Pushkar Mishra, and Ekaterina Shutova. 2022. Scientific and creative analogies in pretrained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2094–2100, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Erik-Lân Do Dinh, Hannah Wieland, and Iryna Gurevych. 2018. Weeding out conventionalized metaphors: A corpus of novel metaphor annotations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1424, Brussels, Belgium. Association for Computational Linguistics.

D. Gentner and A. B Markman. 1997. Structure mapping in analogy and similarity. *American Psychologist*, 52(1):45–56.

Joseph Grady, Todd Oakley, and Seana Coulson. 1999. Blending and metaphor. *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4*, pages 101–124.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding.

Rohan Joseph, Timothy Liu, Aik Beng Ng, Simon See, and Sunny Rai. 2023. NewsMet : A 'do it all' dataset of contemporary metaphors in news headlines. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10090–10104, Toronto, Canada. Association for Computational Linguistics.

Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, Anna Kocoń, Bartłomiej Koptyra, Wiktoria Mieleszczenko-Kowszewicz, Piotr Miłkowski, Marcin Oleksy, Maciej Piasecki, Łukasz Radliński, Konrad Wojtasik, Stanisław Woźniak, and Przemysław Kazienko. 2023. Chatgpt: Jack of all trades, master of none. *Information Fusion*, page 101861.

George Lakoff. 2014. Mapping the brain's metaphor circuitry: metaphorical thought in everyday reason. *Frontiers in Human Neuroscience*, 8.

George Lakoff and Mark Johnson. 1980. *Metaphors we live by*. University of Chicago Press.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

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.

Yucheng Li, Shun Wang, Chenghua Lin, Frank Guerin, and Loic Barrault. 2023. FrameBERT: Conceptual metaphor detection with frame embedding learning. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1558–1563, Dubrovnik, Croatia. Association for Computational Linguistics.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2022. Holistic evaluation of language models.

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the ability of language models to interpret figurative language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4437–4452, Seattle, United States. Association for Computational Linguistics.

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.

Michael Mohler, Mary Brunson, Bryan Rink, and Marc Tomlinson. 2016. Introducing the LCC metaphor datasets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*

(LREC'16), pages 4221–4227, Portorož, Slovenia. European Language Resources Association (ELRA).

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is ChatGPT a general-purpose natural language processing task solver? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1339–1384, Singapore. Association for Computational Linguistics.

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.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel

Habacker, Ramón Risco Delgado, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Timothy Telleen-Lawton, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.

Harshvardhan Srivastava. 2022. Poirot at CMCL 2022 shared task: Zero shot crosslingual eye-tracking data prediction using multilingual transformer models. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 102–107, Dublin, Ireland. Association for Computational Linguistics.

Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna A. Kaal, and Tina Krennmayr. 2010a. Metaphor in usage. *Cognitive Linguistics*, 21(4):765–796.

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

Kevin Stowe, Prasetya Utama, and Iryna Gurevych. 2022. IMPLI: Investigating NLI models' performance on figurative language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5375–5388, Dublin, Ireland. Association for Computational Linguistics.

Xiaoyu Tong. 2021. Metaphor paraphrasing and word-sense disambiguation: toward a new approach to automated metaphor processin. Master's thesis, Universitetit van Amsterdam, the Netherlands.

Xiaoyu Tong, Ekaterina Shutova, and Martha Lewis. 2021. Recent advances in neural metaphor processing: A linguistic, cognitive and social perspective. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4673–4686, Online. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Taylor Webb, Keith J. Holyoak, and Hongjing Lu. 2023. Emergent analogical reasoning in large language models.

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, Jie Zhou, Siming Chen, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models.

Shenglong Zhang and Ying Liu. 2022. Metaphor detection via linguistics enhanced Siamese network. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4149–4159, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert.

A Previous metaphor understanding datasets and tasks

Table 4 summarises the differences between MUNCH and previous datasets.

Example (1) is extracted from MPEC. The correct paraphrase, sentence (1-a), is almost completely different from the original sentence. The two distractor sentences that follow indicate different types of misinterpretation: Sentence (1-b) wrongly interprets the meaning of the original sentence, while the last sentence is based on a literal use of the word *wheels*.

(1) the wheels of justice turn slowly

	#met	met:length	#correct	correct:type	#distractor	distractor:type
MPEC	192	9 (4)	218	$s \rightarrow s$	526	mixed
NewsMet	791	12 (3)	791	$w \rightarrow w$	0	NA
IMPLI	913	16 (10)	1032	$w \rightarrow p$	281	context change
FLUTE	1500	11 (5)	1500	$p \rightarrow p$	1500	opposite meaning
MiQA	150	8 (2)	150	$s \rightarrow s$	150	context change
Fig-QA	10256	9 (3)	10256	$s \rightarrow s$	10256	opposite meaning
MUNCH	2953	26 (15)	10261	$w \rightarrow w$	1492	paraphrase

Table 4: Differences between MUNCH and previous datasets that provide paraphrases for metaphors: MPEC (Bizzoni and Lappin, 2018; github.com/yuri-bizzoni/Metaphor-Paraphrase), NewsMet (Joseph et al., 2023; https://github.com/AxleBlaze3/NewsMet_Metaphor_Dataset/tree/main), IMPLI (Stowe et al., 2022; github.com/UKPLab/acl2022-impli), FLUTE (Chakrabarty et al., 2022b; https://github.com/tuhinjubcse/model-in-the-loop-fig-lang), MiQA (Comșa et al., 2022), and Fig-QA(Liu et al., 2022; https://github.com/nightingal3/Fig-QA/tree/master). We present their differences regarding number of metaphor samples (#met), mean (SD) length of the metaphor samples (met:length, measured by number of words), number of correct paraphrases (#correct), the part of a metaphor sample that is replaced to create correct paraphrases (correct:type; s=sentence, p=phrase, w=word), number of distractors (#distractor), and distractor type. The numbers are calculated from the datasets available on GitHub. Note that our dataset is much more extensive than the previous ones.

a. it might take time but eventually justice prevails
b. ¿ justice prevails in very little time
c. ¿ the wheels of a car turn slowly

The MPEC corpus is employed by two metaphor understanding tasks in BIG-Bench (Srivastava et al., 2022). The metaphor-boolean task uses a binary classification setup: Given a pair of sentences, is the second sentence a paraphrase of the first? GPT-2 only reached 0.41 accuracy on this task in a zero-shot scenario. The metaphor-understanding task consists of two subtasks: metaphor to paraphrase, which asks the model to select the correct paraphrase from 4 candidates; and paraphrase to metaphor, which requires the model to distinguish the metaphorical sentence corresponding to a given paraphrase from 3 other metaphors. GPT-2 large performed poorly on both subtasks: In a zero-shot scenario, the model gave 0.27 accuracy on the metaphor-to-paraphrase task, and 0.67 accuracy on the paraphrase-to-metaphor task.

The metaphor-literal pairs in the NewsMet dataset was created with the help of LLMs. Each news headline has a verb considered as the focus word. They first passed the headlines with the focus words masked to ALBERT (Lan et al., 2020) to obtain the first 200 words that can replace the focus word. These 200 words were then passed to a metaphor detector to obtain the top-6 metaphorical and top-6 literal candidates. Human annotators then identified the best literal counterparts for metaphorical focus words and the best metaphorical counterpart for literal focus words.

In the IMPLI example (2), the correct paraphrase (2-a) uses a phrase, *paid for*, to explain the metaphorically used word *absorbed* in the original sentence. The distractor, on the other hand, is based on the literal meaning of *absorbed*. Fine-tuned RoBERTa base and RoBERTa large achieved high accuracies (> 0.8) on labelling these metaphor-paraphrase and metaphor-distractor sentence pairs.

(2) he absorbed the costs for the accident
a. he paid for the costs for the accident
b. ¿ he absorbed the sunlight after the accident

Example (3) is extracted from the FLUTE dataset; included in the parentheses are explanations for the paraphrase and the contradict respectively. Contrary to the MUNCH dataset, the authors aimed at paraphrases that use more than one word to replace a metaphorically used word. Note that sentence (3-b) is more of a direct contradiction of the reference metaphor than the paraphrase, as it preserves the metaphorically used word *louder*. The difference between the contradict and the reference metaphor may thus be easier to detect as compared to a contradict that is more similar to the paraphrase (e.g., *Actions are not more important than words*).

(3) Actions speak louder than words.

a. Actions are more important than words. (This phrase is used to say that what someone does is more important than what they say.)

b. ¿ Actions are not louder than words. (The metaphor suggests that deeds or actions are more important than words, while the contradiction suggests that words are more important than deeds or actions.)

As example (4) shows, each metaphorical premise in the MiQA dataset is paired with a literal premise exemplifying literal use of the metaphorical word; the dataset also includes implications (the text in parenthesis) of the metaphorical and literal premises:

(4) a. I **see** what you mean (I understand you)

b. I **see** what you are pointing at (My eyes are working well)

Comșa et al. (2022) set up 2 binary-choice tasks using the MiQA dataset: (1) Given a metaphorical premise, select the correct implication; (2) given an implication, select the corresponding premise. They also set up a generative task: Given a metaphorical premise, answer whether it implies the literal conclusion. LLMs performed well on these tasks.

The Fig-QA dataset provides similes of opposite meanings as well as their implications (given in parentheses):

(5) a. The meteor was as bright as New York City (The meteor was very bright)

b. The meteor was as bright as coal (The meteor was not bright at all)

The binary-choice task is similar to MiQA: Given a metaphorical premise, select the correct implication. They also develop a generative task which prompt models to generate implications freely. Liu et al. (2022) found these tasks challenging for LLMs in zero-shot settings.

B Crowdsourcing task

The participant information sheet, which was presented to the crowd workers prior to the consent form, has a section dedicated to potential disadvantages and risks involved in participating in the study—

> The sentences you will paraphrase were from a wide range of sources, including newspapers, fiction, and dialogues. You may occasionally encounter violence or taboo topics (e.g., war, crime, sex), as well as potentially disturbing opinions.
>
> If you are concerned, you do not have to give consent; you can also withdraw anytime during the experiment.

The information sheet also explains how data collected from the study will be used. The workers were informed that their participation would remain confidential, that their response would be anonymised, and that the data would be made open access at the end of the study.

The annotation guidelines are shown in Figure 6. The trial sentence is provided in example (6), where *introduce* is the metaphorical word to be interpreted. Our final list of acceptable answers includes: *address*, *advance*, *clarify*, *convey*, *cover*, *define*, *describe*, *discuss*, *elucidate*, *establish*, *explain*, *mention*, *present*, *propose*, *reveal*, *share*, *show*, *state*, *submit*, *suggest*, *teach*, *unveil*.

(6) I shall now **introduce** the concept of an elementary charge, 1.6 × 10 -19 C, carried by an elementary particle called the electron.

Table 5 presents the 17 sentences for which none of the crowd workers were able to provide single-word substitutions for the metaphorical words. These are mainly highly conventionalised metaphors, for which it is usually difficult to find an alternative expression. There are also cases where the target word is part of a multi-part word (e.g., *carry out*, *point of view*) or a phrase (e.g., *put in an appearance*, *get rid of*). These stem from annotation mistakes in VUA: According to the MIPVU procedure, VUA should have marked the entire word or phrase as a single annotation unit. We still collected paraphrases for these cases as there were no suitable way to filter them out automatically.

C Inapt paraphrase annotation

The guidelines for inapt paraphrase annotation are presented in Figure 7.

D Novelty distribution of MUNCH metaphor samples

Figure 8 presents the novelty distribution of the metaphor samples in MUNCH.

1 The summer's sprawl begins to be oppressive at this stage in the year and trigger fingers are itching to snip back overgrown mallows, clear out the mildewing foliage of **golden** rod and reduce the overpowering bulk of bullyboy ground cover.

2 The red and green of the Aztec necklace links it compositionally with the indigenous plants to the 'south' of the painting, the pink colonial-style dress tonally blending with the skyscrapers to the '**north**'.

3 Nine out of 10 are routine calls, many of which could be **carried** out by mini cabs.

4 This example assumes that a sympathy for motorists with **overwhelm** any tendency to logical analysis.

5 There were, in fact, about a **score**.

6 Mrs Bottomley is convinced the Tory victory provides the opportunity to entrench the reforms — and to give doctors, nurses and managers the confidence to **make** them work.

7 Thus, as with biological theories, crime is seen as pathological (a disease), as something to be looked at from the medical **point** of view.

8 'So you've decided to **put** in an appearance?'

9 He was in **there** twice, at a Wimpole Street number and again at an address in Mill Hill: Rufus H. Fletcher, MB, MRCP.

10 Once again he backtracks and assumes a larger unity in which conflict **takes** place.

11 no I'm alright Ann, I mean, feel a bit ba ah I mean I'm sorry I do have to buy a **feel** a bit of, I feel a bit dizzy you know as if I

12 Mick said to me last night, he said to me you can never **fit** not used to it, but

13 Now if he doesn't get the economy right he's gon na end up with **egg** on his face and

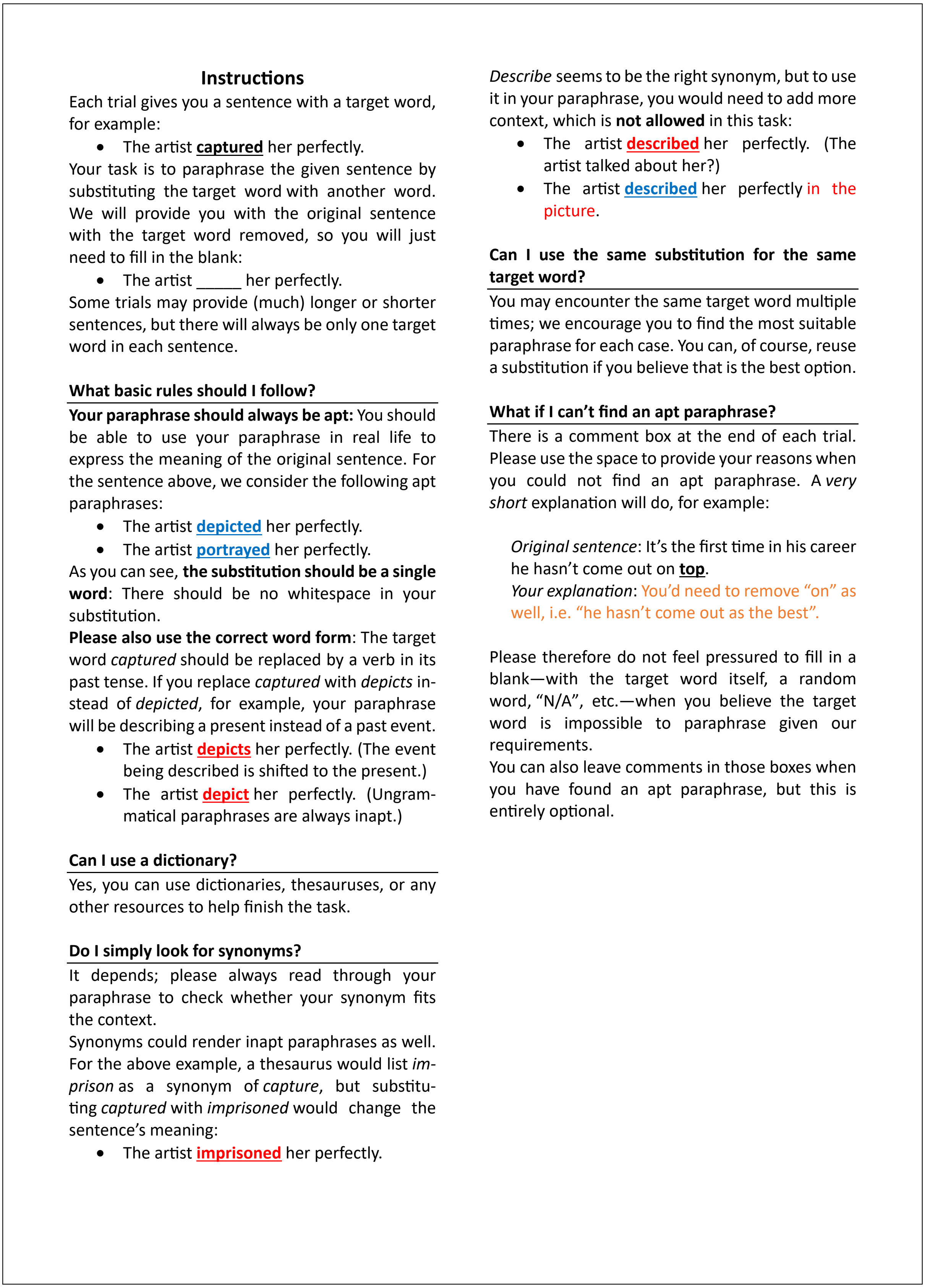
14 That **take** me nearly all the er

15 As this is been shared by **lines** int it?

16 Well seven nines, well ee er, it **takes** you so long

17 Take what you want and leave the rest, your mother'll **get** rid of it.

Table 5: Sentences that did not receive single-word substitutions in the crowdsourcing task.



Instructions

Each trial gives you a sentence with a target word, for example:

- The artist **captured** her perfectly.

Your task is to paraphrase the given sentence by substituting the target word with another word. We will provide you with the original sentence with the target word removed, so you will just need to fill in the blank:

- The artist _____ her perfectly.

Some trials may provide (much) longer or shorter sentences, but there will always be only one target word in each sentence.

What basic rules should I follow?

Your paraphrase should always be apt: You should be able to use your paraphrase in real life to express the meaning of the original sentence. For the sentence above, we consider the following apt paraphrases:

- The artist **depicted** her perfectly.
- The artist **portrayed** her perfectly.

As you can see, **the substitution should be a single word**: There should be no whitespace in your substitution.

Please also use the correct word form: The target word *captured* should be replaced by a verb in its past tense. If you replace *captured* with *depicts* instead of *depicted*, for example, your paraphrase will be describing a present instead of a past event.

- The artist **depicts** her perfectly. (The event being described is shifted to the present.)
- The artist **depict** her perfectly. (Ungrammatical paraphrases are always inapt.)

Can I use a dictionary?

Yes, you can use dictionaries, thesauruses, or any other resources to help finish the task.

Do I simply look for synonyms?

It depends; please always read through your paraphrase to check whether your synonym fits the context.

Synonyms could render inapt paraphrases as well. For the above example, a thesaurus would list *imprison* as a synonym of *capture*, but substituting *captured* with *imprisoned* would change the sentence's meaning:

- The artist **imprisoned** her perfectly.

Describe seems to be the right synonym, but to use it in your paraphrase, you would need to add more context, which is **not allowed** in this task:

- The artist **described** her perfectly. (The artist talked about her?)
- The artist **described** her perfectly in the picture.

Can I use the same substitution for the same target word?

You may encounter the same target word multiple times; we encourage you to find the most suitable paraphrase for each case. You can, of course, reuse a substitution if you believe that is the best option.

What if I can't find an apt paraphrase?

There is a comment box at the end of each trial. Please use the space to provide your reasons when you could not find an apt paraphrase. A *very short* explanation will do, for example:

> *Original sentence*: It's the first time in his career he hasn't come out on **top**.
> *Your explanation*: You'd need to remove "on" as well, i.e. "he hasn't come out as the best".

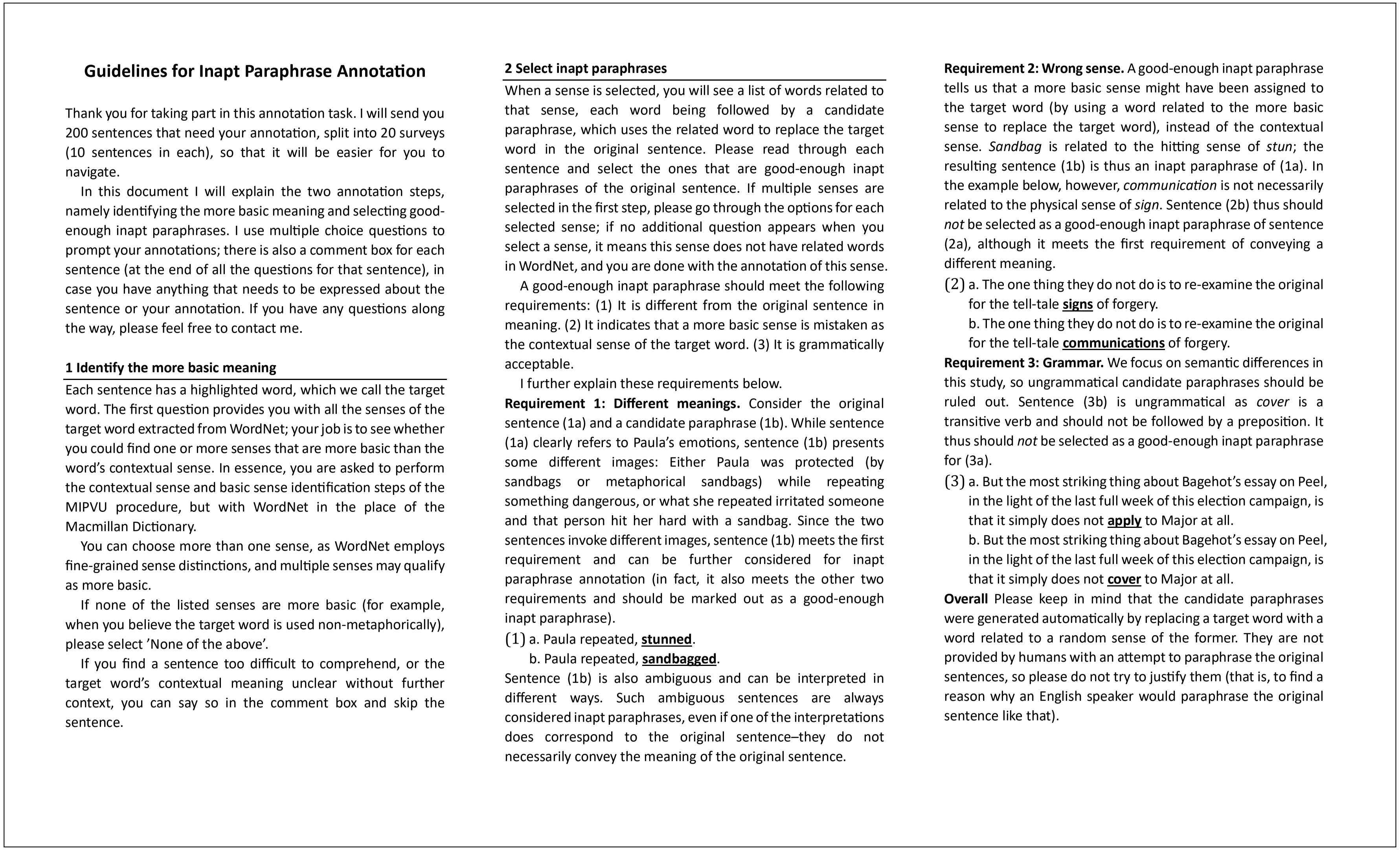
Please therefore do not feel pressured to fill in a blank—with the target word itself, a random word, "N/A", etc.—when you believe the target word is impossible to paraphrase given our requirements.

You can also leave comments in those boxes when you have found an apt paraphrase, but this is entirely optional.

Figure 6: Instructions for the paraphrasing task.

E Model evaluation details

We accessed the LLaMA models through Hugging Face; the queries used ∼880 GPU hours. Our GPT-



Guidelines for Inapt Paraphrase Annotation

Thank you for taking part in this annotation task. I will send you 200 sentences that need your annotation, split into 20 surveys (10 sentences in each), so that it will be easier for you to navigate.

In this document I will explain the two annotation steps, namely identifying the more basic meaning and selecting good-enough inapt paraphrases. I use multiple choice questions to prompt your annotations; there is also a comment box for each sentence (at the end of all the questions for that sentence), in case you have anything that needs to be expressed about the sentence or your annotation. If you have any questions along the way, please feel free to contact me.

1 Identify the more basic meaning

Each sentence has a highlighted word, which we call the target word. The first question provides you with all the senses of the target word extracted from WordNet; your job is to see whether you could find one or more senses that are more basic than the word's contextual sense. In essence, you are asked to perform the contextual sense and basic sense identification steps of the MIPVU procedure, but with WordNet in the place of the Macmillan Dictionary.

You can choose more than one sense, as WordNet employs fine-grained sense distinctions, and multiple senses may qualify as more basic.

If none of the listed senses are more basic (for example, when you believe the target word is used non-metaphorically), please select 'None of the above'.

If you find a sentence too difficult to comprehend, or the target word's contextual meaning unclear without further context, you can say so in the comment box and skip the sentence.

2 Select inapt paraphrases

When a sense is selected, you will see a list of words related to that sense, each word being followed by a candidate paraphrase, which uses the related word to replace the target word in the original sentence. Please read through each sentence and select the ones that are good-enough inapt paraphrases of the original sentence. If multiple senses are selected in the first step, please go through the options for each selected sense; if no additional question appears when you select a sense, it means this sense does not have related words in WordNet, and you are done with the annotation of this sense.

A good-enough inapt paraphrase should meet the following requirements: (1) It is different from the original sentence in meaning. (2) It indicates that a more basic sense is mistaken as the contextual sense of the target word. (3) It is grammatically acceptable.

I further explain these requirements below.

Requirement 1: Different meanings. Consider the original sentence (1a) and a candidate paraphrase (1b). While sentence (1a) clearly refers to Paula's emotions, sentence (1b) presents some different images: Either Paula was protected (by sandbags or metaphorical sandbags) while repeating something dangerous, or what she repeated irritated someone and that person hit her hard with a sandbag. Since the two sentences invoke different images, sentence (1b) meets the first requirement and can be further considered for inapt paraphrase annotation (in fact, it also meets the other two requirements and should be marked out as a good-enough inapt paraphrase).

(1) a. Paula repeated, **stunned**.
b. Paula repeated, **sandbagged**.

Sentence (1b) is also ambiguous and can be interpreted in different ways. Such ambiguous sentences are always considered inapt paraphrases, even if one of the interpretations does correspond to the original sentence–they do not necessarily convey the meaning of the original sentence.

Requirement 2: Wrong sense. A good-enough inapt paraphrase tells us that a more basic sense might have been assigned to the target word (by using a word related to the more basic sense to replace the target word), instead of the contextual sense. *Sandbag* is related to the hitting sense of *stun*; the resulting sentence (1b) is thus an inapt paraphrase of (1a). In the example below, however, *communication* is not necessarily related to the physical sense of *sign*. Sentence (2b) thus should *not* be selected as a good-enough inapt paraphrase of sentence (2a), although it meets the first requirement of conveying a different meaning.

(2) a. The one thing they do not do is to re-examine the original for the tell-tale **signs** of forgery.
b. The one thing they do not do is to re-examine the original for the tell-tale **communications** of forgery.

Requirement 3: Grammar. We focus on semantic differences in this study, so ungrammatical candidate paraphrases should be ruled out. Sentence (3b) is ungrammatical as *cover* is a transitive verb and should not be followed by a preposition. It thus should *not* be selected as a good-enough inapt paraphrase for (3a).

(3) a. But the most striking thing about Bagehot's essay on Peel, in the light of the last full week of this election campaign, is that it simply does not **apply** to Major at all.
b. But the most striking thing about Bagehot's essay on Peel, in the light of the last full week of this election campaign, is that it simply does not **cover** to Major at all.

Overall Please keep in mind that the candidate paraphrases were generated automatically by replacing a target word with a word related to a random sense of the former. They are not provided by humans with an attempt to paraphrase the original sentences, so please do not try to justify them (that is, to find a reason why an English speaker would paraphrase the original sentence like that).

Figure 7: Guidelines for inapt paraphrase annotation.

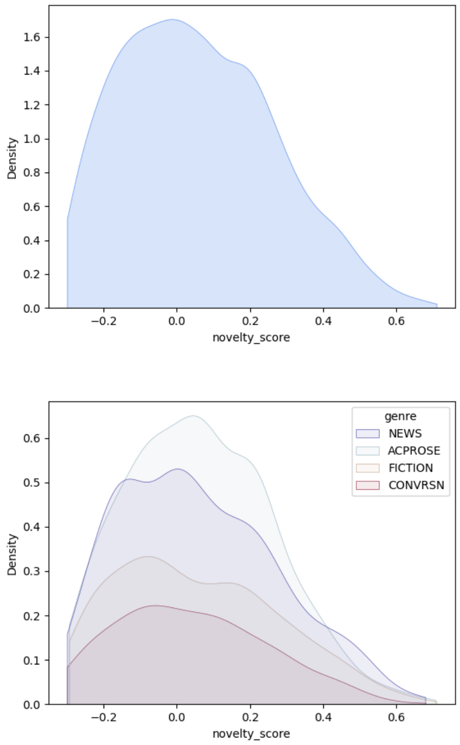


Figure 8: Novelty distribution of the metaphor samples, across all genres (above) and in different genres (below). The novelty scores are extracted from Do Dinh et al. (2018).

3.5 queries through the OpenAI API cost ∼255 USD.

We provide all the prompts used in this study: three prompts for each condition of the paraphrase judgement tasks, including word judgement (Table 6) and sentence judgement (Table 7); and three prompts for the paraphrase generation task (Table 8).

F Error analysis details

F.1 Paraphrase judgement

Table 9 shows the number of each expectation-vs-prediction combination for each model when it achieved the highest accuracy score across all conditions and prompts: The *Metaphor-Word* condition of *Word-judgement* for LLaMA-13B, using the third prompt (see Table 6); the *Metaphot-Word* condition of *Sentence-judgement* for LLaMA-30B, using the second prompt (see Table 7); the *Implicit* condition of *Word-judgement* for GPT-3.5, using the third prompt.

Implicit
(1) Choose the word(s) that can replace the highlighted word in the given sentence without changing the meaning of the sentence. Sentence: (a metaphor sample) Option A: (a substitution word) Option B: (another substitution word) Option C: Both Option A and Option B Option D: Neither Option A nor Option B Correct answer: Option
(2) Select words that can replace the highlighted word in the given sentence without altering the sentence's meaning. (. . .)
(3) Which of the given options can replace the highlighted word in the given sentence without altering the sentence's meaning? (. . .)
M-sent
(1) Choose the word(s) that can replace the highlighted word in the given metaphorical sentence without changing the meaning of the sentence. (. . .)
(2) Select words that can replace the highlighted word in the given metaphorical sentence without altering the sentence's meaning. (. . .)
(3) Which of the given options can replace the highlighted word in the given metaphorical sentence without altering the sentence's meaning? (. . .)
M-word
(1) Choose the word(s) that can replace the highlighted metaphorically used word in the given sentence without changing the meaning of the sentence. (. . .)
(2) Select words that can replace the highlighted metaphorically used word in the given sentence without altering the sentence's meaning. (. . .)
(3) Which of the given options can replace the highlighted metaphorically used word in the given sentence without altering the sentence's meaning? (. . .)

Table 6: Prompts for the word judgement task.

F.2 Factors associated with model performance

Table 10 summarises the novelty scores of the metaphor samples that receive correct versus incorrect answers from the models in the two paraphrase tasks. Table 11 and 12 show model accuracies in different genres and for different POS of the metaphorical word respectively. Like in F.1, the statistics are based on the best performance of each model. In paraphrase generation, the LLaMA models achieve their respective best performance when given the first prompt (see Table 8); for GPT-3.5, it is the second prompt.

Implicit
(1) Choose the correct paraphrase(s) for the given sentence. Sentence: (a metaphor sample) Option A: (a candidate paraphrase) Option B: (another candidate paraphrase) Option C: Both Option A and Option B Option D: Neither Option A nor Option B Correct answer: Option
(2) Select sentences that paraphrase the given sentence. (. . .)
(3) Select sentences that are semantically equivalent to the following sentence. (. . .)
M-sent
(1) Choose the correct paraphrase(s) for the given metaphorical sentence. (. . .)
(2) Select sentences that paraphrase the given metaphorical sentence. (. . .)
(3) Select sentences that are semantically equivalent to the following metaphorical sentence. (. . .)
M-word
(1) You are given a sentence where the highlighted word is metaphorically used. Choose the correct paraphrase(s) for the given sentence. (. . .)
(2) Given a sentence where the highlighted word is metaphorically used, select sentences that paraphrase this sentence. (. . .)
(3) Given a sentence where the highlighted word is metaphorically used, select sentences that are semantically equivalent to this sentence. (. . .)

Table 7: Prompts for the sentence judgement task.

(1) Paraphrase the given sentence by substituting the highlighted word with another word. The substitution should be a single word. Sentence: No golden light *bathed* the red brick of the house. llama: Paraphrase: No golden light *[blank]* the red brick of the house. [blank] should be “___ gpt: Paraphrase: No golden light *___* the red brick of the house.
(2) Use a single word to replace the highlighted word in the given sentence, so that the new sentence and the given sentence mean the same thing. Sentence: (. . .) New sentence: (. . .)
(3) Given a sentence with a highlighted word, replace this word with a different word to make a paraphrase. Sentence: (. . .) Paraphrase: (. . .)

Table 8: Prompts for the paraphrase generation task. The blank (___) denotes the place where models are asked to provide their answers: The LLaMA models append answer after the left quotation mark (“) while GPT-3.5 inserts answer between the two asterisks (*). The blank itself is not part of the prompts.

	llama-13b	llama-30b	gpt-3.5
A/B			
A/B	667	462	373
B/A	212	47	20
C	193	563	641
D	0	0	38
C			
C	10	32	38
A/B	35	13	5
D	0	0	2
D			
D	0	0	101
A/B	241	33	52
C	134	342	222

Table 9: Count for each combination of expected answer and correct or incorrect prediction when each model achieves their highest performance in the paraphrase judgement task. A/B (A or B) means one of the two candidate paraphrases is expected or predicted as the correct answer. The counts are based on the predictions of the models when they reach their respective highest accuracy in our experiments.

	Judgement	Generation
llama-13b	0.07 / 0.07	0.07 / 0.06
llama-30b	**0.05 / 0.08**	0.06 / 0.06
gpt-3.5	0.06 / 0.08	**0.04 / 0.07**

Table 10: Mean novelty scores of metaphor samples that each model gives correct/incorrect answers when it achieves its respective highest performance in the paraphrase judgement and paraphrase generation tasks. All standard deviations are 0.20 ± 0.01. Boldface denotes that the difference between correct and incorrect answers is **statistically significant**.

	ACPROSE	NEWS	FICTION	CONVRSN
Judgement				
llama-13b	.44	.47	.47	-
llama-30b	**.37**	.33	**.24**	-
gpt-3.5	.34	.36	.32	-
Generation				
llama-13b	.15	.17	**.21**	**.13**
llama-30b	.34	.37	**.37**	**.32**
gpt-3.5	**.45**	.41	**.40**	**.40**

Table 11: Model accuracy in different genres when the models achieve their best performance in the paraphrase judgement and paraphrase generation tasks. The metaphor samples for the paraphrase judgement task do not cover the conversation genre. Boldface denotes **statistically significant** difference between the highest and lowest accuracies on the same row.

	N	V	A	R
Judgement				
`llama-13b`	.44	.47	.46	.70
`llama-30b`	.34	.32	.30	.30
`gpt-3.5`	**.38**	**.29**	.31	.30
Generation				
`llama-13b`	.18	.16	.15	.13
`llama-30b`	**.37**	.36	**.32**	.37
`gpt-3.5`	.44	**.41**	**.40**	**.52**

Table 12: Model accuracy per POS of the metaphorical word (**N**oun, **V**erb, **A**djective, and adve**R**b) when each model achieves its best performance in the paraphrase judgement and paraphrase generation tasks. Boldface denotes **statistically significant** difference between the highest and lowest accuracies on the same row. Accuracies for adverb metaphors in the paraphrase judgement task are disregarded as the task only includes 10 adverb samples.