LMs stand their Ground: Investigating the Effect of Embodiment in Figurative Language Interpretation by Language Models

Philipp Wicke
Ludwig-Maximilians-University (LMU)
Institute for Information and Language Processing (CIS)
Munich Center for Machine Learning (MCML)
pwicke@cis.lmu-munich.de

Abstract

Figurative language is a challenge for language models since its interpretation is based on the use of words in a way that deviates from their conventional order and meaning. Yet, humans can easily understand and interpret metaphors, similes or idioms as they can be derived from embodied metaphors. Language is a proxy for embodiment and if a metaphor is conventional and lexicalised, it becomes easier for a system without a body to make sense of embodied concepts. Yet, the intricate relation between embodiment and features such as concreteness or age of acquisition has not been studied in the context of figurative language interpretation concerning language models. Hence, the presented study shows how larger language models perform better at interpreting metaphoric sentences when the action of the metaphorical sentence is more embodied. The analysis rules out multicollinearity with other features (e.g. word length or concreteness) and provides initial evidence that larger language models conceptualise embodied concepts to a degree that facilitates figurative language understanding.

1 Introduction

Infants acquire their first conceptual building blocks by observation and manipulation in the physical world. These primary building blocks enable them to make sense of their perceptions (Mandler and Cánovas, 2014). In return, their embodiment defines the capabilities with which they can explore and understand the world. The early conceptual system is built from spatial schemas, which enables early word understanding (Mandler, 1992). These so-called *Image Schemas* are recurring cognitive structures shaped by physical interaction with the environment. They emerge from bodily experience and motivate subsequent conceptual metaphor mappings (Johnson, 2013). The metaphorical mapping is visible in our everyday language whenever we use figurative language. For example, if we say that *she dances like a turtle*, that is to say, that *she dances poorly*. The metaphor in this phrase is readily interpreted by humans, who would favour the interpretation of *dances poorly* over *dances well*. The *turtle dance* example employs a conceptual mapping in which the *turtle* provides the source domain for the attributes *slow* and *rigid*, which turn the *dance* target domain into *poorly dancing*. This mapping draws from the human, bodily experience of dancing and therefore enables interpretation.

For a language model (LM), the understanding of figurative language is a great challenge (Liu et al., 2022). By nature of their digital implementation as computer algorithms, LMs are non-embodied and do not ground their conceptualisation by physical interaction with the environment. Instead, LMs learn statistical features of language by deep learning vast amounts of data (Vaswani et al., 2017). Whether these learned statistical features allow LMs to mirror or copy natural language understanding (NLU) is subject to discussion (Zhang et al., 2022a). Moreover, Tamari et al. (2020) suggest an *embodied* language understanding paradigm for LMs can benefit NLU systems through grounding by metaphoric inference.

One can argue that most embodied metaphors are heavily conventional (e.g. UP IS GOOD, DOWN IS BAD, KNOWLEDGE IS LIGHT, IGNORANCE IS DARKNESS) and as such, they are lexicalised in a language without an inherent need to understand their bodily basis. This lexicalisation should allow LMs to conceptualise and interpret them correctly and more robustly than less conventional metaphors. Conventionality relates to word frequency and age of acquisition (AoA), i.e. more frequent words and words that are acquired early in life are more conventional. We argue that embodiment has a measurable effect on the interpretation of metaphors that differs from the effect of other linguistic features. Moreover, we investigate whether an interpretation of figurative language with more

Findings of the Association for Computational Linguistics: ACL 2023, pages 4899–4913
July 9-14, 2023

embodied concepts is easier for LMs. Analogously, we investigate conflating factors such as the AoA, word frequency, concreteness and word length. The relation between embodiment and LMs' ability to interpret figurative language has not yet been investigated and is the key contribution of this research.

The following Section (2) starts with a review of language model abilities, more specifically figurative language interpretation abilities. The review identifies a suitable data set for our experiment and describes its formation in Section 3. We use a subset of the *Fig-QA* data set (Liu et al., 2022), a Winograd-style figurative language understanding task, and correlate the performance of various LMs concerning the degree of embodiment of the metaphorical actions that the LMs are tasked to interpret. In Section 4, we identify that models, that can reach a certain performance on our *Fig-QA* subset, shows a significant and positive correlation between the rating of the embodiment of the action involved in the metaphorical phrase and the model's ability to interpret the metaphor correctly. An in-depth analysis of additional features, such as the AoA, word length or frequency, does not indicate multicollinearity among those features. In Section 5 we conclude that the degree of embodiment of the action within the metaphoric phrase is a predictor of the LMs' ability to correctly interpret the figurative language. Lastly, we discuss the limitations and broader implications of the work.

2 Related Works

2.1 Language Model Abilities

The presented work investigates the zero-shot capabilities of LMs of different types and sizes. Arguably, LMs' capabilities to solve language-based tasks, which they have not been trained on, are an emerging property of their complexity and large-scale statistical representation of language. It is a property that makes them unsupervised multi-task learners (Radford et al., 2019; Brown et al., 2020). Despite task-agnostic pre-training and a task-agnostic architecture, LMs can perform various NLP tasks without seeing a single example of the task, albeit with mixed results (Srivastava et al., 2022). This raises the question of whether language models mirror the human conceptual understanding encoded in language or whether they "only" learn statistical features from the underlying training distribution, allowing them to generalise and convincingly solve previously unseen tasks.

Several works have tried to assess to what extent LMs are capable to perform more complex NLP tasks (e.g. logical reasoning or metaphoric inference). For example, Zhang et al. (2022a) investigate the logical reasoning capabilities of BERT (Devlin et al., 2019). For this, the authors define a simplistic problem space for logical reasoning and show that BERT learns statistical features from its training distribution, but fails to generalise when presented with other distributions and drops in performance. According to the authors, this implies that BERT does not emulate a correct reasoning function in the same way that humans would conceptualise the problem. Similarly, Sanyal et al. (2022) evaluate whether the RoBERTa model (Liu et al., 2019) or the T5 model (Raffel et al., 2020) can perform logical reasoning by understanding implicit logical semantics. The authors test the models on various logical reasoning data sets whilst introducing minimal logical edits to their rule base. Consequently, Sanyal et al. (2022) show that LMs, even when fine-tuned on logical reasoning, do not sufficiently learn the semantics of some logical operators. Han et al. (2022) present a diverse data set for reasoning in natural language. An evaluation of the GPT-3 model (Brown et al., 2020) on their data set shows a performance that is only slightly better than random. This indicates that there is a fundamental gap between human reasoning and LM reasoning and their conceptualisation capabilities. Yet, language models have demonstrated emergent abilities (Wei et al., 2022), encompassing enhanced skills and capabilities that are absent in smaller language models. Such abilities cannot be accurately predicted by extrapolating the performance of smaller models. Consequently, investigating the influence of model size on different tasks becomes imperative in comprehending the potentials and constraints of smaller and large language models.

The related works show that, although LMs seem to mirror an aspect of reasoning, e.g. logical reasoning, a closer look at the underlying conceptualisation of these abilities can reveal they are not robust and fail to mirror deeper semantics. Both logical reasoning and figurative language interpretation require an understanding of relationships between words and concepts and the ability to make inferences based on that understanding. This overlap in cognitive processes allows for the development of models that can perform both tasks effectively.

2.2 Figurative Language Interpretation

Liu et al. (2022) are among the first to quantitatively assess the ability of LMs to interpret figurative language. Their *Fig-QA* data set is publicly available[1] and we discuss the construction of our subcorpus in more detail in Section 3.1. In short, the authors present crowdsourced creative metaphor phrases with two possible interpretations of various LMs and check for which interpretation the model returns the higher probability distribution. The main contribution of Liu et al. (2022) is the *Fig-QA* task, which consists of 10,256 examples of human-written creative metaphors that are paired in a Winograd schema. The authors also contribute an assessment of various LMs in zero-shot, few-shot and fine-tuned settings on *Fig-QA*. Moreover, their results indicate that overall, LMs fall short of human performance. On a phrase and word level, the authors find that longer phrases are harder to interpret and that metaphors relying on commonsense knowledge concerning objects' volume, height, mass, brightness or colour are easier to interpret. This indicates that bodily modalities seem to facilitate interpretation success. They also show that larger models (i.e. number of parameters) perform better on the task. All of these findings have been reproduced by our experiments.

Chakrabarty et al. (2022) present FLUTE, a data set of 8,000 figurative NLI instances. Their data set includes the different figurative language categories of metaphor, simile, and sarcasm. In contrast to *Fig-QA*, the authors do not create metaphors in a Winograd scheme as a forced-choice task but create natural language explanations (NLE) using GPT-3 (Brown et al., 2020) and human validation. Their experiments with state-of-the-art NLE benchmark models show poor performance in comparison to human performance. The authors do not differentiate the metaphors, similes and sarcastic phrases concerning linguistic features. Moreover, they include a language model in the creation process, which, as far as our study is concerned, introduces a bias to the data set. Hence, we decide to use the *Fig-QA* data set instead of FLUTE.

2.3 Modelling Embodied Language

It is generally understood that language is grounded in experience based on interaction with the world (Bender and Koller, 2020; Bisk et al., 2020). Hence, there is an interest to leverage LMs' capabilities in interactions with the environment. For example, Suglia et al. (2021) present EmBERT, which attempts language-guided visual task completion. Their model uses a pre-trained BERT stack fused with an embedding for detecting objects from visual input. The model achieves competitive performance on ALFRED, a benchmark task for interpreting instructions (Shridhar et al., 2020).

Huang et al. (2022) investigate if LMs know enough embodied knowledge about the world to ground high-level tasks in the procedural planning of instructions for household tasks. For example, the authors pass a prompt, e.g. "*Step 1: Squeeze out a glob of lotion*" to a pre-trained LM (e.g. GPT-3) and extract actionable knowledge from its response. Their results indicate that large language models (<10B parameters) can produce plausible action plans for embodied agents.

Embodiment In this study, the term *embodiment* relates to cognitive sciences: Humans process a linguistic statement such as "*to grab an apple*" using embodied simulations in the brain. Perceptual experiences activate cortical regions that are dedicated to sensory actions and those regions partially reactivate premotor areas to implement, what Barsalou (1999) calls, *perceptual symbols*. Reading of actions words such as *kick* or *lick* is associated with premotor cortex activation responsible for controlling movements for these actions (Hauk et al., 2004). This effect is diminished by figurative language (Schuil et al., 2013). Therefore, a statement such as "*to grasp the idea*" does not necessarily rely on premotor cortex simulation. The semantic processing of the linguistic statement is therefore linked to its context and degree of embodiment in the sense that the action can be simulated by a brain in a body (Zwaan, 2014). This understanding of the term *embodiment* guides the evaluation of how language models, which do not have a brain in a body, can interpret figurative language phrases with a varying degree of embodied actions.

3 Statistical Evaluation

The review of related works shows that there are abilities of LMs that go beyond mere language generation, e.g. logical reasoning, and action planning. It is unclear how LMs conceptualise actions that humans conceptualise using interaction with the environment. Figurative language acts as a test bed to assess metaphorical conceptualisations since they

[1] https://huggingface.co/datasets/nightingal3/fig-qa

are grounded in embodied experience and interaction with the environment. We take Liu et al. (2022)'s findings as a starting point to focus on the effect of embodiment in figurative language interpretation by langauge models of various sizes.

3.1 Experimental Framework

Embodiment Rating and Data Set To assess the effect of embodiment on the task, we discuss the effects of embodiment in semantic processing and introduce the simplification underlying our study through an example. The *Fig-QA* provides the following item:

(A) *The pants were as faded as ...*

(A.1) ... *the memory of pogs*

(A.2) ... *the sun in June*

with the possible interpretations:

(A.I) *they were very faded*

(A.II) *they were bright*

The LM is prompted with each combination of sentence completion and interpretation (i.e. A.1+A.I, A.1+A.II, A.2+A.I, A.2+A.II). Notably, Liu et al. (2022) have shown that the addition of "*that is to say*" as a concatenation between metaphorical phrase and interpretation phrase elicits better model performance, hence we also include this prompt in our studies. Subsequently, the prediction scores of the language modelling head (scores for each vocabulary token) are retrieved and the highest probability becomes the LMs choice of interpretation (for more details, see (Liu et al., 2022)). We compare this example with a different *Fig-QA* item:

(B) *She dances like a ...*

(B.1) ... *fairy*

(B.2) ... *turtle*

with the possible interpretations:

(B.I) *she dances well*

(B.II) *she dances poorly*

Given our hypothesis that embodiment affects the LMs' ability to interpret these phrases, we score (A) and (B) concerning the embodiment. As a simplification, we limit the rating of embodiment to the actions within the phrase. Every phrase evaluated has at least one word with a score related to an action. Most of the time, these related actions are verbs. Thus, we rate *faded* for (A) and *dances* for (B) with respect to their relative embodiment. For this scoring, we consult data by Sidhu et al. (2014).

In their empirical study, Sidhu et al. (2014) characterise a dimension of a relative embodiment for verbs. In the construction of the data set, "*participants were asked to judge the degree to which the meaning of each verb involved the human body, on a 1–7 scale*" (Sidhu et al., 2014). Their resulting data set consists of ratings for 687 English verbs. Our hypothesis is that embodiment is a semantic component which affects the interpretation ability of LMs concerning figurative language. With their data set, the authors provide evidence that the meaning of a verb has a semantic component linked to the human body in the lexical processing of that verb. They assume that more robust semantic activation is generated by more embodied verbs (Sidhu et al., 2014). This provides us with data set we can apply to our experiment on figurative language. Moreover, their experiment provides additional control variables such as the AoA and word length, which have a known effect on lexical processing (Colombo and Burani, 2002) and are included in our results (Sec. 4).

At the time of conducting our experiment, *Fig-QA* did only provide the *training* and *development* data, which we will refer to as train & dev. Hence, we identify all phrases from the train & dev data set that contain at least one word with an embodiment rating from Sidhu et al. (2014). The process of creating the subcorpus with embodiment ratings (C_{Emb}) begins by identifying verbs using *spaCy* (Honnibal et al., 2020). The lemmatized versions of the verbs for the metaphorical phrases are then matched with embodiment scores, resulting in a subcorpus (C_{Emb}) with 1,438 entries. If more than one verb is present in the metaphorical sentence, the average is assigned. We note that, future work will assess whether a different heuristic for treating multiple actions influences our results. Analogously, we construct a subcorpus of the same size with metaphorical phrases that do not contain an embodied verb (C_{NoE}). For both subcorpora, we only keep phrases in which the verb is contained in the Winograd pair. The resulting subcorpora statistics are listed in Table 1 and further examples from the subcorpus are presented in the Appendix in Section A. The previous examples (A) and (B) are thus augmented as follows:

(A) *The pants were as faded as ...*

Embodiment Rating: 2.36

(B) *She dances like a ...*

Embodiment Score: 6.50

With the annotated *Fig-QA* subcorpus C_{Emb} we now turn to the models we select to assess whether there is a correlation between embodiment score and LM task performance.

Hypotheses The main hypothesis for the statistical evaluation can be summarized as follows:

1. There is a correlation between the LMs' interpretation capabilities of metaphors and the amount of embodiment of the verbs within those metaphorical phrases.

Intuitively, more embodied actions such as *kick*, *move* or *eat* are much more concrete, shorter and basic, when compared to *resonate*, *compartmentalise* or *misrepresent*. Therefore, the analysis of embodied actions must take into account factors such as concreteness, AoA, word length and word frequency. Moreover, common metaphors are conventional and more lexicalised. Consequently, they might simply be more embodied and the effect of embodied verbs might stem from the fact that these verbs are more concrete in the context that they are presented. Hence, the first hypothesis should not stand alone, but will be evaluated along with two additional null hypotheses:

1.I There is no correlation between the LMs' interpretation capabilities of metaphors and the amount of concreteness of the verbs within those metaphorical phrases irrespective of their embodiment rating.

In our evaluation, the concreteness of a word in its context will be scored using an open-source predictor[2] based on distributional models and behavioural norms explained in (Rotaru, 2020). Details of the concreteness scoring with the predictor have been summarized in Sec. 3.2. Concreteness ratings are often subjective ratings (Brysbaert et al., 2014) or determined by other low-level features, such as AoA, word frequency and word length (Rotaru, 2020). To isolate the effect of embodiment, we add the second null hypothesis:

1.II There is no correlation between the LMs' interpretation capabilities of metaphors and other linguistic features, such as AoA, word frequency and word length.

For AoA we obtain scores for each of the actions from (Kuperman et al., 2012) and for word frequency from (Van Heuven et al., 2014). Together with word length and embodiment score we test for variance inflation to respond to 1.II.

Model Selection The selection of our models is based on three criteria: First, we want to reproduce the results by (Liu et al., 2022) having a comparable measure. Second, we want to check whether the effect generalises to other large LMs. Third, we want a variation of different model sizes to account for varying performance on the task as a result of model size. For the latter two criteria, we start with the smallest available models of each type and check intermediate model sizes. We do not consider it necessary to check whether or not scaled, largest versions of each model perform better on the task since this is a general property of LMs (Brown et al., 2020; Srivastava et al., 2022).

In the original *Fig-QA* study, the authors examine three transformer-based LMs with different parameter sizes: GPT-2 (Radford et al., 2019), GPT-3 (Brown et al., 2020) and GPT-Neo (Black et al., 2021). To reproduce the results by Liu et al. (2022), we include GPT-2, GPT-3, GPT-Neo LMs and add OPT LMs (Zhang et al., 2022b). An overview of the models and their specifications is shown in Table 2. Notably, we want to correlate whether the type and number of parameters play a role when it comes to performance concerning the embodiment. Hence, we include pairs of models from each type that are small (<1 billion) and medium to large (>1 billion) in their number of parameters.

3.2 Methodology

We apply the same methodology of evaluation as Liu et al. (2022). In our zero-shot setting, each pretrained LM is prompted with the metaphor sentences combined with one of the interpretation sentences, concatenated with *that is to say*. For *OpenAI* models, the API provides the log probabilities per token as `logprob` return value. We access all other models using `huggingface.co` and its `transformer` library. To create the same evaluation metric as for results by (Liu et al., 2022), we follow (Tunstall et al., 2022) and implement a function that returns the `logprob` based on the prediction scores of the language modelling head. All code and data is publicly available[3].

[2] https://github.com/armandrotaru/TeamAndi-CONcreTEXT

[3] https://osf.io/puhxb/?view_only=15933a2da0a14f07834ba1d479ce9c43

Label	Source	Description	Number of entries
C_{Liu}	Fig-QA test	All phrases from the Fig-QA test set	1,146
C_{Emb}	Fig-QA train/dev	Phrases that have at least one action with embodiment rating	1,438
C_{NoE}	Fig-QA train/dev	Phrases that do not have an action with embodiment rating	1,438

Table 1: C_{Liu} is 100% of the *Fig-QA* test set and 11% of the entire *Fig-QA* data set (Liu et al., 2022). Our selected subsets C_{Emb} and C_{NoE} are mutually exclusive and each composes 14% of the entire *Fig-QA* data set.

Label	#Parameters in Millions	Provider
GPT-3 (small)	∼350	OpenAI
GPT-3 (large)	∼175,000	OpenAI
OPT (small)	350	Facebook
OPT (medium)	13,000	Facebook
GPT-Neo (small)	125	EleutherAI
GPT-NeoX (medium)	20,000	EleutherAI
GPT-2 (small)	355	OpenAI
GPT-2 XL (medium)	1,500	OpenAI

Table 2: Different model types and parameter numbers have been selected for the evaluation. For each type, we have selected a pair of smaller and medium to large model version.

Reproduction and Suffix Prompting In an initial experiment, we reproduce the same experiment by Liu et al. (2022), but instead of performing the zero-shot classification on the test set, we evaluate the performance on C_{Emb} and C_{NoE} with and without suffix prompting (*that is to say*). This allows us to compare our data set against the baseline. The result indicates that GPT-3 models perform slightly worse on our subcorpora, but all conditions benefit from the suffix prompting (see Appendix C). Since both C_{Emb} and C_{NoE} are more difficult for GPT-3, we can rule out that this effect stems solely from the embodiment component present in C_{Emb}, which is not present in C_{NoE}. Moreover, we adopt the suffix prompting for all further experiments.

Concreteness Scoring To determine the concreteness of a verb in context, (Rotaru, 2020) built a predictor based on a combination of distributional models, together with behavioural norms. We adopt the same settings and model choice as presented by the author, but exclude the word frequency behavioural norm, as we investigate it as a separate feature. We evaluate our predictor on the same English test set of the *CONcreTEXT* task at *EVALITA2020* (Gregori et al., 2020) and receive a mean Spearman correlation of 0.87, which is in line with (Rotaru, 2020). The context-dependent models used in the predictor include ALBERT (Lan et al., 2020), BERT and GPT-2.

Statistical Tests For each model, we obtain its performance on the data set with a binary scoring of each figurative phrase as being correctly or incorrectly identified. We correlate this series of binary values with the continuous variable of embodiment ratings by calculating the point biserial correlation coefficient and the associated p-value. Moreover, we assess various other language features to isolate any effect of embodiment. As described in previous sections and based on the work by Liu et al. (2022); Sidhu et al. (2014); Colombo and Burani (2002), we test for the effects of word concreteness, AoA, word frequency and word length. This analysis includes an assessment of the amount of multicollinearity within the regression variables by determination of the variance inflation factor (VIF). Moreover, we conduct linear regressions for all models (and all sizes) with respect to their task performance and the features: embodiment score, AoA, word frequency and word length. For these linear regressions, we include those with and without the embodiment score feature in order to assess whether this feature contributes to a higher coefficient of determination (R^2).

4 Results

Embodiment Correlation The results of all models are listed in Table 3 and visualized in Figure 1. Overall, for each pair of small and larger models, the larger models always perform better on the interpretation task than the smaller version of the model. Moreover, *all* larger versions of the models show a significant correlation ($p < 0.05$) between the embodiment rating and task performance. In two instances, GPT-NeoX (20B) and GPT-2 (1.5B),

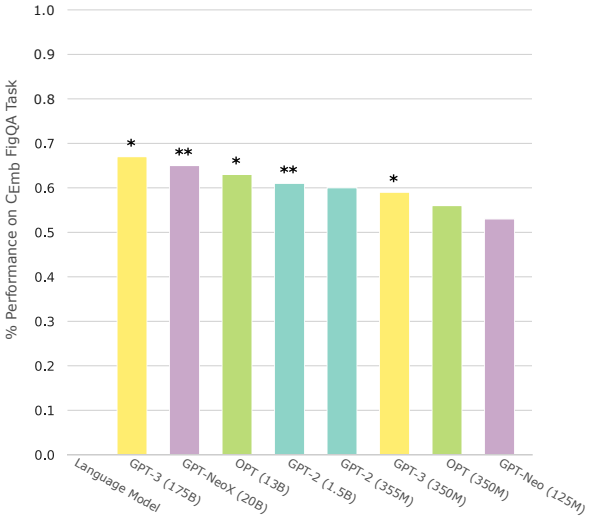

Figure 1: Performance of four language models in two size-variations on C_{Emb}. Significant results of the point biserial correlation between embodiment score and model performance are marked for $p < 0.05$ with * and for $p < 0.01$ with **. Colours correspond to the same model type, and x-labels provide model size.

the p value is < 0.01. In the case of GPT-3, both model variants show a significant correlation. In all correlations, the coefficient is positive, albeit small (<0.1), which indicates that embodiment has a positive effect on task performance. All smaller models (except for GPT-3 with 350M parameters) do not show a significant correlation between embodiment score and task performance.

Concreteness Using the concreteness-in-context predictor, we provide a concreteness value for each verb in C_{Emb} and correlate those predictions with all models' performance. As a result, there is no significant correlation between the concreteness of the action word in its context and the performance of the LM on the interpretation (results in Appendix B). We do not reject hypothesis 1.I.

Regression Analysis The linear regressions for all models and model sizes, both with and without the feature of embodiment score, revealed that the coefficient of determination (R^2) was consistently higher for regressions that included the embodiment score feature. Furthermore, for cases without the embodiment feature, none of the other variables, such as Age of Acquisition (AoA), word frequency, or word length, showed a significant correlation with task performance. Related results and figures are available in Section D of the Appendix.

Model	Accuracy on C_{Emb}	p-Value	Correlation coefficient	* p <.05 ** p <.01
GPT-3 (small)	0.594	0.018	0.062	*
GPT-3 (large)	0.667	0.034	0.056	*
OPT (small)	0.561	0.206	0.033	
OPT (medium)	0.627	0.034	0.056	*
GPT-Neo (small)	0.535	0.399	0.022	
GPT-NeoX (medium)	0.648	0.005	0.073	**
GPT-2 (small)	0.597	0.158	0.037	
GPT-2 XL (medium)	0.606	0.009	0.069	**

Table 3: Experimental results of all model pairs (small and larger versions) on the C_{Emb} corpus. The last column marks significant results of the point biserial correlation between embodiment score and model performance for $p < 0.05$ with * and for $p < 0.01$ with **.

AoA	Word Frequency	Embodiment Rating	Word Length	Constant
1.610	1.345	1.326	1.017	86.387

Table 4: VIF from the four features. Constant denotes the intercept provided for the VIF. A factor close to 1 indicates no correlation with values above 4 regarded as moderate correlation.

Variance Inflation Pairwise correlations between AoA, word frequency, embodiment score and word length are visualized in Figure 2. Intuitively, frequency and AoA are expected to be correlated with each other, because words that are acquired much later in life are often less frequently used words, as they tend to be more complex or specific words. The multicollinearity test through VIF is presented in Table 4. All factors are close to 1.0, which indicates that there is no multicollinearity among predictor variables (if the VIF is between 5 and 10, multicollinearity is likely to present) (James et al., 2013). Given that there is no multicollinearity between embodiment score and other linguistic features, we do not reject hypothesis 1.II.

4.1 Interpretation

The results of the correlation analysis indicate that embodiment affects the LMs' ability to interpret figurative language when the LM achieves a certain level of performance, which depends on the size of the model. The correlation coefficient is positive in all significant cases and those significant correlations occur in all larger (>1B parameter) model ver-

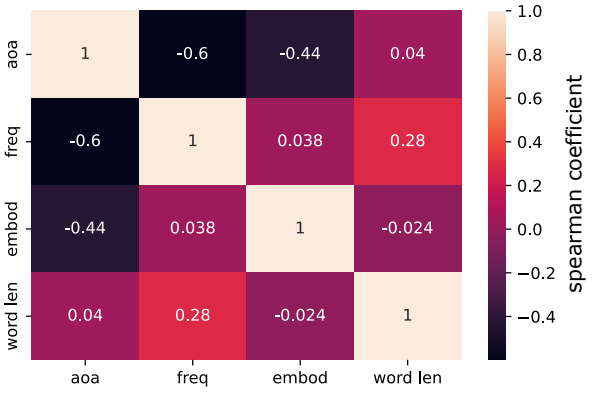

Figure 2: Pairwise correlation for the variables of AoA, word frequency, embodiment score and word length. Positive numbers indicate that increase in one variable correlates with an increase in the other and analogously for negative numbers with a decrease. Value 1 is the perfect correlation of the variable with itself. `aoa`: Age of acquisition, `freq`: Word frequency, `embod`: Embodiment score and `word len`: word length.

sions. Since task performance increases with model size, the effect of embodiment becomes more apparent through more successful interpretations in better-performing models. The fact that concreteness, AoA, word length and word frequency do not inflate this effect, shows that the embodiment rating is not an arbitrary construct that implicitly models another linguistic feature.

There are slight differences in the model types when it comes to the performance of the models. For example, GPT-3 shows a significant correlation for the effect of an embodiment for both, the small (350M) and large (175B) model sizes. Yet, this effect does not occur in the small GPT-2 (355M) model, but in the large GPT-2 (1.5B) version. Notably, *OpenAI* does not explicitly list the *Ada* model with 350M, but its performance ranks close with 350M versions on various tasks (Brown et al., 2020). Hence, this difference has only limited relevance. Nonetheless, we assume that the effect is correlating with model size and that a reliable effect can be seen in larger models with a parameter number of over 1 billion.

5 Conclusion

5.1 Contribution to the Field

We successfully reproduce results that are in line with (Liu et al., 2022). Moreover, we provide a subcorpus with ratings of an embodiment for the *Fig-QA task*. We identify the contribution of embodied verbs to LMs' ability to interpret figurative language. To the best of our knowledge, this study is the first to provide evidence that the psycholinguistic norm of the perceived embodiment has been investigated in an NLP task for LMs.

5.2 Discussion

Benchmarks, such as BIG-bench (Srivastava et al., 2022), show that different types and sizes of LMs can be evaluated on many different tasks to identify potential shortcomings or limitations. This paper takes an entirely different approach by zeroing in on a particular task, which has been augmented with a specific semantic evaluation (embodiment ratings of actions), to highlight how difficult tasks, such as figurative language interpretation, benefit not only from model size but from specific embodied semantics.

Figurative language is difficult for LMs because its interpretation is often not conveyed directly by the conventional meaning of its words. Human NLU is embodied and grounded by physical interaction with the environment (Di Paolo et al., 2018). Consequently, it could be expected that LMs struggle when the interpretation of figurative language depends on a more embodied action. Yet, the opposite has been shown as more embodied concepts are more lexicalised and larger LMs can interpret them better in figurative language. Hence, our study provides valuable insight that raises the question of whether this effect is limited to figurative language or translates to other NLU tasks for LMs.

5.3 Limitations and Future Works

The current results are limited to one specific figurative language task (*Fig-QA*). In future work, we aim to test whether our hypothesis holds for other figurative language interpretation tasks, such as those by (Chakrabarty et al., 2022; Stowe et al., 2022). Moreover, we want to assess BIG-bench (Srivastava et al., 2022) performances on various other tasks concerning embodiment scoring and see whether the bias can be detected in tasks other than figurative language interpretation.

The statistical evaluation has attempted to measure many different linguistic dimensions, e.g. AoA, word length, word frequency and concreteness in context. Empirically, this indicates that the effect of embodiment is not simply explainable by other factors. Theoretically, we argue that this correlation can be causally explained through the lexicalisation of conventional metaphors. We simplify conventionality by assuming that word

frequency and age of acquisition (AoA) are indicators of conventionality, i.e. more frequent words and words that are acquired early in life are more conventional. Nonetheless, a thorough explanation of the effect of embodiment on LMs' capabilities for language tasks requires many more studies.

Ethical Consideration It should be noted that a key component of the experiment is built from (Sidhu et al., 2014) with their ratings of relative embodiment. For their study, the authors have sampled data exclusively from (N=67, 57 female) "*graduate students at the University of Calgary who participated in exchange for bonus credit in a psychology course, had a normal or corrected-to-normal vision, and reported English proficiency*" (Sidhu et al., 2014). Even though *embodiment* is supposed to be a general, human experience, the pool of participants is relatively homogeneous (mostly female, educated and presumably able-bodied). A broader and more diverse set of ratings, specifically concerning differently-abled participants and cultural backgrounds should be targeted.

Computing Cost All model inferences (except *OpenAI*) have been conducted on University servers with 8x *NVIDIA RTX* A6000 (300 W). Each experiment for each model lasted at most 10min with full power consumption. A conservative estimate of 2,400 W (8 GPUs x 300 W) for 20 experiments results in a power consumption of at most 8 kWh, which equals emission of at most ∼3.5 kg CO_2 for all experiments with model inference.

Data and Artefact Usage Existing artefacts used in this research are attributed to their creators and their consent has been acquired before the studies. This concerns the *embodiment ratings* by Sidhu et al. (2014), the *Fig-QA* corpus by Liu et al. (2022) and the *concreteness predictor* by Rotaru (2020).

Acknowledgements

We would like to thank the Pexman Language Processing Lab (University of Calgary) for sharing the embodiment ratings. We would also like to thank Team Andi for publishing and providing their CONcreTEXT submission. Without model access and hosting by *Huggingface* and *OpenAI* the study would not have been possible. We would also like to thank Marianna Bolognesi and Stefan Riegl for their valuable feedback. Moreover, we thank the three ARR reviewers for their valuable feedback.

References

Lawrence W Barsalou. 1999. Perceptual symbol systems. *Behavioral and brain sciences*, 22(4):577–660.

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.

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 EMNLP*, pages 8718–8735. ACL.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46(3):904–911.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. Flute: Figurative language understanding and textual explanations. In *Proceedings of the 2022 Conference on EMNLP*.

Lucia Colombo and Cristina Burani. 2002. The influence of age of acquisition, root frequency, and context availability in processing nouns and verbs. *Brain and Language*, 81(1-3):398–411.

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 NAACL: Human Language Technologies*, page 4171–4186.

Ezequiel A Di Paolo, Elena Clare Cuffari, and Hanne De Jaegher. 2018. *Linguistic bodies: The continuity between life and language*. MIT press.

Lorenzo Gregori, Maria Montefinese, Daniele P Radicioni, Andrea Amelio Ravelli, and Rossella Varvara. 2020. Concretext@ evalita2020: The concreteness in context task. In *EVALITA proceedings*. CEUR.org.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Luke Benson, Lucy Sun, Ekaterina Zubova, Yujie Qiao, Matthew Burtell, et al. 2022. Folio: Natural language reasoning with first-order logic. *arXiv preprint arXiv:2209.00840*.

Olaf Hauk, Ingrid Johnsrude, and Friedemann Pulvermüller. 2004. Somatotopic representation of action words in human motor and premotor cortex. *Neuron*, 41(2):301–307.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. *arXiv preprint arXiv:2201.07207*.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An introduction to statistical learning*, volume 112. Springer.

Mark Johnson. 2013. *The body in the mind: The bodily basis of meaning, imagination, and reason*. University of Chicago press.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 english words. *Behavior research methods*, 44(4):978–990.

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 *Proceedings of ICRL 2020*.

Emmy Liu, Chen 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 NAACL*. ACL.

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. *arXiv preprint arXiv:1907.11692*.

Jean M Mandler. 1992. How to build a baby: Ii. conceptual primitives. *Psychological review*, 99(4):587.

Jean M Mandler and Cristóbal Pagán Cánovas. 2014. On defining image schemas. *Language and cognition*, 6(4):510–532.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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

Armand Stefan Rotaru. 2020. Andi@ concretext: Predicting concreteness in context for english and italian using distributional models and behavioural norms (short paper). In *EVALITA proceedings*. CEUR.org.

Soumya Sanyal, Zeyi Liao, and Xiang Ren. 2022. Robustlr: Evaluating robustness to logical perturbation in deductive reasoning. In *Proceedings of the 2022 Conference on EMNLP*.

Karen DI Schuil, Marion Smits, and Rolf A Zwaan. 2013. Sentential context modulates the involvement of the motor cortex in action language processing: An fmri study. *Frontiers in human neuroscience*, 7:100.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10740–10749.

David M Sidhu, Rachel Kwan, Penny M Pexman, and Paul D Siakaluk. 2014. Effects of relative embodiment in lexical and semantic processing of verbs. *Acta psychologica*, 149:32–39.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

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. ACL.

Alessandro Suglia, Qiaozi Gao, Jesse Thomason, Govind Thattai, and Gaurav Sukhatme. 2021. Embodied bert: A transformer model for embodied, language-guided visual task completion. In *Novel Ideas in Learning-to-Learn through Interaction at EMNLP2021*.

Ronen Tamari, Chen Shani, Tom Hope, Miriam RL Petruck, Omri Abend, and Dafna Shahaf. 2020. Language (re) modelling: Towards embodied language understanding. pages 6268–6281.

Lewis Tunstall, Leandro von Werra, and Thomas Wolf. 2022. *Natural language processing with transformers*. O'Reilly Media, Inc.

Walter JB Van Heuven, Pawel Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2014. Subtlex-uk: A new and improved word frequency database for british english. *Quarterly journal of experimental psychology*, 67(6):1176–1190.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Honghua Zhang, Liunian Harold Li, Tao Meng, Kai-Wei Chang, and Guy Van den Broeck. 2022a. On the paradox of learning to reason from data. *arXiv preprint arXiv:2205.11502*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022b. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Rolf A Zwaan. 2014. Embodiment and language comprehension: Reframing the discussion. *Trends in cognitive sciences*, 18(5):229–234.

A Appendix: Further samples form C_{Emb}

Further samples from C_{Emb} are provided in Table 5. The table depicts a selection of linguistic examples of more embodied and less embodied metaphors. Originally, these samples are form the *Fig-QA* data set by Liu et al. (2022), providing the entries for the columns *Figurative Phrase* and *Target Interpretation*. The *Embod. Score* is the embodiment rating of the action identified in the respective column and taken from Sidhu et al. (2014). A lower embodiment score indicates that the action has been rated as "less embodied". In Table 5, none of the language models shows a statistically significant correlation between the variables ($\alpha < 0.05$).

B Appendix: Concreteness Correlation

Table 6 shows the results of the the point biserial correlation between the concreteness of action in context and performance of LM on the interpretation of figurative language phrases (C_{Emb}). `pointbiserial` of the *scipy stats* package for *Python* has been used to determine the correlation. This function uses a t-test with n-1 degrees of freedom. The value of the point-biserial correlation has been calculated from:

$$r_{\text{pb}} = \frac{(\bar{Y_1} - \bar{Y_0})}{s_y} \sqrt{\frac{N_1 N_2}{N(N-1)}} \qquad (1)$$

With Y_0 and Y_1 as the means of the metric observations; N_1 and N_2 as the number of observations; N as the total number of observations and s_y as the standard deviation of all the metric observations.

C Appendix: Reproduction and Suffix Prompting

Table 7 presents the results of the zero-shot performance of the GPT-3 models *Ada* (∼350M parameters) and *Davinci* (∼175B parameters) on the different corpora (Table 1) with respect to the suffix *that is to say*. On all data sets (C_{Liu}, C_{Emb}, C_{NoE} and $C_{Emb} + C_{NoE}$) performance of both models is better if the suffix is provided.

D Appendix: Linear Regression Data

In addition to the VIF, we performed two linear regression analyses for each of the models (and sizes). These results are visualized in Figure 3. The first includes all features (embodiment score, Age of Acquisition, word frequency and word length). The second excludes the embodiment score and shows a lower coefficient of determination (R^2) for all regressions. Details of these linear regressions are exemplified in the results for *GPT3_350m* in Table 8 (with embodiment score feature) and in Table 9 (without embodiment score feature).

Figurative Phrase	Target Interpretation	Embod. Score	Action
The chihuahua believes it is a wolf	The small dog thinks it is undefeatable	2.83	believe
The chihuahua believes it is a lap blanket	The small dog always stays on your lap	2.83	believe
He knew her like a sister	He knew her very well.	3.00	know
He knew her like a stranger	He didn't know her well.	3.00	know
He guided her like a lighthouse.	He was a good guide.	3.61	guide
He guided her like a broken GPS.	He was a terrible guide.	3.61	guide
The argument appears as a crystal clear spring	The argument makes sense	3.90	appear
The argument appears as a muddy rut	The argument is senseless	3.90	appear
the movie raised your spirits to heaven	The movie was uplifting.	4.29	raise
the movie raised your spirits to the ocean floor	The movie was depressing.	4.29	raise
It was buried as deep as an oil well	It was buried deep	4.87	bury
It was buried as deep as a bathtub	It was not buried deep	4.87	bury
The reporter wrote like a monkey on crack	The reporter wrote badly	5.19	write
The reporter wrote like Hemingway	The reporter wrote well	5.19	write
He should be cooking for Gordon Ramsay	His cooking is good	5.47	cook
He should be cooking for McDonalds	His cooking is bad	5.47	cook
She sings like a nightingale	She sings beautifully	6.03	sing
She sings like an angry crow	She sings horribly	6.03	sing
The food tasted like eating a mother's love	The food tasted amazing	6.26	eat, taste
The food tasted like eating the bottom of a shoe	The food tasted disgusting	6.26	eat, taste
He could sprint like the wind	He was fast.	6.46	sprint
He could sprint like a tortoise	He was slow.	6.46	sprint

Table 5: Linguistic examples of more embodied and less embodied metaphors, sampled from C_{Emb} (derived from Liu et al. (2022)). Every pair of sentences is presented with both target sentences to each model. Embodiment scores retrieved from Sidhu et al. (2014).

Model	**GPT-3**	**GPT-3**	**OPT**	**OPT**	**GPT-Neo**	**GPT-NeoX**	**GPT-2**	**GPT-2 XL**
Parameters	350M	175B	350M	13B	125M	20B	355M	1.5B
Correlation	-0.016	-0.039	-0.037	-0.032	-0.020	-0.026	-0.036	-0.003
p-value	0.554	0.139	0.163	0.227	0.448	0.318	0.176	0.921

Table 6: Results of the point biserial correlation between the concreteness of action in context and performance of LM on the interpretation of figurative language phrases (C_{Emb}). None of the LMs shows a statistically significant correlation between the variables ($\alpha < 0.05$).

	GPT-3 (small)		GPT-3 (large)	
Corpus	w/ suffix	w/o suffix	w/ suffix	w/o suffix
C_{Liu}	0.601	0.591	-	0.684
C_{Emb}	0.594	0.577	0.667	0.661
C_{NoE}	0.583	0.572	0.661	0.659
$C_{Emb} + C_{NoE}$	0.591	0.574	0.665	0.660

Table 7: Comparing the zero-shot performance of the GPT-3 models *Ada* (∼350M parameters) and *Davinci* (∼175B parameters) on the different corpora (Tab. 2). The comparison includes the variable *that is to say* suffix prompting.

feature	**coef**	**se**	**T**	**p-val**	$\mathbf{R^2}$	**adj_r2**	**CI[2.5%]**	**CI[97.5%]**
Intercept	0.45075	0.12647	3.56405	0.00038	0.00393	0.00115	0.20266	0.69884
embodiment rating	0.02906	0.01449	2.00613	**0.04503**			0.00064	0.05748
freq-rating	-0.00000	0.00000	-0.07397	0.94104			-0.00000	0.00000
aoa-rating	0.00163	0.01278	0.12763	0.89846			-0.02344	0.02671
len-rating	0.00073	0.01385	0.05249	0.95814			-0.02645	0.02790

Table 8: Linear Regression results with all features including embodiment score for GPT-3 (*ada*). *coef*: regression coefficients, *se*: standard errors, *T*: T-values, *p-val*: p-values, *r2*: coefficient of determination (R^2), adj_{r2}: adjusted R^2, *CI[2.5%]*: lower confidence intervals, *CI[97.5%]*: upper confidence intervals.

Feature	coef	se	T	pval	R^2	adj_r2	CI[2.5%]	CI[97.5%]
Intercept	0.6659	0.0671	9.9174	0.00000	0.00114	-0.00095	0.5342	0.7976
freq-rating	-0.00000	0.00000	-1.0400	0.29852			-0.00000	0.00000
aoa-rating	-0.00994	0.01142	-0.8700	0.38434			-0.03234	0.01246
len-rating	-0.00227	0.01379	-0.1648	0.86916			-0.02931	0.02477
len-rating	0.00073	0.01385	0.05249	0.95814			-0.02645	0.02790

Table 9: Linear Regression results with all features excluding embodiment score for GPT-3 (*ada*). *coef*: regression coefficients, *se*: standard errors, *T*: T-values, *p-val*: p-values, *r2*: coefficient of determination (R^2), adj_{r2}: adjusted R^2, *CI[2.5%]*: lower confidence intervals, *CI[97.5%]*: upper confidence intervals

.

Figure 3: Coefficients of determination (R^2) for all models and all model sizes through linear regression with (orange) and without (blue) embodiment score as feature. For all regressions, the features Age of Acquisition, word frequency and word length have been included. For all models and sizes, the R^2 value is lower when embodiment is excluded as a feature. This is in line with the VIF analysis. Details of these linear regressions are exemplified in the results for *GPT3_350m* in Table 8 (with embodiment score feature) and in Table 9 (without embodiment score feature).

ACL 2023 Responsible NLP Checklist

A For every submission:

☐ A1. Did you describe the limitations of your work?
Left blank.

☐ A2. Did you discuss any potential risks of your work?
Left blank.

☐ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

☐ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☐ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
Left blank.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Left blank.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Left blank.

C ☐ Did you run computational experiments?

Left blank.

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Left blank.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Left blank.

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Left blank.

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Left blank.

D ☐ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Left blank.