Automatic Scoring of Metaphor Creativity with Large Language Models

Paul V. DiStefano, John D. Patterson, & Roger E. Beaty

Department of Psychology, Pennsylvania State University

Author Note

R.E.B. is supported by grants from the National Science Foundation [DRL-1920653; DUE-2155070].

Correspondence should be addressed to Roger E. Beaty, 140 Moore Building, University Park, PA 16802. Email: rebeaty@psu.edu.

Abstract

Metaphor is crucial in human cognition and creativity, facilitating abstract thinking, analogical reasoning, and idea generation. Typically, human raters manually score the originality of responses to creative thinking tasks—a laborious and error-prone process. Previous research sought to remedy these risks by scoring creativity tasks automatically using semantic distance and large language models (LLMs). Here, we extend research on automatic creativity scoring to metaphor generation—the ability to creatively describe episodes and concepts using nonliteral language. Metaphor is arguably more abstract and naturalistic than prior targets of automated creativity assessment. We collected 4,589 responses from 1,546 participants to various metaphor prompts and corresponding human creativity ratings. We fine-tuned two open-source LLMs (RoBERTa and GPT-2)—effectively "teaching" them to score metaphors like humans—before testing their ability to accurately assess the creativity of new metaphors. Results showed both models reliably predicted new human creativity ratings (RoBERTa r = .72, GPT-2 r = .70), significantly more strongly than semantic distance (r = .42). Importantly, the fine-tuned models generalized accurately to metaphor prompts they had not been trained on (RoBERTa r = .68, GPT-2 r = .63). We provide open access to the fine-tuned models, allowing researchers to assess metaphor creativity in a reproducible and timely manner.

Keywords: automated scoring; creativity; creative thinking; large language models; metaphor generation

Automatic Scoring of Metaphor Creativity with Large Language Models

Evaluating the originality of ideas poses a major challenge for creativity assessment. Traditionally, researchers have relied on time-consuming and subjective human ratings to systematically score large volumes of open-ended responses to creativity tasks. To address this labor bottleneck and potential rating biases, computational methods like semantic distance have been proposed to automate creativity scoring (Beaty et al., 2022; Buczak et al., 2023; Dumas & Dunbar, 2014; Forthmann et al., 2022; Green, 2016; Hass, 2017; Heinen & Johnson, 2018; Kenett, 2019; Landauer et al., 1998; Patterson, Merseal, et al., 2023). To date, however, these tools have mostly focused on standard divergent thinking tasks, like the Alternate Uses Task (Guilford, 1967). Little work has explored the automatic assessment of more abstract creativity tasks, like metaphor production, in which people describe experiences and concepts using nonliteral language. Metaphors are ubiquitous in both everyday speech and literary works (Billow, 1977; Gibbs, 1990), allowing people to convey complex ideas by relating one concept (e.g., the brain) to another (e.g., a computer).

Recent advances in large language models (LLMs; a class of artificial neural network), such as the Generative Pre-trained Transformer (GPT), offer new opportunities to significantly improve automatic creativity scoring, by training LLMs to rate metaphors like humans (Organisciak et al., 2023). Here, we leverage LLMs to automatically score the originality of novel metaphors, using a dataset of human-generated metaphor responses and ratings (N = 4,589). Although LLMs have successfully scored divergent thinking responses (Organisciak et al., 2023), it is unclear whether they would be similarly successful in scoring metaphors, which requires a deeper understanding of nonliteral language.

Automatic Creativity Assessment

Automatic scoring methods are increasingly employed to overcome the challenges of subjective human scoring. In contrast to subjective scoring, these “objective” methods use machine learning and text analysis tools to compute originality metrics. The first automated scoring method for divergent thinking tasks involved text-mining variables such as word count and average word length (Paulus et al., 1970). Decades later, this approach still holds up as effective (Forthmann & Doebler, 2022).

In recent years, the most popular approach has been semantic distance—a natural language processing technique that captures the remoteness or novelty of an idea by mathematically comparing word vectors that are learned from text corpora, most often vectors based on distributional semantics (e.g., Latent Semantic Analysis, LSA, Landauer et al., 1998; and Global Vectors for Word Representation, GloVe, Pennington et al., 2014). For example, the words *coffee-drink* are less semantically distant than the words *coffee-write*. When applied to divergent thinking tasks, such as the AUT, semantic distance is often computed between the prompt/object (e.g., brick) and all words in the response (e.g., grind it up and make a filtering substance; Yu et al., 2023). The reliability and validity of semantic distance have been demonstrated, with several studies reporting high positive correlations with human originality ratings (Hass, 2017; Heinen & Johnson, 2018; Kenett, 2019; Landauer et al., 1998) and other measures, such as creative behavior (Beaty et al., 2022; Beaty & Johnson, 2021; Dumas et al., 2020; Fan et al., 2023; Yu et al., 2023). Anyone can now compute semantic distance on word association and divergent thinking tasks using openly available scoring platforms, SemDis (semids.wlu.psu.edu) and Open Creativity Scoring (openscoring.du.edu).

Another recent automatic scoring method is divergent semantic integration (DSI; Johnson et al., 2022). DSI was developed to analyze semantic distance in narratives, such as short stories. It assesses the extent to which divergent ideas are connected by computing the distance between all pairs of words in a text. Johnson and colleagues applied DSI to short stories, finding that DSI correlated strongly with human creativity ratings—explaining up to 72% of the variance.

A key innovation of the DSI metric compared to previous semantic distance methods was the introduction of transformer neural networks to generate the word vector representations used in semantic distance computations. Transformer-based models—commonly referred to as large language models (LLMs)—include neural network architectures such as the Bidirectional Encoder Representations from Transformers (BERT; Devlin et al., 2019) model. LLMs are pre-trained by predicting missing words within input sequences using extensive training datasets. Various models employ different pre-training regimes and use different text corpora for training. Consequently, these differences in training change the model's capabilities. Additionally, models also vary in terms of their number of parameters, where the greater number of parameters the model has corresponds with a higher model capacity.

Importantly, BERT and other LLMs (e.g., the Generative Pretrained Transformer, GPT; Radford et al., 2019) produce context-dependent language representations. That is, they are able to consider each word in the input to the model in relation to the other words in the input, rather than in isolation, as with legacy word representation approaches such as LSA and GloVe. Transformer models can thus accommodate multiple word meanings (e.g., *instrument-bass* vs. *fish-bass*), thanks to their "self-attention" mechanism, which allows them to dynamically adjust the interpretation of each word based on the context provided by the surrounding words in the input.

Another advantage of LLMs is that they can be "fine-tuned" through supervised learning (human guidance) to significantly enhance their performance on specific tasks. Supervised learning results in predictive models rather than descriptive ones (Lantz, 2013). LLM fine-tuning involves adjusting the pre-trained models (e.g., BERT, GPT) to learn from specific input-output pairs, such as AUT responses and human ratings. In this way, the pre-trained model takes advantage of the knowledge it learned during pre-training and adapts it to the specific task or dataset. Fine-tuned models are thus capable of predicting outputs (creativity ratings) for unseen inputs (responses). The process of fine-tuning can dramatically improve model performance compared to pre-trained models that were not fine-tuned on a specific task, often matching or exceeding human performance (Bakker et al., 2022).

Recently, Organisciak et al. (2023) employed LLMs to automatically score the originality of AUT responses, showing substantial improvement in the prediction of human ratings compared to semantic distance. The authors fine-tuned the models T5-Base and GPT-3 on a large set of AUT responses and human ratings, finding that the fine-tuned models strongly predicted human creativity ratings for new (previously unseen) responses. Moreover, the fine-tuned models generalized to AUT items (objects) that were not included in the training dataset, suggesting they had learned something general about how humans rate creativity on the AUT that extended beyond the specific items they were trained on. However, it is unclear what the models are attending to exactly. LLMs may be influenced by the humans' rating distributions or the prompt that it was trained on. This work built upon other supervised learning methods to automatically score the AUT, such as Buczak et al. (2023) and Stevenson et al. (2020), who used more extensive training approaches to automate creativity scoring on the AUT. To date, however, supervised machine learning of verbal creativity scoring has been primarily focused on the AUT, with little work on other verbal creativity tasks, such as novel metaphor production. One of the few verbal creativity tasks that has been scored using artificial intelligence is the Measure of Original Thinking for Elementary School (MOTES), and a similarly high level of agreement with human raters was found (Acar et al., 2023).

The Present Research

In the present study, we aimed to extend research on automatic creativity scoring to novel metaphor production. Metaphor is a naturalistic form of verbal creativity that is used to describe concepts and episodes using figurative language (e.g., “time is a thief”), and it is a powerful rhetorical tool for creatively conveying complex ideas. Compared to conventional metaphors, which involve recalling culturally familiar expressions (e.g., describing a boring experience as “watching paint dry”), novel metaphors reflect entirely new expressions of figurative language that people produce spontaneously (e.g., describing a boring experience as “watching a turtle sprint”). Metaphor is increasingly studied in creativity research, including behavioral (Stamenković et al., 2023) and neuroimaging (Beaty et al., 2017; Cardillo et al., 2012) studies, and it has been shown to be an effective tool for boosting learning in educational settings (Tiberius, 1986). Automatic scoring of novel metaphors could accelerate the pace of research, which currently relies on subjective human scoring. While automatic scoring of metaphor novelty has been investigated using syntactically related word pairs (Parde & Nielsen, 2018), automated assessment of metaphor creativity, specifically, has yet to be explored. Additionally, LLMs have yet to be used for metaphor creativity evaluation.

However, there is reason to suspect that language models may struggle with scoring novel metaphors, based on how humans process metaphors. Humans use metaphors by relating a “topic” (i.e., the mind) to a “vehicle” (i.e., a machine) that is conceptually but not literally related. How humans understand nonliteral language like metaphor has been studied for several decades, largely focusing on metaphor structure and function (Gibbs, 1994; Glucksberg et al., 1997; Lakoff & Johnson, 2008). The property attribution model of metaphor proposes that metaphor comprehension involves creating an abstract link between a topic and a vehicle with similar characteristics (Glucksberg & McGlone, 2001). Additionally, for a metaphor to be comprehensible, the shared knowledge between the topic and vehicle must be identified (Glucksberg et al., 1997). Humans also tend to select metaphors that are semantically similar to the target concept (Clevenger & Edwards, 1988), yet people prefer metaphors of moderate semantic distance (Katz, 1989). Creative metaphors should therefore be semantically distant yet appropriate enough to be relatable to the topic. If so, a measure of semantic distance alone is unlikely to capture metaphor creativity.

In the context of natural language processing (NLP), metaphor is a relatively under-studied topic, and it remains an open question as to how well language models can interpret nonliteral phrases (Liu et al., 2022). Historically, computational linguistics have worked to understand how both human language and computer models can understand the complexities of syntax and semantics through studying metaphor, wordplay, and garden path sentences. Due to the prevalence of figurative language in naturalistic language, there has been interest in using NLP to understand metaphor (Carbonell, 1982; Shutova et al., 2013; Veale et al., 2016), wordplay (Taylor & Mazlack, 2004), and garden path sentences (Jia-li & Ping-fang, 2013). In humans, it is believed that language comprehension systems create 'good enough' representations of syntax and semantics (Ferreira & Patson, 2007) and make robust predictions for each aspect (Ferreira & Qiu, 2021). This work highlights the importance of context in parsing and processing complex sentence structure and could offer valuable insight into how large language models can begin to understand figurative language more deeply.

Some recent studies have shown that LLMs have a limited capacity to handle figurative language, such as irony, sarcasm, idioms, and metaphors (Chakrabarty et al., 2022; E. Liu et al., 2022). For example, Chakrabarty et al. (2022) found that pre-trained LLMs lagged substantially behind humans in generating or choosing plausible continuations for narratives with figurative expressions. LLMs may therefore lack the necessary world knowledge and commonsense to 'comprehend' nonliteral language, limiting their ability to reliably evaluate metaphor quality—perhaps even after fine-tuning.

Here, we tested the capacity of LLMs to automatically score metaphors using data from previous studies of metaphor production (N = 4,589 responses, N = 1,546 participants). Participants in these studies were presented with open-ended prompts that related to common experiences (e.g., describing a bad movie with a metaphor). We fine-tuned two popular, open-source LLMs—RoBERTa and GPT-2—on these responses and corresponding human creativity ratings. We then tested their ability to predict human ratings for new responses they had not seen before—both those based on metaphor prompts the model was trained on as well as those based on prompts the model was never trained on, as a test of far generalization.

In addition, we scored metaphor responses using word count and DSI. Word count captures the elaboration of a response, and it has been previously related to creativity ratings on other tasks (Forthmann & Doebler, 2022). On the other hand, DSI provides a semantic distance baseline that is more suitable than other approaches to semantic distance given its ability to handle multi-word responses in a context-dependent manner. Although we expected the fine-tuned LLMs to correlate with human ratings more strongly than DSI, based on related fine-tuning research (Organisciak et al., 2023), DSI provides a more interpretable metric of semantic distance. Using context-dependent pairwise word comparisons, DSI can quantify the relatedness of the concepts underlying metaphors.

Methods

The data used in this study was compiled from prior studies and unpublished research. In many of the datasets, participants completed metaphor generation tasks in addition to other cognitive assessments. The specific cognitive tasks and their order differed for each study; we only include metaphor tasks here. All metaphor responses were assessed for creativity by human raters. All data and code are available online at https://osf.io/2dqpj/?view_only=339b6f7febe646309484bb06eb2914e9.

Metaphor Task and Human Scoring

The metaphor generation task presented participants with a prompt and asked them to write a creative metaphor (Beaty et al., 2017; Beaty & Silvia, 2013; Kasirer & Mashal, 2018; Silvia & Beaty, 2012). Participants responded to four different prompts which varied across the datasets in this study: *boring class*: ‘Think of the most boring high-school or college class you’ve ever had. What was it like to sit through?’; *gross food or drink:* ‘Think about the most disgusting thing you ever ate or drank. What was it like to eat or drink it?’; *bad movie*: ‘Think about the worst movie or TV show you have ever seen. What was it like to watch it?'; and *messy room*: 'Think of the messiest room that you've ever had to live in. What was it like to live there?'. Participants were given examples of different types of metaphors (e.g., compound metaphor), as well as optional starting stems to help them (e.g., "Sitting through that movie was…").

Metaphors were rated for creativity using the subjective scoring method (Silvia et al., 2008), which is based on an adapted version of the Consensual Assessment Technique (CAT). Consistent with past work (Beaty & Silvia, 2013; Silvia & Beaty, 2012), metaphors were judged by multiple raters using a 5-point Likert scale, where 1 represented *not at all creative* and 5 represented *very creative*. Each dataset was rated by 2 to 4 trained raters (see rater agreement below), who rated their respective dataset independently. In general, cliche metaphors tended to receive lower ratings (e.g., "Sitting through that movie was like watching paint dry") whereas clever or humorous metaphors received higher ratings (e.g., "That movie was a good-looking guy with no personality").

Datasets

Across the seven datasets, there are 4,589 responses from 1,546 participants (1,058 bad movie; 1,387 boring class; 1,385 gross food; 759 messy room). This study protocol was approved by the Penn State University Internal Review Board (IRB STUDY00010475). Table 1 (below) is an overview of the datasets, including the sample size, prompts, number of raters, and intraclass correlation coefficient (ICC; rater agreement):

Table 1

Summary of Metaphor Datasets

Dataset	**Participants**	**Responses**	**Raters**	***ICC(3, k)***	**Prompt(s)**
1	222	443	4	0.75	*gross food or drink, boring class*
2	164	330	4	0.84	*gross food or drink, boring class*
3	151	302	3	0.73	*bad movie*
4	476	1,888	4	0.77	*gross food or drink, boring class, bad move, messy room*
5	133	266	3	0.60	*gross food or drink, boring class*
6	111	214	2	0.88	*gross food or drink, boring class*
7	289	1,146	4	0.82	*gross food or drink, boring class, bad move, messy room*

Computational Experiments

Our main goal was to test two methods for automatically scoring metaphors: semantic distance and fine-tuned LLMs. To compute semantic distance, we used Divergent Semantic Integration (DSI; Johnson et al., 2022), a baseline semantic model, in contrast to the fine-tuned models. Regarding fine-tuning, we compared two supervised LLM architectures, RoBERTa and GPT-2. Both approaches employ "context-sensitive" models—which can capture nuanced word meanings (e.g., *instrument-bass* vs. *fish-bass*)—and thus should be better suited to capture metaphors than context-independent models (e.g., latent semantic analysis).

Baseline: Divergent Semantic Integration

DSI computes word-to-word semantic distance between all words in a response, reflecting the extent to which a response integrates diverse topics and contexts. DSI does so by extracting vector representations for each word in a response from two early-middle layers of BERT-large (a 24-layer LLM; Devlin et al., 2019)—layers that carry both syntactic and semantic linguistic knowledge (Jawahar et al., 2019). Semantic distance (1 minus cosine similarity) is then computed between all pairs of word vectors; the semantic distance values are then averaged to give the DSI score for the response. For more on DSI—including a tutorial and code for computing it—see Johnson et al. (2022) and its associated Open Science Foundation repository (https://osf.io/ath2s/). The DSI code was unchanged from the repository to predict the creativity of metaphors in this study.

DSI has been used to automatically score multi-word responses (i.e., short stories), and has been shown to explain substantial variance in human creativity ratings across multiple datasets (Johnson et al., 2022). Like short stories, metaphors are often very elaborate and vary considerably across participants. Earlier automated scoring methods that use context-insensitive semantic distance (e.g., latent semantic analysis) would be inappropriate to score metaphors because metaphors rely on abstract relationships that cannot be well represented using context-insensitive models. Since empirical work has shown that the highest-performing variant of DSI uses a context-sensitive LLM to render the word vectors, it stands as a theoretically valid baseline of automated metaphor assessment. The present study thus provides the first test of whether semantic distance (in the form of DSI) relates to human creativity ratings of metaphors.

Fine-Tuning LLMs

We fine-tuned and evaluated two transformer-based LLMs, RoBERTa (Y. Liu et al., 2019) and GPT-2 (Radford et al., 2019), to predict human creativity ratings of metaphors. These models were chosen for four reasons. Most importantly, they are both highly-performant models known to do well on natural language processing benchmarks as well as match human behavior and neural activity (e.g., Caucheteux et al., 2022; Johnson et al., 2022). Second, both models are open-access and free to use. This stands in contrast to proprietary models like GPT-4 (OpenAI) that cost money for researchers to train and use and cannot be downloaded off the proprietary server environment. Consequently, when proprietary LLM companies make upgrades and deprecate older LLMs (e.g., GPT-3 vs. GPT-4), the old models are lost forever—harming reproducibility. A third reason for choosing these two models is because they have comparable parameter counts yet also have distinct model architectures (i.e., design principles) and training regimes (i.e., the pre-training tasks used to build up syntactic and semantic representations of language)—affording an alluring comparison. Finally, and arguably most importantly, we chose these models given their compatibility with computational resources researchers are likely to have access to; the fine-tuned models can be run on standard laptops with 16GB of RAM. This allows for the fruits of this work to impact the greatest number of researchers and practitioners.

The Robustly optimized BERT approach, RoBERTa-base (Y. Liu et al., 2019), is a 125M parameter model based on the Bidirectional Encoder Representations from Transformers (BERT; Devlin et al., 2019) architecture. However, RoBERTa was trained under an improved pre-training regime that was empirically shown to result in more robust language representations (Y. Liu et al., 2019). BERT and subsequently RoBERTa were trained with a word corpus sourced from Wikipedia and BooksCorpus. A key characteristic of RoBERTa is that it falls under a class of transformers known as 'bidirectional' models. During the pre-training process, RoBERTa is trained using a masked language modeling objective—meaning that select words in the input are hidden from the model and the model attempts to predict the masked words with the surrounding context words. Because RoBERTa is a bidirectional model, the context words that occur both before and after the masked words are used to predict the masked words. In other terms, words that occur ‘after’ a masked word can be used to predict a missing word, in addition to words that occur ‘before’ the masked word in the sequence.

The final model to be openly released by OpenAI, GPT-2, is a 137M parameter, pre-trained transformer model. GPT-2 was trained on the WebText dataset, which was created by crawling Reddit and harvesting text from outbound links. While having a comparable number of parameters to RoBERTa, GPT-2 is pre-trained using a different language modeling objective. Unlike the bidirectional learning objective of RoBERTa, where context later in a sequence can be used to predict missing words earlier in a sequence, GPT-2 uses a unidirectional next-word-prediction pre-training objective. This means that the missing word the model is trained to predict during the pre-training process is always the final word in the sequence, and the context used to predict the missing word always precedes the missing word. Some have argued the unidirectional prediction task adopted by GPT-2 (and models like it) is more in line with human predictive processing that is often temporally constrained in a comparable fashion. Indeed, models that perform best at the next-word-prediction objective align more strongly with human neural recordings (Schrimpf et al., 2021).

Data Preprocessing

The creativity ratings provided by a varying number of raters were averaged for each response. The average human ratings were also z-scored within each dataset. The averaged z-scored human ratings were then used in the fine-tuning process. To fine-tune the models, we developed a dedicated data pipeline for each LLM. Consistent with best practices in machine learning (Zhou, 2021), metaphor responses and human ratings were randomly distributed into training, validation, and held-out test datasets, utilizing a 70/10/20 split ratio respectively. The data split was the same for the two models to improve comparison. To be clear, the training set consisted of responses the model learned about scoring metaphors from (i.e., these examples are used to update the weights of the LLM to perform better on the rating task). The validation set served as a pseudo-test set and was used as a barometer of how the model was performing under different settings (a.k.a., hyperparameters; detailed further below). The validation set was used for model selection, but the model was never trained on the validation set. The held-out test set consisted of responses that were neither used for model selection nor were experienced by the model during training. It thus served as a stringent test of the model's ability to generalize to untrained responses. The held-out test set was only employed once the best model settings—as revealed by the validation set—were finalized.

As a more extreme test of the model's generalizability, we also tested the model on responses to a metaphor prompt it was never exposed to during the training process (the 'messy room' prompt). If the models can successfully generalize to new metaphor prompts, this would suggest that they had learned something general about how humans rate metaphor creativity, beyond the specific prompts they saw during training. Successful generalization would also suggest that new metaphor prompts could be continually developed and automatically scored, without needing additional fine-tuning of the model.

The raw responses in each of the datasets (i.e., training, validation, held-out test, held-out prompt) were first inserted into a sentence frame. As noted above, transformers' performance and outputs can be affected by the sentential context. As such, we varied what the critical adjective in the sentence frame ('creative,' 'novel,' 'useful,' 'surprising,' or 'unique') that the model was supposed to assess for. The model's training data consisted of sentences structured as follows: "A [adjective] metaphor for [prompt] is [response]." For instance, the model was trained on combinations of adjectives, prompts, and responses, such as "A creative metaphor for a boring class is my class was like a farmer's market: we were all vegetables." Notably, Organisciak et al. (2023) employed the adjective 'surprising' for all fine-tuned AUT items. The framed responses in each of the four datasets then underwent tokenization; tokenization is the process by which natural language is transformed into a numerical vocabulary that is understandable to LLMs. These tokenized versions of the text were used to train and evaluate the models (via the Huggingface Trainer application programming interface [API] for Python).

Model Training and Hyperparameter Search

To prepare the LLMs for the task of metaphor creativity prediction, we affixed a regression head atop each LLM (via the Hugging Face AutoModelForSequenceClassification API, with the number of output labels set to 1). This projects the high-dimensional output of the LLM onto a single output regression node. This is desirable, as the continuous-valued output of the regression node for a given input corresponds to the model's creativity assessment for that input, and the targets (i.e., ratings) were also continuous-valued.

In order for the computer to understand the text it must be tokenized, or broken up into smaller pieces that can be represented with numbers. The tokenization is dependent on the LLM, in this case either GPT-2 or RoBERTa. To ascertain which model settings (i.e., hyperparameters) optimized the performance of each LLM, we subjected each model to a hyperparameter search using the Optuna package for Python (Akiba et al., 2019). In this fine-tuning process, we used the default Hugging Face optimizer, AdamW, which adjusts the model weights using weight decay (Loshchilov & Hutter, 2019). Across 60 trials, we searched over four hyperparameters. First, we searched over learning rate within the range 5e-07 to 5e-02. Learning rate affects the extent to which training episodes change the weights of the model, where higher learning rates correspond to making bigger changes to the model's weights for each training batch. If the learning rate is too high, the model will not learn and may become progressively worse across training. If the learning rate is too low, the model may not reach a viable solution during training, especially if the training period is brief. Second, we searched over batch size, testing values of 4, 8, 16, and 32. Batch size reflects how many responses the model receives feedback on at a time. Larger batch sizes lead to faster convergence/training, but can also lead to suboptimal solutions. Smaller batch sizes introduce beneficial stochasticity into the training process but take more time. Third, we searched for the optimal number of training epochs; that is, the number of full training passes through the training dataset. We conducted a search within the epoch range of 10 to 150. Last, we searched over the prefix adjective used in the sentence frame (e.g., ‘creative’, ‘novel’, or ‘surprising’; see above for the full list).

The Optuna package assists the search over these hyperparameter settings by employing the Tree-structured Parzen Estimator approach, a Bayesian optimization method. At first, Optuna assumes a uniform prior over values for each hyperparameter setting, but as the algorithm updates its priors—by observing which hyperparameter combinations optimize performance on the validation set—it begins to explore combinations that are most likely to perform optimally. The metric Optuna sought to optimize (minimize) was the mean squared error (MSE) between model-predicted and human-provided ratings on the validation set. The MSE serves as the error signal to update the weights of the model during training based on the training dataset labels. Lower MSE indicates higher prediction accuracy.

After the optimal hyperparameters were identified for each model (individually), we evaluated the best-fitting models on the held-out test set based on correlation with the human ratings. We then assessed the ability of the model to generalize to responses from a held-out prompt ('messy room'). In all analyses, we removed outliers from both the human and model creativity ratings, defined as values beyond 3 standard deviations from the mean.

In summary, the regression head affixed to the LLM predicts the creativity ratings. The model's predictions and performance are influenced by various hyperparameters (model settings), so we conducted tests to identify the optimal combination of hyperparameters to yield the best performance.

To achieve this, we employed Optuna to minimize the MSE. Once the optimal hyperparameters were determined, we assessed the model's performance on both the test set and the held-out prompt.

Results

Baselines: Word Count and Semantic Distance

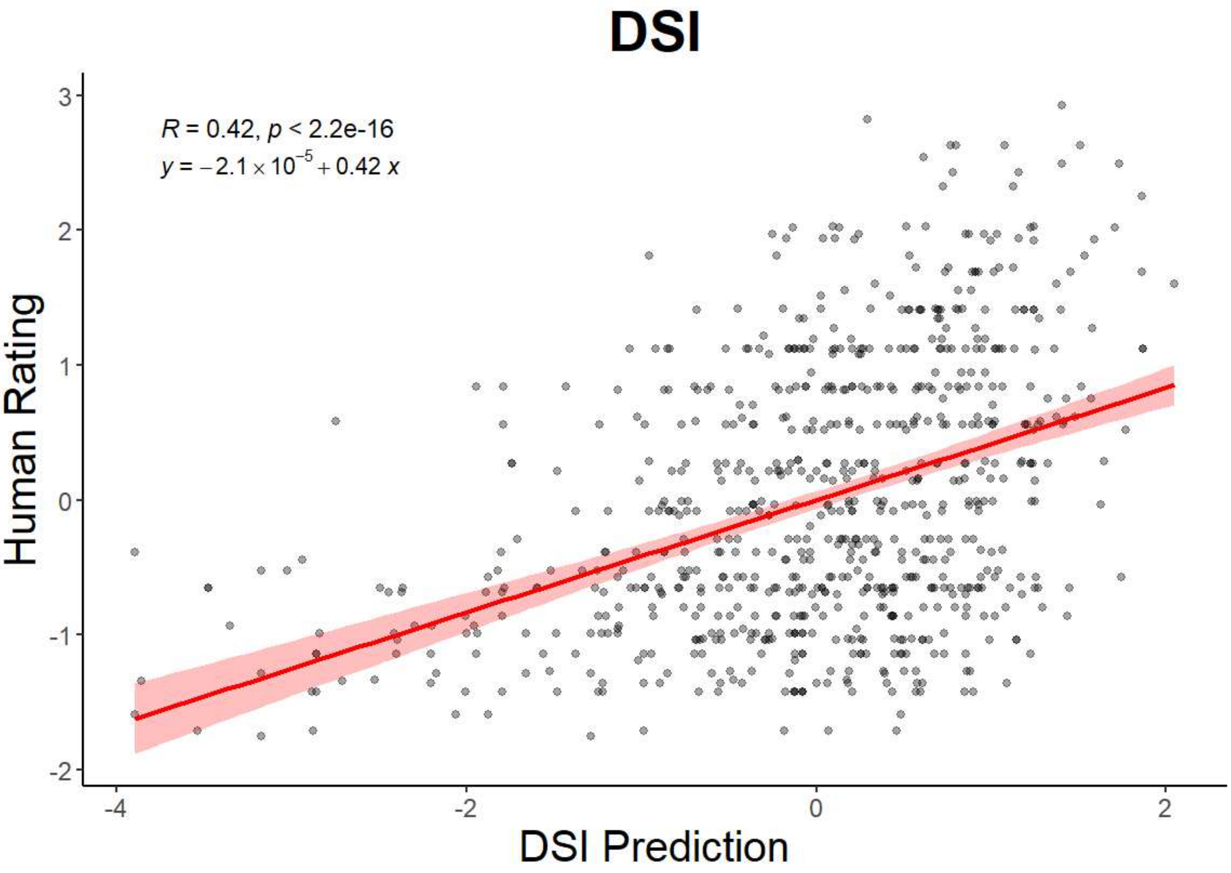
We began by computing Pearson correlations between human creativity ratings of metaphors and two baseline measures: word count (a measure of elaboration) and DSI (a measure of semantic distance). This correlational analysis was conducted three times, once on the full dataset (*N* = 4,589 responses; *n* = 4,509 post outlier removal), once on the held-out test set (*N* = 766; *n* = 753), and once on the held-out prompt (*N* = 759; *n* = 750). The full dataset analysis allows us to extend previous research on word count and DSI for creativity assessment by computing this correlation on a large set of metaphors. The analyses with the held-out test set and held-out prompt allow for a fair comparison between word count, DSI scores, and the fine-tuned LLM predictions.

Regarding word count, we observed a correlation of *r* = 0.51; CI 95% [0.48, 0.53] with human creativity ratings of the full dataset, suggesting that the raters scored more elaborate metaphors as more creative, consistent with past work with other verbal creativity tasks (Dumas et al., 2021; Johnson et al., 2022). We found similar correlations between word count with the human ratings of the held-out test set (*r* = 0.49; CI 95% [0.43, 0.54]) and the held-out prompt (*r* = 0.47; CI 95% [0.42, 0.53]).

Next, we used BERT DSI to assess the extent to which semantic distance correlates with human creativity ratings. Using the full dataset (*n* = 4544), we found that DSI scores were positively correlated with human ratings, *r* = 0.31; 95% CI [0.28, 0.33]. Moreover, calculating the DSI scores using only the held-out test set (*n* = 762) yields a correlation of *r* = 0.42; 95% CI [0.36, 0.47] with human ratings (see Figure 1). We observed the highest correlation between the DSI scores of the held-out prompt and the human ratings, *r* = 0.51; CI 95% [0.46, 0.56]. The magnitude of this effect is moderate, and notably smaller than previous findings with short stories (Johnson et al., 2022). This finding indicates that more semantically distant metaphors were rated as slightly more creative by humans.

Figure 1

Correlation between Human-Rated Creativity and DSI Ratings of the Held-out Test Set



Note. DSI and human ratings are Z-scored. *n* = 762; 4 outliers were removed.

Hyperparameter Results

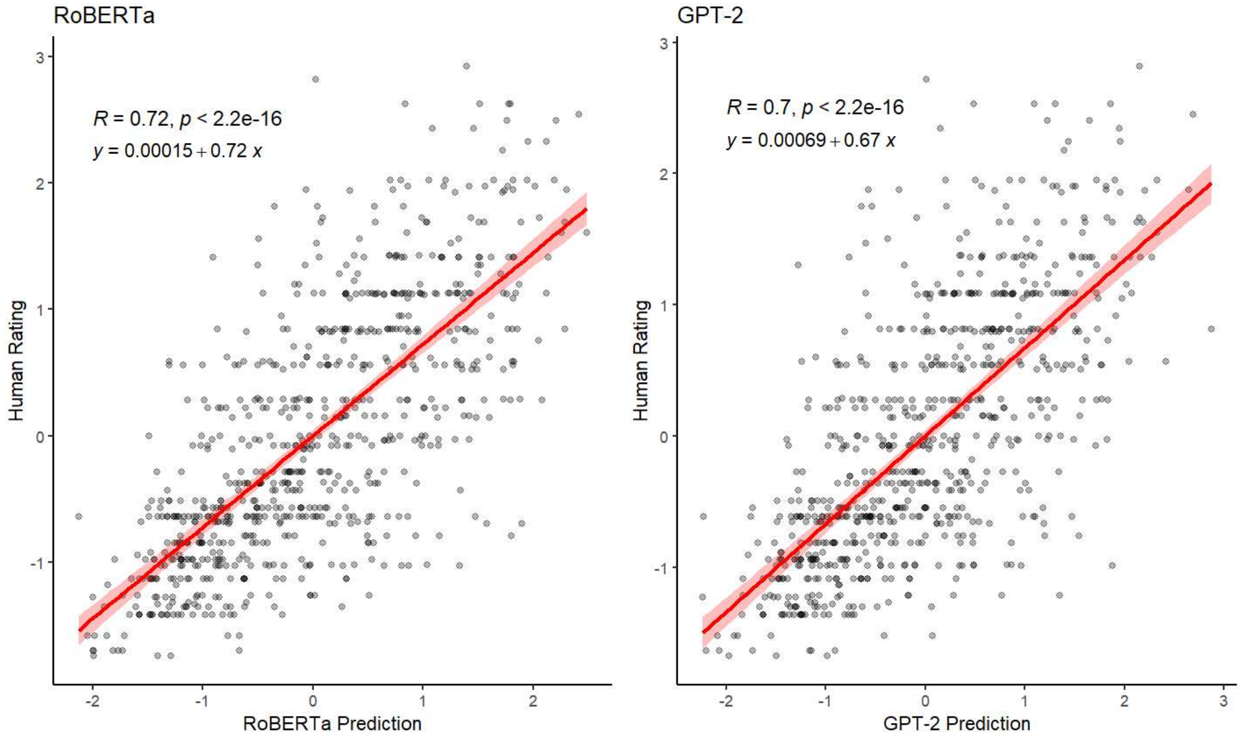
Next, we conducted fine-tuning experiments on the two LLMs: GPT-2 and RoBERTa, to assess their ability to replicate human creativity ratings of metaphors. After 60 training trials of each model, we selected the parameters from the trial with the lowest mean squared error on the validation set (*N* = 383), representing the best-performing configuration.

For GPT-2, the top-performing trial had an MSE of 0.461, utilized a learning rate of 2.47e-05, ran 139 epochs, and employed a training batch size of 16. The best-fitting model used the prefix adjective 'novel' (as opposed to other prefixes; e.g., creative). Regarding RoBERTa, the highest-performing trial had an MSE of 0.450, used a learning rate of 4.00e-06, ran for 111 epochs, and had a training batch size of 4. Unlike GPT-2, the best performing RoBERTa model used the prefix adjective 'surprising.'

Held-out Test Performance

Figure 2

Correlation between Human-Rated Creativity and Model Predictions from the Test Set



Note. Model predictions and human ratings are Z-scored. RoBERTa $n = 765$, GPT-2 $n = 764$; 1 outlier was removed from the RoBERTa predictions, and 2 outliers were removed from the GPT-2 predictions.

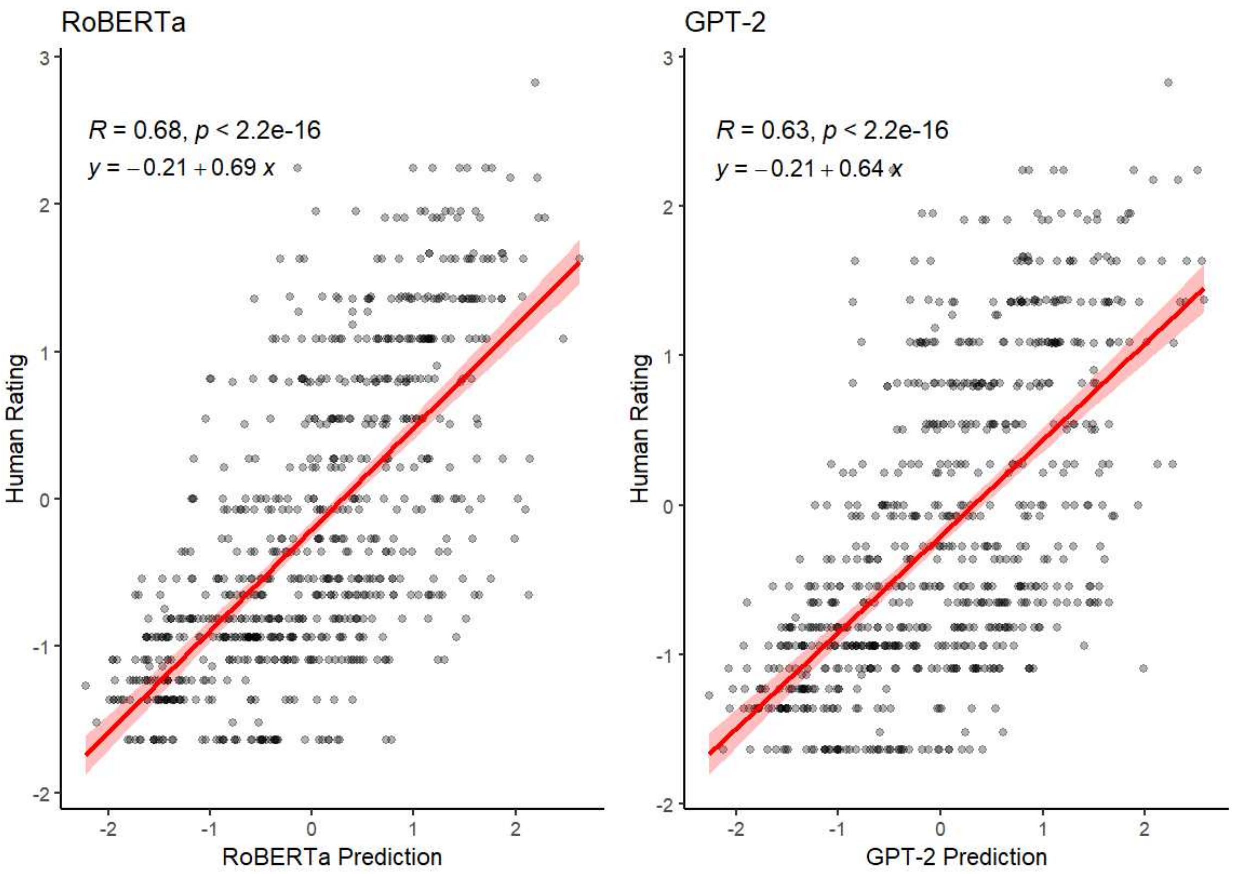
When evaluated on the test data ($n = 764$) with the optimized hyperparameters, GPT-2 positively predicted human creativity ratings, yielding a correlation of $r = 0.70$; 95% CI [0.66, 0.73]—substantially larger than the semantic distance and elaboration baselines. Upon evaluation of the test set ($n = 765$)

with RoBERTa, the predicted creativity scores also demonstrated a strong agreement with human ratings, *r* = 0.72; 95% CI [0.69, 0.76]. Figure 2 (above) shows the correlation of model predictions with the human creativity ratings of the test set. The RoBERTa correlation was numerically larger than GPT-2 and considerably larger than semantic distance. In both models, we also determined that mean prediction errors (MPE) were higher for the responses that had the highest and lowest creativity, measured by two standard deviations from the mean, RoBERTa MPE within 2 *SD*: 0.678, RoBERTa MPE ± 2 *SD*: 1.16; GPT-2 MPE within 2 *SD*: 0.653, GPT-2 MPE ± 2 *SD*: 0.99, suggesting the models consistently underestimate highly creative outputs and overestimate less creative ones. Taken together, the results demonstrate that fine-tuned LLMs, with different neural network architectures, can robustly capture human creativity ratings of novel metaphors.

Held-out Prompt Performance

Figure 3

Correlation between Human-Rated Creativity and Model Predictions from the Held-out Metaphor Prompt



Note. Model predictions and human ratings are Z-scored. *n* = 758; 1 outlier was removed for both models.

To assess the generalizability of the fine-tuned models, we evaluated their ability to predict human creativity ratings of metaphors on a previously unseen prompt: 'messy room' (*N* = 759). When evaluating the responses associated with the untrained metaphor prompt (*n* = 758), GPT-2 exhibited comparable performance to the prompts it was trained on, resulting in a correlation of *r* = 0.63; 95% CI [0.59, 0.67]. RoBERTa also demonstrated transferability to the previously unseen prompt (*n* = 758), with a correlation *r* = 0.68; 95% CI [0.63, 0.71], see Figure 3. We once again found that the MPE for both were higher for the responses that had creativity ratings two standard deviations from the mean, RoBERTa MPE within 2 *SD*: 0.689, RoBERTa MPE ± 2 *SD*: 1.32; GPT-2 MPE within 2 *SD*: 0.651, GPT-2 MPE ± 2 *SD*:

1.15. Thus, both GPT-2 and RoBERTa showed evidence of learning how humans rate the creativity of novel metaphors—beyond their training datasets.

Discussion

The present study introduces a method for automatically assessing metaphor creativity by harnessing the power of large language models (LLMs). Given the complexity of metaphor comprehension in humans, computationally scoring metaphors presented a compelling challenge. Despite this challenge, our results demonstrate that LLMs fine-tuned on human ratings of metaphor creativity can reliably predict the creativity of metaphors on both untrained responses and on metaphors from an untrained prompt. This finding underscores the potential of LLMs to both 'understand' figurative language and automate the assessment of metaphor creativity.

Automatic scoring methods can increase the reproducibility and efficiency of creativity assessment—relieving humans from labor-intensive subjective scoring. Automatic scoring has primarily focused on the AUT, leaving other verbal creativity tasks, such as metaphor production, without dedicated tools. Here, we sought to automate the scoring of the metaphor generation task by leveraging recent advances in LLMs. We fine-tuned two widely used LLMs, RoBERTa and GPT-2, to explore the viability of automated scoring of metaphor creativity. Despite differences in their pre-training regimes, both models robustly predicted human creativity ratings and surpassed the performance of the baseline semantic distance metric used: DSI. Together these findings indicate that fine-tuned LLMs provide a promising alternative to human creativity ratings of metaphors, and suggest that LLMs can be trained to understand the creativity of complex, nonliteral language (cf. Ichien et al., 2023)

Prior research on automatic verbal creativity scoring has focused on semantic distance applied to the AUT (Beaty et al., 2022; Dumas et al., 2020; Dumas & Dunbar, 2014; Forthmann et al., 2022; Hass, 2017; Yu et al., 2023) and short stories (Johnson et al., 2022). Johnson et al. (2022) found that

DSI—which computes semantic distance between all words in a response—has high predictive power for human ratings of creativity in short stories. We applied the same DSI approach to metaphors to test how well they relate to human creativity ratings. Our results yielded a modest correlation with the test set (*r* = 0.42), which is smaller than what was found for short stories. This finding is notable given the theoretical relevance of semantic distance to metaphor (Clevenger & Edwards, 1988; Katz, 1989, 1992; Winter & Strik-Lievers, 2023): by definition, metaphor involves connecting two seemingly unrelated concepts, so one might expect semantic distance-based metrics to be relevant. Yet metaphors must also be apt and meaningful, and previous work suggests people prefer metaphors of moderate semantic distance (Katz et al., 1988). Nevertheless, our study provides a first demonstration that semantic distance captures some variance in human ratings of metaphor creativity.

The main objective of this study was to fine-tune LLMs to predict human creativity ratings for novel metaphors. Recently, Organisciak et al. (2023) fine-tuned two state-of-the-art models (GPT-3 and T5) to predict human creativity ratings of AUT responses, reporting strongly positive correlations that far exceeded semantic distance. We extended this general approach to metaphor generation, using two smaller, widely-used, and open-source LLMs that varied in the design of their self-attention mechanism: RoBERTa (a bidirectional model) and GPT-2 (a unidirectional model). Both RoBERTa and GPT-2 demonstrated the ability to generate creativity scores comparable to human raters, with RoBERTa exhibiting slightly superior performance (*r* = .72) compared to GPT-2 (*r* = .70). Interestingly, the top performing RoBERTa and GPT-2 hyperparameters used different prompt prefixes—‘novel’ for GPT-2 and ‘surprising’ for RoBERTa. ‘Surprising’ is the same prefix used in the AUT fine-tuning study of Organisciak et al. (2023), but not ‘novel.’ Our findings demonstrate the need for a more exhaustive prefix search to elicit the highest performance from LLMs. Alternatively, one could employ a multi-adjective approach, where scores for all the adjectives are averaged.

We further evaluated the performance of these fine-tuned models on a more challenging scenario: metaphors generated using a previously unseen prompt (‘*messy room*’). Remarkably, both RoBERTa and GPT-2 maintained strong performance in replicating human creativity scores for these novel metaphors. Once again, RoBERTa (*r* = .68) narrowly outperformed GPT-2 (*r* = .63). The models thus seemingly learned something general about creativity in figurative language that generalized beyond their training data. Practically, future researchers could use these models to score the creativity of new metaphor prompts that were not included in this study, though model performance may vary, especially if the prompt significantly deviates from the structure and topics in the current work. We recommend testing new prompts with a small sample of human ratings to ensure model scores are reliable and valid. Future work should further explore the far-generalization capabilities and failure modes of these models as a function of linguistic and conceptual distance from the training set.

Although the present research focused on RoBERTa and GPT-2 for automated metaphor rating, it is worth noting that there are continual advances made in the field of LLMs. Our results indicate that these models can be effectively fine-tuned to score the creativity of metaphors. It is important to recognize that ongoing developments in LLMs have led to the emergence of newer, high-performing models. While RoBERTa and GPT-2 have exhibited robust performance, recent progress in the field has given rise to open-source models that can surpass their capabilities while maintaining similar computational efficiency. Importantly, the foundational principles and successful methodologies outlined in this paper remain pertinent, offering enduring insights into the automated scoring of metaphor creativity, even in the face of evolving state-of-the-art LLMs.

Taken together, our results demonstrate the viability of LLMs to score metaphor creativity, with implications for natural language processing research on the ability of LLMs to understand figurative language. Expanding upon recent findings that LLMs can seemingly “understand” figurative language (Ichien et al., 2023; E. Liu et al., 2022), these results show LLMs can be fine-tuned to pick up figurative language well enough to evaluate metaphor creative quality. Automatic scoring further helps to lower the labor cost and increase the reproducibility of creativity research, allowing researchers and practitioners to efficiently assess creative thinking in academic or industry settings.

Strengths, Limitations, and Future Directions

Our results demonstrate the robust capacity of LLMs to automatically score metaphor creativity. While previous work has shown that LLMs can automatically score divergent thinking tasks (i.e., the AUT; Organisciak et al., 2023), our work extends this capability to metaphor—a naturalistic form of figurative language. Notably, we show that it is possible to closely match human creativity ratings for metaphors, despite using relatively small open-source models (i.e., GPT-2 and RoBERTa). The development of efficient, yet powerful, open-source models for metaphor assessment represents a key strength of the present work. The resulting models are compatible with computational resources researchers or educators are likely to have access to and are free to use—broadening potential impact. The models will also exist in perpetuity on an Open Science Foundation repository, which facilitates reproducibility in the future. In contrast, very large closed-source models (e.g., GPT-4) cost researchers money to train and use, and are subject to depreciation as technology advances—precluding long-term reproducibility.

Despite the promising results obtained in this study, there are a number of opportunities for future improvement. One such avenue for future research involves investigating additional LLM architectures, models, and a larger hyperparameter search. We tested a limited set of adjective prefixes to fine-tune the models (i.e., 'creative,' 'novel,' 'useful,' 'surprising,' and 'unique'); however, this is hardly an exhaustive list. Perhaps the models would perform even better with a different adjective or a different sentence frame altogether. The quantified differences between the effect of these prompts on the model performance has also yet to be investigated; in the present study we treated the prefix as a hyperparameter, but future work could compare the effect of prompt directly.

As is often the case in machine learning, increasing the volume of training data and/or increasing the diversity of metaphor tasks the model is trained on could enhance model generalizability and yield more robust representations of creative metaphors. Another opportunity for improvement could be improving the baseline; DSI itself could be improved or an alternate baseline better suited for figurative language could be utilized. Additionally, this work could be limited by the quality of human ratings of metaphor creativity. The datasets used to train the LLMs had various intraclass correlations and the rater disagreement could translate to the model. Moreover, this work would benefit from a measure of uncertainty in the model outputs. Recently, researchers have developed methods to generate uncertainty measures in LLMs that specialize in natural language generation (Lin et al., 2023). By extending this work to automated assessment ratings, researchers could gain confidence in the model predictions and identify which responses may need to be rated by a human. For example, any new words would not be understood by the LLM and as a result, would not be accurately scored by the model and an uncertainty rating could help identify such an issue.

Future research can also investigate more thoroughly the shortcomings of LLM's using this approach. From our research, it is unclear how context-dependent the model is when generating creativity ratings. Human raters are able to incorporate the prompt with their understanding of the world when evaluating responses, while the process of the LLM is not fully clear. As a result, when predicting the creativity for a novel prompt, the model may be biased towards highly-creative responses from the training data. Additional investigation is required to understand the role of context in automating the scoring of novel prompts.

An outstanding question that arises from this study pertains to the applicability of LLMs to score metaphors in languages other than English. Since these models are primarily pre-trained on English corpora, their effectiveness for scoring figurative language should be expected to differ across languages (with likely performance degradation). Understanding how these models perform under different linguistic contexts, and how to rectify cases of underperformance, is necessary to ensure optimal performance and to improve the accessibility of these tools beyond the English language. Recent work has explored semantic distance scoring of the AUT via multilingual LLMs (Patterson, Merseal, et al., 2023) and showed encouraging performance across 12 different languages in terms of predicting human ratings.

In this study, we found the fine-tuned LLMs both predicted metaphor creativity ratings better than word count alone, suggesting that the models are attending to something more than just elaboration of the responses. Although, LLMs still may be victim to elaboration bias despite outperforming the correlation with word count (Forthmann et al., 2019). Further research is needed to identify what exactly the model is attending to. For example, the models could be making an association between responses with higher frequency and lower human ratings of creativity. By examining the frequency of types of responses, beyond semantic distance, we could begin to understand how the models are able to predict the creativity of figurative language. Upon calculating the MPE in the held-out test and held-out prompt datasets, we determined that both models had higher residuals for the responses that were rated two standard deviations away from the mean. This observation suggests that the models were overestimating the ratings of least creative responses and underestimating the ratings of the most creative responses. This finding has been demonstrated previously with the automated scoring of figural creativity tasks (cf. Patterson, Barbot, et al., 2023). A lack of data at these margins could be the result of these less accurate model predictions.

Another aspect of this study worth further exploring is the distinction between metaphor creativity and metaphor aptness: the degree to which a metaphor is semantically coherent and meaningful given the context. The human raters in our study scored metaphors based on creativity, but there is a long history of research on metaphor aptness (Blasko & Connine, 1993; Holyoak & Stamenković, 2018; Katz, 1989, 1989; Stamenković et al., 2023; Tourangeau & Sternberg, 1981). Since aptness improves the comprehension of metaphors in humans (Stamenković et al., 2023), future research could explicitly include aptness in the human ratings of metaphor creativity to see if that construct improves the automated scoring of metaphors in LLMs. One consideration regarding aptness is whether creativity ratings already encompass aptness to some extent. Although the raters in this study were instructed to rate metaphors for creativity/originality, they may have tacitly considered their aptness (or "usefulness"), which could account for some of the variance in ratings.

Conclusion

In this study, we introduced a novel approach to scoring the metaphor generation task using fine-tuned LLMs and compared it to two baselines: word count/elaboration and semantic distance (i.e., DSI; Johnson et al., 2022). The results showed that LLM-generated creativity scores correlated strongly with human creativity ratings—far exceeding both baselines. These results highlight the ability of LLMs to accurately score products from complex and abstract creative thinking tasks that go beyond the alternate uses task. Our research offers a reliable and efficient alternative to labor-intensive and subjective human ratings—improving the reproducibility and scalability of creative assessment. We provide yet another advancement towards the automation of creativity assessment, empowering researchers to understand and improve creative thinking.

References

Acar, S., Dumas, D., Organisciak, P., & Berthiaume, K. (2023). *Measuring original thinking in elementary school: Development and validation of a computational psychometric approach*. https://doi.org/10.13140/RG.2.2.19804.56968

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2623–2631. https://doi.org/10.1145/3292500.3330701

Bakker, M., Chadwick, M., Sheahan, H., Tessler, M., Campbell-Gillingham, L., Balaguer, J., McAleese, N., Glaese, A., Aslanides, J., Botvinick, M., & Summerfield, C. (2022). Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in Neural Information Processing Systems*, *35*, 38176–38189. https://doi.org/10.48550/arXiv.2211.15006

Beaty, R. E., & Johnson, D. R. (2021). Automating creativity assessment with SemDis: An open platform for computing semantic distance. *Behavior Research Methods*, *53*(2), 757–780. https://doi.org/10.3758/s13428-020-01453-w

Beaty, R. E., Johnson, D. R., Zeitlen, D. C., & Forthmann, B. (2022). Semantic Distance and the Alternate Uses Task: Recommendations for Reliable Automated Assessment of Originality. *Creativity Research Journal*, *34*(3), 245–260. https://doi.org/10.1080/10400419.2022.2025720

Beaty, R. E., & Silvia, P. J. (2013). Metaphorically speaking: Cognitive abilities and the production of figurative language. *Memory & Cognition*, *41*(2), 255–267. https://doi.org/10.3758/s13421-012-0258-5

Beaty, R. E., Silvia, P. J., & Benedek, M. (2017). Brain networks underlying novel metaphor production. *Brain and Cognition*, *111*, 163–170. https://doi.org/10.1016/j.bandc.2016.12.004

Billow, R. M. (1977). Metaphor: A review of the psychological literature. *Psychological Bulletin*, *84*(1), 81–92. https://doi.org/10.1037/0033-2909.84.1.81

Blasko, D. G., & Connine, C. M. (1993). Effects of familiarity and aptness on metaphor processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(2), 295–308. https://doi.org/10.1037/0278-7393.19.2.295

Buczak, P., Huang, H., Forthmann, B., & Doebler, P. (2023). The Machines Take Over: A Comparison of Various Supervised Learning Approaches for Automated Scoring of Divergent Thinking Tasks. *The Journal of Creative Behavior*, *57*(1), 17–36. https://doi.org/10.1002/jocb.559

Carbonell, J. G. (1982). Metaphor: An Inescapable Phenomenon in Natural-Language Comprehension. In *Strategies for Natural Language Processing*. Psychology Press.

Cardillo, E. R., Watson, C. E., Schmidt, G. L., Kranjec, A., & Chatterjee, A. (2012). From novel to familiar: Tuning the brain for metaphors. *Neuroimage*, *59*(4), 3212–3221. https://doi.org/10.1016/j.neuroimage.2011.11.079

Caucheteux, C., Gramfort, A., & King, J.-R. (2022). Deep language algorithms predict semantic comprehension from brain activity. *Scientific Reports*, *12*(1), Article 1. https://doi.org/10.1038/s41598-022-20460-9

Chakrabarty, T., Choi, Y., & Shwartz, V. (2022). It’s not Rocket Science: Interpreting Figurative Language in Narratives. *Transactions of the Association for Computational Linguistics*, *10*, 589–606. https://doi.org/10.1162/tacl_a_00478

Clevenger, T., & Edwards, R. (1988). Semantic distance as a predictor of metaphor selection. *Journal of Psycholinguistic Research*, *17*(3), 211–226. https://doi.org/10.1007/BF01686356

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. https://doi.org/10.48550/arXiv.1810.04805

Dumas, D., & Dunbar, K. N. (2014). Understanding Fluency and Originality: A latent variable perspective. *Thinking Skills and Creativity*, *14*, 56–67. https://doi.org/10.1016/j.tsc.2014.09.003

Dumas, D., Organisciak, P., & Doherty, M. (2020). Measuring Divergent Thinking Originality with Human Raters and Text-Mining Models: A Psychometric Comparison of Methods. *Psychology of Aesthetics Creativity and the Arts*. https://doi.org/10.1037/aca0000319

Dumas, D., Organisciak, P., Maio, S., & Doherty, M. (2021). Four Text-Mining Methods for Measuring Elaboration. *The Journal of Creative Behavior*, *55*(2), 517–531. https://doi.org/10.1002/jocb.471

Fan, L., Zhuang, K., Wang, X., Zhang, J., Liu, C., Gu, J., & Qiu, J. (2023). Exploring the behavioral and neural correlates of semantic distance in creative writing. *Psychophysiology*, *60*(5), e14239. https://doi.org/10.1111/psyp.14239

Ferreira, F., & Patson, N. D. (2007). The 'Good Enough' Approach to Language Comprehension. *Language and Linguistics Compass*, *1*(1–2), 71–83. https://doi.org/10.1111/j.1749-818X.2007.00007.x

Ferreira, F., & Qiu, Z. (2021). Predicting syntactic structure. *Brain Research*, *1770*, 147632. https://doi.org/10.1016/j.brainres.2021.147632

Forthmann, B., Beaty, R. E., & Johnson, D. R. (2022). Semantic Spaces Are Not Created Equal – How Should We Weigh Them in the Sequel? *European Journal of Psychological Assessment*. https://doi.org/10.1027/1015-5759/a000723

Forthmann, B., & Doebler, P. (2022). Fifty years later and still working: Rediscovering Paulus et al's (1970) automated scoring of divergent thinking tests. *Psychology of Aesthetics, Creativity, and the Arts*, No Pagination Specified-No Pagination Specified. https://doi.org/10.1037/aca0000518

Forthmann, B., Oyebade, O., Ojo, A., Günther, F., & Holling, H. (2019). Application of latent semantic analysis to divergent thinking is biased by elaboration. *The Journal of Creative Behavior*, *53*(4), 559–575. https://doi.org/10.1002/jocb.240

Gibbs, R. W. (1990). *The Process of Understanding Literary Metaphor*. *19*(2), 65–79. https://doi.org/10.1515/jlse.1990.19.2.65

Gibbs, R. W. (1994). *The Poetics of Mind: Figurative Thought, Language, and Understanding*. Cambridge

University Press.

Glucksberg, S., & McGlone, M. S. (2001). *Understanding Figurative Language: From Metaphor to Idioms*. Oxford University Press, USA.

Glucksberg, S., McGlone, M. S., & Manfredi, D. (1997). Property Attribution in Metaphor Comprehension. *Journal of Memory and Language*, *36*(1), 50–67. https://doi.org/10.1006/jmla.1996.2479

Green, A. E. (2016). Creativity, Within Reason: Semantic Distance and Dynamic State Creativity in Relational Thinking and Reasoning. *Current Directions in Psychological Science*, *25*(1), 28–35. https://doi.org/10.1177/0963721415618485

Guilford, J. P. (1967). *The nature of human intelligence*. McGraw-Hill.

Hass, R. W. (2017). Tracking the dynamics of divergent thinking via semantic distance: Analytic methods and theoretical implications. *Memory & Cognition*, *45*(2), 233–244. https://doi.org/10.3758/s13421-016-0659-y

Heinen, D. J. P., & Johnson, D. R. (2018). Semantic distance: An automated measure of creativity that is novel and appropriate. *Psychology of Aesthetics, Creativity, and the Arts*, *12*(2), 144–156. https://doi.org/10.1037/aca0000125

Holyoak, K. J., & Stamenković, D. (2018). Metaphor comprehension: A critical review of theories and evidence. *Psychological Bulletin*, *144*(6), 641–671. https://doi.org/10.1037/bul0000145

Ichien, N., Stamenković, D., & Holyoak, K. J. (2023). *Large Language Model Displays Emergent Ability to Interpret Novel Literary Metaphors* (arXiv:2308.01497). arXiv. https://doi.org/10.48550/arXiv.2308.01497

Jawahar, G., Sagot, B., & Seddah, D. (2019). What Does BERT Learn about the Structure of Language? *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3651–3657. https://doi.org/10.18653/v1/P19-1356

Jia-li, D. U., & Ping-fang, Y. U. (2013). A computational linguistic approach to natural language processing with applications to garden path sentences analysis. *International Journal of Advanced Computer Science and Applications (IJACSA)*, *3*(9), Article 9. https://doi.org/10.14569/IJACSA.2012.030909

Johnson, D. R., Kaufman, J. C., Baker, B. S., Patterson, J. D., Barbot, B., Green, A. E., van Hell, J., Kennedy, E., Sullivan, G. F., Taylor, C. L., Ward, T., & Beaty, R. E. (2022). Divergent semantic integration (DSI): Extracting creativity from narratives with distributional semantic modeling. *Behavior Research Methods*. https://doi.org/10.3758/s13428-022-01986-2

Kasirer, A., & Mashal, N. (2018). Fluency or Similarities? Cognitive Abilities that Contribute to Creative Metaphor Generation. *Creativity Research Journal*, *30*(2), 205–211. https://doi.org/10.1080/10400419.2018.1446747

Katz, A. N. (1989). On choosing the vehicles of metaphors: Referential concreteness, semantic distances, and individual differences. *Journal of Memory and Language*, *28*(4), 486–499. https://doi.org/10.1016/0749-596X(89)90023-5

Katz, A. N. (1992). Psychological Studies in Metaphor Processing: Extensions to the Placement of Terms in Semantic Space. *Poetics Today*, *13*(4), 607–632. https://doi.org/10.2307/1773291

Katz, A. N., Paivio, A., Marschark, M., & Clark, J. (1988). Norms for 204 Literary and 260 Nonliterary Metaphors on 10 Psychological Dimensions. *Metaphor and Symbol - METAPHOR SYMB*, *3*, 191–214. https://doi.org/10.1207/s15327868ms0304_1

Kenett, Y. N. (2019). What can quantitative measures of semantic distance tell us about creativity? *Current Opinion in Behavioral Sciences*, *27*, 11–16. https://doi.org/10.1016/j.cobeha.2018.08.010

Lakoff, G., & Johnson, M. (2008). *Metaphors We Live By*. University of Chicago Press.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2–3), 259–284. https://doi.org/10.1080/01638539809545028

Lantz, B. (2013). *Machine Learning with R* (3rd ed.). Packt Publishing.

https://www.packtpub.com/product/machine-learning-with-r-third-edition/9781788295864

Lin, Z., Trivedi, S., & Sun, J. (2023). *Generating with Confidence: Uncertainty Quantification for Black-box Large Language Models* (arXiv:2305.19187). arXiv. https://doi.org/10.48550/arXiv.2305.19187

Liu, E., Cui, C., Zheng, K., & Neubig, G. (2022). Testing the Ability of Language Models to Interpret Figurative Language. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4437–4452. https://doi.org/10.18653/v1/2022.naacl-main.330

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (arXiv:1907.11692). arXiv. https://doi.org/10.48550/arXiv.1907.11692

Loshchilov, I., & Hutter, F. (2019). *Decoupled Weight Decay Regularization* (arXiv:1711.05101; Version 3). arXiv. https://doi.org/10.48550/arXiv.1711.05101

Organisciak, P., Acar, S., Dumas, D., & Berthiaume, K. (2023). Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. *Thinking Skills and Creativity*, *49*, 101356. https://doi.org/10.1016/j.tsc.2023.101356

Parde, N., & Nielsen, R. (2018, May). A Corpus of Metaphor Novelty Scores for Syntactically-Related Word Pairs. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. LREC 2018, Miyazaki, Japan. https://aclanthology.org/L18-1243

Patterson, J. D., Barbot, B., Lloyd-Cox, J., & Beaty, R. E. (2023). AuDrA: An automated drawing assessment platform for evaluating creativity. *Behavior Research Methods*. https://doi.org/10.3758/s13428-023-02258-3

Patterson, J. D., Merseal, H. M., Johnson, D. R., Agnoli, S., Baas, M., Baker, B. S., Barbot, B., Benedek, M., Borhani, K., Chen, Q., Christensen, J. F., Corazza, G. E., Forthmann, B., Karwowski, M., Kazemian, N., Kreisberg-Nitzav, A., Kenett, Y. N., Link, A., Lubart, T., … Beaty, R. E. (2023). Multilingual semantic distance: Automatic verbal creativity assessment in many languages. *Psychology of Aesthetics, Creativity, and the Arts*, *17*(4), 495–507. https://doi.org/10.1037/aca0000618

Paulus, D. H., Renzulli, J. S., & Archambault, F. X. (1970). Computer simulation of human ratings of creativity. *Final Report*, *(No. 9-A-032)*. https://files.eric.ed.gov/fulltext/ED060658.pdf

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners*. https://api.semanticscholar.org/CorpusID:160025533

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*(45), e2105646118. https://doi.org/10.1073/pnas.2105646118

Shutova, E., Teufel, S., & Korhonen, A. (2013). Statistical Metaphor Processing. *Computational Linguistics*, *39*(2), 301–353. https://doi.org/10.1162/COLI_a_00124

Silvia, P. J., & Beaty, R. E. (2012). Making creative metaphors: The importance of fluid intelligence for creative thought. *Intelligence*, *40*(4), 343–351. https://doi.org/10.1016/j.intell.2012.02.005

Silvia, P. J., Winterstein, B. P., Willse, J. T., Barona, C. M., Cram, J. T., Hess, K. I., Martinez, J. L., & Richard, C. A. (2008). Assessing creativity with divergent thinking tasks: Exploring the reliability and validity of new subjective scoring methods. *Psychology of Aesthetics, Creativity, and the Arts*, *2*(2), 68–85. https://doi.org/10.1037/1931-3896.2.2.68

Stamenković, D., Milenković, K., Ichien, N., & Holyoak, K. J. (2023). An Individual-Differences Approach to Poetic Metaphor: Impact of Aptness and Familiarity. *Metaphor and Symbol*, *38*(2), 149–161. https://doi.org/10.1080/10926488.2021.2006046

Stevenson, C., Smal, I., Baas, M., & Grasman, R. (2020). *Putting GPT-3's Creativity to the (Alternative Uses) Test*. https://doi.org/10.48550/arXiv.2206.08932

Taylor, J. M., & Mazlack, L. J. (2004). *Computationally Recognizing Wordplay in Jokes*.

Tiberius, R. G. (1986). Metaphors Underlying the Improvement of Teaching and Learning. *British Journal of Educational Technology*, *17*(2), 144–156. https://doi.org/10.1111/j.1467-8535.1986.tb00504.x

Tourangeau, R., & Sternberg, R. J. (1981). Aptness in metaphor. *Cognitive Psychology*, *13*(1), 27–55. https://doi.org/10.1016/0010-0285(81)90003-7

Veale, T., Shutova, E., & Klebanov, B. B. (2016). *Metaphor: A Computational Perspective*. Morgan & Claypool Publishers.

Winter, B., & Strik-Lievers, F. (2023). Semantic distance predicts metaphoricity and creativity judgments in synesthetic metaphors. *Metaphor and the Social World*, *13*(1), 59–80. https://doi.org/10.1075/msw.00029.win

Yu, Y., Beaty, R. E., Forthmann, B., Beeman, M., Cruz, J. H., & Johnson, D. (2023). A MAD method to assess idea novelty: Improving validity of automatic scoring using maximum associative distance (MAD). *Psychology of Aesthetics, Creativity, and the Arts*. https://doi.org/10.1037/aca0000573

Zhou, Z.-H. (2021). *Machine Learning*. Springer. https://doi.org/10.1007/978-981-15-1967-3