Automated Assessment of Creativity in Multilingual Narratives

Simone A. Luchini[1], Muhammad Moosa[1], John D. Patterson[1], Dan Johnson[2], Matthijs Baas[3], Baptiste Barbot[4], Iana Bashmakova[5], Mathias Benedek[6], Qunlin Chen[7], Giovanni E. Corazza[8], Boris Forthmann[9], Benjamin Goecke[10], Sameh Ibrahim[4], Maciej Karwowski[11], Yoed N. Kenett[12], Izabela Lebuda[11], Todd Lubart[13], Kirill G. Miroshnik[14], Felix-Kingsley Obialo[15], Marcela Ovando-Tellez[16], Ricardo Primi[17], Rogelio Puente[18], Claire Stevenson[3], Emmanuelle Volle[16], Aleksandra Zielińska[11], Janet G. van Hell[1], Wenpeng Yin[1], & Roger E. Beaty[1]

[1] Pennsylvania State University, USA

[2] Washington and Lee University, USA

[3] University of Amsterdam, Netherlands

[4] UCLouvain, Belgium

[5] University of British Columbia, Canada

[6] University of Graz, Austria

[7] Southwest University, China

[8] University of Bologna, Italy

[9] University of Münster, Germany

[10] University of Tübingen, Germany

[11] University of Wroclaw, Poland

[12] Technion - Israel Institute of Technology, Israel

[13] Université Paris Cité and Univ Gustave Eiffel, LaPEA, Boulogne-Billancourt, France,

[14] Saint Petersburg State University, Russia

[15] Dominican University Ibadan, Nigeria

[16] Sorbonne University, Paris Brain Institute, INSERM, France

[17] Universidade São Francisco, Brazil

[18] Universidad Anáhuac México Norte, Mexico

Abstract

Researchers and educators interested in creative writing need a reliable and efficient tool to score the creativity of narratives, such as short stories. Typically, human raters manually assess narrative creativity, but such subjective scoring is limited by labor costs and rater disagreement. Large language models (LLMs) have shown remarkable success on creativity tasks, yet they have not been applied to scoring narratives, including multilingual stories. In the present study, we aimed to test whether narrative originality—a component of creativity—could be automatically scored by LLMs, further evaluating whether a single LLM could predict human originality ratings across multiple languages. We trained three different LLMs to predict the originality of short stories written in 11 languages. Our first monolingual model, trained only on English stories, robustly predicted human originality ratings (r = .81). This same model—trained and tested on multilingual stories translated into English—strongly predicted originality ratings of multilingual narratives ($r \geq .73$). Finally, a multilingual model trained on the same stories, in their original language, reliably predicted human originality scores across all languages ($r \geq .72$). We thus demonstrate that LLMs can successfully score narrative creativity in 11 different languages, surpassing the performance of the best previous automated scoring techniques (e.g., semantic distance). This work represents the first effective, accessible, and reliable solution for the automated scoring of creativity in multilingual narratives.

Keywords: Automated Scoring; Creativity Assessment; Creative Writing; Narratives; Large Language Models

Multilingual Automated Assessment of Creativity in Narratives

Evaluating the creativity of human-generated narratives has been of consistent interest to researchers. Creativity assessments of narratives generally involve the combined judgment of several human raters such as in the Consensual Assessment Technique (CAT; Amabile, 1982). Typically, humans tend to agree on the underlying creativity of a narrative (D'Souza, 2021; Kaufman et al., 2013; Mozaffari, 2013; Vaezi & Rezaei, 2019). Nevertheless, any assessment relying on human ratings will inevitably face certain limitations associated with human labor. These include, but are not limited to: high costs in terms of resources and effort (human raters will generally require training and payment), occasional disagreements between raters, and a slow pace of research (Barbot, 2018; Forthmann et al., 2017; Reiter-Palmon et al., 2019).

Recently, computational tools have been employed to automate the creativity scoring of narratives, leading to methods that are free from the limitations associated with human rating (Johnson et al., 2023; Zedelius et al., 2019). However, to date, no attempt has been made to train large language models (LLMs; Conneau et al., 2019; Devlin et al., 2018; Liu et al., 2019; Radford et al., 2018; Vaswani et al., 2017) to predict the creativity assessment of narratives—a naturalistic form of creative expression central to literary creativity—even though trained LLMs have demonstrated a strong performance in scoring laboratory-based creativity tasks, such as with the alternate uses task (Organisciak et al., 2023; Yang et al., 2023; Zielińska et al., 2023), creative problem-solving task (Luchini et al., 2023), alternate questions task (Raz et al., 2024), scientific creative thinking task (Goecke et al., 2024), and metaphor generation task (DiStefano et al., 2023). Further, previous attempts at automatically rating the originality of narratives have only focused on English stories (Johnson et al., 2023; Toubia et al., 2021), overlooking whether such methods could be used in the scoring of multilingual data. In the present paper, we trained a

series of LLMs to rate the originality of narratives on a multilingual dataset of short stories in 11 languages, with the goal of providing an efficient, accurate, and accessible tool to assess narrative creativity for the education and research communities.

Human Scoring of Narrative Creativity

Evaluating narrative creativity requires tools to distinguish between more and less creative stories. This task becomes particularly demanding in situations where, given a large sample size, researchers find themselves having to rate the creativity of hundreds or thousands of stories. The most common approach for rating narrative creativity involves the use of human raters, including the Consensual Assessment Technique (CAT). The CAT involves recruiting "experts" in a given field to judge the creativity of products or ideas (Amabile, 1982). Although useful in demonstrating the reliability of subjective scoring (e.g., Hennessey, 1994), the approach was criticized for the often costly nature of recruiting domain experts, and the lack of guidance given to such raters (Kaufman et al., 2008). Building on the CAT, the subjective scoring method extended its use to trained non-expert raters (Silvia et al., 2008). Thus, up until recently, most creativity researchers depended on trained research assistants to meet their needs for creativity assessment.

Subjective scoring techniques, while effective, nevertheless carry their own sets of limitations. One of the most notable issues with subjective scoring is the labor cost associated with human raters. Recruiting and retaining a team of trained raters can often be hard and resource intensive, and re-training a new team of raters can prove costly and impractical (Benedek et al., 2013). Moreover, qualified human raters may not always be available to researchers, or may be inaccessible due to lack of funding for their remuneration. A second issue

with subjective scoring is that of rater disagreement, as raters do not always agree on what is or is not creative (Ceh et al., 2022; Forthmann et al., 2017). Finally, the diversity in training approaches employed by different research groups introduces its own challenge, making it hard to compare or compile data from different studies. To overcome these issues, creativity researchers have been exploring automated methods for scoring creative responses, including unsupervised machine learning and supervised training of large language models.

Automated Creativity Scoring with Semantic Distance

Unsupervised machine learning methods—requiring no human involvement—have recently been applied to automatically score the originality of divergent thinking tasks (Beaty et al., 2022; Beaty & Johnson, 2021; Chrysikou et al., 2021; Dumas et al., 2021; Murray et al., 2021; Patterson et al., 2023), creative word association tasks (Beaty et al., 2021; Gray et al., 2019; Heinen & Johnson, 2018; Johnson et al., 2021; Merseal et al., 2023), remote associates test (Beisemann et al., 2019; Smith & Vul, 2015), narratives (Johnson et al., 2023), and sentences more broadly (Weinstein et al., 2022). The most common unsupervised approach has been using computational models of semantic distance (Günther et al., 2019; Mandera et al., 2017).

Semantic distance relates to the associative theory of creativity, which holds that highly original ideas will tend to arise from the combination of semantically "distant" concepts (Beaty & Kenett, 2023; Bendetowicz et al., 2018; Kenett, 2019; Mednick, 1962; Ovando-Tellez et al., 2023). Thus, the dissimilarity between meanings of words or concepts should determine the originality of the ideas they compose. By compiling the co-occurrence frequencies for a given word in natural language, one can generate a word embedding—a vector representation of a word in high-dimensional space—which can then be employed in mathematical operations such

as the comparison of word meanings (Lund & Burgess, 1996). Other ways of generating word embeddings have since been developed, such as with latent semantic analysis (LSA; Deerwester et al., 1990) or through computational models including Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), or any LLM (Reimers & Gurevych, 2020). Semantic distance scores extract the dissimilarity between the meanings of two words from their word embeddings (Beaty & Johnson, 2021).

Semantic distance scores have been shown to predict human ratings of originality for responses to the alternate uses task (AUT), a divergent thinking task where participants are required to come up with unusual uses for everyday objects (Acar & Runco, 2019; Beaty et al., 2023; Dickson et al., 2022; Dumas & Runco, 2018; Forthmann et al., 2023; Maio et al., 2020; Patterson et al., 2023). For a single AUT response, a semantic distance score is computed by taking the converse cosine angle between the word embedding of the AUT object (e.g., box) and each word contained in the participant's response (e.g., seat cushion), separately. The semantic distance is a composite score generated from all pairwise comparisons, or the largest converse cosine angle amongst all comparisons with the object (Yu et al., 2023). While semantic distance scores are suitable for short-form responses such as with the AUT, they may not be appropriate to determine the originality of long-form responses, such as short stories.

To this purpose, divergent semantic integration (DSI) was developed as a method to leverage semantic distance scores for the originality assessment of long-form responses (Johnson et al., 2023). A DSI score is the average of all semantic distances computed between all possible pairs of words in a response. DSI uses unsupervised word embeddings that are sensitive to context, which is particularly useful to address cases of polysemy—words with multiple meanings depending on the surrounding context (e.g., computer-server, restaurant-server). The

increased sensitivity to context makes DSI a better method for evaluating the originality of long-form responses than previous approaches relying on superficial linguistic features (Zedelius et al., 2019), or other non-context dependent semantic distance approaches (e.g., LSA; Jawahar et al., 2019). Johnson et al. (2023) found that DSI scores robustly predicted human-rated originality scores of short stories, explaining 72% of the variance for human ratings. Further, DSI retained strong predictions even with responses from non-native English speakers, compared to those from native English speakers, indicating that DSI can generalize across individuals with varying linguistic backgrounds.

Automated Creativity Scoring with Language Models

More recently, LLMs have been applied to the scoring of creative tasks (DiStefano et al., 2023; Luchini et al., 2023; Organisciak et al., 2023; Raz et al., 2024). LLMs are deep neural networks which undergo extensive pre-training procedures. Pre-training is an unsupervised learning process (i.e., without human feedback) wherein the LLM is exposed to large amounts of text data, allowing the model to develop representations of language and outperform other natural language processing (NLP) tools even on tasks it was never trained on (Vaswani et al., 2017). The typical pre-training procedure involves automated pattern detection, wherein the LLM learns from unlabeled text data—text without any tag or number for the model to predict—via word prediction problems, requiring the model to predict a missing word from the context or vice versa (Jiang et al., 2020). Pre-training can then be supplemented with a process termed fine tuning—additional training where human input guides the models' predictions through supervised learning. Fine-tuning serves to specialize the model to perform a specific task, such as creativity scoring, adjusting the internal weights (i.e., parameters) of the model by exposing it to several labeled examples (e.g., AUT responses with human originality ratings). Indeed, fine-

tuned LLMs have displayed impressive performance on a variety of creativity tasks, compared to previous automated scoring methods which only used static word embeddings (DiStefano et al., 2023; Luchini et al., 2023; Organisciak et al., 2023; Raz et al., 2024).

For instance, LLMs surpassed semantic distance when predicting the human originality ratings of AUT responses (Organisciak et al., 2023). After fine-tuning a series of LLMs on 27,217 AUT responses, Organisciak and colleagues found the best performing model to be GPT-3, which could robustly predict the human originality ratings of new responses (r = .81). Strong performance (r > .75) was achieved even after reducing the number of responses (to 6,000) shown to the model during fine-tuning. The model also performed strongly (r = .70) after in-context learning—providing a small number of examples to the model as part of the input, without any adjustment of model parameters—with only 20 example responses, and no fine-tuning. Researchers evaluated the performance of the fine-tuned model against that achieved by semantic distance scores, which were found to poorly correlate with human originality ratings at the response level (r = .20). Of note, response level correlations are a stricter test of human-rated originality prediction than participant level correlations—involving the prediction of a single score per participant. As such, this work indicated that LLM fine-tuning is indeed a viable technique for the automated scoring of creativity. Nevertheless, the best performing model was found to be GPT-3, a large closed-source LLM. Of note, such closed-source models have limited accessibility, due to costs and limitations associated with their use. When employing an open-sourced model, T5-base, the model displayed weaker but nevertheless robust predictions of human-rated originality scores (r = .75).

Extending such work, smaller LLMs than those employed in Organisciak et al. (2023), RoBERTa-base (Liu et al., 2019) and GPT-2 (Radford et al., 2018), have been shown to reliably

predict both originality and quality (i.e., effectiveness) scores of responses to the creative problem-solving task (Luchini et al., 2023). There are several benefits of smaller-sized, open-weight models (such as RoBERTa-base or GPT-2) for promoting accessibility, reproducibility, and data protection. For one, by virtue of their open-weight nature, the models do not incur costs, in contrast to proprietary models such as GPT-3 or Claude that charge for usage—presenting a potential financial barrier. Second, because the models are downloaded locally, they can be used anywhere on a standard laptop, even in settings without an internet connection. Third, because users have control of the model weights, results are reproducible; closed-source models are often changed or discontinued without notice (e.g., GPT-3), precluding reproducibility. Finally, because smaller open-weight models can be used without uploading participant data to the cloud, they offer enhanced data protection and are more likely to be compatible in localities with data restrictions. Luchini et al. (2023) reported that, after training on only 2,000 responses, both LLMs performed better when predicting quality scores (r = .83) than when predicting originality scores (r = .79 ~ .80). The same LLMs have then been reported to successfully predict the originality of metaphors after having been trained on a similarly small dataset (r = .70 ~ .72; DiStefano et al., 2023). Further, RoBERTa-base has also been shown to predict the complexity of responses to a question-asking task (r = .80; Raz et al., 2024); in this case, complexity was derived from human-rated Bloom taxonomy scores, a metric that correlates with the originality of such questions (Raz et al., 2023). Together, these studies indicate that smaller LLMs, trained on datasets of only a few thousand responses, have a broad applicability to the automated scoring of a variety of creativity tasks involving longer responses and figurative language.

Multilingual Automated Creativity Scoring

In an attempt to increase the accessibility of unsupervised automated creativity scoring methods, researchers have recently attempted a multilingual validation of semantic distance scores across 12 languages (Patterson et al., 2023). They gathered 102,672 AUT responses in multiple languages and computed semantic distance scores for each individual response. Multilingual LLMs were used to extract word embeddings for the calculation of semantic distance scores. Multilingual LLMs possess the same architecture as monolingual models, as they achieve their multilingual capabilities by being pretrained on data from multiple different languages. Medium to large correlations were observed between the semantic distance scores and the human-rated creativity scores at the participant level on seven out of twelve languages ($r \geq .30$). Researchers further noted that an English bias did not emerge in their results, with the model performing roughly the same between the other languages as with English.

In a previous attempt at scoring non-English creativity responses with LLMs, researchers employed a monolingual LLM (GPT-3), trained only on English responses, and tested how well it predicted the originality of Polish AUT responses that were translated to English (Zielińska et al., 2023). They reported medium to large latent correlations between model predictions and human-rated originality, depending on which AUT prompt was analyzed ($r = .56 \sim .95$). Translating responses to English allowed the authors to employ a monolingual model for non-English data, providing an alternative to training a new model on the non-English data. However, it remains unclear whether this approach could be employed for languages other than Polish, such as ideographic languages which employ a different writing system than alphabetic languages such as Polish and English. It is also possible that, due to translating errors, the meaning of responses may have been altered after translation. Finally, it should be noted that researchers employed a large closed source model, GPT-3, leaving an open question as to

whether smaller models can achieve the same performance on such tasks. It thus remains unclear whether translating multilingual responses to English is the best approach to increase the accessibility of automated creativity scoring tools. One could train a multilingual model to predict the originality of responses in their original language, without the need for translation. In one such attempt, it was shown that XLM-RoBERTa (Conneau et al., 2019), a smaller-sized multilingual LLM, predicted the creativity of responses to a German scientific creative thinking task (r = .81; Goecke et al., 2024).

Nevertheless, all such works considered a single language in their analyses, thus facing the same limitation of other monolingual investigations of automated creativity assessment in English language. In order to determine whether LLMs can handle multiple languages—without biases or significant reductions in accuracy—further work is required to extend these findings to other languages.

The Present Research

Evaluations of narrative creativity have typically relied on human raters. Extensive research has validated this method, demonstrating good inter-rater agreement on judgments of narrative creativity (D'Souza, 2021; Kaufman et al., 2013; Mozaffari, 2013; Vaezi & Rezaei, 2019). Nevertheless, such methods are cost-intensive, relying on the recruitment, training, and labor of research assistants (Benedek et al., 2013; Forthmann et al., 2023). Thanks in part to recent advancements in natural language processing, automated methods for scoring narrative creativity may replace the need for human raters (Johnson et al., 2023). Yet no work to date has validated the use of fine-tuned LLMs, some of the more powerful models in natural language processing, for the purpose of narrative creativity scoring. Given past work indicating the

superior performance of such models in scoring other creativity tasks (DiStefano et al., 2023; Luchini et al., 2023; Organisciak et al., 2023; Raz et al., 2024), LLMs present a promising avenue for the creativity scoring of narratives. Additionally, researchers have yet to evaluate whether language models can be trained to score multilingual narratives. Ensuring that LLMs can score multilingual narratives will extend their accessibility to researchers from diverse backgrounds, who may study creativity across different languages than English.

The present research evaluates whether RoBERTa, a monolingual LLM, can be trained to successfully predict the creativity of English narratives, and multilingual narratives translated to English. Further, we test whether a multilingual LLM, XLM-RoBERTa, can successfully be trained to predict the creativity of narratives in their original languages. We include a LLM trained on English-translated data given the overrepresentation of English language responses in our training data, as well as reports that multilingual LLMs like XLM-RoBERTa typically demonstrate an English bias (Choudhury & Deshpande, 2021). Thus, an English-translated model may outperform any multilingual model in automated narrative scoring. Our trained models are further evaluated by comparing their predictions against semantic distance (DSI) and wordcount—the amount of words in a response—which have been shown to explain substantial variance in human creativity ratings (Johnson et al., 2023; Taylor & Barbot, 2021).

Methods

The datasets, models and analysis code are available online on OSF (https://osf.io/x79fm/?view_only=66c2c803c797467bb03bc53ce58d890f).

Datasets

English dataset. The largest dataset, ENG1, included 153 participants (68 male, 82 female, 3 non-binary; mean age = 38.62 years, age range = 22 ~ 70 years) who completed the English version of the five-sentence story task (FSST; Prabhakaran et al., 2013). Each participant provided a response to each of 7 prompts, amounting to 7 responses per participant and a total of 1,071 responses in the entire dataset. This dataset was extracted from study 2 of the Johnson et al. (2023) validation of DSI scores and included human ratings from 4 raters. A total of 79 responses for which ratings from 2 or more raters were missing were removed from the final dataset, leading to a final sample of 997 responses.

Multilingual dataset. The second dataset, our multilingual dataset, was collected specifically for this study and consisted of stories spanning 11 languages: Arabic, Chinese (Mandarin), Dutch, English, French, German, Hebrew, Italian, Polish, Russian, and Spanish. The multilingual dataset included between 140 and 152 responses for each language, for a single writing prompt, for a total of 1638 responses. All participants included in the multilingual dataset wrote a single short story and were fluent in the language in which they wrote their story.

Five-Sentence Story Task (FSST)

The FSST is a creative short story writing task and is a measure of verbal creativity (Prabhakaran et al., 2014). Participants were presented with a three-word prompt (e.g., *stamp-letter-send*) and were asked to write (type) a creative short-story that includes all words in the prompt and is five sentences long. The *pen-paper-write* prompt was used in the Johnson et al. (2023) validation study of DSI for scoring short stories. Coauthors of this manuscript, who speak the 11 languages represented in the study, reviewed all writing prompts used in Johnson et al. (2023) and agreed the *pen-paper-write* prompt was the most linguistically and culturally accessible. Trained human raters from each lab/language involved in the project evaluated the

creativity of stories in accordance with the subjective scoring method (Silvia et al., 2008), an approach derived from Amabile's (1982) consensual assessment technique. Evaluators used a scale ranging from 1 (very uncreative) to 5 (very creative) to rate each story. For the multilingual dataset, a different set of raters was recruited for each of the languages (Appendix C). Multilingual stories were thus scored in their original language by raters who were fluent in that language. All raters, across languages, were provided with the same rating instructions (Appendix F). The intra-class correlation (ICC; Shrout & Fleiss, 1979) between the raters for each language demonstrated a strong reliability (ICC's = .75 ~ .88) for the absolute agreement on the average ratings using a two-way random model (Appendix C).

Baseline Measures

DSI

We calculated DSI scores for all responses in the ENG1 dataset, as well as all English-translated responses in the multilingual dataset. DSI scores served as one of our baseline measures to evaluate LLM performance, given they are derived from unsupervised learning (i.e., the model does not see any human-rated scores or undergo any additional training). Here, we use the variant of DSI that best matched human creativity ratings in Johnson et al. (2023). Specifically, word embeddings were first extracted from layers 6 and 7 of BERT-large, a pre-trained LLM. BERT is a predecessor to RoBERTa-base and XLM-RoBERTa-base, which were employed for fine-tuning and later prediction of the ENG1 and multilingual datasets. BERT-large is the biggest version of this model, containing even more parameters than both RoBERTa-base and XLM-RoBERTa-base. The cosine of the angle between a pair of word embeddings gave us a measure of the similarity in the meanings of the two words (Salton et al., 1975). As

such, this value was subtracted from 1 to compute semantic distance scores. After all possible semantic distances were calculated between pairs of words in a response, the average of these scores was taken to derive the DSI score for the response.

Wordcount

We employed wordcount as a further baseline measure that could be applied to both the English and multilingual datasets. Given that DSI scores have only been validated in the English language (Johnson et al., 2023), wordcount served as our sole baseline measure for our untranslated multilingual dataset. For Chinese (Mandarin), an ideographic language, we instead employed a character count as mere wordcount would not apply to its writing system.

English Large Language Models

We ran two different fine-tuning procedures using RoBERTa-base, a monolingual English LLM. One fine-tuning procedure, involving our English-only model, included data from the ENG1 dataset and served to test whether an LLM could robustly predict the originality of English narratives, without any involvement of non-English data. Then, for our English-translated model, we included both the training data from the ENG1 dataset, as well as our English-translated multilingual dataset, in the fine-tuning procedure.

RoBERTa

We selected the RoBERTa-base model for both our English-only and English-translated models, as it was shown to perform well in the automated scoring of other creativity tasks (DiStefano et al., 2023; Luchini et al., 2023; Raz et al., 2024). RoBERTa is a pretrained LLM, a transformer model (see Vaswani et al., 2017) which underwent extensive self-supervised

learning, without any human labeling, during its pretraining procedure. RoBERTa-base is an advancement over the BERT model (Liu et al., 2019), which was introduced by Google in 2018 (Devlin et al., 2018), and possesses a larger architecture of 123 million parameters. Pretraining the RoBERTa-base model was conducted on an extensive corpus of English data, including collection of data from the Toronto BookCorpus (800M words), the English portion of Wikipedia (2,500M words), the CC-News English dataset (63M news articles), the OpenWebText dataset of Reddit posts, and the Stories dataset which contains a subset of the CommonCrawl repository.

English-Only Language Model

Our first fine-tuning procedure was used to develop our English-only model and select hyperparameter settings for all our fine-tuned models. This involved the RoBERTa-base model with an added regression head—the output layer of a model which transforms the high-dimensional output into a single continuous value—so the model would output predictions as a single originality score.

For our English-Only model, we only considered the data from the ENG1 dataset. The ENG1 dataset was randomly split into training, validation, and holdout-responses sets following a 70/10/20 ratio respectively. All responses associated with one of the three-word prompts (*year-week-embark*) were withheld from this split and assigned to a holdout-prompt set for later generalization testing. Thus, the holdout-prompt set contained responses to a prompt that was never shown to the model during training, to test whether the model could extend its predictions to this unseen prompt. This test is particularly useful in determining whether the model may be applicable to external data involving the same task but different prompts. The split led to 598

responses included in the training set, 83 in the validation set, 172 in the holdout-responses set, and 144 responses in the holdout-prompt set.

During training, the training set was inputted to the model which would then predict a single continuous creativity score for each response. The mean squared error between these predicted scores and the human creativity ratings would then dictate the degree of adjustment to the model parameters during each training episode. The validation set was instead used to guide hyperparameter tuning (i.e., adapting the model's settings to optimize predictive accuracy). That is, periodically during the training process the model would be tested on the validation set and, if unsatisfactory, changes to the model settings were made and the process was repeated. Only when performance on the validation set was satisfactory the model was used for prediction of ratings on the holdout-responses and -prompt set. In this way, the holdout sets allowed for a true measure of model generalizability to unseen data. Generalizability is an important criterion for the evaluation of LLM performance, as it indicates whether the model can be used for the prediction of novel data and story prompts not used in this analysis.

Prior to training our English-only model, and in accordance with best practices for tuning BERT-like models, we manually searched through an arbitrary hyperparameter space over the number of epochs, the learning rate, and the training batch size (Maity & Takahashi, 2021). The number of epochs reflects the number of complete training passes the model makes through the entire training dataset during training. In contrast, the learning rate influences the speed at which the model learns from the data, thereby affecting the degree of weight adjustment at each training step. Batch size denotes the number of responses the model receives corrective feedback for at a time. Training with larger batch sizes provides more stable estimates of error and faster training

times. Smaller batch sizes are more costly in terms of training time but introduce noise in the weight updates that can help the model find a better solution.

English-Translated Language Model

Our multilingual dataset was first repurposed into an English-translated dataset, by translating all the non-English responses into English. We automatically translated the multilingual narratives with the use of Google Translate, one of the best performing machine translation tools (Groves & Mundt, 2015). Translating narratives into English allows for a normally monolingual model to be employed with multilingual data, although this adds some potential noise depending on the accuracy of translations.

We thus fine-tuned our English-Translated model, again using the RoBERTa-base model with an appended regression head, except we combined the training set of the ENG1 data with the translated version of the multilingual dataset. While the training set from ENG1 was only employed in training, a k-fold validation technique with 5 splits was run over the multilingual dataset. A k-fold validation involves a more rigorous testing of model performance on the entire dataset, by iteratively training and testing the model over the entire dataset. This technique was selected for fine-tuning over the multilingual dataset to avoid any spurious correlations that may arise given the low number of responses associated with each individual language. In accordance with the k-fold validation technique, data for each language was first randomly split into 5 equal folds (20% of responses each). Across all languages, the amount of responses in a single fold ranged from 28 to 31. One of these 5 folds was then selected as a holdout-responses set, and the model was trained on the remaining 4 folds in addition to the ENG1 training set.

Once the training process was complete, the model was tested on the multilingual holdout-responses set. This process was then repeated, cycling through each of the 5 folds to serve as the holdout-responses set. In this way, the model iterated through a total of 5 training and testing runs, providing a more exhaustive assessment of model performance over the multilingual dataset. Of note, hyperparameter tuning was not performed for our English-translated model and instead we employed the same hyperparameter settings we used for our English-only final model.

Multilingual Large Language Models

We finally ran a single fine-tuning procedure using XLM-RoBERTa, a multilingual LLM, including the training data from our ENG1 dataset and the untranslated multilingual dataset. This fine-tuning procedure was run to test whether a single LLM could reliably predict the originality of narratives in 11 different languages, without the need for any translation to English.

XLM-RoBERTa

For our multilingual model, we employed the base version of XLM-RoBERTa with an added regression head for originality score prediction. We selected XLM-RoBERTa as it shares the same architecture of the monolingual RoBERTa-base model, differing only in its pretraining process. This made any comparisons between the two models more informative of the role of multilingual training data in driving performance in automated creativity scoring. XLM-RoBERTa is a multilingual version of RoBERTa that was released by Facebook AI in 2020 (Conneau et al., 2019). While RoBERTa was pretrained on English data alone, XLM-RoBERTa was pretrained on 100 different languages, including the 11 in our multilingual dataset.

Specifically, the multilingual dataset used for the pretraining of XLM-RoBERTa included 2.5 Terabytes of data from the CommonCrawl repository.

Multilingual Model

Our third fine-tuning procedure involved our multilingual model. We used the XLM-RoBERTa-base model and a combination of the training set from the ENG1 dataset and the entirety of the untranslated multilingual dataset. All other steps in this fine-tuning were identical to those employed for the translated multilingual model. Of note, each set in the k-fold validation included an equal number of untranslated responses from each language.

Results

All Pearson's correlations in our analysis were computed at the response-level. We further report root mean squared error (RMSE) values alongside each correlation.

English-Only Model

We first ran a series of Pearson's correlations between human-rated originality scores for the ENG1 dataset and our baseline variables (DSI and wordcount). Across all sets and responses in the ENG1 dataset, the mean wordcount was 70.8, the mean DSI was 0.75, and the mean human-rated originality score was 3.08. Regarding DSI, we observed small to large correlations between DSI scores and human-rated originality on the holdout-responses set (r = .34; p < .001; 95% CI: [.20, .47]; RMSE = 1.28), holdout-prompt set (r = .53; p < .001; 95% CI: [.40, .64]; RMSE = 1.19), and validation set (r = .38; p < .001; 95% CI: [.17, .55]; RMSE = 1.06). Regarding wordcount, we observed medium to large correlations with human-rated originality on the holdout-responses set (r = .55; p < .001; 95% CI: [.44, .65]; RMSE = 74), holdout-prompt set

(r = .63; p < .001; 95% CI: [.52, .72]; RMSE = 71.76), and validation set (r = .48; p < .001; 95% CI: [.29, .63]; RMSE = 79.69).

Manually tuned hyperparameter settings for our final English-only RoBERTa model included 25 epochs and a learning rate of 5e-5. Once fine-tuned, our RoBERTa model was used to predict human-rated originality scores for each story in the ENG1 dataset. A series of linear regressions were then run between the predicted model scores and the human-rated scores. Critically, we found that the model performed well on a test of generalizability, with large correlations between predicted and human-rated scores for the holdout-responses set (r = .81; p < .001; 95% CI: [.76, .86]; RMSE = .58) and holdout-prompt set (r = .80; p < .001; 95% CI: [.74, .86]; RMSE = .61), far exceeding the baseline of DSI and wordcount. As expected, given this set was used in the tuning of the model, we then found that our model strongly predicted the human-rated scores on the validation set (r = .88; p < .001; 95% CI: [.81, .92]; RMSE = .44).

English-Translated Model

Given that we employed the same RoBERTa-base model as the English-Only model, hyperparameter tuning was not performed for our English-translated model. The number of epochs and learning rate were taken from those used for our English-Only model, and all other settings were left to the default values for each model. A k-fold validation was run for both the English-translated and multilingual models, and correlations between predicted and human-rated scores were calculated for each fold (Figure 1).

We first computed Pearson's correlations between the DSI scores of English-translated stories and human-rated originality scores, to serve as a baseline to compare model performance (Figure 3). Across all languages, we found small to large correlations between DSI and human-

rated originality (r = .21 ~ .76; $p \leq .01$; RMSE = 1.27 ~ 1.28; See Appendix B). We then computed Pearson's correlations between the wordcount of English-translated stories and human-rated originality scores. We found small to large correlations between wordcount and human-rated originality across all languages (r = .26 ~ .63; $p \leq .001$; RMSE = 60.82 ~ 78.94; See Appendix B).

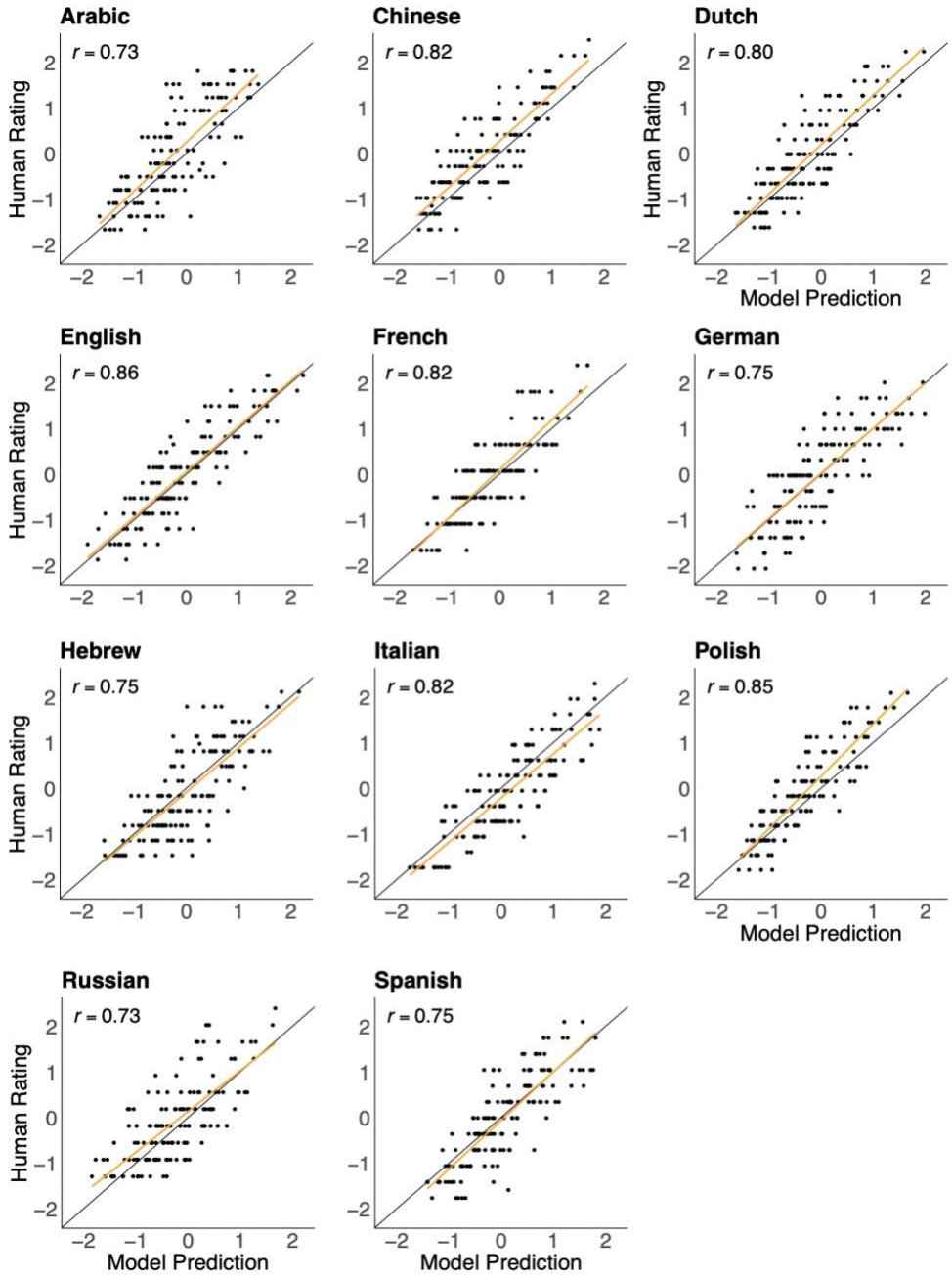
We finally ran a series of Pearson's correlations for all languages between our English-translated model's predictions and human-rated originality scores (Figure 1): Arabic (r = .73; p < .001; 95% CI: [.65, .80]; RMSE = .71), Chinese (r = .82; p = .001; 95% CI: [.76, .87]; RMSE = .64), Dutch (r = .80; p < .001; 95% CI: [.74, .85]; RMSE = .64), English (r = .86; p < .001; 95% CI: [.81, .89]; RMSE = .52), French (r = .82; p < .001; 95% CI: [.76, .87]; RMSE = .59), German (r = .75; p < .001; 95% CI: [.67, .82]; RMSE = .66), Hebrew (r = .75; p < .001; 95% CI: [.67, .81]; RMSE = .67), Italian (r = .82; p < .001; 95% CI: [.76, .87]; RMSE = .59), Polish (r = .85; p < .001; 95% CI: [.80, .89]; RMSE = .59), Russian (r = .73; p < .001; 95% CI: [.65, .80]; RMSE = .71), and Spanish (r = .75; p < .001; 95% CI: [.66, .81]; RMSE = .67). For each language, the performance of our English-translated model far exceeded those achieved by DSI scores or wordcount alone.

Multilingual Model

Having demonstrated robust correlations across languages for English-translated stories, we turned to testing multilingual models trained on the untranslated stories. This approach allows us to determine whether similarly strong performance could be achieved without having to translate the stories into English.

Figure 1

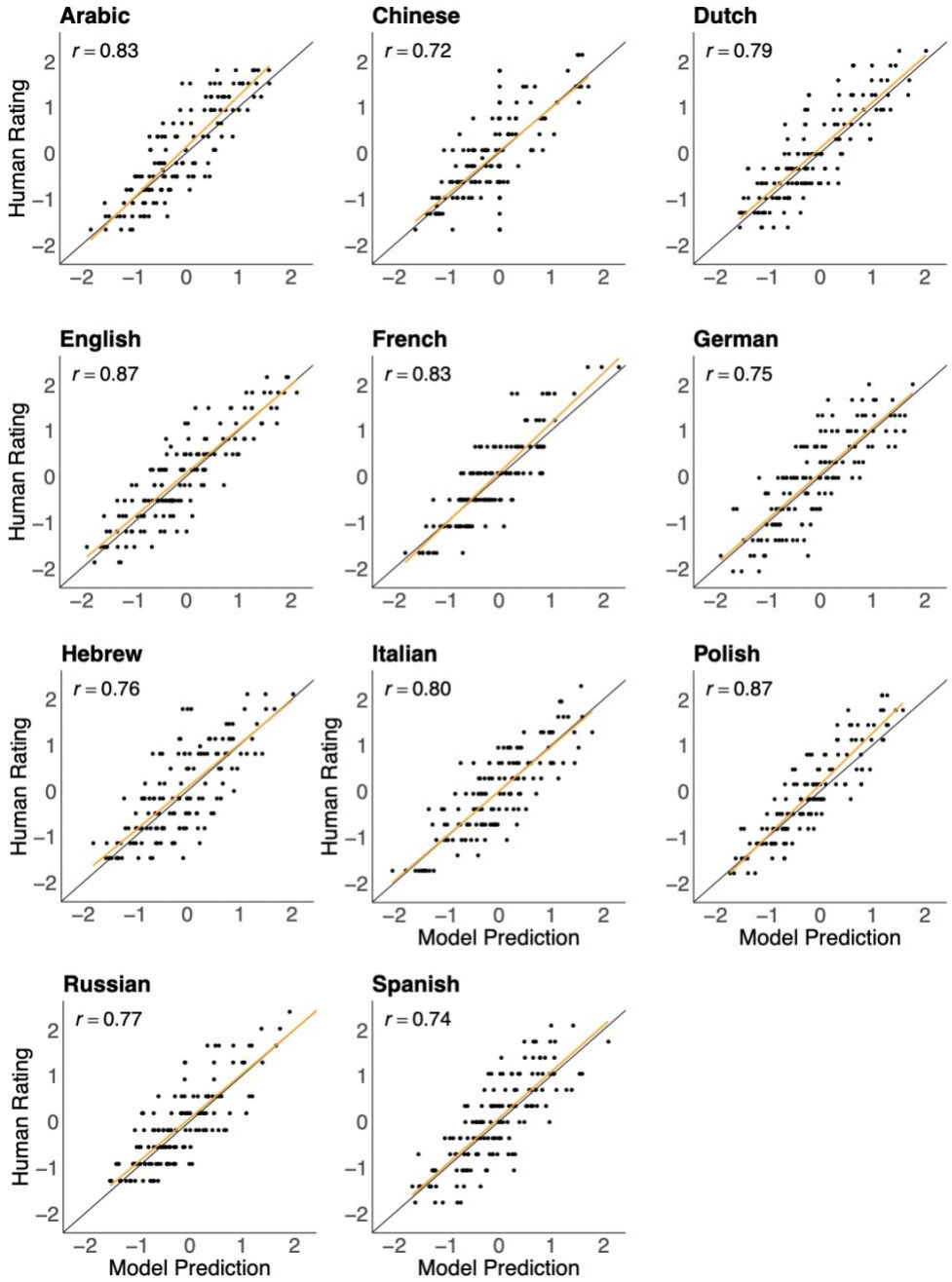
Linear Regressions Between Human-Rated Originality Scores and English-Translated Model Predictions



Note: Single responses are denoted by the black dots. Ideal performance (r = 1.00) is denoted by the black line, while the orange line is the line of best fit. All p values < .001.

Figure 2

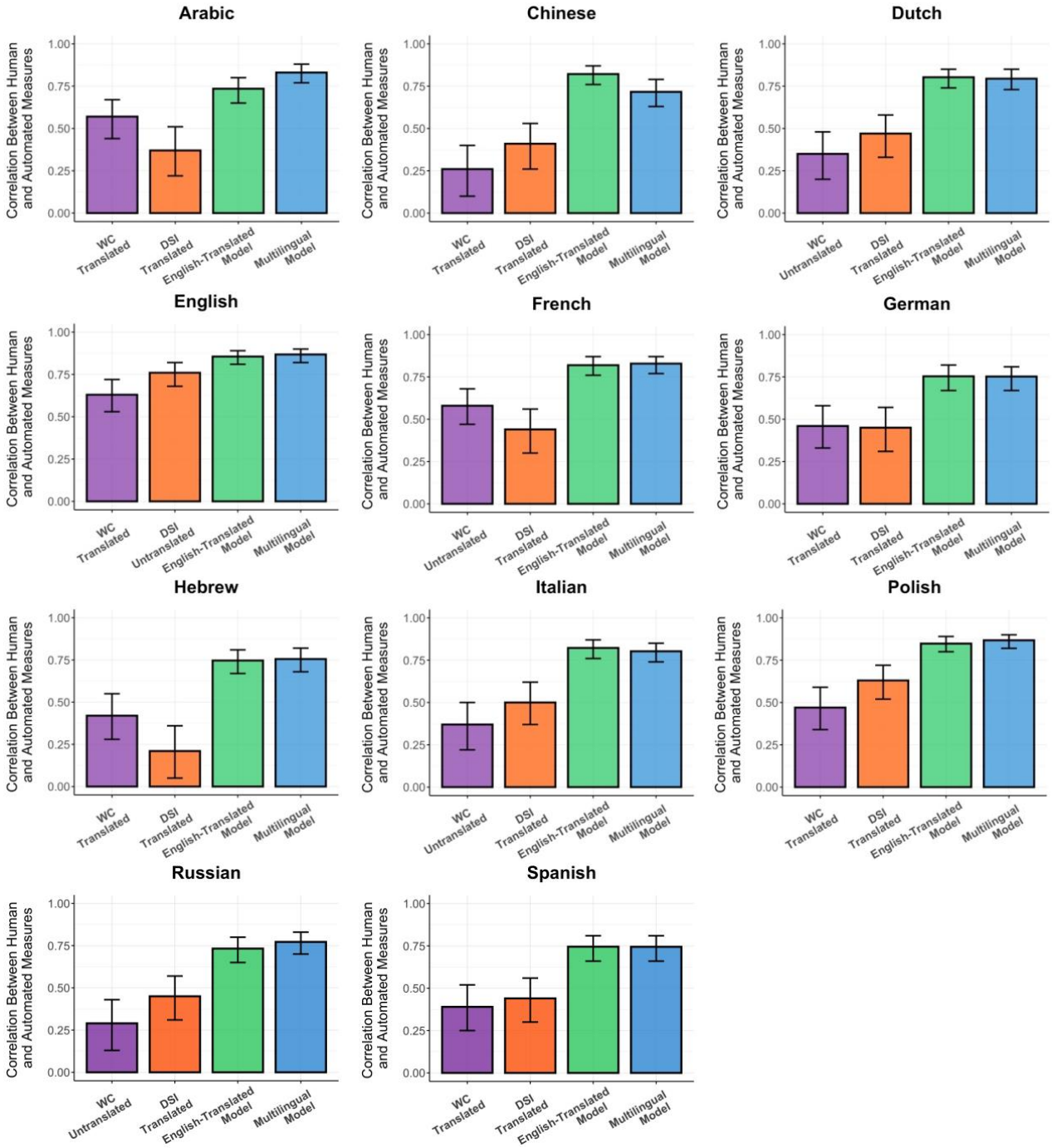
Linear Regressions Between Human-Rated Originality Scores and Multilingual Model Predictions



Note: Single responses are denoted by the black dots. Ideal performance (r = 1.00) is denoted by the black line, while the orange line is the line of best fit. All p values < .001.

Figure 3

Correlations Between Human-Rated Originality Scores and Automated Measures Across Languages



Note: WC, wordcount; DSI, divergent semantic integration. For each language, only the highest correlation between translated or untranslated wordcounts is displayed. Correlations refer to Pearson's product moment correlation coefficients (*r*). Error bars represent 95% confidence intervals.

Hyperparameter tuning was not performed for the multilingual model and the settings were instead taken from those selected for our English-Only model, as both models rely on a similar basic RoBERTa architecture. We ran a series of Pearson's correlations between wordcount (character count for Chinese) and human-rated originality scores for the untranslated multilingual stories. We observed small to large correlations across all languages (r = .26 ~ .63; $p \leq .001$; RMSE = 45.1 ~100.3; See Appendix B), consistent with the English-translated analysis.

Then, we ran Pearson's correlations between predictions from our multilingual model and human-rated originality scores, and observed large associations across all languages (Figure 2): Arabic (r = .83; p < .001; 95% CI: [.77, .88]; RMSE = .57), Chinese (r = .72; p < .001; 95% CI: [.63, .79]; RMSE = .71), Dutch (r = .79; p < .001; 95% CI: [.73, .85]; RMSE = .63), English (r = .87; p < .001; 95% CI: [.82, .90] ; RMSE = .5), French (r = .83; p < .001; 95% CI: [.77, .87]; RMSE = .57), German (r = .75; p < .001; 95% CI: [.67, .81]; RMSE = .67), Hebrew (r = .76; p < .001; 95% CI: [.68, .82]; RMSE = .66), Italian (r = .80; p < .001; 95% CI: [.74, .85]; RMSE = .6), Polish (r = .87; p < .001; 95% CI: [.82, .90]; RMSE = .52), Russian (r = .77; p < .001; 95% CI: [.70, .83]; RMSE = .65), and Spanish (r = .74; p < .001; 95% CI: [.66, .81]; RMSE = .67). Similarly to our English-translated model, our multilingual model far outperformed the predictions achieved by DSI scores and wordcount, across all languages.

Discussion

Creativity researchers have recently turned to LLMs in an attempt to overcome the limitations associated with subjective scoring, such as the costly nature of human raters and a slow pace of research (Barbot, 2018; Forthmann et al., 2017; Reiter-Palmon et al., 2019). To this

end, LLMs have been reported to surpass previous best practices for automated originality scoring of AUT responses (Dumas et al., 2021; Organisciak et al., 2023), metaphors (DiStefano et al., 2023), questions (Raz et al., 2024), and creative problem solving responses (Luchini et al., 2023). No such work to date has investigated whether LLMs can successfully predict the creativity of narratives, nor whether they are fit to assess the creativity of multilingual datasets.

To test whether LLMs could successfully predict human-rated originality scores on English narratives, we first trained a monolingual LLM, RoBERTa, on a dataset consisting of only English narratives. The model robustly predicted human-rated originality scores ($r \geq .80$), and did so better than DSI scores ($r \leq .53$) or wordcount ($r \leq .63$). We then tested whether we could train a model to predict human-rated originality scores on a set of multilingual narratives, irrespective of the language. To do this, we first trained the RoBERTa model on a compiled dataset including both English narratives and multilingual narratives that were translated into English. We observed strong correlations between model predicted scores and human-rated originality scores across all languages ($r = .73 \sim .86$), surpassing those achieved when employing DSI scores ($r = .21 \sim .76$) or wordcount alone ($r = .26 \sim .63$). Finally, we trained a multilingual model, XLM-RoBERTa, on the same compiled dataset of English and multilingual narratives, but kept the multilingual narratives in their original language. The multilingual model demonstrated similarly robust correlations between its predictions and the human-rated originality scores across all languages ($r \geq .72$), again displaying stronger correlations than those achieved with mere wordcount ($r \leq .63$).

The present work demonstrates, across three different iterations and 11 different languages, that LLMs can be employed successfully for the prediction of multilingual narrative creativity. Further, our findings indicate that LLMs provide the best solution to date for the

automatic prediction of narrative creativity, surpassing the previously best performing method employing DSI scores. In part, this is likely due to the static nature of the word embeddings that are used in the computation of DSI scores, lacking any dynamic learning or adaptation to the task of rating the creativity of narratives. Further, LLMs are the currently most advanced natural language processing tool, which have previously been noted to display a strong performance across a variety of linguistics tasks (Vaswani et al., 2017). The models also surpassed the correlations with human-rated originality scores that were achieved by wordcount alone, indicating that LLMs measure more than simply length of the narratives in their predictions. We found that our fine-tuned LLMs generally performed best, as indicated by lower prediction errors, on stories associated with average originality scores, as compared to stories with extreme scores, both high and low (Appendix E). Of note, this trend was also reported in past work on automated creativity scoring, likely pointing to a general weakness of fine-tuned LLMs (DiStefano et al., 2023; Luchini et al., 2023; Raz et al., 2024). Further, these findings align with recent accounts of creativity which posit that highly creative products and ideas tend to cause the largest disagreement amongst judges and are often not recognized until later (Corazza et al., 2022).

For both our English-translated and multilingual models we found the models performed best on the English (English-translated, r = .86, 95% CI: [.81, .89]; multilingual, r = .87, 95% CI: [.82, .90]) and Polish (English-translated, r = .85, 95% CI: [.80, .89]; multilingual, r = .87, 95% CI: [.82, .90]) narratives, compared to the other languages (English-translated, $r \leq .82$; multilingual, $r \leq .83$). Of note, both languages also displayed the largest correlations between DSI scores and human-rated originality scores (English, r = .76, $p < .001$; Polish, r = .63, $p < .001$), when compared to the other translated languages ($r \leq .50$). It is thus possible that English

and Polish narratives were predisposed to being more accurately scored by LLMs due to a higher number of easily traceable linguistic features, captured by DSI, which correlated strongly with human-rated originality scores in these narratives. This is further supported for the English narratives as they displayed the strongest correlation between wordcount and human-rated originality scores (r = .63) than the narratives in any of the other languages, both translated ($r \leq$ .57) and untranslated ($r \leq .58$). It is also possible, given the overrepresentation of English narratives in training data, that the models developed a bias towards English written stories. If this were the case, increasing the representation of the other languages in our training data would likely flatten the performance differences between languages.

We further explored the possibility that differences in story complexity across language datasets could be driving the differences in model performance. We thus calculated mean wordcount and entropy—a measure of linguistic randomness—for the stories in each language (Appendix D). We observed no clear pattern between model predictions and either mean wordcount or entropy, suggesting that differences in model performance across languages are actually driven by other factors. We further explored this possibility, investigating whether differences in model performance between languages were due to the prototypicality or commonality of the responses within each dataset, relative to the dataset itself. If responses within the dataset for a particular language are more frequently conceptually similar to each other, this would lead to a higher overlap between the training and holdout sets for that language. Ultimately, this might advantage the model at predicting originality scores within that dataset without providing a stringent enough test of model performance for the language in question.

To test this, we computed a series of within-dataset cross response embedding dissimilarity (CRED; see Appendix A for a detailed description of the approach and results)

scores. Of note, the Polish dataset displayed the smallest within-dataset CRED score (CRED = 0.55) compared to any of the other language datasets (CRED-within ≥ 0.57), indicating that the Polish narratives possessed a large amount of linguistic similarities between them, when compared to the other languages. In these terms, it is likely that the Polish dataset presented an easier challenge for the LLMs to successfully predict the originality of narratives, given what was likely a stronger similarity between narratives in the training and holdout sets. It is nevertheless curious that while the LLMs performed best on the English dataset, this dataset also possessed the largest within-dataset CRED score (CRED-within = 0.66), indicating that the English stories were highly dissimilar from each other. One possibility is that, given the overrepresentation of English stories in the training data, the LLMs gained an advantage towards the English language. Another possibility, then, is that the LLMs were able to overcome any issues arising from scoring dissimilar stories by relying on more superficial linguistic features from the English stories, which in turn were strongly correlated with human-rated originality scores, as suggested by DSI scores and wordcount.

Our work provides the first evidence that LLMs can be successfully employed for the automated prediction of creativity in human-generated narratives. We further demonstrate that a single LLM can robustly predict the creativity of narratives written in 11 different languages, both when translated to English and when in their original language. We thus extend past work indicating that Polish AUT responses translated to English could be automatically scored for their originality by GPT-3 (Zielińska et al., 2023). Indeed, we demonstrate that the same can be done for narratives, across 11 different languages, and by employing a smaller-sized, open-access LLM. While our English-Translated model performed comparatively with our multilingual model, being able to score narratives in their original language increases access to

automated creativity scoring tools for creativity researchers worldwide, and avoids any risks associated with improper translations that may distort the meaning of the original text. Multilingual automated creativity scoring tools will ensure researchers won't need to rely on subjective scoring, increasing the possibility for larger-scale multilingual studies to be conducted in the field of creativity. Critically, all the LLMs solutions we explored in the present paper surpassed the performance of previous best practices in the automated scoring of narrative creativity.

Of note, the concept of divergence is only relevant to our use of DSI scores as a baseline originality measure. DSI scores measure conceptual divergence in text data, with the underlying assumption that this will correlate with text originality. Our fine-tuned LLMs are trained on human ratings of originality, which concern a much broader definition of narrative originality than mere divergence. Nevertheless, the current research only considered originality scores, overlooking other components of creativity, such as quality (i.e., effectiveness). According to the standard definition, originality and quality are both necessary for creativity (Runco & Jaeger, 2012; Stein, 1953). Quality, which is thought to subsume the concepts of aesthetics and value (Corazza, 2016), has been found to strongly correlate with originality on some creativity tasks (Acar et al., 2017). It has nevertheless been noted that originality and quality are distinct components of creativity, each possessing unique associations to other constructs (Morral-Robinson & Reiter-Palmon, 2013; Reiter-Palmon et al., 1997; 2009). Indeed, past work has indicated that LLMs are better at predicting human-rated quality scores, than originality scores (Luchini et al., 2023). Nevertheless, future research should test whether a multilingual LLM can also reliably predict the quality of narratives across a variety of languages.

Further, it is important that further research should replicate the present findings across levels of individual differences, such as language fluency. As we only recruited fluent speakers for our multilingual dataset, it is possible that our findings will not perfectly extend to less fluent second language speakers. Nevertheless, past work employing DSI to predict narrative creativity, a method employing word-embeddings derived from a BERT model, has indicated that the measure performed equally well with stories written by second language English speakers (Johnson et al., 2023). Given that fine-tuned LLMs are a more powerful, and context sensitive tool than word-embedding based approaches, it is thus likely that any such bias should also not be observed with the present approach.

Finally, while we selected two models from the RoBERTa family of LLMs for the present work, many other models have been developed since their release, some of which have been shown to outperform RoBERTa models on several benchmarks (Liu et al., 2023a; 2023b). Further, this trend of gradual LLM improvement is likely to continue in the future, as more advanced and better tuned models are developed. While we provide the current best-method for the automated creativity scoring of multilingual narratives, the challenge of developing a flawless tool for this purpose is not yet over. Indeed, perfect model performance is likely unachievable, in part due to rater disagreement in the human ratings which adds unpredictable noise to the data. Nevertheless, as the field of natural language processing advances, so should creativity researchers strive to update their methods. To this end, we make our data openly available, so that researchers may test whether better predictions can be achieved with different techniques.

This evidence confirms one of the possible forms of human-machine cooperation in creative activity (Vinchon et al., 2023), but it also opens the way to further extensions into

scenarios in which AI takes on a more active role in co-creation. For instance, our models could be used to provide real-time feedback to individuals that are taking part in the FSST. This would allow researchers to study the effects of feedback on the creative process, as seen in past work employing LLMs to provide real-time feedback on the AUT (de Chantal & Organisciak, 2023). Other models could then be developed to provide story writing support for more realistic tasks such as longer-form story writing. Leveraging the generative nature of LLMs, qualitative feedback could also be provided to writers, such as by outlining how a story could be elaborated to maximize its creativity. In one such example, researchers found that individuals who collaborated with GPT-4 wrote more creative, but less thematically diverse narratives (Doshi & Hauser, 2023). Nevertheless, given that expert human writers still outperform modern LLMs in creative story writing, it remains unclear whether such models might indeed aid more experienced writers (Chakrabarty et al., 2023). Cyber-human creative writing with reinforcement learning generated by both sides of the collaboration is a line of research worth pursuing in the future. Cyber-human creative writing with reinforcement learning generated by both sides of the collaboration is a line of research worth pursuing in the future.

References

Acar, S., Burnett, C., & Cabra, J. F. (2017). Ingredients of creativity: Originality and more. *Creativity Research Journal, 29*(2), 133-144. https://doi.org/10.1080/10400419.2017.1302776

Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology*, *43*(5), 997–1013. https://doi.org/10.1037/0022-3514.43.5.997

Barbot, B. (2018). The dynamics of creative ideation: Introducing a new assessment paradigm. *Frontiers in Psychology*, *9*, 2529. https://doi.org/10.3389/fpsyg.2018.02529

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., & Kenett, Y. N. (2023). Associative thinking at the core of creativity. *Trends in Cognitive Sciences*, *27*(7), 671–683. https://doi.org/10.1016/j.tics.2023.04.004

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., Kenett, Y. N., Hass, R. W., & Schacter, D. L. (2023). Semantic memory and creativity: The costs and benefits of semantic memory structure in generating original ideas. *Thinking & Reasoning*, *29*(2), 305–339. https://doi.org/10.1080/13546783.2022.2076742

Beisemann, M., Forthmann, B., Bürkner, P. C., & Holling, H. (2020). Psychometric evaluation

of an alternate scoring for the remote associates test. *The Journal of creative behavior*, *54*(4), 751-766. https://doi.org/10.1002/jocb.394

Bendetowicz, D., Urbanski, M., Garcin, B., Foulon, C., Levy, R., Bréchemier, M. L., ... & Volle, E. (2018). Two critical brain networks for generation and combination of remote associations. *Brain*, *141*(1), 217-233. https://doi.org/10.1093/brain/awx294

Benedek, M., Mühlmann, C., Jauk, E., & Neubauer, A. C. (2013). Assessment of divergent thinking by means of the subjective top-scoring method: Effects of the number of top-ideas and time-on-task on reliability and validity. *Psychology of Aesthetics, Creativity, and the Arts*, *7*(4), 341–349. https://doi.org/10.1037/a0033644

Ceh, S. M., Edelmann, C., Hofer, G., & Benedek, M. (2022). Assessing raters: What factors predict discernment in novice creativity raters? *The Journal of Creative Behavior*, *56*(1), 41–54. https://doi.org/10.1002/jocb.515

Choudhury, M., & Deshpande, A. (2021). How linguistically fair are multilingual pre-trained language models? *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(14), 12710–12718. https://doi.org/10.1609/aaai.v35i14.17505

Chrysikou, E. G., Morrow, H. M., Flohrschutz, A., & Denney, L. (2021). Augmenting ideational fluency in a creativity task across multiple transcranial direct current stimulation montages. *Scientific Reports*, *11*(1), 8874. https://doi.org/10.1038/s41598-021-85804-3

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2019). *Unsupervised Cross-lingual Representation Learning at Scale* (arXiv:1911.02116). arXiv. http://arxiv.org/abs/1911.02116

Corazza, G. E. (2016). Potential originality and effectiveness: The dynamic definition of

creativity. *Creativity research journal*, *28*(3), 258-267. https://doi.org/10.1080/10400419.2016.1195627

Corazza, G. E., Agnoli, S., & Mastria, S. (2022). The dynamic creativity framework. *European Psychologist*, *27*(3). https://doi.org/10.1027/1016-9040/a000473

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

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

Dickson, D., Jonczyk, R., Gunay, E., Van Hell, J., & Siddique, Z. (2022, August). Utilization of automatized creativity ratings in linguistically diverse populations: Automated scores align with human ratings. In *2022 ASEE Annual Conference & Exposition*. http://dx.doi.org/10.18260/1-2--41126

DiStefano, P. V., Patterson, J. D., & Beaty, R. (2023). Automatic scoring of metaphor creativity with large language models. *PsyArXiv*. https://doi.org/10.31234/osf.io/6jtxb

Dumas, D., & Runco, M. (2018). Objectively scoring divergent thinking tests for originality: A re-analysis and extension. *Creativity Research Journal*, *30*(4), 466-468.

Dumas, D., Organisciak, P., & Doherty, M. (2021). Measuring divergent thinking originality with human raters and text-mining models: A psychometric comparison of methods. *Psychology of Aesthetics, Creativity, and the Arts*, *15*(4), 645–663. https://doi.org/10.1037/aca0000319

D'Souza, R. (2021). What characterises creativity in narrative writing, and how do we assess it? Research findings from a systematic literature search. *Thinking Skills and Creativity*, *42*, 100949. https://doi.org/10.1016/j.tsc.2021.100949

Forthmann, B., Beaty, R. E., & Johnson, D. R. (2023). Semantic spaces are not created equal – how should we weigh them in the sequel? On composites in automated creativity scoring. *European Journal of Psychological Assessment*, *39*(6), 449–461. https://doi.org/10.1027/1015-5759/a000723

Forthmann, B., Goecke, B., & Beaty, R. E. (2023). Planning missing data designs for human ratings in creativity research: A practical guide. *Creativity Research Journal*, 1-12. https://doi.org/10.1080/10400419.2023.2250976

Forthmann, B., Holling, H., Zandi, N., Gerwig, A., Çelik, P., Storme, M., & Lubart, T. (2017). Missing creativity: The effect of cognitive workload on rater (dis-)agreement in subjective divergent-thinking scores. *Thinking Skills and Creativity*, *23*, 129–139. https://doi.org/10.1016/j.tsc.2016.12.005

Groves, M., & Mundt, K. (2015). Friend or foe? Google Translate in language for academic purposes. *English for Specific Purposes*, *37*, 112–121. https://doi.org/10.1016/j.esp.2014.09.001

Goecke, B., DiStefano, P. V., Aschauer, W., Haim, K., Beaty, R., & Forthmann, B. (2024). Automated Scoring of Scientific Creativity in German: A Brief Report of Study Design and Statistical Properties. *PsyArXiv*. https://doi.org/10.31234/osf.io/be7yr

Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science*, *14*(6), 1006-1033. https://doi.org/10.1177/17456916198613

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

Hennessey, B. A. (1994). The consensual assessment technique: An examination of the relationship between ratings of product and process creativity. *Creativity Research Journal*, *7*(2), 193–208. https://doi.org/10.1080/10400419409534524

Jawahar, G., Sagot, B., & Seddah, D. (2019). What does BERT learn about the structure of language?. In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*. https://doi.org/10.18653/v1/P19-1356

Johnson, D. R., Cuthbert, A. S., & Tynan, M. E. (2021). The neglect of idea diversity in creative idea generation and evaluation. *Psychology of Aesthetics, Creativity, and the Arts*, *15*(1), 125–135. https://doi.org/10.1037/aca0000235

Johnson, D. R., Kaufman, J. C., Baker, B. S., Patterson, J. D., Barbot, B., Green, A. E., ... & Beaty, R. E. (2023). Divergent semantic integration (DSI): Extracting creativity from narratives with distributional semantic modeling. *Behavior Research Methods*, *55*(7), 3726-3759. https://doi.org/10.3758/s13428-022-01986-2

Kaufman, J. C., Baer, J., Cole, J. C., & Sexton∗, J. D. (2008). A comparison of expert and nonexpert raters using the consensual assessment technique. *Creativity Research Journal*, *20*(2), 171–178. https://doi.org/10.1080/10400410802059929

Kaufman, J. C., Baer, J., Cropley, D. H., Reiter-Palmon, R., & Sinnett, S. (2013). Furious activity vs. understanding: How much expertise is needed to evaluate creative work? *Psychology of Aesthetics, Creativity, and the Arts*, *7*(4), 332–340. https://doi.org/10.1037/a0034809

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

Liu, H., Ning, R., Teng, Z., Liu, J., Zhou, Q., & Zhang, Y. (2023a). Evaluating the logical reasoning ability of chatgpt and gpt-4. *arXiv*. https://doi.org/10.48550/arXiv.2304.03439

Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., ... & Ge, B. (2023b). Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-Radiology*, *1*(2). 100017. https://doi.org/10.1016/j.metrad.2023.100017

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*. https://doi.org/10.48550/ARXIV.1907.11692

Luchini, S., Maliakkal, N. T., DiStefano, P. V., Patterson, J. D., Beaty, R., & Reiter-Palmon, R. (2023). Automatic scoring of creative problem-solving with large language models: A comparison of originality and quality ratings. *PsyArXiv*. https://doi.org/10.31234/osf.io/g5qvf

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior research methods, instruments, & computers*, *28*(2), 203-208. https://doi.org/10.3758/BF03204766

Maio, S., Dumas, D., Organisciak, P., & Runco, M. (2020). Is the reliability of objective originality scores confounded by elaboration? *Creativity Research Journal*, *32*(3), 201–205. https://doi.org/10.1080/10400419.2020.1818492

Maity, D., & Takahashi, T. (2021). Existence and uniqueness of strong solutions for the system of interaction between a compressible Navier–Stokes–Fourier fluid and a damped plate

equation. *Nonlinear Analysis: Real World Applications*, *59*, 103267. https://doi.org/10.1016/j.nonrwa.2020.103267

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, *92*, 57-78. https://doi.org/10.1016/j.jml.2016.04.001

Mednick, S. (1962). The associative basis of the creative process. *Psychological Review*, *69*(3), 220–232. https://doi.org/10.1037/h0048850

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *ArXiv*. http://arxiv.org/abs/1301.3781

Morral-Robinson, E., & Reiter-Palmon, R. (2013). The interactive effects of self-perceptions and job requirements on creative problem solving. *Journal of Creative Behavior*, *47*, 200-214. https://doi.org/10.1002/jocb.31

Mozaffari, H. (2013). An analytical rubric for assessing creativity in creative writing. *Theory and Practice in Language Studies*, *3*(12), 2214–2219. https://doi.org/10.4304/tpls.3.12.2214-2219

Murray, S., Liang, N., Brosowsky, N., & Seli, P. (2021). What are the benefits of mind wandering to creativity? *Psychology of Aesthetics, Creativity, and the Arts*. https://doi.org/10.1037/aca0000420

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

Ovando-Tellez, M., Kenett, Y. N., Benedek, M., Bernard, M., Belo, J., Beranger, B., Bieth, T., &

Volle, E. (2023). Brain connectivity-based prediction of combining remote semantic associates for creative thinking. *Creativity Research Journal*, 35(3), 522–546.

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

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

Prabhakaran, R., Green, A. E., & Gray, J. R. (2014). Thin slices of creativity: Using single-word utterances to assess creative cognition. *Behavior research methods*, *46*, 641-659. https://doi.org/10.3758/s13428-013-0401-7

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. *OpenAI*. https://openai.com/blog/language-unsupervised/

Raz, T., Luchini, S., Beaty, R. E., & Kenett, Y. N. (2024). Automated scoring of question complexity: A large language model approach. *Researchsquare*.

Raz, T., Reiter-Palmon, R., & Kenett, Y. N. (2023). The role of asking more complex questions in creative thinking. *Psychology of Aesthetics, Creativity, and the Arts.* Advance online publication. https://doi.org/10.1037/aca0000658

Reimers, N., & Gurevych, I. (2020). Making monolingual sentence embeddings multilingual

using knowledge distillation. *ArXiv*. http://arxiv.org/abs/2004.09813

Reiter-Palmon, R., Forthmann, B., & Barbot, B. (2019). Scoring divergent thinking tests: A review and systematic framework. *Psychology of Aesthetics, Creativity, and the Arts*, *13*(2), 144–152. https://doi.org/10.1037/aca0000227

Reiter-Palmon, R., Illies, M. Y., Kobe Cross, L., Buboltz, C., & Nimps, T. (2009). Creativity and domain specificity: The effect of task type on multiple indexes of creative problem-solving. *Psychology of Aesthetics, Creativity, and the Arts*, *3*(2), 73. https://doi.org/10.1037/a0013410

Reiter-Palmon, R., Mumford, M. D., Boes, O. J., & Runco, M. A. (1997). Problem construction and creativity: The role of ability, cue consistency, and active processing. *Creativity Research Journal, 10*, 9-24. https://doi.org/10.1207/s15326934crj1001_2

Runco, M. A., & Jaeger, G. J. (2012). The standard definition of creativity. *Creativity research journal, 24*(1), 92-96. https://doi.org/10.1080/10400419.2012.650092

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18*(11), 613-620. https://doi.org/10.1145/361219.361220

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

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

Smith, K. A., & Vul, E. (2015). The role of sequential dependence in creative semantic search.

Topics in Cognitive Science, *7*(3), 543-546. https://doi.org/10.1111/tops.12152

Stein, M. I. (1953). Creativity and culture. *The journal of psychology*, *36*(2), 311-322. https://doi.org/10.1080/00223980.1953.9712897

Taylor, C. L., & Barbot, B. (2021). Dual pathways in creative writing processes. *Psychology of Aesthetics, Creativity, and the Arts*. https://doi.org/10.1037/aca0000415

Toubia, O., Berger, J., & Eliashberg, J. (2021). How quantifying the shape of stories predicts their success. *Proceedings of the National Academy of Sciences*, *118*(26), e2011695118. https://doi.org/10.1073/pnas.201169511

Vaezi, M., & Rezaei, S. (2019). Development of a rubric for evaluating creative writing: A multi-phase research. *New Writing*, *16*(3), 303–317. https://doi.org/10.1080/14790726.2018.1520894

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *ArXiv*. http://arxiv.org/abs/1706.03762

Vinchon, F., Lubart, T., Bartolotta, S., Gironnay, V., Botella, M., Bourgeois-Bougrine, S., Burkhardt, J.M., Bonnardel, N., Corazza, G.E., Glăveanu, V. & Hanchett Hanson, M. (2023). Artificial intelligence & creativity: A manifesto for collaboration. *The Journal of Creative Behavior*, *57*(4), 472-484. https://doi.org/10.1002/jocb.597

Weinstein, T. J., Ceh, S. M., Meinel, C., & Benedek, M. (2022). What's creative about sentences? A computational approach to assessing creativity in a sentence generation task. *Creativity Research Journal*, *34*(4), 419-430. https://doi.org/10.1080/10400419.2022.2124777

Yang, T., Zhang, Q., Sun, Z., & Hou, Y. (2023). Automatic Assessment of Divergent Thinking in Chinese Language with TransDis: A Transformer-Based Language Model Approach.

arXiv. https://doi.org/10.48550/arXiv.2306.14790

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

Zedelius, C. M., Mills, C., & Schooler, J. W. (2019). Beyond subjective judgments: Predicting evaluations of creative writing from computational linguistic features. *Behavior Research Methods*, *51*(2), 879–894. https://doi.org/10.3758/s13428-018-1137-1

Zielińska, A., Organisciak, P., Dumas, D., & Karwowski, M. (2023). Lost in translation? Not for Large Language Models: Automated divergent thinking scoring performance translates to non-English contexts. *Thinking Skills and Creativity*, *50*, 101414. https://doi.org/10.1016/j.tsc.2023.101414

Appendix

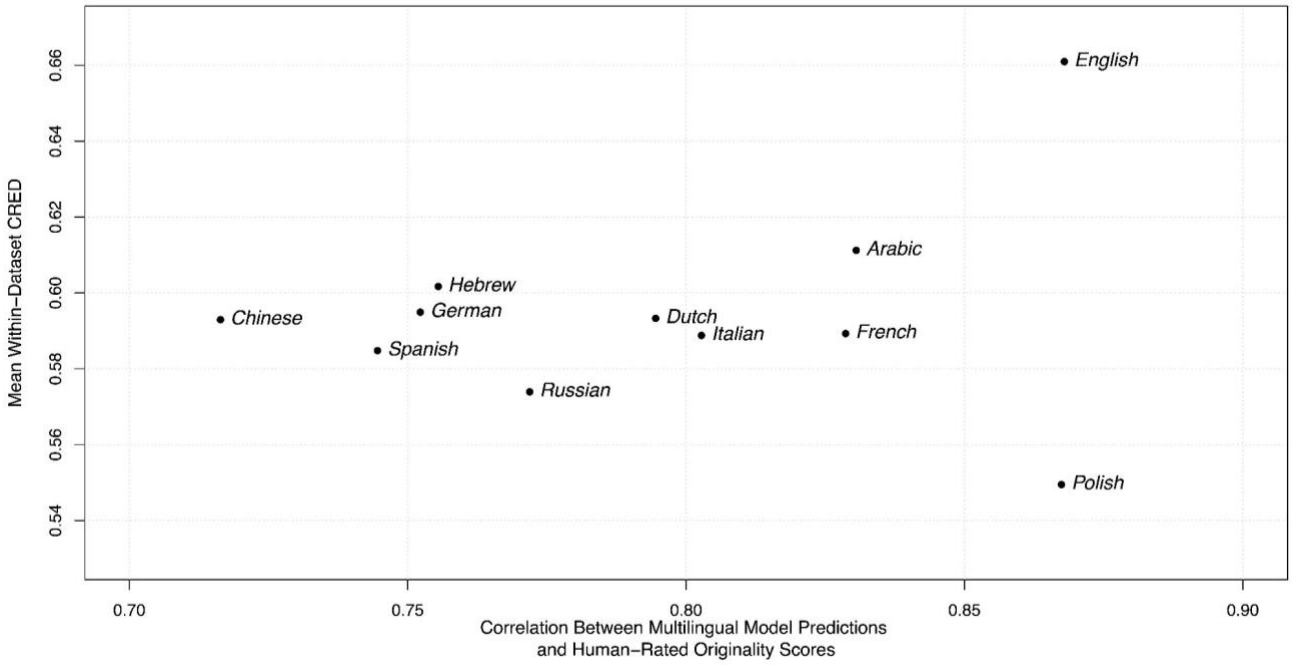
Appendix A.

Cross Response Embedding Dissimilarity (CRED).

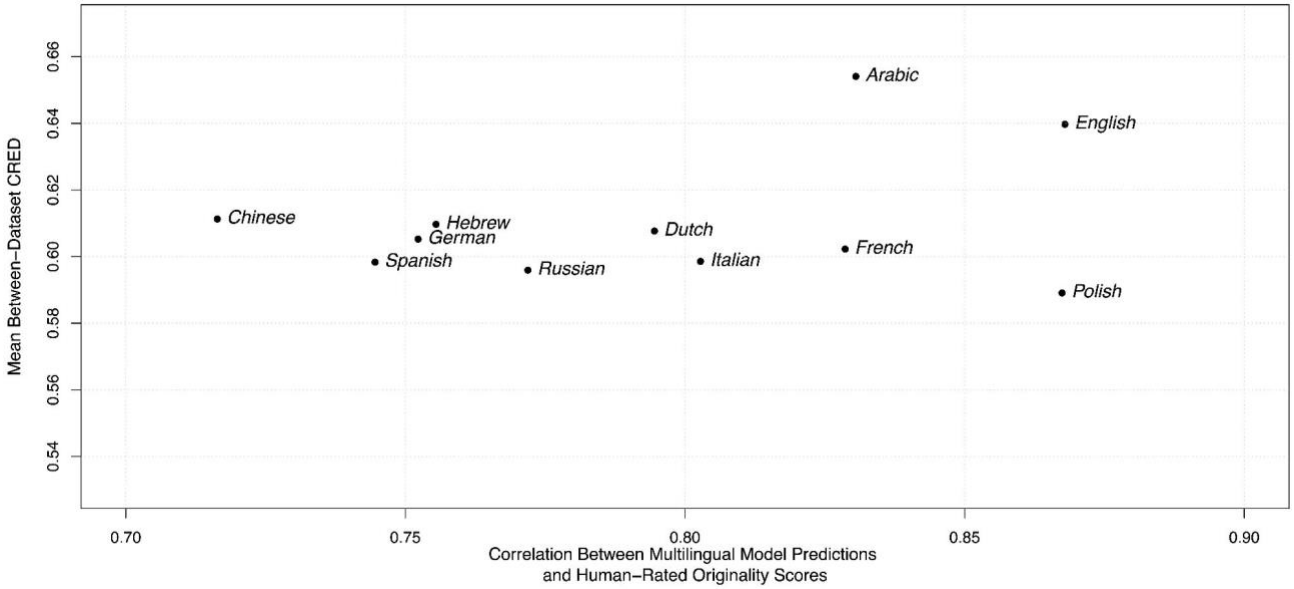
We computed cross response embedding dissimilarity (CRED) scores both within and between each language dataset within our multilingual collection. Similarly to DSI scores, CRED scores are also derived from an unsupervised machine learning approach. CRED leverages response-level embeddings to calculate the relative dissimilarity between text responses. Applied to short stories data, CRED can provide an indication of how dissimilar the stories within a dataset, or between two datasets, are to each other. We extracted response-level embeddings from the paraphrase-xlm-r-multilingual-v1 model (Reimers & Gurevych, 2020), which is a multilingual version of the paraphrase-distilroberta-base-v1 model based on the RoBERTa model (Devlin et al., 2018). The paraphrase-xlm-r-multilingual-v1 shares the same architecture as RoBERTa but has been fine-tuned to generate response-level embeddings—high-dimensional vector representations of sentences or paragraphs—across more than 50 languages, including all of those present in our multilingual dataset. Response-level embeddings are more sensitive to context than word-embeddings, capturing both semantic and syntactic features of text data. As such, response-level embeddings are a good candidate to compute the dissimilarity of two excerpts of text. The similarity between two response-level embeddings is derived from the cosine of the angle between them (i.e., cosine similarity; Salton et al., 1975). To calculate the dissimilarity between two responses one can then simply take the converse of the cosine similarity (1 – cosine similarity). Within-dataset CRED scores are calculated by repeating this comparison for all possible pairs of responses in a dataset and averaging the resulting values. In this way, within-dataset CRED scores reflect the semantic and syntactic diversity of a dataset.

Between-dataset CRED scores are instead derived by first calculating the dissimilarity across all possible pairs of responses between two datasets and taking the average of all such pairs. Repeating the process for all pairings of datasets, a single between-dataset CRED score is derived by averaging all scores associated with a particular dataset. In this way, between-dataset CRED scores reflect the semantic and syntactic distance between one dataset and all other datasets.

Appendix A1. *Within-Dataset CRED Score by Multilingual Model Performance Across Languages*



Appendix A2. *Between-Dataset CRED Score by Multilingual Model Performance Across Languages*



Appendix B. *Correlations Between Baseline Measures and Human-Rated Originality Scores*

	Translated Responses		Untranslated Responses
Language	DSI	Wordcount	Wordcount
Arabic	r = .37 (p < .001), RMSE = 1.27	r = .57 (p < .001), RMSE = 60.82	r = .48 (p < .001), RMSE = 45.1
Chinese	r = .41 (p < .001), RMSE = 1.27	r = .26 (p = .001), RMSE = 61.38	r = .26 (p = .001), RMSE = 100.3
Dutch	r = .47 (p < .001), RMSE = 1.27	r = .33 (p < .001), RMSE = 64.97	r = .35 (p < .001), RMSE = 66.28
English	r = .76 (p < .001), RMSE = 1.27	r = .63 (p < .001), RMSE = 79.78	r = .63 (p < .001), RMSE = 79.78
French	r = .44 (p < .001), RMSE = 1.28	r = .56 (p < .001), RMSE = 65.45	r = .58 (p < .001), RMSE = 66.48
German	r = .45 (p < .001), RMSE = 1.28	r = .46 (p < .001), RMSE = 69.27	r = .45 (p < .001), RMSE = 68.66
Hebrew	r = .21 (p = .01), RMSE = 1.28	r = .42 (p < .001), RMSE = 78.94	r = .41 (p < .001), RMSE = 58.13
Italian	r = .50 (p < .001), RMSE = 1.28	r = .35 (p < .001), RMSE = 78.92	r = .37 (p < .001), RMSE = 74.9
Polish	r = .63 (p < .001), RMSE = 1.27	r = .47 (p < .001), RMSE = 69	r = .46 (p < .001), RMSE = 55.97
Russian	r = .45 (p < .001), RMSE = 1.27	r = .27 (p < .001), RMSE = 63.22	r = .29 (p < .001), RMSE = 50.09
Spanish	r = .44 (p < .001), RMSE = 1.28	r = .39 (p < .001), RMSE = 69.93	r = .39 (p < .001), RMSE = 69.99

Note: RMSE = Root mean squared error.

Appendix C. *Correlations Between Model Predictions and Human-Rated Originality Scores, Number of Raters, and Average Rater Intraclass Correlations (ICCs)*

Language	Raters	ICC	English-Translated Model	Multilingual Model
Arabic	3	0.84 [0.79, 0.88]	*r* = .73, RMSE = .71	*r* = .83, RMSE = .57
Chinese	3	0.83 [0.77, 0.87]	*r* = .82, RMSE = .64	*r* = .72, RMSE = .71
Dutch	3	0.87 [0.83, 0.9]	*r* = .8, RMSE = .64	*r* = .79, RMSE = .63
English	3	0.8 [0.73, 0.85]	*r* = .86, RMSE = .52	*r* = .87, RMSE = .5
French	2	0.78 [0.7, 0.84]	*r* = .82, RMSE = .59	*r* = .83, RMSE = .57
German	3	0.75 [0.67, 0.81]	*r* = .75, RMSE = .66	*r* = .75, RMSE = .67
Hebrew	3	0.79 [0.73, 0.84]	*r* = .75, RMSE = .67	*r* = .76, RMSE = .66
Italian	3	0.87 [0.83, 0.9]	*r* = .82, RMSE = .59	*r* = .8, RMSE = .6
Polish	3	0.88 [0.84, 0.91]	*r* = .85, RMSE = .59	*r* = .87, RMSE = .52
Russian	3	0.77 [0.7, 0.83]	*r* = .73, RMSE = .71	*r* = .77, RMSE = .65
Spanish	3	0.78 [0.71, 0.83]	*r* = .75, RMSE = .67	*r* = .74, RMSE = .67

Note: All *p* values < .001. Intra-class correlations refer to ICC3k for all languages except ICC2k for French. RMSE = Root mean squared error.

Appendix D. *Wordcounts and Entropy Scores*

Language	Wordcount – Original *M* (SD)	Wordcount – Translated *M* (SD)	Entropy – Original *M* (SD)	Entropy – Translated *M* (SD)
Arabic	42.6 (15.5)	56.5 (23.2)	4.22 (.2)	4 (.11)
Chinese	89.9 (45)	56.3 (24.7)	5.4 (.48)	4.09 (.09)
Dutch	63.4 (19.6)	62.1 (19.4)	4.09 (.09)	4.1 (.08)
English	75.7 (25.9)	75.7 (25.9)	4.09 (.08)	4.09 (.08)
French	63.8 (19.4)	62.8 (19)	4.16 (.09)	4.07 (.06)
German	65.4 (21.5)	65.9 (21.7)	4.14 (.08)	4.09 (.07)
Hebrew	55.1 (19.1)	74.5 (26.6)	4.17 (.08)	4.06 (.08)
Italian	71.5 (22.6)	75.5 (23.5)	4.05 (.07)	4.07 (.08)
Polish	53.5 (16.8)	66.3 (19.5)	4.46 (.07)	4.08 (.07)
Russian	47.4 (16.6)	59.6 (21.4)	4.38 (.1)	4.06 (.1)
Spanish	66.8 (21.3)	66.7 (21.3)	4.13 (.08)	4.08 (.07)

Note: Entropy was calculated via the Shannon Entropy formula.

Appendix E. *Mean Prediction Errors*

Language	± 2 SD – Original	Within 2 SD – Original	± 2 SD – Translated	Within 2 SD – Translated
Arabic	NA	.46	NA	.57
Chinese	.96	.5	1.07	.46
Dutch	.9	.46	1	.47
English	.36	.4	.43	.41
French	1.01	.4	1.13	.41
German	.63	.52	.76	.52
Hebrew	.83	.51	.81	.52
Italian	1.35	.45	.96	.49
Polish	1.06	.37	1.24	.41
Russian	1.19	.43	1.35	.49
Spanish	1.43	.49	1.39	.48

Note: Errors are averaged in the form of absolute values, between two standard deviations of the mean, and above or below two standard deviations of the mean of human originality ratings. NA: No ratings are above or below 2 standard deviations.

Appendix F. *Instructions Provided to Human Raters for Multilingual Dataset Scoring*

In this study, participants were given 3 words -- "pen, paper, write" -- and asked to write a story that includes all 3 words. They were told to be imaginative and creative when writing their stories. The stories are about 5 sentences long.

You will rate the creativity of each story using a 1 (very uncreative) to 5 (highly creative) scale. Here is the full rating scale:
1: Very Uncreative
2: Uncreative
3: Undecided
4: Creative
5: Very Creative

When rating the stories, try not to focus too much on the length of the story, or how good the language is. Consider the overall creativity of the story.
You could consider how creatively the 3 words were used; how emotive, descriptive, or humorous the story was; or how much it "came alive."

Here's a few things to keep in mind when rating the stories:

- The stories were written by participants in a psychology experiment.
- Compare the stories to one another, not to another standard (what is a great story).
- Try to use the full 1-5 scale. For example, don’t rate all stories either 1, 2, or 3.
- Read a few stories before you start your rating (or go back and review your first ratings after you get a sense of the level of the stories).
- You can change your ratings as much as you wish, but you don't need to agonize over each rating -- just give your best expert judgment of the creativity of each story.
- If you see any invalid stories, where the story is only a few words, or people didn't take the task seriously, just put N/A as your rating.
- Occasionally, you may see a story that ends mid-sentence. This is likely because people ran out of time. You should still rate these stories.
- Feel free to take a break whenever you feel like it.