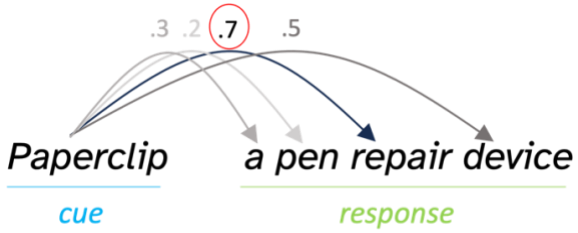
Understanding Semantic Distance (SemDis)

What is SemDis? Semantic distance reflects the relation between ideas or concepts based on how frequently their words appear in similar contexts. For instance, 'coffee' and 'mug' have a low semantic distance as they are often associated together, while 'coffee' and 'astronaut' have a high distance due to their unrelated contexts. Semantic distance scores range from 0 to 1, with higher scores indicating a greater semantic distance between words.

How does SemDis work? We employ the multilingual SemDis approach in Patterson, Merseal et al. (2023). The approach selects the appropriate the model (either multilingual BERT or RoBERTa) that performs best for the detected language in the uploaded file. The model then returns the maximum associative distance (MAD; Yu et al., 2023) between the cue and each word in the response. This approach correlates more strongly with human creativity ratings compared to alternative methods that average across semantic distances between cue and response words.



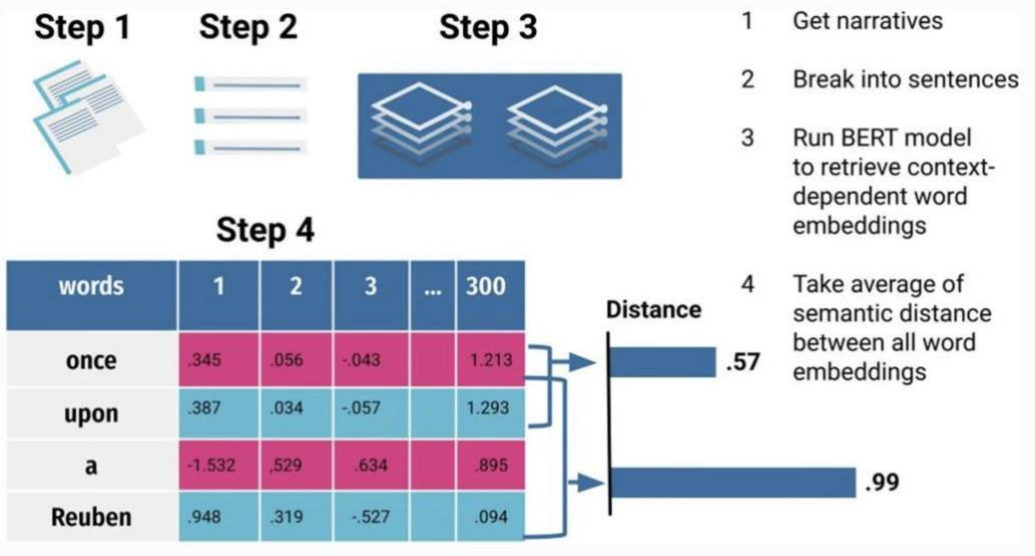
What languages does SemDis support? The BERT and RoBERTa models incorporated in our app have multilingual capabilities and support 12 languages. The languages supported include Arabic, Chinese, Dutch, English, Farsi, French, German, Hebrew, Italian, Polish, Russian, and Spanish.

AE CN NL GB IR FR DE IL IT PL RU ES

What steps does CAP take to calculate semantic distance? Once a file is uploaded and its language is identified, CAP computes the MAD using one of two multilingual language models: Multilingual Bidirectional Encoder Representations from Transformers (MBERT) or Cross-lingual Language Model RoBERTa (XLMR). The choice between MBERT and XLMR, as well as the specific layers within these models employed for computing semantic distance, hinges on the language in question. For more information, see the study by Patterson et al. (2023), which validated semantic distance in 12 languages and identified the model and layer combinations that exhibited the strongest correlation with human creativity ratings for each language. It's worth noting that MAD performance can vary among languages, and the models have only undergone validation for the AUT, not for other tasks.

Understanding Divergent Semantic Integration (DSI)

What is DSI? DSI is an extension of the semantic distance (SemDis) to longer form writing, proposed in Johnson et al. (2022). DSI captures the extent to which a story connects divergent ideas. Mathematically, it is the mean semantic distance between all pairwise comparisons of words within a body of text, where higher scores indicate that the story connects more divergent ideas. DSI scores tend to fall in the range of .70 - .90, but 0 - 1 is the full range of possible DSI values.



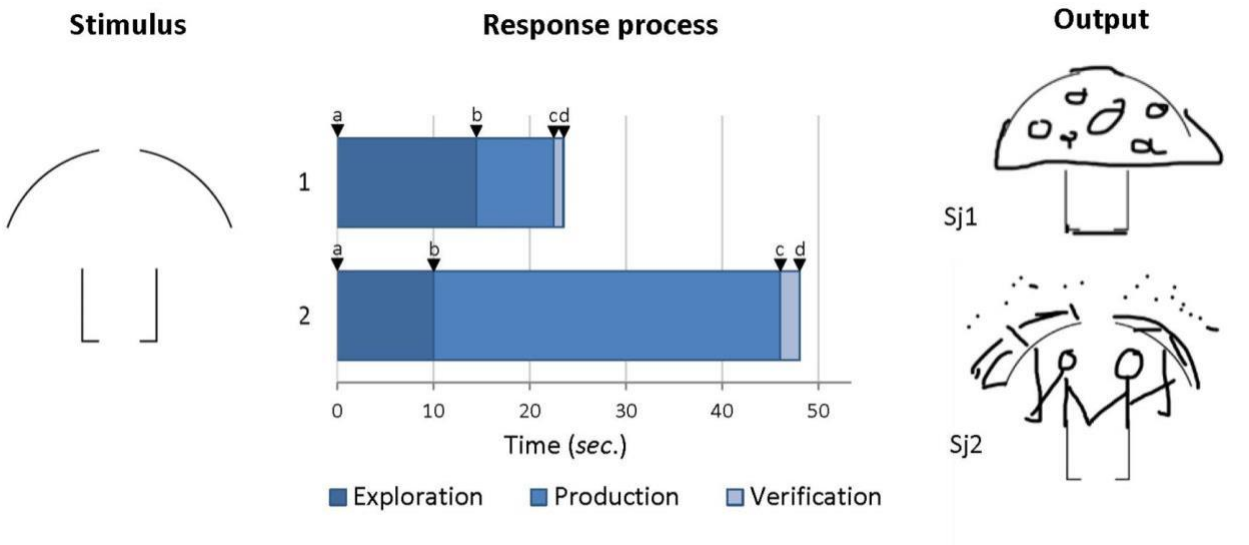
How does DSI work? DSI scores on CAP are generated using word vector representations from BERT large—a pre-trained 24-layer language model in which each word provided to the model is represented in 24 different ways, one way per layer. CAP employs the computational approach that, empirically, best matches human creativity ratings (Johnson et al., 2022). Specifically, early to middle layers in BERT have been shown to represent syntactic and semantic information (Jawahar et al., 2019), and layers 6 and 7 have been shown to match human creativity ratings particularly well in short narratives (Johnson et al., 2022, Supplemental Material). Concordant with empirical findings, CAP calculates DSI scores based on word representations from both layer 6 and layer 7 of BERT, where each word is represented twice in the DSI computation (once by layer 6 and again by layer 7).

What languages does DSI support? At this time, only English. As such, we strongly recommend against using DSI for other languages unless the inputs are first translated into English (e.g., via Google Translate).

How well does DSI perform? DSI has been shown to correlate strongly at the person level (*r* ~= .7) and at the individual response level (*r* ~= .6).

Understanding Automated Drawing Assessment (AuDrA)

What is AuDrA? AuDrA is a finetuned convolutional neural network that has been trained to assess drawings that result from the 'Incomplete Shapes' variant of the Multi Trial Creative Ideation (MTCI) task of Barbot (2018; shown below).



How does AuDrA work? AuDrA generates a creativity prediction for each input image that falls in the range of 0-1, though scores tend to fall in the 0.25-.75 range. The higher the rating, the more creative AuDrA believes the drawing to be. AuDrA has been shown to correlate strongly with human creativity ratings at the response level (i.e., individual drawings; see Patterson et al., 2023), with a Pearson correlation of .8 on the 'Incomplete Shapes' task it was trained to perform assessment on.

Can I use AuDrA to score other drawings? AuDrA will take in any drawings you upload. However, that does not mean you *should* upload drawings from other tasks. Finetuned convolutional neural networks, like AuDrA, learn domain/task specific knowledge that often does not transfer to other tasks/domains. Patterson et al. (2023) showed that AuDrA correlated much less with human ratings (r = .5) compared to the task it was trained on (r = .8). Accordingly, it is recommended to use CAP's Drawing task (i.e., a recreation of Barbot's MTCI task) for accurate results.

What languages does AuDrA support? AuDrA scores depend only on the contents of the uploaded drawing; language model representations were not used to train AuDrA. The drawings and human creativity ratings used to train AuDrA were obtained by participants in multiple countries: mainly from Belgium and the United States. Although it is possible cultural variables may affect the content of the drawings (and possibly the model's accuracy), we are unaware of any effects of language, directly, on AuDrA's accuracy.

How well does AuDrA perform? On the drawing task it was trained on (and CAP's Drawing Task), AuDrA correlates strongly with human creativity ratings for drawings it was never trained on (r = .8).

Understanding Multilingual Assessment of Short Stories (MAoSS)

What is MAoSS? MAoSS is an approach to automatically assessing the creativity of short stories written in multiple languages using large language models (LLMs). It was developed and validated using stories generated from the Short Story Task, where participants are presented with a three-word prompt (e.g., stamp-letter-send) and asked to write a creative short story that includes all the prompt words and is five sentences long.

How does MAoSS work? MAoSS employs fine-tuned LLMs to predict human creativity ratings for short stories. The LLMs undergo supervised training on datasets containing short stories along with their corresponding human creativity ratings. During this fine-tuning process, the model learns to associate certain linguistic features in the stories with higher or lower creativity scores. Once trained, the model can then predict creativity scores for new, unseen stories.

What languages does MAoSS support? MAoSS has been validated on short stories written in 11 different languages: Arabic, Chinese (Mandarin), Dutch, English, French, German, Hebrew, Italian, Polish, Russian, and Spanish.

AE CN NL GB FR DE IL IT PL RU ES

Two different approaches were tested:

An English-only model (RoBERTa-base) was fine-tuned on a dataset containing English stories and multilingual stories machine translated into English. This model can score the creativity of stories in any of the 11 languages as long as they are first translated into English.

A multilingual model (XLM-RoBERTa-base) was fine-tuned on the same dataset, but with the multilingual stories kept in their original languages. This model can score creativity directly in any of the 11 languages without needing translation.

As the multilingual model yielded stronger results, we provide access to the multilingual version on CAP.

How well does MAoSS perform? Both the English-only and multilingual models demonstrated strong performance in predicting human creativity ratings across all 11 languages (correlations ranging from r=.72 to r=.87). This exceeded the performance of other automated methods like semantic distance scores and word count.

The English-only model performed best on English (r=.86) and Polish (r=.85) stories that were machine translated into English. The multilingual model also had the highest correlations for English (r=.87) and Polish (r=.87) stories in their original languages.

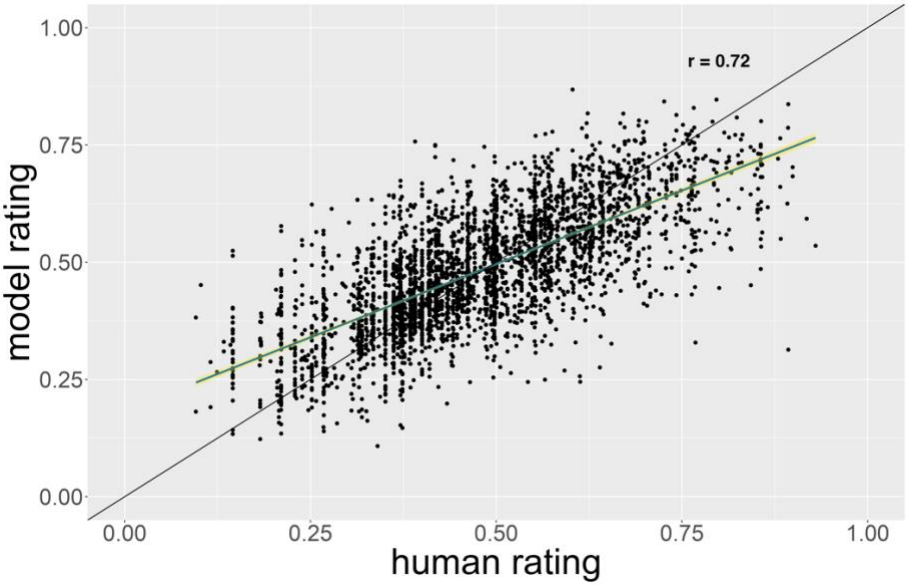
Understanding Automated Scoring of the Scientific Creative Thinking Test (SCTT-AI)

What is SCTT-AI? SCTT-AI is an approach to automatically assessing scientific creative thinking using the RoBERTa-base language model. It was developed and validated using responses to the Scientific Creative Thinking Test (SCTT), where participants are presented with open-ended scientific problems and asked to generate novel and plausible solutions in three areas: hypotheses, research questions, and experimental designs.

How does SCTT-AI work? SCTT-AI employs a fine-tuned version of the RoBERTa-base language model to predict human originality ratings for SCTT responses. The model undergoes supervised training on a dataset containing SCTT responses along with their corresponding human creativity ratings. During this fine-tuning process, the model learns to associate certain linguistic features in the responses with higher or lower creativity scores. Once trained, the model can then predict creativity scores for new, unseen SCTT responses.

What languages does SCTT-AI support? SCTT-AI currently only supports responses in English, but multilingual support is planned for the future.

How well does SCTT-AI perform? The fine-tuned RoBERTa-base model demonstrated strong predictive accuracy for human creativity ratings. On a held-out test set, the model achieved a correlation of r = .72 with human ratings, substantially exceeding a word count baseline (r = .31). Critically, the model also generalized reasonably well to SCTT items that were not included in its training data, with a correlation of r = .46 on these untrained items—still surpassing the word count baseline (r = .29). This suggests that SCTT-AI has learned to capture the underlying features that drive human creativity judgments, rather than simply overfitting to specific items or relying on superficial cues like response length.



Understanding Cross-Lingual Alternate Uses Scoring (CLAUS)

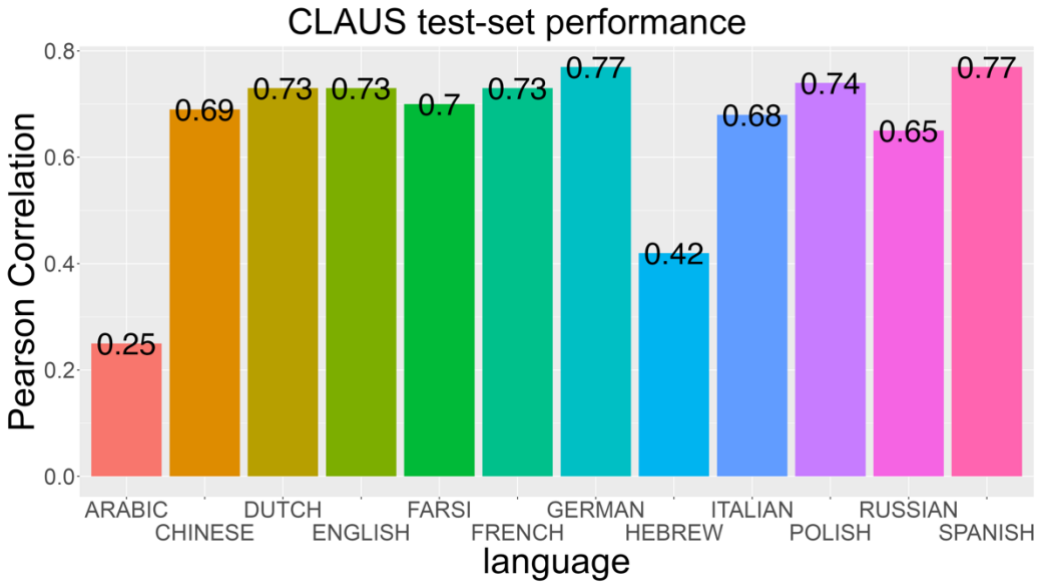
What is CLAUS? CLAUS is an approach to automatically scoring Alternate Uses Task (AUT) responses across multiple languages using fine-tuned large language models (LLMs). It was developed and validated using AUT responses from 12 languages, where participants are presented with a common object (e.g., a brick) and asked to generate as many unusual uses for that object as possible within a given time limit. CLAUS was trained with an English task description but with the AUT item and response in the native language.

How does CLAUS work? CLAUS employs a fine-tuned version of the XLM-RoBERTa language model to predict human creativity ratings for AUT responses. The models undergo supervised training on datasets containing AUT responses along with their corresponding human creativity ratings across 12 languages. During this fine-tuning process, the models learn to associate certain linguistic features in the responses with higher or lower creativity scores. Once trained, the models can then predict creativity scores for new, unseen AUT responses in multiple languages.

What languages does CLAUS support? CLAUS has been validated on AUT responses in 12 languages: Arabic, Chinese, Dutch, English, Farsi, French, German, Hebrew, Italian, Polish, Russian, and Spanish. However, due to smaller training sets, CLAUS performs much worse on Arabic and Hebrew. We encourage researchers to translate to English for these languages.

AE CN NL GB IR FR DE IL IT PL RU ES

How well does CLAUS perform? The fine-tuned RoBERTa-base model demonstrated strong predictive accuracy for human creativity ratings on an English-translated dataset across all 12 languages. The XLM-RoBERTa model, which was trained on untranslated responses, also showed good performance across languages, although there was some variation in accuracy between languages. Importantly, both models were able to generalize to new AUT responses that were not included in their training data, suggesting that they have learned general features of creative language use rather than overfitting to specific responses.



Understanding Artificial Intelligence for Design Evaluation (AIDE)

What is AIDE? AIDE is an approach to automatically score responses from the Design Problem task via fine-tuned large language models (LLMs). It was developed and validated using English responses to the Design Problems task.

How does AIDE work? AIDE employs a fine-tuned version of the RoBERTa-base language model to predict human creativity ratings for Design Problem task responses. The model undergoes supervised training on datasets containing Design Problem responses along with their corresponding human creativity ratings. During this fine-tuning process, the models learn to associate certain linguistic features of the responses with higher or lower creativity scores. Once trained, the model can then predict creativity scores for new, unseen Design Problem responses.

What languages does AIDE support? Currently, AIDE supports only English but multilingual support is planned. For responses in other languages, we recommend using Google Translate to translate responses into English.

How well does AIDE perform? The fine-tuned RoBERTa-base model demonstrated strong predictive accuracy for human creativity ratings across all prompts the model was trained on (r = .78). Critically, AIDE was also shown to accurately predict human creativity ratings for prompts the model was never trained on (red bars below)—indicating the model's robust ability to generalize far beyond the training data.