Published Citation

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

[Preprint – June 2023]

Beyond Semantic Distance: Automated Scoring of Divergent Thinking Greatly Improves with Large Language Models

Peter Organisciak[1], Selcuk Acar[2], Denis Dumas[3], Kelly Berthiaume[2]

[1]University of Denver

[2]University of North Texas

[3]University of Georgia

Author Note

Peter Organisciak https://orcid.org/0000-0002-9058-2280

Correspondence concerning this article should be addressed to Peter Organisciak, Department of Research Methods and Information Science, University of Denver, 1999 E. Evans Ave, Denver, 80208, United States.

Email: peter.organisciak@du.edu

Preprint Version History
v.1 – August 2022
v.2 – March 2023 – update with ChatGPT and GPT-4
v.3 – June 2023

Abstract

Automated scoring for divergent thinking (DT) seeks to overcome a key obstacle to creativity measurement: the effort, cost, and reliability of scoring open-ended tests. For a common test of DT, the Alternate Uses Task (AUT), the primary automated approach casts the problem as a semantic distance between a prompt and the resulting idea in a text model. This work presents an alternative approach that greatly surpasses the performance of the best existing semantic distance approaches. Our system, *Ocsai*, fine-tunes deep neural network-based large-language models (LLMs) on human-judged responses. Trained and evaluated against one of the largest collections of human-judged AUT responses, with 27 thousand responses collected from nine past studies, our fine-tuned large-language-models achieved up to $r = .81$ correlation with human raters, greatly surpassing current systems ($r = .12 – .26$). Further, learning transfers well to new test items and the approach is still robust with small numbers of training labels. We also compare prompt-based zero-shot and few-shot approaches, using GPT-3, ChatGPT, and GPT-4. This work also suggests a limit to the underlying assumptions of the semantic distance model, showing that a purely semantic approach that uses the stronger language representation of LLMs, while still improving on existing systems, does not achieve comparable improvements to our fine-tuned system. The increase in performance can support stronger applications and interventions in DT and opens the space of automated DT scoring to new areas for improving and understanding this branch of methods.

Keywords: divergent thinking; alternate uses test; large-language models; automated scoring

Introduction

Historically, divergent thinking (DT) research has been restrained by measurement challenges. By their nature, tests of DT are formulated in an open-ended way, which increases the time, effort, and cost of measurement. Recent advances in DT research, however, have found that automated methods can reliably score at least one type of DT task, the Alternate Uses Task (AUT; Beaty & Johnson, 2021; Dumas & Dunbar, 2014; Dumas et al., 2020). These methods capitalize on the natural property of text-mining models to calculate semantic distance or relationships between words as a measurable distance and use that distance as a proxy for how divergent an idea is from a prompt. An elegant feature of this approach is that it is effectively *unsupervised*, in that it does not require learning from training examples.

However, there are limitations to the current semantic distance approach to automated DT scoring. First, the specific semantic models that are used have been outperformed in most other applications within natural language processing (Wang et al., 2018; Wang et al., 2019). New advancements in a class of deep-neural network-based models have shown a consistently stronger and more robust grasp of the relationships between word concepts (Devlin et al., 2018; Liu et al., 2019; Radford et al., 2018; Raffel et al., 2020; Vaswani et al., 2017; Yang et al., 2019). Further, the semantic models currently used in automated scoring only understand text as a set of independent words, whereas newer models account for the complexities of context when words are used together to communicate a sentence or passage.

In this paper, we demonstrate a significant improvement in performance over existing AUT scoring methods by fine-tuning Large Language Models (LLMs)—a class of neural network-based approaches to modeling text—to score originality of AUT responses based on learned examples. We also measure the ability of this approach to scale to new prompts, the effect of training size on model strength, and the strength of LLMs without any fine-tuning. Our system, *Ocsai*, is available for free online.

The new approach that we introduce here is *supervised*, in that it is given a set of input-output pairs to learn from, toward being able to predict outputs for never-before-seen inputs. Supervised learning was previously applied to the AUT by Buczak et al. (2022), who used feature engineering to extract salient information for use with machine learning regression, and Stevenson et al. (2020), who paired clustering on semantic model embeddings with human judged responses, assigning new responses the score representative of their closest cluster. In contrast, our approach uses fine-tuning, where neural network-based model that has already been trained–in this case, a Large Language Model trained on general texts–is further trained on task-specific data. In doing so, a classifier can build from a strong foundational knowledge of language in learning how to interpret the relationship between an input and out – such as that between an AUT response and a multi-rater judgement of its originality.

The improvements presented here are a strong leap over the current state-of-the-art approaches and provides evidence suggesting that they can be further improved with additional training data, larger models, and more iteration of the approach. LLMs such as the ones applied here—T5 (Raffel et al., 2020) and GPT-3 (Brown et al., 2020)—better reflect the current best

approaches in other text-based domains, in tasks such as sentiment analysis (Socher et al., 2013), commonsense reasoning (Roemmele et al., 2011), and question answering (Rajpurkar et al., 2016). LLMs also introduce new challenges to the study of automated DT scoring. For one, deep neural network models, with millions or even billions of parameters, require novel approaches toward explaining their complex and nuanced internal logic (Barredo Arrieta et al., 2020; Gunning et al., 2019; Kojima et al., 2022). Additionally, the use of training data requires large, high-quality item-level ground truth, and may require more data-sharing coordination within the DT measurement community.

This study asks the following research questions representing *performance, robustness,* and *transferability*:

RQ1: How do LLMs affect performance for automated AUT scoring?

RQ2: What is the performance effect of different training size, including prompt-based few-shot and zero-shot contexts?

RQ3: How well do trained LLMs transfer to unseen prompts?

These three questions study whether LLMs improve on existing automated scoring models, to what degree and in what contexts, and how well the perform in previously unseen situations. In addition to fine-tuning LLMs, we also measure a prompt-based approach without fine-tuning.

Background

Divergent Thinking and the Alternate Uses Task

Tests of DT date back to the cognitive assessment efforts that emerged in the early 20th century (Plucker et al., 2022). Beginning with Guilford (1950), along with the seminal works by Torrance (1966, 1980; see Kim, 2006 for a review) and Runco (1991; 2013), DT tests have become the most popular method of creativity assessment (Snyder et al., 2019) in both psychoeducational settings and research. Quite a few longitudinal (Cramond et al., 2005; Runco et al., 2010; Torrance, 1972; Zaccaro et al., 2015) and meta-analytic studies (Kim, 2008; Said-Metwaly et al., 2022) have provided evidence of their predictive power. DT tests are often used to predict creative potential (Runco & Acar, 2012) and are sometimes used for gifted identification.

There are various types of DT tests besides AUT such as Consequences, Instances, Similarities, Realistic, Problem-Generation, Line Meanings, Picture Construction, Picture Completion, Pattern Meanings, and Asking Questions (Runco et al., 2016; Torrance, 1966; Wallach & Kogan, 1965). Due to the open-ended nature of these tasks, there are many different methods of scoring, but the conventional indices comprise fluency (number of produced responses), flexibility (diversity of responses), originality (unusual, uncommon responses), and elaboration (level of elegance and detail). Torrance Tests of Creative Thinking is one example of a creativity assessment that builds on the principles and structure of DT tests where several DT tasks with varying type of task structures are integrated (Acar, 2023). For the past several decades, AUT has been the most popular type of DT test used in research, though not necessarily the strongest (Runco et al., 2016). In AUT, participants are asked to generate uses for everyday objects. Besides this essential structure, AUT has some variations in terms of explicit

instructions and specific prompts used. For example, the AUT in the verbal form of the Torrance Tests of Creative Thinking (referred to by Torrance as the Unusual Uses Test) uses the prompts tin can and cardboard box, whereas Wallach and Kogan's test (1965) uses newspaper and knife. Like other DT tasks, AUT has been scored manually using human judges until recently. The next section presents these novel methods that are successful in quickly scoring responses produced for DT tests.

Automated AUT Scoring

Computational approaches to scoring creativity have existed for over fifty years (Forthmann & Doebler, 2022; Paulus, 1970; Paulus & Renzuli, 1968). The modern study of automated AUT scoring, however, has its roots in a method known as Latent Semantic Analysis (LSA; Deerwester et al., 1990; Landauer & Dumais, 1997). LSA is a method from information retrieval research that finds relationships between words by analyzing their cooccurrence in texts. It does so by building a term-document matrix of word counts and performing Singular Value Decomposition to reduce it to a lower-dimensional representation. A component of this decomposition is a semantic term-document space where similar words are closer together in a measurable, geometric sense (i.e., they are a similar mix of latent topics). Landauer and Dumais (1997) popularized LSA in psychology, noting that this approach approximated how humans themselves parse information from a relatively terse language. Since LSA, there have been several similar types of modeling approaches (e.g., pLSA, Hofmann, 2013; Non-Negative Matrix Factorization, Lee & Seung, 1999; Latent Dirichlet Allocation, Blei et al., 2003). More recently, a class of 'word embedding models' have trained semantic models on co-occurrences in word context windows (e.g., Word2Vec, Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013; GloVe, Pennington et al., 2014; fastText, Bojanowski et al., 2017). With word embedding models, it also became more commonplace to share models pre-trained on massive corpora of generalized English texts.

The current state-of-the-art automated AUT scoring approach is a clever use of semantic scoring models like LSA. In a well-trained semantic model similar concepts hold council near each other, while disjoint concepts are further apart. The AUT seeks to measure thinking that is divergent, surprising, or disjoint from a given prompt: a goal that aligns neatly with distance within a semantic model. For automated scoring, each prompt and response are projected to vectors in the semantic space. The cosine of the angle is then used to calculate the distance between these vectors, or the semantic distance. This is the underlying approach taken by currently utilized automated AUT systems like Open Creativity Scoring (OCS, Organisciak & Dumas, 2020; https://openscoring.du.edu) and SemDis (Beaty & Johnson, 2021; http://semdis.wlu.psu.edu/). OCS and SemDis differ in terms of how texts are preprocessed, how phrases are handled, the training texts used to inform the models, and the semantic space that they use, but both operate under the same semantic space distance principle.

Overall, these approaches have sought similar goals with varying methods. In this paper, *semantic model* is used to refer to the general body of modeling approaches that allows linearly comparable distances between words as a stand-in for similarities of semantics between those

words, whether by LSA or a word embedding approach. This approach to automated AUT scoring is unsupervised, as word distance can be calculated without any underlying knowledge of the task or responses. Recent work has also explored *supervised approaches*. As previously mentioned, Buczak et al. (2022) took a feature engineering and classification approach, augmenting word embeddings with other features and training predictive models using regressors such as Random Forests and XGBoost. Stevenson et al. (2020) clustered word embeddings of human-judged responses, assigning each cluster a mean human score and scoring new responses based on the cluster they best fit with. Finally, though not pursued as a supervised learning application, Beaty and Johnson (2021) found that raw semantic scores could be improved by a latent factor weighting, suggesting that tuning based on prior knowledge may provide a much higher ceiling for explaining originality scores than semantic distance alone.

The present study approaches automated AUT scoring directly as supervised learning. Our underlying premise is the same, though the methods differ from Buczak et al.'s (2022) approach. Prior work has used the paradigm of *fully supervised learning with feature engineering* (Liu et al. 2021), which extracts salient information in responses and trains classifiers with that information. Our work follows the paradigm of *pre-train and finetune* (Liu et al., 2021), where a neural network model is pre-trained on a great deal of general data and is then trained to a task-specific objective with training label. In this study, we use pretrained large language models from Text-to-Text Transfer Transformer (T5; Raffel et al., 2020) and Generative Pre-trained Transformer-3 (GPT-3; Brown et al., 2020).

Recent Innovations in Natural Language Processing

Recent years have seen a great deal of change in natural language processing approaches. Improvements to deep neural networks in the machine learning community have unlocked ways to model text with more nuance and complexity. One major innovation in text modeling is the *transformer* architecture*,* which utilizes a concept called *attention* (Vaswani et al., 2017). Modeling language as sequences of words has been a long-time goal in natural language processing, but traditional recurrent neural networks have run into limits due to computational complexity. Attention makes it computationally tractable for a transformer model to consider a long sequence of text, by selecting parts of the sequence that are most important. This allows training large models on not just words, but the complex contexts in which those words occur. Two notable early transformer-based architectures were Bidirectional Encoder Representations from Transformers (BERT; Devlin et al., 2018) and GPT (Radford et al., 2018). BERT was a landmark model and other architectures subsequently optimized or improved upon it including a Robustly Optimized BERT (RoBERTa; Liu et al., 2019), XLNet (Yang et al., 2019), and T5 (Raffel et al., 2020). Transformer-based models—sometimes called *Large Language Models* (LLMs)—generally outperform word embedding models (WEMs) on standard tasks in natural language processing, and often by large margins (Wang et al., 2018; Wang et al., 2019). This includes tasks which are notably similar to AUT scoring, such as semantic textual similarity (e.g., Raffel et al., 2020).

With LLMs, a process called *transfer learning* is often practiced in which the models are trained on extraordinary quantities of text and are released publicly for use so that other researchers and practitioners can build on their model of language rather than rebuilding it. In applications, a pre-trained model undergoes a form of supervised learning where it is *fine-tuned* to a given problem. In fine-tuning, some or all the neural network layers are unfrozen so the model can continue being trained for specific tasks that individual researchers choose, except examples represent a task-specific objective rather than generic texts. In some cases, an additional classification layer is appended to the end of the network.

In addition to transfer learning, larger LLMs have been found to be few-shot learners and even effective at some zero-shot tasks (Brown et al., 2020). Few-shot and zero-shot tasks do not fine-tune a pre-trained LLM, and instead directly query the 'out-of-the-box' or 'vanilla' model to respond from its general understanding of language. This works well with generative models (e.g., GPT-3; Brown et al., 2020) or text-to-text-models (e.g., T5; Raffel et al. 2020), which can take a plain text input and provide a text response. For zero-shot, no examples of correct answers are provided (e.g., an input might ask, 'how original is this use for a brick: {user response}?'. Few-shot functions similarly but shows a small number of examples of correctly scored responses, still in the input to a vanilla model. Both zero-shot and few-shot are sensitive to the manner of constructing the input prompt.

This study focuses on the efficacy of LLM approaches for AUT scoring, measuring the value of both fine-tuning and few-shot, prompt-based scoring. In fine-tuning, a standard LLM model is fine-tuned with examples of human creativity judgments to gain a better sense of originality. Later, we measure few- and zero-shot approaches with GPT-3, ChatGPT, and GPT-4, finding that the raw model without fine-tuning does have some sense of originality out of the box, but it is improved with training.

Data

The data used in this study was compiled from several prior studies with available item-level human judgments of AUT responses as well as recent research with elementary-aged participants from the Measure of Original Thinking in Elementary Students (MOTES) project (Dumas et al. 2023; Acar et al. 2023). The federated data in this study is potentially the largest collection of human-judged item-level AUT responses compiled, with 27,217 responses from 2,039 participants across nine datasets.

The overview of the datasets is as follows, organized by the identifier for each that is used for reporting.

1. betal18 (Beaty et al., 2018): This dataset used AUT prompts for *box* and *rope*, administered to 171 adult participants, resulting in n = 2,918 total responses. Responses were judged by four raters, with an averaged random Intraclass correlation coefficient (*ICC2k*) of .81.[1]

[1] Data is available at https://osf.io/gz4fc/

2. bs12 (Beaty & Silvia, 2012): This dataset used a single prompt, *brick*, with 133 college-aged adults. Responses were judged by three raters (n = 1,807, *ICC2k* = .72).[1]
3. dod20 (Dumas et al., 2020): This dataset consists of 10 AUT prompts - *book, bottle, brick, fork, pants, rope, shoe, shovel, table, tire*—administered to 92 participants, and comprises 5,435 total ratings. It was scored by three raters (n = 5,435, *ICC2k* = .85).
4. hmsl (Hofelich Moer et al., 2016): This data comprises 638 participants and two AUT prompts, for *paperclip* and *brick*. Four judges rated the responses (n = 3,843, *ICC2k* =.67).[2]
5. motesf: This dataset is a previously unreleased dataset associated with the Measuring Original Thinking in Elementary Students (MOTES) project, a study developing a DT test for elementary-aged students. The data used here is spelling-corrected data from an AUT portion of the measure, with 8 prompts administered to 385 participants and judged by 4 raters (n = 2,924, *ICC2k* = .73).
6. motesp: This data corresponds to a pilot version of the *motesf* data, with 35 participants and the same prompts, as well as *backpack* and *shoe*. (n = 339, *ICC2k* = .81).
7. setal08 (Silvia et al., 2008): This research studied DT through six tasks, including consequences, instances, and AUT. This study uses the AUT, which asked 241 participants for creative uses for a *brick* and a *knife*. Three judges rated the originality of responses, with (n = 3,425, *ICC2k* = .48).[3]
8. snb17 (Silvia & Beaty, 2017): In this data, 142 college students were administered two AUT prompts: *box* and *rope*. Responses were judged by three raters (n = 2,272, *ICC2k* = .67).[1]
9. snbmo09 (Silvia et al., 2019): Finally, in this dataset, 202 college-aged students were asked to develop alternate uses for three tasks: *brick*, *knife*, and *box*. In the originating study, 13 participants were removed for low engagement; this study uses all data available. Responses were judged by four raters (n = 4,099, *ICC2k* = .69).

Table 1

Counts of Rated AUT Responses Prior to De-duplication

Dataset	Responses
motesp	963
bs12	1807
snb17	2372
betal18	2918
motesf	2924

[2] Data is available at https://conservancy.umn.edu/handle/11299/172116
[3] Data is available is at https://osf.io/9dnx7.

setal08	3425
hmsl	3843
snbmo09	4099
dod20	5490

In all datasets, individual AUT responses were coded by multiple human raters. In past research, this process involved multiple raters who scored the responses individually (Hass et al., 2018; Silvia, 2011), selected maximal subset (Benedek et al., 2013; Shaw, 2021; Silvia, 2011) or as a total response set (Acar et al., 2023; Runco & Mraz, 1992; Silvia, 2008). Scoring originality can be considered as more of a *normative* task than an objective one. What this means is that if enough raters are asked, they generally move toward a consensus on originality ratings, even if well-trained individual raters may differ on the difference between highly and moderately original responses. We considered the breadth of different raters from different projects to be a benefit to the flexibility of the trained model. However, despite the contextual diversity, raters tend to be students or scholars, and future work could benefit from a more deliberately approach to demographic diversity among human judges. This study used the mean of multiple ratings for a ground truth judgment of each response's originality. Data was rounded to the nearest 0.1. Datasets were remapped to a five-point scale if they did not originally use one, scaling linearly between the minimum and maximum score.

Data items were deduplicated, so only one of each prompt/response pair was preserved. In total, 7,015 responses were removed for being repeated responses, 25.8% of the data, resulting in a final data size of 20,202 responses. The most repeated response was to use a brick as a paperweight; the ten most common are shown in Table 2. Deduplication was done only on exact duplicates, so for example 'build a house' and 'make a house' were not considered duplicates.

Table 2

Most Repeated Responses to Prompt Items

prompt	response	count
brick	paperweight	169
brick	weapon	124
brick	paper weight	93
brick	door stop	89
rope	belt	84
box	hat	76
brick	build a house	67

In applied work with the AUT, deduplication would not necessarily be done, because it is common to see different participants give the same response. Some tests rely on past information about the common responses to help score originality (Torrance 1966, Torrance 1972). The motivation for excluding duplicate responses here is that duplicates greatly advantage supervised learning without revealing much about the underlying robustness of the tool. It is a form of *data leakage*, where some of the testing data makes its way into training data, and a serious challenge to reproducible results commonly seen in fields newly adopting machine learning (Kapoor & Narayanan, 2022). While in this case it is not necessarily an unrealistic advantage, this study aims to focus more on the robustness of LLMs in learning the bounds of the AUT and extending it to new responses or even new prompts. To put it another way, without de-duplication our system would look much stronger–but the source of that strength would tell us less about the generalizability of the approach or how well the system understands the *task*. It would also make the findings more biased to our specific corpus, because the level of duplication seen in our corpus is not particularly what other studies may have. We briefly report on an un-deduplicated model, however, for comparison with prior work.

Ground truth scores (i.e., human judged originality ratings) for repeating responses were averaged. For example, each participant that had 'paperweight' as a use for 'brick' was rated by a panel of human judges, and the ground truth presented in this study combined all those judgments into a single consistent score prior to de-duplication. In this way, although the repeated responses did not appear in our training dataset more than once, all the trained human judges who rated those repeated responses provided equally weighted information about what the true originality of that response was, because we averaged those ratings.

Input data was randomized and split into training, cross-validation, and test data. The response-level randomization and deduplication are incongruous with participant-level metrics, since no participants are wholly represented within the test data, and only response-level judgements are reported.

Computer Experiments

Data was prepared for experiments in *performance*, *robustness*, and *transferability*. These experiments align to the three research questions.

Performance: How do LLMs affect performance for automated AUT scoring?

To determine the measure a general performance of different scoring approaches, an 80-5-15 fully randomized training-cross-validation-testing split was used, where supervised learning methods were trained with 80% of the entire data and evaluation was performed on 15% of the dataset. The remaining cross-validation data is optionally used for seeing progress against a held-out set during training without compromising the testing data.

Robustness: What is the performance effect of increases in training data size?

Supervised learning requires training examples, but how many examples? Where the primary performance evaluation was split from all ground truth, it is also worth measuring the

robustness of supervised learning relative to training dataset size. Using the same split as used above, models were trained with smaller subsets of the training data, to see how the availability of training data affects the quality of the model.

Additionally, we evaluated prompt-based approaches (Liu et., 2022) which do not use any fine-tuning. Rather, they explicitly (i.e., in plain English) ask a pre-trained model to score the originality of a response. We compared prompts with five example scores provided, as well as with none.

Transferability: How do trained LLMs transfer to unseen prompts?

The above experiments focus on how the model learns to score new responses of known prompts. In considering transferability, this study also looks at how well LLMs can generalize the AUT task, to score new responses for never-before-seen prompts.

For this experiment, input data is split by prompt. The prompts represented in the training data and test data are mutually exclusive. Training prompts included *brick, rope, box, knife, book, table, tire, ball, lightbulb, pencil, shoe, sock, fork*, *hat, toothbrush*, and *backpack*. Prompts in the test set were *paperclip, spoon*, *bottle, shovel*, and *pants*.

Methods

This study compared supervised learning with two LLM architectures—T5 (Raffel et al, 2020) and GPT-3 (Brown et al., 2020) with a typical unsupervised semantic model baseline, using the implementations from Dumas et. al., (2020) and Beaty and Johnson (2021). Additionally, an embedding-distance unsupervised application of GPT-3 and T5 more in-line with semantic approaches is compared.

Baseline: Semantic Distance Methods

As a baseline, the current state-of-the-art automated scoring models are applied, which use distance metrics in semantic spaces. Specifically, scoring is applied from Open Creativity Scoring (OCS, Organisciak & Dumas 2020; Dumas et al., 2020), and SemDis (Beaty & Johnson, 2021).

SemDis includes five trained models, as well as an ensemble score which takes the mean of all five scores (Beaty & Johnson, 2021). The ensemble is chosen for reporting here, as *SemDis_MEAN,* as it is the authors' recommendation and the best performer. Additional parameter recommendations from Beaty and Johnson (2021) are also followed: using text cleaning with stoplist word removal and applying multiplicative rather than additive composition of words into phrases.

The Open Creativity Scoring baseline is based on work from Dumas et al. (2020) and uses the GloVe-based model recommended in that paper. GloVe is a form of word embedding model first described by Pennington et al. (2014) and released with a set of pretrained models. Here, the 300-dimension Gigaword 6B pre-trained model is used (Pennington et al., 2014), which was trained on 6 billion words from Wikipedia and the Gigaword 5 corpus (Parker et al., 2011). As with SemDis, the author-recommended parameters are followed: the system removes function words (*stoplisting)*, composes phrases with a mean of word vectors weighted by their

relative importance (*term weighting*), and avoids penalizing responses that reuse a prompt word by excluding the prompt word from responses.

Both baseline systems are available online: SemDis at http://semdis.wlu.psu.edu and Open Creativity Scoring at https://openscoring.du.edu.

Semantic Distance with Large Language Models

LLMs are generally not successful at out of the box semantic distance (Reimers & Gurevych, 2019). However, it is possible to build downstream semantic distance models on top of LLMs, learned from rated pairs of sentences (Neelakantan et al., 2022; Ni et al., 2021; Reimers & Gurevych, 2019). LLM-based embedding could potentially leverage the stronger internal representation of language seen in an LLM and better handling of phrases while still allowing for unsupervised use in the tradition of earlier automated DT scoring systems.

Here, embedding-based semantic distance scores are reported from GPT-3 Embeddings (Neelakantan et al., 2022) and Sentence-T5 (Ni et al., 2021). Each of these models have different sizes, as listed in Table 3. Note that the architecture of the Sentence-T5 models only include half of their corresponding T5 model, which is why the model named st5-3B has half of the 3 billion parameters that its naming suggests.

Table 3

Size of Embedding Models Trained on LLMs

Model	Model size	Embedding size
gpt-text-similarity-ada	300M parameters	1024
gpt-text-similarity-babbage	1.2B parameters	2048
st5-base	110M parameters	768
st5-large	335M parameters	768
st5-3b	1.24B parameters	768

Fine-Tuned Large Language Models

Large Language Models are the primary focus of this study. Here, two architectures are evaluated: T5 (Raffel et al., 2020) and GPT-3 (Brown et al., 2020).

T5: T5 is set of architectures introduced by Raffel et al. (2020). T5 was the product of a study comparing various improvements to BERT-like models. Since LLMs can often be improved by larger models and more training texts, it is sometimes difficult to determine where improvements are coming from. Raffel et al. (2020) compared different modeling approaches while controlling for computational resources and data contexts, releasing models for the best-performing architectures.

This study uses the *T5-Base* model, which has 220 million parameters in its network. There are three larger released T5 models, up to 11 billion parameters, which would likely improve task performance with a greater cost to implementation and transferability.

T5 is a text-to-text model, which casts all use into a format where text is given as input, and the model in turn provides output as text[4]. This may seem an ill-fit for AUT scoring, where the intent is to generate a continuous quantified variable, not unrestrained text generation. We apply no constraint to make the output a number; T5 could generate a soliloquy in the style of Hamlet if it desired to. Nonetheless, after fine-tuning, it learns that the response is expected to be numerical.

To accommodate the text-to-text format, training and inference inputs were formatted to the follow template:

`{prefix} {prompt} {response}`

where the prefix is '`autscore:`', the prompt is structured as "`question: What is a surprising use for X`" and the response is structured as "`response: Y`". For training, the target output was the ground truth score, with one precision point, input as a string of the number. For inference, this number was predicted and interpreted as a number. Demonstrative examples are shown in Table 4.

Table 4

Example Inputs and Outputs for T5

Input	Output
autscore question: What is a surprising use for a BOOK response: relay race marker	5.0
autscore question: What is a surprising use for a PANTS response: take them off	1.5
autscore question: What is a surprising use for a SHOE response: fungus grower	5.0
autscore question: What is a surprising use for a FORK response: utensil	1.0

GPT-3: GPT-3 is a model from OpenAI (Brown et al., 2020) that prioritizes text generation, aiming to predict what text follows an input. Generating realistic human-like text requires some variability, which would make for poor classification. However, it is possible to turn down the 'temperature', the parameter which affects how variable the sampling of new tokens is. With the temperature at zero, the model outputs best guess text, allowing it to be used in a similar text-to-text manner as T5.

The largest GPT-3 models have up to 175B parameters, while smaller released models have approximately 350M, 1.3B, and 6.7B parameters. This study fine-tunes a version of each model, which are referred to – from smallest to largest—as *ada*, *babbage*, *curie*, and *davinci*. GPT-3 is only available through a paid application programming interface (API) from OpenAI. This means its use and tuning have some costs, but processing occurs on a hosted server, lowering the complexity and computational needs of applying it.

[4] The title of this manuscript was suggested by an LLM in this manner, an example of the text-to-text principle that seemed appropriately demonstrative at the time of manuscript preparation in early 2022, but which has likely grown familiar to most readers in light of the widely-popular release of ChatGPT at the end of that year.

Zero-Shot GPT-3: Finally, a novel zero-shot approach is compared to test the inherent knowledge of a GPT-3, without fine-tuning. Zero-shot is a type of unsupervised learning where, despite not seeing prior examples, a model can infer sensible classes from a briefly described context. For example, one might ask "What is the originality of [sentence x] on a scale of 1-10", and a model may infer the nature of the task and the criteria (1-10). However, this approach does benefit from prompt engineering (i.e., carefully wording the input to the model to get the most useful output) to determine what types of questions are best understood by the model.

Results

Overall Performance for Replicating Human Judgements

The overall performance of the large language model methods, on randomized, deduplicated AUT responses, is presented in the ALL column of Table 5. Pearson correlation between the human-judged ground truth and the model prediction is reported. Additionally, correlation with humans on each sub-dataset is shown. Performance ranges from r = .12 to r = .81. Since the fine-tuned approaches are on the same scale as the human judgements, Root Mean Square Error (RMSE) is also provided. In addition to overall correlation, a mean of per-prompt correlations is provided, $\sum_{p \in P} \frac{r^p}{n(P)}$ for all prompts P (Table 6). Mean prompt correlation is less sensitive to difficulties with individual prompts and different prompt sample sizes. Mean prompt correlations range from r = .19 to .80.

The *hmsl* (Hofelich Moer et al., 2016) and *setal08* (Silvia et al., 2018) datasets were previously used for automated scoring by Buczak et al (2022), achieving response-level correlations of up .73 and .57, respectively, and RMSE of up to .51 and .45. Their work did not do a deduplication step, so for comparability, we trained a GPT-3 Davinci-sized model on data without deduplication, achieving respective performance on hmsl and setal08 of r= .82 and r=.77, with RMSE of .48 and .38. Though we caution again about data leakage, the overall performance of this un-deduplicated model does offer a glimpse into how it might perform in practice, where previously seen responses are typical. This model's overall correlation across all datasets is *r =.86 (RMSE=.45).*

Table 5

Overall Performance of Each Model

	ALL	**betal18**	**bs12**	**dod20**	**hmsl**	**motesf**	**motesp**	**setal08**	**snb17**	**snbmo09**
Baseline										
semdis-mean	.120	.210	.167*	.243	.155	.191	-.045†	.064†	.157*	-.020†
ocs-main	.256	.319	.178	.371	.364	.257	.337*	.328	.193	.295
LLM Embeddings										
st5-base	.223	.443	.227	.451	.278	.221	.421	.257	.295	.253
st5-large	.227	.425	.219	.385	.309	.226	.403*	.245	.307	.267
st5-3b	.204	.393	.260	.349	.280	.199	.455	.234	.288	.231
gpt3_emb-ada	.285	.396	.284	.429	.396	.382	.480	.314	.369	.254

gpt3_emb-babbage	.173	.320	.214	.352	.252	.214	.340*	.243	.319	.151
LLM Fine-tuned										
t5-base	.756 (.595)	.602	.593	.716	.725	.661	.484	.660	.627	.655
gpt3-ada	.764 (.573)	.627	.611	.715	.724	.686	.594	.694	.615	.686
gpt3-babbage	.792 (.541)	.704	.671	.758	.729	.730	**.755**	**.723**	.619	.727
gpt3-curie	.791 (.543)	.749	.651	.780	.725	.732	.616	.648	.643	.739
gpt3-davinci	**.813** (**.518**)	**.762**	**.712**	**.802**	**.730**	**.801**	.717	.697	**.698**	**.744**

Note: Performance measured in Pearson Correlation. Root Mean Squared Error of overall performance in parentheses for fine-tuned models. Overall performance presented as ALL; other columns present individual datasets. Best results, per condition, are marked in boldface. All results are significant at $p < .01$, except: *$p < .05$ and †$p > .05$.

Table 6 shows the performance of each prompt per model, along with the mean and standard deviation. The LLM models have lower standard deviation, while the semantic distance methods vary per prompt.

Table 6

Performance of Each Model Per Prompt

	backpack	ball	book	bottle	box	brick	fork	hat	knife	lightbulb	pants	paperclip	pencil	rope	shoe	shovel	sock	table	tire	toothbrush	M	SD
semdis-mean	.09	.09	.22	.10	.11	.11	.19	.30	.01	.28	.34	.14	.33	.19	.24	.10	.04	.34	.36	.16	.19	.10
ocs-main	.11	.38	.45	.46	.24	.30	.26	.41	.30	-.07	.41	.28	.30	.21	.17	.34	.35	.48	.61	.26	.31	.14
st5-base	.63	.51	.41	.43	.32	.22	.40	.08	.36	.23	.52	.34	.25	.33	.62	.54	.36	.34	.34	.27	.36	.15
st5-large	.35	.43	.34	.36	.33	.22	.46	.11	.31	.11	.48	.30	.23	.35	.54	.44	.37	.29	.30	.44	.33	.12
st5-3b	.39	.39	.23	.34	.29	.25	.38	.10	.28	.08	.41	.29	.24	.30	.47	.25	.38	.27	.38	.45	.30	.11
gpt3_emb-ada	.35	.60	.42	.45	.38	.29	.37	.42	.38	.29	.42	.45	.57	.31	.71	.38	.39	.39	.40	.53	.42	.11
gpt3_emb-babbage	.29	.47	.28	.33	.28	.21	.42	.24	.28	.26	.38	.37	.47	.30	.65	.42	.36	.30	.28	.55	.34	.12
t5-base	.55	.70	.77	.82	.61	.61	.53	.54	.72	.57	.86	.75	.76	.49	.70	.70	.51	.85	.82	.71	.68	.12
gpt3-ada	.40	.78	.68	.80	.65	.60	.79	.66	.75	.75	.83	.77	.73	.45	.73	.80	.53	.84	.85	.75	.70	.13
gpt3-babbage	.62	.81	.70	.88	.69	**.64**	.76	.65	.78	.72	.88	.73	**.85**	.48	.87	**.83**	.69	.69	.83	.86	.73	.75
gpt3-curie	.17	.79	**.78**	.86	.73	.61	.79	.60	.77	.71	.90	.75	.80	.54	.86	.81	**.71**	.87	.86	.80	.73	.16
gpt3-davinci	**.80**	**.84**	.71	**.88**	**.74**	**.64**	**.83**	**.77**	**.81**	**.83**	**.91**	**.79**	.85	**.56**	**.91**	.79	.69	**.90**	**.92**	**.81**	**.80**	**.09**

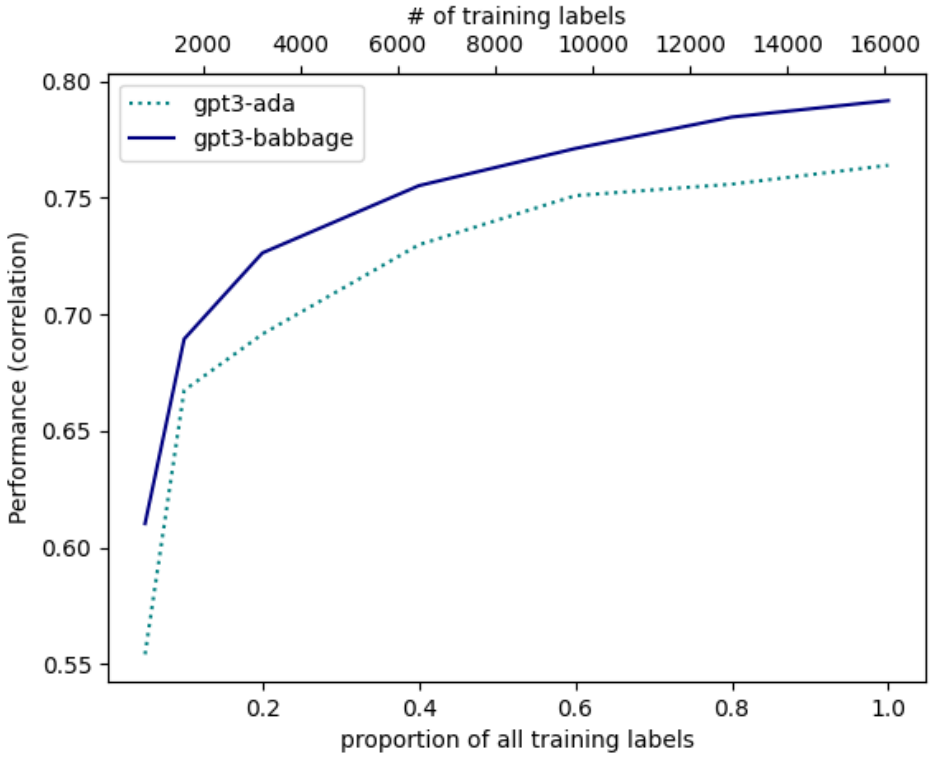
Note. Per-prompt mean correlation offers an alternative overall measure of performance. Best results, per condition, marked in boldface.

Robustness to Size of Training Data

How does training size affect quality? It has been noted that LLMs are few-shot learners (Brown et al., 2020), meaning they can learn a task from very few examples. Figure 1 shows the performance of the GPT-3 *ada* and *babbage*-sized models fine-tuned with different proportions of rating data. With 5% of the data (804 training labels), $r = .61$ for *babbage* and $r = .55$ for *ada* are still notable improvements over the baseline models. Additional training data is important early but begins to level off. For *babbage*, four-fifths of the total improvements seen in Figure 1 occur with just 40% of the data.

Figure 1

Effect of increasing training size



The larger model learns more effectively with fewer training examples. This is particularly apparent with even less training data: 160 labels, or 1% of the full set. With that low count of training labels, r = .48 with *babbage* while *ada* has a performance of r = .31, .64 of the performance of the bigger model. That relative performance ratio grows quickly to .91 with 804 labels, then is relatively stable between .95 – .97 from 10 – 100% of the training labels.

Prompt-based Few-shot Learning

With less than 200 training labels, the performance of a large language model is still notably stronger than the baseline models from SemDis and OCS. This begs two questions: how little data is needed to match baseline performance (few-shot), and how good can an LLM be without *any* training data (zero-shot), solely on the strength of its understanding of language.

We approached this question without any fine-tuning. Rather, we used an out-of-the-box 'vanilla' model and its own internal understanding of language. Using GPT-3's largest model size as of August 2022, as well as March 2023 snapshots of ChatGPT (based on Ouyang et al. 2022) and GPT-4 (OpenAI 2023). we constructed two prompt types. For few-shot engineering, we constructed a text prompt asking for a rating of how original each use is, then enumerated 5 training examples and 10 test examples, and provided ratings for the training examples. The choice to use 5 training examples in few shot was motivated by a conservative consideration of

costs per scored response, allowing for both training examples and multiple unscored responses to be contained within the prompt. As LLMs grow to allow longer input texts as well as lowering in cost, a worthwhile avenue for future research will be to measure few-shot with more training examples. This change is already being seen with ChatGPT (referred to as GPT 3.5), which has costs at a fraction of other models, and GPT-4, which allows 8,192 token in its base model. We present a brief example of prompt-based scoring with GPT-4 with 20 training examples and 20 test examples. Table 7 shows an example prompt in this style. The scale was multiplied by 10, due to how text generation functions: full numbers count as a single token whereas decimal numbers are three tokens, which means three predictive decisions for each score rather than one.

Table 7

Example of a Few-Shot Text-to-Text Prompt and Completion

Example Prompt	**Model Completion**
Below is a list of uses for a SOCK. On a scale of 10-50, judge how original each use for a sock is, where 10 is 'not at all creative' and 50 is 'very creative': USES 1. to use it like a puppet. 2. You can put googly eyes and make a sock puppet show. 3. You can color it and maybe make a snake. 4. a cool and funny puppet. 5. maybe you could put it on your hands and pretend to have superpowers. 6. using it as gloves. 7. you could use it for ASMR 8. Cut them and make a 3D sculpture. 9. you can make a dress for your doll 10. to use it like a backpack or store money in it RATINGS 1. 27 2. 27 3. 32 4. 24 5. 36 6.	20 7. 40 8. 50 9. 45 10. 35

Note: The prompt is what is provided to the model, including ratings for the first five items here. The completion is how the model continues from the prompt, starting after '*6.*' in this example.

For zero-shot, a similar prompt was used, but with no examples whatsoever. Rather, the model had to rely entirely on its own interpretation of 'how original each use for *X*' is. Since 10-50 is an unusual scale, we used 1-10 and divided by 2, meaning the few-shot completions were at a half-point step. Results are shown in Table 8.

Table 8

Zero-shot and Few-shot performance (GPT-3 DaVinci, ChatGPT, GPT-4), Overall and By Dataset

N	Model	ALL	betal18	bs12	dod20	hmsl	motesf	motesp	setal08	snb17	snbmo09
0	*GPT-3*	.13	.15	.17	.19	.17	.25	.09	.10	.31	.18
	ChatGPT	.19	.24	.17	.29	.26	.37	.52	.15	.28	.22
	GPT-4	**.53**	**.57**	**.52**	**.71**	**.64**	**.71**	**.67**	**.62**	**.52**	**.68**
5	*GPT-3*	.42	.18	.38	.43	.43	.38	.37	.24	.09	.36
	ChatGPT	.43	.32	.28	.52	.30	.50	.43	.12	.27	.34
	GPT-4	**.66**	**.64**	**.61**	**.72**	**.56**	**.72**	**.71**	**.62**	**.49**	**.62**
20	*GPT-4*	**.70**	**.63**	**.63**	**.75**	**.71**	**.73**	**.68**	**.62**	**.57**	**.67**

Transferability to Unseen Prompts

In addition to scoring unseen responses, can large language models learn the format of the alternate uses task itself, and be applied to never-before-seen items? Table 9 addresses this question, reapplying models trained on one set of unique prompt items and evaluating on another, entirely unseen set. This was conceptualized as the strongest test of the usefulness of supervised LLMs, because this is the use case that the previously established unsupervised learning approach (i.e., the semantic distance approach) should excel at since it never uses any training data to begin with. The best model performed at an overall r = .63 (.66 at prompt-level), while baselines showed r = .14 –.28 (mean of prompt r = .19 –.32).

Table 9

Per-Prompt Performance for Held-Out Prompts (Pearson Correlation)

model	bottle	pants	paperclip	shovel	spoon	ALL	M	SD
semdis-mean	.24	.20	.14	.15	.21	.14	.19	.04
ocs-main	.46	.27	.20	.44	.23	.28	.32	.11
t5-base	.49	.43	.46	.34	.29	.47	.40	.08
gpt3-ada	.71	.65	.55	.45	.54	.60	.58	.09
gpt3-babbage	**.78**	.69	**.56**	.67	.52	.63	.64	**.09**
gpt3-curie	.76	**.69**	.54	**.72**	**.58**	**.63**	**.66**	.08

Discussion

Fine-tuned Large Language Models Greatly Outperform Current Automated Scoring Approaches

The results demonstrate that fine-tuned large-language models outperform the current state-of-the-art approaches from Open Creativity Scoring (Dumas et al., 2020) and SemDis (Beaty & Johnson, 2021). The magnitude of the improvement is extraordinary. Evaluated against what may be the largest multi-human-judged dataset of AUT responses, the best SemDis model

showed performance of r = .12 *(mean of prompts* = .19*)*, while OCS performed at r = .26 *(mean of prompts* = .31*).* The fine-tuned LLMs presented here, T5 and GPT-3, ranged from r = .76 to r = .81. This held true not only across datasets, each with variation in raters and participants, but also across prompts.

Previous work has shown the value of supervised learning for AUT scoring (Buczak et al., 2022). This study supports those findings with a different style of supervised learning, fine-tuning deep neural network text models to learn the style of the AUT task and prompts. Presented is an initial exploration of these methods, and we expect there is a good deal of potential improvements yet to explore.

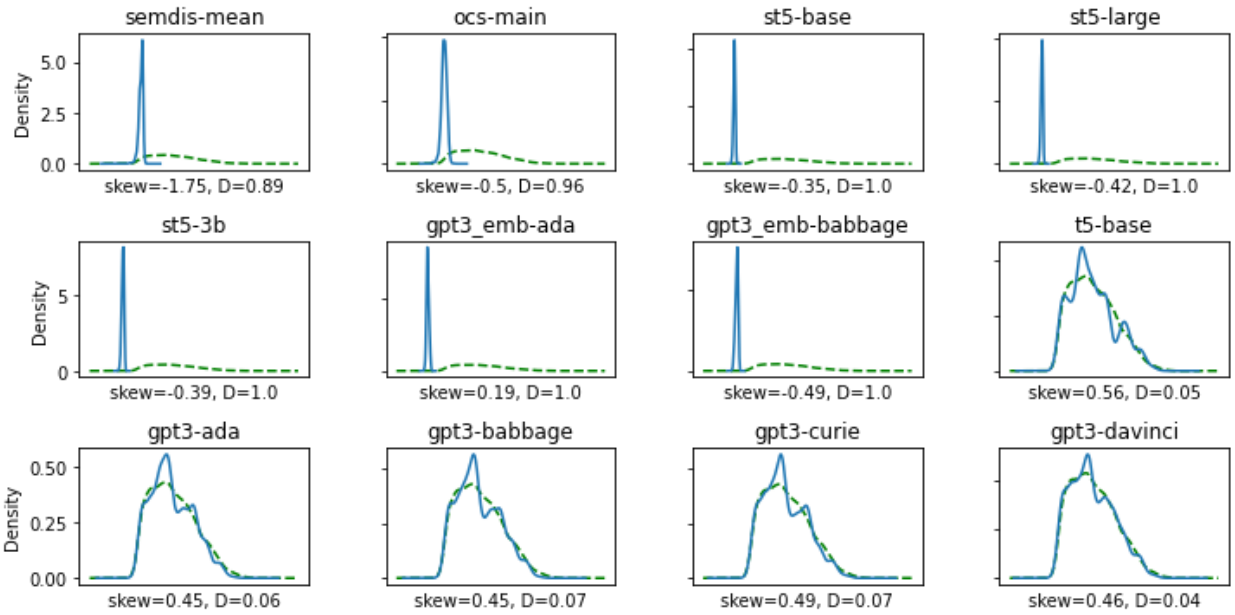
For context, the limit at which humans correlated with themselves is likely not much higher than the correlation between the LLMs and the humans in this study. For example, we examined responses given more than once, and found that among randomized pairings of duplicated response, human-judges correlated with other human judges at an average correlation of r = .83. Comparing a single response judgement to the less noisy mean of all judgments of its duplicates, the correlation was r = .88. This value might be interpreted as the approximate ceiling at which we could expect a model to correlate with human judgements.

The question remains: *why* are LLMs so capable in this context? Most of the benefit is likely from the apparent source: that LLMs have a strong and very robust understanding of language. The robustness of the models with few training labels supports this view. Yet, some of the improvements may also follow from other hidden patterns, beyond just understanding the task. For example, even though we removed exact duplicates, there are inexact duplicates, where the same concept is expressed with different words, punctuation, or spelling. For example, our deduplication removed 260 responses of ‘paperweight’ or ‘paper weight’, but one of each remained, as well as ‘weight to hold objects in place’ and ‘use as a weight’.

It's also likely that LLMs find another hidden pattern: the behaviors of given rater groups. Some raters may be more averse to the lowest or highest range, some may distribute their ratings more while others may strongly adhere to the mode. For the sake of discussion, Figure 2 shows the distribution of model predictions, along with their skew and Kolmogorov–Smirnov D to show goodness of fit. It shows how the LLM fine-tuning models hew much more closely to the human distribution than the baselines and LLM embedding approaches. Interestingly, the entropy of human judgments for a given item – a measure of how unpredictable the distribution is – does *not* correlate with our approach’s performance on that item (r =.004). Further work could benefit from embracing rater variance or disagreement, challenging supervised learning models with more of the unpredictability of humans to encourage more generalizable models.

Figure 2

Plots of Model Prediction Distribution Density Compared to Ground Truth Human Judgments



Note: Ground truth judgments shown in dotted line. Kernel density estimation used to compare distribution with different scales. Skewness and Kolmogorov-Smirnov statistic *D* noted. Ground truth skew is 0.51.

Large Language Models Can Be Robust with Only a Small Number of Training Examples

A benefit of unsupervised semantic-distance-based approaches such as those used by OCS and SemDis is that they do not need training. Yet, the advantage is small: *Ocsai* surpasses the performance of semantic distance models with only a small number of training examples, as low as 1% of our full training data. Our experiments trained for 18 AUT prompts simultaneously; in settings with only a handful of prompts, even fewer labels would be needed.

Further, the internal understanding of language in LLMs is competitive without any fine-tuning, particularly with new models such as ChatGPT and GPT-4. For a text-to-text model, crafting an appropriate prompt for text completion, referred to as *prompt engineering* (Liu et al., 2021), may be sufficient. Simply *asking* – literally, in plain English – a non-finetuned model to rate originality without having shown it examples of a good rating, we found a statistically significant but relatively low correlation for GPT-3 ($r = .13$, $p < .001$) and ChatGPT (r=.19, p<.001). comparable to the baseline models. The March 2023 release of GPT-4 (OpenAI) showed remarkable progress, far exceeding semantic models even for zero-shot (r=.53). Providing five examples of good scores in the question raised that performance to as high as r=.66 (GPT-4). Given its larger input limit, GPT-4 can take more examples in its prompt; with 20 prompt-based examples, performance is yet higher (r=.70*)*. LLMs are sensitive to prompt

engineering tweaks, and there are likely adjustments which improve the performance of this approach.

There is still a benefit to adding more labels, even though the performance improvements brought by each new label begin to level off. Observing so, it will be valuable for more DT researchers to share their response-level coded items, as done by the authors of the datasets studied here (Beaty et al, 2018; Beaty and Silvia, 2012; Dumas et al., 2020; Hoeflich Mohr et al., 2016; Silvia et al., 2008; Silvia et al, 2017; Silvia et al., 2019). Further, we argue that the community would benefit from a common benchmark dataset, normalized and doublechecked for consistency and quality, and with standardized response splits for training, cross-validation, and testing. Similar benchmarks are used in different communities (e.g., TREC for information retrieval, Voorhees & Harman 2005; SemEval, GLUE, and SuperGLUE for various natural language processing tasks, Wang et al., 2018; Wang et al., 2019; MIREX for digital musicology, Downie 2008), and would allow future supervised learning work to be comparable across publications.

There are Limits to the Semantic Distance Theoretical Model

In considering the future of semantic distance models, perhaps the finding of greatest import was what we observed in applying semantic distance embedding models based on LLMs. The LLM embeddings did improve on the baselines, but only marginally: an average performance of .22, compared to an average of .19 for the baselines. However, within the different models compared, a surprising trend occurred.

Among LLMs, the most important factor for performance is the size of the model (Kaplan et al., 2020). While the scale and quality of input data are also important (Hoffman et al., 2022, Raffel et al., 2020), when using the same data and architecture, we nearly always observe performance improvements with large models (Kaplan et al., 2020). This pattern did not hold for LLM embeddings in this paper: *st5-base* outperformed *st5-3b*, gpt-3's *ada* outperformed *babbage*.

This result challenges the assumption underlying the use of semantic distance as a proxy for DT, suggesting an upper limit to how effective semantic distance can be. As a model's understanding of language improved, the relationship of its semantic distance to originality began to weaken. The reason is yet unclear but the consequences for the automated scoring community are significant, and it bears future investigation to understand these limits, by studying how the smaller and larger models differ.

Transparency and Explainability of Machine Learning Models

A challenge of all automated scoring, semantic distance models and LLMs alike, is the risk of importing various biases from the training texts, or from the human judges used as a ground truth. In training the underlying models, the goal is generally to reflect the language seen in the originating text as realistically as possible. However, any broad corpus of texts will contain a litany of pervasive cultural biases. Some such biases are subtly applied in the English language, such as gender bias in discussing or representing various professions yet are undeniably harmful if codified going forward. In building automated models, particularly ones

which may one day be used in high-stakes educational settings (i.e., identifying gifted students in schools), we should hope for something better: more equitable and less biased than a general cross-section of human-written language.

There are some theoretical benefits of automated systems in tracking bias, though they need attention and action from researchers. Unlike open-ended tests which need to be scored by various trained judges, they can operate consistently across all collected datasets, and can be inspected for biases more directly. Whether a model is biased and to what degree is out in the open, and able to be published. In practice however, how the field would inspect for biases and what we would do about them could be a logistic challenge.

Inspecting for biases as well as developing 'explainable' artificial intelligence systems are active areas of study (e.g., Barredo Arrieta et al., 2020; Gunning et al., 2019). Generally, the reduced complexity of semantic approaches has been a benefit in this respect: cosine similarity is an explainable process rather than a tangle of complex decisions—a 'black box'—and the models themselves are easier to inspect for undesirable correlations. LLMs work much more closely to how we hope they would, but their complexity may allow for undesired biases to escape notice (Rudin 2019), or for other stakeholders (e.g., parents of respondents) to develop distrust in the artificially intelligent judgements. Some newer large language models even pre-empt high stakes use, out of concern for uninspected biases (BigScience, 2022). There are emerging approaches for making LLM rationale more explicit. Recent work reported that when presented with a reasoning task, asking LLMs to 'think it through step by step' not only results in a description of how they arrived at their conclusion, but the conclusions themselves tend to be more accurate (Kojima et al., 2022). A challenge, however, is that an LLM's explanation of its thought process can be 'hallucinated': it can offer a description, but that does not necessarily mean that is how the LLM made the choice. Another approach would be a regression analysis, to see how much of the LLM's score can be explained by a model of textual and contextual features. For example, elaboration has confounded semantic models (Forthmann et al., 2019) and the treatment for varying numbers of words has remained in debate, from term weighting (Dumas et al., 2020) to multiplicative composition (Beaty & Johnson 2021). If the confound still remains in LLM scoring, and to what degree relative to its effect on human judges, would be a valuable question for further investigation.

In the area of fairness, the supervised learning approaches investigated here present progress over semantic models. First, they show a greatly increased ability to mimic human raters—specifically, a composite of multiple raters, which softens the challenges of validity and possible biases that come with individual graders on open-ended tests. This means the models increasingly look more like our best alternative: multiple trained judges. Indeed, it offers us a way of thinking of these systems: they are an extra judge, one that hews closer to a panel of multiple raters with less inter-response variability than seen with individual raters. At the same time, as a reflection of human judgment, LLM automated scoring should also be critically approached and inspected for adverse bias in the same way that we do with humans. Secondly, supervised learning can continue to learn and improve. New human judgments or corrections can

be used to improve models, where errors found in semantic model scoring do not have easy correctives.

Explainability and fairness will be an important part of the discussion moving forward, and it is worth considering the activities of other communities in adapting machine learning models in creativity research. For example, the European Union's proposed AI Act outlines protocols for using machine learning in high-stakes decision making, including educational contexts. This includes standardized assessment of risks and human oversight or possibility for human intervention (Veale & Borgesius, 2021).

Release of Materials

The best performing models from this study are available as *Ocsai*, a free web-based tool, at the Open Creativity Scoring website[5]. In the experiments performed for this study, duplicated responses were removed, so that the model cannot 'cheat' by seeing the exact same response in evaluation as it saw in training. For future applications, we acknowledge the value of the model having this knowledge and trained an *ALL* split without duplicate removal that is also available for free online. This condition used a greater portion, 95%, of the full dataset. Code and results for experiments are available at GitHub[6]. This includes the dataset of AUT responses, to encourage scholars hoping to work with the same composite of prior AUT datasets, as well as applying our normalizations and splits to the data. Experiments and analysis are prepared as scientific notebooks which can be run in a web browser.

Further Work

The results presented here are an initial investigation into the performance, robustness, and transferability of LLMs for automated evaluation of DT tests. The results establish the practicality of the approach, but also open the door to a good deal of new inquiry. There remain many potential improvements to study, effects that need further investigation, and new considerations opened by this body of approaches. There are also additional tests of DT (see (Runco et al., 2016 for a comparison) beyond the popular AUT such as such as consequences ('what would happen if…', Guilford, Christensen, & Merrifield, 1958) and instances ('list things that are…' Wallach & Kogan, 1965). which may be similarly evaluated with our approach. Particularly intriguing are applications on much longer responses that can benefit from the natural language understanding of LLMs. For example, Johnson et al. (2022) applied an LLM, BERT, to scoring creative narrative writing. Their approach used a pre-trained BERT model to extract embeddings from narrative sentences for distance comparison, akin to the way semantic models are used in OCS (Dumas et al., 2020) and SemDis (Beaty & Johnson 2021).

Toward improved performance, a greater breadth of model architectures and sizes may be compared, as well as collection of more training data. Auditing our datasets for sources of error, toward cleaner data, may also benefit performance. It has also been shown that transformer-

[5] https://openscoring.du.edu
[6] https://github.com/massivetexts/llm_aut_study

based models benefit from domain- and task- adaptive pretraining (Gururangan et al., 2020). That is, prior to teaching them how to complete a given classification task, it is valuable to adjust the general-by-design language model to specialize more on the language of the domain. For example, Organisciak et al. (2023) found that for scoring on elementary-aged children, a model that has been pre-trained on children's and child-facing text (domain-adaptive) performs better than a general model.

This study opens new questions about the limits of semantic models as well as the characteristics of how LLMs work on the AUT, as discussed earlier. Future work may also expand to consider the tractability of LLMs for additional DT scoring issues scoring, such as robustness to cheating, scoring of instances and consequences tasks not just the AUT, working with poor spelling, or identifying sexual or violent responses. Another intriguing question is whether an LLM's underlying measure of confidence in its prediction be used to indicate responses which need human intervention, an important step toward trustworthy applications in high-stakes settings. Further, there has been study to understand the limits of semantic distance methods. For example, Beaty and Johnson (2021) found that it primarily is a measure of novelty, removed from usefulness. LLMs should likewise be inspected for their quirks and limitations: do they improve on those challenges, or can they be modified to do so? Finally, the current study is entirely situated in English, but it would be valuable to evaluate if the methods translate to scoring in other languages too.

Conclusion

In this study, we presented a new approach to automated scoring of the alternate uses task, a test of DT. Our approach, *Ocsai*, applied finetuned large language models and compared them to the current state-of-the-art semantic distance models (Beaty & Johnson, 2021; Dumas et al., 2020).

On overall performance, *Ocsai* greatly improved over existing baselines, where the various supervised learning approaches showed an average performance of $r = .783$ with human judges versus $r = .188$ across the baseline semantic distance systems. It also improved over feature-based supervised learning approaches, such as presented in Buczak et al. 2022. LLM-based semantic distance approaches did not show the same gains as our finetuned models.

Looking at robustness to the amount of training data, our approach improved with more data but already showed gains over the baselines with as little as 1% of the training data. We also showed an application where a five-example prompt on an entirely *untrained* LLM will outperform semantic models, which we believe vibrantly illustrate the promise of supervised learning in this area. Particularly promising is the recent pace of improvements on prompt-based few-shot and zero-shot applications: ChatGPT (Ouyang et al 2022) improved on GPT-3 slightly (at greatly reduced cost), while GPT-4 (OpenAI 2023) showed tremendous gains. Though it underperformed fine-tuned models, GPT-4 offers a strong alternative that even outperformed semantic models with not training examples (i.e. zero-shot). Finally, the transferability of the *Ocsai* approach was evaluated, which still showed significant gains on AUT items that it had never seen before.

We also compiled and deduplicated a large dataset comprised of nine past studies, with each AUT response judged by at least three human judges, which was used to train and evaluate our models. It has been noted that individual human raters are imperfect in their own ways (Beaty & Johnson, 2021; Dumas et al., 2022), but we observed that the outcome of multiple raters can be stable enough to be predictable. As our findings are strongly encouraging for supervised learning methods, large composite datasets will be increasingly important in this line of research.

The primary contribution of this paper is in presenting a new avenue for automated DT scoring. By showing a very strong ability to align with multi-rater human judgements, our results help move automated DT scoring toward applications in creativity research, where automated response scorings are more reliable proxies for multiple trained judges. In addition to avoiding the challenges associated with human raters, it may be applied in places where a judge is not tractable, such as interventions with real-time feedback given to students. The results also re-calibrate debate and inquiry on the limits of automated scoring, introducing a departure from existing methods which merits its own analysis and study within DT research. In all, this current work contributes to the ongoing effort within creativity research to develop methods to scale up original thinking measurement in a valid, replicable, and affordable way.

Acknowledgements

This work was funded by the Institute of Education Sciences (IES), grant #R305A200199.

References

Acar, S. (2023). Does the task structure impact the fluency confound in divergent thinking? An investigation with TTCT-Figural. *Creativity Research Journal, 35*(1), 1-14. https://doi.org/10.1080/10400419.2022.2044656

Acar, S., Berthuiaume, K., Grajzel, K., Dumas, D., Flemister, T., & Organisciak, P. (2023). Applying automated originality scoring to the verbal form of Torrance Tests of Creative Thinking. *Gifted Child Quarterly, 67*(1), 3-17. https://doi.org/10.1177/00169862211061874

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

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., Christensen, A. P., Rosenberg, M. D., Benedek, M., Chen, Q., Fink, A., Qiu, J., Kwapil, T. R., Kane, M. J., & Silvia, P. J. (2018). Robust prediction of individual creative ability from brain functional connectivity. *Proceedings of the National Academy of Sciences*, *115*(5), 1087–1092. https://doi.org/10.1073/pnas.1713532115

Beaty, R. E., & Silvia, P. J. (2012). Why do ideas get more creative across time? An executive interpretation of the serial order effect in divergent thinking tasks. *Psychology of Aesthetics, Creativity, and the Arts*, *6*(4), 309–319. https://doi.org/10.1037/a0029171

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

BigScience. (2021). *BigScience Language Open-science Open-access Multilingual (BLOOM) Language Model*. https://huggingface.co/bigscience/bloom.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022. https://dl.acm.org/doi/10.5555/944919.944937

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. https://doi.org/10/gfw9cs

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901. https://doi.org/10.48550/arXiv.2005.14165

Buczak, P., Huang, H., Forthmann, B., & Doebler, P. (2022). The machines take over: A comparison of various supervised learning approaches for automated scoring of divergent thinking tasks. *The Journal of Creative Behavior*. Advance online publication. https://doi.org/10.1002/jocb.559

Cramond, B., Matthews-Morgan, J., Bandalos, D., & Zuo, L. (2005). A report on the 40-year follow-up of the Torrance Tests of Creative Thinking: Alive and well in the new millennium. *Gifted Child Quarterly*, *49*(4), 283–291. https://doi.org/10.1177/001698620504900402

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/db4ft5

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding*. arXiv. https://arxiv.org/abs/1810.04805v2

Downie, J. S. (2008). The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research. *Acoustical Science and Technology, 29*(4), 247–255. https://doi.org/10.1250/ast.29.247

Dumas, D., & Dunbar, K. N. (2014). Understanding fluency and originality: A latent variable perspective. *Thinking Skills and Creativity*, *14*, 56–67. https://doi.org/10/f6wb79

Dumas, D., Organisciak, P., & Doherty, M. D. (2020). 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/ghcsqq

Dumas, D., Acar, S., Berthiaume, K., Organisciak, P., Eby, D., Grajzel, K., Vlaamster, T., Newman, M., & Carrera, M. (2023). What Makes Children's Responses to Creativity Assessments Difficult to Judge Reliably? *The Journal of Creative Behavior*. https://doi.org/10.1002/jocb.588

Forthmann, B., Oyebade, O., Ojo, A., Günther, F., & Holling, H. (2019). Application of Latent Semantic Analysis to Divergent Thinking is Biased by Elaboration. *The Journal of Creative Behavior*, *53*(4), 559–575. https://doi.org/10.1002/jocb.240

Forthmann, B., & Doebler, P. (2022). *Fifty years later and still working: Rediscovering Paulus et al.'s (1970) automated scoring of divergent thinking tests*. [Pre-print]. https://doi.org/10.31234/osf.io/byj8c

Guilford, J. P. (1950). Creativity. *American Psychologist*, *5*(9), 444–454. https://doi.org/10.1037/h0063487

Guilford, J. P., Christensen, P. R., & Merrifield, P. R. (1958). *Consequences: Manual for administration, scoring, and interpretation*. Sheridan Psychological Services.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). XAI—Explainable artificial intelligence. *Science Robotics*, *4*(37), eaay7120. https://doi.org/10.1126/scirobotics.aay7120

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). *Don't stop pretraining: Adapt language models to domains and tasks.* arXivhttp://arxiv.org/abs/2004.10964

Hass, R. W., Rivera, M., & Silvia, P. J. (2018). On the dependability and feasibility of layperson ratings of divergent thinking. *Frontiers in Psychology*, *9*. https://www.frontiersin.org/articles/10.3389/fpsyg.2018.01343

Hofelich Mohr, A., Sell, A., & Lindsay, T. (2016). Thinking inside the box: Visual design of the response box affects creative divergent thinking in an online survey. *Social Science Computer Review*, *34*(3), 347–359. https://doi.org/10.1177/0894439315588736

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. de L., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., Driessche, G. van den, Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., … Sifre, L. (2022). *Training compute-optimal large language models.* arXiv. https://doi.org/10.48550/arXiv.2203.15556

Hofmann, T. (1999). Probabilistic latent semantic indexing. *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 50–57. https://doi.org/10.1145/312624.312649

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). *Scaling laws for neural language models.* arXiv. https://doi.org/10.48550/arXiv.2001.08361

Kapoor, S., & Narayanan, A. (2022). *Leakage and the reproducibility crisis in ML-based science*. arXiv. https://doi.org/10.48550/arXiv.2207.07048

Kim, K. H. (2006). Can we trust creativity tests? A review of the Torrance Tests of Creative Thinking (TTCT). *Creativity Research Journal*, *18*(1), 3. https://doi.org/10.1207/s15326934crj1801_2

Kim, K. H. (2008). Meta-analyses of the relationship of creative achievement to both IQ and divergent thinking test scores. *The Journal of Creative Behavior*, *42*(2), 106–130. https://doi.org/10.1002/j.2162-6057.2008.tb01290.x

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). *Large language models are zero-shot reasoners.* arXiv. https://doi.org/10.48550/arXiv.2205.11916

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211. https://doi.org/10/dcpw35

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788–791.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2021). *Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing.* arXiv. https://doi.org/10.48550/arXiv.2107.13586

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. http://arxiv.org/abs/1907.11692

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

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 26* (pp. 3111–3119). Curran Associates, Inc. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

Neelakantan, A., Xu, T., Puri, R., Radford, A., Han, J. M., Tworek, J., Yuan, Q., Tezak, N., Kim, J. W., & Hallacy, C. (2022). *Text and code embeddings by contrastive pre-training*. arXiv. https://doi.org/10.48550/arXiv.2201.10005Ni, J., Ábrego, G. H., Constant, N., Ma, J., Hall, K. B., Cer, D., & Yang, Y. (2021). *Sentence-T5: Scalable sentence encoders from pre-trained text-to-text models.* arXiv. https://doi.org/10.48550/arXiv.2108.08877

Ni, J., Ábrego, G. H., Constant, N., Ma, J., Hall, K. B., Cer, D., & Yang, Y. (2021). *Sentence-T5: Scalable Sentence Encoders from Pre-trained Text-to-Text Models* (arXiv:2108.08877). arXiv. https://doi.org/10.48550/arXiv.2108.08877

OpenAI. (2023). *GPT-4 Technical Report* (arXiv:2303.08774). arXiv. https://doi.org/10.48550/arXiv.2303.08774

Organisciak, P., & Dumas, D. (2020). *Open creativity scoring* [Computer software]. University of Denver. https://openscoring.du.edu

Organisciak, P., Newman, M., Eby, D., Acar, S., & Dumas, D. (2023). How do the kids speak? Improving educational use of text mining with child-directed language models. *Information and Learning Sciences*, *124*(1/2), 25–47. https://doi.org/10.1108/ILS-06-2022-0082

Parker, R., Graff, D., Kong, J., Chen, K., & Maeda, K. (2011). *English Gigaword Fifth Edition* [Data set]. Linguistic Data Consortium. https://doi.org/10.35111/WK4F-QT80

Paulus, D. H. (1970). *Computer Simulation of Human Ratings of Creativity. Final Report.* (No. 9-A-032). https://files.eric.ed.gov/fulltext/ED060658.pdf

Paulus, D. H., & Renzuli, J. S. (1968). Scoring creativity tests by computer. *Gifted Child Quarterly*, *12*(2), 79–83. https://doi.org/10.1177%2F001698626801200202

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/gfshwg

Plucker, J. A., Meyer, M. S., & Liu, P. (2022). Divergent thinking: Early views. In M. A. Runco & S. Acar (Eds.), *Handbook of creativity assessment*. Edward Elgar.

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

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*(140), 1–67. https://jmlr.org/papers/v21/20-074.html

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). *SQuAD: 100,000+ questions for machine comprehension of text*. arXiv. https://doi.org/10.48550/arXiv.1606.05250

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence embeddings using siamese BERT-Networks*. arXiv. https://arxiv.org/abs/1908.10084v1

Roemmele, M., Bejan, C. A., & Gordon, A. S. (2011). Choice of plausible alternatives: An evaluation of commonsense causal reasoning. *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, 90–95.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Runco, M. A. (1991). *Divergent thinking*. Ablex Publishing Corporation Norwood, NJ.

Runco, M. A. (2008). Creativity and education. *New Horizons in Education*, *56*(1). https://eric.ed.gov/?id=EJ832901

Runco, M. A., Abdulla, A. M., Paek, S. H., Al-Jasim, F. A., & Alsuwaidi, H. N. (2016). Which test of divergent thinking is best? *Creativity. Theories – Research - Applications*, *3*(1), 4–18. https://doi.org/10.1515/ctra-2016-0001

Runco, M. A., & Acar, S. (2012). Divergent thinking as an indicator of creative potential. *Creativity Research Journal*, *24*(1), 66–75. https://doi.org/10.1080/10400419.2012.652929

Runco, M. A., Millar, G., Acar, S., & Cramond, B. (2010). Torrance Tests of Creative Thinking as predictors of personal and public achievement: A fifty-year follow-up. *Creativity Research Journal*, *22*(4), 361–368. https://doi.org/10.1080/10400419.2010.523393

Runco, M. A., & Mraz, W. (1992). Scoring divergent thinking tests using total ideational output and a creativity index. *Educational and Psychological Measurement*, *52*(1), 213–221. https://doi.org/10.1177/001316449205200126

Said-Metwaly, S., Taylor, C. L., Camarda, A., & Barbot, B. (2022). Divergent thinking and creative achievement—How strong is the link? An updated meta-analysis. *Psychology of Aesthetics, Creativity, and the Arts*. Advance online publication. https://doi.org/10.1037/aca0000507

Shaw, A. (2021). It works…but can we make it easier? A comparison of three subjective scoring indexes in the assessment of divergent thinking. *Thinking Skills and Creativity*, *40*, 100789. https://doi.org/10.1016/j.tsc.2021.100789

Silvia, P. J. (2011). Subjective scoring of divergent thinking: Examining the reliability of unusual uses, instances, and consequences tasks. *Thinking Skills and Creativity*, *6*(1), 24–30. https://doi.org/10.1016/j.tsc.2010.06.001

Silvia, P. J., Nusbaum, E. C., & Beaty, R. E. (2017). Old or new? Evaluating the old/new scoring method for divergent thinking tasks. *The Journal of Creative Behavior*, *51*(3), 216–224. https://doi.org/10.1002/jocb.101

Silvia, P. J., Nusbaum, E. C., Berg, C., Martin, C., & O'Connor, A. (2009). Openness to experience, plasticity, and creativity: Exploring lower-order, high-order, and interactive effects. *Journal of Research in Personality*, *43*(6), 1087–1090. https://doi.org/10.1016/j.jrp.2009.04.015

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

Snyder, H. T., Hammond, J. A., Grohman, M. G., & Katz-Buonincontro, J. (2019). Creativity measurement in undergraduate students from 1984–2013: A systematic review. *Psychology of Aesthetics, Creativity, and the Arts*, *13*(2), 133–143. https://doi.org/10.1037/aca0000228

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1631–1642. https://aclanthology.org/D13-1170

Stevenson, C., Smal, I., Baas, M., Dahrendorf, M., Grasman, R., Tanis, C., Scheurs, E., Sleiffer, D., & van der Maas, H. (2020). *Automated AUT scoring using a big data variant of the consensual assessment technique: Final technical report*.

Torrance, E. P. (1966). *Torrance test of creative thinking: Norms-technical manual research edition-verbal Tests, forms A and B-figural tests, forms A and B*. Princeton: Personnel Press.

Torrance, E. P. (1972). Predictive validity of the Torrance Tests of Creative Thinking. *The Journal of Creative Behavior*, *6*(4), 236–252. https://doi.org/10.1002/j.2162-6057.1972.tb00936.x

Torrance, E. P. (1980). Growing up creatively gifted: A 22-yr longitudinal study. *Creative Child & Adult Quarterly*, *5*(3), 148–158, 170.

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

Veale, M., & Borgesius, F. Z. (2021). Demystifying the Draft EU Artificial Intelligence Act—Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, *22*(4), 97–112. https://doi.org/10.9785/cri-2021-220402

Voorhees, E. M., & Harman, D. K. (Eds.). (2005). *TREC: Experiment and evaluation in information retrieval*. MIT Press. https://mitpress.mit.edu/9780262220736/trec/

Wallach, M. A., & Kogan, N. (1965). *Modes of thinking in young children*. New York.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2019). SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *Advances in Neural Information Processing Systems*, *32*. https://papers.nips.cc/paper/2019/hash/4496bf24afe7fab6f046bf4923da8de6-Abstract.html

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 353–355. https://doi.org/10.18653/v1/W18-5446

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems*, *32*.

Zaccaro, S. J., Connelly, S., Repchick, K. M., Daza, A. I., Young, M. C., Kilcullen, R. N., Gilrane, V. L., Robbins, J. M., & Bartholomew, L. N. (2015). The influence of higher order cognitive capacities on leader organizational continuance and retention: The mediating role of developmental experiences. *The Leadership Quarterly*, *26*(3), 342–358. https://doi.org/10.1016/j.leaqua.2015.03.007