A New Dataset and Method for Creativity Assessment Using the Alternate Uses Task

Luning Sun[1][0000−0002−2470−4278], Hongyi Gu[2][0009−0006−4885−1412], Rebecca Myers[1], and Zheng Yuan[3,2][0000−0003−2406−1708]

[1] University of Cambridge
{ls523,rm804}@cam.ac.uk
[2] NetMind.AI
hongyi.gu@netmind.ai
[3] King's College London
zheng.yuan@kcl.ac.uk

Abstract. Creativity ratings by humans for the alternate uses task (AUT) tend to be subjective and inefficient. To automate the scoring process of the AUT, previous literature suggested using semantic distance from non-contextual models. In this paper, we extend this line of research by including contextual semantic models and more importantly, exploring the feasibility of predicting creativity ratings with supervised discriminative machine learning models. Based on a newly collected dataset, our results show that supervised models can successfully classify between creative and non-creative responses even with unbalanced data, and can generalise well to out-of-domain unseen prompts.

Keywords: Creativity · Alternate uses task · Automated scoring

1 Introduction

Creativity, defined as the production of novel and useful products [24], is one of the most important skills for student and young people development [3], and a valuable employee outcome associated with organisational sustainability and innovation [13]. A core element of creativity is divergent thinking in problem solving [15,20]. One of the most widely used divergent thinking tests is the alternate uses task (AUT) [14,31], which asks respondents to list as many uses for common items (e.g. newspaper) as possible, and usually within a time limit. The responses are then rated on dimensions such as *fluency*, *originality*, *flexibility*, and *elaboration* [1]. Similar to many other creativity tests, it requires human raters to score the responses manually, rendering the results subjective, unreliable, and undermining their validity [17]. Consequently, education and training in creativity are severely constrained by the lack of an objective and efficient measurement of creativity [30].

To automate the scoring process of the AUT, researchers have capitalised on recent developments in natural language processing (NLP) and proposed that semantic distance could be calculated to predict human creativity ratings. For

instance, [12] found that GloVe [26], among a number of publicly available word embeddings models, produced the most reliable and valid *originality* scores on the AUT. [5] constructed a latent semantic distance factor based on five non-contextual semantic spaces, and found strong correlations between the semantic distances and the respondent-level (i.e. person-level) human ratings of creativity in the AUT responses.

Unlike previous work, we propose to address the AUT scoring as a supervised discriminative machine learning problem and particularly as a binary classification problem: classifying between creative and non-creative responses. In addition to examining the relationship between semantic distance variables and human ratings of creativity in the AUT responses, we explore supervised machine learning models for the prediction of creativity ratings. Our results show that the proposed method generalises well to unseen tasks and prompts. We also compare the performance of our proposed models to that of OpenAI's ChatGPT,[4] and discuss its potential application in creativity assessment.

This paper makes the following contributions. First, we introduce a new dataset of AUT responses, the Cambridge AUT Dataset,[5] and make it publicly available to facilitate future research on creativity assessment. Second, to our knowledge, we present the first comparison between the application of contextual and non-contextual semantic spaces in the context of creativity assessment. Finally, as far as we know, this is the first attempt to apply a supervised learning model to the scoring of AUT responses, which demonstrates performance improvement across a set of different prompts.

2 The Cambridge AUT Dataset

2.1 Data collection

The AUT data used in this study was collected as part of a larger project on creativity assessment [25] that received ethics approval from both the Faculty of Education, University of Cambridge and Cambridge Judge Business School. Two common objects were implemented as prompts for the AUT, namely *bowl* and *paperclip*. For each prompt, participants were given 90 seconds to come up with as many different uses as possible (see Section A).

A total of 1,297 participants (Gender: 693 female, 567 male, 14 other, 23 missing; Age: mean 26.26 years, SD 9.68 years, 13 missing; Ethnicity: 883 White, 54 Asian, 54 Black, 110 mixed, 124 other, 72 missing), who were recruited through Cambridge University mailing lists, social media, and a testing website,[6] took part in the task online between April 2020 and January 2021.[7] 1,027 of them provided non-empty answers for *bowl* (each with an average of 7.40 uses; SD:

[4] `https://chat.openai.com/`

[5] `https://github.com/ghydsgaaa/Cambridge-AUT-dataset`

[6] `https://discovermyprofile.com/`

[7] Participants were not paid but given the opportunity to opt into a draw to win one of ten £10 Amazon vouchers.

Table 1: Response examples of different average ratings for each prompt.

Average rating	Prompt: bowl	Prompt: paperclip
1.0	fish holder	drawing
2.0	doing an inhalation	make a logo
3.0	space ship	pasta mold
4.0	sending mail through river	holding nose while swimming

3.49) and 1,020 for *paperclip* (each with an average of 6.23 uses; SD: 2.94). For each object, all uses (referred to as *responses* below) were pooled together and only the English ones were subject to annotation.

2.2 Annotation

We applied the subjective scoring method based on the Consensual Assessment Technique [2,10]. A group of psychology students were trained on how to evaluate the responses on their *originality*, using a Likert scale from 0 to 4, where 0 indicates a not valid or not relevant use, 1 a common use without any originality, 2 an uncommon use with limited originality, and 3 and 4 original uses with moderate and extreme creativity, respectively.

Three raters were initially recruited to annotate the AUT responses. Each of them was tasked with a random sample of the responses. The assignment of the responses among the raters ensured that each unique response would be rated by at least two raters. Due to time constraints, one of the raters had to quit midway and the remaining annotation was completed by a fourth rater (their ratings were combined in the dataset).

After removing duplicate responses, a total of 3,380 responses for *bowl* and 3,650 for *paperclip* were annotated. Both objects received the same average rating (1.27, SDs: 0.49 for *bowl* and 0.45 for *paperclip*). 95 responses for *bowl* and 86 for *paperclip* received average ratings of below 1, which means that at least one of the raters rated the responses as invalid uses, hence being removed from the subsequent analyses. Response examples of different average ratings for each prompt are presented in Table 1.

Notably, the dataset is severely unbalanced, with more than half responses rated 1 and only a few responses rated 3 and above - see Table 2. This is expected, as creative responses are less frequent by nature. Nonetheless, less frequent responses may not necessarily be creative. The creativity ratings in this work focus on the absolute originality in the responses rather than their relevant frequency. It is also worth noting that the inter-rater agreement is not particularly high (correlations range from 0.39 to 0.58 - see Table 3) compared to other assessment tasks such as essay scoring [4]. This is likely due to the nature of human ratings in creativity assessment, which are based on their own subjective perception of creativity [10,23].

Table 2: Number of responses per average rating in the Cambridge AUT dataset. Responses with an average rating below 1 (i.e. a not valid or not relevant use) are excluded from the analyses.

Average rating	#responses combined	#responses (bowl)	#responses (paperclip)
< 1.0	181	95	86
1.0	4,167	2,096	2,071
1.5	1,717	638	1,079
2.0	666	392	274
2.5	188	104	84
3.0	92	48	44
3.5	15	6	9
4.0	4	1	3
Total	7,030	3,380	3,650

3 Semantic Models

Following previous work [12,5], we analyse the AUT responses collected in our dataset and test whether combining multiple models of semantic distance into a single latent variable can approximate human creativity ratings.

3.1 Semantic distance

Pre-trained semantic models are used to compute the semantic distance (i.e. cosine distance) between the prompt and the response. We employ four contextual models: Universal Sentence Encoder [9],[8] Sentence-Transformers [27],[9] DistilRoBERTa [28],[10] and GPT-3 [8];[11] and three non-contextual models: GloVe [26],[12] Word2vec [21],[13] and fastText [7].[14]

For non-contextual models, we first extract embeddings for each word in the response, and then take the multiplicative composition as suggested by [22,5]. For contextual models, we extract the sentence embeddings directly.

3.2 Confirmatory factor analysis

Table 3 presents zero-order correlations among human ratings and semantic distance variables. Confirmatory factor analysis (CFA) is performed to investigate

[8] `https://tfhub.dev/google/universal-sentence-encoder/4`
[9] `https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2`
[10] `https://huggingface.co/distilroberta-base`
[11] `https://beta.openai.com/docs/models/gpt-3`
[12] glove-wiki-gigaword-300
[13] word2vec-google-news-300
[14] fasttext-wiki-news-subwords-300

Table 3: Correlations among human ratings and semantic distance variables: polychoric correlations between human raters, polyserial correlations between human raters and semantic distances, and pearson correlations between semantic distances. r1-3: rater1-3; USE: Universal Sentence Encoder; ST: Sentence-Transformers.

	r1	**r2**	**r3**	**USE**	**ST**	**RoBERTa**	**GPT-3**	**GloVe**	**Word2vec**	**fastText**
r1	1.00	-	-	-	-	-	-	-	-	-
r2	0.58	1.00	-	-	-	-	-	-	-	-
r3	0.39	0.44	1.00	-	-	-	-	-	-	-
USE	0.14	0.16	0.16	1.00	-	-	-	-	-	-
ST	0.19	0.17	0.31	0.62	1.00	-	-	-	-	-
RoBERTa	0.12	0.16	0.22	0.57	0.76	1.00	-	-	-	-
GPT-3	0.08	0.20	0.11	0.51	0.50	0.55	1.00	-	-	-
GloVe	0.05	0.02	0.06	0.12	0.05	-0.08	-0.14	1.00	-	-
Word2vec	0.03	0.04	0.05	0.22	0.17	0.09	0.21	0.40	1.00	-
fastText	-0.03	-0.01	0.03	0.15	0.11	0.07	0.14	0.35	0.24	1.00

the latent correlation between human ratings and a semantic distance factor underlying different semantic models.[15]

We specify two models to examine the relationship between the response-level human ratings and the semantic distance factors built upon contextual ($\mathbf{Model_{contextual}}$) and non-contextual semantic models ($\mathbf{Model_{non\text{-}contextual}}$), respectively. Both contextual and non-contextual models yield good model fit to the data.[16] The contextual model reveals a higher correlation between the latent semantic distance factor and the human ratings than the non-contextual model ($r = 0.065$, $p < .001$ for the non-contextual model - see Figure 1, Section B; and $r = 0.293$, $p < .001$ for the contextual model - see Figure 2, Section B).

Nevertheless, these latent correlations between the response-level human ratings and the semantic distance factors are still considerably low, in comparison to those correlations reported in previous studies based on the respondent-level data [12,5], suggesting that these semantic distance variables cannot be used reliably as an unsupervised model to predict human creativity ratings.

4 Binary Classification Models

Since the semantic distance variables reported above fail to adequately predict the human creativity ratings, in this section we turn to supervised machine learning methods. In light of the availability of a labeled dataset, we conduct experiments, where we fine-tune pre-trained language models to improve their prediction accuracy. Since the dataset is severely unbalanced (see Table 2), we

[15] CFA is a statistical technique used to verify the factor structure of a set of observed variables and test if the relationship between observed variables and their underlying latent constructs exist.

[16] Detailed CFA results are presented in Table 6, Section B.

Table 4: Micro-average F1 scores on the AUT test sets. The highest scores for each prompt are in bold.

Tested on	$\textbf{Model}_{\textbf{bowl}}$	$\textbf{Model}_{\textbf{paperclip}}$	$\textbf{Model}_{\textbf{bowl+paperclip}}$	ChatGPT	Baseline
Bowl	**0.79**	0.73	0.76	0.65	0.70
Paperclip	0.61	0.65	**0.67**	0.56	0.60
Combined	0.69	0.68	**0.72**	0.60	0.65

cast the task as a binary classification between creative (average rating > 1, i.e. at least one of the raters assigned 2 or above) and non-creative (average rating $= 1$) responses. Take prompt *bowl* as example, "mixing stuff" is considered a non-creative response with average rating 1 and "knee caps" is considered a creative response with average rating 3. We further split the dataset into a training set (90%) and a test set (10%).

4.1 Fine-tuned models

Fine-tuning pre-trained language models via supervised learning is key to achieving state-of-the-art performance in many NLP tasks. Adopting this approach, we experiment with three transformer-based pre-trained language models: BERT [11], RoBERTa [19], and GPT-3 [8].

To fine-tune BERT and RoBERTa, we use them as the underlying language model and add a linear layer on the top, which allows for binary classification. We construct the input by concatenating the prompt w and the response $R = r_1, r_2, ..., r_n$:

$$[CLS]; w; [SEP]; r_1, r_2, ..., r_n; [SEP] \tag{1}$$

where the $[CLS]$ representation is then fed into the output layer for classification. During training, the model is optimised in an end-to-end manner. We fine-tune *bert-base-uncased*[17] and *roberta-base*[18] on the AUT data, with a batch size of 32 and a learning rate of $3 \times e^{-05}$ for 5 epochs.

For GPT-3, we fine-tune the GPT-3 babbage model using the OpenAI's API.[19]

In our experiments, 5-fold cross validation is performed and detailed results are presented in Table 7, Table 8 and Table 9, Section C. The fine-tuned BERT models are chosen for later experiments due to their superior micro-average F1 scores.

4.2 Results

Prediction results of our fine-tuned BERT models on the test sets for each prompt as well as both prompts combined are reported in Table 4. Three binary classification models trained on different data are compared: $\textbf{Model}_{\textbf{bowl}}$ is trained

[17] `https://huggingface.co/bert-base-uncased`
[18] `https://huggingface.co/roberta-base`
[19] `https://openai.com/blog/openai-api`

Table 5: Micro-average F1 scores on the dataset from [6]. The highest scores for each prompt are in bold.

Tested on	$\textbf{Model}_{\textbf{bowl}}$	$\textbf{Model}_{\textbf{paperclip}}$	$\textbf{Model}_{\textbf{bowl+paperclip}}$	ChatGPT	Baseline
Box	0.64	**0.72**	0.69	0.54	0.58
Rope	**0.62**	0.61	**0.62**	0.51	0.51

on responses for *bowl* only; $\textbf{Model}_{\textbf{paperclip}}$ is trained on responses for *paperclip* only; and $\textbf{Model}_{\textbf{bowl+paperclip}}$ is trained on the data for both prompts.

Using the majority class as **Baseline**, we observe an increase in the F1 scores on the prompt-specific level (i.e. in-domain) and the same for the cross-prompt predictions (i.e. out-of-domain). The best model for prompt *bowl* is the prompt-specific model $\textbf{Model}_{\textbf{bowl}}$, achieving a micro-average F1 score of 0.79. Notably, $\textbf{Model}_{\textbf{bowl+paperclip}}$ yields the best performance when tested on prompt *paperclip*, outperforming its prompt-specific model $\textbf{Model}_{\textbf{paperclip}}$ (0.67 vs. 0.65). These results suggest that given more data (even from out-of-domain prompts), the model is able to improve the overall performance on different prompts, hence showing a potential to serve as prompt-independent filters for creative responses in the AUT.

4.3 A case study with new AUT prompts

In order to explore the generalisability of our models, we apply our classification models to the AUT responses collected in a previous study with different prompts than those here, namely *rope* and *box* [6]. Since a different annotation scheme was used - a scale from 1 (not at all creative) to 5 (very creative), we split their data into two classes: non-creative (responses with an average human rating of 1), and creative (those with an average human rating of 2 or above).

In Table 5 we report the prediction results of our models on the responses to prompts *box* and *rope*.[20] In general, all our models outperform the majority class baseline, indicating a prompt independence and a cross-dataset applicability. The result suggests that using training data from only a few prompts (even just one or two), it is possible to develop supervised machine learning models that can work as a generic, automated scoring tool for the AUT with any unseen prompt.

4.4 Comparison with ChatGPT predictions

Inspired by recent progress on using generative, pre-trained large language models as evaluators in tasks like machine translation [18], code generation [32] and grammatical error correction [29], we explore how these models can be applied in creativity assessment. We apply ChatGPT (gpt-3.5-turbo at temperature 0) to the same task on both our dataset and that from [6],[21] and report results

[20] Per-class precision, recall and F1 scores are reported in Table 10, Section D.
[21] The prompt we used for experiments with ChatGPT is provided in Section E.

in Table 4 and Table 5. We can see that **ChatGPT** underperforms the majority **Baseline** on both datasets, revealing its limitation in evaluating abstract concepts like creativity.

Detailed per-class analysis reveals that **ChatGPT** achieves high precision, yet considerably low recall for non-creative responses on both datasets, while an opposite pattern is observed for creative responses.[22] As its performance seems complementary to that of our fine-tuned models, we see a potential of integrating both methods, which may result in further performance gains in creativity assessment.[23]

5 Conclusions

In this paper, we performed confirmatory factor analysis to investigate the latent correlations between the semantic distance factors and the human ratings of creativity in a newly collected AUT dataset, the Cambridge AUT Dataset. On the response level, we observed significant but lower correlations than those on the respondent level as reported in previous studies. It was also noted that contextual semantic models appear to show greater resemblance to the human ratings than non-contextual models. One step further, we experimented with several fine-tuned models, which showed encouraging performance improvement in classifying between creative and non-creative responses under both in-domain and out-of-domain settings. When applied to an external dataset with new prompts, the models trained on our dataset exhibited reasonably well predictions, showing promising generalisability.

With the above findings, we see a possibility of developing an automated scoring tool for the AUT using supervised machine learning models. To extend this line of research, we plan to examine different model architecture and gather more data with different prompts, in order to better understand the generalisability of the supervised models in the general creativity assessment.

6 Limitations

We notice relatively low agreement among the annotators. One possible explanation is that the annotators come from different countries (e.g. the UK, India, and China) with different native languages and cultural backgrounds. Past literature [16] found cross-cultural differences in both the idea generation and the idea evaluation phases of the divergent thinking task. It is likely that the annotators do not share entirely the same conceptual framework for creative ideas

[22] Per-class precision, recall and F1 scores are reported in Table 11 and Table 12, Section F.

[23] One viable solution is employing a voting ensemble technique, which involves assigning weights to results of both models and striking a balance between precision and recall. Alternatively, we could prompt ChatGPT to generate quantified results and establish a threshold for comparing its outputs with those of the fine-tuned models.

around the prompts, resulting in inconsistent ratings. Future work is warranted to confirm this.

Due to data imbalance and sparsity, this paper addresses the AUT scoring as a binary classification between creative and non-creative responses. The proposed approach may therefore fail to evaluate creativity at detailed levels of granularity. It would be ideal to collect more responses with higher ratings so as to develop an automated creativity assessment system with greater precision. Moreover, to address the concern of overfitting in our experiments, we used 5-fold cross validation and applied our models to unseen data, which showed comparable results.

The results with regard to ChatGPT is based on preliminary experiments. A more thorough investigation using different parameters, prompts, and models is warranted. We are excited to see how large language models like ChatGPT may help with creativity assessment in the future.

Acknowledgements We would like to thank all participants who took part in the AUT and all raters who annotated the responses. LS acknowledges financial support from Invesco through their philanthropic donation to Cambridge Judge Business School.

References

1. Amabile, T.M.: Social psychology of creativity: A consensual assessment technique. Journal of Personality and Social Psychology **43**(5), 997–1013 (1982)
2. Amabile, T.M.: The social psychology of creativity: A componential conceptualization. Journal of Personality and Social Psychology **45**(2), 357–376 (1983)
3. Ananiadou, K., Claro, M.: 21st century skills and competences for new millennium learners in oecd countries. (OECD Education Working Papers (41) (2009), `https://www.oecd-ilibrary.org/content/paper/218525261154`
4. Andersen, Ø.E., Yuan, Z., Watson, R., Cheung, K.Y.F.: Benefits of alternative evaluation methods for automated essay scoring. In: Proceedings of the 14th International Conference on Educational Data Mining (EDM 2021). Paris, France (2021)
5. Beaty, R.E., Johnson, D.R.: Automating creativity assessment with SemDis: An open platform for computing semantic distance. Behavior Research Methods **53**(2), 757–780 (2021)
6. 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., et al.: Robust prediction of individual creative ability from brain functional connectivity. Proceedings of the National Academy of Sciences **115**(5), 1087–1092 (2018)
7. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the association for computational linguistics **5**, 135–146 (2017)
8. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)

9. Cer, D., Yang, Y., Kong, S.y., Hua, N., Limtiaco, N., John, R.S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., et al.: Universal sentence encoder. arXiv preprint arXiv:1803.11175 (2018)
10. Cseh, G.M., Jeffries, K.K.: A scattered CAT: A critical evaluation of the consensual assessment technique for creativity research. Psychology of Aesthetics, Creativity, and the Arts **13**(2), 159–166 (2019)
11. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
12. Dumas, D., Organisciak, P., Doherty, M.: 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 (2021)
13. George, J.M., Zhou, J.: Dual tuning in a supportive context: Joint contributions of positive mood, negative mood, and supervisory behaviors to employee creativity. Academy of Management Journal **50**(3), 605–622 (2007), https://doi.org/10.5465/AMJ.2007.25525934
14. Guilford, J.P.: The nature of human intelligence. McGraw-Hill, New York, NY (1967)
15. Guilford, J.P.: Creative talents: Their nature, uses and development. Bearly Limited, Buffalo, NY (1986)
16. Ivancovsky, T., Shamay-Tsoory, S., Lee, J., Morio, H., Kurman, J.: A dual process model of generation and evaluation: A theoretical framework to examine cross-cultural differences in the creative process. Personality and Individual Differences **139**, 60–68 (2019). https://doi.org/https://doi.org/10.1016/j.paid.2018.11.012, https://www.sciencedirect.com/science/article/pii/S0191886918306081
17. Kim, K.H.: Can We Trust Creativity Tests? A Review of the Torrance Tests of Creative Thinking (TTCT). Creativity Research Journal **18**(1), 3–14 (2006), https://doi.org/10.1207/s15326934crj1801_2
18. Kocmi, T., Federmann, C.: Large language models are state-of-the-art evaluators of translation quality. arXiv preprint arXiv:2302.14520 (2023)
19. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
20. McCrae, R.R.: Creativity, divergent thinking, and openness to experience. Journal of Personality and Social Psychology **52**(6), 1258–1265 (1987)
21. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems **26** (2013)
22. Mitchell, J., Lapata, M.: Composition in distributional models of semantics. Cognitive science **34**(8), 1388–1429 (2010)
23. Mouchiroud, C., Lubart, T.: Children's original thinking: An empirical examination of alternative measures derived from divergent thinking tasks. Journal of Genetic Psychology **162**(4), 382–401 (2001)
24. Mumford, M.D.: Where have we been, where are we going? taking stock in creativity research. Creativity Research Journal **15**(2-3), 107–120 (2003). https://doi.org/10.1080/10400419.2003.9651403, https://doi.org/10.1080/10400419.2003.9651403
25. Myers, R.J.: Measuring creative potential in higher education: The development and validation of a new psychometric test (2020), Unpublished Master's Dissertation, University of Cambridge

26. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
27. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084 (2019)
28. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019)
29. Sottana, A., Liang, B., Zou, K., Yuan, Z.: Evaluation Metrics in the Era of GPT-4: Reliably Evaluating Large Language Models on Sequence to Sequence Tasks. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2023)
30. Susnea, I., Pecheanu, E., Costache, S.: Challenges of an e-learning platform for teaching creativity. In: Proceedings of the 11th International Scientific Conference eLearning and Software for Education. Bucharest, Romania (Apr 2015)
31. Torrance, E.P.: Torrance tests of creative thinking - norms technical manual research edition - verbal tests, forms A and B - figural tests, forms A and B. Personnel Press, Princeton, NJ (1966)
32. Zhuo, T.Y.: Large language models are state-of-the-art evaluators of code generation (2023)

A The instructions used for the AUT

General instruction: For the next four questions, there will be a time limit. For each task, please read the instructions and enter each possible answer separately by pressing the enter key after each one. If you run out of answers you may move on by pressing the next button, otherwise your question will automatically change after the allocated time.

Each task requires you to come up with as many different answers as possible. Try to be creative as there is no right or wrong answer.

Prompt 1: List as many different uses of a bowl as you can think of.

Prompt 2: Think of many different uses of a paperclip.

B Detailed CFA results

Table 6: Latent correlations between human creativity ratings and semantic distance factors ($\mathbf{Model_{non\text{-}contextual}}$ and $\mathbf{Model_{contextual}}$) on the Cambridge AUT dataset.

Tested on	$\mathbf{Model_{non\text{-}contextual}}$	$\mathbf{Model_{contextual}}$
Bowl	0.127	0.278
Paperclip	-	0.296
Combined	0.065	0.293

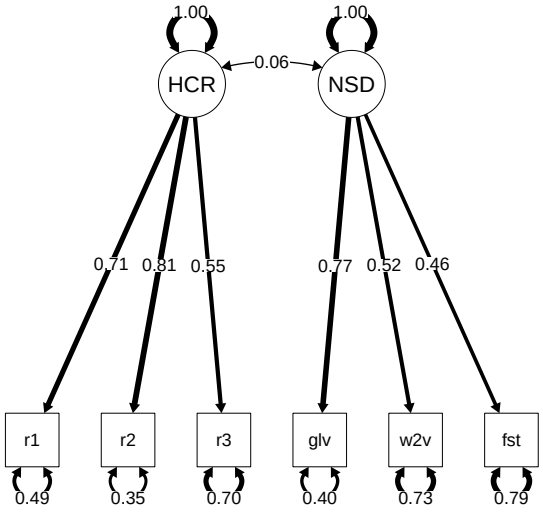

Fig. 1: CFA diagram of $\mathbf{Model_{non\text{-}contextual}}$ on the Cambridge AUT dataset. r1-3: rater1-3; glv: GloVe; w2v: Word2vec; fst: fastText; HCR: human creativity rating factor, NSD: non-contextual semantic distance factor.

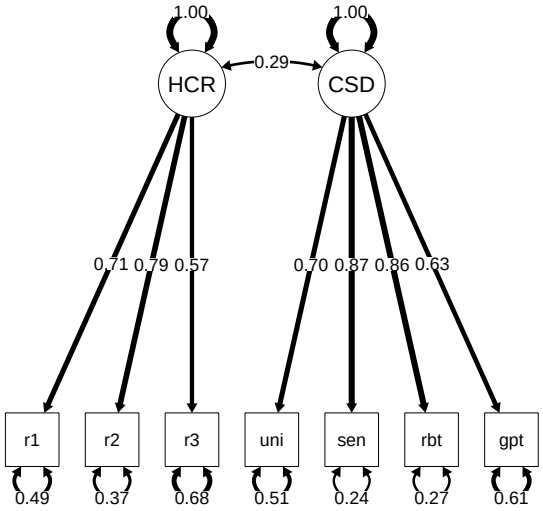

Fig. 2: CFA diagram of $\mathbf{Model_{contextual}}$ on the Cambridge AUT datseta. r1-3: rater1-3; uni: Universal Sentence Encoder; sen: Sentence-Transformers; rbt: RoBERTa; gpt: GPT-3; HCR: human creativity rating factor, CSD: contextual semantic distance factor.

C Cross validation results

Table 7: Fine-tuned BERT cross validation results on the Cambridge AUT training sets. P: precision; R: recall.

Model	Non-creative			Creative			Micro-average
	P	R	F1	P	R	F1	F1
BERT$_{\text{bowl}}$	0.86	0.88	0.87	0.72	0.68	0.70	0.82
BERT$_{\text{paperclip}}$	0.85	0.69	0.76	0.44	0.66	0.53	0.69
BERT$_{\text{bowl+paperclip}}$	0.91	0.82	0.86	0.53	0.72	0.61	0.80

Table 8: Fine-tuned RoBERTa cross validation results on the Cambridge AUT training sets. P: precision; R: recall.

Model	Non-creative			Creative			Micro-average
	P	R	F1	P	R	F1	F1
RoBERTa$_{\text{bowl}}$	0.83	0.86	0.85	0.66	0.60	0.63	0.79
RoBERTa$_{\text{paperclip}}$	0.70	0.81	0.76	0.63	0.49	0.55	0.68
RoBERTa$_{\text{bowl+paperclip}}$	0.80	0.74	0.77	0.58	0.67	0.62	0.71

Table 9: Fine-tuned GPT-3 babbage cross validation results on the Cambridge AUT training sets. P: precision; R: recall.

Model	Non-creative			Creative			Micro-average
	P	R	F1	P	R	F1	F1
GPT-3$_{\text{bowl}}$	0.87	0.87	0.87	0.69	0.71	0.70	0.82
GPT-3$_{\text{paperclip}}$	0.74	0.76	0.75	0.62	0.61	0.62	0.70
GPT-3$_{\text{bowl+paperclip}}$	0.80	0.79	0.80	0.64	0.63	0.63	0.71

D Model performance on the dataset from [6]

Table 10: Prediction performance on the dataset from [6]. P: precision; R: recall.

Tested on	Model	Non-creative			Creative			Micro-average
		P	R	F1	P	R	F1	F1
Box	$\textbf{Model}_{\textbf{bowl}}$	0.57	0.65	0.61	0.70	0.63	0.66	0.64
	$\textbf{Model}_{\textbf{paperclip}}$	0.60	0.78	0.68	0.84	0.69	0.75	0.72
	$\textbf{Model}_{\textbf{bowl+paperclip}}$	0.75	0.66	0.70	0.64	0.73	0.68	0.69
Rope	$\textbf{Model}_{\textbf{bowl}}$	0.67	0.62	0.65	0.57	0.62	0.60	0.62
	$\textbf{Model}_{\textbf{paperclip}}$	0.41	0.71	0.52	0.83	0.57	0.67	0.61
	$\textbf{Model}_{\textbf{bowl+paperclip}}$	0.60	0.64	0.62	0.64	0.60	0.62	0.62

E ChatGPT prompt

You are a judge in the alternate uses task, where respondents are asked to list different uses for a common object. You will be presented with the object and a response that illustrates one of its uses. Please judge if the response is creative or non-creative. Inappropriate, invalid, irrelevant responses, and responses with common uses are considered non-creative, whereas appropriate, valid, novel and unusual uses are considered creative.
The object is: {prompt}
The response is: {response}
Please give your answer in "creative" or "non-creative".
Your answer:

F ChatGPT classification results

Table 11: ChatGPT results on the Cambridge AUT dataset. P: precision; R: recall.

Tested on	Non-creative			Creative			Micro-average
	P	R	F1	P	R	F1	F1
Bowl	0.85	0.60	0.70	0.45	0.75	0.56	0.65
paperclip	0.82	0.35	0.48	0.48	0.88	0.62	0.56
Combined	0.84	0.48	0.61	0.46	0.83	0.60	0.60

Table 12: ChatGPT results on the dataset from [6]. P: precision; R: recall.

Tested on	Non-creative			Creative			Micro-average
	P	R	F1	P	R	F1	F1
Box	0.77	0.28	0.42	0.48	0.88	0.62	0.54
Rope	0.69	0.35	0.47	0.42	0.76	0.54	0.51