Are Language Models Unsupervised Multi-domain CC Systems?

Robert Morain, Branden Kinghorn and **Dan Ventura**
Computer Science Department
Brigham Young University
Provo, UT 84602 USA
rmorain2@byu.edu, brando3034king@hotmail.com, ventura@cs.byu.edu

Abstract

Recently, ChatGPT has grown in popularity due to its ability to generate high quality text in a wide variety of contexts. In order to determine whether ChatGPT threatens to undermine the need for traditional CC systems, ChatGPT's ability to generate textual creative artifacts needs to be formally analysed. To do this, we constructed a survey that compares artifacts generated by traditional CC systems with corresponding artifacts generated by ChatGPT. Both types of artifacts are also evaluated independently on how well they possess certain desirable characteristics. Overall, the survey shows that artifacts generated by ChatGPT are preferred 36.84% ($p = 0.014$) more often and rated higher by 0.5 mean Likert scale points ($p = 0.0004$). These results indicate a need to reconsider the purpose and approach of traditional CC systems going forward.

Introduction

Computational creativity (CC) researchers often create applications that address creativity in specific domains such as stories (Pérez and Sharples 2001), poetry (Boggia et al. 2022), or puns (Ritchie 2003). These CC systems often introduce novel methods for generating creative artifacts such as templates, rules, or machine learning models. The authors then evaluate these generated artifacts either automatically or by way of a user survey. Recently, ChatGPT (OpenAI 2023) has demonstrated impressive text generation abilities. In this paper, we aim to evaluate ChatGPT's ability to generate creative artifacts by comparing ChatGPT's artifacts to artifacts generated by domain specific CC systems. While the scope of these experiments could include other modalities such as images (Ramesh et al. 2021), this paper focuses on textual creative artifacts.

This paper uses a definition of creativity that focuses on the generated artifact rather than on the process by which is created (Wiggins 2006).

Motivation

As statistical large language models improve, the need for domain-specific CC systems requires further consideration. If traditional CC systems are to remain relevant, they must offer distinct advantages over models like ChatGPT and its successors. ChatGPT implicitly learns many language related tasks through the general tasks of autoregressive language modeling (Radford et al. 2018) and fine-tuning with human feedback (Ouyang et al. 2022). The extent to which these abilities overlap with traditional CC systems is the central focus of this paper. Another topic related to this work is the philosophical question surrounding the creative limitations of language models.

All of the data used for this paper including prompts, survey results, and analysis is available online.[1]

Methods

The survey used in this paper includes artifacts from a selection of CC systems covering four textual domains: puns, six word stories, poetry, and short stories. These systems are JAPE (Ritchie 2003), PAUL BOT (Tyler, Wilsdon, and Bodily 2020), Movie Title Puns (Hämäläinen and Alnajjar 2019), MICROS (Spendlove, Zabriskie, and Ventura 2018), HIEROS (Spendlove and Ventura 2020), a GPT-2 model fine-tuned on six-word stories (Chien 2020), mBART for poetry generation (Boggia et al. 2022), TwitSong 3.0 (Lamb and Brown 2019), PoeTryMe (Oliveira and Alves 2016), MEXICA (Pérez and Sharples 2001), INES (Concepción, Gervás, and Méndez 2019), and Chronicle (Pickering and Jordanous 2017).

Notice that this survey does not discriminate against CC systems that use language models. The focus of the survey is not to determine whether language models are useful in CC systems, but rather to evaluate ChatGPT's creative text generation abilities compared to existing CC systems regardless of the methods used in the system.

While many potential systems could have been selected for the survey, these systems were chosen based on a combination of factors: the availability of artifacts present in the paper; the popularity of the system in the CC community; suitability to the domain; and recency of the paper's publication. Interested readers should refer to these papers for more details about their methods and implementation.

A set of twelve artifacts was compiled by selecting one artifact from each paper, to reduce the possibility of a single system under performing and thus skewing the results. Because the highest scoring artifact is not reported in the pa-

[1] `https://tinyurl.com/s4k5pwry`

CC system artifact:	How is a nice girl like a sugary bird? They are both sweet chicks.
Prompt:	Think like a punster. How is a nice girl like a sugary bird?
ChatGPT:	Here's a possible punning answer to your riddle: How is a nice girl like a sugary bird? They're both tweethearts!

Table 1: To create an artifact using ChatGPT with the same subject as a CC system artifact, a prompt with the appropriate domain and subject matter is provided to the model. The generated artifact is manually extracted. ChatGPT does not receive the original artifact in the prompt.

pers, the artifact expected to perform best according to the authors' subjective judgement was chosen.

Given the set of selected artifacts from CC systems, corresponding artifacts with similar subject matter were generated using ChatGPT. [2] ChatGPT was prompted to create an artifact from a particular domain (pun, six-word story, etc.) that included the same subject matter as the original artifact. Table 1 provides an example for how these artifacts were generated. This process facilitates the comparison of artifacts based on quality rather than other factors such as preference of subject. In some cases, when the generated artifact was too long or did not possess the correct subject matter, ChatGPT was iteratively prompted to generate a suitable artifact. Otherwise, the first artifact generated was selected. Artifacts were also screened for plagiarism by searching the web for exact copies.

Next, a survey was created to evaluate the artifacts based both on reviewers' preferences and characteristics used by various authors to evaluate the corresponding CC systems. To evaluate preferences, reviewers are asked to choose between a CC system artifact and the corresponding ChatGPT artifact, in a side by side comparison. Reviewers also had the option to mark "no preference". The reviewers were not made aware of which artifact came from a specialized CC system and which came from ChatGPT. To evaluate artifacts based on their characteristics, reviewers rated each artifact based on how well they possessed each characteristic on a Likert scale (1: strongly disagree, 2: somewhat disagree, 3: neither agree nor disagree, 4: somewhat agree, 5: strongly agree). For puns, the evaluation characteristics are "funny," and "surprising"; for six-word stories, "coherent" and "impactful"; for poems, "meaningful" and "emotional"; and for short stories, "entertaining" and "surprising". These characteristics were selected from the evaluation criteria used by the original authors to evaluate the CC systems. In addition, artifacts from all four domains are also rated on how creative they are perceived to be.

The survey was distributed online through Facebook, Instagram, Twitter, and Reddit. On Reddit, the survey was sent to the r/ArtificialIntelligence, r/MachineLearning, r/deeplearning, and r/ChatGPT subreddits. The survey does not ask for respondents to identify themselves or to rate their own knowledge of AI or CC; therefore it is unknown whether the reviewers are experts or not. The survey is randomized such that the questions and answers appear in random order.

[2] At the time of this experiment in April 2023, ChatGPT uses GPT 3.5 (See release notes: `https://help.openai.com/en/articles/6825453-chatgpt-release-notes`).

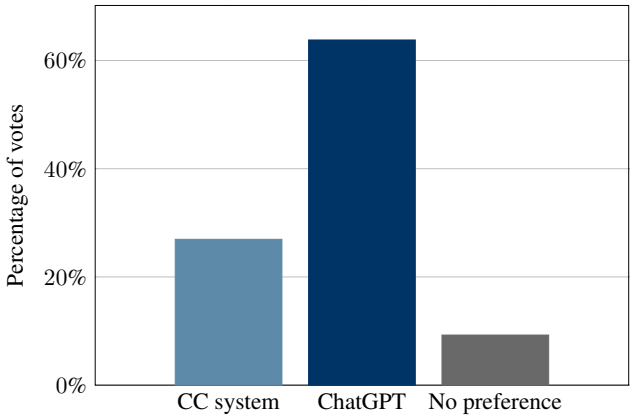

Figure 1: Reviewers' preferences in a one-to-one comparison between CC system generated artifacts and corresponding ChatGPT generated artifacts. These votes are aggregated across all domains and systems.

Results

Responses from 148 individuals resulted in an average of 39.5 responses to each question in the survey. Figure 1 shows reviewers' overall preferences across all domains and systems. The artifact produced by ChatGPT is preferred over the related CC system artifact 63% ($p = 0.014$)[3] of the time. However, the difference in terms of the characteristic evaluation of the two types of artifacts is relatively small. Figure 2 shows a difference of 0.50 Likert scale points ($p = 0.0004$), favoring the ChatGPT artifacts. Using the common significance threshold of 0.05, both of these results are statistically significant.

Figure 3 shows reviewers' preferences broken down by the four domains and aggregated across the three systems in each. For each domain, ChatGPT gains at least 61% of the votes. ChatGPT received the lowest percentage of votes in the poetry domain and the highest in the short story category with 77% of the votes. However, Figure 4 shows that the characteristic scores for the ChatGPT artifacts are relatively close to those for the original CC system artifacts. ChatGPT's lowest mean Likert scale score is in the pun domain with a score of 2.93 which is 0.10 points lower than the CC systems' score. The domain with the largest difference is the six-word story category with a margin of 0.62 points in favor of the ChatGPT artifacts.

The preferences for each artifact generated by their respective CC system along with the ChatGPT generated

[3] Significance is calculated using a paired sample t-test.

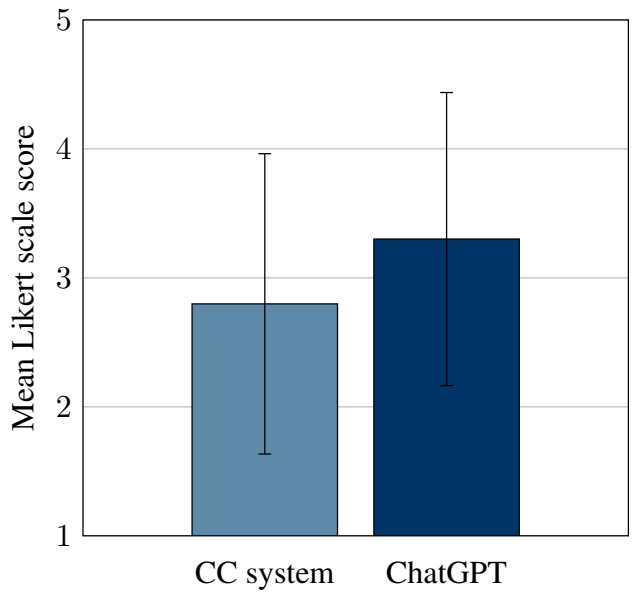

Figure 2: Characteristic evaluation of generated artifacts aggregated across all domains and systems.

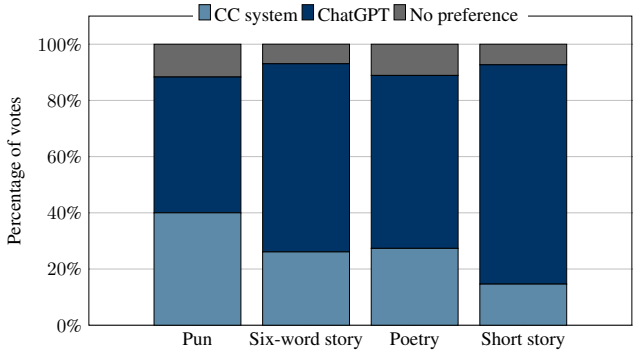

Figure 3: Reviewers' preferences aggregated across systems but broken down by domain.

counterpart is shown in Figure 5. For each system, the ChatGPT artifacts are preferred, with the exception of artifacts produced by PAUL BOT and Chien 2020. It is interesting to note that the INES system did not receive a single vote.

Figure 6 shows the mean Likert scale score for each artifact. The highest score overall belongs to (Chien 2020) which was generated by GPT-2 fine-tuned on a dataset of six-word stories. The characteristic evaluation scores usually correlate with the reviewers' preferences in that preferred artifacts have a higher score, with the exception of Chronicle which is preferred less but has a higher characteristic evaluation score than its ChatGPT counterpart.

For the characteristic evaluations, we can measure agreement between reviewers as a way to further assess our ability to be confident in the survey results, and this inter-rater agreement can be measured using Krippendorff's alpha (Krippendorff 2013). Across all systems and domains (cf.

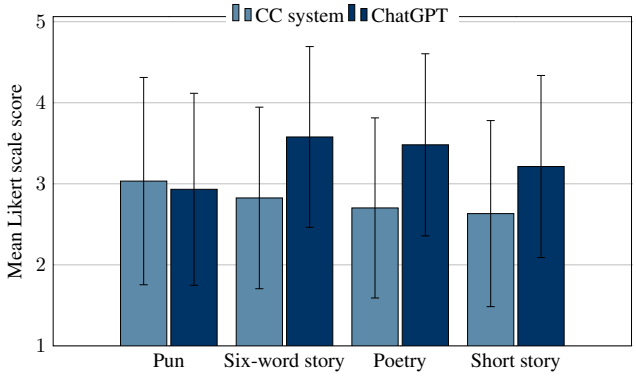

Figure 4: Reviewers' characteristic evaluation of artifacts in each domain, aggregated across systems.

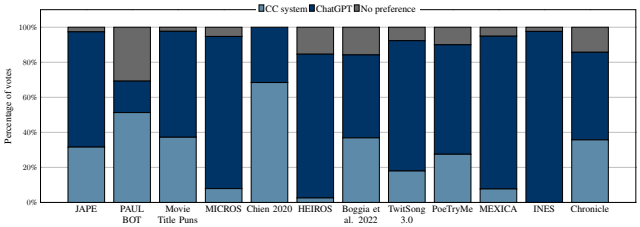

Figure 5: Reviewers' preferences broken down by the system that generated each artifact.

Fig 2), reviewer agreement produces $\alpha = 0.291$. Figures 7 and 8 show reviewer agreement broken down by domain (cf. Fig 4) and system (cf. Fig 6). Each of these values fall well below the recommended threshold of $\alpha \geq 0.8$ that would suggest reliable inter-rater agreement on preference for one system over another.

Discussion

The results seem to indicate that ChatGPT is able to generate artifacts that are just as good or better than the CC systems. This is similar to results found in (Radford et al. 2018) which shows that training a model on a general task like autoregressive language modeling leads to improved zero-shot performance on several downstream tasks as well. In this case, the data show that ChatGPT generalizes to creative tasks by outperforming CC systems overall, as well as at the domain and individual system level. The statistical significance of these results suggests that ChatGPT artifacts are likely to be preferred to and rated higher than (current/traditional) CC system artifacts.

While the results show that ChatGPT is capable of matching or surpassing CC systems in terms of the characteristic evaluation across all domains (Figure 4), the relative difference between CC system and ChatGPT artifacts is not as large as in the direct preferences analysis (Figure 3). In addition, the inter-rater agreement at the overall, domain, and system level is well below the recommended threshold for reviewer agreement, suggesting that characteristic evaluation does not completely explain reviewers' preferences.

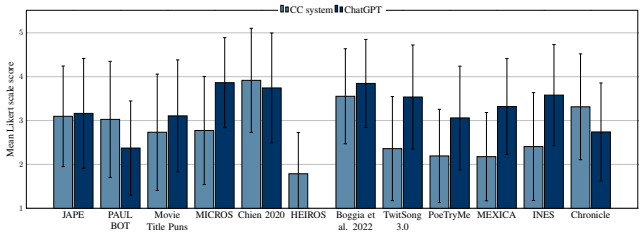

Figure 6: Reviewers' characteristic evaluation of each artifact.

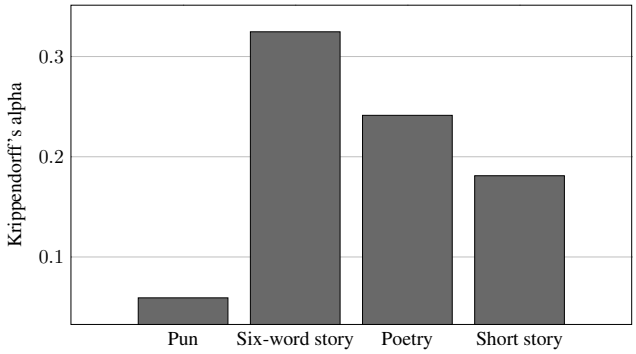

Figure 7: Agreement between reviewers by domain.

One reasonable explanation for this is that an artifact only has to be slightly better in order to be preferred. Although, the presence of a "no preference" option provides confidence that there is a real difference in preference between the artifacts, even if that preference is small.

It is also possible that the criteria used in the characteristic evaluation fail to capture all of the reasons why reviewers prefer an artifact. For example, large language models like ChatGPT are very capable of generating fluent text even if the content of the text is nonsense. In addition, there may be other positive characteristics that ChatGPT includes in its artifacts, such as accessibility to a general audience or even other domain specific characteristics.

It is also reasonable to conclude that the characteristic evaluation is reliable—reviewers generally prefer the ChatGPT artifacts, and while the difference between the artifacts in terms of their character evaluation is not large, the significance testing provides confidence that this difference is, in fact, real. Also, it is important to remember that the artifacts selected for the survey that came from CC systems are (presumably) the best those systems have to offer. On the other hand, ChatGPT's artifacts are not cherry picked and most of the artifacts were generated with a single non-engineered prompt. Therefore, it may be argued that these results may represent a comparison of the floor of ChatGPT's abilities to the ceiling of (traditional) CC systems' abilities.

Implications and Future Work

The findings of this survey do not discount the work of CC researchers. Rather, their accomplishments with significantly fewer resources indicate that many of these traditional

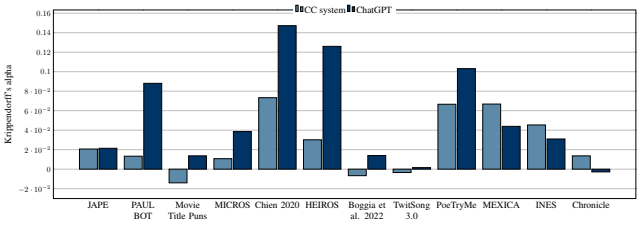

Figure 8: Agreement between reviewers by system.

CC systems are truly ahead of their time. It is also possible that the methods demonstrated by these systems applied at the scale of ChatGPT may outperform ChatGPT.

The purpose of this paper is to spark debate about the creative limitations of language models like ChatGPT and CC systems in general. Given that this level of performance comes from a general language model like ChatGPT means that the purpose and approach of domain-specific CC systems needs to be carefully considered. At the very least, ChatGPT should be used as a baseline when evaluating CC systems going forward.

ChatGPT represents a paradigm shift in terms of interactivity in creative systems. In these experiments, interactive prompts serve to constrain the system to produce corresponding artifacts that are comparable to their CC system counterparts. ChatGPT's ability to do this successfully demonstrates the system's robustness and ease of use. It also suggests a possible move away from fully autonomous systems towards more co-creative solutions (though this certainly doesn't preclude fully autonomous systems in any way, of course.)

These results also highlight an opportunity to improve the performance of language models on creative tasks. While the ChatGPT artifacts are preferred, the overall characteristic evaluation shows that reviewers still have a generally neutral attitude toward the artifacts. It is not yet clear from where these improvements will come, but it is possible that some help may be found in traditional CC approaches.

References

Boggia, M.; Ivanova, S.; Linkola, S.; Kantosalo, A.; and Toivonen, H. 2022. One line at a time — generation and internal evaluation of interactive poetry. In *Proceedings of the International Conference on Computational Creativity*, 7–11. Association for Computational Creativity.

Chien, G. 2020. Generating six-word stories. `http://cs230.stanford.edu/projects_fall_2020/reports/55790134.pdf`.

Concepción, E.; Gervás, P.; and Méndez, G. 2019. Evolving the INES story generation system: From single to multiple plot lines. In *Proceedings of the International Conference on Computational Creativity*, 220–227. Association for Computational Creativity.

Hämäläinen, M., and Alnajjar, K. 2019. Modelling the socialization of creative agents in a master-apprentice setting: The case of movie title puns. In *Proceedings of the Inter-*

national Conference on Computational Creativity, 266–273. Association for Computational Creativity.

Krippendorff, K. 2013. *Content Analysis: An Introduction to Its Methodology*. Sage, 3rd edition.

Lamb, C., and Brown, D. G. 2019. Twitsong 3.0: Towards semantic revisions in computational poetry. In *Proceedings of the International Conference on Computational Creativity*, 212–219. Association for Computational Creativity.

Oliveira, H. G., and Alves, A. O. 2016. Poetry from concept maps–yet another adaptation of PoeTryMe's flexible architecture. In *Proceedings of the International Conference on Computational Creativity*, 246–253. Sony Computer Science Laboratories.

OpenAI. 2023. Introducing ChatGPT. `https://openai.com/blog/chatgpt`. Accessed 2023-4-11.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P. F.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, 27730–27744. Curran Associates, Inc.

Pérez, R. P. Y., and Sharples, M. 2001. MEXICA: A computer model of a cognitive account of creative writing. *Journal of Experimental & Theoretical Artificial Intelligence* 13(2):119–139.

Pickering, T., and Jordanous, A. 2017. Applying narrative theory to aid unexpectedness in a story generation system. In *Proceedings of the International Conference on Computational Creativity*, 213–220. Association for Computational Creativity.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2018. Language models are unsupervised multitask learners. `https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf`. Accessed 2023-4-11.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *Proceedings of the International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 8821–8831.

Ritchie, G. 2003. The JAPE riddle generator: technical specification. Technical Report EDI-INF-RR-0158, School of Informatics, University of Edinburgh.

Spendlove, B., and Ventura, D. 2020. Creating six-word stories via inferred linguistic and semantic formats. In *Proceedings of the International Conference on Computational Creativity*, 123–130. Association for Computational Creativity.

Spendlove, B.; Zabriskie, N.; and Ventura, D. 2018. An HBPL-based approach to the creation of six-word stories. In *Proceedings of the International Conference on Computational Creativity*, 161–168. Association for Computational Creativity.

Tyler, B.; Wilsdon, K.; and Bodily, P. 2020. Computational humor: Automated pun generation. In *Proceedings of the International Conference on Computational Creativity*, 181–184. Association for Computational Creativity.

Wiggins, G. A. 2006. A preliminary framework for description, analysis and comparison of creative systems. *Knowledge Based Systems* 19(7):449–458.