Comparing ChatGPT and Human Raters in Creative Idea Evaluation

Delaney Chiarelli, Mikayla Luskey, Ava Rico, Seth Vallee
Department of Psychology, University of Texas at Arlington

Graduate Student Mentor: Ellie Tran
Faculty Mentor: Logan Watts, Ph.D.

pelican lab

Introduction

- Data coding is a time-consuming research task that can be affected by human biases and cognitive constraints such as fatigue and attention lapses (Kovacs et al., 2021).
- AI software could address the issues of human fatigue and biases. However, its accuracy requires further investigation, as previous research found AI offers consistency while humans are better at nuanced understanding (Prescott et al., 2024).
- ChatGPT, developed by OpenAI, could automate creative evaluation, but its ability to assess creativity requires further testing.
- The purpose of this study is to evaluate the performance of free and paid versions of ChatGPT in coding creativity dimensions and compare their results to those of human coders.
- This study will contribute to understanding how AI models can be used in creative assessments and whether they can match or complement human evaluators in creativity tasks.

Research Questions

1. Are there differences in ratings of creativity dimensions (originality, usefulness, and elegance) among the free version of ChatGPT, paid Plus version of ChatGPT, and human raters?
2. How often do human raters and ChatGPT give extreme ratings (score of either 1 or 5) on creativity dimensions?

Methods

Sample

- A total of 328 creative plans, coded by human raters on creativity dimensions in a previous study, were also provided to free ChatGPT and ChatGPT Plus for rating on the same dimensions.

Measures

- Each creative plan was rated on three dimensions of creativity using a 5-point Likert scale (1 = poor, 5 = excellent)
  1. Originality: How original and novel the plan is.
  2. Usefulness: Overall quality and feasibility of the plan.
  3. Elegance: How well the plan is designed and flow.

Procedures

- The coding manual used to train human raters was adapted as the prompt for ChatGPT.
- This prompt was input into both the free and paid versions of ChatGPT, and the creative plans were provided one at a time for evaluation.

Analyses

Results

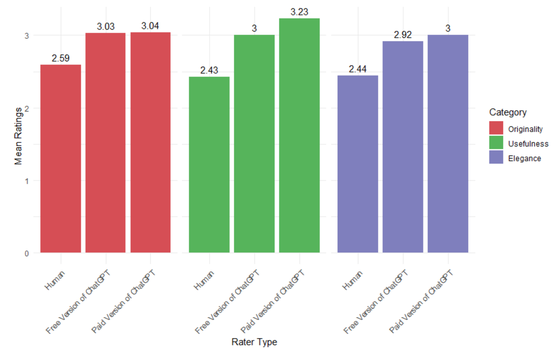
Table 1

Means, Standard Deviations, and ANOVA Results for Ratings by Human, Free ChatGPT, Paid ChatGPT Plus

Variables	Human		Free Version of ChatGPT		Paid Version of ChatGPT		ANOVA Results			
	M	*SD*	*M*	*SD*	*M*	*SD*	*df*		*F*	*p*
							Between	Within		
Originality	2.59	0.84	3.03	0.96	3.04	0.99	2	810	20.29	<.001
Usefulness	2.43	0.93	3.00	0.99	3.23	1.04	2	810	46.94	<.001
Elegance	2.44	1.00	2.92	1.01	3.00	1.06	2	810	24.20	<.001

Figure 1

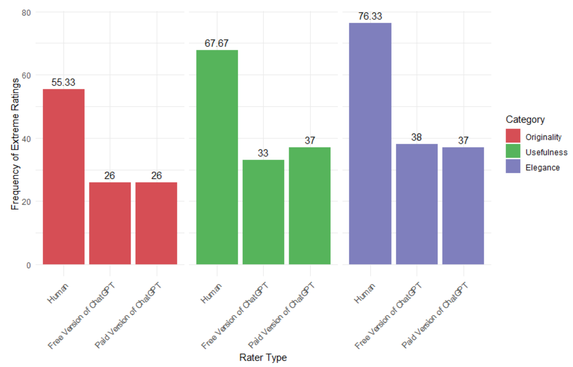
Mean Ratings by Human Raters, Free ChatGPT, and Paid ChatGPT Plus



- There were significant differences between human and AI ratings in all creativity dimensions: originality, usefulness, and elegance.
- **Originality**: Human rating scores were lower than scores from both free ChatGPT and ChatGPT Plus.
- **Usefulness**: Human rating scores were lower than scores from free ChatGPT. Free ChatGPT scores were lower than ChatGPT Plus scores.
- **Elegance**: Human rating scores were lower than scores from both free ChatGPT and ChatGPT Plus.

Figure 2

Frequency of Extreme Ratings by Human Raters, Free ChatGPT, and Paid ChatGPT Plus



- **Originality**: Human raters provided extreme ratings (1 or 5) an average of 55.33 times, while AI did so an average of 26 times.
- **Usefulness**: Human raters provided extreme ratings (1 or 5) an average of 67.67 times, while AI did so an average of 35 times.
- **Elegance**: Human raters provided extreme ratings (1 or 5) an average of 76.33 times, while AI did so an average of 37.5 times.

Discussion

- ChatGPT's significantly higher ratings than human ratings suggest that AI may offer more lenient evaluations.
- Fewer extreme ratings by ChatGPT compared to human raters may result from its reliance on patterns and algorithms for consistency, as previous research has highlighted algorithmic bias as a common issue in AI systems across domains (Min, 2023).

Implications

- This study contributes to the growing literature on AI use in psychology research.
- Despite AI's efficiency, rating differences between humans and ChatGPT, as well as between two ChatGPT versions, raise concerns about AI's validity and reliability in assessing creativity, indicating the need for caution of AI use in data coding until further validation.
- AI's limited context understanding may lead to neutral evaluations (Prescott et al., 2024). This suggests a need for better training to improve context comprehension.

Limitations and Future Decisions

- While the same instructions were given to ChatGPT as to human raters, the lack of clarity or precision in the prompts could lead to coding errors. Future research could refine prompt design and explore prompt engineering techniques to improve AI accuracy.
- This study only used ChatGPT models, which may limit the generalizability of the findings. Future research could compare multiple AI models to assess which provides the most accurate and consistent ratings.

References

1. Kovacs, M., Hoekstra, R., & Aczel, B. (2021). The Role of Human Fallibility in Psychological Research: A Survey of Mistakes in Data Management. *Advances in Methods and Practices in Psychological Science, 4*(4).
2. Min, B. (2023). Artificial Intelligence and Bias: Challenges, Implications, and Remedies. *Journal of Social Research, 2*(11), 3808-3817.
3. Prescott, Maximo R et al. "Comparing the Efficacy and Efficiency of Human and Generative AI: Qualitative Thematic Analyses." *JMIR AI.* 3 (2024): n. pag. Web.

DISCOVER