HOW CREATIVE IS AI WRITING? GENERATIVE AND COLLABORATIVE AI IN JAPANESE FICTION

Yuki Asano
City St George's, University of London

Marco Bastos
University College Dublin & City St George's, University of London

Extended abstract

The Author's Guild disclosed that 23% of writers have reported using generative AI in their creative process (Milliot, 2023). The growing presence of human-AI collaboration in professional writing is embodied in Sean Michaels' novel 'Do You Remember Being Born?' which incorporates AI-generated sentences (Michel, 2023). AI-generated content plays a meta-fictional role in the novel, as a poet collaborates with another poet using AI. Other examples of award-winning literary works produced through human-AI collaboration include Japanese author Rie Qudan, who received one of Japan's most prestigious awards for early-career writers for her novel Tokyo-to Dojo-to, in which approximately 5% of the text was generated by AI (Ha, 2024). About 10% of submissions to the Hoshi Shinichi Award—a Japanese literary prize that permits AI-generated submissions—consist of AI-generated sentences, an epochal change that illustrates a model of human-AI collaboration in which AI-generated text is based on human-written drafts and subsequently revised.

Several studies have sought to distinguish between AI-generated and human-written texts (Hitsuwari et al., 2023; Köbis & Mossink, 2021; Proksch et al., 2024; Sardinha, 2024). Cognate literature has questioned the authenticity of AI-generated literature (Hutson & Schnellmann, 2023; Pretsch, 2023). Similarly, Niloy et al. (2024) rolled out an experimental study with 600 students across 10 universities that incorporated ChatGPT into creative writing and found a detrimental association between ChatGPT use and university students' creative writing abilities, whereas Yan (2025) devised a narrative experiment with ChatGPT and argued that the tool supports posthuman creativity. There is also a sizable body of research on algorithm aversion, an effect grounded on bias against algorithms compared to human agents. This informs a central component of our study, which is the assumption that readers are likely to associate higher-quality writing with human authorship due to algorithm aversion. In the study carried out by

Köbis and Mossink (2021), participants exhibited a mild aversion to AI-generated poetry regardless of whether they were informed of the poem's algorithmic origin. Sardinha (2024) found that readers can distinguish AI-generated texts from human-written texts, often exhibiting algorithm aversion. Similarly, Proksch et al. (2024) found that texts labeled as human-written were consistently rated as being of higher quality.

It is against this backdrop that we examine the quality of generative and collaborative AI for fictional prose in Japanese. In our experiment with 60 participants, the first half of each story is always written by professional or amateur writers with minor literary awards. It consists of approximately 300-500 Japanese syllables extracted from contemporary fiction written within the past 30 years, with minimal use of dialects or traditional language. The samples include literary fiction, entertainment fiction, and light novels (a type of young adult fiction native to Japan). The second half of each story would randomly present one of the following options: A: Original human-written text; B: Zero-shot-AI-generated text; C: AI-generated with prompt specification; D: Human-AI collaboration using multiple-shot and prompt specification. The experiment includes a total of 100 pieces of Japanese fiction. After collecting data from the experiments, we rolled out focus groups to elicit responses from a subset of the participants.

The results of the experiment show that version D: Human-AI collaboration is the most favorably perceived by the highest number of readers (Table 1). Version D achieved consistently high scores while avoiding the extreme highs and lows observed in other versions. This outcome is likely due to the nature of AI-generated text, which is produced based on phrase patterns learned from internet sources. Version A received the second-highest average score; however, its ratings show significant variation with. Figure 1 shows a broad range of responses for A characterized by a large interquartile range and extended whiskers that indicate the scores span from very low to very high. This suggests that participants evaluated version A in markedly different ways. Notably, while some readers perceived version A as being of exceptionally high quality, others rated it much lower.

Table 1: Summary of average scores of A-D

Version	Average score/respondant	Standard Deviation	Minimum	1st Qunatile	Median	3rd Qunatile	Maximum
A	139.6	32.6	81.0	115.8	133.0	161.0	211.0
B	130.7**	29.1	69.0	112.0	130.0	151.0	203.0
C	134.6	22.5	76.0	121.0	133.0	148.0	187.0
D	145.0***	20.0	90.0	133.0	143.0	157.0	202.0

Note: T-test on null hypothesis that sample mean is greater/less than expected average score per respondent (i.e. 137.5). ***p<0.01, **p<0.05, *p<0.1
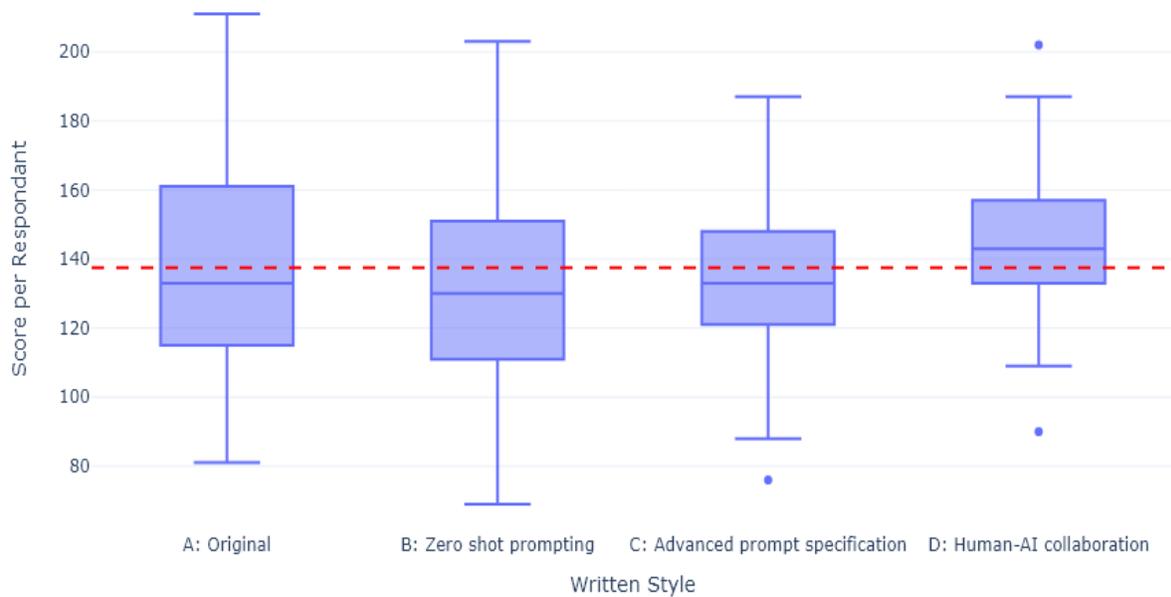
Figure 1: Boxplot of scores assigned by participants for each written output

Version C: Detailed prompt specification received the third-highest average score with a narrower range of variation comparable with version D, but also with a lower score range. Figure 1 shows a compact interquartile range and shorter whiskers similar to version D, albeit with a lower mean. The distribution scores shown in Figure 2 further corroborate this, showing a bell-shaped curve centered around the mean, with fewer extremely high or low scores compared to version D. These findings suggest that version C is a less favorable iteration of version D. When comparing version C to version B, version C exhibits a higher mean and a smaller range of variability (Figure 2). Version B: Zero-shot prompt received the lowest average score, with a wide range of variability concentrated in the lower scoring range. This suggests that without structured prompts, AI-generated text is more likely to result in lower perceived quality.

Perceptions of AI-generated text varied between participants of the first and second focus group sessions. Participants of the first session agreed with the view that AI-generated fiction was difficult to distinguish from human-written text. This perception was attributed to the similarity between AI-generated wording and human language, making it difficult to set them apart (see focus group transcripts in the supplementary materials). Conversely, participants of the second focus group reached a consensus that distinguishing between human- and AI-generated fiction varies between samples, with one participant describing it as 'fifty-fifty.' Nevertheless, all participants perceived AI-generated texts as somewhat 'unnatural' or containing tasteless content. These findings are in line with previous research suggesting that readers can differentiate between human- and AI-generated texts but are less likely to be able to distinguish these differences in shorter works such as poetry.
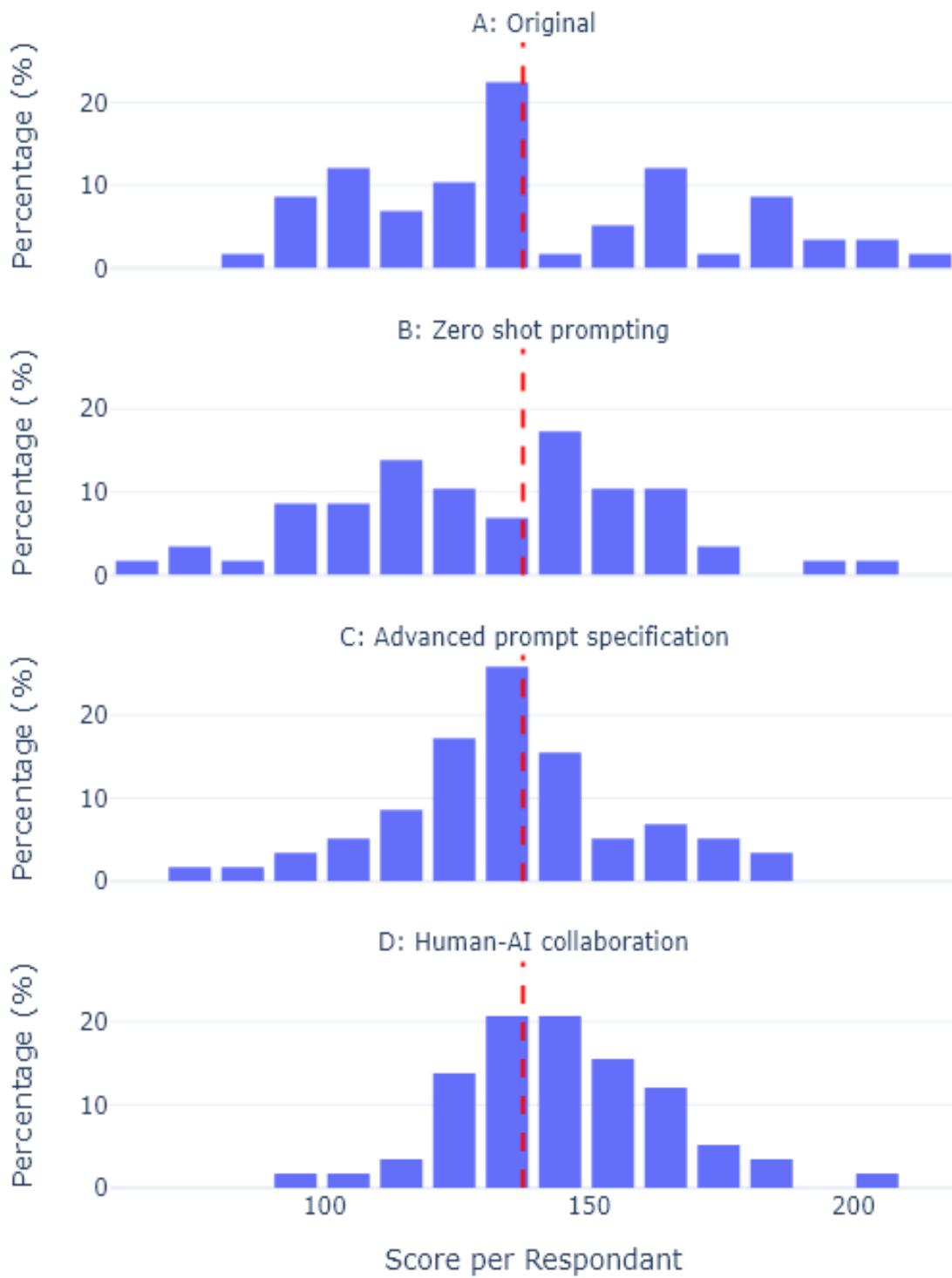
Figure 2: Distribution of scores for each output

When discussing the factors that determine the quality of fiction during the focus groups, participants mentioned narrative flow, story transitions, dialogue, the expression of human emotion, and spontaneity. Participants highlighted conversational tone, proper nouns, and poetic expressions as particularly reflective of human emotion. Notably, several participants reported that they assessed the texts by focusing on linguistic expressions and scene transitions rather than the overall theme, as they were limited to evaluating short passages. Consistent with the literature on algorithm aversion, participants exhibited a tendency to react negatively to texts that were clearly AI-generated. Such responses were primarily driven by perceptions of uncanny, incoherence, and reliance on formulaic expressions. This strategy for identification of AI-generated text was mentioned in both focus group sessions.

The results reported in this study underscore the potential of human-AI collaboration in fiction writing, particularly through human editing of AI-generated content. These findings also highlight the growing role of AI in professional creative writing and raise critical discussions on authorship, originality, and the evolving nature of literary production. The results of the experiment, in particular, indicate that human-AI collaboration received the highest average score, a finding corroborated by the focus groups where participants described D as well-structured and free from notable flaws.

References

Ha, T.-H. (2024). Akutagawa Prize draws controversy after winning for work that used ChatGPT. The Japan Times. https://www.japantimes.co.jp/culture/2024/01/19/books/akutagawa-prize-book-chatgpt/

Hitsuwari, J., Ueda, Y., Yun, W., & Nomura, M. (2023). Does human–AI collaboration lead to more creative art? Aesthetic evaluation of human-made and AI-generated haiku poetry. Computers in Human Behavior, 139.

Hutson, J., & Schnellmann, A. (2023). The poetry of prompts: the collaborative role of generative artificial intelligence in the creation of poetry and the anxiety of machine influence. Global Journal of Computer Science and Technology: D, 23(1).

Köbis, N., & Mossink, L. D. (2021). Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. Computers in Human Behavior, 114.

Michel, L. (2023). Will A.I. Change Art? A New Novel Uses AI to Explore Just That. The New York Times. https://www.nytimes.com/2023/09/04/books/review/do-you-remember-being-born-sean-michaels.html

Milliot, J. (2023). Authors Guild issues contract clause changes to account for AI. Publishers Weekly. https://www.publishersweekly.com/pw/by-topic/industry-news/industry-deals/article/92464-authors-guild-issues-contract-clause-changes-to-account-for-ai.html

Niloy, A. C., Akter, S., Sultana, N., Sultana, J., & Rahman, S. I. U. (2024). Is ChatGPT a menace for creative writing ability? An experiment. Journal of Computer Assisted Learning, 40(2), 919-930.

Pretsch, E. (2023). Artificial Intelligence and creativity in poetry: effect of AI-written poems on human emotions. Journal of Creativity and Inspiration, 1(1).

Proksch, S., Schühle, J., Streeb, E., Weymann, F., Luther, T., & Kimmerle, J. (2024). The impact of text topic and assumed human vs. AI authorship on competence and quality assessment. Frontiers in Artificial Intelligence, 7.

Sardinha, T. B. (2024). AI-generated vs human-authored texts: A multidimensional comparison. Applied Corpus Linguistics, 4(1).

Yan, D. (2025). Posthuman creativity: Unveiling cyborg subjectivity through ChatGPT. Qualitative Inquiry, 31(2), 253-264.