City Research Online

City St George's, University of London

Citation: Asano, Y. & Bastos, M. (2026). How Creative is AI writing? Generative and Collaborative AI in Japanese Fiction. Journal of Creative Communications,

This is the accepted version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/37728/

Copyright and Reuse: Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to City Research Online policy.

City Research Online: http://openaccess.city.ac.uk/ publications@citystgeorges.ac.uk

How Creative is AI writing? Generative and Collaborative AI in Japanese Fiction

Yuki Asano (yuki.asano@city.ac.uk) & Marco Bastos (marco.bastos@ucd.ie)

Accepted for publication in *Journal of Creative Communications*

Abstract

This study examines readers' perception of quality and originality in AI-generated and human-AI collaborated Japanese literary fiction. Fifty-eight participants with backgrounds in creative writing ranked the endings of 25 stories across four conditions. In all cases, the first half of each story was authored by professional or amateur writers; the second half varied by condition: original human-authored text, zero-shot AI generation, AI generation with detailed prompt specification, and human-AI collaboration combining multiple-shot prompting with human editing. Focus group sessions followed to explore the criteria underlying the ranking provided by participants. Human-AI collaboration received the highest mean quality rankings, followed by original human writing, detailed prompt specification, and zero-shot generation. While the differences between conditions were small relative to within-condition variance, the contrast between human-AI collaboration and zero-shot generation was statistically significant. In addition to that, collaboration was rated highest in perceived quality but lowest in originality, a dissociation that previous research had not clearly identified: AI-assisted editing produces polished, coherent prose that readers find accessible, but at the cost of the stylistic expression that defines literary fiction. These findings suggest that the contribution of Generative AI to creative writing is better characterized as a reduction of variance than a tangible elevation of quality.

Keywords: ChatGPT; Generative AI; Fiction Writing; Creative Writing; Originality; Literature

Introduction

The public launch of OpenAI's generative artificial intelligence chatbot ChatGPT in November 2022 garnered extensive media attention by highlighting its ability to produce remarkably intricate and human-like text. Built on top of OpenAI's GPT-3.5 families of Large Language Models (LLMs), ChatGPT was a flashpoint for wider social anxieties about the role of human agents in the writing process, particularly in creative writing like literature, poetry, and screenwriting (Halperin and Rosner, 2024). A plethora of AI-powered novel-writing tools emerged following the release of ChatGPT, including Novel AI, Sudowrite, Sassbook, Rytr AI, and Squibler, which catered both to amateur and professional creative writers. In the following years, AI-generated fiction had a profound impact on the practice of fiction writing and the broader creative writing industry by introducing new modes of content production, reshaping authorship practices, redefining conceptions of creativity, and altering the fundamental nature of creative process and artistic engagement.

ChatGPT generates text by estimating probable token sequences, drawing on training across large volumes of internet text data. This computational process requires no human creativity in the conventional sense: the model produces fluent, contextually appropriate language through statistical pattern-matching rather than intentional expression. Built on Transformer architecture and trained using self-supervised learning, ChatGPT is designed to sustain natural dialogue with users (Oka and Hashimoto, 2024). The Transformer model is based on the concept of self-attention, which captures the relationships between tokens within a sequence and represents them as weighted connections (Vaswani, 2017). Multi-head self-attention extends this by processing the same sequence through multiple attention layers simultaneously and allowing the model to represent different types of linguistic relationships in parallel. Self-supervised learning enables the model to train itself without human-labeled

data by generating prediction tasks directly from existing text and comparing the output against the correct answers. Through repeated exposure to large volumes of internet text, ChatGPT learns to identify statistical regularities in token sequences. A key consequence of this is that output quality varies with the frequency of relevant patterns in the training data: when a requested style or topic is well-represented, responses tend to be fluent and consistent; when it is not, outputs may be unpredictable or incoherent. This has direct implications for literary fiction, where distinctive and infrequent stylistic patterns are central to quality writing (Oka and Hashimoto, 2024).

The integration of AI into creative writing has reached a stage where its influence can no longer be overlooked. It is increasingly assisting human writers in their work regardless of their willingness to adopt it. Public discourse about this set of tools is nonetheless divided on its implications for creative practice. While observers have described ChatGPT as a disruptive innovation threatening human creativity by automating processes that were previously the exclusive domain of human expression, others have argued that it functions as a tool for creativity amplification that extends rather than replaces what human writers can do (Clark, 2022). This tension between AI as a threat to authentic creative authorship and AI as an instrument for expanding creative possibilities remains unresolved in both public discourse and academic research, and it informs the research design and the central question driving this study.

The Authors Guild disclosed that 23% of writers report using generative AI in their creative process (Milliot, 2023). Among these, 47% utilize AI for grammatical assistance, 29% for generating plot ideas and character development, 14% for structuring or organizing drafts, and 7% for direct text generation (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. There are also examples of award-winning literary works produced through human-AI collaboration. Japanese author Rie Qudan, who received one of Japan's most prestigious awards for early-career writers for her novel Tokyo-to Dojo-to, revealed post-award that approximately 5% of the text was generated by AI (Ha, 2024).

There are more extreme cases for awards that allow for using Generative AI. The winner of the Hoshi Shinichi Award—a Japanese literary prize that permits AI-generated submissions—described using AI extensively in his creative process. Kamome (2022) generated approximately one hundred short stories by collaborating with AI, refining drafts using GPT-2, and then editing the outputs. Ultimately, about 10% of the award-winning work consisted of AI-generated sentences, illustrating a model of human-AI collaboration in which AI-generated text is based on human-written drafts and subsequently revised. Similarly, Yang's 'The Land of Machine Memory' won a national competition in China that placed no restrictions on AI-generated content. The piece was generated using AI within three hours. It employed 66 prompts that yielded 40,000 Chinese characters, ultimately edited down to 5,900 characters for the final version (Chik, 2023).

It is against this backdrop that this study examines readers' perceptions of quality and originality in fiction written by humans, generated by AI, and produced through human-AI collaboration. The study was designed with ecological validity in mind: the human-authored stories used in the experiment were written by professional and amateur fiction writers, one of whom is a member of the research team. We address three specific questions: whether readers can distinguish human-written narratives from AI-generated and AI-assisted content; which versions are perceived as higher quality; and how readers articulate their evaluative criteria

when the provenance of the content is not disclosed.

For the purposes of this study, fiction refers to literary prose narratives—a category that excludes essays, reviews, non-fiction, scripts, poetry, and short-form verses such as Haiku and Tanka. The corpus is limited to works written in Japanese and produced specifically for this research. We distinguish between two types of AI involvement: AI-generated fiction, in which the narrative is produced entirely by AI—with or without prompt specification—and without any human editing; and AI-collaborated fiction, in which AI-generated text is incorporated into a narrative that has been subsequently edited or restructured by a human author.

Previous Work

The literature on creative writing and Generative AI has investigated three central effects that inform this study: 1) human-AI collaboration consistently improves measurable output quality in professional writing contexts; 2) algorithm aversion is a real but variable effect, attenuated in shorter literary forms and when AI authorship is not perceptible; and 3) AI's generative capacity is limited by its inability to replicate the emotional depth, cultural specificity, and narrative originality that define literary fiction. While most of the extant research has been conducted in English and focused on poetry, short reports, or technical writing, our study examines readers' perceptions of AI-generated and AI-collaborated literary fiction in Japanese using a four-condition experimental design that maps quality perception across the full spectrum of AI involvement. In the following, we unpack the literature that informed our study and the central concepts tested with our experimental design.

Generative AI and Creative Writing

The proliferation of AI-generated text has had a pronounced disruptive effect on the publishing industry, with a marked uptake in self-published, low-quality works from amateur writers. ChatGPT, in particular, triggered a surge in self-published content on platforms such as Amazon's Kindle Direct Publishing (KDP), where financial incentives rather than literary merit drive production (Bensinger, 2023). By January 2024, more than 4,000 English-language e-books listed on Amazon credited AI as sole author or co-author, bypassing the editorial controls that established publishers carefully maintained. A stark illustration of this problem is the publication of an AI-generated book about the 2023 Maui wildfires within days of the disaster. Despite the poor quality and sensationalist nature of the book, it went on to become an Amazon bestseller (Quach, 2023). The corporation responded to this problem by requiring authors to disclose any use of AI and limiting publication to no more than three AI-generated books per day.

Concerns about intellectual property have further intensified the debate: technology companies have reportedly used nearly 200,000 books to train AI systems without informing or compensating their authors (Asmelash, 2023). Due to AI's reliance on pre-existing literary works, writers have promptly identified their own work within AI-generated texts (Mauran, 2023; Tapper, 2023). This prompted copyright infringement lawsuits by prominent fiction writers (Alter and Harris, 2023) and calls from the Authors Guild (2023) and the European Writers Council (2023) for legal safeguards against unauthorized extraction of copyrighted text (Milliot, 2023). These developments establish the broader industrial and legal context within which this study is situated.

Scholarly debate on AI's role in creative writing has examined the distinguishability between AI-generated and human-written texts (Sardinha, 2024; Proksch et al., 2024;

Hitsuwari et al., 2023; Köbis and Mossink, 2021) and the authenticity of AI-generated literature (Pretsch, 2023; Hutson and Schnellmann, 2023; Çelik, 2023), with critics pointing that such tools dehumanize creativity by hindering the labor dimension of creative work (Lee and Lee, 2022). Niloy et al. (2024) conducted an experiment with 600 students across ten universities and found that ChatGPT use was negatively associated with creative writing ability, with human evaluators confirming its adverse impact on originality, elaboration, and content quality. Fang et al. (2023) systematic review of AI in story writing, on the other hand, reported that AI technologies increased writing proficiency, student engagement, and motivation, with potential benefits for language acquisition. Similarly, Yan (2025) devised a narrative experiment with ChatGPT and argued that the tool supports a form of posthuman creativity in which human and machine agency are productively entangled.

Productivity Gain and Algorithm Aversion

Research in the area also tested the 'productivity gain' hypothesis by probing whether Generative AI improves the efficiency and output quality of professional writing in controlled settings. Noy and Zhang (2023) ran randomized controlled trials with 444 college-educated professionals in human relations and marketing who performed 20-30 minutes tasks like writing news releases and brief reports. Participants using ChatGPT completed the tasks 37% faster than those who did not have access to the tool, the reported productivity effect, which also boosted job satisfaction by 20%. Peng et al. (2023) found comparable effects among software developers using GitHub Copilot, who completed entry-level tasks 55% faster than the control group. While these studies have established that human-AI collaboration yields measurable performance gains in technical writing, it is unclear if this productivity advantage extends to the domain of literary fiction where aesthetics and subjective evaluations prevail over speed or accuracy measurements.

Conversely, a substantial body of experimental research has examined whether readers can reliably distinguish AI-generated from human-written text, and whether the knowledge of AI authorship triggers a negative response—an effect termed 'algorithm aversion'—compelling readers to rate texts lower whenever they are identified or suspected to have been generated by AI. Sardinha (2024) found that readers can distinguish AI-generated texts from human-authored ones along multiple dimensions, with aversion effects following identification. Proksch et al. (2024) reported that texts labeled as human-written were consistently rated higher in quality than identical texts labeled as AI-generated, suggesting that authorship attribution shapes perceived quality independently of the text itself. Similarly, Köbis and Mossink (2021) found a mild aversion to AI-generated poetry even when participants were not explicitly told about its provenance. This body of work suggests that human authorship carries an implicit quality premium in readers' perceptions and informs another component of our study: the assumption that readers are likely to associate higher-quality writing with human authorship due to algorithm aversion.

Other studies complicate the algorithm aversion thesis, particularly for shorter literary forms. Porter and Machery (2024) found that non-expert readers performed below chance in identifying AI-generated poems by well-known human poets, and that AI-generated poems were rated more favorably in terms of rhythm and aesthetic quality and that this contributed to their mistaken identification as human-authored. Köbis and Mossink (2021) reported that participants presented with randomly selected poems could not reliably distinguish between AI-generated and human-authored content. Hitsuwari et al. (2023) found similar results for Haiku, a traditional Japanese poetic form consisting of 17 syllables: participants could not reliably distinguish AI-generated Haiku from human-authored ones. Kuwahara's (1946) seminal study showed that even Haiku by celebrated and amateur poets are often

indistinguishable from one another, suggesting that the brevity and structural constraints of the form—rather than any equivalence of poetic expression—renders this format more challenging to classification tasks.

A recurring concern in the literature reviewed hitherto is whether AI-generated content can be genuinely creative or is inexorably lacking in personal experience, emotional consciousness, and lived human subjectivity—the very qualities that give literary fiction resonance. This speaks to a central question driving this study: whether readers can detect the absence of genuine creative agency in AI-generated literary fiction, and if so, through which textual features this is perceived. Another debate in the literature that informs our study is the tension between two conflicting positions: AI as amplifier versus AI as substitute, a contention that underpins much of the 'productivity gain' hypothesis. Lastly, the studies reviewed above also suggest that algorithm aversion is genre-dependent: the effect is weaker in shorter, more formulaic literary forms, and likely stronger in longer prose with greater scope for narrative inconsistency and stylistic individuality.

Data & Methods

Prompt engineering refers to the practice of crafting inputs to large language models in ways that maximize the relevance and quality of their outputs. In the context of fiction writing, Fang et al. (2023) identify three core AI approaches: script learning and generation, which uses story corpora to produce narratives; story completion, which generates a continuation when provided with an opening context; and story generation, which supports users in producing stories through a range of natural language processing tasks (Fang et al., 2023). Effective prompting requires inputs that fully engage the model's knowledge and reasoning capabilities (Oka and Hashimoto, 2024). To this end, Oka and Hashimoto (2024) identify

several prompt patterns relevant to fiction writing. The persona pattern instructs the model to adopt a specific role (e.g., ‘please behave as persona X and complete the task’); audience persona pattern (e.g., ‘please explain X, assuming I am persona Y’); question refinement (e.g., ‘when I ask questions, please always propose improved questions’); flipped interaction (e.g., ‘to accomplish X, please ask me questions until the objective is accomplished’); and trigger prompt (e.g., ‘let us solve the problem step by step’ or ‘please think through this step by step.’)

Applied to fiction creation, prompt engineering allows writers to specify style, audience, genre, and narrative context when querying an LLM—for instance, by requesting a story written in the style of a particular author, aimed at readers in their twenties and thirties, or set within a defined situation or backdrop. However, there remain significant limitations in current AI systems for their effective use in this domain. Gholami et al. (2024) identify restricted memory capacity as one of the primary challenges of AI-assisted fiction writing: as a story progresses, the model tends to lose track of earlier plot points and character details, which result in narrative inconsistencies. Addressing this requires carefully designed prompting strategies that reinforce contextual continuity throughout the generation process. This limitation also implies that Generative AI is currently most effective for short stories or discrete fictional passages rather than extended narratives.

Our expectation is that ChatGPT-4 with zero-shot prompting will produce boilerplate prose of limited literary appeal, given the tendency of LLMs to default to generic, formulaic language in the absence of detailed guidance. We further expect that sophisticated prompt specification will improve on this baseline, and that human editing of AI-generated content will produce output of quality comparable to or exceeding that of the original human-authored text. To test this hypothesis, we construct four versions of each story’s second half. Version A

is the original ending written by the human author, and it serves as the baseline. Version B is generated by ChatGPT-4 using only the first half of the story as input, with no additional prompt specification. Version C is generated by ChatGPT-4 with detailed prompt specification, instructing the model to write as an experienced fiction writer. Version D is ChatGPT-4 output that has subsequently been edited for clarity and style by the story's original author. We expect this last sample to be difficult to identify as it was composed with support from ChatGPT-4 but amended through human editing. Each participant reads and ranks all four versions for each story without being informed about which version was produced by which method.

Stories

We proceed by creating a corpus of 25 stories, each with four alternative endings, thereby yielding 100 story versions in total. The use of 25 stories is intended to provide sufficient variation across narrative styles, genres, and authors to allow for generalizable conclusions rather than findings contingent on the properties of a single text. In other words, by drafting a large number of stories, we expect to be able to establish whether the results are consistent or within a margin of error. For each story, participants read the shared first half followed by all four endings, presented in randomized order to avoid order effects, and rank them from best to worst. The full corpus, comprising over 73,000 Japanese syllables, is publicly available at the repository of the project.

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 novels with minimal use of dialects or traditional language. The samples include 11 pieces of pure literary fiction, 10 pieces of entertainment fiction, and 4 light novels (a type of young adult fiction native to Japan), including fantasy, classified based on the major

categories of contemporary Japanese fiction. The second half of each story breaks down as follows: A) **Original human-written text**; B) **Zero-shot prompt**: AI-generated content created by querying ChatGPT with the first half of the story and instructing it to write the latter half with no additional guidance; C) **Detailed prompt specification**: AI-generated content created using an elaborated query with detailed prompt engineering techniques as discussed above. Instructions include various patterns, such as persona, audience persona, question refinement, flipped interaction, and trigger prompts. In some cases, the research team specifies genres such as mystery or horror or includes specific words; D) **Human-AI collaboration**: this output results from close collaboration between the author and ChatGPT-4, including editing, selecting, adding, or providing drafts. Collaboration can range from ChatGPT-4 writing most of the content with the research team editing it afterward, the research team writing most of the content with ChatGPT-4 editing it, or a combination thereof. In some cases, the research team asked ChatGPT-4 for advice on how to continue the story.

To summarize the experimental design: Version A contains no AI input; Versions B and C each contain approximately 50% AI-generated content; and Version D contains approximately 25% AI-generated content, with the remainder having been edited or rewritten by the original human author. The experiment was rolled out using Google Forms, which required the research team to make adjustments for line breaks and consistency. Completing the survey required participants to read and rank all 25 stories, taking between 60 and 150 minutes. This experimental part of the project included a survey with demographic information and a supplementary section on Japanese literature and creative writing. The resulting data, available at the repository of the project, identifies how participants ordered A-D in each of the stories and their endings, which yielded 100 different stories.

Recruitment

A total of 60 individuals were recruited by targeting professionals and students engaged in creative writing and related fields in Japanese. We recruited individuals with a BA or MA in Japanese literature, a BA or MA in Creative Writing in Japanese, or who were active in the creative writing industry, including in novel writing and editorial work. While participation is not restricted by nationality or ethnicity, participants are either Japanese native speakers or with professional fluency of the language. They are also avid readers of Japanese novels and thus in a position to make an informed assessment of the material provided. Participation in the experimental portion of this project took place between July and August 2024. Participant's input was registered using Google Forms, as most competing platforms do not support the inclusion of large texts in Japanese. Informed consent was provided in Japanese and English. Experimental survey data and reports are available at the repository of the project.

Two weeks after the experimental phase we rolled out two focus group sessions to gain insights into participants' reasoning and the criteria underlying their rankings (Bryman, 2015). Participants were selected based on their active participation in the creative writing or literature-related sector. Focus group participants were drawn from the experimental sample, selecting individuals who were actively engaged in creative writing or literature-related fields. To minimize social desirability effects and encourage candid discussion, participants within each session were unknown to one another. Focus group sessions were carried out in mid-August following the traditional Obon holiday to maximize recruitment, as this is a period when many Japanese nationals return to their hometowns. The first focus group session was held at 8 PM Japan Time on August 15, 2024, with four participants. The following session took place at 8 PM on August 16, 2024, with three participants.

During the focus group sessions we probed the reasons behind their personal rankings, including the criteria employed, and queried about the distinguishability of the story samples. Key to the focus groups was to enable in-depth exploration by engaging several participants interactively and simultaneously. A member of the research team facilitated the focus group sessions by asking questions and prompting discussions with follow-up questions and summaries, while also encouraging all participants to share their thoughts (Bryman, 2015). Participants were informed that the focus group sessions were recorded for transcription purposes and signed a participant consent form. Data from focus groups 1 and 2 are available at the repository of the project.

Each focus group session began with participants signing a consent form, followed by the provision of all four versions of each story so that participants could refer back to the texts during discussion. A 50-minute guided discussion then addressed three broad themes: first, participants' ability to identify AI-authored content, the cues they used to do so, and the factors that made identification difficult; second, the criteria they applied when evaluating the quality of each version; and third, their preferences among versions A–D and how these compared to their rankings from the experimental phase. Focus group discussions were transcribed, anonymized, and translated into English. All materials, including writing samples, experimental data, and transcripts of focus groups 1 and 2, are available from the project repository.

Results

After removing invalid responses, the final sample comprised 58 participants: 50% male, 43.1% female, and 6.9% identifying as other. The largest age group was participants in their thirties (37.9%), followed by those in their fifties (19%); the remaining age groups (twenties,

forties, and sixties) were broadly similar, each accounting for between 12% and 13% of the sample. Native Japanese speakers made up 86.2% of participants, with 12.1% being non-native speakers. In the open-ended section about literature background, 62% reported active engagement with Japanese fiction or creative writing, either as amateur or professional writers, or held undergraduate, postgraduate, or academic qualifications in Japanese literature or a related field. Full demographic details are shown in Figure 1.

Experiment

We begin by reviewing the scores produced by each respondent through the ranking of versions A-D. We transform this ordinal vector into a numeric vector by assigning a score of 10 to top-ranked stories, 7 to second best, 4 to third best, and 1 point for the version ranked worst. We then calculate the average score per version (A-D) by respondent and adjust for statistical significance using a t-test. Scores for A-D were calculated from the raw data of respondents (Figure 2). Table 1 shows that the average scores favored version D: Human-AI collaboration, with an average score of 145.0 (scores statistically significant at $p < .01$), followed by A: Original human-written text, which scored an average of 139.6; then C: Detailed prompt specification, which averaged 134.6, and finally B: Zero-shot prompt with 130.7 ($p < .05$). The t-test shows that the sample mean of 137.5 is significantly greater than the expected average score per respondent (Figures 2 and 3).

INSERT FIGURE 1 HERE

The results of the experiment show that version D: Human-AI collaboration is on average the most favorably perceived by the readers. 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, as discussed earlier in this study. Additionally, these findings are consistent

with previous work that found human-AI collaboration to improve human writing. One plausible explanation is that AI lacks the capability to assess textual quality and struggles to maintain coherence due to limitations in the computational resources available. Consequently, AI-generated fiction achieves the highest perceived quality when produced in collaboration with a human author responsible for editing, selecting, inputting, and evaluating the text. In other words, human-AI collaboration clearly improves the perceived quality of the text compared with fully AI-generated versions.

INSERT TABLE 1 HERE

Version A received the second-highest average score; however, its ratings show significant variation with some participants assigning it the highest scores observed in the study. The boxplot in Figure 2 shows a broad range of responses characterized by a large interquartile range and extended whiskers that indicate variability outside the upper and lower quartiles. Similarly, Figure 3 shows the distribution of scores spanning from very low to very high and highlighting the high variability and relatively high mean. This suggests that participants evaluated version A: Original human-written text in markedly different ways. Notably, while some readers perceived version A (the original) as being of exceptionally high quality, others rated it much lower. As such, genuine human authored text attracts both the strongest enthusiasts and the harshest critics.

INSERT FIGURE 2 HERE

Version C: Detailed prompt specification received the third-highest average score among versions A-D, with a narrower range of variation in the data comparable to version D, but also with a lower score range. The boxplot shown in Figure 2 indicates a relatively small range of scores, with a compact interquartile range and shorter whiskers similar to version D, albeit with a lower mean. The distribution scores shown in Figure 3 further corroborate this,

showing a bell-shaped curve centered on 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. In other words, the human-AI collaborative process in version D improves the perceived quality through human intervention, thereby explaining the differences in scores between versions C and D. When comparing version C to version B: Zero-shot prompt, version C exhibits a higher mean and a smaller range of variability (Figure 3). Indeed, AI-generated texts (versions C and B) often exhibited inconsistencies, suggesting that the perceived quality of AI-generated works can be improved by using detailed prompts. Nonetheless, inherent weaknesses associated with AI-generated text remain evident.

INSERT FIGURE 3 HERE

Version B: Zero-shot prompt received the lowest average score among all versions, with a wide range of variability concentrated in the lower scoring range. Figure 2 shows that version B has the lowest mean score, the longest whisker extending into the lower range, and a relatively large interquartile range, indicating broad variability skewed towards lower ratings. The distribution of scores for version B (Figure 3) reveals a pattern similar to that of version A: Original human-written text, but with greater weight in the lower score range. While both versions A and B exhibit wide variability in scoring, version B lacks the high scores attained by version A and has a greater proportion of lower ratings. This suggests that without structured prompts, AI-generated text is more likely to result in lower perceived quality, failing to reach the standard achieved through human-AI collaboration, the original text, or AI-generated work guided by detailed prompt engineering.

Modeling

Lastly, we conduct a mixed-effects analysis to establish whether the pairwise differences between versions are statistically significant. While the averages are clearly ranked, with D

(Human-AI collaboration) averaging 145.0, A (Original) 139.6, C (Detailed prompt) 134.6, and B (Zero-shot) 130.7, the total possible range in the scoring spans from a minimum of roughly 106 to a maximum of around 169 per respondent. The four conditions are therefore clustered within a 14.3-point band across this range.

We run a set of mixed models using the lme4 and ordinal packages for R and found that once the above condition was accounted for, there was no systematic variance attributable to individual participants rating differently from one another, nor to particular vignettes being systematically harder or easier to rank. All the variance sits in the residual, which is the condition-by-vignette interaction: each story responds differently to each condition, and that noise overwhelms the condition main effects. The likelihood ratio test is significant at $p = .022$ in the participant model, but only marginal at $p = .066$ in the vignette model. Critically, these tests are driven almost entirely by the B (Zero-shot prompt) contrast. Looking at the pairwise comparisons between versions, we find that D vs B is statistically significant, and that B vs A is marginally significant, but that D vs A and C vs A are not. This is because the between-condition variance in vignette means is .061, while the average within-condition variance across vignettes is higher at .643—indeed, roughly an order of magnitude higher.

As such, the likelihood ratio test comparing the full model to the null model is significant ($\chi^2 = 10.16$, df = 3, $p = .017$) but driven almost entirely by the B vs A contrast, as the three-way comparison among A, C, and D is largely affected by noise. The effect sizes, however, tell a nuanced story: D vs B gives Cohen's $d = .56$, which is statistically significant, but D vs A gives $d = .20$, and D vs C gives $d = .39$—both small, and neither significant. The model supports the conclusion that zero-shot AI generation (B) is reliably rated lower than human-AI collaboration (D), with a moderate effect size. But the assumption that human-AI collaboration is superior to original human writing is not supported by the model, even if the

mean score for human-AI collaboration is higher than that of the original human writing. This presents a narrower but more interesting empirical import: original human writing, prompt-engineered AI, and human-AI collaboration are statistically indistinguishable in their mean quality—as what actually differentiates them is variance, not central tendency.

The more striking finding in the data, as presented in Figures 2 and 3, is not the hierarchy of means but the difference in distributional shape. The original human-written text (A) shows very high variance—it attracts both the strongest enthusiasts and the harshest critics. The AI-generated and collaborated versions (B, C, D) show progressively tighter distributions around their means. Human-AI collaboration (D) is essentially the safest, most generic option: nobody loves it passionately, but very few readers reject it strongly either. This hints to the prospect of AI and human-AI collaboration producing consistent mediocrity, while human writing produces polarized responses. In other words, AI optimizes for acceptability and inoffensiveness rather than literary excellence—a point that is echoed by the observations of focus group participants about version D lacking originality.

Focus groups

The focus group sessions were led by a facilitator who is a native Japanese speaker with a background in Creative Writing in Japanese. Participants in the first sessions include S: a PhD student in French Literature working with Japanese fiction writing; K: an editor with several published novels, poems, Tanka, and Haiku; T: a professional writer in the gaming industry with a background in Creative Writing; and Z: a student at a vocational school for animation and illustration in Japan with N-1 fluency in Japanese primarily engaged in manga creation. The second session included participant L: a salesperson in Japan with N-1 fluency with a background in Creative Writing; Y: a former editor working for a Japanese publishing company with fiction writing experience; and M: an undergraduate student of English

Literature in Japan with extensive familiarity with Japanese literature. All participants in the first session were male except for the facilitator, whereas participants in the second session were evenly divided between males and females. Table 2 provides an overview of participants' responses to the task of identifying AI-generated content.

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' (Table 2). 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 literary forms.

INSERT TABLE 2 HERE

When discussing the factors that determine the quality of fiction, participants mentioned narrative flow, story transitions, dialogue, the expression of human emotion, and spontaneity. Some participants highlighted aspects such as conversational tone, proper nouns, and poetic expressions as particularly reflective of human emotion. Notably, several participants reported that, unlike their usual reading habits, they assessed the texts by focusing more on linguistic expressions and scene transitions 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 of this study show that human-AI collaboration (Version D) received the highest mean score (145), followed by the original human-authored text (Version A, 139.6), AI generation with detailed prompt specification (Version C, 134.6), and zero-shot AI generation (Version B, 130.7). However, these differences should be interpreted cautiously. The mixed-effects analysis reveals that condition accounts for only a small proportion of total variance, and pairwise comparisons indicate that the differences between Versions A, C, and D are not statistically distinguishable. The only contrast that reaches substantive magnitude is between Version D and Version B, suggesting that the reliable finding is not that human-AI collaboration necessarily outperforms human writing, but that zero-shot AI generation is rated lower than all other conditions. Notably, Version D also displays the lowest variance across participants, while Version A displays the highest—indicating that original human writing produces more polarized responses, attracting both stronger enthusiasm and stronger rejection than AI-assisted content. These findings suggest that Generative AI's primary effect on perceived quality is a reduction in variance rather than an elevation of the average quality.

Focus group discussions reinforced these findings, with participants describing D as well-structured and free from notable flaws, making it a 'safe' choice. Indeed, some participants noted that while D was generally polished and natural, it lacked distinctive originality. Participants attributed the human-like quality of D to its effective use of proper nouns and rewritten dialogue. Observations suggest that human modifications—such as refining dialogue and incorporating proper nouns—contributed to its perceived authenticity (see focus group transcripts in the supplementary materials). AI-generated texts were

criticized for excessive or unnatural dialogue, which lacked emotional depth and included formulaic expressions. Notably, nearly all of D's dialogues were rewritten in the human-AI collaborative process, further compounding this perception.

Discussion

This study examined reader perceptions of AI-generated and AI-collaborated fiction by comparing texts produced with varying degrees of AI involvement. Combining an experimental ranking task with focus group discussions, it investigated both readers' ability to distinguish AI-generated content and their quality assessments of each condition. The experimental results show that human-AI collaboration (Version D) received the highest mean score, followed by original human-authored text (Version A), detailed prompt specification (Version C), and zero-shot prompt (Version B). However, a mixed-effects analysis indicates that condition accounts for only a small proportion of total variance, and that the only robust contrast is between Version D and Version B. This suggests that zero-shot AI generation is reliably rated lower than content involving human authorship or editing, while the differences among the remaining conditions remain within the range of noise. These findings partially align with the broader literature: the relative weakness of zero-shot generation is consistent with research on AI's limitations in sustaining narrative coherence (Gholami et al., 2024), while the absence of a clear quality advantage for human-AI collaboration over original human writing complicates more optimistic accounts of AI's creative potential (Pretsch, 2023; Hutson and Schnellmann, 2023; Çelik, 2023).

The focus group findings largely corroborate and enrich the experimental results. Participants described Version D as polished and natural in tone, free of the contradictions and incoherence that marked the AI-generated versions, though they consistently noted its limited

originality. Version A elicited the most divided responses: some participants found the style and genre compelling, while others actively disliked it—a pattern that mirrors the high variance observed in the experimental scores. Versions B and C were both criticized for the limitations typical of AI-generated prose, including unnatural phrasing, emotional shallowness, and narrative inconsistency, with Version C rated somewhat better than Version B. These observations are consistent with Gholami et al.'s (2024) findings on AI's restricted memory capacity and its tendency to produce narrative incoherence in longer texts.

On the question of distinguishability, the two focus group sessions produced diverging responses: participants in the first session reported that they could not reliably identify AI-generated content, while participants in the second suggested that identification was possible but genre-dependent. Rather than contradicting each other, these responses point to a common underlying finding: distinguishability is not a stable capacity but varies with the stylistic conventions of the genre and the degree of human editing applied to the AI output. This contrasts with prior research on short-form poetry, where AI-generated content has been found to be particularly difficult to distinguish from human-authored work (Hitsuwari et al., 2023; Porter and Machery, 2024)—a difference likely attributable to the greater length and narrative complexity of prose fiction, which provides more opportunities for the limitations of AI generation to become apparent.

Our findings offer only partial support for the 'productivity gain' hypothesis previously reported in technical writing contexts (Noy and Zhang, 2023; Peng et al. 2023): while human-AI collaboration produces consistently acceptable prose with low variance, it does not demonstrably outperform original human writing in perceived quality, suggesting that the gains associated with AI assistance in professional and technical domains do not translate straightforwardly into literary fiction, where originality and stylistic distinctiveness

are central to quality judgments. More importantly, the potential gain noted in version D is primarily attributable to the human editorial layer rather than the AI's generative capacity alone.

One likely explanation for the relative success of Version D is that AI systems lack the metacognitive capacity to detect and correct their own failures—inconsistent phrasing, narrative incoherence, and formatting irregularities persist precisely because the model cannot evaluate its own output against literary standards. Human editorial intervention addresses this gap through selection, rewriting, and refinement. Focus group participants specifically noted that the restoration of proper nouns and the revision of dialogue were among the most effective editorial interventions, contributing to the perception of naturalness and human-like tone in Version D that avoid algorithm aversion effects. At the same time, the ceiling that human-AI collaboration appears to reach is telling: while it avoids the weaknesses of AI generation, it does not attain the distinctive qualities that the best human-written texts occasionally achieved. Participants identified subtle and implied emotional expression as a defining characteristic of original human writing—one that AI-generated and AI-collaborated content consistently failed to replicate, regardless of the degree of prompt specification or human editing applied.

Our findings offer partial support for the algorithm aversion hypothesis. Consistent with Köbis and Mossink (2021) and Proksch et al. (2024), focus group participants reacted negatively to texts they identified as AI-generated, citing unnatural phrasing, formulaic expressions, and emotional shallowness. However, the experimental results complicate this picture. Version D, despite containing AI-generated content, received the highest mean scores—suggesting that human editing effectively obscured its AI origins and neutralized the aversion response. This is consistent with Porter and Machery (2024), who found that readers

could not reliably identify AI-generated poetry and rated it more favorably than human-authored verse. Taken together, these findings suggest that algorithm aversion is not a stable disposition but a contextually contingent response: it is activated when AI authorship is detectable through textual markers and attenuated when human editing has smoothed those markers away.

Several focus group participants explicitly noted this difficulty. A related issue is that some Version A excerpts were drawn from early in their source narratives, which may have introduced apparent incoherence where scenes appeared to skip key plot developments; this could partly explain why some participants rated Version A as the least coherent despite it being the only fully human-authored condition. Future studies should address these limitations by using longer and more self-contained text samples, which would allow for a more ecologically valid assessment of how readers perceive AI involvement across the full arc of a narrative.

Conclusion

This study examined how Japanese readers with creative writing backgrounds perceive the quality and originality of fiction produced under four conditions: original human authorship, zero-shot AI generation, prompt-specified AI generation, and human-AI collaboration. The human-AI collaboration received the highest mean quality ratings in the experimental phase, and the focus group sessions shed considerable light on these results, particularly for Version A. Several participants acknowledged that their rankings reflected subjective taste, and some rated the original human-written text lowest simply because they disliked its style. While the texts hold strong reputations within their respective genres, their stylistic distinctiveness appealed to niche audiences and conflicted with some participants' preferences—a pattern that

explains the high variance observed in Version A's experimental scores. These shortcomings are offset by the more generic prose of AI systems, though often at the cost of uninspiring writing.

The central strength of original human-authored text was its capacity to convey subtle and culturally specific emotional subtext. Focus group participants consistently identified nuanced implied emotion—such as shifts in tone or culturally loaded word choices—as a defining characteristic of human writing and one that AI-generated content conspicuously lacked. Examples included using the term 'Jyanshi' to disdainful refer to a professional Mahjong player, whereas AI prose tended to be more positive. Participants noted that AI-generated text generally lacked such implicit emotional depth, rarely incorporating conflicting or implied commentary, and favoring instead idealized, conflict-free prose that participants found emotionally shallow. This finding is consistent with Pretsch (2023), Çelik (2023), and Hutson and Schnellmann (2023), though with an important qualification: the same qualities that make human writing emotionally rich also make it polarizing, as stylistic distinctiveness divides readers in ways that generic AI prose does not.

Version C's limitations centered on a tendency to prioritize prompt requirements over narrative flow, resulting in incoherence and, in some cases, unintentionally surreal content. Version B's weaknesses were more fundamental: without prompt refinement or human intervention, zero-shot generation produced contextual inconsistencies, structural breakdowns, poor transitions, and formatting errors that made AI authorship readily apparent and triggered strong negative reactions among participants commonly referred to as algorithm aversion. The contrast between B and C confirms that prompt engineering offers meaningful but incomplete mitigation of AI's literary limitations.

Taken together, the findings of this study support three main conclusions. First, the

more robust experimental finding is that zero-shot AI generation is reliably rated lower than all other conditions; the differences in mean quality scores between original human writing, prompt-specified generation, and human-AI collaboration are small and not statistically distinguishable. Second, the more meaningful difference between conditions lies in variance rather than means: human writing produces polarized responses while AI-assisted content produces consistently acceptable but undistinguished prose, suggesting that Generative AI's primary effect is anodyne prose expressed through a reduction in variance rather than a tangible elevation of quality. Third, the study identifies a dissociation between perceived quality and originality that previous work had not clearly articulated: human-AI collaboration was rated highest in consistency and polish yet lowest in originality, confirming that AI-assisted editing optimizes for acceptability at the expense of the stylistic distinctiveness that defines literary fiction. Finally, the study refines the concept of algorithm aversion, showing it to be contingent on the perceptibility of AI authorship through textual markers rather than a stable reader disposition—an effect that human editing can effectively neutralize.

Ultimately, the four-condition experimental design employed in this study—comparing zero-shot generation, prompt-specified generation, human-AI collaboration, and original human writing—is more granular than the binary designs and allows for isolating the distinct contributions of prompt engineering, human editorial intervention, and original human authorship. A key finding of this design is that collaborated text received the highest mean quality ratings yet the lowest originality ratings—a result that decouples two concepts previous studies had conflated. By combining quantitative ranking with qualitative focus groups, this study also refines the concept of algorithm aversion, showing it to be contingent on the perceptibility of AI authorship through textual markers rather than a stable reader disposition.

Acknowledgments

The corresponding author also acknowledges support from the University College Dublin and OBRSS scheme (grants R21650 and R20825) and the National Council for Scientific and Technological Development (grant 406504/2022-9).

References

Alter A and Harris EA (2023) Franzen, Grisham and Other Prominent Authors Sue OpenAI. *The New York Times*.

Asmelash L (2023) These books are being used to train AI. No one told the authors. *CNN*, 8 October 2023.

Bensinger G (2023) ChatGPT launches boom in AI-written e-books on Amazon. *Reuters*.

Bryman A (2015) *Social research methods*. Oxford: Oxford University Press.

Çelik MA (2023) Death of the author: A survey on artificial intelligence in literature. *İletişim Bilimi Araştırmaları Dergisi* 3(2): 142-154.

Chik H (2023) A Chinese professor used AI to write a science fiction novel. Then it was a winner in a national competition. *South China Morning Post*.

Clark L (2022) Towards "creativity amplification": Or, AI for writers, or beating the system. *Writing in Practice*.(7).

European Writers Council (2023) Analysis: The success of Generative AI in the book sector is based on theft. Reportno. Report Number|, Date. Place Published|: Institution|.

Fang X, Ng DTK, Leung JKL, et al. (2023) A systematic review of artificial intelligence technologies used for story writing. *Education and Information Technologies 2023 28:11* 28(11).

Gholami A, Yao Z, Kim S, et al. (2024) AI and memory wall. *IEEE Micro*.

Ha T-H (2024) Akutagawa Prize draws controversy after winning for work that used ChatGPT. *The Japan Times*.

Halperin BA and Rosner DK (2024) 'AI is Soulless': Hollywood Film Workers Strike and Emerging Perceptions of Generative Cinema. *ACM Transactions on Computer-Human Interaction*. DOI: 10.1145/3716135.

Hitsuwari J, Ueda Y, Yun W, et al. (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 and 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).

Kamome A (2022) 3 週間で 100 篇小説を書いて、AI を利用した小説で史上初めて星新一賞に入選した話. In: Ashizawa Kamome. Available at: https://note.com/ashizawakamome/n/n35bdb0486a6d#44ef08d4-5a86-4a43-b2a8-cf6e8c8687b6.

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

Kuwahara T (1946) 第二芸術 —現代俳句について—, 世界, 岩波書店. *世界* 1(1).

Lee H-K and Lee H-K (2022) Rethinking creativity: creative industries, AI and everyday creativity. *Media, Culture & Society* 44(3).

Mauran C (2023) Author finds AI books falsely written under her name for sale on Amazon, .

Mashable.

Michel L (2023) Will A.I. Change Art? A New Novel Uses AI to Explore Just That. *The New York Times*.

Milliot J (2023) Authors Guild issues contract clause changes to account for AI. *Publishers Weekly*.

Niloy AC, Akter S, Sultana N, et al. (2024) Is ChatGPT a menace for creative writing ability? An experiment. *Journal of Computer Assisted Learning* 40(2): 919-930.

Noy S and Zhang W (2023) Experimental evidence on the productivity effects of generative artificial intelligence. *Science* 381: 187-192.

Oka M and Hashimoto T (2024) AI 時代の質問力 プロンプトリテラシー 「問い」と「指示」が生成 AI の可能性を最大限に引き出す. 翔泳社. In: AI and the Art of Novel Writing: A Case Study in Creative AI Applications. Available at: https://www.linkedin.com/pulse/ai-art-novel-writing-case-study-creative-applications-pentzek-%E6%BD%98%E6%8B%A9%E7%A7%91-qfixc/.

Peng S, Kalliamvakou E, Cihon P, et al. (2023) The impact of AI on developer productivity: Evidence from GitHub Copilot. *arXiv preprint arXiv:2302.06590*.

Porter B and Machery E (2024) AI-generated poetry is indistinguishable from human-written poetry and is rated more favorably. *Scientific Reports* 14(1): 26133.

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, et al. (2024) The impact of text topic and assumed human vs. AI authorship on competence and quality assessment. *Frontiers in Artificial Intelligence* 7.

Quach K (2023) AI-written history of Maui wildfire becomes Amazon bestseller, fuels conspiracies. 18 August 2023.

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

Tapper J (2023) Authors shocked to find AI ripoffs of their books being sold on Amazon. *The Guardian*.

The Authors Guild (2023) More than 15,000 Authors Sign Authors Guild Letter Calling on AI Industry Leaders to Protect Writers. Reportno. Report Number|, Date. Place Published|: Institution|.

Vaswani A (2017) Attention is all you need. *Advances in neural information processing systems*.

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

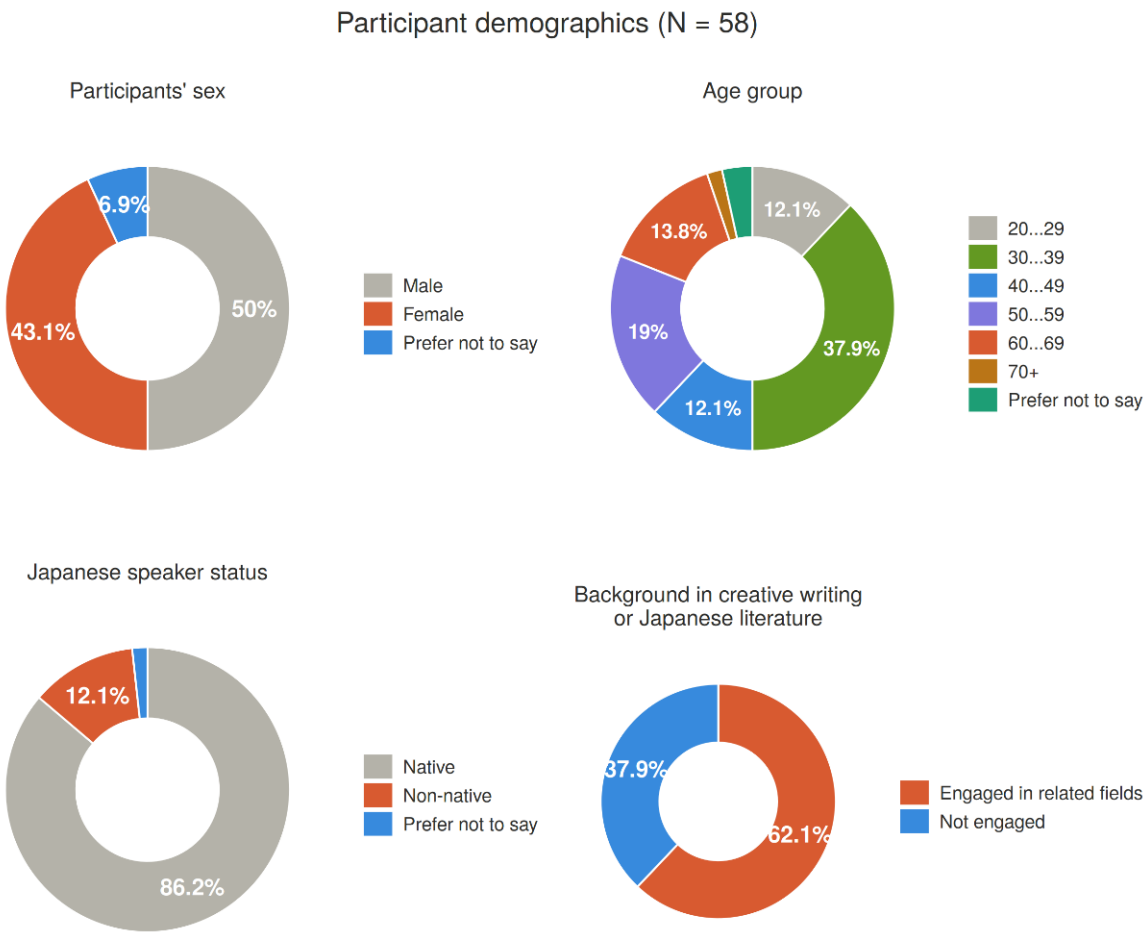

Figure 1: Demographic breakdown of participants

Table 1: Summary of average scores of A-D

Version	Mean	SD	Min	Q1	Median	Q3	Max
A: Original	139.9	32.8	81	115.0	133.0	161.0	211
B: Zero-shot	130.4**	29.2	69	111.0	130.0	151.0	203
C: Detailed prompt	134.7	22.7	76	121.0	133.0	148.0	187
D: Human-AI	145.0***	20.2	90	133.0	142.0	157.0	202

** p<.05 *** p<.01 (one-sample t-test vs. expected mean of 137.5)
N = 58. Scores = aggregate ranking points per participant over 25 vignettes (max = 175).

Table 2: Participants' responses about identifying AI-written literary work

Session 1	*'I could not distinguish AI-generated novels with any certainty. I looked at many articles in this experiment, and I was always like, I cannot distinguish them.'* (K)
	'The working favored by AI wording resembles human prose, so it took time to judge the quality, to see the flow or transition of the story.' (Z)
	'It was difficult to identify the AI text, but the articles with more dialogue were easier to flag as obviously ChatGPT written.' (S)
	'Light novels were relatively difficult to distinguish because AI can create them. This type of fiction was difficult to judge.' (S)
Session 2	*'I felt I could tell which ones were AI-written about half the time. There were cases where the beginning and the end were clearly not connected, but some of them seemed quite natural.'* (L)
	'It was easy to identify the original one with a distinctive writing style in some parts. For example, it is like, let us see... the one with a loud writing style. In such cases, it was very easy to know which one was the original second half… I also thought it was AI when the sentence expressed something very cheesy… like just saying typical positive things. Those were very much like AI.' (M)

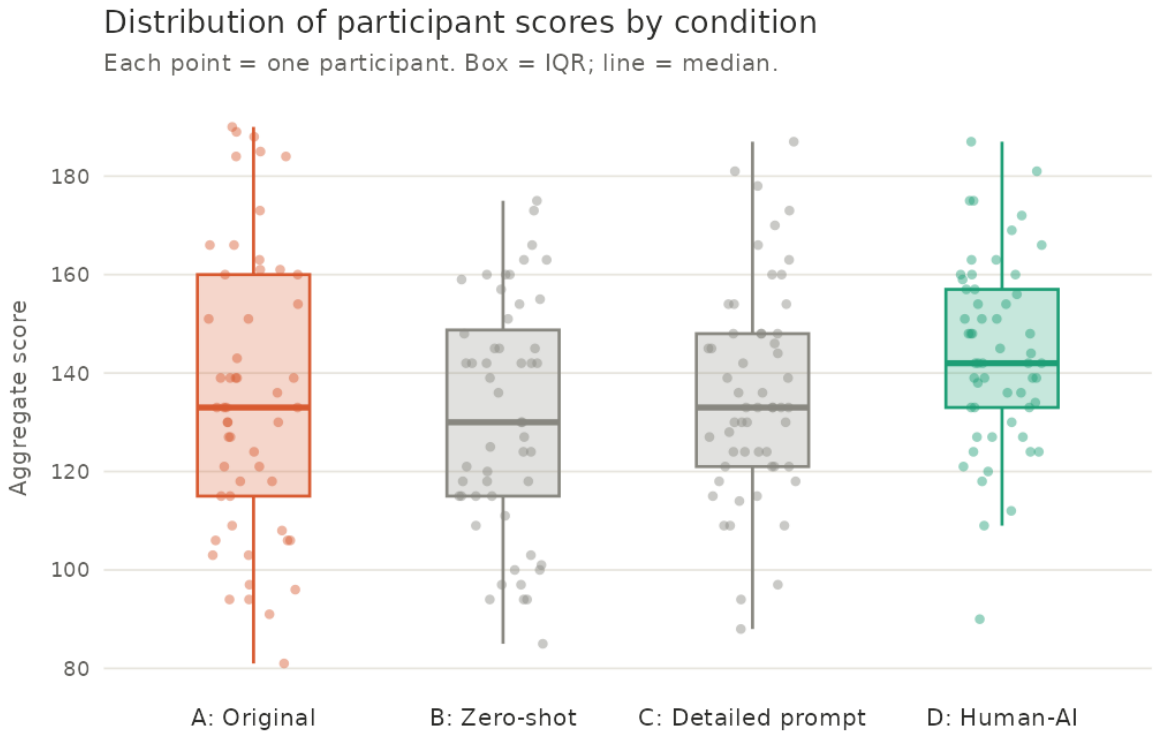

Figure 2: Boxplot of scores assigned by participants for each written output

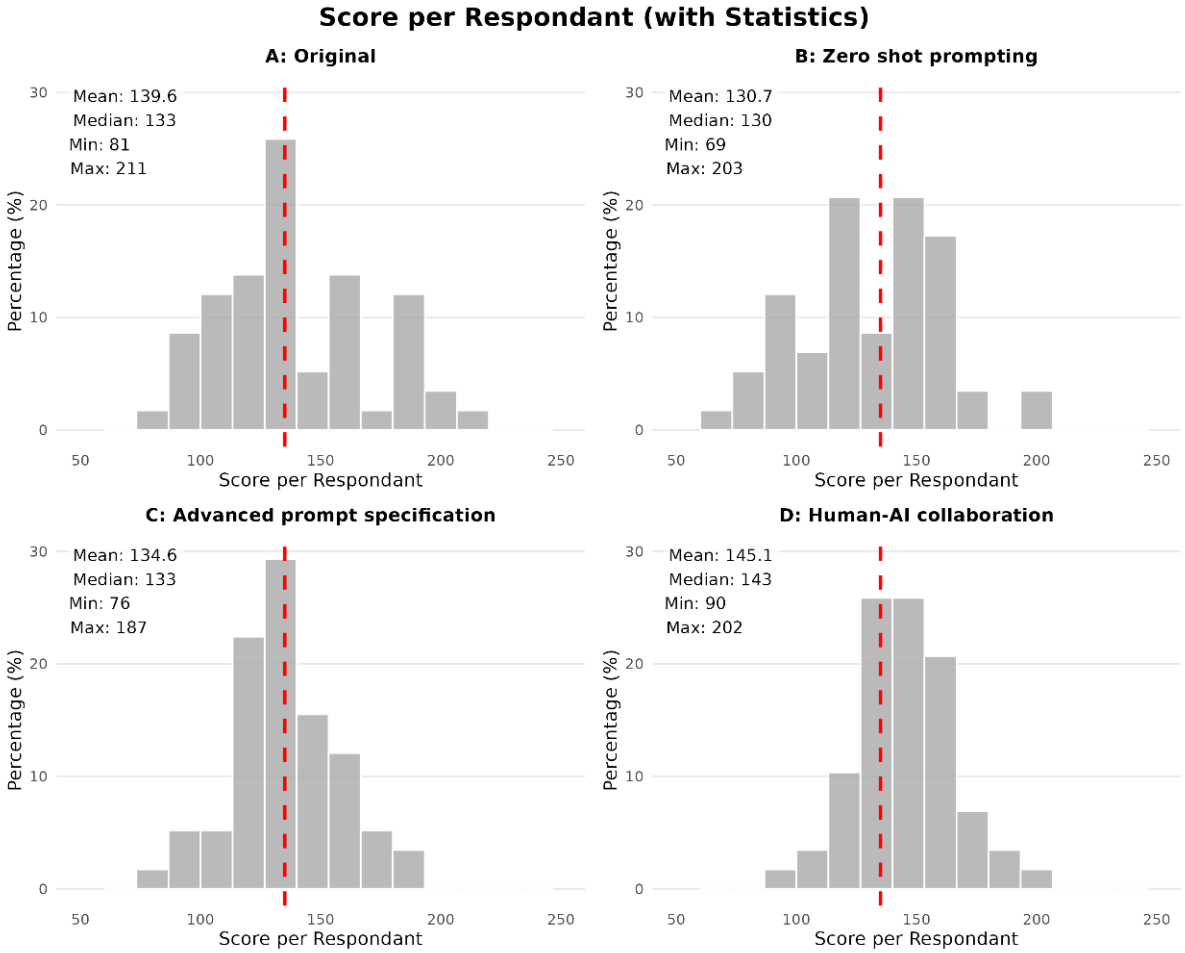

Figure 3: Distribution of scores for each written output