

EMPIRICAL ARTICLE

Bot or not: Can people tell the difference between stories written by a human or by an AI system?

Sydney Sears and Deena Skolnick Weisberg 

Psychological and Brain Sciences, Villanova University, USA

Corresponding author: Deena Skolnick Weisberg; Email: deena.weisberg@villanova.edu

Received: 4 February 2026 UTC; **Revised:** 4 May 2026 UTC; **Accepted:** 8 June 2026 UTC

Keywords: artificial intelligence; Turing test; fictional stories; creative writing

Abstract

How do people perceive artificial intelligence (AI)-generated fictional stories? In Study 1, participants (1,682 adults recruited from Prolific) read either a human-written or AI-generated story. Half were told that their story was written by a human, and half were told that their story was written by ChatGPT. Participants rated the AI-generated story as higher quality and more absorbing, but they rated the story better on both measures when they were told it was written by a human. In Studies 2 and 3, participants (905 adults recruited from Prolific) read both types of story and guessed each story's origins. Participants were no better than chance at differentiating between the human-written and AI-generated stories. Self-reported expertise with AI, but not with fictional literature, was positively correlated with correct story identification. AI programs thus can generate creative work that is viewed as being at least as good as human-generated work but are not perceived as being able to do so.

1. Introduction

Artificial intelligence (AI) is changing the way work is being done in many fields. While AI programs can be beneficial in many ways, some critics have argued that using AI to replace human work can be harmful (e.g., Eloundou et al., 2024). This was one of the major issues in the 2023 strike by the Writers Guild of America, as writers feared that their work would be devalued if it could simply be replaced by AI-generated scripts. But are people even able to recognize the difference between AI-authored and human-authored creative works? And how do people's perceptions of AI's capabilities affect these judgments? The current studies aim to answer those questions, focusing specifically on AI's abilities to generate creative work (see Beguš, 2024; Shidiq, 2023), as this is a domain where human experience is perceived to be particularly important to creating high-quality outputs.

1.1 Failing the Turing test

The Turing test was initially proposed by Alan Turing as a partial metric for whether computer systems are truly intelligent. It tests whether the responses of such systems are distinguishable from human responses in a blind comparison. AI systems are already able to pass such tests (e.g., Mei et al., 2024) and have been able to do so for a while. Indeed, a 1972 study coded an AI system to imitate a patient with schizophrenia, and a panel of judges correctly identified which transcript was machine-generated only 48% of the time (Colby et al., 1972).

© The Author(s), 2026. Published by Cambridge University Press on behalf of Society for Judgment and Decision Making and European Association for Decision Making. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial licence (<https://creativecommons.org/licenses/by-nc/4.0>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original article is properly cited. The written permission of Cambridge University Press or the rights holder(s) must be obtained prior to any commercial use.

Several recent studies have examined the differences in moral decision-making between AI and humans (Aharoni et al., 2024; Dillion et al., 2025; Osborne and Bailey, 2025). For example, participants were asked to rate advice given by ChatGPT or a human ethicist and then to determine if a human or a computer was the one to give the advice. While participants performed significantly better than chance at differentiating between the human-given and AI-given advice (Aharoni et al., 2024), these studies found that participants rated the quality of the AI's moral reasoning as superior to humans' moral reasoning on dimensions such as virtuousness, intelligence, and trustworthiness. The same is true for people's judgments of AI-generated content in a variety of other domains, including student essays (Herbold et al., 2023), advertising campaigns (Zhang and Gosline, 2023), and Instagram posts (Park et al., 2024).

There are several possible reasons why AI-generated content is perceived as indistinguishable from—or, frequently, better than—human-written content (see review in Coutanche and Weisberg, 2026). For instance, AI-generated content is often more fluent and easier to parse than human-written content (Markowitz, 2024), and people have a well-known tendency to prefer material with those characteristics (e.g., Reber et al., 2004). Additionally, AI-generated content tends to have a more positive emotional tone (Muñoz-Ortiz et al., 2024), which people also perceive more positively.

Investigating fictional stories can provide a different perspective on these issues, since the features that characterize high-quality creative writing (e.g., character complexity) are somewhat different than the features that characterize high-quality expository writing (e.g., soundness of arguments). Crucially, and in contrast to expository writing, a degree of obscurity around a story's central thesis can sometimes be desirable and is often seen as a characteristic of literary (rather than popular) fiction (Castano et al., 2023; Zunshine, 2022). AI-generated creative writing thus might not be subject to the same preference structure as other forms of AI-generated writing.

Contrary to this prediction, however, recent work has also found a preference for AI-generated creative writing. One study paired poems by famous authors (e.g., Shakespeare, Plath) with poems generated by ChatGPT 3.5 in the style of these authors. Participants in this study were slightly worse than chance at distinguishing human-written from AI-generated poems; specifically, human-written poems were slightly more likely to be characterized as AI (Porter and Machery, 2024). A similar study also asked participants to differentiate human-written from AI-generated poems, finding that participants were able to do so only when the authors paired the human poem with a randomly generated poem from ChatGPT 2.0; when the authors selected the best AI-generated poem from a group of poems generated by ChatGPT 2.0, participants were no longer reliably able to choose which was written by a human (Köbis and Mossink, 2021). More generally, AI systems outperform humans on standard creativity tasks, like the alternative uses task (Wenger and Kenett, 2026).

These studies suggest that AI systems like ChatGPT are already able to generate human-like creative work, at least in the domain of poetry. The current studies extend this work to a different creative domain—fictional stories—hypothesizing that participants will again be unable to tell which of two stories was generated by an AI (see Orwig et al., 2024).

1.2 *Perceptions of AI*

Although people cannot always tell the difference between AI-generated and human-created creative works, they nevertheless tend to believe that AI-generated works are worse than human-created ones ('algorithm aversion'; see Dietvorst et al., 2015). For example, a study on AI-generated advice found that participants tended to rate the advice more poorly when they knew it was AI-generated (Osborne and Bailey, 2025). Similarly, in one of the studies on poetry cited above (Porter and Machery, 2024), poems received lower rating on a variety of qualitative metrics (e.g., how beautiful the poem was; how well it conveyed a specific emotion) when participants were told that the poems were generated by an AI system—regardless of the actual author of these poems. In a related study (Ragot et al., 2020), participants were told before critiquing a set of artworks whether the pieces were AI-generated

or human-created. Paintings described as being drawn by humans were evaluated significantly more highly than those described as being made by AI (see also Hong and Curran, 2019). Similar effects have been found for fictional stories, the topic of the current study. When participants were told that an AI program had written the story that they read, participants reported less narrative transportation into the story (Messingschlager and Appel, 2022). Conversely, if the machine that generated the creative work was more anthropomorphized (i.e., presented as more human-like), participants had fewer negative prejudices against it (Chamberlain et al., 2018). And a recent highly powered study found that people's aversion to AI-generated creative work was so 'sticky' that the authors' attempts to mitigate this bias (e.g., by providing participants with information about how impressive AI's capacities could be) failed to improve people's ratings of various kinds of writing created by ChatGPT (Raj et al., 2026).

Perceptions of AI and its capabilities may thus matter to people's abilities to tell the difference between human-written and AI-generated works. Specifically, although people judge AI-generated output positively, as reviewed above, they generally have negative attitudes toward AI-generated work and do not believe that AI systems can create human-level work. For the current study, we predict that participants will rate AI-generated stories as worse than human-generated stories, regardless of those stories' true origin.

1.3. *Role of expertise*

The participants in the studies described above did not have particular expertise in either AI or the artistic domain they were asked to judge (e.g., poetry). But expertise could affect people's judgments. Typically, the more experience or expertise an individual has in a specific area, the more general success they have in that area (see Ericsson and Lehmann, 1996, for review). Experts tend to perceive situations related to their domain of expertise differently from novices, leading to improved decision-making.

In the domain of art in particular, art experts reported that a set of images was significantly less confusing and more interesting than novices (Silvia, 2013). Similarly, one study that examined how experts and students in English literature courses described a complex narrative found differences in these groups' response to this prompt: Students tended to paraphrase the text, while literary experts tended to comment on the language used in the text, the dialogue's structure and function, and the relationship between the author and reader (Graves and Frederiksen, 1991). This work illustrates that experts and nonexperts may perceive the same task differently, leading to differences in performance. With respect to the current studies, we aimed to investigate whether expertise either with AI systems or with fictional stories might affect participants' abilities to discriminate between AI-generated and human-written texts.

1.4. *The present study*

The present study examines whether people can differentiate between human-written and AI-generated short fictional stories, their judgments about them, and whether having prior experience with these subject areas has any effects. In Study 1, we manipulate both which type of story participants receive and what they are told about it, so that half of our participants are told the truth about their story's author (human or AI) and the other half are falsely told that it was written by the other type of author. After reading their assigned story, participants are asked a series of questions about the story's quality and about how absorbed they were in the story. To begin to examine the role of people's beliefs about AI, we also ask about their familiarity with AI programs and their perception of these programs.

In Studies 2 and 3, all participants receive one human-written story and one AI-generated story, but they are not told the origins of the stories. After reading their assigned story pair, participants are asked to determine which story is which. They are also asked a series of questions about their prior experiences with and expertise in both AI software and fictional literature, to test whether either form of expertise affects their judgments.

Table 1. Number of participants per condition.

	Told story was written by a person	Told story was written by ChatGPT
Read human-written story	417 participants	416 participants
Read AI-generated story	423 participants	426 participants

2. Study 1

2.1. Methods

2.1.1. Participants

The final sample in this study consisted of 1,682 adult participants recruited from Prolific (843 female, 809 male, 26 nonbinary/third gender, 2 other, and 2 preferred not to answer; mean age = 45.9 years; age range = 18–81 years). This sample was recruited to be representative of the population of the United States on the characteristics of gender, age, and race. Participants were paid \$3 for a 15-minute study. Table 1 provides details about the number of participants in the final sample assigned to each condition.

An additional 242 participants were recruited but removed from the final sample for not consenting to participate ($n = 4$), failing to complete the study in its entirety ($n = 110$), failing at least two out of the three memory questions about their assigned story ($n = 52$), or providing incorrect responses to any of the attention checks (details below; $n = 76$).

2.1.2. Materials

This study was not preregistered. All study materials and measures are available on the Open Science Framework (<https://osf.io/d8f6m/>).

Stories. We used three fictional short stories written by human authors, which were published in a reputable literary journal or short story collection. We then used ChatGPT to generate three corresponding fictional short stories. All stories had word counts of approximately 1,000 words (equating to approximately a five-minute read). This length was long enough for participants to have a sense of the writing style and be potentially absorbed by the story and also short enough for participants to be fully engaged throughout the whole story.

The first human-written story was titled ‘FISH’. It was published in *Longleaf Review* and was written by Emilie Fox in 2021. For the corresponding AI-generated story, ChatGPT 4.0 was prompted with the following: ‘Write a ~1000-word short story in a style that would be published in a literary magazine that is about generational life and death and uncertainty and uses koi fish as a symbol written from the point of view of the grown child of a woman’. ChatGPT gave this story the title ‘Reflections in Still Water’. The story was generated on November 15, 2023, at approximately 9:37 p.m. In the final sample, 385 participants received ‘FISH’ and 374 participants received ‘Reflections in Still Water’.

The second human-written story was titled ‘High Heels’. It was one story published as a part of a short story collection titled *The Short Hello* by Susie Maguire, published in 2000. For the corresponding AI-generated story, ChatGPT 4.0 was prompted with the following: ‘Write a ~1,000-word short story in a style that would be published in a short story collection that is about a man observing a woman dancing naked that emphasizes how both feel weak when looking at themselves but make the other feel stronger and loved, written from a third person point of view’. ChatGPT gave this story the title ‘The Dance of Being’. This story was generated on December 2, 2024, at approximately 2:35 p.m. In the final sample, 221 participants received ‘High Heels’ and 230 participants received ‘The Dance of Being’.

The third human-written story was titled ‘Inisfree’. It was one story published as a part of a short story collection titled *Whipping the Cat: Tales from Lough Neagh’s Banks* by Patrick Smyth, published in 1993. For the corresponding AI-generated story, ChatGPT 4.0 was prompted with the following: ‘Write a ~1,000-word short story in a style that would be published in a short story collection that takes place in the late 1910s to early 1920s and is about a man who questions his purpose as a government

worker during an air raid strike, ending with him leaving for the countryside with his family to find peace, written from a third person point of view, using British language'. ChatGPT gave this story the title 'The Weight of the Sky'. This story was generated on December 4, 2024, at approximately 4:00 p.m. In the final sample, 227 participants received 'Inisfree' and 245 participants received 'The Weight of the Sky'.

Story Absorption Scale. To determine how absorbed and engaged participants were in the story, we used an edited version of the Story World Absorption Scale (Kuijpers et al., 2014). The original scale is 18 items long and covers different aspects of story absorption, including attention, transportation, emotional engagement, and mental imagery. Since the original scale was designed for longer pieces of fiction, we used 12 of the items, since these were the ones that fit best with short stories. For example, 'When I was reading the story, I was focused on what happened in the story' and 'It was difficult to connect to the main character in the story' (reverse-coded). Participants ranked these statements on a scale from 'strongly disagree' (−3) to 'strongly agree' (3). Since none of the items from the original scale were reverse-coded (higher rankings always indicated higher absorption), we changed the wording of three of the items so that they would be reverse-coded. Cronbach's alpha for this scale was 0.91.

We included an attention check in this set of statements, which read, 'Please select "strongly disagree" if you are reading this'. Participants who failed this attention check were excluded from analysis.

Perceived Story Quality Scale. We developed 10 statements asking about the story's plot (e.g., 'The plot of the story was boring to me', reverse-coded), characters (e.g., 'The character(s) in the story had depth'), theme (e.g., 'The story had a theme that was difficult to discern', reverse-coded), and style (e.g., 'The story effectively employed literary devices (such as metaphor, symbolism, etc.)'). Three of these items were reverse-coded. Participants rated these statements on a scale from 'strongly disagree' (−3) to 'strongly agree' (3). As with the absorption scale, we included an attention check statement, which read, 'Please select "strongly agree" if you are reading this'. Participants who failed this attention check were excluded from analysis. Cronbach's alpha for this scale was also 0.91.

Memory questions. We created three multiple-choice memory questions for each of the stories as a further check to ensure that participants had fully read and comprehended their assigned story. These memory questions asked about events or themes mentioned in each story. For example, a question given to participants who read the story 'FISH' was 'What was the "incident" that the narrator's mother had?' with the answer choices of a) 'She sees hallucinations and fights to get out of bed by herself' (correct answer), b) 'She accidentally wets the bed', or c) 'She yells obscenities and punches a nurse'. Participants who answered two or three of these memory questions incorrectly were excluded from analysis.

Deception question. After the main dependent measures, we confronted participants with the possibility that what they had originally been told about the origin of the story they read (human-written or ChatGPT-written) was false and that its origin was potentially the opposite of what they had been told. Specifically, we told participants, 'At the beginning of this survey, we told you that the story was written [by a human/by an AI program, ChatGPT]. But that may not have been true. For this study, there were two possible stories you could have been given: a story written by a human, or a story generated by ChatGPT. You were then randomly either told the truth about who wrote your story (that is, you were told it was written [by a human/by ChatGPT] and it actually was) or you were lied to about who wrote your story (that is, you were told it was written [by a human/by ChatGPT] when it was actually written [by ChatGPT/by a human])'.

Participants were then asked, 'With this knowledge, do you think that what we told you about your story's author ([human/ChatGPT]) is true?' They responded to this question on a 5-point Likert scale, with answers ranging from 'It was definitely written by the other kind of author' (coded as −2) to 'I was definitely told the truth' (coded as 2), with 'I'm not sure' as the midpoint (coded as 0).

Attitudes toward AI. To measure participants' perceptions of AI systems, we created a set of four questions asking about participants' familiarity with AI programs (e.g., 'Have you used ChatGPT, or

another type of AI program, before?') and about their feelings about these programs (e.g., 'Do you feel that AI programs like ChatGPT are beneficial technologies?'). These questions were all set up in a 3-point Likert scale format, with the answer choices being 'Yes', 'Maybe', and 'No' (coded 1, 0.5, and 0, respectively). This set of items additionally included an attention check, asking participants to select 'yes'; participants who selected any other answer option were excluded from analysis.

2.1.3. Procedure

We used a Qualtrics survey to administer the study. Participants first read one of two possible introductions, telling them that they were about to read a story written by a contemporary author or by an AI program, depending on their condition. They next read one of the six short stories. The 'next' button on the page with the story text was programmed to only appear after one minute to ensure that participants spent at least that long on the page.

After reading, participants first completed the Story Absorption Scale and the Perceived Story Quality Scale (order randomized between participants). The items on each of these scales (including the attention checks) were presented in a random order for each participant.

Next, participants were asked the memory questions about their story; these questions and their answer choices were presented in a random order.

Then, participants¹ received a manipulation check question: 'At the beginning of the survey, what did we tell you about the author of the story that you read?' They were asked to choose if the story had a human or AI author. Participants were also asked how confident they were about their answer to this question on a 3-point scale: 'very confident', 'somewhat confident', and 'not at all confident'. We did not use this item as an exclusion criterion since it was presented after the treatment, but we note that 94 participants responded incorrectly. Our results do not change with these participants excluded.

After this, participants responded to the question about deception. Finally, they responded to questions about their perceptions of AI, with the order of the statements on this scale being randomized for each participant.

Participants then reported their age, gender, race, current job, and highest level of education.

At the end of the procedure, participants were debriefed and told the true origin of the story that they had read.

2.2. Results

All data and analysis scripts are available on the Open Science Framework (<https://osf.io/d8f6m/>).

2.2.1. Story absorption

We used a mixed-effects model to compare the mean story absorption scores by condition, with story condition and introduction condition as fixed effects (see [Figure 1](#)). To account for variance attributable to the specific story that a participant read, we included story as a random intercept. We found a significant effect of story condition, whereby participants who read an AI-generated story ($M = 1.42$, standard deviation [SD] = 0.99) rated their story as more absorbing than participants who read a human-written story ($M = 1.00$, SD = 1.14; $b = -0.43$, $p = .05$). We also found a significant effect of introduction condition, whereby participants who were told that they were reading a human-written story ($M = 1.32$, SD = 1.07) rated their story as more absorbing than participants who were told that they were reading an AI-generated story ($M = 1.10$, SD = 1.10; $b = 0.23$, $p = .001$). The interaction effect was not statistically significant, and the random intercept for story accounted for a small proportion of the total variance (0.034). See [Supplementary Material](#) for detailed analyses of responses by story.

¹This question was added after data collection had already begun, so we do not have data for this question from a subset of our participants ($n = 306$).

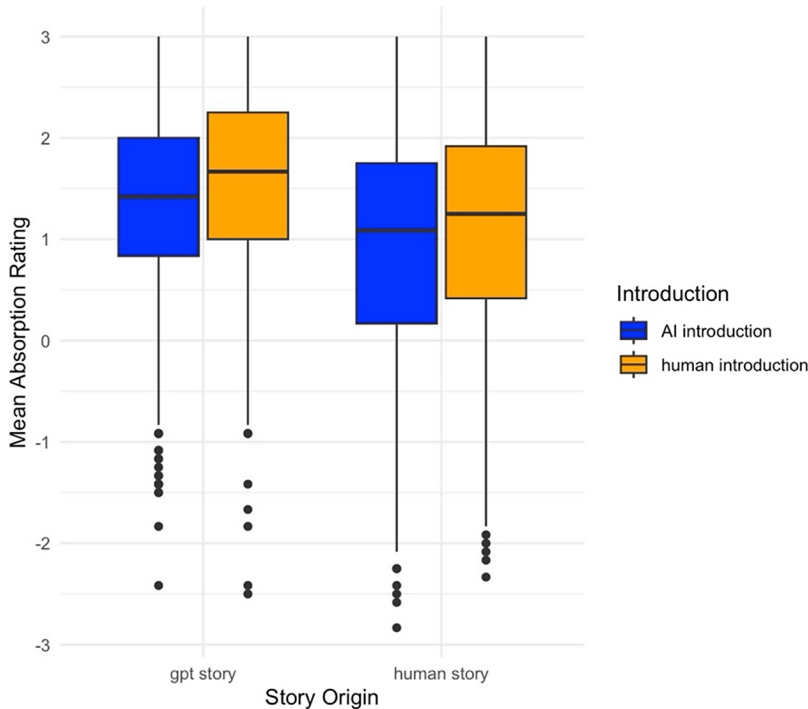


Figure 1. Mean story absorption ratings by condition.

2.2.2. Perceived story quality

We again used a mixed-effects model with story as a random intercept to compare the mean perceived story quality scores by condition (see Figure 2). Results aligned with our findings for absorption. Again, we found a main effect of story condition, whereby participants who read an AI-generated story ($M = 1.54$, $SD = 1.00$) rated their story more highly than participants who read a human-written story ($M = 0.97$, $SD = 1.23$; $b = -0.62$, $p = .01$). We also found a main effect of introduction condition, whereby participants who were told that they were reading a human-written story ($M = 1.40$, $SD = 1.11$) rated their story more highly than participants who were told that they were reading an AI-generated story ($M = 1.12$, $SD = 1.19$; $b = 0.22$, $p = .003$). The interaction effect was not statistically significant. Again, the random intercept for story accounted for a small proportion of the total variance (0.029), and detailed analyses of the different responses for each story can be found in the [Supplementary Material](#).

2.2.3. Effect of attitudes toward AI

We next examined the impact of participants' attitudes toward AI on their ratings of their assigned story's absorption and quality, taking into account their condition assignment. To do so, we used two analyses of covariance (ANCOVAs), one for each dependent measure.

For story absorption, we again found significant main effects of the story's actual author ($F(1,1674) = 75.22$, $p < .001$, $h^2 = .03$) and the story's introduction ($F(1,1674) = 19.70$, $p < .001$, $h^2 = .01$). There was also a main effect of participants' AI attitudes, whereby participants with more positive attitudes toward AI gave higher absorption ratings ($F(1,1674) = 117.94$, $p < .001$, $h^2 = .06$). Interestingly, there was a significant interaction between the story introduction and AI attitudes: Participants with more positive attitudes toward AI gave even higher ratings if they were told the story was written by ChatGPT ($F(1,1674) = 13.86$, $p < .001$, $h^2 = .007$; see Figure 3). No other interactions were significant.

The same results were obtained for perceived story quality (main effect of story author: $F(1,1674) = 120.57$, $p < .001$, $h^2 = .06$; main effect of story introduction: $F(1,1674) = 29.21$, $p < .001$, $h^2 = .01$;

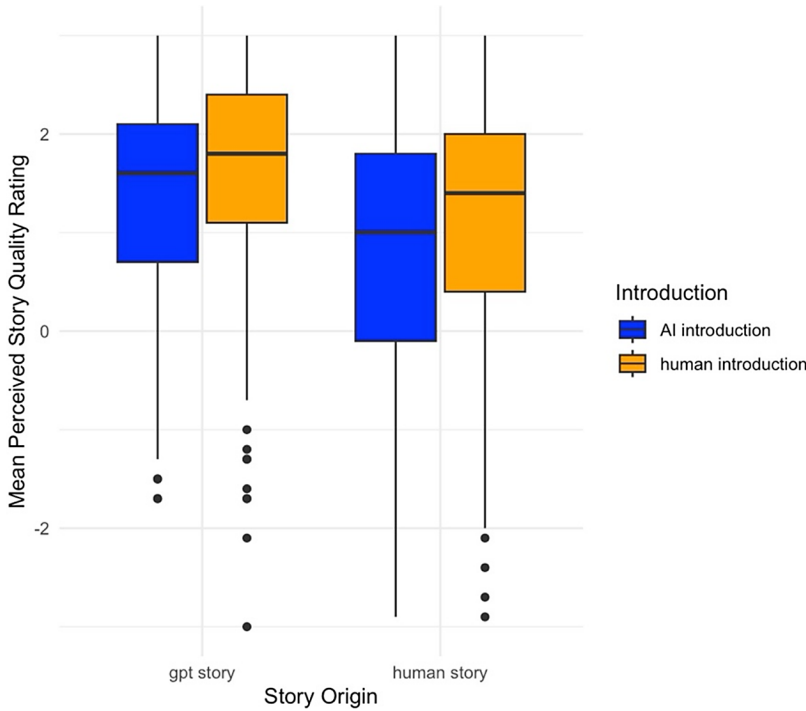


Figure 2. Mean perceived story quality ratings by condition.

main effect of AI attitudes: $F(1,1674) = 164.12, p < .001, h^2 = .08$), including the significant interaction effect between the story's introduction and participants' AI attitudes ($F(1,1674) = 13.80, p < .001, h^2 = .007$). Again, no other interactions were significant.

2.2.4. Reactions to deception

We wanted to determine whether participants would be more likely to doubt what we told them about their story's origin in the introduction depending on their condition. A 2×2 analysis of variance (ANOVA) on participants' level of confidence in their story's author revealed no main effects of introduction condition, $F(1,1678) = 0.065, p = 0.80$, nor of which story they actually read ($F(1,1678) = 0.083, p = 0.77$). There was a significant interaction ($F(1,1678) = 7.52, p = .006, h^2 = .004$), whereby participants who were told the truth about their stories (e.g., read a ChatGPT-generated story and were told that it was written by ChatGPT) were less likely to believe what we had originally told them than participants who had been deceived about the origin of their stories (e.g., read a ChatGPT-generated story and were told that it was written by a human). However, it is worth noting that all scores on this measure were close to 0, indicating that participants were generally unsure about the author of their story now that we had told them that they might have been deceived.

2.3. Discussion

Can large language models like ChatGPT generate human-quality fictional stories? Our participants seem to think that it can: They rated a story generated by ChatGPT as higher quality and more absorbing than a human-written story. This aligns with prior work showing that participants are generally unable to distinguish the outputs of AI systems from human outputs (e.g., Hong and Curran, 2019; Köbis and Mossink, 2021).

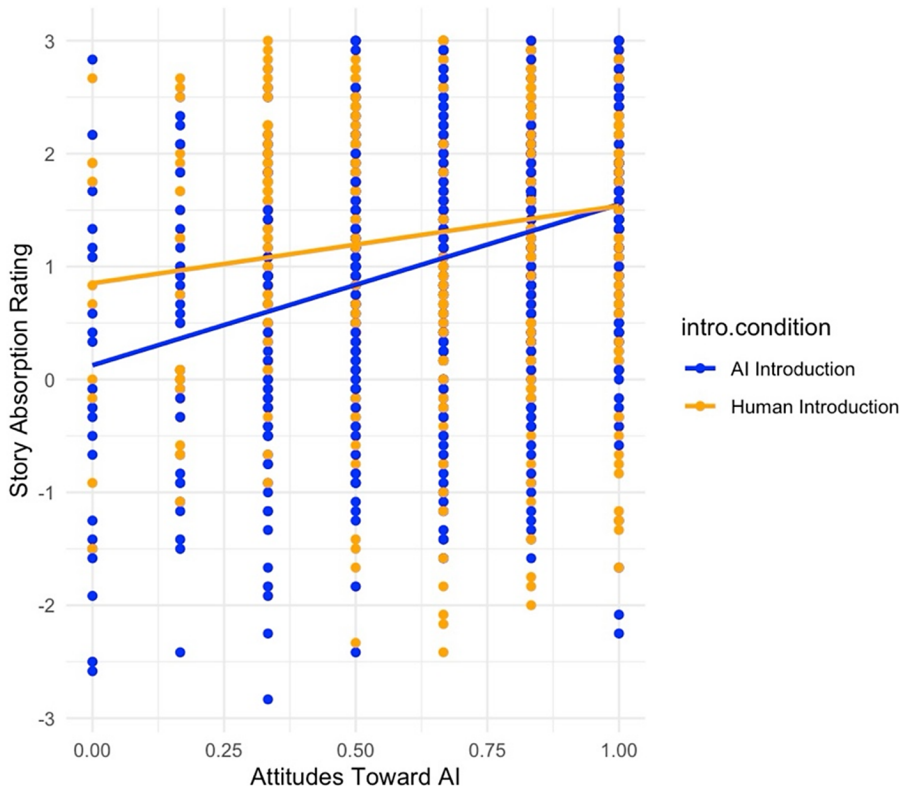


Figure 3. Interaction between attitudes toward AI and reported author on story absorption ratings.

Also, in line with prior work, our participants displayed a bias against AI, rating stories more favorably if they had been told they were written by a human. These findings correspond nicely to those of Porter and Machery (2024), who used the same 2×2 design in their second study and who also found that ChatGPT-generated creative work (poetry in their case) was rated higher than human-generated work, but especially when it was (falsely) described as being written by a human.

Although we obtained virtually identical results with the absorption scale as with the perceived quality scale, it is important to note that these constructs can (and perhaps should) come apart. One of the features of high-quality literature is that it is often difficult to absorb; literary fiction aims to challenge readers' abilities to find simple meanings in the text, while genre fiction is often easier to read and presents more predictable events and characters (see Castano et al., 2023). The converse is also true: A highly absorbing story may be very low-quality, like much of online clickbait. So, while our measure of people's perceptions of quality may have aligned with their level of absorption, different measures may yield different results. This point is particularly important as our perceived quality measure was not previously validated, although it does have some face validity.

Additionally, responses may vary depending on an individual's level of expertise with literary fiction, or with AI systems, as an issue that we address more directly in Study 2. It is important to know whether these findings also obtain for people who know more about how AI systems work, about the kind of output that these systems tend to generate, or about the structure of fictional stories and are potentially more likely to be able to discriminate between the two types of stories. Indeed, professional writers have been shown to rate AI-generated stories as significantly less creative than stories written by humans (Chakrabarty et al., 2023). And results from Study 1 suggest that participants' attitudes may affect their perceptions of AI-generated stories. Study 2 addresses this issue directly, asking participants for self-reports of their expertise and experience both with AI systems and with literature.

Further, in Study 1, participants received only one story and were asked to judge it. While we believe that this makes the task more ecologically valid, since people in the real world will most likely be reacting to individual stories, it is still important to know whether people can tell the difference between human-written and AI-generated stories when asked directly. It is possible that some features of these stories may seem more obviously AI-generated when put in contrast with a human-written story. Therefore, Study 2 uses this design. Examining participants' responses to this forced-choice task allows us to determine whether people can recognize AI-generated stories when directly contrasted with their human counterparts.

We acknowledge at the outset that one drawback of this design is that it lacks some external validity, in that our participants were told that one of their stories was written by a human and the other was generated by AI. When encountering content outside of a laboratory setting, these base rates are not made explicit—and indeed may be constantly fluctuating as the proportion of human-written to AI-generated work continues to change. Our forced-choice task may thus be easier than the task of spotting AI-generated work in the real world, but this could be seen as a feature rather than a bug, since this design gives our participants a better chance at performing well than they may have under more naturalistic circumstances.

3. Study 2

3.1. Method

This study was not preregistered.

3.1.1. Participants

The final sample consisted of 424 adults recruited from Prolific (265 female, 146 male, 10 nonbinary/third gender, 1 preferred not to answer, and 2 left the option blank; mean age = 38.5 years; age range = 18–77 years). As in Study 1, Prolific recruited this sample to be demographically representative of the United States with respect to age, gender, and race. Participants were paid \$4.

An additional 58 participants were recruited but removed from the final sample for failing to complete the study in its entirety ($n = 32$) or failing one or more memory questions from each story ($n = 26$).

3.1.2. Materials

Stories. The same six stories used in Study 1 were used for this study. However, unlike in Study 1, each participant received two stories (one human-written story, one corresponding ChatGPT-generated story). In this sample, 142 received the pair of 'FISH' 'Reflections in Still Water', 142 received the pair of 'High Heels' and 'The Dance of Being', and 140 received the pair of 'Inisfree' and 'The Weight of the Sky'.

Story origin question. Our main dependent measure in this study was whether participants could tell which story was generated by ChatGPT and which story was human-written. Each participant received one of two versions of this question. They were asked either which story they believed was human-written or which story they believed was ChatGPT-generated.

They were then asked to rate their confidence in their decision, on a 3-point Likert scale: 'Not at all confident' (coded 1), 'Somewhat confident' (coded 2), and 'Confident' (coded 3).

Quality checklist. Participants were given a checklist of six items that asked them which aspects of the stories helped them make their decision about which story was which. The items on this checklist were based on the quality questions used in Study 1 (e.g., 'The use of symbolism, literary devices, etc.', 'My level of enjoyment when reading the story', 'The story's theme'). Participants selected as many of these items as they wished. There was also an 'other' option where participants could provide their own responses about a different aspect of the story that helped them to make their decision.

Memory questions. The same memory questions created for Study 1 were used for this study. Because participants were now reading two stories instead of one, we only asked them two memory questions for each story instead of three. If participants failed at least three of the four memory questions, and/or only answered the memory questions correctly for one of the two stories, they were excluded from analysis.

AI expertise questions. We created a set of five self-report questions to examine participants' expertise and experience with ChatGPT and other AI software (see Table S1 in the Supplementary Material). For example, we asked, 'How often do you use ChatGPT or other AI software?' and 'Do you have an educational degree in a field related to Computer Science, Technical Engineering, Machine Learning, or Robotics?' These questions were scored as indicated in the table and were summed to form an AI expertise score for each participant (maximum possible score = 20).

Fictional literature expertise questions. We created a set of five self-report questions to ask about participants' experiences with and expertise in reading, writing, and/or analyzing fictional literature (see Table S2 in the Supplementary Material). For example, we asked, 'How often do you read fictional literature for pleasure/in your free time?' and 'How often do you have to read, analyze, and/or write fictional texts as part of your profession?' The scores for these questions were summed to form an overall fictional expertise score for each participant (maximum possible score = 14).

AI attitudes. We asked two questions to assess participants' attitudes about AI: 'Do you believe that AI programs like ChatGPT can generate human-quality work?' and 'Do you believe that ChatGPT or other AI software could potentially take over human jobs/roles?' As in Study 1, participants responded on a 3-point scale: 'Yes' (scored 1), 'Maybe' (scored 0.5), and 'No' (scored 0), which we averaged to create an overall attitude score.

3.1.3. Procedure

Participants first read one pair of stories, consisting of one human-written story and one ChatGPT-generated story (order randomized). Then, they were asked to select which story they believed to be written by a human (or which story they believed was written by AI, depending on the wording of their question). After making this choice, they rated their level of confidence and selected which aspects of the story helped them to make their decision.

Next, participants completed the memory questions about their stories and the AI attitudes questions. Participants then answered the AI expertise questions and the fictional literature expertise questions (order counterbalanced across participants). The items on each of these scales appeared in a random order for each participant. Participants then responded to demographic questions asking about their age, gender, race, highest level of education, and occupation.

Finally, participants were debriefed and told the true origins of the stories they had read.

3.2. Results

All data and analysis scripts are available at <https://osf.io/d8f6m/>.

3.2.1. Can participants differentiate human and AI writing?

The primary research question of Study 2 was whether participants would be able to correctly differentiate between the two stories at a better-than-chance level, when given two stories back to back. To test this, we conducted a simple numeric count of the individuals who correctly identified each story. We found that 167 participants (39.39%) were able to correctly select which story was human-written or which was AI-generated. This number was found to be significantly less than 50% by an exact proportion test ($p < .001$). That is, participants were worse than chance at guessing which story was written by each type of author. Performance differed slightly by story (see Supplementary Material), but participants were never above-chance at guessing the story's author.

We also examined the role of the question phrasing on participants' success. For participants who were asked, 'Which story do you believe was written by a human author?' ($n = 211$), 80 of them (37.91%) responded correctly, significantly less than 50% (exact proportion test, $p < .001$). For

Table 2. Number and percent of participants who selected features of the stories as helping them in their responses in Study 2, by correctness of response.

Story feature selected	Correct response	Incorrect response
Language	93 (33.09%)	188 (66.90%)
Character	66 (40.24%)	98 (59.76%)
Theme	52 (48.60%)	55 (51.40%)
Plot	53 (38.41%)	85 (61.59%)
Symbolism	90 (40.54%)	132 (59.46%)
Enjoyment	43 (30.71%)	97 (69.29%)

Table 3. Descriptive statistics for AI expertise, fictional literature expertise, and AI attitude scales.

	Mean	Standard deviation	Median	Range
AI expertise	5.02	3.04	5	Possible: 0–20 Actual: 0–14
Fictional literature expertise	4.77	2.62	4	Possible: 0–14 Actual: 0–14
AI attitudes	0.68	0.27	0.75	Possible: 0–1 Actual: 0–1

participants who were asked, ‘Which story do you believe was written by ChatGPT?’, 87 of them (40.8%) responded correctly, again significantly less than 50% (exact proportion test, $p < .001$). There was no difference in correct responding between these two phrasings of the question, $\chi^2(1) = 0.38$, $p = .54$.

We also tested whether participants’ confidence levels in their response related to the accuracy of their response using a logistic regression, predicting correct performance from confidence level. There was no significant relation ($\beta = -.17$, $p = .28$). On average, participants rated themselves around the middle of the confidence scale ($M = 2.19$, $SD = 0.65$).

We next tested whether the amount of time participants spent reading their assigned stories related to their success at guessing which story was which. Participants spent an average of 569.88 seconds ($SD = 339.60$) reading both stories. Reading time did not relate to correct performance, as shown by a logistic regression ($\beta = -0.0001$, $p = .64$).

We also examined which features of the story participants selected as having helped them to make their decision to see if any of these features tended to be more associated with a correct response to the main test question (see Table 2). Because we were conducting multiple comparisons, we adjusted the alpha level for these tests to .008. We found that participants were more likely to respond incorrectly if they reported relying on the story’s language ($\chi^2(1) = 13.81$, $p < .001$). There were no other statistically significant effects.

3.2.2. Relations with AI expertise, fictional literature expertise, and AI attitudes

Table 3 shows the descriptive statistics of participants’ scores on the two self-report expertise measures. Table 4 shows a correlation matrix for the two self-report expertise measures and the AI Attitudes measure.

To examine any effects of self-reported AI expertise on participants’ judgments, we ran a logistic regression analysis, predicting whether participants correctly guessed the story’s author from their AI expertise scores. We found a significant positive effect of AI expertise on participants’ correct responses

Table 4. Correlation coefficients for relations among AI expertise, fictional literature expertise, and AI attitudes.

	Fictional literature expertise	AI expertise	AI attitudes
Fictional literature expertise	X	X	X
AI expertise	0.11*	X	X
AI attitudes	−0.08	0.20**	X

* $p < .05$.** $p < .001$.

($\beta = .13$, $p < .001$); participants with higher AI expertise scores were more likely to correctly guess the author.

We ran a second logistic regression using participants' self-reported fictional literature expertise scores as a predictor. This analysis found no effect of fictional literature expertise ($\beta = -.006$, $p = .87$).

A third logistic regression predicted participants' correct responses about the identity of their author from their AI attitudes score. We found a significant positive effect, $\beta = .1.004$, $p = .008$.

There were significant positive correlations between AI attitudes and scores on the AI expertise scale, $r(422) = .20$, $p < .001$, and between AI expertise and fictional literature expertise, $r(422) = .11$, $p = .027$.

3.3. Discussion

The primary question of this study was whether participants could tell the difference between an AI-generated short story and a human-written short story. We found that, overall, participants were significantly worse than chance at correctly differentiating between stories written by a human author and stories generated by ChatGPT. Interestingly, we found no difference in the general pattern of responding when participants were asked to select which story was human-written and which story was AI-generated. That is, participants were equally likely to mistakenly select that a human-written story was AI-generated as the reverse.

One of the story features that participants said they relied on when making their decision (the ease of understanding the story's language) was related to worse performance on the main test question. Together with our main finding that participants were worse than chance at determining which story was human-written and which was AI-generated, this result suggests that people may be focusing on the wrong features when making these decisions, which may lead them astray.

Additionally, we explored how self-reported expertise in different fields may impact participants' accuracy at selecting which story was which. Although self-reported expertise with fictional literature did not relate to accuracy, we found evidence that self-reported expertise with AI systems did positively relate to participants' performance on our task. This result suggests that it may be possible to improve people's abilities to determine whether a text is AI-generated by providing them with more experience with these systems and the kinds of outputs that they tend to generate.

In order to more robustly examine the effects of participants' expertise with and attitudes toward AI, Study 3 replicates Study 2 but using a previously validated scale, the AI Literacy Scale (AILS; Wang et al., 2023). We decided to use this scale based on recommendations from a review article that evaluated 16 scales (Lintner, 2024). This review outlined each scale's structural validity and internal consistency, and it recommended the AILS for use with the general population.

4. Study 3

4.1. Method

This study was not preregistered.

4.1.1. Participants

The final sample consisted of 481 adults recruited from Prolific (232 men, 239 women, 8 nonbinary, 1 other, and 1 preferred not to answer; average age = 46.22 years, SD age = 15.66 years, age range = 19–80 years). As in the prior two studies, the sample was recruited to be demographically representative of the United States with respect to age, gender, and race. Participants were paid \$4.

An additional 56 participants were recruited but were not included in the final sample because they either failed to complete the study in its entirety ($n = 37$), failed to answer the question asking which story was which ($n = 1$), or failed one or more memory questions from each story ($n = 18$).

4.1.2. Materials and procedure

We used the same materials and procedure as in Study 2, except that we removed the AI expertise and AI attitudes questions and replaced them with the AILS (Wang et al., 2023). The AILS is a self-report scale that consists of twelve statements that fall equally into four factors: awareness (e.g., ‘I can distinguish between smart devices and non-smart devices’), usage (e.g., ‘I can skillfully use AI applications or products to help me with my daily work’), evaluation (e.g., ‘I can evaluate the capabilities and limitations of an AI application or product after using it for a while’), and ethical standards (e.g., ‘I always comply with ethical principles when using AI applications or products’). Participants rate their level of agreement with these statements on a 7-point Likert scale, which we converted into a –3 to 3 scale for analysis. Three of the items are reverse-coded. We added an attention-checking item to this scale that asked participants to select ‘Strongly disagree’ if they were reading the item. Because four participants failed this attention check, we exclude these participants from any analyses that include the AILS.

4.2. Results

All data and analysis scripts are available at <https://osf.io/d8f6m/>.

4.2.1. Can participants differentiate human and AI writing?

We began by counting the number of participants who correctly identified the author of their stories. There were 250 participants (51.97%) who responded correctly and 231 participants (48.02%) who responded incorrectly. This number was no different from 50%, exact proportion test, $p = .41$. This pattern held for two out of the three stories, but participants who received ‘FISH/Reflections in Still Water’ responded correctly at above-chance levels (see analyses in [Supplementary Material](#)).

We next tested whether there was an effect of question phrasing. Participants who were asked to select which story was written by a human ($n = 239$) were correct 50.21% of the time ($n = 120$), no different from chance (exact proportion test, $p = 1$). Participants who were asked to select which story was written by ChatGPT ($n = 242$) were correct 54.72% of the time ($n = 130$), no different from chance (exact proportion test, $p = .27$). There was no difference in responding between these two question phrasings, $\chi^2(1) = 0.59$, $p = .44$.

We also tested for any effect of participants’ reported confidence in their decision ($M = 2.43$, $SD = .51$), using a logistic regression to predict correct performance from confidence level. There was no significant relation ($\beta = 0.13$, $p = .48$).

We next looked for relations between correct performance and the amount of time participants spent reading their two assigned stories. On average, participants spent a total of 612.72 seconds ($SD = 354.12$) reading both stories (range = 123.71–2048.23 seconds). There was a statistically significant (though trivial) effect of total reading time ($\beta = -0.001$, $p < .001$), such that participants who spent *less* time reading their assigned stories were more likely to choose correctly.

We also looked at which features of the stories were selected as helping participants to make their decisions (see [Table 5](#)), again adjusting our alpha level to .008 to account for multiple comparisons. Participants were significantly more likely to respond incorrectly if they said that they relied on the story’s language ($\chi^2(1) = 9.11$, $p = .003$) or their level of enjoyment ($\chi^2(1) = 9.57$, $p = .002$). Conversely,

Table 5. Number and percent of participants who selected features of the stories as helping them in their responses in Study 3, by correctness of response.

Story feature selected	Correct response	Incorrect response
Language	148 (46.98%)	167 (53.02%)
Character	109 (51.90%)	101 (48.10%)
Theme	87 (60.84%)	56 (39.16%)
Plot	104 (53.33%)	91 (46.67%)
Symbolism	133 (58.33%)	95 (41.67%)
Enjoyment	53 (40.46%)	78 (59.54%)

participants were marginally significantly more likely to respond correctly if they said that they relied on the story's use of symbolism ($\chi^2(1) = 7.02, p = .008$). There was no effect of reliance on character, theme, or plot.

4.2.2. Relations with AI expertise and fictional literature expertise

Participants' average score on the AILS measure was 1.11 ($SD = 0.66$). We ran a logistic regression to predict participants' correct responses about the story's author from these scores, and we found that there was a significant effect ($\beta = .29, p = .041$)—participants with higher scores were more likely to respond correctly.

Participants' average score on the fictional literature expertise measure was 4.78 ($SD = 2.43$). We ran a second logistic regression on participants' correct responses and their fictional literature scores but found no effect ($\beta = .004, p = .91$).

There was a marginally significant positive correlation between participants' scores on these two measures, $r(475) = 0.08, p = .09$.

4.3. Discussion

The results of Study 3 largely confirm what we found in Study 2: There is no general evidence that participants were able to distinguish human-written from AI-generated stories. One interesting difference between these studies is that performance in Study 2 was significantly worse than chance (i.e., participants tended to mistakenly judge their given story as being written by the opposite type of author), while here, performance was at chance (i.e., participants made errors equally in both directions). We do not have a ready explanation for this difference. However, given that Study 3 was run several months after Study 2, it is possible that there was more individual or societal exposure to AI-generated content, making participants less poor at making a judgment about this content. Importantly, though, in neither study were participants generally better than chance at making this judgment. To investigate this issue further, future studies could give participants stories one at a time and ask them to judge each story individually, allowing for greater insight into the types of errors that people make.

Additionally, at least some of the story features that they relied on to make this decision led to worse performance. This suggests that the ways in which participants were interacting with the stories to make this decision, while they may have seemed sensible, were not actually helpful; participants generally seemed to have poor instincts about how to tell which story was AI-generated. Relatedly, participants who spent more time reading were less successful at this task. A quick assessment rather than a more detailed examination of these stories may thus be more effective, as is sometimes the case with using 'thin slices' of behavior to make accurate predictions about people (Ambady, 2010). However, we should be cautious in our interpretations of this finding, since the effect size was extremely small. Finally, as in Study 2, participants with higher self-reported levels of AI literacy (as measured by the AILS) performed better on this task, but there was no relation with expertise with fictional literature.

5. General discussion

The three studies reported here aimed to investigate how people respond to AI-generated short fictional stories and whether they can differentiate these stories from human-written ones. Study 1 found not only that people tended to rate AI-generated stories more favorably than human-written stories but also that they tended to rate stories that were described as human-written more favorably. That is, people actually prefer AI-generated stories while believing that these stories are worse.

Studies 2 and 3 found that people's performance was at or below chance when asked which of two stories was human-written. Individuals with higher self-reported levels of experience with or more positive attitudes toward AI were more successful at this task; this was not the case for individuals with higher self-reported levels of experience with fictional stories. Taken together, these results suggest that creative outputs generated by AI systems are perceived as at least comparable to human work, if not better.

Given that people tend to show a bias against AI-generated creative work, why are AI-generated stories judged more favorably? One possibility (explored in more detail in Porter and Machery, 2024) is that AI-generated work is seen as easier to interpret—people generally like these works better because they feel as though they are easier to read or understand. People may assume that poems or stories that they like better must be written by humans due to preexisting biases against AI. These two findings thus reinforce each other: ChatGPT-written stories seem more human because people respond positively to them, and insofar as people respond positively to them, they are judged to be written by humans. While there are measurable differences between texts generated by humans and by AI systems (Berber Sardinha, 2024), future work should investigate in more detail exactly which differences consumers rely on when making judgments about a work's origin. Our results from Studies 2 and 3 indicate that the features of stories that people say they rely on in making their judgments may be the wrong features to use. This may be one way in which greater familiarity with AI systems could be helping people to differentiate these stories; this greater expertise may help people to focus on the features of text that actually are different between human and AI writing.

A second, related, possibility relies on the way that AI systems like ChatGPT assemble new work out of data culled from millions of existing pieces of data. This process in some sense averages together the tendencies that exist in the data. In this way, the very nature of the way in which ChatGPT generates creative works may smooth out the rough edges of human-generated works into something that people find more appealing, much in the same way that images of faces generated by averaging real human faces together are judged to be more attractive (Langlois and Roggman, 1990).

Our participants only judged the output of a generative AI system, and we found that these products are seen as being at least as good as human-created products. But our work does not speak to how people might perceive differences in the creative process between humans and AI systems. Broadly, human creativity involves two distinct steps: idea generation (or brainstorming) and selection of a promising direction to pursue. However, generative AI works differently. Although the field is changing rapidly, the models that we used for the current study included only the first step (generation) but not the second (reflection/selection). This difference highlights one possible reason why AI output is generally seen as higher quality, which relies solely on how they complete the first step: AI systems have massively larger amounts of material to draw from and combine into new content, and conceptual combination is a foundational process in creativity (Beaty and Kenett, 2023; Coutanche et al., 2020; Shtulman, 2023).

However, the creative output generated by AI systems tends to be relatively homogeneous (Wenger and Kenett, 2026). Individual human responses may be judged as less creative, but the body of human responses is significantly more variable. That is, humans explore more of the possibility space than AI systems, which potentially means that humans require more reflection to select an appropriate response. This suggests that future work in this area should more closely examine the second step of the creative process, since many AI systems do not incorporate this selection step. Indeed, when humans are in the loop, selecting the best AI output from several options, participants judge AI-generated work even more positively (Köbis and Mossink, 2021). This further suggests that our evaluation of creative work

may need to shift to considering larger bodies of output than just individual stories. Doing so would provide a more holistic understanding of how the different creative processes at work in humans and AI systems lead to qualitatively different outcomes. Finally, people's judgments could be affected by seeing not just the output itself but also information about the process by which a person or AI system has generated that output could affect people's judgments; responses will likely vary if the output is seen as generated by a process involving authentic (human) creativity (Jackson et al., 2023).

6. Limitations

The primary limitation of the present study is the nature of the stories. They were very brief, taking only a few minutes to read in their entirety. This is also true of much of the prior work on this topic, as the stimuli are designed to be read during a brief experiment. It is possible that longer creative works, such as entire novels, may show a different pattern of responses, as AI programs may be less able to develop characters and plot points over the course of a longer work, but this remains an open question. This would mean that the effects that we observed here may not occur with longer narratives.

Relatedly, all of the short stories used here were realistic fiction. But participants may have different reactions to other genres, perhaps particularly if those genres align with their own preferences. Specifically, future work should look at whether AI systems might be less successful at generating convincingly human-like stories in genres like romance, which deal with human relationships and emotions, or whether they might be more successful in such domains due to the highly stereotyped characters and plots that tend to characterize such stories. A relevant prior study compared participants' responses to a realistic fiction story and a science fiction story, both of which were presented as either human-written or AI-generated. Surprisingly, participants experienced less engagement with the realistic fiction story when they thought it was AI-generated, but there was no such difference for the science fiction story (Messingschlager and Appel, 2022). This suggests that people may view AI programs as being more capable of generating certain types of content, which, in turn, affects their responses to this content.

As noted earlier, the forced-choice task we used in Studies 2 and 3 explicitly told participants that one of their stories was human-written and one was AI-generated, information that is not generally available in the real world as people are making judgments about content. Given that people performed poorly at differentiating the stories even with this information, we believe that they would perform even more poorly under more naturalistic circumstances, but this is an open question. Future research could investigate this issue by presenting stories one at a time and asking participants to select the author type. Such a design could also manipulate how many human-written vs. AI-generated stories a particular participant received and could vary whether people were told how many of each they would receive, to see if that affected their accuracy.

With the use of AI software raising concerns in academic and work settings, studying work that would be typically done in these environments, such as homework assignments or occupational writing, would be beneficial for employers, school faculty, and policymakers. Some work has already begun to study these forms of writing (e.g., Herbold et al., 2023), and a recent review explores why people prefer AI-generated content across a wide variety of domains (Coutanche and Weisberg, 2026). More systematic examinations of a wider variety of writing forms are needed to determine the true extent of this preference.

Another limitation is the use of self-report measures for our expertise scales, which may not have yielded the most accurate responses, even though we used a validated measure in Study 3. Future work should include more objective measures of expertise to more fully explicate how this variable relates to task performance.

The final limitation is about our participants. Being an online survey that is not completed in a lab setting, it is possible that the participants may have been distracted or using outside resources while completing the survey—or could be AI bots themselves! Although Prolific has been used as a

reliable source of participants across many areas of research, and it has the advantage of allowing us to collect a representative sample, future research should also conduct a similar survey in the controlled environment of a lab to see if the results replicate.

Additionally, the population studied in this paper is adults in the United States between the ages of 18 and 81. Although this represents a broad population with respect to demographic variables like age and education level, responses to AI-generated materials may differ across countries and cultures, depending on local traditions of how fictional narratives are structured and how AI-generated content is perceived. Because many current AI tools, including ChatGPT, have been developed largely for and by a global northern and western market, our choice of a U.S.-based subject population fits well with these tools' capabilities. But tools from different cultures (like DeepSeek, developed in China) might generate a different pattern of responses, and future research should investigate this issue in more detail.

7. Conclusion

The current results indicate that AI programs based on large language models are capable of generating short stories that cannot be distinguished from human-written stories and that tend to be perceived as higher quality and more absorbing. Given that the abilities of AI programs will only increase over time, these results raise difficult questions about whether and how we use such programs in making creative work and about the nature of creativity itself.

Supplementary material. The supplementary material for this article can be found at <http://doi.org/10.1017/jdm.2026.10042>.

Data availability statement. All materials, data, and analysis scripts are available at <https://osf.io/d8f6m/>.

Acknowledgments. Thanks to everyone at the STAR Lab, especially Brooke Gonzalez, for their assistance with this study. A version of this manuscript was presented at the annual meeting of the Psychonomic Society (November 2024).

Funding statement. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing interest. The authors declare none.

References

- Aharoni, E., Fernandes, S., Brady, D. J., Alexander, C., Criner, M., Queen, K., . . . Crespo, V. (2024). Attributions toward artificial agents in a modified moral Turing test. *Scientific Reports*, 14(1), 8458. <https://doi.org/10.1038/s41598-024-58087-7>.
- Ambady, N. (2010). The perils of pondering: Intuition and thin slice judgments. *Psychological Inquiry*, 21(4), 271–278. <https://doi.org/10.1080/1047840X.2010.524882>.
- Beaty, R. E., & Kenett, Y. N. (2023). Associative thinking at the core of creativity. *Trends in Cognitive Sciences*, 27(7), 671–683. <https://doi.org/10.1016/j.tics.2023.04.004>.
- Beguš, N. (2024). Experimental narratives: A comparison of human crowdsourced storytelling and AI storytelling. *Humanities and Social Sciences Communications*, 11(1), 1–22. <https://doi.org/10.1057/s41599-024-03868-7>.
- Berber Sardinha, T. (2024). AI-generated vs human-authored texts: A multidimensional comparison. *Applied Corpus Linguistics*, 4(1), 100083. <https://doi.org/10.1016/j.acorp.2023.100083>.
- Castano, E., Zanella, J., Saeedi, F., Zunshine, L., & Ducceschi, L. (2023). On the complexity of literary and popular fiction. *Empirical Studies of the Arts*, 027623742311634. <https://doi.org/10.1177/02762374231163483>.
- Chakrabarty, T., Laban, P., Agarwal, D., Muresan, S., & Wu, C. S. (2023). Art or artifice? Large language models and the false promise of creativity. In *Proceedings of the conference on human factors in computing systems* (pp. 1). <https://doi.org/10.1145/3613904.3642731>.
- Chamberlain, R., Mullin, C., Scheerlinck, B., & Wagemans, J. (2018). Putting the art in artificial: Aesthetic responses to computer-generated art. *Psychology of Aesthetics, Creativity, and the Arts*, 12(2), 177–192. <https://doi.org/10.1037/ACA0000136>.
- Colby, K. M., Hilf, F. D., Weber, S., & Kraemer, H. C. (1972). Turing-like indistinguishability tests for the validation of a computer simulation of paranoid processes. *Artificial Intelligence*, 3, 199–221. [https://doi.org/10.1016/0004-3702\(72\)90049-5](https://doi.org/10.1016/0004-3702(72)90049-5).
- Coutanche, M. N., & Weisberg, D. S. (2026). The cognitive biases underlying our preference for AI. *Manuscript under Review*.

- Coutanche, M. N., Solomon, S. H., & Thompson-Schill, S. L. (2020). Conceptual combination. In D. Poeppel, G. R. Mangun, & M. S. Gazzaniga (Eds.), *The cognitive neurosciences* (6th ed., pp. 827–836). MIT Press. <https://doi.org/10.7551/mitpress/4071.003.0008>.
- Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1), 114–126. <https://doi.org/10.1037/xge0000033>.
- Dillion, D., Mondal, D., Tandon, N., & Gray, K. (2025). AI language model rivals expert ethicist in perceived moral expertise. *Scientific Reports*, 15(1), 4084. <https://doi.org/10.1038/s41598-025-86510-0>.
- Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2024). GPTs are GPTs: Labor market impact potential of LLMs. *Science*, 384(6702), 1306–1308. <https://doi.org/10.1126/science.adj0998>.
- Ericsson, K. A., & Lehmann, A. C. (1996). Expert and exceptional performance: Evidence of maximal adaptation to task constraints. *Annual Review of Psychology*, 47, 273–305. <https://doi.org/10.1146/annurev.psych.47.1.273>.
- Graves, B., & Frederiksen, C. H. (1991). Literary expertise in the description of a fictional narrative. *Poetics*, 20(1), 1–26. [https://doi.org/10.1016/0304-422X\(91\)90031-J](https://doi.org/10.1016/0304-422X(91)90031-J).
- Herbold, S., Hautli-Janisz, A., Heuer, U., Kikteva, Z., & Trautsch, A. (2023). A large-scale comparison of human-written versus ChatGPT-generated essays. *Scientific Reports*, 13(1), 18617. <https://doi.org/10.1038/s41598-023-45644-9>.
- Hong, J. W., & Curran, N. M. (2019). Artificial intelligence, artists, and art. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 15(2s). <https://doi.org/10.1145/3326337>.
- Jackson, J. C., Yam, K. C., Tang, P. M., Liu, T., & Shariff, A. (2023). Exposure to robot preachers undermines religious commitment. *Journal of Experimental Psychology: General*. <https://doi.org/10.1037/xge0001443>.
- 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, 106553. <https://doi.org/10.1016/j.chb.2020.106553>.
- Kuijpers, M. M., Hakemulder, F., Tan, E. S., & Doicaru, M. M. (2014). Exploring absorbing reading experiences: Developing and validating a self-report scale to measure story world absorption. *Scientific Study of Literature*, 4(1), 89–122. <https://doi.org/10.1075/SSOL.4.1.05KUI>.
- Langlois, J. H., & Roggman, L. A. (1990). Attractive faces are only average. *Psychological Science*, 1(2), 115–121. <https://doi.org/10.1111/j.1467-9280.1990.tb00079.x>.
- Lintner, T. (2024). A systematic review of AI literacy scales. *NPJ Science of Learning*, 9(1), 50. <https://doi.org/10.1038/s41539-024-00264-4>.
- Markowitz, D. M. (2024). From complexity to clarity: How AI enhances perceptions of scientists and the public's understanding of science. *PNAS Nexus*, 3(9), 387. <https://doi.org/10.1093/pnasnexus/pgae387>.
- Mei, Q., Xie, Y., Yuan, W., & Jackson, M. O. (2024). A Turing test of whether AI chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences of the United States of America*, 121(9). <https://doi.org/10.1073/pnas.2313925121>.
- Messingschlager, T. V., & Appel, M. (2022). Creative artificial intelligence and narrative transportation. *Psychology of Aesthetics, Creativity, and the Arts*. <https://doi.org/10.1037/aca0000495>.
- Muñoz-Ortiz, A., Gómez-Rodríguez, C., & Vilares, D. (2024). Contrasting linguistic patterns in human and LLM-generated news text. *Artificial Intelligence Review*, 57(9), 265. <https://doi.org/10.1007/s10462-024-10903-2>.
- Orwig, W., Edenbaum, E. R., Greene, J. D., & Schacter, D. L. (2024). The language of creativity: Evidence from humans and large language models. *Journal of Creative Behavior*, 58(1), 128–136. <https://doi.org/10.1002/jocb.636>.
- Osborne, M. R., & Bailey, E. R. (2025). Me vs. the machine? Subjective evaluations of human- and AI-generated advice. *Scientific Reports*, 15(1), 3980. <https://doi.org/10.1038/s41598-025-86623-6>.
- Park, J., Oh, C., & Kim, H. Y. (2024). AI vs. human-generated content and accounts on Instagram: User preferences, evaluations, and ethical considerations. *Technology in Society*, 79, 102705. <https://doi.org/10.1016/j.techsoc.2024.102705>.
- Porter, B., & Machery, E. (2024). AI-generated poetry is indistinguishable from human-written poetry and is rated more favorably. *Scientific Reports*, 14(1), 26133. <https://doi.org/10.1038/s41598-024-76900-1>.
- Ragot, M., Martin, N., & Cojean, S. (2020). AI-generated vs. human artworks: A perception bias towards artificial intelligence? In *Proceedings of the conference on human factors in computing systems*. <https://doi.org/10.1145/3334480.3382892>.
- Raj, M., Berg, J. M., & Seamans, R. (2026). The artificial intelligence disclosure penalty: Humans persistently devalue AI-generated creative writing. *Journal of Experimental Psychology: General*. <https://doi.org/10.1037/xge0001889>.
- Reber, R., Schwarz, N., & Winkielman, P. (2004). Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience? *Personality and Social Psychology Review*, 8(4), 364–382. https://doi.org/10.1207/s15327957pspr0804_3.
- Shidiq, M. (2023). The use of artificial intelligence-based ChatGPT and its challenges for the world of education; from the viewpoint of the development of creative writing skills. In A. Mundiri(Ed.), *Proceedings of the international conference on education, society and humanity*. Nurul Jadid University, Probolinggo, Indonesia, (pp. 353–357).
- Shulman, A. (2023). *Learning to imagine: The science of discovering new possibilities*. Harvard University Press.
- Silvia, P. (2013). Interested experts, confused novices: Art expertise and the knowledge emotions. *Empirical Studies of the Arts*, 31(1), 107–115. <https://doi.org/10.2190/em.31.1.f>.

- Wang, B., Rau, P. L. P., & Yuan, T. (2023). Measuring user competence in using artificial intelligence: Validity and reliability of artificial intelligence literacy scale. *Behaviour & Information Technology*, 42(9), 1324–1337. <https://doi.org/10.1080/0144929x.2022.2072768>.
- Wenger, E., & Kenett, Y. N. (2026). Large language models are homogeneously creative. *PNAS Nexus*, 5(3). <https://doi.org/10.1093/pnasnexus/pgag042>.
- Zhang, Y., & Gosline, R. (2023). Human favoritism, not AI aversion: People’s perceptions (and bias) toward generative AI, human experts, and human–GAI collaboration in persuasive content generation. *Judgment and Decision Making*, 18, e41. <https://doi.org/10.1017/jdm.2023.37>.
- Zunshine, L. (2022). *The secret life of literature*. <https://doi.org/10.7551/mitpress/13964.001.0001>.