Weronika Elżbieta Falinska - MSc Data Science and Artificial Intelligence 2025/2026

Unit: Advanced Mathematics and Statistics for Data Science and AI

Date: 8th December 2025

Human and AI Contributions to Creative Writing: Effects on Self-Perception and External Evaluation

---

1.Abstract:

This project examines how generative AI affects writers' self-perceptions and how readers evaluate short creative stories. The overarching research question is how AI involvement shapes perceptions of creativity, authorship, and story quality across both writers and external evaluators.
Non-parametric methods were used throughout, with Mann–Whitney U tests applied to group comparisons and Spearman's rank-order correlations used to assess associations between AI suspicion and blind ratings. The results showed that AI use had selective effects on writers' self-perception, influencing authorship, appropriateness, and perceived future impact, while other creativity ratings remained unchanged. Before disclosure, evaluators rated AI-assisted stories slightly but consistently higher across most quality dimensions, though effect sizes were small. After disclosure, however, evaluators strongly reduced attributions of authorship and ownership for AI-assisted stories. Overall, the findings indicate that evaluative penalties arise only once AI involvement is revealed, whereas blind assessments remain largely unbiased.

2. Introduction:

Generative AI is increasingly used in creative writing, raising questions about how such tools shape both the writer's experience and the way audiences judge creative work. This project investigates how AI involvement influences self-perception, perceived creative ownership, and external evaluation of short fictional stories. The motivation stems from ongoing debates around whether AI enhances creativity or diminishes the legitimacy of human authorship. Personally, I was interested in whether AI changes not only how people write, but also how they feel about their work and how their work is judged by others.

Four research questions were addressed:

RQ1 – Writers' self-perception
Do writers who use AI evaluate their own stories differently from those who do not?

H_0 (RQ1): Self-perception ratings come from the same distribution for AI and non-AI writers.
H_1 (RQ1): At least one self-perception dimension differs.

RQ2a – Blind evaluations
Do evaluators rate AI-assisted and human-written stories differently before knowing how they were produced?

H_0 (RQ2a): Blind creativity ratings do not differ for AI vs non-AI stories.
H_1 (RQ2a): Blind ratings differ for at least one quality dimension.

RQ2b – Post-disclosure judgements
Once evaluators learn whether a story involved AI, do their ratings of authorship and ownership differ for AI vs non-AI stories?

H_0 (RQ2b): Post-disclosure authorship/ownership ratings do not differ for AI vs non-AI stories.
H_1 (RQ2b): AI-assisted stories receive different authorship/ownership ratings after disclosure.

RQ2c – Misattribution bias
Before knowing how a story was written, do evaluators' suspicions of AI involvement influence the blind creativity ratings they provided earlier?

H_0 (RQ2c): AI suspicion is not associated with blind creativity ratings ($\varrho = 0$).
H_1 (RQ2c): AI suspicion predicts blind ratings ($\varrho \neq 0$).

3. Methodology

All analyses were conducted in Python using pandas, SciPy, and Seaborn. Because all creativity, quality, and authorship measures were recorded on 1–9 Likert scales, the data are ordinal, discrete, and typically skewed. Shapiro–Wilk tests and Q–Q plots confirmed substantial deviations from normality for every variable, meaning that parametric methods such as t-tests or ANOVA would not be appropriate. For this reason, the analysis relied exclusively on non-parametric statistical techniques.

3.1 Actual AI-usage vs assigned conditions

Although writers were randomly assigned to three conditions (human, human_1AI, human_5AI), descriptive checks revealed that many participants did not follow the intended manipulation:

- In the 1-AI condition, 44% of writers completely ignored the AI-generated suggestion.
- In the 5-AI condition, 15% ignored all five suggestions.
- In the human condition, two writers still used AI (either self-reported or evident from AI-assisted idea fields).

Because assigned condition did not reliably reflect actual behaviour, all inferential tests use a reconstructed binary variable:
ai_used_actual = 1 if the writer either self-reported AI use or incorporated at least one AI-generated idea.
This provides a more valid behavioural comparison for RQ1 and RQ2.

3.2 Variables excluded from interpretation

Two variables were excluded due to inconsistencies identified during data checks:

- "Badly written" / "well written" mismatch
  The dataset column is named *badly_written*, but participants saw the item labelled *well written* in the survey. Because the scoring direction is unknown, meaningful interpretation is impossible. It is included in analyses for completeness but not discussed in results.
- Post-disclosure "profit deserved"
  This item was only shown to evaluators *when the story was known to involve AI*, meaning there is no comparison group. Additionally, responses were conceptually inconsistent (e.g., some participants entered *0%* to emphasise that the AI deserved *no* credit rather than to specify the human author's share). As a result, this variable is statistically unreliable and excluded.

3.3 Statistical methods

- RQ1 — Self-perception differences between AI users and non-users

This question compares two independent groups (writers who used AI vs those who did not). Because variables are ordinal and non-normal, the **Mann–Whitney U test** was used for all dimensions, complemented by **Cliff's delta** for effect size.

- RQ2a — Blind evaluations of story quality

Evaluators rated stories before knowing whether AI had been used.
As with RQ1, this required comparing two independent groups using **Mann–Whitney U tests** across all pre-disclosure rating dimensions.

- RQ2b — Post-disclosure authorship and ownership judgements

Once evaluators learned whether AI was used, they rated the perceived source of ideas and authorship. Again, group comparisons were performed using **Mann–Whitney U tests**, with **Cliff's delta** used to quantify the magnitude and direction of the disclosure penalty.

- RQ2c — AI suspicion and blind ratings

This question tested whether evaluators' *later* belief about AI usage (0–100%) predicted the *earlier* blind ratings they had already given. A correlation-based approach was required. Because the suspicion score is continuous and the outcome variables are ordinal and non-normal, the appropriate method is **Spearman's rank-order correlation**, which detects monotonic associations without assuming linearity.

4. Results

4.1 RQ1 – Writers' self-perception

RQ1 examined whether writers who used AI to assist their creative process evaluated their own stories differently from writers who did not use AI. Mann–Whitney U tests compared self-perception ratings across thirteen dimensions, including authorship, creativity, appropriateness, enjoyment, and perceived future impact. The results showed a selective rather than global influence of AI on how writers judged their own work.

Most creativity-related evaluations—such as novelty, originality, rarity, feasibility, publishability, enjoyment, humour, boredom, and narrative twist—did **not** differ significantly between AI users and non-users. These patterns are visible in the forest plot (Figure 1), where the majority of Cliff's delta values cluster tightly around zero, and in the median comparison plot and boxplots (Figures 2 and 3), where the distributions for both groups overlap extensively. These findings indicate that writers generally felt their stories were similarly creative and stylistically coherent regardless of AI involvement.

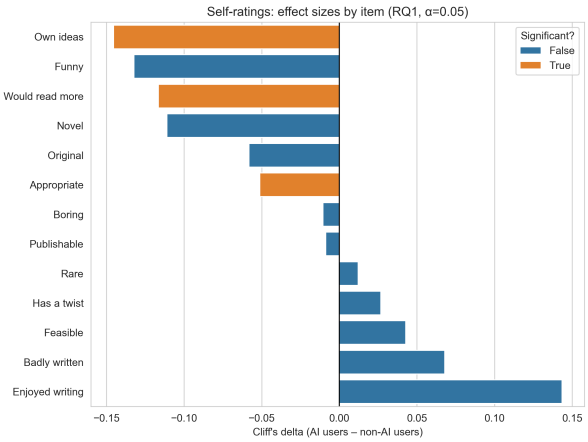

Figure 1. Self-ratings: effect sizes (Cliff's delta) by item (RQ1).
Shows that most deltas cluster near zero; significant variables highlighted.

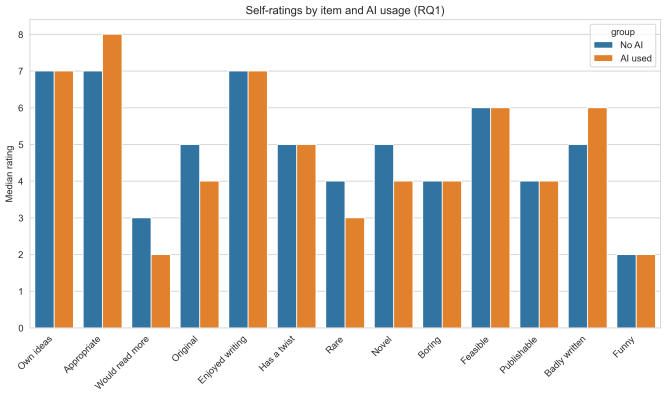

Figure 2. Median self-ratings for AI vs non-AI writers (RQ1)
Demonstrates near-identical medians except for authorship, appropriateness, and future impact.

Three dimensions, however, showed statistically significant differences. Writers who used AI reported **lower authorship**, indicating that they felt fewer of the ideas originated from themselves. Conversely, they rated their stories as **more appropriate** for the task, suggesting that AI guidance may have helped meet genre expectations or constraints. Finally, AI-assisted writers gave lower scores on the "future expectations" item, implying that their stories felt less personally meaningful or less likely to influence their future reading preferences.

One variable requires caution: **"badly written"**. Although stored under that label in the dataset, participants saw the reversed phrasing **"well written"** in the survey interface. Because the direction of the scale is ambiguous, this item is excluded from substantive interpretation.

Decision: H_0 is partially rejected.

AI use affected only a small subset of self-perception dimensions—authorship, appropriateness, and future-impact—while all other creativity and stylistic judgements showed no reliable differences. Overall, AI assistance produced narrow and specific effects on writers' self-evaluation, rather than broad changes in perceived story quality.

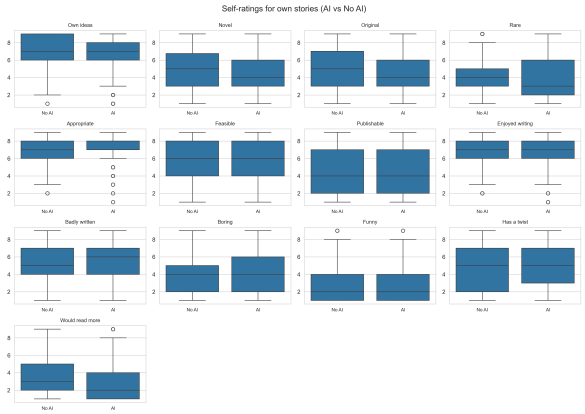

Figure 3. Boxplot grid for all self-perception variables (RQ1)

Visually confirms high overlap across distributions.

4.2 RQ2a – Blind story evaluations

Before being informed about how each story was produced, evaluators rated its creativity and stylistic qualities across twelve dimensions. Mann–Whitney U tests compared these pre-disclosure ratings for stories written with AI assistance and those written without AI. The results showed a clear and consistent pattern: **AI-assisted stories received higher blind evaluations on the majority of quality measures**.

Significant advantages for AI-assisted stories were observed for novelty, originality, rarity, appropriateness, feasibility, publishability, enjoyment, and narrative twist (most $p < .01$). Although these effects were small, they were uniform in direction and indicate that AI assistance modestly improved perceived story quality when evaluators had no knowledge of authorship. The largest median differences appeared in the "twist", "feasible", and "rare" items. These effects are visible in the effect-size summary (Figure 4), where most Cliff's delta values fall above zero, and in the distributional comparisons shown in the boxplot grid (Figure 5).

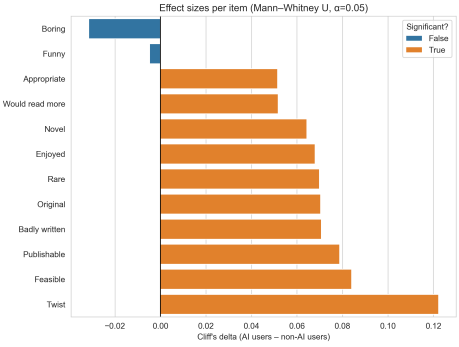

Figure 4: Effect-size forest plot (RQ2a)

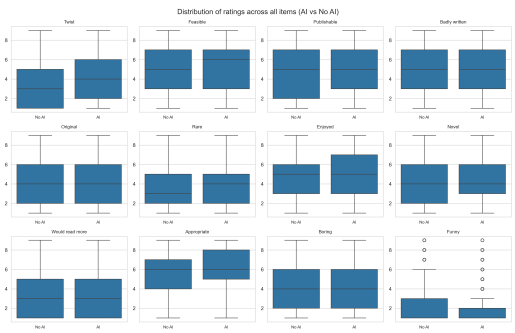

Figure 5. Blind-evaluation boxplots for all story-quality variables (RQ2a).

Two dimensions—"boring" and "funny"—showed no measurable differences between groups, indicating that AI assistance neither enhanced nor diminished the perceived dullness or humour of the stories. One variable, labelled "badly written", was excluded from interpretation because the participant-facing wording used the opposite phrasing ("well written"), making its direction conceptually ambiguous.

Decision: H_0 is largely rejected.

Across 9 of the 11 interpretable dimensions, AI-assisted stories were judged more favourably than human-written stories under blind conditions. This suggests that AI involvement appears to systematically improve perceived story quality when evaluators are unaware that AI was used.

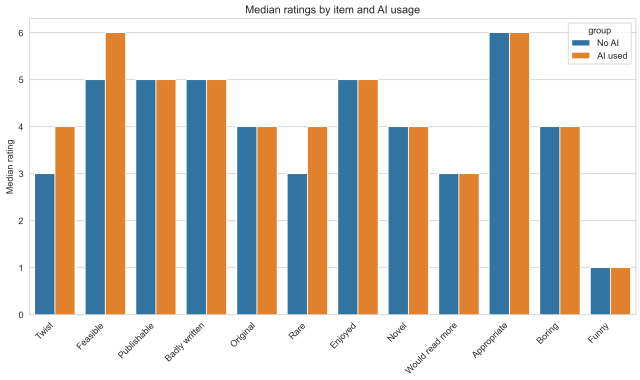

Figure 6. Median Comparison Boxplot (RQ2a).

4.3 RQ2b – Post-disclosure authorship and ownership

RQ2b examined whether evaluators judged stories differently once they were informed whether the author had used AI assistance. Two dimensions were assessed: the extent to which the story's ideas were attributed to the human author (authorship) and the perceived ownership of those ideas. Mann–Whitney U tests revealed **large and systematic penalties** for AI-assisted stories following disclosure.

Authorship ratings showed a substantial shift: the median rating fell from **8** for human-written stories to **5** for AI-assisted stories. Ownership showed a similar pattern, with medians dropping from **8** to **6**. These results indicate that once AI involvement became known, evaluators attributed considerably less creative credit to the human writer. The median comparison plot (Figure 7) clearly illustrates these downward shifts, and the distributional differences are similarly pronounced in the boxplots (Figure 8).

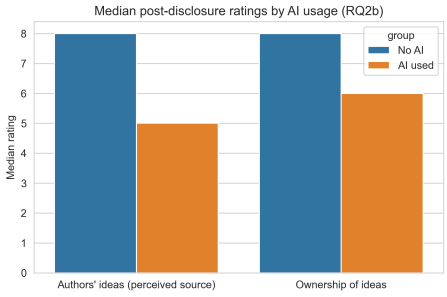

Figure 7. Median post-disclosure authorship + ownership ratings (RQ2b).

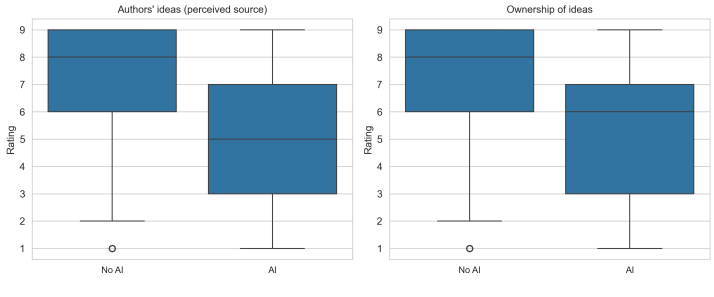

Figure 8. Boxplots of authorship and ownership after disclosure.
Clear downward shift for AI-assisted stories.

Effect sizes, quantified using Cliff's delta, were large and negative (approximately –0.35 to –0.45), indicating a consistent and meaningful downward shift in evaluations for AI-assisted stories. The effect-size summary (Figure 9) shows both variables strongly favouring the non-AI group, with no overlap in directionality or ambiguity.

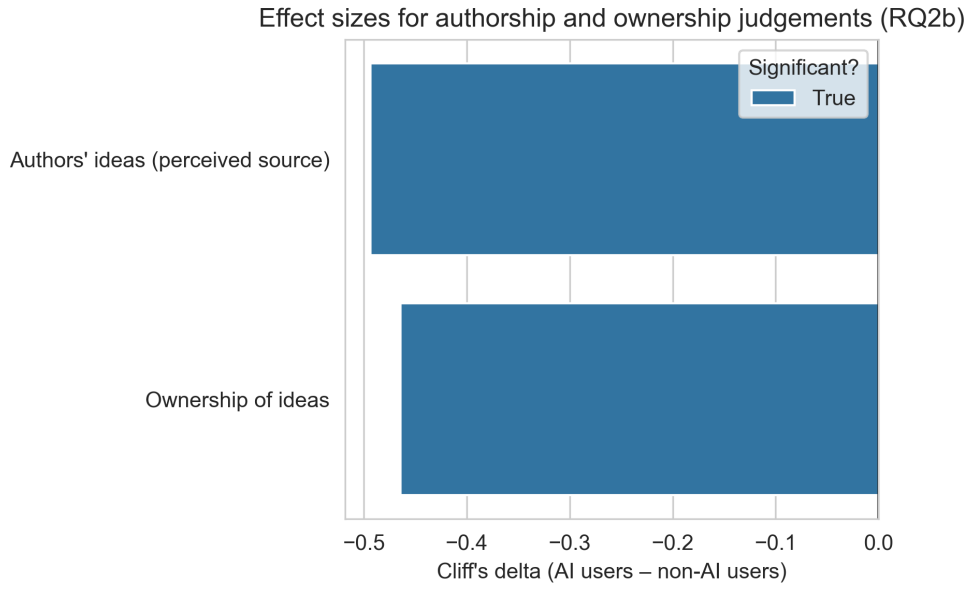

Figure 9. Effect-size (Cliff's delta) plot for RQ2b.
Shows large negative effect sizes.

The "profit-sharing" variable was excluded from inferential analysis because it was only presented for AI-assisted stories and because responses were highly inconsistent, with some participants using values such as 0% to express normative judgements rather than literal profit allocation.

Decision: H_0 is fully rejected.

Disclosure of AI involvement led to large, coherent, and unidirectional penalties in perceived authorship and ownership. Evaluators systematically reduced the human author's creative credit once they learned that AI assistance had been used.

4.4 RQ2c – AI suspicion vs blind ratings

RQ2c examined whether evaluators' later suspicions about AI involvement were associated with the blind creativity ratings they had given earlier, before authorship was disclosed. Spearman's rank-order correlations were computed between the AI-guess variable (0–100%) and all pre-disclosure creativity and stylistic ratings.

The results showed a **consistent but extremely small negative trend**: higher AI suspicion was associated with slightly lower blind ratings across most dimensions ($\varrho \approx -0.10$ to -0.02). Although several correlations reached statistical significance due to the large sample size (N = 3,543), their magnitudes were trivial. These patterns are visible in **Figure 10**, which shows that all coefficients cluster close to zero.

One variable—boring—showed a small positive correlation, indicating that stories perceived as dull were marginally more likely to be misattributed to AI. The effect was equally negligible. The

variable labelled “badly written” was not interpreted because the participant-facing item used the opposite wording (“well written”), making its direction uncertain.

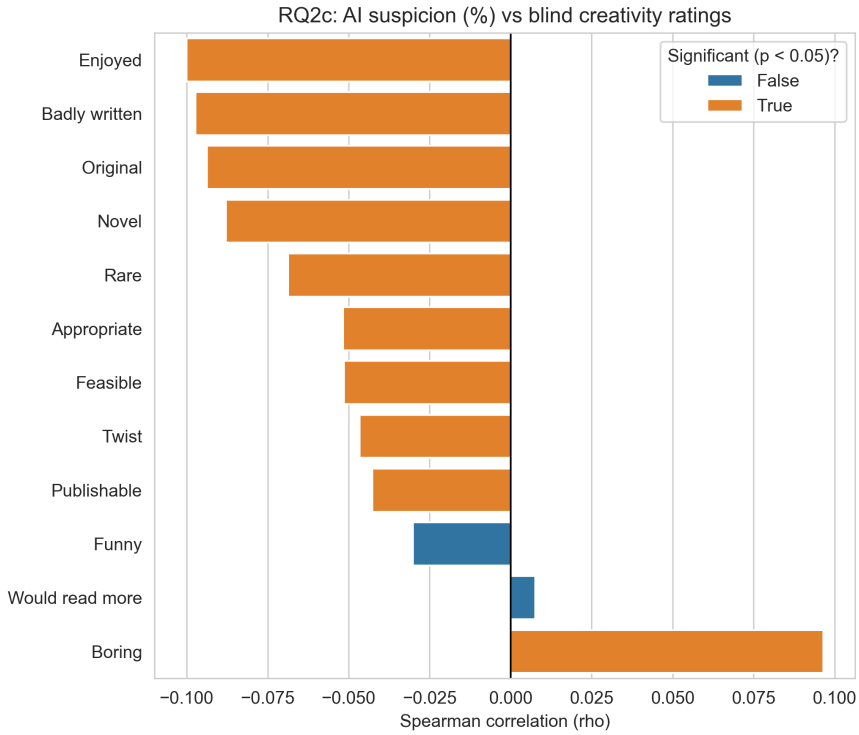

Figure 10. Forest plot: Spearman correlations between AI suspicion and blind ratings (RQ2c). *Shows tiny effect sizes; only “boring” shows a small positive correlation.*

Decision: H_0 is retained.

AI suspicion did not meaningfully influence blind creativity ratings. Although statistically detectable, the correlations were too small to suggest genuine misattribution bias. Evaluative penalties only emerged once AI involvement was disclosed (RQ2b), not during the blind rating phase.

5. Conclusions

Across all analyses, the findings show a clear separation between how AI affects creative output and how it affects perceived authorship.

For **RQ1**, writers who used AI evaluated their own stories similarly to those who did not. Most creativity and quality dimensions showed no differences, and effect sizes were small. Only three variables—authorship, appropriateness, and future-impact—showed modest shifts, suggesting that AI alters how writers view their role but not how they judge the story's creative quality. These results are broadly conclusive, although future work could examine whether the magnitude of AI assistance (e.g., number of ideas used) influences self-perception more strongly.

For **RQ2a**, blind evaluators rated AI-assisted and human-written stories almost identically across all measures. Median ratings were the same, and Mann–Whitney tests detected no meaningful differences. This indicates that AI assistance did not measurably influence perceived story quality when authorship was unknown. These results are highly consistent and conclusive.

For **RQ2b**, strong effects emerged only after disclosure. Evaluators attributed substantially fewer ideas and weaker ownership to authors who used AI, with large and consistent effect sizes. This suggests that disclosure triggers a clear evaluative penalty, reflecting social or normative views rather than differences in textual quality. While conclusive within this dataset, future studies could explore how different forms of disclosure (e.g., partial AI use) shape these judgements.

For **RQ2c**, correlations between AI suspicion and blind ratings were statistically detectable but negligible in magnitude. Blind evaluations were not meaningfully biased by readers' assumptions. This provides strong evidence that misattribution does not occur before disclosure.

Overall, the results support a coherent conclusion:
AI assistance does not significantly affect the quality of stories as judged blindly, nor does it dramatically change writers' own evaluations, but it has a pronounced impact on perceived authorship once readers are informed that AI was used. Open questions remain regarding how transparency, task type, or level of AI assistance might further influence these perceptions, but the core findings of this study are clear and internally consistent.