



An Explainable AI Analysis of Automated Creativity Scoring Models

Ennio Geniblazo and Antonio Laverghetta Jr., MPCC Lab at Wesleyan University

1. Introduction

- **WHAT IS CREATIVITY?** “Creativity,” historically defined, requires both originality and effectiveness (Runco & Jaeger, 2012).
- **EXPERTS ARE EXPENSIVE.** Using experts to rate creativity are expensive and also time-consuming. Automated approaches are effectively “black boxes” whose predictions we do not fully understand.
- **MODELS READ DIFFERENTLY.** Automated scoring models may not use the same evidence as human raters to justify scoring decisions. We need to understand how models rate to ensure fair evaluations.
- **WHY DO WE USE SHAP?** SHapley Additive exPlanations is our method of estimating how each token contributes to the creativity score (Lundberg & Lee, 2017).

Make the logo pink and sparkly then the company will be successful
+0.48 +0.40 +0.53 +0.35 +0.25 +0.43 0 +0.12 +0.03 -0.14 0 -0.40
Prediction: All the values summed=2.48.

2. Methods

- Do tokens at different positions contribute differently to creativity scores?
- Do tokens derived from correctly spelled and misspelled words have different creativity scores?
- Do semantically rare tokens contribute more to the overall creativity score?

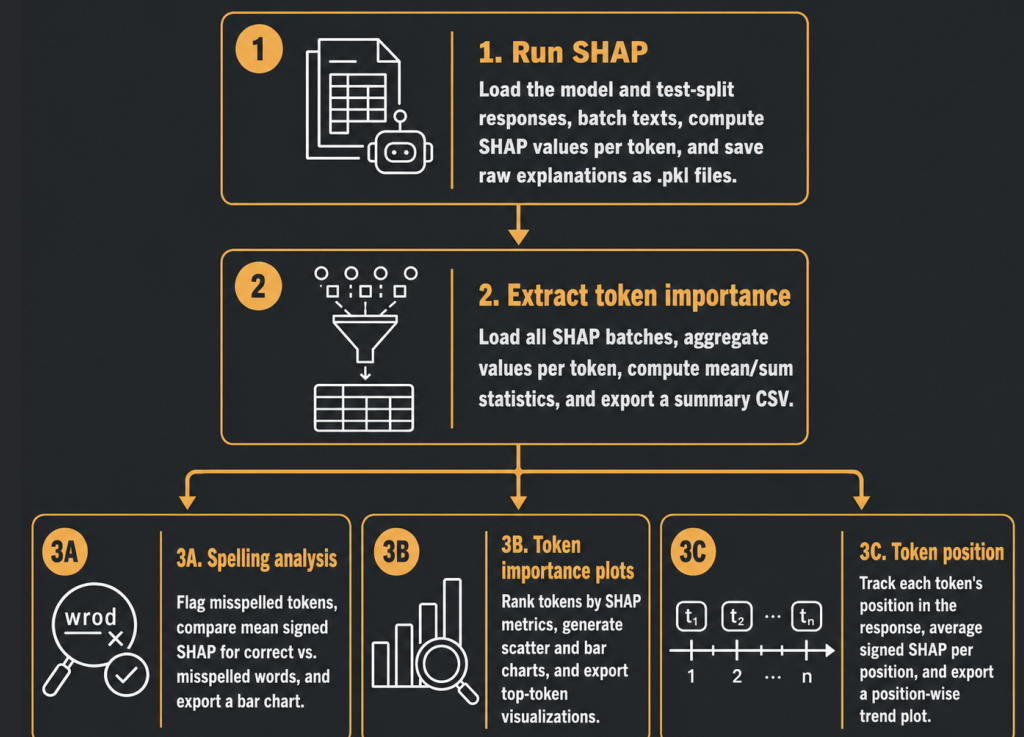
AIDE — Artificial Intelligence for Design Evaluation, n = 7391

MAoSS — Multilingual Assessment of Short Stories, n = 853

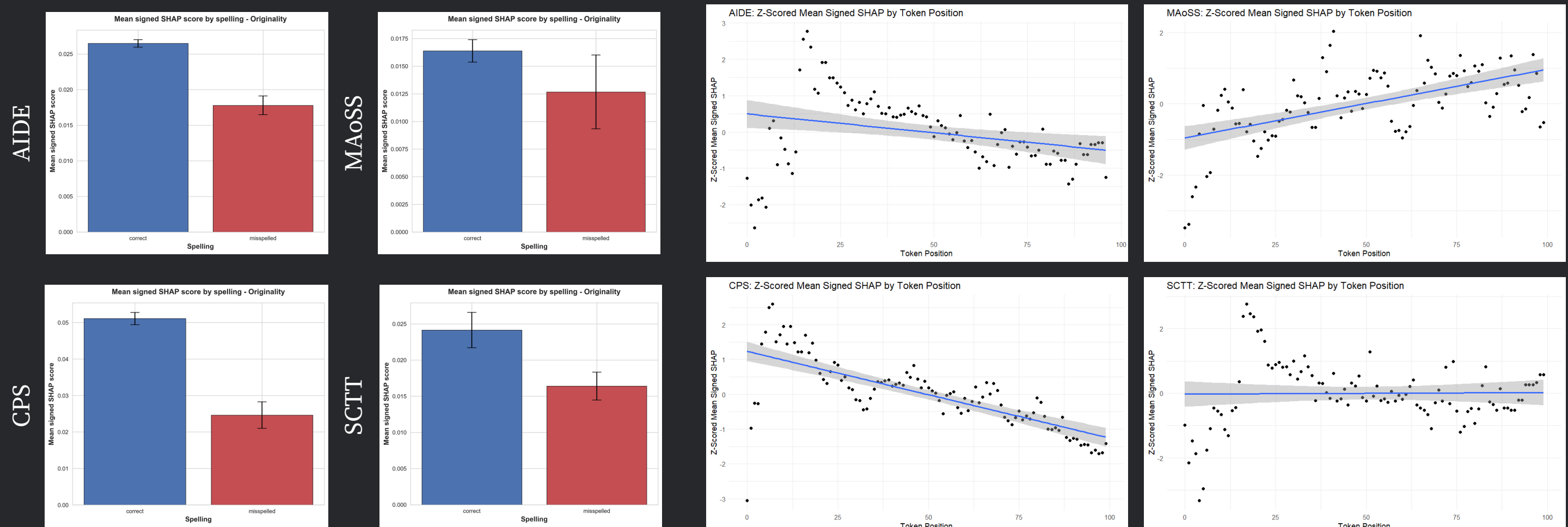
CPST — Creative Problem Solving Task, n = 3056

SCTT — Scientific Creative Thinking Test, n = 18161

SHAP Interpretability Pipeline



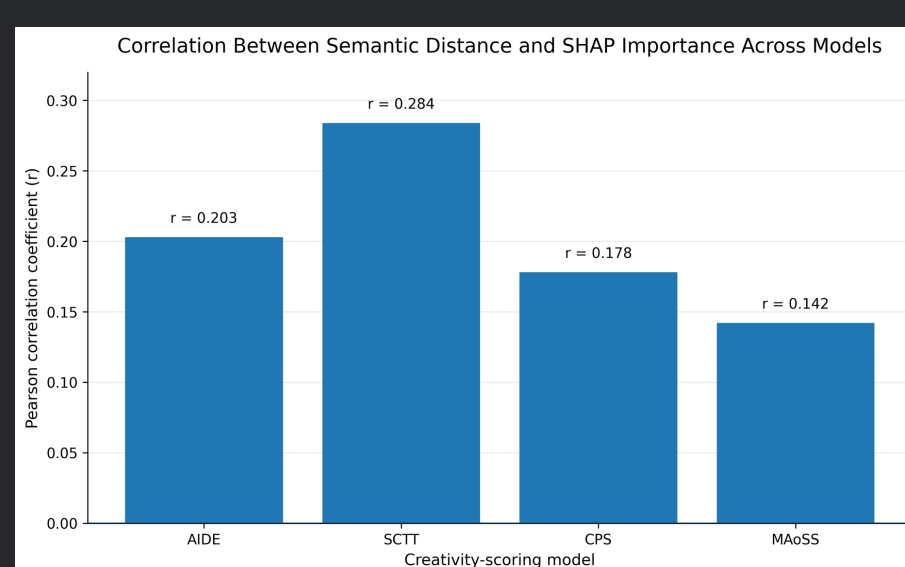
3. Results



- **CORRECTLY SPELLED TOKENS SHOW HIGHER MEAN SIGNED SHAP VALUES** than misspelled tokens across all four models. This suggests that tokens from correctly spelled words tend to contribute more positively to creativity scores.
- **THE SIZE AND CERTAINTY OF THE SPELLING EFFECT VARIES BY MODEL.**
- **ALL MODELS SHOW AN EARLY POSITION BIAS.** In AIDE, CPS, and SCTT, the first few token positions begin relatively low or negative, then rise sharply into positive SHAP values before trending downward.
- **THE MAIN PATTERN IS NOT LINEAR.** Token position effects appear nonlinear; early positions may be unstable, middle positions often show declining positive contributions, and later positions even become negative.

4. Conclusion

- **SEMANTIC DISTANCE WAS POSITIVELY CORRELATED WITH SHAP SCORES** across all four models, suggesting that tokens that are more semantically unusual tended to have stronger influence on creativity predictions.



- **THIS PROJECT FOUND THAT SPELLING, TOKEN POSITION, AND RELATIVE SEMANTIC DISTANCE WERE ASSOCIATED WITH SHAP CONTRIBUTIONS.**
- Correctly-spelled tokens showed higher mean signed SHAP values
- Most models showed an early rise followed by a decline in position-based SHAP values,
- Semantically rarer tokens have positive correlations with SHAP scores.
- These results identify measurable patterns in behavior, but they do not prove causal effects.
- Future work should control for response length, examine larger samples, and test whether these patterns are applicable across other creativity tasks.

Creativity Assessment Platform. (n.d.). Creativity Metrics: Artificial Intelligence scoring models. Pennsylvania State University.

Johnson, D. R., Green, A. E., & Beaty, R. E. (2025). Thematic profile analysis reveals the explanatory power of themes in creative ideas. *Psychology of Aesthetics, Creativity, and the Arts*. Advance online publication. <https://dx.doi.org/10.1037/aca0000822>

Luchini, S. A., Mallikar, N. T., DiStefano, P. V., Laverghetta, A., Jr., Patterson, J. D., Beaty, R. E., & Reiter-Palmon, R. (2025). Automated scoring of creative problem solving with large language models: A comparison of originality and quality ratings. *Psychology of Aesthetics, Creativity, and the Arts*. Advance online publication. <https://dx.doi.org/10.1037/aca0000736>

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, 30.

Orwig, W., Edenbaum, E.R., Greene, J.D. and Schacter, D.L. (2024). The Language of Creativity: Evidence from Humans and Large Language Models. *J Creat Behav*, 58: 128-136. <https://doi.org/10.1002/jocb.636>