Validating Originality Intensity (OI) using an External Validation Approach

EXECUTIVE SUMMARY

The goal of this report is to determine the validity of Originality Intensity (OI), an automated computational index of how far a single text diverges from common AI output, by examining its performance on an independently collected dataset from a peer-reviewed article in *Nature Human Behaviour*. Neither the data nor the external diversity metrics originate with us: the responses were collected by independent researchers, and the diversity metrics were defined by independent researchers. The externally-derived diversity metrics capture how much each individual idea increases or decreases the collective diversity of ideas in the pool – which is the same the goal of the OI metric, but OI is focused on measuring how much an idea expands the diversity of a pool ideas coming from common AI output.

The external validation diversity metrics come from Meincke, Nave, and Terwiesch, in *Matters Arising in Nature Human Behaviour* (2025, vol. 9, 1107–1109). Meincke and colleagues found AI-assisted idea pools were substantially ***less diverse*** than ideas generated by humans unassisted by AI. Critically and paradoxically, expert human raters rated the same AI-assisted ideas as ***more creative*** than ideas generated by humans unassisted by AI. The results showed that while an AI-generated idea may appear creative in isolation, the homogeneity or sameness that comes from AI-assisted brainstorming was completely invisible to human experts. Put simply, using AI helped many individuals generate a creative idea, but they all anchored on the same AI-generated ideas, thereby decreasing the collective diversity of ideas in the pool. Automated computationally-derived metrics are needed to make the invisibility of AI homogeneity visible to humans – this is what OI captures.

External validation results indicated OI reproduced the Meincke and colleagues' diversity metrics results including, the condition ordering was identical, and statistical significance and effect sizes were of comparable magnitude. Both OI and Meincke's diversity metrics showed AI-assisted ideas were substantially less diverse than ideas generated by humans unassisted by AI – despite the fact expert human raters rated AI-assisted ideas higher in creativity. It is important to note that the external diversity metrics must see the entire response pool to be computed, while OI needs only the single text itself. This means the external diversity metrics are always sample-specific and a single text's score is only meaningful when situated within a large pool of other texts. OI can score a single text without an accompanying sample, task prompt, or any additional context – only the text itself is needed. The agreement between OI and external diversity metrics supports using OI as a single-text measurement of how much a text contributes to collective idea diversity, that is, how much it expands the collective idea space.

Is an idea creative because of what it is, or because of where it sits relative to the ideas around it? Torrance's (1968) classic formulation ties originality to the statistical infrequency of a response, which implies that originality is a property of a set as much as of an individual contribution. Meincke, Nave, and Terwiesch (2025) exploit this distinction in their reanalysis of Lee and Chung (2024). Where Lee and Chung scored each idea in isolation and concluded that ChatGPT assistance elevates creativity, Meincke and colleagues scored the *pool*, operationalizing diversity as the conceptual distance between ideas, using metrics focused on cosine-similarity of sentence embeddings. Meincke and colleagues found that ChatGPT assistance decreases the collective diversity of the idea pool. We reanalyze two of these experiments by scoring each idea with Meincke's diversity metrics and with OI. Experiments 2b and 5 were chosen for two reasons. First, both show the same pattern across all three external diversity metrics, which gives an unambiguous criterion to validate against rather than one whose direction depends on which metric is selected. Second, they use entirely different creativity tasks, so agreement in both speaks to generalizability across tasks rather than to the particulars of a single prompt or single creativity task. In experiment 2b, participants invented a toy that a seven-year-old could play with, built from a paper bag, leftover construction bricks, and an unused fan. In experiment 5, participants invented a new use for two broken household objects, a flashlight and a hairspray bottle. Both are open-ended divergent-thinking tasks in which a person must generate a novel, useful idea under fixed object constraints, and in both, the ideas were scored for creativity by expert human raters. All ideas were then scored by OI and 3 external diversity metrics to test for agreement between the two sets of results.

Experiment 2b: human-only, chatbot-assisted, and chatbot-only ideas

The experiment 2b responses (N = 298) reproduce both halves of the paradox. Human raters judged chatbot-assisted ideas substantially more creative than unaided ideas (chatbot-only M = 5.37, chatbot + human M = 5.32, human-only M = 3.76; Figure 1C), while every index of a response's contribution to the collective idea space moves in the opposite direction. Centroid distance was highest for unaided humans (M = 0.267), intermediate under chatbot assistance (M = 0.208), and lowest for chatbot-only ideas (M = 0.119) – see Figure 1A. The ordering repeats for mean pairwise distance (0.458, 0.369, 0.222) and mean minimum-spanning-tree edge distance (0.256, 0.171, 0.097). Originality Intensity (OI) recovers the identical condition ordering (M = 62.55, 4.45, and 0.00 for human only, chatbot + human, and chatbot only, respectively) - see Figure 1B.

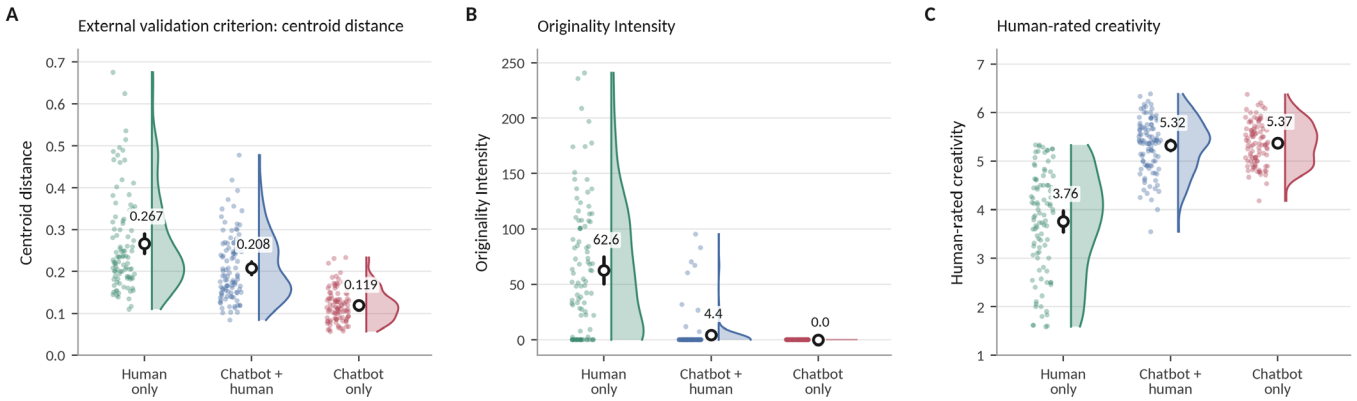


Figure 1. Experiment 2b. Contribution to idea-space diversity (A), Originality Intensity (B), and human-rated creativity (C) by condition. Each panel shows the kernel density of the raw responses (right), every response jittered (left), and the condition mean with its 95% confidence interval (center, open circle). Panels A and B order the conditions identically; panel C reverses that ordering. Every chatbot-only response scores 0 on Originality Intensity, so panel B shows a point mass rather than a density.

Table 1 quantifies the human-only versus chatbot + human contrast, using Cohen's d statistic and p-value from a t-test, which isolates what is lost when a person ideates with an AI model rather than alone. Positive values indicate that the human-only condition scored higher than the AI-assisted condition; df = 197 (see Appendix A5 for estimation details).

Table 1. Experiment 2b: human-only versus chatbot-assisted human.

Measure	*d* [95% CI]	*p-value*	*r* with OI
Human-rated creativity	−1.86 [−2.20, −1.53]	9.8×10^{-29}	−.46
Originality Intensity (OI)	1.34 [1.03, 1.64]	1.2×10^{-17}	1.00
Centroid distance	0.60 [0.32, 0.89]	3.4×10^{-5}	.32
Pairwise distance	1.21 [0.90, 1.51]	4.2×10^{-15}	.44
MST mean edge distance	1.14 [0.84, 1.44]	6.7×10^{-14}	.43

Experiment 5: web search versus chatbot assistance

Experiment 5 (N = 364) replaces the human-only unaided comparison with a google web-search-assisted one, which is the stronger test: both conditions have external help, and only the source of that help differs (AI vs. web-search). The same double dissociation appears. Chatbot-assisted responses were rated by human experts as more creative than Google-assisted responses (M = 5.01 versus 3.77; Figure 2C), yet contributed less to the idea space on every geometric index—centroid distance (M = 0.252 versus 0.370), mean pairwise distance (0.439 versus 0.600), and mean minimum-spanning-tree edge distance (0.215 versus 0.311). Originality Intensity again follows the geometric indices, with 78% of chatbot-assisted responses at the floor, compared to 32% of Google-assisted responses. Humans brainstormed with assistance in both conditions, but only brainstorming with AI reduced the collective idea diversity of responses – according to both external diversity metrics and OI.

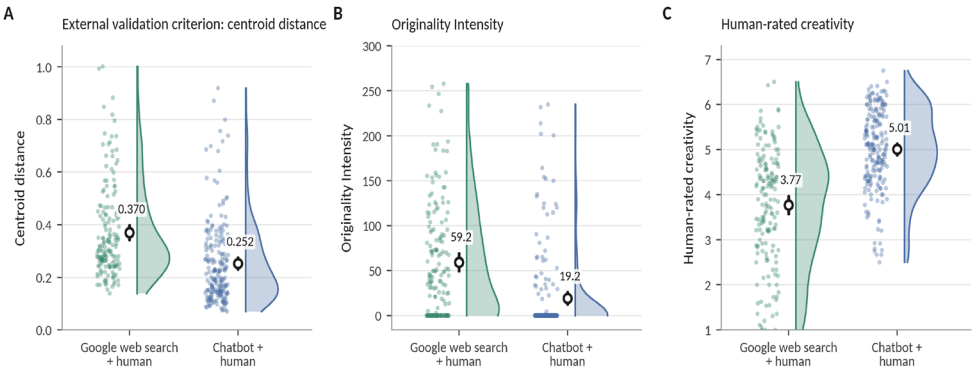


Figure 2. Experiment 5. Contribution to idea-space diversity (A), Originality Intensity (B), and human-rated creativity (C) by condition, plotted with the conventions of Figure 1. The chatbot condition is labeled chatbot + human throughout, since every response in it was produced by a participant working with the model.

Table 2 reports the corresponding effect sizes, $df = 362$; positive values indicate that the Google web search + human condition scored higher than AI-assisted brainstorming.

Table 2. Experiment 5: Google web search + human versus chatbot + human.

Measure	*d* [95% CI]	*p-value*	*r* with OI
Human-rated creativity	−1.15 [−1.37, −0.93]	3.8×10^{-24}	−.21
Originality Intensity	0.73 [0.51, 0.94]	2.3×10^{-11}	1.00
Centroid distance	0.71 [0.49, 0.92]	8.3×10^{-11}	.19
Pairwise distance	1.39 [1.16, 1.62]	1.1×10^{-32}	.28
MST mean edge distance	0.84 [0.62, 1.05]	2.5×10^{-14}	.26

Summary

Originality Intensity (OI) tracks the external diversity indices in both datasets, replicating the ordering of conditions and correlating positively with all three external diversity indices, thereby validating its use as a single-text measurement of idea-space expansion.

Appendix A: Methodology

A1. Creativity task, experiment 2b

Participants were asked to propose a toy that a seven-year-old child could play with, built from three ordinary objects: a paper bag, leftover bricks from construction, and an unused fan. The design has three conditions. In the human-only condition participants generated the idea unaided. In the chatbot + human condition participants worked with ChatGPT and submitted the result of that collaboration. In the chatbot-only condition the model's output was submitted directly. Each participant contributed one idea, and external raters scored it on originality and appropriateness dimensions that were averaged into the creativity score analyzed here.

A2. Creativity task, experiment 5

Participants were asked to propose a new use for two broken household objects, a flashlight and a hairspray bottle, repurposing them into something useful or appealing. Both conditions received external assistance and differ only in its source: participants in the Google web search + human condition used web search, and participants in the chatbot + human condition worked with ChatGPT. The chatbot condition is labeled chatbot + human throughout this report because every response in it was produced by a participant working with the model, not by the model alone.

A3. Inclusion criteria and preprocessing

Originality Intensity requires a minimum amount of text to be estimated reliably, so all responses were filtered to a length of at least ten words before any statistical analysis. The filter was applied once, to the full response pool, prior to computing the diversity metrics, so that every metric is defined over the same set of responses. After filtering, the analyzed samples comprise $N = 298$ responses in experiment 2b (human-only $n = 97$, chatbot + human $n = 102$, chatbot-only $n = 99$) and $N = 364$ in experiment 5 (Google $n = 161$, chatbot + human $n = 203$). No further exclusions were applied.

A4. Semantic diversity metrics

The three diversity metrics follow the definitions Meincke and colleagues used, with one substitution: semantic similarity is computed from MiniLM sentence embeddings (sentence-transformers/all-MiniLM-L6-v2) rather than from Google's Universal Sentence Encoder. Each metric is computed within a condition. That is, the pool against which a response is scored consists of the other responses produced under the same experimental condition of the

same experiment, not of all responses to the prompt. Scoring within condition matches Meincke and colleagues' set-level analysis.

The notation below uses the following symbols. The subscript c identifies a condition; S_c is the set of responses belonging to that condition and n_c is how many responses it contains. The subscript i identifies the single response currently being scored, and j ranges over the other responses in that same condition. Each response is represented by an embedding vector, e_i for response i and e_j for response j. The building block shared by all three metrics is the cosine distance between a pair of embeddings, written $w_{ij} = 1 - \cos(e_i, e_j)$, which is 0 when two responses are semantically identical and grows as they diverge. Every metric is computed once per response rather than once per set.

Centroid distance places the response against the average position of its own condition. The condition centroid $\bar{e}_c$ is the mean of the embeddings in S_c, so a large value means the response sits far from the typical idea produced under that condition:

$$CD_i = 1 - \cos(e_i, \bar{e}_c), \qquad \bar{e}_c = (1/n_c) \sum_{j \in Sc} e_j \qquad (1)$$

Mean pairwise distance measures how far the response sits from each other response in its condition individually rather than from their average, making it sensitive to clustering a centroid can conceal. The sum runs over the $n_c - 1$ responses in S_c other than i itself, since comparing a response to itself would contribute a guaranteed distance of zero:

$$PD_i = 1/(n_c - 1) \sum_{j \in Sc,\, j \neq i} w_{ij} \qquad (2)$$

The third metric uses a minimum spanning tree built within the condition. Treat every response in S_c as a node of a complete weighted graph whose edge weights are the cosine distances w_{ij}, and let T_c be a minimum spanning tree of that graph. Only a response's tree neighbors enter the average, written $N_{Tc}(i)$, the set of responses joined to i by an edge of T_c, with $|\cdot|$ denoting how many such neighbors there are:

$$MST_i = 1/|N_{Tc}(i)| \sum_{j \in NTc(i)} w_{ij} \qquad (3)$$

The tree retains only the $n_c - 1$ shortest edges needed to connect every response in the condition, so a response whose incident tree edges are long is one the rest of its condition can only reach by crossing a large semantic gap. Meincke and colleagues reported this as a set-level robustness check; here it, centroid distance, and pairwise distance are all evaluated per response, which is what allows them to serve as a criterion for a response-level index.

A5. Statistical procedures

Condition means are reported with 95% confidence intervals from the t distribution. Cohen's d uses the pooled within-condition standard deviation, and its confidence interval is obtained by inverting the noncentral t distribution. Significance tests are two-tailed independent-samples t tests assuming equal variances; though Welch-corrected tests give the same conclusions throughout. Correlations are Pearson coefficients pooled across conditions. One caveat applies to Originality Intensity: its distribution has a substantial portion of responses at the floor (i.e., OI = 0) in every condition, so its d values summarize a variable that departs markedly from normality and should be interpreted together with the distributions in Figures 1B and 2B.

Appendix B: The remaining external validation criteria

The main text plots centroid distance as the external validation criterion because it is the most direct expression of how far a response sits from the center of the idea pool. This appendix repeats the analysis for the other two, mean pairwise distance and mean minimum-spanning-tree edge distance, plotted alongside Originality Intensity for direct comparison. The conventions match Figures 1 and 2.

B1. Experiment 2b

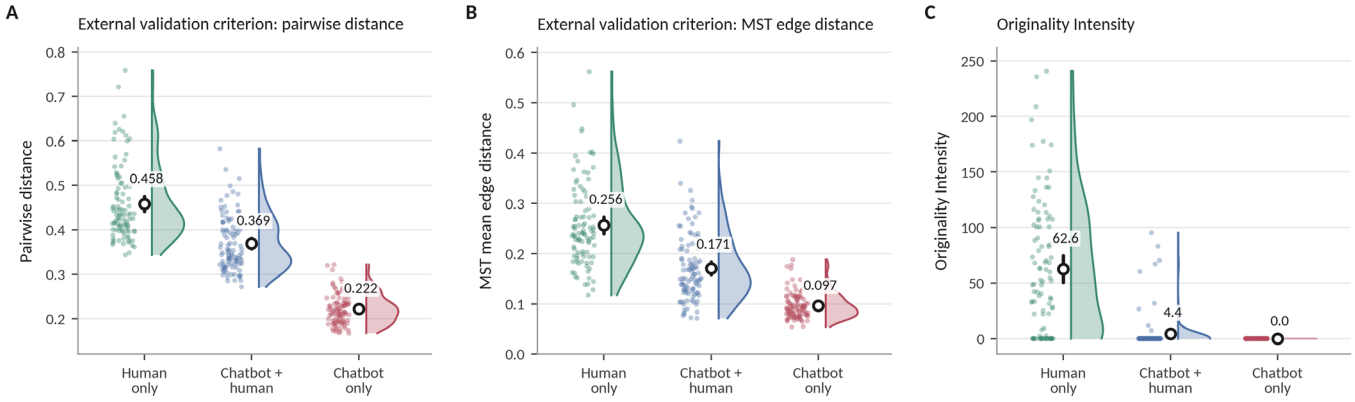


Figure B1. Experiment 2b. The two remaining external validation criteria, pairwise distance (A) and minimum-spanning-tree mean edge distance (B), alongside Originality Intensity (C). All three place the conditions in the same order.

Table B1. Experiment 2b: human-only versus chatbot + human, remaining criteria and Originality Intensity.

Measure	*d* [95% CI]	*p-value*
Pairwise distance	1.21 [0.90, 1.51]	4.2×10^{-15}
MST mean edge distance	1.14 [0.84, 1.44]	6.7×10^{-14}
Originality Intensity	1.34 [1.03, 1.64]	1.2×10^{-17}

B2. Experiment 5

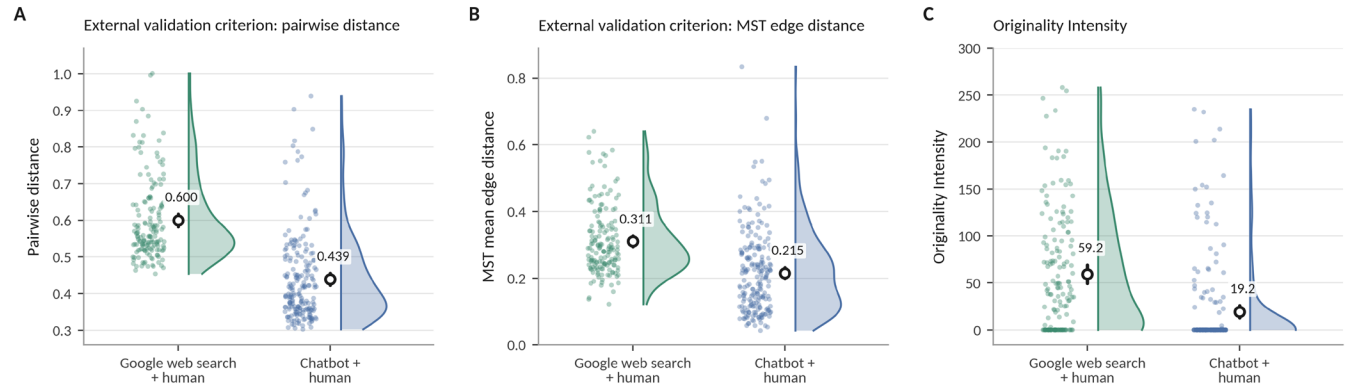


Figure B2. Experiment 5. The two remaining external validation criteria, pairwise distance (A) and minimum-spanning-tree mean edge distance (B), alongside Originality Intensity (C), for the Google web search + human and chatbot + human conditions.

Table B2. Experiment 5: Google web search + human versus chatbot + human, remaining criteria and Originality Intensity.

Measure	*d* [95% CI]	*p-value*
Pairwise distance	1.39 [1.16, 1.62]	1.1×10^{-32}
MST mean edge distance	0.84 [0.62, 1.05]	2.5×10^{-14}
Originality Intensity	0.73 [0.51, 0.94]	2.3×10^{-11}

B3. Interpretation

In each experiment Originality Intensity separates the conditions in the same direction and the same order as both remaining criteria, so the agreement is one of direction and ordering rather than of exact magnitude. Its correlations

with the two criteria are .44 and .43 in experiment 2b and .28 and .26 in experiment 5. Two features of the measures bear on the difference in magnitude. Pairwise and minimum-spanning-tree distance are continuous and unbounded below, whereas Originality Intensity has a hard floor at zero that a large share of chatbot-assisted responses reaches, which limits the separation it can express. The two criteria are also computed from the same embedding space as one another and correlate highly between themselves (.84 in experiment 2b, .87 in experiment 5), so their close agreement is partly shared method rather than independent confirmation.

Appendix C: Text length as a competing explanation

Every measure in this report is computed from text, and the conditions differ sharply in how much text they produced. Text length is therefore a competing explanation for the condition effects in the main text, and the question this appendix addresses is which measures are most susceptible to it and how much of each condition difference could be attributed to length alone. Word count is the number of whitespace-delimited tokens in the response.

C1. Text length by condition

The length difference between conditions is large. In experiment 2b, unaided responses averaged 55 words against 326 for chatbot + human, $d = -2.37$ [−2.73, −2.00], $p = 1.3 \times 10^{-39}$, with an omnibus condition effect of $F(2, 295) = 226.84$, $p = 2.2 \times 10^{-60}$. In experiment 5, Google web search + human responses averaged 46 words against 205 for chatbot + human, $d = -1.40$ [−1.63, −1.17], $p = 3.5 \times 10^{-33}$. Length is not a minor nuisance variable in these data; it is the single largest difference between the conditions.

Table C1. Response length by condition, in words.

Condition	*M*	*SD*	*Mdn*
Experiment 2b: human-only	55.2	31.5	46
Experiment 2b: chatbot + human	325.5	156.3	292
Experiment 2b: chatbot-only	277.9	32.4	279
Experiment 5: Google web search + human	46.0	41.2	36
Experiment 5: chatbot + human	205.2	147.3	230

C2. Impact of text length on each measure, within condition

Correlations between word count and each measure are reported within condition. In experiment 2b no coefficient exceeds .37 in absolute value, and most fall below .20. Originality Intensity has no coefficient in the chatbot-only condition, where every response scores 0 and the measure has no variance. In experiment 5 the associations are larger in the chatbot + human condition, between .43 and .49 in absolute value, than in the Google web search + human condition, where they range from .06 to .23. Across both experiments, Originality Intensity is not systematically more impacted by text length than the three external criteria: it is the least impacted of the four in the human-only condition of experiment 2b and comparable to centroid and pairwise distance elsewhere. Human-rated creativity is the only measure positively related to length in every condition where it can be estimated, meaning longer responses were rated more creative.

Table C2. Experiment 2b: within-condition Pearson correlation between response word count and each measure.

Measure	Human-only	Chatbot + human	Chatbot-only
Human-rated creativity	+.37	+.17	+.00
Originality Intensity	+.07	−.24	—
Centroid distance	−.17	−.00	−.19
Pairwise distance	−.17	−.00	−.19
MST mean edge distance	−.01	+.16	−.09

Table C3. Experiment 5: within-condition Pearson correlation between response word count and each measure.

Measure	Google web search + human	Chatbot + human
Human-rated creativity	+.23	+.43
Originality Intensity	+.06	−.44
Centroid distance	−.19	−.49
Pairwise distance	−.19	−.49
MST mean edge distance	−.13	−.47

C3. Pairwise analyses of covariance with word count as the covariate

Every pair of conditions within an experiment was compared in an analysis of covariance fitted to all conditions of that experiment simultaneously, with word count as a single covariate and a common slope across conditions. The contrast is the difference between two *adjusted* means, each condition's mean evaluated at the grand mean of word count rather than at its own. This is what removes the condition difference in length from the numerator: a condition

whose responses happen to be long is no longer credited or penalized for that length, because every condition is evaluated at the same covariate value.

The residual standard deviation is the denominator. It is computed from the residuals of that same model: for each response, the value the model predicts from its condition and its word count is subtracted from the observed value; these residuals are squared, summed, divided by $N - G - P$ degrees of freedom, and the square root is taken. Because word count is a term in the model, the variance in the outcome that word count accounts for has already been removed from these residuals. The residual standard deviation therefore describes variation among responses that share both a condition and a length, and dividing the adjusted contrast by it expresses the condition difference in units of that length-controlled variability. This quantity is the standardized contrast ψ* of Shieh (2023), whose exact confidence intervals are obtained by inverting the noncentral *t* distribution rather than by a normal approximation; Shieh's simulations show that the exact intervals attain their nominal coverage while the simpler direct-division intervals do not.

The unadjusted column is computed from the same model with the covariate omitted, so the only difference between the two columns is word count and not the choice of denominator. These unadjusted values are consequently not identical to the Cohen's *d* values in Tables 1 and 2, which pool only the two conditions being compared, whereas the denominator here pools every condition in the experiment; the two agree exactly in experiment 5, which has only two conditions. One limitation should be noted. Because the residual standard deviation is smaller than the raw within-condition standard deviation whenever the covariate explains outcome variance, ψ* is not on the same scale as a conventional *d* and should not be compared directly with effect sizes from studies that do not adjust for the same covariate (Olejnik & Algina, 2000).

Table C4. Experiment 2b: standardized contrasts without and with word count as a covariate, v = 295 unadjusted and 294 adjusted. Positive values indicate the first-named condition scored higher.

Measure	Contrast	ψ* unadjusted [95% CI]	ψ* adjusted [95% CI]	*p* adjusted
Human-rated creativity	Human-only vs chatbot + human	−2.14 [−2.47, −1.81]	−1.78 [−2.23, −1.33]	8.2×10^{-15}
	Human-only vs chatbot-only	−2.20 [−2.53, −1.87]	−1.91 [−2.33, −1.49]	2.0×10^{-19}
	Chatbot + human vs chatbot-only	−0.06 [−0.34, +0.22]	−0.13 [−0.41, +0.16]	.377
Originality Intensity	Human-only vs chatbot + human	+1.64 [+1.33, +1.94]	+1.49 [+1.04, +1.93]	5.3×10^{-11}
	Human-only vs chatbot-only	+1.76 [+1.45, +2.07]	+1.64 [+1.23, +2.05]	4.1×10^{-15}
	Chatbot + human vs chatbot-only	+0.13 [−0.15, +0.40]	+0.15 [−0.13, +0.43]	.294
Centroid distance	Human-only vs chatbot + human	+0.71 [+0.42, +0.99]	+0.60 [+0.17, +1.03]	.006
	Human-only vs chatbot-only	+1.78 [+1.46, +2.09]	+1.69 [+1.28, +2.10]	6.8×10^{-16}
	Chatbot + human vs chatbot-only	+1.07 [+0.78, +1.36]	+1.09 [+0.79, +1.38]	6.2×10^{-13}
Pairwise distance	Human-only vs chatbot + human	+1.40 [+1.10, +1.70]	+1.30 [+0.86, +1.74]	7.5×10^{-9}
	Human-only vs chatbot-only	+3.73 [+3.32, +4.14]	+3.64 [+3.15, +4.12]	6.3×10^{-51}
	Chatbot + human vs chatbot-only	+2.32 [+1.99, +2.66]	+2.34 [+2.00, +2.68]	8.2×10^{-43}
MST mean edge distance	Human-only vs chatbot + human	+1.36 [+1.06, +1.66]	+1.61 [+1.16, +2.06]	1.6×10^{-12}
	Human-only vs chatbot-only	+2.54 [+2.19, +2.89]	+2.75 [+2.30, +3.20]	3.0×10^{-34}
	Chatbot + human vs chatbot-only	+1.18 [+0.89, +1.48]	+1.14 [+0.84, +1.44]	4.6×10^{-14}

Table C5. Experiment 5: standardized contrasts for Google web search + human versus chatbot + human, v = 362 unadjusted and 361 adjusted. Positive values indicate the Google web search + human condition scored higher.

Measure	ψ^* unadjusted [95% CI]	ψ^* adjusted [95% CI]	p adjusted
Human-rated creativity	−1.15 [−1.37, −0.93]	−0.76 [−1.02, −0.50]	8.1×10^{-9}
Originality Intensity	+0.73 [+0.51, +0.94]	+0.37 [+0.12, +0.63]	.004
Centroid distance	+0.71 [+0.49, +0.92]	+0.20 [−0.06, +0.45]	.127
Pairwise distance	+1.39 [+1.16, +1.62]	+0.91 [+0.65, +1.17]	9.2×10^{-12}
MST mean edge distance	+0.84 [+0.62, +1.05]	+0.32 [+0.07, +0.58]	.013

C4. Interpretation

In experiment 2b the adjustment changes little. The largest shift is for human-rated creativity, where the human-only versus chatbot + human contrast moves from −2.14 to −1.78, and the minimum-spanning-tree contrasts increase rather than decrease. Every contrast that was statistically reliable before adjustment remains statistically reliable after it. In experiment 5 the adjustment is substantial: every diversity metric falls, and centroid distance is reduced to +0.20 [−0.06, +0.45], no longer statistically reliable, while pairwise distance, minimum-spanning-tree distance and Originality Intensity remain so.

The adjusted values should be read as a lower bound rather than as a corrected estimate. Word count here is not a pre-existing covariate but a product of the manipulation, since chatbot assistance produces longer text by its nature. Conditioning on a variable that the treatment itself causes removes part of the treatment effect along with the confound, so the adjustment is biased against the condition difference. The unadjusted values are correspondingly an upper bound, because nothing in these data separates the claim that AI assistance narrows the idea space from the claim that longer text scores lower on these measures.

For the validation argument specifically, the relevant finding is that Originality Intensity is no more impacted by length than the external criteria it is being validated against are. Length therefore does not offer an explanation of the agreement between them that would not apply equally to the criteria themselves. And it is important to note all responses were of relatively short text lengths, barely over 300 words, so it is also possible for longer text lengths, word count may play an even smaller role in scoring.

References

Lee, B. C., & Chung, J. J. (2024). An empirical investigation of the impact of ChatGPT on creativity. *Nature Human Behaviour, 8*, 1906–1914. https://doi.org/10.1038/s41562-024-01953-1

Meincke, L., Nave, G., & Terwiesch, C. (2025). ChatGPT decreases idea diversity in brainstorming. *Nature Human Behaviour, 9*, 1107–1109. https://doi.org/10.1038/s41562-025-02173-x

Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology, 25*, 241–286. https://doi.org/10.1006/ceps.2000.1040

Shieh, G. (2023). Assessing standardized contrast effects in ANCOVA: Confidence intervals, precision evaluations, and sample size requirements. *PLOS ONE, 18*(2), e0282161. https://doi.org/10.1371/journal.pone.0282161

Torrance, E. P. (1968). A longitudinal examination of the fourth grade slump in creativity. *Gifted Child Quarterly, 12*, 195–199.