Evaluating Creativity Support Tools via Homogenization Analysis

Barrett R. Anderson
Independent Researcher
Santa Cruz, California, USA
barrettrees@gmail.com

Jash Hemant Shah
Santa Clara University
Santa Clara, California, USA
jshah5@scu.edu

Max Kreminski
Santa Clara University
Santa Clara, California, USA
mkreminski@scu.edu

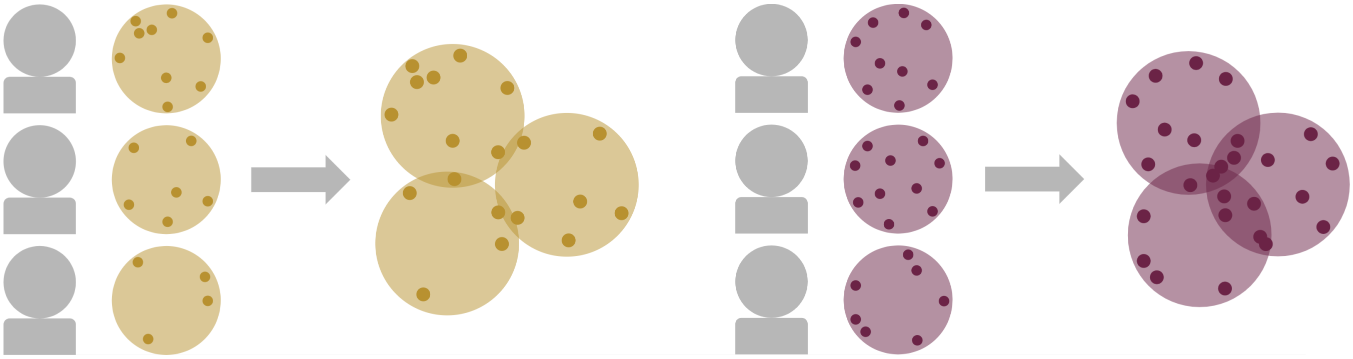

Figure 1: Homogenization analysis involves semantic similarity comparisons between artifacts produced by users of a creativity support tool (CST). Here, users of the CST on the left (in yellow) and the CST on the right (in purple) each produce similarly homogenous sets of artifacts as *individuals*—but collectively, users of the CST on the right produce a more homogenous set of artifacts at the *group* level (as shown by the higher degree of overlap between the sets of artifacts produced by each user).

ABSTRACT

The evaluation of creativity support tools (CSTs)—software systems intended to support human creativity—remains an open problem, in part because creativity is fundamentally difficult to define and in part because different CSTs target a wide variety of different creative domains. We propose a new, general-purpose evaluation criterion for CSTs: namely, the extent to which a CST homogenizes the creative output of its users. We also demonstrate one way to conduct this kind of homogenization analysis, leveraging semantic similarity between embeddings of users' creative outputs to quantify the degree of homogenization that a CST induces.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; • **Applied computing** → *Arts and humanities*; • **Computing methodologies** → Natural language processing.

KEYWORDS

creativity support tools, divergent ideation, large language models, user study

ACM Reference Format:
Barrett R. Anderson, Jash Hemant Shah, and Max Kreminski. 2024. Evaluating Creativity Support Tools via Homogenization Analysis. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3613905.3651088

1 INTRODUCTION

Success in creative contexts (such as creative writing and product design) often hinges on the ability to come up with ideas that are—in line with prominent creativity researcher Margaret Boden's definition of creativity—simultaneously "new, surprising, and valuable" to some extent [7]. Research in *creativity support tools* (CSTs) has long aimed to produce software systems that can support parts of human creative processes [14, 23, 49]. However, evaluating CSTs has been considered an open problem since the founding of the field [27], and recent reviews of CSTs research confirm that the evaluation problem remains one of the thorniest issues facing the field today [46].

Meanwhile, following the incorporation of modern generative AI technologies (such as large language models and text-to-image models) into CSTs, researchers have begun to express concern that the widespread use of a small number of highly centralized, data-driven AI systems may lead to decreased diversity in the outputs of creative processes that incorporate these tools [9, 22, 26, 35]. These concerns resonate with earlier work that has proposed diversity of output as a potential evaluation criterion for AI-based CSTs [10, 33]. Despite these earlier discussions, however, it is only very recently that researchers have begun to directly study the question of whether the use of AI-based CSTs leads to homogenization of human creative output [3, 19, 28, 33, 41].

We propose that the degree of homogenization induced by a CST on its users can be quantified and used as an explicit, cross-domain evaluation criterion for CSTs, mitigating to some degree the

CHI EA '24, May 11–16, 2024, Honolulu, HI, USA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0331-7/24/05.
https://doi.org/10.1145/3613905.3651088

difficulty of comparing CSTs across different creative domains and enabling evaluators to more directly investigate a CST's effects on one fundamental facet of (divergent) creativity. Broadly speaking, this kind of *homogenization analysis* (Figure 1) can be conducted by means of a comparative user study in which participants are asked to produce creative outputs using several different CSTs. These outputs can then be embedded and the semantic similarity between outputs compared by means of cosine similarity between embedding vectors. Finally, these cosine similarity values can be aggregated by CST to quantify the homogeneity of creative outputs produced by users of each CST.

We piloted this approach in the context of a 36-participant comparative user study of ChatGPT [39] and an alternative, non-AI-based CST [21], both of which were used for divergent ideation in two different creative domains (product design and fictional scenario development). Each participant completed four divergent thinking tasks—half with support from ChatGPT, and half with support from the non-AI CST—yielding 1271 ideas in total. Based on the resulting data, we apply homogenization analysis and find that use of ChatGPT leads to greater *group-level* homogeneity, while both CSTs are comparable in terms of *individual-level* homogeneity. This, in turn, suggests that homogenization effects of ChatGPT stem largely from the tool's influencing different users to think along similar lines, rather than from the tool's tendency to promote creative fixation [1, 16, 29, 43] in users. These findings are supported by qualitative evidence from participant survey responses. Furthermore, in Appendix A, we validate the use of our chosen sentence embedding model for homogenization analysis by comparing it to a human baseline in the context of an idea categorization task.

2 BACKGROUND AND RELATED WORK

2.1 Prior Studies of Homogenization

To date, there have been relatively few direct studies of creative homogenization resulting from the use of CSTs. Early results include Arnold et al.'s finding that "predictive text encourages predictable writing" [3] in the context of single-word suggestions given by smartphone keyboards and Kreminski et al.'s evaluation of whether a (rules-based) AI poetry composition tool caused users to produce more or less similar poems over time [33].

A form of homogenization analysis by means of sentence embeddings has been performed in the context of two recent studies of LLM-supported writing. Padmakumar and He evaluate the baseline GPT-3 model [8] versus the instruction-finetuned variant InstructGPT [40] in the context of a short-form argumentative essay writing task and find a homogenization effect from LLM assistance at both the lexical and content levels, but only for the instruction-tuned LLM [41]. Meanwhile, Doshi and Hauser evaluate GPT-4 in the context of short-form fictional narrative writing and observe a similar homogenization effect [19]. However, both of these studies stop short of attempting to distinguish between individual-level and group-level homogenization, and the authors of these articles do not note the potential connection between their work and the evaluation of CSTs in general.

While not addressing homogenization effects directly, several other recent studies examine how LLMs change the writing of humans who use them for writing support. Lee et al. [34] find that LLM support tends to increase the diversity of a user's vocabulary but may reduce their feelings of ownership for the text they produce. Jakesch et al. [28] find that an opinionated LLM-based CST influences the opinions expressed by its users in argumentative writing. Similarly, Bhat et al. [6] find that LLM-supplied next-phrase suggestions may alter the form and content of a human user's writing even when the user dislikes these suggestions, while Roemmele [47] finds that observation of LLM-generated text can shape ideas expressed by human writers even when they do not incorporate LLM-generated text directly into their writing.

Finally, in psychology, there have also been several recent attempts to evaluate creativity via semantic similarity. The SemDis platform [5] uses aggregated word embeddings [42] to determine the semantic distance between an ideation prompt and ideas generated in response to that prompt, ultimately using this as a quantitative proxy for originality: one of the dimensions of creativity evaluated by the Torrance Tests of Creative Thinking [50]. Another, similar platform is provided and validated by Dumas et al. [20]. However, these studies focus on evaluating the creativity of people rather than the effects of CSTs, and they do not investigate the overall homogeneity of sets of creative outputs.

2.2 Evaluating Creativity and CSTs

Evaluation of CSTs has remained an open problem since essentially the beginning of CST research [27], due in part to the ambiguous and multifaceted nature of "creativity" as a phenomenon, in part to the wide range of (sometimes contradictory) user needs associated with different creative contexts, and in part to the lack of a clear consensus around what aspects of CSTs should be evaluated [46]. Broadly speaking, approaches to the evaluation of CSTs can be divided into two categories: those that primarily evaluate aspects of the creative *process* when the CST is used, and those that primarily evaluate the creative *products* that emerge from this process.

On the process side, CSTs are most frequently evaluated by means of subjective self-reports of experience from tool users. The Creativity Support Index (CSI) [12] is a widely used and psychometrically validated survey instrument that attempts to standardize some aspects of this experience reporting process across different CSTs. Other (often bespoke) survey instruments are also deployed in CST evaluation, either as a supplement or an alternative to the CSI (e.g., [13, 25, 31, 32, 52]). Process is sometimes also evaluated via observation of user *actions* during the creative process (e.g., [2, 17, 31, 33, 52]).

On the product side, CSTs can be evaluated by examining the quantity, quality, or other characteristics of the artifacts that their users produce. The Torrance Tests of Creative Thinking (TTCT) [50] evaluate the creativity of test-takers according to four facets of creative output: *fluency*, or sheer quantity of artifacts created; *flexibility*, or quantity of distinct categories of artifacts created; *originality*, or dissimilarity of created artifacts to others' creations; and *elaboration*, or level of detail in created artifacts. These same criteria can also be applied to the evaluation of CSTs by comparing users of two or more different tools along these lines. Notably, the TTCT does not include artifact *quality* as an evaluation criterion; where studies of CSTs have attempted to evaluate output quality, human raters have usually been employed to judge the results (e.g., [15, 18, 24, 36]).

Our comparison of ChatGPT to a non-AI CST makes use of both process and product data, with a particular focus on the assessment of homogenization effects via examination of creative products: the ideas that study participants produce. Homogenization effects are most closely linked to the *originality* dimension of the TTCT; like other parallel studies of homogenization effects [19, 41], we investigate originality primarily by means of semantic similarity, using a well-performing sentence embedding model [45] whose direct predecessors [42] have been found to agree well with human judgments of originality in creativity research [5, 20]. Incidental results on other aspects of creativity (including fluency and self-reported user experience) provide additional support for our findings.

3 METHODS

We conducted a within-subjects experiment to evaluate the effects of using two different CSTs for idea generation: ChatGPT and the Oblique Strategies Deck.

3.1 Participants

3.1.1 Recruitment. Participants were recruited from academic mailing lists, forums, and solicitations posted by the experimenters on social media. Our study protocol and recruitment materials were approved by the Santa Clara University IRB. Participation was incentivized with a $17.50 gift card for a one hour session. Participants were required to have a stable Internet connection, a device capable of screen-sharing, and access to a quiet place for the duration of the session in order to participate in the study. Of the 36 participants originally recruited, three were excluded from all analyses for failure to follow direction (e.g., not using ChatGPT when directed).

3.1.2 Demographics. Our sample included 33 participants, ranging in age from 22 to 44 (M=28.36, SD=6.84), and including 63.63% (n=21) men and 33.36% (n=12) women. We had 39.39% Black or African American (n=13), 36.36% Asian (n=12), and 24.24% (n=8) White participants. Participant occupations included 36.36% students (n=12), 30.30% creative professionals (e.g. game designer, writers, n=10), and 33.33% other professionals (e.g registered nurse, social worker, customer service, n=11). Educational experience included 15.15% high school graduates (n=5), 15.15% participants with some college (n=5), 48.48% college graduates (n=16), and 21.21% participants with a Master's degree or higher (n=7). Experience with text-based generative AI (e.g. ChatGPT) was varied, with 72.72% of participants reporting daily (n=16) or weekly (n=8) usage of LLM tools, and 27.27% of participants reporting that they had used them once a month or less (n=3), or never used them before the study (n=6). Experience with generative art AI tools (e.g. Midjourney, Stable Diffusion) was less common, with 63.63% of participants reporting that they had either never used such tools (n=13), or used them once a month or less (n=9), and 33.33% of participants reporting that they used them about once a week (n=5) or daily (n=6).

3.2 Materials

3.2.1 ChatGPT. ChatGPT is a popular LLM-based tool trained to respond to text instructions [39]. Participants in this study used the versions of ChatGPT 3.5 released on May 3rd 2023 (n=6, 16.6%) and on August 3rd, 2023 (30, 83.3%).

3.2.2 Oblique Strategies Deck. The Oblique Strategies deck, originally created by the artists Brian Eno and Peter Schmidt [21], consists of a collection of cards with prompts designed to support creative work. Example prompts include "Turn it upside down", "Don't avoid what is easy", "Destroy the most important thing", and "How would someone else do it?" We used a web app version of the deck [48] as an alternative CST in our control condition.

3.2.3 Creative Ideation Prompts. Participants were asked to respond to creative ideation prompts for two types of divergent thinking tasks: *Product Improvement* (PI) and *Improbable Consequences* (IC). These prompts included:

- How could you make a stuffed toy animal more fun to play with? (PI)
- How could you make a jigsaw picture puzzle more interesting and engaging? (PI)
- Suppose a great fog has fallen over the earth, and all we can see of people is their feet. What would happen? (IC)
- Suppose gravity suddenly became incredibly weak, and objects can float away easily. What would happen? (IC)

For each prompt, participants were instructed to generate as many ideas as possible and to try to come up with ideas that no one else would think of.

3.2.4 Creativity Support Index. The Creativity Support Index (CSI) is a survey instrument for assessing the ability of a CST to assist a user engaged in creative work [12]. We administered the CSI to capture participant experiences with each CST after they used that CST to complete a creative ideation task.

3.3 Procedure

All experimental sessions were remote-moderated over videoconferencing software. In each session, participants were asked to generate ideas in response to specific prompts, first while using one of the two CSTs (ChatGPT, or the Oblique Strategies deck), and then while using the other tool. Participants were instructed to generate as many ideas as they could, and to try to come up with ideas that no one else would think of.

During the session, participants were encouraged to think aloud, to the degree that they felt doing so would not interfere with their performance. The time was held constant at 8 minutes per prompt, and participants responded to two prompts with each support tool. With each tool, the first prompt asked participants to come up with ideas for improving an existing product. The second prompt asked them to consider an impossible situations and imagine as many possible consequences as they could think of. The order of CSTs and prompts was randomized per participant and balanced across the entire experiment.

After using each tool the participants responded to the Creativity Support Index questionnaire, and indicated the degree to which they felt personally responsible for their output, or that they felt their output came from the tool that they used. Each session concluded with an open-ended discussion of each participant's experience with both ChatGPT and the Oblique Strategies deck.

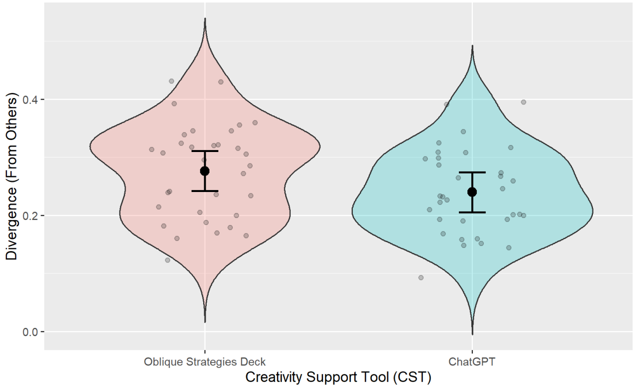

Figure 2: Participant responses were more homogenous at the group level (i.e., more semantically similar to the average embedding of *all* participant ideas) when using ChatGPT.[1]

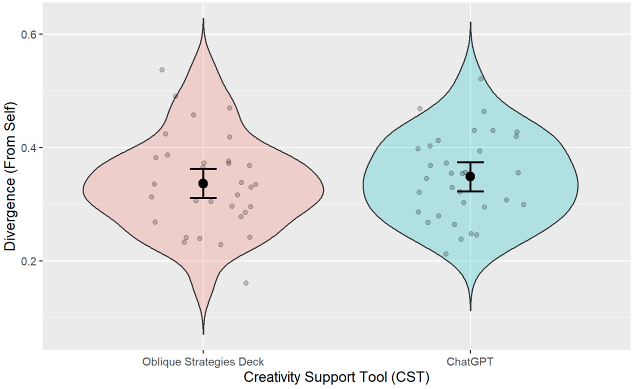

Figure 3: Participant responses were not observably more or less homogenous at the individual level (i.e., more semantically similar to the average embedding of *this participant's* ideas) when using ChatGPT.

4 RESULTS

We report on two categories of data that we gathered about each CST: product and process. Product data includes the list of ideas generated with each CST, which we evaluated for homogeneity as well as for other dimensions of creativity (e.g., fluency). Process data includes Creativity Support Index (CSI) ratings, ratings of how responsible participants felt for the ideas they generated (vs. crediting the CST), and a brief open-ended discussion with each participant about the experience of using each CST, conducted at the end of each experimental session.

4.1 Evaluating Homogenization

Our homogenization analysis was primarily based on semantic similarity assessment via sentence embeddings [45]. Sentence embeddings allow us to quantify homogenization by comparing the similarity of each idea a participant generated to the average embedding of all participant ideas. This enables assessment of homogeneity at the group level. We can also perform a similar analysis on the sets of ideas produced by each individual participant, to gauge homogeneity at the individual level. By assessing both forms of homogeneity in parallel, we can better judge whether a CST's homogenization effects stem from promotion of individual-level creative fixation [1, 16, 29, 43] (i.e., increasing the propensity of users to become stuck in creative "ruts") or from influencing all users to ideate along similar lines.

Our semantic similarity-based approach to homogenization analysis closely follows recent psychological studies of creativity [5, 20]. However, the specific sentence embeddings that we used for our homogenization analysis (though high-performing at semantic similarity tasks in general) have not previously been validated for creativity assessment. As a result, we performed an experiment to validate the agreement of these embeddings with human judgments of semantic similarity on our dataset. See Appendix A for details.

4.1.1 Group-Level Homogenization. When participants used ChatGPT as a CST, the ideas they produced were less divergent from the average embedding of all ideas generated for that task, (M = .24, SD = .07), compared to the ideas that they produced when using a traditional CST, (M = .28, SD = .08), $t(32)$ = 2.154, p = 0.038, d = .47, 95% CI [.00,.07]. See Figure 2. At the group level, ideas produced with the help of ChatGPT were more homogenous.

4.1.2 Individual-Level Homogenization. When participants used ChatGPT as a CST, the ideas they produced were not observably more or less divergent from the average embedding of all of the other ideas that they themselves generated for that task, (M = .65, SD = .07), compared to semantic divergence for ideas that they produced when using a traditional CST, (M = .66, SD = .08), $t(32)$ = .944, p = 0.352, d = .12, 95% CI [-.04,.01]. See Figure 3. At the individual level, we did not observe a difference in homogeneity for ideas generated using ChatGPT.

4.2 Supporting Results

4.2.1 Fluency. Participants generated roughly one additional idea—approximately a 15% increase—when using ChatGPT as a CST (M = 8.39, SD =3.39) compared to the number of ideas they generated when using the Oblique Strategies Deck (M = 7.32, SD = 3.22), $t(32)$ = 2.10, p = 0.044, d = .32, 95% CI [.03,2.11].

4.2.2 Sense of Responsibility. Participants assigned less responsibility to themselves (and more to the tool) for ideas generated while using ChatGPT (M =48.17%, SD =26.22%), compared to ideas generated while using the Oblique Strategies deck (M=63.63%, SD =17.36%), $t(32)$ = 3.21, p = 0.003, d = .67, 95% CI [-24.60%, -5.51%].

4.2.3 Interview Responses. Participants were asked to discuss their own experiences using both tools, and several themes emerged from those discussions (see Table 1). The single most common theme was that ChatGPT felt easier to use, but the Oblique Strategies deck felt more rewarding (27.78%, n = 10). The second most common theme was a positive sentiment regarding the speed and accuracy of ChatGPT responses (25.00%, n = 9). Some participants also reported

[1] All error bars are 95% Confidence Intervals, with Cousineau-Morey (2008) corrections for within-subjects data [37].

Theme	***Example Responses***	***n***	**%**
Effort/Reward Tradeoff (Oblique Strategies) The non-LLM CST was more challenging, but also more rewarding.	*Oblique Strategies got me thinking more creatively, but I got more responses with ChatGPT.* *It was harder to use Oblique Strategies, but it was more fun and it got me to more interesting places.*	*10*	27.78%
Speed/Accuracy (ChatGPT) ChatGPT was fast, and its responses were accurate.	*ChatGPT gave me the right answers.* *Using ChatGPT is a very nice experience... It's very fast and accurate.*	*9*	25.00%
Low Engagement (ChatGPT) Using ChatGPT was less engaging.	*ChatGPT allowed me to turn my brain off. It did more of the heavy lifting.* *ChatGPT reduced the confidence I had to come up with creative things on my own.*	*8*	30.56%
Low Task Relevance (Oblique Strategies) Responses from the traditional CST were less task-relevant.	*I didn't really understand Oblique Strategies. It didn't relate to most of the questions.* *The cards were inspirational, but most of them were just random thoughts.*	*7*	19.44%
Repetitive Responses (ChatGPT) ChatGPT responses were repetitive.	*ChatGPT is a more research-based tool. ChatGPT is a bit repetitive, but it has a lot of data.* *When I asked for more [ChatGPT] repeated half... When I want more, I want different more.*	*3*	8.33%
High Engagement (Oblique Strategies) A traditional CST was more engaging.	*I got into a flow with Oblique Strategies.* *[Oblique Strategies cards] were more interesting than ChatGPT.*	*3*	8.33%
Premature Closure (ChatGPT) The ChatGPT responses became too specific too quickly.	*ChatGPT feels like it can go really specific really quickly. Almost more than you need.* *With ChatGPT, I felt like it was more guided and way more specific.*	*2*	5.56%

Table 1: Reflections on experiences with idea generation using both CSTs (ChatGPT and the Oblique Strategies deck).

finding ChatGPT to be less engaging (22.22%, n =8). A few participants felt that ChatGPT's responses were too repetitive (8.33%, n = 3) and that ChatGPT's responses became too specific too quickly (5.56%, n = 2).

4.2.4 *Creativity Support Index.*

We did not observe any differences in Creativity Support Index (CSI) ratings for ChatGPT (M=78.03%, SD =18.82%) and for the Oblique Strategies deck (M=73.98%, SD =15.35 %), $t(35)$ = 1.028, p = 0.312, d = .24, 95% CI [-3.94%, 12.02%]. We also observed no differences for any of the CSI sub-scales (*Exploration, Engagement, Effort/Reward Tradeoff, Tool Transparency, Expressiveness*). We did not collect responses for the *Collaboration* subscale, which is irrelevant and often omitted in exclusively single-user contexts like that of our study (e.g., [4, 51]).

5 DISCUSSION AND CONCLUSION

The greater group-level homogeneity of ideas produced with ChatGPT, combined with the lack of any significant difference in individual-level homogeneity between ideas produced with either CST, suggests that ChatGPT leads different users to think along similar lines (for instance by suggesting similar outputs in the context of similar prompts) *without* contributing measurably to individual-level creative fixation (i.e., tendency of users to become creatively "stuck"). This is supported by users' lower feelings of responsibility for their creative output when using ChatGPT (i.e., they felt that the ideas belonged more to the CST than to them—perhaps because ChatGPT's output played a larger role in shaping the specific ideas that ChatGPT users suggested) and their anecdotal statements about the subjective differences between the two CSTs: ChatGPT "did more of the heavy lifting" and was perceived as giving "accurate" responses or "the right answers", whereas the Oblique Strategies deck "got [users] to more interesting places".

One potential alternative explanation of the apparent homogenization effect of ChatGPT—that users simply produced more ideas, and therefore necessarily more similar ideas, with ChatGPT—is not borne out by the comparison between individual and group-level homogenization results. Although fluency data does show that ChatGPT users tended to produce slightly more ideas per ideation prompt overall, a fluency effect on homogeneity would be expected to appear at both the individual and group level, but individual-level homogeneity remained consistent across both tools. This is another reason to consider both individual and group-level homogeneity in future homogenization analyses of CSTs.

Limitations of our study include the use of ChatGPT 3.5 (rather than later, better-performing language models) and imperfect correlation between embedding-based and human judgments of semantic similarity. We selected ChatGPT 3.5 both for easy availability to participants (at the time of our study, ChatGPT 3.5 was available to users for free, whereas ChatGPT 4 and other LLMs of comparable quality remained paywalled) and for ecological validity (we expect a majority of LLM users to adopt the most readily available model that they judge to be of acceptable quality). Nevertheless, future studies should likely take improvements in LLM performance into account, as well as potential differences between homogenization effects of instruct-tuned and base models [41]. For a broader discussion of how embedding-based judgments of semantic similarity compare to those of human raters, see Appendix A.

Altogether, we believe that homogenization analysis (as modeled here) should be taken up as a useful additional tool in the CST evaluation toolbox. The CST evaluation problem remains complex and multifaceted, and no one evaluation technique can be expected to serve as a "silver bullet"—but some CSTs (e.g., ChatGPT) can be shown to increase homogeneity of output in some creative contexts, and the ability to assess homogenization effects is important for understanding a CST's overall impact on human creativity. The absence of any clear difference between ChatGPT and the Oblique Strategies deck on any of the criteria evaluated by the Creativity Support Index suggests that homogenization analysis is able to capture an important dimension of creativity that the CSI is not equipped to address. Furthermore, homogenization analysis shows promise as a domain-general technique: although this study specifically made use of sentence embeddings [45] to assess the homogeneity of sets of short texts, a similar approach could be applied in any creative domain for which embedding models exist, including images [44] and audio [30]. For all of these reasons, we argue for the adoption of homogenization analysis in future evaluation of CSTs.

ACKNOWLEDGMENTS

This work was supported by Hackworth Grant GR102981, "Investigating Homogenization of Imagination by Generative AI Models", from the Markkula Center for Applied Ethics.

REFERENCES

[1] Leyla Alipour, Mohsen Faizi, Asghar Mohammad Moradi, and Gholamreza Akrami. 2018. A review of design fixation: Research directions and key factors. *International Journal of Design Creativity and Innovation* 6, 1-2 (2018), 22–35.

[2] Alberto Alvarez, Jose Font, and Julian Togelius. 2022. Toward Designer Modeling Through Design Style Clustering. *IEEE Transactions on Games* 14, 4 (2022), 676–686.

[3] Kenneth C Arnold, Krysta Chauncey, and Krzysztof Z Gajos. 2020. Predictive text encourages predictable writing. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 128–138.

[4] Ilhan Aslan, Katharina Weitz, Ruben Schlagowski, Simon Flutura, Susana Garcia Valesco, Marius Pfeil, and Elisabeth André. 2019. Creativity support and multimodal pen-based interaction. In *2019 International Conference on Multimodal Interaction*. 135–144.

[5] Roger E Beaty and Dan R Johnson. 2021. Automating creativity assessment with SemDis: An open platform for computing semantic distance. *Behavior Research Methods* 53, 2 (2021), 757–780.

[6] Advait Bhat, Saaket Agashe, Parth Oberoi, Niharika Mohile, Ravi Jangir, and Anirudha Joshi. 2023. Interacting with Next-Phrase Suggestions: How Suggestion Systems Aid and Influence the Cognitive Processes of Writing. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 436–452.

[7] Margaret A Boden. 2004. *The Creative Mind: Myths and Mechanisms*. Routledge.

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, Vol. 33. 1877–1901.

[9] Daniel Buschek, Lukas Mecke, Florian Lehmann, and Hai Dang. 2021. Nine Potential Pitfalls when Designing Human-AI Co-Creative Systems. In *Joint Proceedings of the ACM IUI 2021 Workshops*.

[10] Alex Calderwood, Vivian Qiu, Katy Ilonka Gero, and Lydia B Chilton. 2020. How Novelists Use Generative Language Models: An Exploratory User Study. In *Workshop on Human-AI Co-Creation with Generative Models (HAI-GEN 2020)*.

[11] Kathy Charmaz. 2006. *Constructing Grounded Theory: A Practical Guide through Qualitative Analysis*. Sage.

[12] Erin Cherry and Celine Latulipe. 2014. Quantifying the creativity support of digital tools through the Creativity Support Index. *ACM Transactions on Computer-Human Interaction (TOCHI)* 21, 4 (2014).

[13] John Joon Young Chung and Eytan Adar. 2023. PromptPaint: Steering Text-to-Image Generation Through Paint Medium-like Interactions. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*.

[14] John Joon Young Chung, Shiqing He, and Eytan Adar. 2021. The intersection of users, roles, interactions, and technologies in creativity support tools. In *Designing Interactive Systems Conference 2021*. 1817–1833.

[15] John Joon Young Chung, Wooseok Kim, Kang Min Yoo, Hwaran Lee, Eytan Adar, and Minsuk Chang. 2022. TaleBrush: Sketching stories with generative pretrained language models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*.

[16] Nathan Crilly. 2019. Creativity and fixation in the real world: A literature review of case study research. *Design Studies* 64 (2019), 154–168.

[17] Nicholas Davis, Chih-Pin Hsiao, Kunwar Yashraj Singh, Brenda Lin, and Brian Magerko. 2017. Creative sense-making: Quantifying interaction dynamics in co-creation. In *Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition*. 356–366.

[18] Nicholas Davis, Alexander Zook, Brian O'Neill, Brandon Headrick, Mark Riedl, Ashton Grosz, and Michael Nitsche. 2013. Creativity support for novice digital filmmaking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 651–660.

[19] Anil R Doshi and Oliver Hauser. 2023. Generative artificial intelligence enhances creativity. *Available at SSRN* (2023).

[20] Denis Dumas, Peter Organisciak, and Michael Doherty. 2021. Measuring divergent thinking originality with human raters and text-mining models: A psychometric comparison of methods. *Psychology of Aesthetics, Creativity, and the Arts* 15, 4 (2021), 645.

[21] Brian Eno and Peter Schmidt. 1975. Oblique Strategies, hand produced deck of cards. http://www.rtqe.net/ObliqueStrategies/

[22] Ziv Epstein, Aaron Hertzmann, and Investigators of Human Creativity. 2023. Art and the science of generative AI. *Science* 380, 6650 (2023), 1110–1111.

[23] Jonas Frich, Lindsay MacDonald Vermeulen, Christian Remy, Michael Mose Biskjaer, and Peter Dalsgaard. 2019. Mapping the landscape of creativity support tools in HCI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*.

[24] Katy Ilonka Gero and Lydia B Chilton. 2019. Metaphoria: An algorithmic companion for metaphor creation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*.

[25] Matthew Guzdial, Nicholas Liao, Jonathan Chen, Shao-Yu Chen, Shukan Shah, Vishwa Shah, Joshua Reno, Gillian Smith, and Mark O Riedl. 2019. Friend, collaborator, student, manager: How design of an AI-driven game level editor affects creators. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*.

[26] Jeffrey T Hancock, Mor Naaman, and Karen Levy. 2020. AI-mediated communication: Definition, research agenda, and ethical considerations. *Journal of Computer-Mediated Communication* 25, 1 (2020), 89–100.

[27] Tom Hewett, Mary Czerwinski, Michael Terry, Jay Nunamaker, Linda Candy, Bill Kules, and Elisabeth Sylvan. 2005. Creativity support tool evaluation methods and metrics. In *Creativity Support Tools: A workshop sponsored by the National Science Foundation*. 10–24.

[28] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*.

[29] David G Jansson and Steven M Smith. 1991. Design fixation. *Design Studies* 12, 1 (1991), 3–11.

[30] Eunjeong Koh and Shlomo Dubnov. 2021. Comparison and analysis of deep audio embeddings for music emotion recognition. *arXiv preprint arXiv:2104.06517* (2021).

[31] Max Kreminski, Melanie Dickinson, Joseph Osborn, Adam Summerville, Michael Mateas, and Noah Wardrip-Fruin. 2020. Germinate: A mixed-initiative casual creator for rhetorical games. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Vol. 16. 102–108.

[32] Max Kreminski, Melanie Dickinson, Noah Wardrip-Fruin, and Michael Mateas. 2022. Loose Ends: a mixed-initiative creative interface for playful storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Vol. 18. 120–128.

[33] Max Kreminski, Isaac Karth, Michael Mateas, and Noah Wardrip-Fruin. 2022. Evaluating mixed-initiative creative interfaces via expressive range coverage analysis. In *IUI Workshops*. 34–45.

[34] Mina Lee, Percy Liang, and Qian Yang. 2022. CoAuthor: Designing a human-AI collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*.

[35] Isabelle Levent and Lila Shroff. 2023. The Model is the Message. In *The Second Workshop on Intelligent and Interactive Writing Assistants*.

[36] Piotr Mirowski, Kory W Mathewson, Jaylen Pittman, and Richard Evans. 2023. Co-Writing Screenplays and Theatre Scripts with Language Models: Evaluation by Industry Professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*.

[37] Richard D Morey. 2008. Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology* 4, 2 (2008), 61–64.

[38] Dmitry Nikolaev and Sebastian Padó. 2023. Representation biases in sentence transformers. *arXiv preprint arXiv:2301.13039* (2023).

[39] OpenAI. 2023. ChatGPT: A Large-Scale Conversational Language Model. https://www.openai.com/research/chatgpt. Accessed: August 3rd, 2023.

[40] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, Vol. 35. 27730–27744.

[41] Vishakh Padmakumar and He He. 2023. Does Writing with Language Models Reduce Content Diversity? *arXiv preprint arXiv:2309.05196* (2023).

[42] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543. http://www.aclweb.org/anthology/D14-1162

[43] A Terry Purcell and John S Gero. 1996. Design and other types of fixation. *Design Studies* 17, 4 (1996), 363–383.

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[45] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. https://arxiv.org/abs/1908.10084

[46] Christian Remy, Lindsay MacDonald Vermeulen, Jonas Frich, Michael Mose Biskjaer, and Peter Dalsgaard. 2020. Evaluating creativity support tools in HCI research. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. 457–476.

[47] Melissa Roemmele. 2021. Inspiration through observation: Demonstrating the influence of automatically generated text on creative writing. In *Proceedings of the 12th International Conference on Computational Creativity (ICCC '21)*.

[48] Matt Ruten. 2012. Oblique Strategies App. https://obliquestrategies.ca

[49] Ben Shneiderman. 2007. Creativity support tools: accelerating discovery and innovation. *Commun. ACM* 50, 12 (2007), 20–32.

[50] E Paul Torrance. 1966. Torrance tests of creative thinking. *Educational and Psychological Measurement* (1966).

[51] Qian Wan and Zhicong Lu. 2023. GANCollage: A GAN-Driven Digital Mood Board to Facilitate Ideation in Creativity Support. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. 136–146.

[52] Chuan Yan, John Joon Young Chung, Yoon Kiheon, Yotam Gingold, Eytan Adar, and Sungsoo Ray Hong. 2022. FlatMagic: Improving flat colorization through AI-driven design for digital comic professionals. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*.

A VALIDATING SENTENCE EMBEDDINGS FOR HOMOGENIZATION ANALYSIS

Our primary homogenization analysis uses a transformer-based sentence embedding model—`all-MiniLM-L6-v2`, one of the standard general-purpose sentence embedding models provided by the Python SentenceTransformers library [45]—to evaluate the semantic similarity between participant ideas expressed as short strings of text. Our methodology here is similar to that employed in several other recent psychological studies of creativity. Broadly speaking, semantic similarity approaches to creativity research involve the use of some algorithm to produce a numeric score representing the similarity of a pair of creative artifacts; the originality of multiple different artifacts can then be determined and compared relative to a fixed reference point. The SemDis platform [5], a key example of this approach, automates large-scale determination of semantic similarity scores between ideas (expressed as short strings of text) and the creative ideation prompt in response to which these ideas were generated; similarity scores are determined by means of cosine similarity between aggregated word embeddings [42], a metric which has been found to agree well with human judgments of semantic similarity [5, 20].

Aggregated word embeddings are generally outperformed on semantic similarity tasks by more recent transformer-based sentence embeddings, which (unlike aggregated word embeddings) take sentence structure into account. However, the rapid pace of progress in machine learning research means that transformer-based sentence embeddings have not yet been validated against human judgments of semantic similarity in the context of creativity research specifically. Therefore, in order to validate our use of `all-MiniLM-L6-v2`, we conducted a small experiment to determine whether this model agrees strongly with human judgments of semantic similarity on our participant ideas dataset.

Our experiment took a set of human-constructed idea categories as a source of ground truth for semantic similarity judgments and evaluated several candidate embedding models in terms of their agreement with the human coders' manual classification of ideas. Construction of categories followed an iterative grounded theory approach [11]: coders reviewed the ideas generated for each ideation prompt and iteratively formed groups of similar ideas, then refined the ideas over the course of several successive coding passes. This process ultimately generated 181 distinct idea categories from 1271 individual responses. Each participant's responses were then tagged with all relevant idea categories. Coders were kept unaware of which ideas were generated by users of which CST.

Model	IC_A	IC_B	PI_A	PI_B	Average
all-MiniLM-L6-v2	60.94%	58.61%	51.48%	64.63%	58.92%
all-mpnet-base-v2	58.91%	57.08%	51.30%	59.78%	56.77%
GloVe 840B	46.59%	47.25%	41.16%	40.02%	43.76%
Random	4.31%	2.37%	4.14%	2.94%	3.44%

Table 2: Percentage agreement of several different embedding models with human idea categorization judgments. Columns IC_A, IC_B, PI_A, and PI_B report performance on ideas generated in response to specific ideation tasks (IC = "Improbable Consequences", PI = "Product Improvement").

To evaluate the alignment of sentence embedding models with human-constructed idea categories, we first produced a category embedding for each category by averaging together the individual embeddings of the ideas belonging to this category. We then iterated over each idea in the dataset, sorted the category embeddings by their cosine similarity to the embedding of the idea being categorized, and assigned the idea to the *n* categories represented by the *n* most similar category embeddings (where *n* = the number of categories human coders assigned to this idea). To avoid producing artificially high similarity scores between ideas and their actual human-assigned categories across the board, we also excluded each idea's own embedding from the average category embedding when testing similarity to the idea's actual human-assigned categories.

We then compared the model-assigned categories for each idea to the actual categories human coders assigned to this idea, and noted the percentage of overlap between these category sets. Finally, we repeated this process for several different embedding models—as well as a pessimistic baseline "model" that assigned each idea to *n* categories at random—and computed the human-agreement percentage of each model on participant ideas generated in response to each of our four creativity tasks.

Results are reported in Table 2. Notably, our chosen sentence embedding model (`all-MiniLM-L6-v2`) agrees with human idea categorizations more than half the time across all four creativity tasks; it therefore substantially outperforms both GloVe 840B (an aggregate word embedding model previously assessed as state-of-the-art for creativity research [5, 20]) and the random baseline (which GloVe itself beats by more than an order of magnitude). It also consistently outperforms `all-mpnet-base-v2`, the theoretically best overall general-purpose SentenceTransformers model, by a small margin.

Sentence embedding models remain imperfect arbiters of semantic similarity. In our categorization experiment, even the best-performing embeddings model achieved only 59% agreement with human coders on average. There also exists some evidence that cosine similarity between sentence embeddings is more strongly influenced by overlap in the set of nouns than by other similarities [38], suggesting that these models do not take all of the nuances of sentence meaning into account when computing similarity scores. However, the agreement between models like `all-MiniLM-L6-v2` and human judgments of semantic similarity strikes us as high enough to justify the use of these models for homogenization analysis in creativity research, in particular for the increased scale of analysis that these models enable.