CRAFT: Exploring AI as a Critical Partner in Design Problem Solving

Marcus Plach*, Gustav Gutsche, Jan Conrad, Dieter Wallach
Hochschule Kaiserslautern - University of Applied Sciences, Kaiserslautern, Germany
***marcus.plach@hs-kl.de**

Abstract. Generative AI is commonly positioned as an idea generator in design problem solving, although direct suggestions can also anchor designers or reduce their engagement with problem framing. CRAFT explores a complementary role: an AI agent that enters only after unaided ideation and challenges a provisional design through critical feedback. In a three-condition between-participants study, 30 students from 12 academic degree programs completed a time-bounded ill-structured design task involving the development of an interactive, adaptive piece of furniture and then reviewed their concepts either without support (A), with a short generic checklist (B), or with design-specific feedback from CRAFT (C; n = 10 each). Outcomes included think-aloud verbalizations, an adapted version of the Creative Product Semantic Scale (CPSS) rated blindly by two human experts and two independent AI systems, design-change ratings, workload, and a semantic-divergence pretest. The small, heterogeneous sample make this an exploratory, formative study rather than a confirmatory test. No omnibus comparison met the .05 threshold for statistical significance. Nevertheless, several descriptive patterns converged: C produced the highest mean CPSS composite for all four raters and the highest qualitative-change rating for all four raters; participants in C also verbalized more requirements and self-criticism, but fewer self-directed process questions and planning statements. Rater rank-order showed convergence in general but varied substantially. The reported observations do not establish an intervention effect. They instead identify candidate mechanisms and controls for a larger preregistered study of whether deliberately critical AI feedback can support reflective design without taking over ideation.

Keywords: human-AI interaction; design critique; metacognition; reflective practice; generative AI

1 Introduction

Generative AI (GenAI) is rapidly becoming part of creative and engineering work, yet its contribution depends on where and how it enters the process. Current tools are often used to supply concepts, images, or solution fragments during ideation. Such support can improve individual outcomes, but the evidence is mixed. AI-provided story ideas increased judged creativity while making outputs more similar to one another [1]. In a visual design task, exposure to AI-generated examples increased fixation (premature commitment in favor of one solution variant) and reduced fluency, variety, and originality [2]. A study with engineering students likewise found no uniform peer-rated advantage for GenAI-supported ideation despite generally favorable attitudes [3]. Together, these results caution against treating more generated content as equivalent to better design thinking.

Design tasks with an open solution space require more than just the production of alternative solutions. Designers must clarify requirements, revisit assumptions, evaluate trade-offs, and decide what to revise. Comparisons of engineering students and expert practitioners show that expertise is associated with richer problem scoping and information gathering, not merely faster solution generation [4]. This is in line with general notions of reflective practice similarly depending on a conversation with an evolving problem and artifact [5]. Design problem solving therefore creates *metacognitive demands*: users need to formulate intent, monitor their own

reasoning, evaluate model output, and decide when reliance is warranted [6]. These demands also matter in generative AI-based support systems, as language models may accommodate towards a user's stated view even when challenge would be more useful, a behavior studied as sycophancy [7].

Consequently, an alternative design role for AI is to withhold solutions and create productive friction. Cognitive forcing functions can reduce overreliance on AI, although sometimes at a usability cost [8]. In information search, metacognitive prompts have been associated with broader inquiry and more deliberate evaluation [9]. Closest to the present work, Gmeiner et al. used a formative Wizard-of-Oz study with 20 mechanical designers to examine agents that supported intent formulation, problem exploration, and evaluation; their findings are promising, but they also show that support strategy and timing matter [10]. A broader review likewise describes AI as both co-creator and design material, with opportunities for inspiration and evaluation alongside risks of overload and dependency [11].

CRAFT (Critical Agent for Formative Thinking), the AI underlying condition C in the reported study, was developed based on the premise that support strategy and timing are crucial for evaluating potential benefits and harms of AI-supported design. It delays AI involvement until participants have independently generated alternatives, asks for a chosen concept and its rationale, and then offers design-specific criticism and probing questions rather than replacement of ideas. The study asks: What process and product patterns appear when this critical-agent review is compared with unsupported review and a static critique checklist? This investigation is intentionally framed as exploratory and formative. With only ten participants per condition, heterogeneous disciplinary backgrounds, and multiple outcome measures, inferentially stable effects are to be considered unlikely. The aim was thus not to prove efficacy, but to assess feasibility, triangulate tentative patterns across measures and raters, and specify a more discriminating subsequent experiment.

2 Method

2.1 Design and Participants

Thirty students participated in a three-condition between-participants study (n = 10 per condition). Participants were assigned sequentially in a repeating A-B-C order, without stratification or matching based on demographic or educational characteristics. The sample included 22 men and eight women; 15 bachelor's, 14 master's, and one doctoral student from 12 reported degree programs. Computer Science was the largest program (n = 11), while the remaining 19 participants were distributed across 11 programs with no more than four from any one program. On average, the participants were in their 9th semester (M = 8.97 semesters, SD = 3.37); mean birth year was 1999 (SD = 3.83, range 1991-2006). Thus, equal cell size in the conditions should not be mistaken for a homogeneous or demographically balanced sample. The available records do not establish design-domain expertise. The Synthetic Divergent Association Task (S-DAT) [12], a refined and multilingual version of the DAT [13], was administered as a five-minute pretest of semantic divergence, not as a design-expertise measure.

2.2 Task, Apparatus, and CRAFT Agent

Participants addressed a written design brief to develop smart interactive, adaptive furniture. A custom web application synchronized instructions, audio, interaction events, and artifacts. Participants sketched with *tldraw* on an iPad Pro and used a laptop for task content; the

experimenter used a separate computer. Audio was transcribed locally using Whisper-large-v3. The application server used Python, HTTP/REST, Socket.IO, and SQLite.

In condition C, CRAFT processed a participant's explanations and sketches with a locally hosted Qwen3.5-122B-A10B vision-language model. Its role prompt instructed the model to challenge assumptions, identify blind spots and conceptual weaknesses, and ask questions or offer provocations that could guide reflection. It was not intended to generate substitute design solutions. Prior to the actual study, prompt design went through several iterations: critique outputs were tested to ensure an adequate degree of criticality (i.e. critical but not offensive) and the absence of specific design suggestions by the language model until a level adequate in the view of the experimenters was reached for the current study. However, because the study did not include independent coding of each of the agent's utterances, this procedure is not equivalent to a comprehensive manipulation check.

2.3 Procedure and Conditions

After the S-DAT and task briefing, participants thought aloud while developing at least three design alternatives during a 25-minute unaided design phase. They then explained their ideas (sketches and annotations) and were asked to identify an early favorite. The review phase varied by condition (Figure 1). Participants in condition A received no external structure and reviewed their results freely. Condition B used a short list of generic textual hints to guide self-critique (e.g. “have you considered all requirements from the design brief?”). Participants in condition C received design-specific critical feedback from CRAFT on each of their developed design alternatives. Subsequent to reading the critique, participants made a final decision on which alternative they wanted to proceed with and had 15 minutes to iterate it before completing the Raw NASA Task Load Index (NASA-TLX) [14]. Participants again thought aloud during the review and iteration phase, respectively. Review was self-paced in all conditions. Median logged review duration was 6.13 minutes in A, 6.66 minutes in B, and 16.31 minutes in C. The longer duration in C may partly reflect the time required to read and process the AI critique, but it may also represent a meaningful component of the intervention: CRAFT provided additional material for reflection and elicited more sustained engagement.

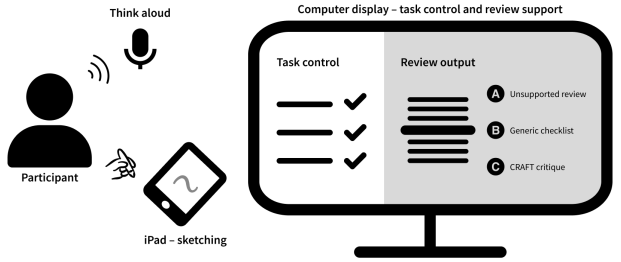


Figure 1. Experimental setup and review conditions. Participants sketched on an iPad while the computer display presented task content and condition-specific review support: either no support in condition A or generic checklist items in B or CRAFT critique in C.

2.4 Measures and Coding

One coder classified think-aloud statements into five categories: self-criticism, self-affirmation, articulation of requirements, self-directed questions, and design planning. These counts were treated as overt indicators of the degree of reflective engagement: higher values suggest that a larger number of distinct considerations became cognitively salient and were verbalized during the task. Each substantively distinct consideration was counted once; repetitions of the same

consideration were not counted again. The resulting counts therefore represent the number of nonredundant considerations articulated within each category rather than the total frequency of utterances. Because the counts were not time-normalized and coding reliability was not assessed, they do have to be interpreted with caution. Favorite-switching recorded whether the final selection differed from the participant's early favorite. Workload used the six 0-20 Raw NASA-TLX items.

Final artifacts were rated blind to condition and in randomized order by two human experts with university and industry experience and by two independent frontier large language models (Claude Opus 5 and GPT 5.6, both in high-effort mode). These systems were separate from the Qwen model used by CRAFT. An adapted Creative Product Semantic Scale (CPSS) [15] yielded common 0-4 scores for Novelty, Resolution, and Elaboration, summed into an additive 0-12 index. Raters also scored quantitative and qualitative change from the initial to the iterated design on 0-4 scales. Human judgments remained the substantive reference; AI ratings were included as additional perspectives, not as ground truth.

2.5 Analysis Strategy

The analysis retained each rater separately because calibration differed. Kruskal-Wallis tests were used for verbalization counts and rater-specific CPSS comparisons, with epsilon-squared as an omnibus effect-size estimate. Pairwise comparisons used Bonferroni correction. Spearman correlations described rank-order convergence among raters and associations with S-DAT scores. Exact p values are unadjusted unless a correction is explicitly stated. The analyses were not preregistered, and the number of exploratory outcomes precludes confirmatory interpretation. Under idealized one-way ANOVA assumptions, N = 30 across three equal groups provides 80% power only for approximately f = 0.60 (eta-squared about .26) at alpha = .05. Accordingly, non-significance is not evidence of equivalence or absence.

3 Results

3.1 Process Measures

No omnibus comparison of the five coded verbalization categories met the .05 threshold (Table 1). Mean self-criticism counts increased from A (4.6) to B (5.7) to C (7.1), $H(2) = 2.98$, $p = .225$, epsilon-squared = .103. Requirement statements were also more frequent in C (4.4) than in A (1.5) or B (1.3), $H(2) = 5.39$, $p = .067$, epsilon-squared = .186. The A–C comparison for requirements articulation yielded an unadjusted $p = .035$ but did not remain significant after Bonferroni correction for the three pairwise comparisons ($p_{adj} = .105$). The comparatively large descriptive difference should therefore be regarded as a formative signal rather than confirmatory evidence. In contrast, self-directed questions were most frequent in B (A = 1.0, B = 2.0, C = 0.6), $H(2) = 4.36$, $p = .113$, epsilon-squared = .150, and planning statements decreased descriptively across conditions (1.2, 0.9, 0.6), $H(2) = 0.86$, $p = .652$. Self-affirmation was similar across conditions (6.6, 6.9, 5.4), $H(2) = 0.51$, $p = .775$.

Nine participants changed away from their early favorite: four in A, three in B, and two in C. Stronger external critique therefore did not descriptively increase abandonment of the initial preference. This could reflect refinement of an initially preferred concept, anchoring, or simple sampling variation; the present data does not allow to distinguish these explanations. Raw NASA-TLX responses suggested moderate mental demand.

Table 1. Descriptive process and product outcomes by condition, incl. raw think-aloud category means, Kruskal-Wallis epsilon-squared estimates, rater-specific CPSS means, and qualitative-change ratings.

Outcome / rater	A	B	C	H(2)	p	epsilon²
Panel A. Think-aloud verbalizations (raw frequency means)						
Self-criticism	4.6	5.7	7.1	2.98	.225	.103
Self-affirmation	6.6	6.9	5.4	0.51	.775	.018
Requirements articulation	1.5	1.3	4.4	5.39	.067	.186
Self-directed questions	1.0	2.0	0.6	4.36	.113	.150
Design planning	1.2	0.9	0.6	0.86	.652	.030
Panel B. Adapted CPSS composite (0–12; means)						
AI Rater 1	6.6	7.2	7.6	—	—	—
AI Rater 2	5.2	5.6	7.0	—	—	—
Human Rater 1	7.0	7.1	7.5	—	—	—
Human Rater 2	5.5	4.6	5.7	—	—	—
Panel C. Qualitative change (0–4; means)						
AI Rater 1	2.0	2.2	2.6	—	—	—
AI Rater 2	1.0	2.0	2.4	—	—	—
Human Rater 1	1.9	2.4	2.7	—	—	—
Human Rater 2	1.4	0.7	1.5	—	—	—

Note. Values are descriptive group means (n = 10 per condition). A = unsupported review; B = checklist-guided review; C = CRAFT. Think-aloud counts are unnormalized raw frequencies. epsilon² = H/(N - 1). Dashes indicate that exact product-test statistics were unavailable in the aggregated analysis.

3.2 Product and Change Ratings

For the CPSS composite, condition C had the highest mean for all raters (Figure 2). Means for A, B, and C were 6.6, 7.2, and 7.6 for AI Rater 1; 5.2, 5.6, and 7.0 for AI Rater 2; 7.0, 7.1, and 7.5 for Human Rater 1; and 5.5, 4.6, and 5.7 for Human Rater 2. The C-A difference was therefore directionally positive in all four rating series, whereas the checklist condition exceeded A for only three of four raters. No rater-specific omnibus test reached $p < .05$. The shared ordering is a descriptive cross-rater pattern, not evidence of a population effect.

Qualitative-change means likewise placed C highest for all four raters: 2.0, 2.2, and 2.6 (AI Rater 1); 1.0, 2.0, and 2.4 (AI Rater 2); 1.9, 2.4, and 2.7 (Human Rater 1); and 1.4, 0.7, and 1.5 (Human Rater 2). For quantitative change, C was highest for three of four raters. These ratings suggest that CRAFT-supported artifacts were often perceived as more substantively revised.

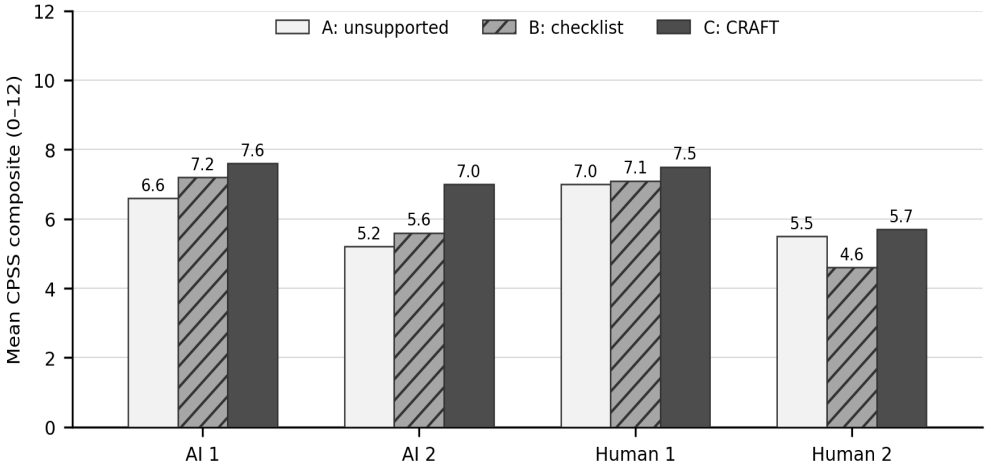


Figure 2. Mean adapted CPSS composite scores (0-12) by review condition and rater. Bars show descriptive group means (n = 10 per condition). The two AI raters were independent of the Qwen model used in condition C; the full-scale axis is retained to avoid exaggerating differences.

3.3 Rater Convergence and Semantic Divergence

Rank-order convergence varied markedly. Spearman correlations were rho = .844 between the two AI raters ($p < .001$), .546 between AI Rater 1 and Human Rater 1 ($p = .002$), and .634 between AI Rater 2 and Human Rater 1 ($p < .001$). Correlations involving Human Rater 2 were smaller (.220-.270, ps >= .148), as was the correlation between the two human raters, rho = .270, $p = .148$. This is evidence about relative ordering, not absolute agreement or interchangeable calibration.

S-DAT scores did not show a clear condition imbalance. Their correlations with CPSS ranged from rho = .260 to .363 across the four raters. Only the association with Human Rater 1 reached the unadjusted threshold, rho = .363, $p = .049$; it would not survive correction across four correlations. The uniformly positive but uncertain estimates are compatible with semantic divergence contributing to performance, while also underscoring that the pretest is not a substitute for baseline design skill.

4 Discussion

This formative study examined a deliberately constrained use of generative AI: critique after independent ideation rather than solution generation during it. The strongest observation is not a significant group difference; none was established. It is the convergence of several descriptive signals despite four differently calibrated raters. Condition C (AI critique) had the highest mean CPSS composite and qualitative-change score for every rater, and it showed more recorded requirements and self-criticism. The pattern is sufficiently consistent to motivate a better-powered study, but it is not sufficiently pronounced to claim that CRAFT improved creativity or reflection.

One candidate interpretation is a shift in the object of reflection. The checklist elicited the most self-directed questions, whereas C produced more statements about the artifact, its requirements, and weaknesses. An interactive critic may have externalized some process control while drawing attention to content-level gaps. This interpretation fits accounts of metacognitive

support as help with intent formulation, exploration, and evaluation [6,10], and with findings that prompts can broaden inquiry [9]. However, the think-aloud coding captured only explicit utterances. The longer self-paced review episodes in C should not be interpreted only as a limitation of the study design. They may indicate that design-specific criticism provided additional reasons to inspect and reconsider the emerging concept and thereby sustained participants' engagement. At the same time, elapsed time includes both reading and active reflection. The present study therefore cannot determine whether the observed outcome patterns were attributable specifically to the critical content, to the additional engagement elicited by CRAFT, or to their combination.

The intervention's timing may nevertheless be theoretically useful. By requiring multiple human-generated alternatives before AI contact, CRAFT avoids placing a salient model-generated example at the start of ideation, where fixation and homogenization risks are most direct [1, 2]. Critical feedback may also preserve authorship better than an agent that supplies complete answers. Yet *favorite-switching* was least common in C, which is ambiguous. Participants may have repaired an initial concept in response to challenge, or they may have remained anchored despite it. Before concluding that CRAFT reduces fixation or preserves designers' agency, future studies should examine whether participants explore genuinely different types of ideas, how substantially successive design variants differ from one another, and how concepts evolve throughout the revision process.

The checklist condition with generic review hints provides a useful but incomplete comparator. Its CPSS mean exceeded unsupported review for three of four raters, suggesting that structure alone may matter. CRAFT combined personalization, interactivity, multimodal access to sketches, different wording, and substantially more elapsed time. The study cannot attribute the C pattern specifically to criticality. A factorial follow-up should separate generic versus design-specific content, and static versus conversational delivery while matching exposure time.

A further question is whether CRAFT's value depends on design experience. Because experienced designers typically engage more extensively in problem scoping and information gathering [4], novice designers may benefit especially when a critical agent externalizes review questions that are not yet part of their routine. More experienced designers may obtain less added value from CRAFT because some review routines are already internalized, although they may be better able to judge and selectively use challenging feedback. Future studies should therefore measure design experience directly and test it as a moderator, for example by comparing novice and advanced designers within each review condition.

The rater results provide convergent as well as complementary perspectives. The two independent AI raters showed strong rank-order agreement and each converged moderately with Human Rater 1, whereas Human Rater 2 used a stricter, more compressed scale and produced a less similar ordering. Despite these calibration differences, condition C obtained the highest mean CPSS score from all four raters – a noteworthy directional convergence in a formative study. Because rank correlations capture relative ordering rather than absolute agreement, the ratings should not be regarded as interchangeable. At the same time, the blinded assessment by two human experts and two AI systems that were separate from the intervention model provides useful triangulation and avoids direct self-evaluation by CRAFT. Future work could consolidate this approach through a preregistered rubric, a larger human panel, and absolute-agreement indices alongside rank correlations.

The findings should be interpreted within the intended formative scope of the study. The sample was small for a three-condition, multi-outcome design and heterogeneous in discipline, degree level, and likely design experience. This heterogeneity provided an initial view across

different student backgrounds but limits the precision of condition estimates, and the use of a single furniture task constrains generalization to other design contexts. Statistical sensitivity was limited to very large effects; accordingly, non-significant results should not be interpreted as evidence of no condition differences. The S-DAT captured semantic divergence rather than task-specific design expertise. Importantly, the think-aloud counts represented distinct, nonredundant considerations, as repetitions of the same point were not counted again. They therefore provide an indicator of some degree of reflective engagement. Because the counts were not time-normalized and coding was conducted by one coder, they describe the total articulated breadth within the observed episode rather than the rate or qualitative depth of reflection, and classification reliability was not independently estimated. The adapted CPSS and change ratings combined two blind human assessments with two independent AI perspectives. Agent responses were not separately coded for fidelity to the intended critical role, and the study focused on immediate process and artifact outcomes rather than longer-term learning or transfer.

A confirmatory follow-up could build directly on these findings by preregistering its hypotheses, sample-size rationale, and analysis plan. In addition, recruiting a larger, more homogeneous or explicitly stratified sample would provide more statistical power. Baseline design experience should be measured directly alongside semantic divergence. Because longer self-paced engagement may be one pathway through which CRAFT operates, review duration need not be treated solely as a variable to be matched. Future studies could instead distinguish reading time, interaction time, and active reflection. In addition, process analysis could be strengthened through independent coding of the rate and depth of reflection. Product assessment would benefit from a larger blind human rater panel with formal reliability estimates. A later, unsupported design task could test whether participants retain and transfer the critical-review strategies prompted by CRAFT, rather than benefiting only while CRAFT is present.

5 Conclusion

CRAFT offers a plausible alternative to AI-led ideation: let people generate first, then use AI to question what they may have overlooked. In this exploratory study, the critical-agent condition showed a consistent descriptive advantage in product and qualitative-change means across four raters and a different profile of verbalized review behavior. However, no inferentially secured effect could be substantiated. The appropriate contribution is therefore formative. The study identifies delayed, design-specific critique as a promising design hypothesis and specifies the controls, fidelity checks, and measurements required to test it credibly.

References

1. Doshi, A. R., & Hauser, O. P. (2024). Generative AI enhances individual creativity but reduces the collective diversity of novel content. Science Advances, 10, eadn5290. https://doi.org/10.1126/sciadv.adn5290
2. Wadinambiarachchi, S., Kelly, R. M., Pareek, S., Zhou, Q., & Velloso, E. (2024). The effects of generative AI on design fixation and divergent thinking. In Proceedings of CHI 2024. https://doi.org/10.1145/3613904.3642919
3. Balta-Salvador, R., El-Madafri, I., Braso-Vives, E., & Pena, M. (2025). Empowering engineering students through artificial intelligence: Blended human-AI creative ideation processes with ChatGPT. Computer Applications in Engineering Education, 33, e22817. https://doi.org/10.1002/cae.22817

4. Atman, C. J., Adams, R. S., Cardella, M. E., Turns, J., Mosborg, S., & Saleem, J. (2007). Engineering design processes: A comparison of students and expert practitioners. Journal of Engineering Education, 96(4), 359-379. https://doi.org/10.1002/j.2168-9830.2007.tb00945.x
5. Schön, D. A. (1983). The Reflective Practitioner: How Professionals Think in Action. Basic Books.
6. Tankelevitch, L., Kewenig, V., Simkute, A., Scott, A. E., Sarkar, A., Sellen, A., & Rintel, S. (2024). The metacognitive demands and opportunities of generative AI. In Proceedings of CHI 2024, Article 680. https://doi.org/10.1145/3613904.3642902
7. Sharma, M., et al. (2024). Towards understanding sycophancy in language models. In Proceedings of ICLR 2024. arXiv:2310.13548
8. Bucinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1). https://doi.org/10.1145/3449287
9. Singh, Y. D., Guan, Y., & Rieh, S. Y. (2025). Enhancing critical thinking in generative AI search with metacognitive prompts. Proceedings of the Association for Information Science and Technology, 62, 672-684. https://doi.org/10.1002/pra2.1287
10. Gmeiner, F., Luo, K., Wang, Y., Holstein, K., & Martelaro, N. (2025). Exploring the potential of metacognitive support agents for human-AI co-creation. In Proceedings of DIS 2025. https://doi.org/10.1145/3715336.3735785
11. Yu, W. F. (2025). AI as a co-creator and a design material: Transforming the design process. Design Studies, 97, 101303. https://doi.org/10.1016/j.destud.2025.101303
12. Haase, J., Hanel, P.H.P., Pokutta, S. (2025) S-DAT: A Multilingual, GenAI-Driven Framework for Automated Divergent Thinking Assessment. Proceedings of the Eighth AAAI/ACM Conference on AI, Ethics, and Society (AIES-25) - Main Track II / Main Track II, Vol. 8, Nr. 2. https://doi.org/10.1609/aies.v8i2.36622
13. Olson, J. A., Nahas, J., Chmoulevitch, D., Cropper, S. J., & Webb, M. E. (2021). Naming unrelated words predicts creativity. Proceedings of the National Academy of Sciences, 118(25), e2022340118. https://doi.org/10.1073/pnas.2022340118
14. Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX: Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), Human Mental Workload (pp. 139-183). North-Holland. https://doi.org/10.1016/S0166-4115(08)62386-9
15. O'Quin, K., & Besemer, S. P. (1989). The development, reliability, and validity of the revised Creative Product Semantic Scale. Creativity Research Journal, 2(4), 267-278. https://doi.org/10.1080/10400418909534323