Proceedings of the ASME 2026
International Design Engineering Technical Conferences and
Computers and Information in Engineering Conference
IDETC-CIE2026
August 23-26, 2026, Houston, TX

DETC2026-193771

AI-ASSISTED IDEATION IN MISSION-CRITICAL ENGINEERING: A PILOT STUDY

J. Justin Klingenberger
University of Virginia
Charlottesville, VA

Jon Rask
NASA Ames Research Center
Mountain View, CA

Eric Reynolds Brubaker
Virginia Tech
Blacksburg, VA

Reid Bailey
University of Virginia
Charlottesville, VA

ABSTRACT

This pilot study examines how access to a large language model (LLM) influences professional engineers' ideation on a mission-critical aerospace design task. Participants generated concepts for landing a fixed-wing aircraft on Mars, either with or without LLM support. Outcomes were assessed using idea counts, NASA SME ratings of desirability, viability, and feasibility (DVF), design-space coverage, process timeline coding, and post-session surveys. Across 8 participants and 66 ideas, AI-assisted ideas scored higher on viability, feasibility, and composite DVF at the idea level, while the manual group produced more ideas and broader category coverage. Process coding showed that AI-assisted participants shifted time away from ideating and toward prompting and reviewing. Perception differences were limited, though qualitative responses suggested text-modality friction and lower process enjoyment in the AI-assisted condition. Given the small sample and nested data structure, findings are preliminary and motivate a larger follow-on study.

Keywords: AI-assisted Ideation, Design Process, Mission-Critical Engineering, Engineering Design

1. INTRODUCTION

The advent of widespread access to powerful generative Artificial Intelligence (AI) has begun to reshape a wide range of industries [1,5,10,11,37]. Large Language Models (LLMs), capable of understanding and generating human-like text, are no longer confined to natural language tasks but are increasingly being leveraged as active partners in a variety of creative fields [6,18,19,34-36,39-43]. Recent studies have demonstrated the potential of LLMs to support a wide spectrum of design activities, including early-stage brainstorming and concept generation [7,18,19,34,35,40-42]. Recently, they have also been explored as proxy interview participants for eliciting user needs in design research [38]. This rapid adoption is driven by the promise of multiple compounding benefits: enhanced efficiency and faster time to concept, the ability to explore a broader design space, access to design support for practitioners across a wide range of skill and expertise levels, and the potential to amplify individual creative output [7,10,11,18,19,34,37,40]. As these tools become more accessible, deployed, and integrated into professional workflows, understanding their impact on engineering design in practice is not merely an academic curiosity but a practical necessity.

A substantial body of research posits that by acting as an inexhaustible source of inspirational stimuli, AI tools can enhance divergent thinking, helping designers generate a greater volume and variety of ideas [2,10,18,19,32-35,37,40-42]. This is particularly valuable in breaking through creative blocks and encouraging the exploration of unconventional solution paths. Conversely, a growing number of studies urge caution, highlighting the potential cognitive pitfalls of human-AI collaboration. A primary concern is the phenomenon of design fixation, where exposure to initial ideas narrows the designer's cognitive search, leading to less original or varied outcomes [12,15,18]. Furthermore, research suggests that while AI may enhance the creative output of individuals, it can paradoxically reduce the collective diversity of ideas among groups as individuals converge on similar themes suggested by the AI [34]. This tension between AI as a creativity enhancer and as a potential influencer of design fixation and convergence motivates the need for more empirical investigations that synthesize and extend emerging studies to clarify these competing effects across tasks and contexts [7,11,15,18,19,34,40].

To begin understanding the practical utility of AI in engineering ideation, the choice of context is critical. Complex aerospace projects require both domain and engineering knowledge due to their uniquely demanding constraints around technical feasibility, mission success, and resilience to extreme

This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

environments. Crucially, professional engineers bring specialized domain expertise and real-world engineering judgment that differs substantially from the student populations that dominate the existing literature on AI-assisted ideation, making this context particularly valuable for understanding how AI tools function under professional, mission-critical conditions [29,37]. These projects provide a more ecologically valid testbed to evaluate how an LLM can be applied to ideation for engineering design over student population studies [6,16,17,40]. This study utilizes NASA's Mars Aerial Explorer (MAE) landing challenge as its design prompt, a representative pre-Phase A conceptual design problem that demands solutions that are both highly innovative and fundamentally practicable.

Due to the professional mission-critical context, we focus the evaluation framework on idea quality and on how the ideation process unfolds [4,13,14,21,24]. Idea quality is assessed by Subject Matter Experts (SMEs) on three dimensions: Desirability, Viability, and Feasibility. The allocation of time across ideation activities, which provides insight into whether AI access shapes the behavioral process of design is tracked using process timelines. Full details are in the Methods section.

The overall aim of this research is to characterize how access to an LLM influences both the outcomes and process of ideation among professional engineers. To address this aim, this paper reports on a controlled pilot experiment comparing AI-assisted and human-only ideation in a high-constraint aerospace design context. The resulting empirical evidence informs both academic theory on Human-AI design collaboration and the integration of AI tools into professional engineering workflows.

1.1 Research Questions

This pilot study on the use of LLMs for creative tasks is guided by research questions organized into two primary areas of focus: outcomes of ideation and the process of ideation.

Ideation Outcomes Research Questions (RQs)

- RQ1 (Quality): To what extent do AI-assisted ideas differ from manually generated ideas in SME rated Desirability, Viability, and/or Feasibility [4-6,16,17,21,23]?
- RQ2 (Design Space): To what extent do AI-assisted ideas span a different portion of the idea space than manually produced ideas [12,15,18]?

Ideation Process Research Questions

- RQ3 (Process): How do professional engineers using an LLM allocate their time during ideation activities [2,8,19,20,28-30,37,40]?
- RQ4 (Perception): How do participants perceive the usefulness and usability of AI, and how do these perceptions relate to their ideation outcomes [2,7,8,11,22,31,37,40]?

2. LITERATURE REVIEW

This research builds upon existing work in two interconnected areas: AI-assisted design ideation (Section 2.1), which draws on both engineering design and human-computer interaction (HCI) research, and the measurement of ideation outcomes (Section 2.2).

2.1 AI In Ideation in Design

Prior, empirical studies suggest that LLMs can be used to broaden the space of candidate concepts and inspire creativity [7,10,16,18,19,32-35,37,40-42]. However, the effectiveness of these tools depends heavily on how designers interact with them. In a comparative study of LLM-generated and human crowd-sourced solutions to engineering design problems, Ma et al. [34] found that although LLMs produced a comparable number of concepts, those concepts clustered more tightly in the design space than human solutions, exhibiting lower semantic diversity and a narrower spread of functional strategies. This tendency is consistent with broader findings that users often anchor on early model outputs, leading to convergence and premature narrowing unless they intentionally prompt for variation [18,19,41]. Structured prompting, such as specifying stakeholder roles, constraints, system-level requirements, or desired exploration strategies, has been shown to increase novelty, technical plausibility, and design-space coverage. In contrast, unstructured or limited prompts tend to yield generic or infeasible concepts [2,8-10,20,37,42]. Studies in engineering design [18,19] and HCI [43-45] demonstrate that prompt specificity, constraint framing, and iterative refinement improve the usefulness of LLM outputs for creative work. Interaction studies [34, 41] further show that designers oscillate between accepting, rejecting, or re-prompting AI tool suggestions, with these interaction modes producing distinct fixation, exploration, and divergence-convergence patterns. In an interview-based study of engineering and design professionals, Ma et al. [40] identified three dominant strategies: generation, evaluation, and domain translation, through which practitioners integrate generative AI into their workflows, each reflecting different balances of human initiative and AI contribution leading to qualitatively different design outcomes. Collectively, the literature indicates that how designers structure their prompts, select interaction strategies, and iterate on AI responses are central drivers of whether LLMs broaden or restrict the design space.

The evaluation of ideation is often concerned with both the quantity and quality of ideas [4,13,21,24,31]. Recent studies suggest that while AI can increase individual creativity and output volume, it may also shift the distribution of ideas across solution categories, that is, either cluster around similar concepts or span a broader range of approaches [7,10,19,34]. This is a critical consideration during concept exploration and screening [4,13,14,21,24,34] because solutions that appear individually strong may collectively cluster around similar technical approaches, limiting the breadth of the design space explored and reducing the probability of identifying truly novel solutions or stalling in less productive areas of the design space.

Recent work extends this support beyond ideation into structured codesign processes [37,40] and even multi-agent pipelines that connect natural language inputs to CAD models [36,43], or to systems-level combinatorial optimization tasks such as Design Structure Matrix sequencing [39]. These broader

This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

studies demonstrate that the role of LLMs in engineering design exists far beyond early-stage brainstorming.

However, there are identified gaps and challenges. Much of the existing evidence comes from studies with students or non-experts [5-7,11,18,19, 37], while firsthand evidence on how AI specifically changes idea quality and the user experience in professional engineering settings remains comparatively rare. Notable exceptions include Ma et al. [40], who interviewed engineering and design professionals and identified the three dominant integration strategies, generation, evaluation, and domain translation, mentioned above. This research vector was continued in Ma et al. [34] where it was found that while LLM-generated solutions matched human-crowdsourced ones in quantity, they clustered more tightly in the design space. This study aims to further address this gap by focusing on practicing engineers in a mission-critical context. Furthermore, reliance on AI tools is not without risk. Studies have warned that AI-assistance can increase design fixation, lead to premature abandonment of promising paths, or add unnecessary complexity to a design if used without considered human oversight [12,15,18,20,36,42]. However, these risks have mainly been explored up until now in lower-stakes or student design studies; with limited evidence on whether, and to what extent, these risks arise in professional or mission-critical engineering contexts.

Critically, existing studies have focused almost exclusively on ideation outcomes or where access to LLMs was engaged in the design process [6,20], overlooking how LLM access changes the moment-to-moment process of ideation itself. Understanding these shifts in cognitive effort and process is just as vital as measuring the final ideas, especially for professional practitioners. This study addresses this gap by simultaneously examining the quality of generated ideas and the behavioral patterns that produce them. By combining idea quality evaluations with process-timeline coding in a highly constrained aerospace context, we explore human-AI design holistically.

2.2 Measuring Ideation Success

Research on engineering ideation traditionally evaluates outcomes using four metrics: quantity, novelty, variety, and overall quality [13, 14, 24, 44]. Quantity captures the number of non-redundant ideas, novelty reflects uncommonness relative to a reference set, and variety assesses breadth across solution categories. Quality is typically treated as a single expert-rated dimension reflecting expected usefulness or performance. AI-assisted ideation studies generally adopt this same framework, reporting how LLMs influence idea counts, novelty distributions, and variety across categories, as well as overall expert-rated quality [7, 10, 18, 19, 34, 35, 41, 42]. While valuable, the quality metrics do not differentiate why an idea is judged as high quality or low quality, an important limitation in high-constraint engineering contexts where an idea may be novel yet infeasible, or technically plausible but operationally unrealistic.

To make idea quality interpretable for such mission-critical settings, this study adopts the Desirability, Viability, Feasibility (DVF) framework [4,21,23]. DVF has roots in innovation and design practice, but here is adapted for aerospace concept evaluation to yield diagnostic insight into the specific strengths and weaknesses of ideas generated with and without AI assistance. In this study, Desirability reflects alignment with mission value; Viability captures programmatic and operational realism; and Feasibility assesses technical plausibility. For instance, a lunar rover concept might be Desirable for its scientific potential yet score low on Viability due to budget constraints or Feasibility because of immature power technologies. Operationalization of the DVF scoring is described in the methodology and results sections.

3. METHODOLOGY

This study employed a comparative post-test only with control group experiment to investigate the influence of LLM access on ideation for a complex engineering problem. The research was centered on a real problem faced by a NASA project: developing concepts for landing a fixed-wing aircraft on the surface of Mars.

3.1 Participants

Participants were professional engineers with backgrounds in aerospace, mechanical, or systems engineering. All had direct experience with early-stage design or technology development projects within NASA or comparable federal research environments. Participants were recruited through internal NASA engineering mailing lists and targeted outreach to engineering teams engaged in conceptual design work. Half of the participants were affiliated with the MAE project, and the other half were recruited from outside the project. Eligibility required (1) at least one year of post-undergraduate engineering experience, (2) familiarity with conceptual design or system development, and (3) affiliation with NASA as either a permanent employee, term employee, or contractor. Experience ranged from 1 year of post-graduate experience to 28 years, with an AI mean of 10.75 years and a manual mean of 6.5 years. A total of N = 8 engineers participated.

3.2 Experimental Design and Data Collected

Participants were assigned to conditions using a constrained randomization procedure. To ensure balanced representation of domain familiarity, participants affiliated with the MAE project were randomly split between conditions (two per condition), as were participants recruited from outside the project (two per condition). Within these constraints, assignment was random.

Control Group (Human-Only Ideation)

Participants in the control group generated ideas using traditional individual ideation methods. They were instructed to produce as many distinct concepts as possible within the allotted time and to document each concept on a separate worksheet, which had space for a title, free drawing, and notes. Twenty worksheets were provided to each participant. Training documents provided prior to the session emphasized breadth, clarity, and independence of ideas, mirroring standard individual-brainstorming protocols used in engineering design. Participants could access online tools, such as search engines, if

This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

they chose to, but could not use any kind of LLM. Following the ideation session, participants completed a survey on their experience.

Experimental Group (AI-Assisted Ideation)

Participants in the experimental condition were given access to a designated LLM, ChatGSFC, an approved LLM operating within the NASA firewall structure, running OpenAI's ChatGPT-3.1 model. They received guidance on how to use the LLM, as well as guidance on prompting strategies aligned to the training documents provided to the control group. They were asked to use or abandon the tool as they saw fit to generate ideas, elaborate concepts, or explore alternatives, and to document each chosen concept on its own worksheet. Participants' prompt transcripts were captured within the LLM environment for later process analysis. Following the ideation session, participants completed a survey (provided in the supplemental materials) on their experience.

Task Duration and Data Collection

All participants were given 60 minutes to complete the ideation task. Ideation tasks were completed during live sessions that were conducted remotely or in person in the same manner via a secure Microsoft Teams channel. Audio and screen recordings were maintained in a private Teams channel accessible only to the research team. Data collected included:

- The full set of written ideas from each participant
- AI-assisted prompt transcripts (experimental group only)
- A post-task survey capturing perceived usefulness, self-reported quality, and interaction experience
- Observational notes and timestamps on idea entry order, the initiation of each new worksheet, and participant behavior

3.3 Prompt

All participants received the same problem prompt (SM-2), which described the mission context, key constraints, and desired outcomes for landing the Mars Airborne Explorer fixed-wing vehicle. The prompt was adapted from NASA's own concept-development materials and selected because it provided clear operational requirements while still leaving meaningful room for creativity. This ensured that any differences between groups were driven by tool use rather than differences in problem framing.

3.4 Subject-Matter Expert Review of Ideas

A panel of NASA subject-matter experts (SMEs) with professional experience in aerospace vehicle design, planetary mission operations, and systems engineering individually evaluated all ideas. SMEs were recruited from the MAE program leadership for their domain knowledge, were required to have at least five years of relevant domain knowledge, and at least ten years of professional experience post-baccalaureate. None had participated in the project's ideation phase.

SMEs were blind to the condition, meaning they did not know whether ideas originated from the human-only or AI-assisted group. Ideas were presented in randomized order to mitigate sequence effects. Ten ideas were randomly selected to appear twice, once early in the randomized order and a second time to assess if any intra-rater drift occurred in the scoring. Scoring took place in a single structured remote session; SMEs received training on the evaluation rubric and completed approximately 240 minutes of scoring with interspersed breaks. Each idea was rated independently by the four SMEs on the three DVF scales.

3.5 Idea Quality Assessment Using DVF

Each distinct idea was rated on a 1-5 anchored scale for each dimension (TABLE 1). Ratings from the four SME raters were averaged to produce a dimension-level score, and the three dimensions were summed to create an overall DVF composite score. Inter-rater reliability was assessed using Cohen's quadratic weighted kappa [27]. Inter-rater reliability yielded: κ = 0.307 (D, Fair), 0.232 (V, Fair), and 0.481 (F, Moderate) agreement across raters. Disagreements were noted, and all four rater scores were retained and averaged rather than adjudicated, as the averaged score was used for all subsequent analysis. Ideas scored twice by the same rater were assessed for drift, but only the first score was utilized for analysis. Intra-rater reliability was assessed using Cohen's quadratic weighted kappa on the 10 duplicate ideas per rater, comparing each rater's original score to their score on the same idea encountered a second time in a randomized position. The mean κ = 0.565 (D, Moderate), 0.610 (V, Moderate/Substantial), and 0.709 (F, Substantial). Three of four raters demonstrated substantial overall self-consistency (mean $\kappa > 0.61$, while one rater showed lower intra-rater agreement on Desirability ($\kappa = 0.085$), attributable in part to a single item scored 5 on first encounter and 1 on re-encounter.

TABLE 1: DEFINITIONS AND ANCHORED SCALES

	Desirability	Viability	Feasibility
Score	***Is the concept likely to deliver value against the mission requirements?***	***Is the concept likely to be built, tested, and flown within the expected time, budget, and operational structure?***	***Is the technology behind the concept plausible to develop and implement?***
1	Minimal mission value	Poor fit	Implausible given physics or current technology
3	Partial mission value	Conditional fit	Plausible but uncertain
5	High mission value	Strong fit	Plausible with current or near-term technology

This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

3.6 Analyzing Data to Address Research Questions

To contextualize subsequent research questions total number of distinct ideas per participant was analyzed. For RQ1, idea quality, the quantitative DVF scores for each dimension were analyzed to compare the AI assisted and non-assisted groups. Both idea-level (treating each idea as an independent observation) and participant-aggregated (each participant's mean DVF across their ideas) analyses were conducted. The participant-level aggregation served as a robustness check to confirm that idea-level patterns were not driven by a single prolific participant, though with only four participants per condition these comparisons are interpreted as descriptive rather than inferential. Because ideation success often depends on peak quality rather than average quality, a participant need only produce one high -value concept among many, we also examined each participant's top D, V, F, and composite scores as a complimentary metric. At the idea level, Mann-Whitney U tests were used given the non-normal distributions with effect sizes reported as rank biserial correlations. The rank-biserial correlation measures from –1 to +1, where the sign indicates direction and the magnitude reflects the degree of separation between the groups. The Holm-Bonferroni correction, a stepwise adjustment that is less strict than a simple Bonferroni correction while managing the familywise error rate, was applied across DVF metrics and process codes respectively. At the participant level, the same non-parametric tests were applied, though these are interpreted primarily as directional indicators given the low power due to small sample size. In order to address the nested structure of the data, a one-way nested ANOVA was applied.

To assess how each group covered the landing-concept design space, and address RQ2, all 66 ideas were categorized by landing approach. Categorization followed a two-phase process. First, a single reviewer conducted an inductive (open) coding pass over all ideas, generating category labels grounded in the data. This yielded nine categories: *Aerodynamic*, *Anchor*, *Deforming Structure*, *Mechanism*, *Parachute*, *Payload*, *Rockets*, *Runway*, and *Other*. Ideas combining two distinct approaches were labeled *Composite* and their constituent categories were recorded. The full category list and their definitions are provided in the supplemental materials (SM-4).

The resulting deductive category list and definitions were then provided to a second independent reviewer, who categorized all 66 ideas using the given list. Inter-rater agreement between the two reviewers was 50.0% (Cohen's $\kappa = 0.432$, moderate agreement). All disagreements were adjudicated through discussion, and a final consensus categorization was established. The consensus categories are used in all reported analyses. For composite ideas, each constituent category was credited to the respective category total (e.g., an idea coded as Composite: Parachute + Anchor counts once toward Parachute and once toward Anchor). This ensures that the category counts reflect full design-space coverage rather than undercounting approaches that appear within multi-method concepts.

To address RQ3, process timelines were constructed from timestamps embedded in the ideation workflow. For each participant, the research team recorded where each new idea was initiated (marked by a participant beginning a new worksheet) and for AI-assisted participants when prompting and reviewing activities began and ended. This was derived from prompt transcript timestamps and screen recordings. These timestamps were used to construct process timelines [28] for each participant, which were then coded into five mutually exclusive categories. These codes were defined in the following way: *General Considerations* refers to time spent on task orientation, reading the prompt, or transitioning between activities; *Sketching* refers to translating or annotating a visual or written representation of an idea on a worksheet; *Ideating* refers to unstructured time where the participant was considering solutions, but not sketching or performing another activity; *Prompting* refers to composing and submitting a query to the LLM; and *Reviewing* refers to reading and evaluating LLM-generated outputs. Per-participant proportions of session time in each code were computed rather than absolute durations because all participants worked in the same 60-minute window. Proportions thus controlled for minor variations in active engagement time and facilitate direct comparison of relative time allocations between conditions.

RQ4 is addressed primarily through qualitative analysis of five open-ended survey prompts covering *method or tool challenges*, *method or tool benefits*, *aspects of the process that were enjoyed*, *aspects of the process that were challenging*, and *general comments* using an inductive thematic coding [26]. Each response was segmented into meaning units and then grouped into higher-order themes. A total of 37 coded segments across the eight participants yielded seven emergent themes: three exclusive to the AI-assisted condition and four shared across conditions but with differing frequency. The codebook is provided in SM-11. Likert-scale responses from the participants were analyzed descriptively [2,15]. Given the small sample size, formal statistical comparisons between groups are reported in the supplemental materials (SM-10) for completeness but are not the primary basis for interpretation.

4. RESULTS

Results are organized around the study research questions (RQ1-RQ4). The pilot sample comprised N = 8 professional engineers (AI-assisted: N= 4; manual: N = 4), yielding n = 66 distinct ideas (AI-assisted: n = 29; manual: n = 37).

4.1 Idea Quantity and Completeness

At the group level, the manual condition produced more distinct ideas (37) than the AI-assisted condition (29) (Table 2). However, at the participant level, the difference in distinct idea counts was not statistically significant (Mann-Whitney $U = 6.5$, $p = 0.770$; rank biserial effect = -0.188). Participants in the AI-assisted group generated a mean of 7.25 distinct ideas per participant (SD = 1.89; median = 6.5) compared to 9.25 (SD = 4.79; median = 8.0) in the manual group (Table 2 and SM-F.1).

This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

TABLE 2: OUTCOMES PER PARTCIPANT BY GROUP

Metric	Group	Mean	Median	SD
Distinct Ideas	AI	7.25	6.5	1.89
	Manual	9.25	8.0	4.79
Complete Ideas	AI	2.50	1.5	3.32
	Manual	5.75	5.0	2.22

A second production metric was collected, self-rated "complete" ideas, counts only those ideas marked complete by participants on their worksheets. The manual condition produced more complete ideas overall (23) than the AI-assisted condition (10). Participant-level complete idea counts also favored the manual group (AI-assisted mean = 2.50, SD 3.32; manual mean = 5.75, SD 2.22), but the difference was not statistically significant in this pilot sample (U = 3.0, p = 0.189; rank-biserial effect = -0.625) (SM-F.2). Descriptively, the manual group shows substantially greater spread in distinct ideas (SD = 4.79 vs 1.89), suggesting greater individual differences in output among manual participants.

4.2 RQ1 (Idea Quality): D, V, F, and DVF Composite

Idea Quality was operationalized using DVF, scored on 1-5 scales and aggregated across four SME raters, consistent with the DVF rubric in Table 1. For each idea, mean D, V, and F scores were computed and summed to a DVF composite (D + V + F).

At the idea level, AI-assisted ideas scored higher on all DVF metrics (Table 3; Fig. 1; SM-F.3). The largest difference was observed for Viability (AI-assisted mean = 2.67 vs manual mean = 2.19), which was statistically significant (U = 742.0, p = 0.0077; effect 0.383). Feasibility was also higher for AI-assisted ideas (2.99 vs 2.63; U = 692.5, p = 0.0434; effect 0.291). Desirability trended higher for AI-assisted ideas (2.64 vs 2.34) but did not reach conventional significance thresholds in this pilot (U=676.0, p = 0.0708; effect = 0.260). The DVF composite was higher for AI-assisted ideas (8.30 vs. 7.16) and statistically significant (U = 728.5, p = 0.0132; effect = 0.358).

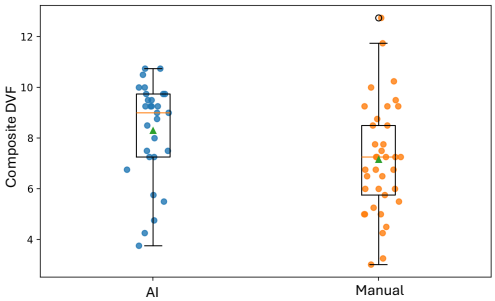


FIGURE 1: COMPOSITE DVF SCORES BY GROUP

TABLE 3: IDEA-LEVEL DESCRIPTIVES AND TESTS

Metric	AI-Assisted Mean (SD)	Manual Mean (SD)	U	p	Effect
D	2.64 (0.65)	2.34 (0.73)	676.0	0.0708	0.260
V	2.67 (0.72)	2.19 (0.78)	742.0	0.0077	0.383
F	2.99 (0.77)	2.63 (0.86)	692.5	0.0434	0.291
DVF Composite	8.30 (1.95)	7.16 (2.18)	728.5	0.0132	0.358

Because ideas are nested within participants, DVF was reanalyzed using a one-way nested ANOVA in which ideas served as the unit of analysis, and participants were treated as a random factor nested within condition. This model (1) tests treatment effects against the subject-within-treatment mean square, accounting for the non-independence of ideas from the same participant.

$$y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk} \quad (1)$$

AI-assisted means were higher than manual means across all metrics (Table 4), but none of the differences were statistically significant: Desirability F(1,6) = 1.657, p = 0.245; Viability F(1,6) = 3.154, p = 0.126; Feasibility F(1,6) = 1.911, p = 0.216; Composite DVF F(1,6) = 2.408, p = 0.172 Holm-Bonferroni adjustment across the four DVF metrics yielded adjusted near p ≥ 0.5 for each comparison.

TABLE 4: IDEA-LEVEL DESCRIPTIVES AND TESTS

Metric	F (1,6)	p	Holm-Bonferroni p
Mean D	1.657	0.245	0.515
Mean V	3.154	0.126	0.504
Mean F	1.911	0.216	0.515
DVF Composite	2.408	0.172	0.515

To test whether the AI-assisted advantage holds for the strongest concepts from each participant, we averaged each participant's N highest scoring ideas by DVF composite and compared N = 1 through 5 (Table 5; Fig. 2).

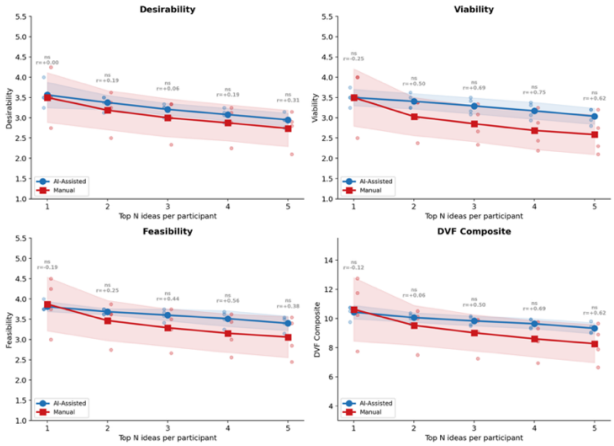


FIGURE 2: D, V, F, AND DVF COMPOSITE OF TOP N IDEAS

This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

TABLE 5: TOP N ANALYSIS OF DVF COMPOSITE

N	AI Mean	Manual Mean	Difference	r (Effect)
1	10.44	10.63	-0.19	-0.125
2	10.06	9.53	+0.53	+0.062
3	9.83	9.02	+0.81	+0.500
4	9.64	8.61	+1.03	+0.688
5	9.32	8.28	+1.05	+0.625

At N = 1, DVF composites were nearly identical (AI mean = 10.44; manual mean = 10.63; U = 7.0, p = 0.885, r = -0.125), and the highest scoring idea in the dataset originated from the manual group (Idea 7, DVF = 12.75). As N increased, the AI advantage emerged and grew monotonically. By N = 3 the composite difference was +0.81 (r = +0.50), and by N = 4 the difference reached +1.03 (r = +0.69). This pattern was consistent across Viability and Feasibility, while Desirability showed negligible separation at all values of N. Because all comparisons are at the participant level, none reached conventional significance thresholds, but the fact that the uniformity of the monotonic effect trend appears across N and across all DVF dimensions strengthens confidence in the directional pattern.

4.3 RQ2 (Design Space Coverage): Idea Categorization

To examine how each group explored the design space, all 66 ideas were categorized by landing approach using the consensus categories described in Section 3.6. Table 6 and Figure 3 summarize the category counts by group, with composite ideas credited to each constituent category.

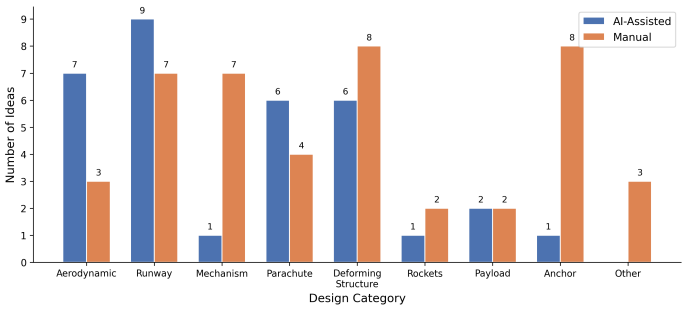


FIGURE 3: GROUP IDEA COUNT BY DESIGN CATEGORY

TABLE 6: IDEA COUNT BY DESIGN CATEGORY AND GROUP

Category	AI-Assisted	Manual	Total
Aerodynamic	7	3	10
Runway	9	7	16
Mechanism	1	7	8
Parachute	6	4	10
Deforming Structure	6	8	14
Rockets	1	2	3
Payload	2	2	4
Anchor	1	8	9
Other	0	3	3
Total instances	**33**	**44**	**77**

At the group level, both conditions covered a broad range of categories, but with notably different emphases. The AI-assisted group produced the most ideas in the Aerodynamic (n = 7) and Runway (n = 9) categories, with Parachute and Deforming Structure also well represented (n = 6 each). In contrast, the manual group showed higher counts in Anchor (n = 8), Deforming Structure (n = 8), Mechanism (n = 7), and Runway (n = 7). Several categories showed pronounced between-group asymmetries: Aerodynamic ideas were more than twice as prevalent in the AI-assisted group (7 vs. 3), while Mechanism (1 vs. 7) and Anchor (1 vs. 8) ideas were substantially more common in the manual group. Three categories that appeared only or predominantly in the manual group were Other (0 vs. 3), Anchor (1 vs. 8), and Mechanism (1 vs. 7).

The AI-assisted group's ideas were more concentrated in fewer categories. The top two AI categories (Runway and Aerodynamic) accounted for 48.5% of AI-attributed category instances, compared to 36.4% for the manual group's top two (Anchor and Deforming Structure). Shannon entropy, a measure of distributional evenness, was lower for the AI group (H = 2.58) than the manual group (H = 2.99; maximum possible H = 3.17 for nine categories), indicating a less even spread across the design space.

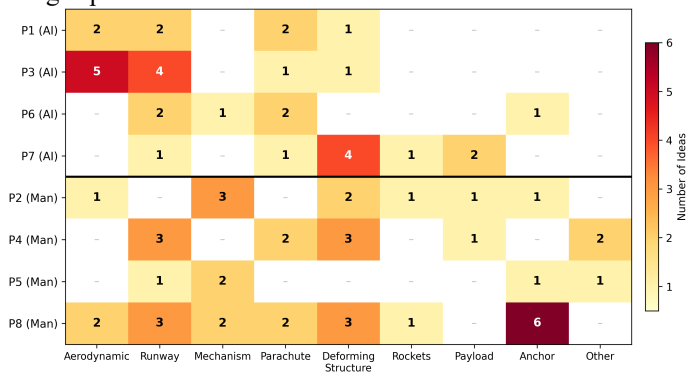


FIGURE 4: DESIGN SPACE COVERAGE HEATMAP

At the individual level (Fig. 4) manual participants covered more categories on average (mean = 5.50 unique categories, SD = 1.29) than AI-assisted participants (mean = 4.25, SD = 0.50). Two AI-assisted participants (P1 and P3) showed particularly concentrated profiles, with all or nearly all of their ideas falling in Aerodynamic and Runway. In contrast, several manual participants (P2, P4, P8) distributed ideas across six or more categories. P8 (Manual) exhibited the broadest coverage (7 of 9 categories), with a notably high count in Anchor (n = 6), a category almost absent in the AI group. Twelve ideas (18.2% of the total) were classified as Composite, indicating that participants in both groups sometimes combined two landing approaches into a single concept; five composites originated from the AI group and seven from the manual group.

4.4 RQ3 (Process): Time Allocation

Process was captured using design-timeline-style coding and summarized as the proportion of session time spent in each code for each participant into process timelines (see Section 3.6

This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

for code definitions). The groups exhibited markedly different time allocation profiles (Table 7; Figs. 5-6). Manual participants spent most of their time *Sketching* (mean = 0.527) and *Ideating* (mean = 0.407), thinking through concepts rather than sketching or performing another activity, with no time spent on *Prompting* or *Reviewing*, as these activities are unique to the AI-assisted workflow. In contrast, AI-assisted participants spent substantial time on *Sketching* (0.399) and *Reviewing* (0.334), with additional time on *Prompting* (0.175) and comparatively less time on *Ideating* (0.038). Because *Prompting* and *Reviewing* are only available in the AI-assisted workflow, differences in those codes primarily reflect condition affordances. Accordingly, we interpret the process results primarily in terms of how time shifted among shared activities, most notably the reduction in Ideating time in the AI-assisted condition.

Mann-Whitney tests on per-participant proportions indicated significant unadjusted differences for *Ideating* (U = 0.0, p = 0.0286, effect = -1.0), *Prompting* (U = 16.0, p = 0.0211; effect 1.0), and *Reviewing* (U = 16.0, p = 0.0211; effect = 1.0). Chi-square tests on pooled counts confirmed the same pattern across all five codes (Table 8), though with greater apparent precision due to unequal participant contributions. Given the sample size, these tests are treated as directional indicators rather than definitive results, and the patterns are discussed further in Section 6.1.

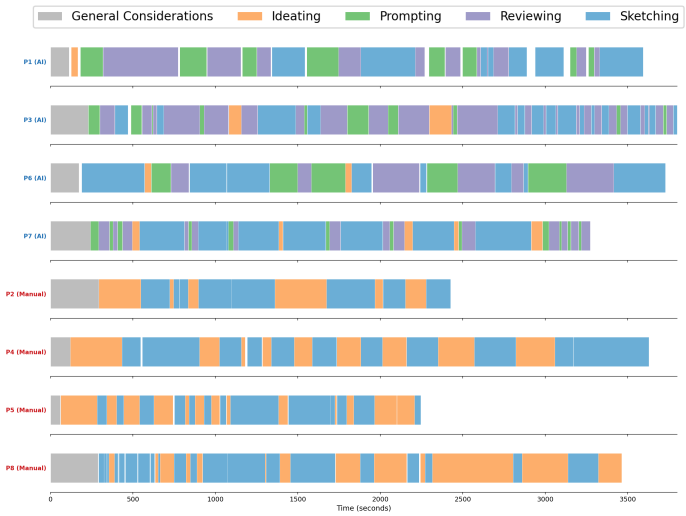


FIGURE 5: PARTICIPANT PROCESS TIMELINES

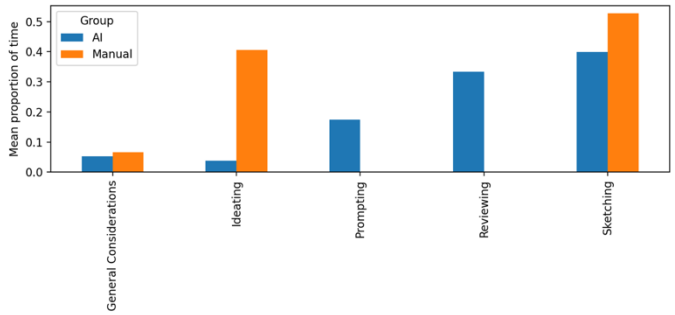


FIGURE 6: MEAN TIME ALLOCATION BY PROCESS CODE

TABLE 7: TIME ALLOCATIONS BY CODE AND GROUP

Code	AI-Assisted Mean (SD)	Manual Mean (SD)
General Considerations	0.053 (0.018)	0.066 (0.045)
Ideating	0.038 (0.026)	0.407 (0.053)
Prompting	0.175 (0.084)	0 (0)
Reviewing	0.334 (0.109)	0 (0)
Sketching	0.399 (0.110)	0.527 (0.049)

TABLE 8: RAW AND HOLM ADJUSTED TESTS

Code	U	MW p	X^2	X^2 p
General Considerations	4.5	0.384	0.03	0.864
Ideating	0	0.029	41.54	<0.001
Prompting	16	0.021	26.22	<0.001
Reviewing	16	0.021	49.05	<0.001
Sketching	3	0.200	17.33	<0.001

4.5 RQ4 (Perception): Participant Ratings and Themes

Inductive coding of free-response items yielded nine codes under seven thematic families (Fig. 7, SM-11). Three thematic families were fully exclusive to the AI-assisted condition, and two thematic families were split into sub-codes to capture differences within a single construct.

			AI-Assisted (n = 4)				Manual (n = 4)					
Code	Theme	Scope	P1	P3	P6	P7	P2	P4	P5	P8	Σ_{AI}	Σ_{M}
T1	LLM as knowledge resource	*AI-only*	■	■	■	■	□	□	□	□	4	0
T2	Prompt literacy learning curve	*AI-only*	□	■	■	■	□	□	□	□	3	0
T3	AI output convergence & filtering	*AI-only*	□	□	■	■	□	□	□	□	2	0
T4a	Sketching as generative asset (+)	*Shared*	□	□	□	□	□	■	■	■	0	3
T4b	Text-based modality friction (–)	*Shared*	□	■	■	□	□	□	□	□	2	0
T5	Cognitive fatigue & sustaining divergence	*Shared*	■	□	□	□	□	■	■	□	1	2
T6	Task & constraint framing	*Shared*	■	□	□	□	■	□	■	■	1	3
T7a	Intrinsic process enjoyment	*Shared*	□	□	□	■	■	■	■	■	1	4
T7b	Instrumental tool satisfaction	*AI-only*	■	■	■	□	□	□	□	□	3	0

Note. ■ = code present in participant's open-ended responses; □ = absent. Shaded rows denote AI-exclusive codes. T4 is split into sub-codes capturing positive visual–spatial affordance (T4a) and negative text-modality friction (T4b). T7 is split into intrinsic process enjoyment (T7a) and instrumental tool satisfaction (T7b). Σ_{AI} and Σ_{M} report the number of participants endorsing each code within the AI-assisted and manual groups, respectively.

FIGURE 7: QUALITATIVE SURVEY CODES AND THEMES

All four AI-assisted participants described the LLM as a *knowledge resource and analytical partner*, citing domain-specific technical information and on-demand feasibility evaluations that they incorporated into their concepts (T1). Three of the four reported a prompt literacy learning curve, and P3-AI noted a "language and vocabulary gap," while P6-AI described spending time "trying to learn the tool... rather than using it as a design tool" (T2). Two participants described *AI output convergence and filtering burden*, P7-AI noted responses were "overly specific & locked into variations of the same/overall concept" and that some outputs had to be discarded (T3).

The perceived relationship between modality and ideation quality divided sharply between condition. Three manual

This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

participants valued *sketching as a generative asset* (T4a), where P4-Manual linked "drawing things out" to natural analogies, and P5-Manual described iterating through drawing and annotation as enjoyable and productive. Conversely, two AI-assisted participants reported *text-modality friction* (T4b), P3-AI noted being a "visual person" who found it "difficult to understand the info given back," and P6-AI cited interface difficulties, "hard to scroll, hard to interact," on two separate survey items.

Cognitive fatigue (T5) was more prominent in the manual group (2 of 4 vs. 1 of 4 in AI), with P4-Manual reporting "my brain got tired; ideas stemmed from a couple main ideas." *Task and constraint framing* (T6) was raised by three manual participants who perceived the constraints as overly restrictive, versus one AI participant who discussed time pressure, possibly because the prompting and reviewing of LLM prompts became a limiting factor. Engagement and effect split into two sub-codes: all four manual participants expressed *intrinsic process enjoyment* (T7a), exemplified by P8-Manual's "It was fun!!!!", whereas the AI-assisted participants' positive comments were *largely instrumental* (T7b), focused on tool capability rather than the experience of ideation itself such as from P6-AI "I think this could turn into a really useful tool." From the AI-assisted condition, only P7-AI expressed intrinsic enjoyment, notably framing it as a personal disposition rather than driven by the tool.

Statistical comparisons between conditions were conducted for completeness and are provided in the supplemental materials (SM-10). The largest descriptive difference is for Ease of Sketching, with lower ratings in the AI-assisted group (mean = 3.0) than the manual group (mean = 4.5).

5. DISCUSSION

This pilot study explored how access to an LLM shapes professional engineers' ideation on a mission critical aerospace design task. Overall, the results suggest that LLM access may change how engineers spend time, even when it does not change idea production. Given the small sample (N = 8), findings are best treated as preliminary signals rather than definitive effects.

5.1 Interpretation by Research-Question Family

Production. In this pilot study LLM access did not increase idea or self-rated completeness counts at the participant level. This is noteworthy given that much of the prior literature on AI-assisted ideation has reported AI-related gains in idea quantity [10,18,19]. The absence of such a gain in a professional context may reflect the more deliberate, quality-focused approach that experienced engineers bring in a highly constrained aerospace context, or the time cost of the prompt-and-review cycles within a fixed session window.

Idea Quality (RQ1). AI-assisted ideas scored higher, on average, on Viability, Feasibility, and the DVF composite, with Desirability trending in the same direction but not to statistical significance. These findings extend prior work suggesting that LLMs can enhance the technical plausibility [18,19,34] into a professional, mission-critical context. Analysis of the Top N ideas qualifies this finding by elaborating that at N = 1, no AI advantage was observed, and that the single best idea in the dataset was produced by the manual group. The AI advantage only emerged, and grew monotonically, as N increased, indicating that the effect is driven by a higher quality floor or lower variability rather than superior peak performance. The LLM appears to function not as a generator of exceptional concepts but rather to consistently reduce technically underdeveloped ideas. This interpretation is reinforced by the near-zero AI advantage in Desirability across all values of N, compared with the observed improvements in Viability and Feasibility.

Design Space Coverage (RQ2). The pattern observed in the design space coverage is consistent with the broader finding in the literature that LLM-generated ideas tend to cluster more tightly in the design space than human-generated solutions [34], possibly because the LLM's training corpus overrepresents conventional or well-documented approaches. The narrower per-participant category breadth (4.25 vs. 5.50 categories) provides quantitative evidence of this clustering effect in a professional engineering context. This finding complements the production and quality findings: AI-assisted ideas scored higher on DVF metrics, whereas manual ideation showed greater variability alongside broader exploration, a pattern that may be valuable in creativity contexts where weaker ideas can still act as stepping stones to more original concepts.

Process (RQ3). AI-assisted participants shifted time away from the Ideating code and into *Prompting*, *Reviewing*, and *Sketching*. Because *Prompting* and *Reviewing* are only possible in the AI condition, the most informative result is the reduction in ideating time and the corresponding redistribution of effort among shared activities. This aligns with prior work showing that interaction strategy and iterative refinement shape human-AI ideation behaviors [18,19,40,41]. The redistribution observed here may therefore help explain lower category breadth in the AI-assisted condition: spending less time in open ideation and more time working through generated suggestions may steer participants toward a narrower region of the design space. While this pilot cannot separate effects of model training from effects related to the use of an AI tool, it suggests both may contribute.

Perception (RQ4). These signals are tentative at this sample size. However, the qualitative coding reported in section 4.5 reveals differences in participant experience that the Likert items alone do not capture. The most striking pattern is the affective split coded as T7a/T7b where all four manual participants expressed spontaneous intrinsic enjoyment of ideation itself, while AI-assisted participants' positive comments were predominantly centered around what the tool could accomplish rather than the experience of creating. Only one AI-assisted participant (P7) expressed intrinsic enjoyment and notably framed it as a personal disposition rather than a tool-mediated experience. This divergence suggests that LLM access while functionally useful, may not enhance, and may even dampen, process satisfaction. This could be a consideration for adoption in professional practice. The modality sub-codes (T4a/T4b) reinforce this pattern. Manual participants described sketching as generative, while two AI assisted participants reported friction with text-heavy outputs. This matches the lower Ease of

This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

Sketching ratings in the AI group (3.0 vs. 4.5) and its positive correlation with self-rated completeness ($\rho = 0.684$, $p = 0.061$). These findings suggest that perception measures in human-AI ideation studies should distinguish between task-outcome satisfaction and process-experience satisfaction, as the two may diverge when an LLM is introduced.

5.2 Potential Mechanisms

From this limited pilot, three mechanisms plausibly explain the observed pattern of unchanged idea count but higher Viability and Feasibility in the AI-assisted condition, with the contrast between these dimensions and the non-significant Desirability result providing additional interpretive texture.

Externalized elaboration and feasibility checking. LLMs rapidly generate technically grounded elaborations of nascent concepts, prompting designers to record ideas with greater technical specificity and fewer untested assumptions, directly improving programmatic and technical plausibility. The process timeline data provide partial support for this mechanism; the substantial shift from *Ideating* toward *Reviewing* in the AI-assisted condition suggests the designers spent time evaluating and incorporating AI-generated technical details rather than generating concepts from scratch.

Time-budget trade-off. Within the fixed 60-minute session, time spent prompting the LLM and reviewing its outputs competed with time available for generating distinct concepts. This is consistent with the observation that AI-assisted participants produced fewer ideas despite producing higher-quality ideas on average.

Training data anchoring. LLMs tend to surface concepts with precedent in their training corpus, meaning AI-assisted ideas may disproportionately reflect technically validated, and thus more feasible and viable, approaches. This also predicts the weaker Desirability impact as mission-specific value is defined by MAE's unique requirements which are nonexistent in general training data. The dominant AI-assisted categories, *Aerodynamic* and *Runway*, correspond to conventional fixed-wing landing approaches, so they are well represented in the aerospace literature and likely in the LLM's training data. Underrepresented categories in the AI group, particularly *Mechanism* and *Anchor*, tend to involve more unconventional or Mars-specific solutions that may have fewer precedents in the training corpus. This effect may partly explain why AI-assisted ideas scored higher on Viability and Feasibility: the LLM may have directed participants to approaches with robust engineering precedent, producing more technically grounded but less diverse concepts.

5.3 Limitations

The limitations of this pilot are best understood in the context of its goals: to generate preliminary signal on AI-assisted ideation in a professional engineering setting and to refine methods for a larger study, not to establish causal claims or broad generalizability. With that framing, the most relevant limitations are as follows.

With N = 8 participants (four per condition), participant-level analyses are substantially underpowered, meaning null results should be interpreted cautiously rather than as evidence of no effect. Ideas are nested within participants, violating the independence assumption of simple group comparisons, this was partly addressed by conducting analysis at both the idea level and the participant-aggregated level, but a mixed effects model approach would handle the nesting more rigorously in a larger study. Condition affordances also present an interpretive challenge. *Prompting* and *Reviewing* activities are definitionally only available in the AI-assisted condition rather than behaviorally emergent.

Uneven prior domain familiarity with the MAE project and unmeasured prior LLM experience may have introduced participant-level confounds, though the constrained randomization was designed to distribute this equally across conditions. The DVF and "completeness" ratings are judgment-based measures, a reality addressed through well-anchored scales and multiple SME raters, though reliability and drift reporting should be strengthened in future work. Similarly, the moderate inter-rater agreement on design-space category assignment ($\kappa = 0.432$) reflects the inherent complexity of categorizing ideas that span multiple approaches; the consensus adjudication process mitigated this, and additional raters in a larger study would further strengthen confidence in category boundaries. To support ecological validity the study used a real NASA design problem and recruited participants across the full range of eligible MAE and non-MAE affiliations.

5.4 Future Work

This pilot study provides preliminary evidence that LLM support can meaningfully shift ideation process and outcomes, with several directions remaining for future research. A larger follow on study with stronger methods, deeper understanding of participant demographics, and greater statistical power could test whether these patterns hold more broadly. Additionally, a more direct examination of how LLM responses shape outcomes, through transcript analysis of prompting and reviewing behaviors would be valuable.

6. SUMMARY OF PILOT FINDINGS

In this pilot dataset, AI assistance did not significantly increase the production of ideas. Still, AI-assisted ideas scored higher on SME-rated Viability, Feasibility, and DVF composite at the idea level (RQ1), though those ideas clustered more narrowly in the design space (RQ2). Process results indicate a substantial reallocation in time from ideating to prompting and reviewing in the AI-assisted condition (RQ3). At the same time, (RQ4) perceptions noted that an LLM can be a useful knowledge resource, but that interacting with an LLM may reduce process enjoyment and increase friction with a text-based interface.

ACKNOWLEDGEMENTS

We would like to thank the MAE Team, NASA Convergent Aeronautics Solutions, and the NASA Langley Engineering Directorate. This study was approved by NASA IRB #STUDY00000558/ MOD00002372 and UVA IRB 7725.

This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

REFERENCES

[1] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., and Ziegler, D., 2022, "Training Language Models to Follow Instructions with Human Feedback," arXiv preprint, arXiv:2203.02155.

[2] Braun, M., Greve, M., Kegel, F., Kolbe, L., and Beyer, P. E., 2024, "Can AI Have a Word With You? A Taxonomy on the Design Dimensions of AI Prompts," *Proceedings of the 57th Hawaii International Conference on System Sciences (HICSS-57)*, Honolulu, HI.

[3] Brown, T. B., Mann, B., Ryder, N., et al., 2020, "Language Models Are Few-Shot Learners," *Advances in Neural Information Processing Systems*, **33**, pp. 1877–1901.

[4] Dean, D. L., Hender, J. M., Rodgers, T. L., and Santanen, E. L., 2006, "Identifying Quality, Novel, and Creative Ideas: Constructs and Scales for Idea Evaluation," *Journal of the Association for Information Systems*, **7**(10), pp. 649–699.

[5] McElheran, K., Li, J. F., Brynjolfsson, E., Kroff, Z., Dinlersoz, E., Foster, L., and Zolas, N., 2024, "AI Adoption in America: Who, What, and Where," *Journal of Economics and Management Strategy*, **33**(2), pp. 375–415. doi:10.1111/jems.12576

[6] Khanolkar, P. M., Vrolijk, A., and Olechowski, A., 2023, "Mapping Artificial Intelligence Based Methods to Engineering Design Stages: A Focused Literature Review," *AI EDAM*, **37**, e10. doi:10.1017/S0890060422000476

[7] Berni, A., Borgianni, Y., Rotini, F., Gonçalves, M., and Thoring, K., 2024, "Stimulating Design Ideation With Artificial Intelligence: Present and (Short-Term) Future," *Proceedings of the Design Society*, **4**, pp. 1939–1948. doi:10.1017/pds.2024.196

[8] Dang, H., Mecke, L., Lehmann, F., Goller, S., and Buschek, D., 2022, "How to Prompt? Opportunities and Challenges of Zero- and Few-Shot Learning for Human-AI Interaction in Creative Applications," arXiv preprint, arXiv:2209.01390.

[9] Sansanee, H., and Kiattisin, S., 2024, "The Current State of Generative AI Prompt Framework Design for Enhancing Utility in Organizational Decision Making," *Proceedings of the 2024 Technology and Innovation Management in Engineering and Science International Conference (TIMES-iCON 2024)*, Bangkok, Thailand. doi:10.1109/TIMES-iCON61890.2024.10630713

[10] Doshi, A., and Hauser, O., 2024, "Generative AI Enhances Individual Creativity but Reduces the Collective Diversity of Novel Content," *Science Advances*, **10**(38), eadn5290. doi:10.1126/sciadv.adn5290

[11] Figoli, F., Bruggeman, R., Rampino, L., and Ciuccarelli, P., 2025, "AI Against Design Map: A Systematic Review of 20 Years of AI Focused Studies in Design Research," *Design Studies*, **96**, p. 101279. doi:10.1016/j.destud.2024.101279

[12] Jansson, D. G., and Smith, S. M., 1991, "Design Fixation," *Design Studies*, **12**(1), pp. 3–11.

[13] Shah, J. J., Vargas-Hernandez, N., and Smith, S. M., 2003, "Metrics for Measuring Ideation Effectiveness," *Proceedings of the ASME 2003 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, IDETC/CIE 2003, Chicago, IL, Paper No. DETC2003/DTM-48630.

[14] Verhaegen, P. A., Peeters, J., Vandevenne, D., Dewulf, S., and Duflou, J. R., 2013, "Refinements to the Variety Metric for Idea Evaluation," *Design Studies*, **34**(2), pp. 243–263. doi:10.1016/j.destud.2012.08.003

[15] Wadinambiarachchi, S., Kelly, R. M., Pareek, S., Zhou, Q., and Velloso, E., 2024, "The Effects of Generative AI on Design Fixation and Divergent Thinking," arXiv preprint, arXiv:2403.11164.

[16] Pierson, K. C., and Ha, M. J., 2024, "Usage of ChatGPT for Engineering Design and Analysis Tool Development," NASA Technical Reports Server, Report No. 20230015977, NASA.

[17] Graydon, M., and Lehmen, S., 2025, "Examining Proposed Uses of LLMs to Produce or Assess Assurance Arguments," NASA Technical Memorandum, NASA-TM-20250001849, NASA.

[18] Ege, D. N., Øvrebø, H. H., Stubberud, V., Berg, M. F., Elverum, C., Steinert, M., and Vestad, H., 2025, "ChatGPT as an Inventor: Eliciting the Strengths and Weaknesses of Current Large Language Models Against Humans in Engineering Design," *AI EDAM*, **39**, e5. doi:10.1017/S0890060424000337

[19] Chen, L., Wu, Q., and Luo, X., 2025, "How Generative AI Supports Human in Conceptual Design," *Design Science*, **11**, e7. doi:10.1017/dsj.2025.4

[20] Qin, P., Yang, C., Li, J., Wen, J., and Lee, Y. C., 2025, "Timing Matters: How Using LLMs at Different Timings Influences Perceptions and Ideation Outcomes in AI-Assisted Ideation," arXiv preprint, arXiv:2502.06197.

[21] Kudrowitz, B., and Wallace, D., 2013, "Assessing the Quality of Ideas From Prolific, Early-Stage Product Ideation," *Journal of Engineering Design*, **24**(2), pp. 120–139. doi:10.1080/09544828.2012.676633

This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

[22] Fetters, M. D., Curry, L. A., and Creswell, J. W., 2013, "Achieving Integration in Mixed Methods Designs—Principles and Practices," *Health Services Research*, **48**(6 Pt 2), pp. 2134–2156. doi:10.1111/1475-6773.12117

[23] IDEO.org, 2015, *The Field Guide to Human-Centered Design*, IDEO.org, San Francisco, CA.

[24] Reinig, B. A., Briggs, R. O., and Nunamaker, J. F., Jr., 2007, "On the Measurement of Ideation Quality," *Journal of Management Information Systems*, **23**(4), pp. 143–161.

[25] Ericsson, K. A., and Simon, H. A., 1993, *Protocol Analysis: Verbal Reports as Data*, Rev. ed., MIT Press, Cambridge, MA.

[26] Braun, V., and Clarke, V., 2006, "Using Thematic Analysis in Psychology," *Qualitative Research in Psychology*, **3**(2), pp. 77–101.

[27] Cohen, J., 1968, "Weighted Kappa: Nominal Scale Agreement With Provision for Scaled Disagreement or Partial Credit," *Psychological Bulletin*, **70**(4), pp. 213–220.

[28] Atman, C. J., 2019, "Design Timelines: Concrete and Sticky Representations of Design Process Expertise," *Design Studies*, **65**, pp. 125–151. doi:10.1016/j.destud.2019.06.005

[29] Atman, C. J., Adams, R. S., Cardella, M. E., Turns, J., Mosborg, S., and Saleem, J., 2007, "Engineering Design Processes: A Comparison of Students and Expert Practitioners," *Journal of Engineering Education*, **96**(4), pp. 359–379. doi:10.1002/j.2168-9830.2007.tb00945.x

[30] Turns, J., Bailey, R., Howe, S., Patel, K., Kotys-Schwartz, D., Lande, M., Patten, E., Batrouny, N. A., and Atman, C. J., 2025, "Using Design Signatures to Make the Invisible Visible: Designing Activities for Design Process Learning," *International Journal of Engineering Education*, **41**(4), pp. 928–946.

[31] Likert, R., 1932, "A Technique for the Measurement of Attitudes," *Archives of Psychology*, **140**, pp. 1–55.

[32] Kwon, E., Goucher-Lambert, K., and Wood, K. L., 2022, "Enabling Multimodal Search for Inspirational Design Stimuli Using Deep Learning," *AI EDAM*, **36**, e22. doi:10.1017/S0890060422000130

[33] Goucher-Lambert, K., and Cagan, J., 2019, "Crowdsourcing Inspiration: Using Crowd-Generated Inspirational Stimuli to Support Designer Ideation," *Design Studies*, **61**, pp. 1–29. doi:10.1016/j.destud.2019.01.001

[34] Ma, K., Grandi, D., McComb, C., and Goucher-Lambert, K., 2025, "Do Large Language Models Produce Diverse Design Concepts? A Comparative Study With Human-Crowdsourced Solutions," *Journal of Computing and Information Science in Engineering*, **25**(2), p. 024501. doi:10.1115/1.4067332

[35] Zhu, Q., and Luo, J., 2023, "Generative Transformers for Design Concept Generation," *Journal of Computing and Information Science in Engineering*, **23**(4), p. 041003.

[36] Ocker, F., Menzel, S., Sadik, A., and Rios, T., 2025, "From Idea to CAD: A Language Model-Driven Multi-Agent System for Collaborative Design," arXiv preprint, arXiv:2503.04417. doi:10.48550/arXiv.2503.04417

[37] Wang, N., Kim, H., Peng, J., and Wang, J., 2025, "Exploring Creativity in Human–AI Co-Creation: A Comparative Study Across Design Experience," *Frontiers in Computer Science*, **7**, p. 1672735. doi:10.3389/fcomp.2025.1672735

[38] Das, M., Li, X., Fabunmi, O., and Hölttä-Otto, K., 2025, "Opportunities and Challenges in Using Large Language Models as Interviewees for User Needs Elicitation in Design Research," *Proceedings of the ASME 2025 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, IDETC/CIE 2025, Paper No. DETC2025-168769, V004T06A041. doi:10.1115/DETC2025-168769

[39] Jiang, S., Xie, M., and Luo, J., 2025, "Large Language Models for Combinatorial Optimization of Design Structure Matrix," *Proceedings of the Design Society: International Conference on Engineering Design (ICED25)*, **5**, pp. 2201–2210. doi:10.1017/pds.2025.10234

[40] Ma, K., Moore, G., Shyam, V., Villarrubia, J., Goucher-Lambert, K., and Brubaker, E. R., 2024, "Human–AI Collaboration Among Engineering and Design Professionals: Three Strategies of Generative AI Use," *Proceedings of the ASME 2024 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, IDETC/CIE 2024, Paper No. DETC2024-143560.

[41] Ma, K., Grandi, D., McComb, C., and Goucher-Lambert, K., 2023, "Conceptual Design Generation Using Large Language Models," *Proceedings of the ASME 2023 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, IDETC/CIE 2023, Vol. 87349, V006T06A021.

[42] Wang, B., Zuo, H., Cai, Z., Yin, Y., Childs, P., Sun, L., and Chen, L., 2023, "A Task-Decomposed AI-Aided Approach for Generative Conceptual Design," *Proceedings of the ASME 2023 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, IDETC/CIE 2023, Vol. 87349, V006T06A009.

This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

[43] Li, X., Sun, Y., and Sha, Z., 2024, "LLM4CAD: Multi-Modal Large Language Models for 3D Computer-Aided Design Generation," *Proceedings of the ASME 2024 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, IDETC/CIE 2024, Paper No. DETC2024-143740.

[44] Linsey, J. S., Clauss, E. F., Kurtoglu, T., Murphy, J. T., Wood, K. L., and Markman, A. B., 2011, "An Experimental Study of Group Idea Generation Techniques: Understanding the Roles of Idea Representation and Viewing Methods," *Journal of Mechanical Design*, **133**(3), pp. 031008-1–031008-15. doi:10.1115/1.4003498

This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

Supplemental Materials

1. SM-1 AI and Manual Survey

AI (Tool)

Section 1: Ideation Process

1. **How would you rate the overall effectiveness of the ideation process you used?**
    - Very Ineffective, Ineffective, Neutral, Effective, Very Effective
2. **How clear were the instructions provided to you before the session?**
    - Very Unclear, Unclear, Neutral, Clear, Very Clear
3. **How many of your ideas to you feel were complete versus being incomplete as defined in the Objective and Prompt?**
4. **How easy was it to translate your ideas into sketches and annotations?**
    - Very Difficult, Difficult, Neutral, Easy, Very Easy

Section 2: Ideation Methods

1. **How helpful was the tool in generating ideas?**
    - Not Helpful at All, Slightly Helpful, Neutral, Helpful, Very Helpful
2. **Do you feel the ideas generated using the tool align with the mission requirements and constraints?**
    - Not at All, Slightly, Neutral, Mostly, Completely
3. **Do you think that additional ideation tools would have been useful to your process?**
    - Not at All, Slightly, Neutral, Mostly, Completely
4. **What challenges did you face while using the ideation tool?**
5. **What aspects of the ideation tool did you find most beneficial?**

Section 3: Perception of Output Quality

1. **How innovative do you feel your ideas were?**
    - Not Innovative, Slightly Innovative, Neutral, Moderately Innovative, Highly Innovative
2. **How feasible do you feel your ideas were?**
    - Not Feasible, Slightly Feasible, Neutral, Moderately Feasible, Highly Feasible
3. **How well did your ideas address the challenge requirements?**
    - Not at All, Slightly, Neutral, Mostly, Completely

Section 4: General Feedback

1. **What did you enjoy most about the ideation process?**
2. **What did you find most challenging about the ideation process?**
3. **Any additional comments or thoughts?**

This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

Manual

Section 1: Ideation Process

1. **How would you rate the overall effectiveness of the ideation process you used?**
    - Very Ineffective, Ineffective, Neutral, Effective, Very Effective
2. **How clear were the instructions provided to you before the session?**
    - Very Unclear, Unclear, Neutral, Clear, Very Clear
3. **How many of your ideas to you feel were complete versus being incomplete as defined in the Objective and Prompt?**
4. **How easy was it to translate your ideas into sketches and annotations?**
    - Very Difficult, Difficult, Neutral, Easy, Very Easy

Section 2: Ideation Methods

1. **How helpful were the ideation methods you used in generating ideas?**
    - Not Helpful at All, Slightly Helpful, Neutral, Helpful, Very Helpful
2. **Do you feel the ideas generated align with the mission requirements and constraints?**
    - Not at All, Slightly, Neutral, Mostly, Completely
3. **Do you think that additional ideation tools would have been useful to your process?**
    - Not at All, Slightly, Neutral, Mostly, Completely
4. **What challenges did you face during ideation?**
5. **What aspects of the type of ideation you engaged in did you find most beneficial?**

Section 3: Perception of Output Quality

1. **How innovative do you feel your ideas were?**
    - Not Innovative, Slightly Innovative, Neutral, Moderately Innovative, Highly Innovative
2. **How feasible do you feel your ideas were?**
    - Not Feasible, Slightly Feasible, Neutral, Moderately Feasible, Highly Feasible
3. **How well did your ideas address the challenge requirements?**
    - Not at All, Slightly, Neutral, Mostly, Completely

Section 4: General Feedback

1. **What did you enjoy most about the ideation process?**
2. **What did you find most challenging about the ideation process?**
3. **Any additional comments or thoughts?**

This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

2. SM-2 Landing Study Prompt

Landing Study Objective and Prompt

Objective: Generate as many complete ideas as possible that address the prompt in the given time. An idea is complete when you feel that you have articulated (drawn and labeled) all the aspects you believe are important to addressing the prompt. You may go back and forth between ideas or complete each one before moving on.

Prompt: You are an **aero-mechanical engineer** working on a team designing an aircraft that will fly on Mars. Other members of the team have designed how it will get to Mars, deploy into the Martian atmosphere, and achieve flight.

Your team lead has given **you** the task of figuring out how to land the vehicle on the Martian Southern Highlands after its mission. It will **not** need to take off again. It is imperative that the vehicle be able to transmit data to satellites orbiting Mars, which will require several days post landing, as well as continue to collect atmospheric data.

Initial conditions

- Mass = 105kg
- Velocity = 70m/s
- Stall Speed = 54m/s
- CL Max = 1.2

Vehicle Characteristics

- Wingspan = 10m
- Wing Area = 12.7m^2
- Engine Output = 22.5kW nominal (30hp nominal)

Considerations

- The general shape of the aircraft must remain the same (blended wing, no tail) as well as the location of the rear mounted propeller. Otherwise, there are no design constraints on modifications to the aircraft.
- Overall damage to airframe should be minimized
- The vehicle must land right side up
- Engine w/ Propeller can run harder for a brief period
- Engine can put propeller in Beta

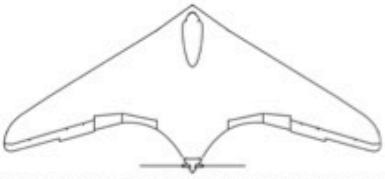


Figure 1: Artistic Representation of the Vehicle

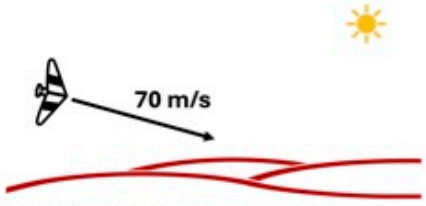


Figure 2: Artistic Representation of Vehicle Path

This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

3. SM-T.1 Full Provided Definitions of Desirability, Viability, and Feasibility and Anchored Scales (Expanded from Table 1)

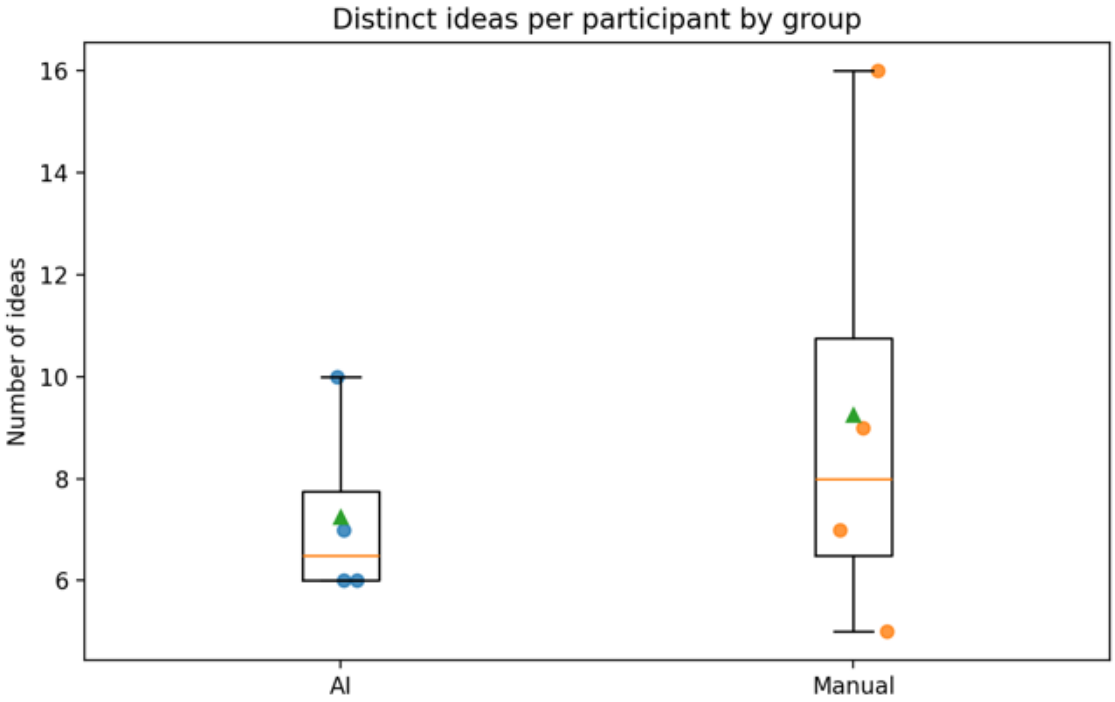
	Desirability	Viability	Feasibility
Score	***Is the concept likely to deliver value against the mission requirements?***	***Is the concept likely to be built, tested, shipped, and flown within the expected time, budget, and operational structure?***	***Is the technology behind the concept plausible to develop and implement?***
1	Minimal mission value; misses core outcomes (upright landing in Southern Highlands, days of comms and atmospheric sensing).	Poor fit; incompatible with available processes/partners, credible cost/schedule for class, and operations load.	Implausible given physics or current technology; no credible integration, assembly, or test path.
3	Partial mission value; covers landing but leaves gaps in post-landing science or comms value.	Conditional fit; plausible with changes, but major cost/schedule or integration unknowns remain.	Plausible but uncertain; partial path exists with key unknowns or major development needed.
5	High mission value; clearly maps approach to all core outcomes and articulates why the science/ops value is strong.	Strong fit; manufacturable with available processes/partners, credible cost/schedule for class, and operations load.	Plausible with current or near-term technology; clear integration, assembly, and test path with known risks and mitigations.

4. SM-T.2 Category Definitions (Corresponds to section 3.5)

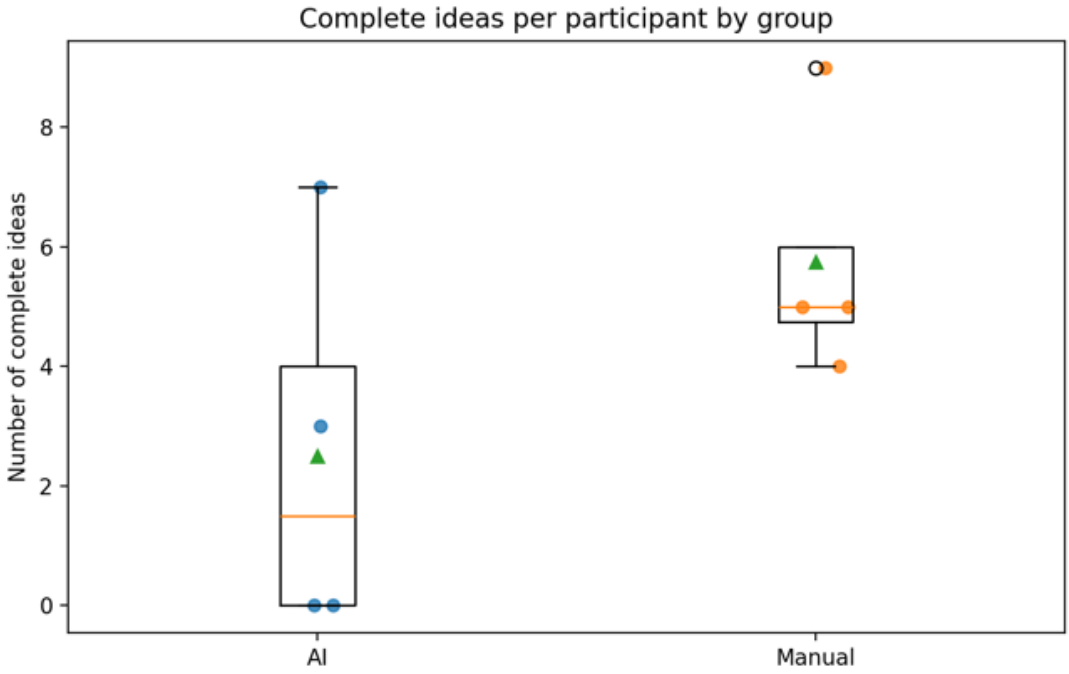
Category	Definition
Aerodynamic	Idea utilizes aerodynamic control surfaces and principles to land
Runway	Idea utilizes a landing gear, skids, or body slide over a distance
Mechanism	Idea utilizes some kind of additional mechanism to land
Parachute	Idea utilizes a parachute or similar device to significantly increase drag
Deforming Structure	Idea utilizes an inflatable or crushable structure to absorb impact energy
Rockets	Idea utilizes rocket or other chemical/jet propellant to arrest speed
Payload	Idea jettisons the main payload structure from the airframe
Anchor	Idea utilizes an anchor to arrest speed
Other	Idea does not explore landing method to point of being able to categorize
Composite	Idea utilizes two categories of landing method

This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

5. SM-F.1 DISTINCT IDEAS PER PARTICIPANT BY GROUP

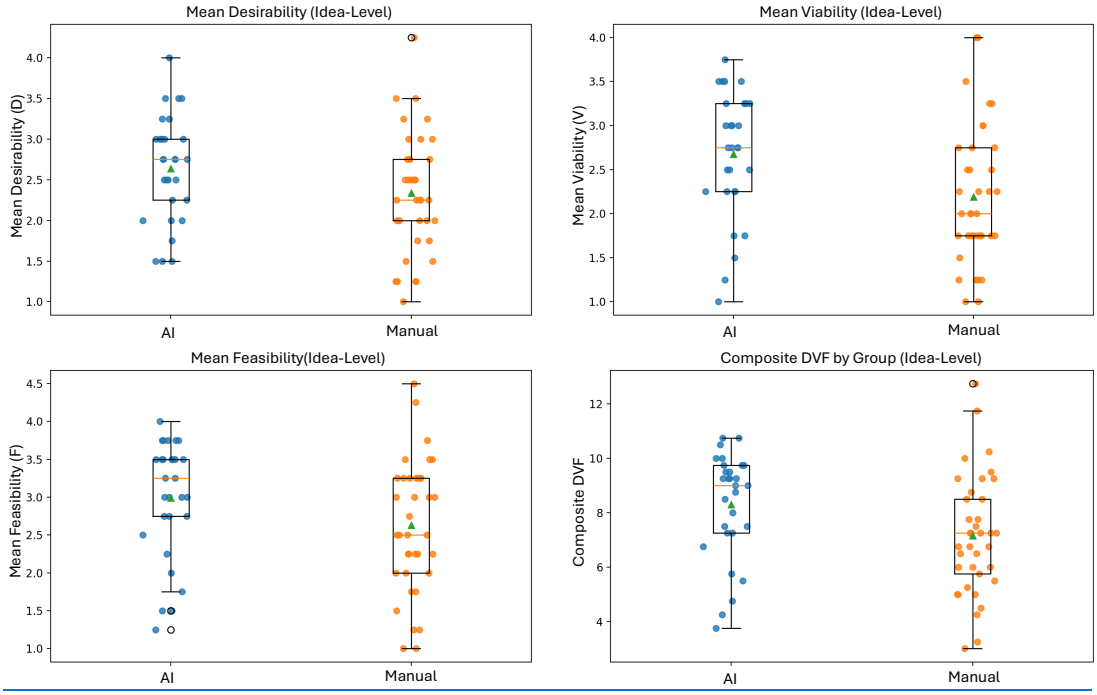


6. SM-F.2 COMPLETE IDEAS PER PARTICIPANT BY GROUP



This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

7. SM-F.3 DISTINCT IDEAS PER PARTICIPANT BY GROUP



8. SM-T.3 Top 5 Ideas Analysis (Expanded from Table 6, Corresponds to section 4.2)

Overall Rank	Idea ID	Participant ID	Group	DVF Composite	Mean D	Mean V	Mean F	Self Perceived Complete
1	7	2	Manual	12.75	4.25	4	4.5	True
2	51	8	Manual	11.75	3.5	4	4.25	
3	16	3	AI	10.75	3.25	3.75	3.75	
4	3	1	AI	10.75	3.5	3.5	3.75	
5	49	7	AI	10.5	3	3.5	4	True

9. SM-T.4 Top N Analysis of DVF Composite by Condition

N	AI Mean	Manual Mean	Difference	U	p	r (Effect)
1	10.44	10.63	-0.19	7.0	0.885	-0.125
2	10.06	9.53	+0.53	8.5	1.000	+0.062
3	9.83	9.02	+0.81	12.0	0.343	+0.500
4	9.64	8.61	+1.03	13.5	0.144	+0.688
5	9.32	8.28	+1.05	13.0	0.200	+0.625

This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

10. SM-T.5 Spearman Correlations Between Perceptions and Outcomes (All Participants)

Perception	Outcome	rho	p	n
Process Effectiveness	Total Ideas	-0.653	0.079	8
Process Effectiveness	Self Perceived Complete Count	0.168	0.691	8
Process Effectiveness	DVF Total By Participant	0.064	0.881	8
Instruction Clarity	Total Ideas	0.000	1.000	8
Instruction Clarity	Self Perceived Complete Count	0.342	0.407	8
Instruction Clarity	DVF Total By Participant	-0.282	0.499	8
Ease of Sketching	Total Ideas	0.380	0.353	8
Ease of Sketching	Self Perceived Complete Count	0.684	0.061	8
Ease of Sketching	DVF Total By Participant	-0.050	0.906	8
Was Method or Tool Helpful	Total Ideas	-0.235	0.576	8
Was Method or Tool Helpful	Self Perceived Complete Count	0.248	0.553	8
Was Method or Tool Helpful	DVF Total By Participant	0.355	0.389	8
Mission Alignment	Total Ideas	0.195	0.643	8
Mission Alignment	Self Perceived Complete Count	-0.312	0.451	8
Mission Alignment	DVF Total By Participant	-0.231	0.581	8
Would an Additional Tool Be Useful	Total Ideas	-0.607	0.110	8
Would an Additional Tool Be Useful	Self Perceived Complete Count	0.166	0.695	8
Would an Additional Tool Be Useful	DVF Total By Participant	0.000	1.000	8
Self Perceived Innovation	Total Ideas	-0.276	0.508	8
Self Perceived Innovation	Self Perceived Complete Count	0.456	0.257	8
Self Perceived Innovation	DVF Total By Participant	0.464	0.247	8
Self Perceived Output Quality	Total Ideas	-0.543	0.165	8
Self Perceived Output Quality	Self Perceived Complete Count	-0.250	0.550	8
Self Perceived Output Quality	DVF Total By Participant	0.000	1.000	8

This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

11. SM-T.6 Open Ended Survey Response Codebook

Code	Label	Scope	Definition	AI n = 4	Manual n = 4
T1	LLM as Knowledge Resource	AI-only	Participant describes the LLM as providing domain-relevant technical knowledge, on-demand feasibility evaluation, or analytical computations that informed their concepts.	4	0
T2	Prompt Literacy Learning Curve	AI-only	Participant reports difficulty formulating effective prompts, unfamiliarity with the text-based interaction paradigm, or time spent learning the tool rather than ideating.	3	0
T3	AI Output Convergence & Filtering Burden	AI-only	Participant reports that LLM responses converged on variations of a single concept or required effortful discarding of inapplicable ideas.	2	0
T4a	Sketching as Generative Asset (+)	Shared	Participant describes drawing, sketching, or visual annotation as positively contributing to idea generation, refinement, or enjoyment.	0	3
T4b	Text-Based Modality Friction (−)	Shared	Participant describes difficulty arising from the text-only AI interface, including trouble processing text-heavy output or missing visual-spatial reasoning affordances.	2	0
T5	Cognitive Fatigue & Sustaining Divergence	Shared	Participant describes mental fatigue, difficulty generating new ideas over time, or challenge of moving beyond initial concept families.	1	2
T6	Task & Constraint Framing	Shared	Participant comments on design constraints being too restrictive, the problem being over-specified, or time pressure limiting the design space.	1	3
T7a	Intrinsic Process Enjoyment	Shared	Participant expresses spontaneous enjoyment of the ideation experience itself—the act of thinking, creating, or engaging with the problem.	1	4
T7b	Instrumental Tool Satisfaction	AI-only	Participant expresses positive sentiment about what the AI tool could accomplish or produce, rather than about the experience of ideation itself.	3	0

This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.