Productive vs. Reflective: How Different Ways of Integrating AI into Design Workflows Affect Cognition and Motivation

Xiaotong (Tone) Xu
University of California, San Diego
La Jolla, CA, USA
xt@ucsd.edu

Arina Konnova*
University of California, San Diego
La Jolla, CA, USA
arkonnova@ucsd.edu

Bianca Gao*
University of California, San Diego
La Jolla, CA, USA
bigao@ucsd.edu

Cindy Peng*
University of California, San Diego
La Jolla, CA, USA
cipeng@ucsd.edu

Dave Vo*
University of California, San Diego
La Jolla, CA, USA
dtvo@ucsd.edu

Steven P. Dow
University of California, San Diego
La Jolla, CA, USA
spdow@ucsd.edu

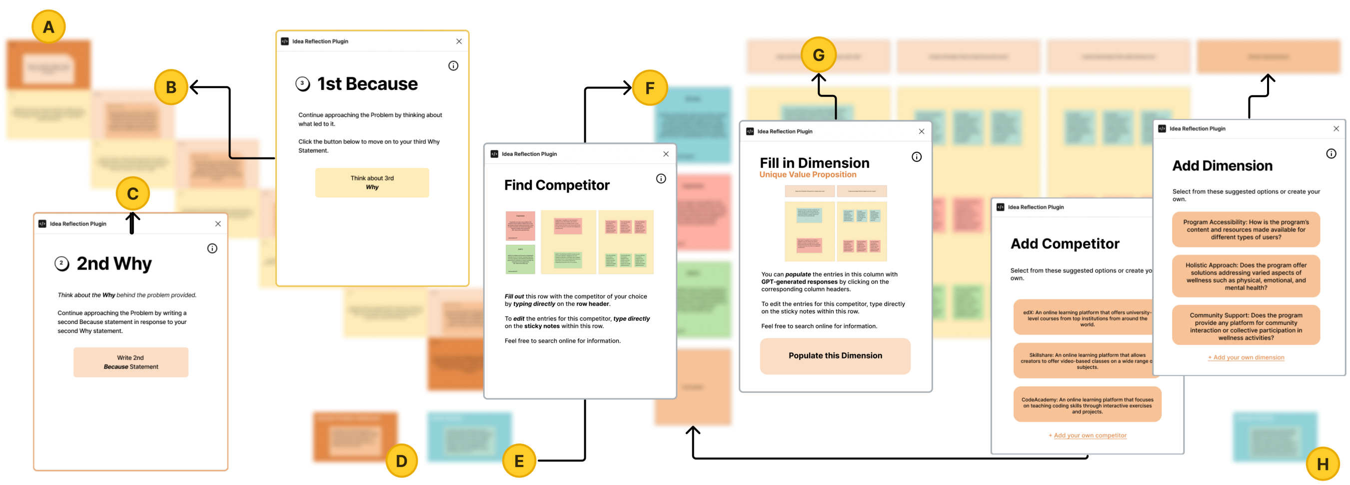

Figure 1: Overview of the Co-led condition. A. Participants were given a problem statement at the start of the study and saw a AI-generated reflective question on it in the box below. B. Participants then filled in the 'Because' box to address the reflective 'Why' questions. C. The 'Why' box prompted participants to reflect on the reasons for their prior 'Because' answer. (D) Participants wrote an iteration on the original problem statement, and then, (E) proposed a novel solution to address it. F. The AI offered competing solutions; participants could add competitors manually. G. AI offered key comparative dimensions, and allowed users to populate the columns or add new dimensions for comparison. H. Participants proposed a final solution after analyzing the competitive space.

Abstract

An increasing number of tools now integrate AI support, extending the ability of users—especially novices—to produce creative work. While AI could play various roles within such tools, less is known about how the positioning of AI affects an individual's cognitive processes and sense of agency. To examine this relationship, we built a collaborative whiteboard plugin that integrates an LLM into design templates to facilitate reflective brainstorming activities. We conducted a between-subjects experiment with N=47 participants assigned to one of three versions of AI-support—No-AI, AI input provided incrementally (Co-led) and AI provided all at once (AI-led)—to compare the allocation of cognitive resources. Results show that the positioning of AI scaffolds shifts the underlying cognition: AI-led participants devoted more time to comprehension and synthesis, which yielded more topically diverse problems and solutions. No-AI and Co-led participants spent more time revising content and reported higher confidence in their process.

*Four authors contributed equally to this research.

CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713649

CCS Concepts

• **Human-centered computing** → **Laboratory experiments**; **Empirical studies in HCI**; **Graphical user interfaces**; **Natural language interfaces**; *Computer supported cooperative work*.

Keywords

Creativity, Critical Thinking, Self-reflection, Learning, Brainstorming, Human-AI Collaboration, Agency, Co-piloting, Steerable AI, LLMs

ACM Reference Format:
Xiaotong (Tone) Xu, Arina Konnova, Bianca Gao, Cindy Peng, Dave Vo, and Steven P. Dow. 2025. Productive vs. Reflective: How Different Ways of Integrating AI into Design Workflows Affect Cognition and Motivation. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3706598.3713649

1 Introduction

The recent evolution of generative AI has fueled innovation in creativity-support tools across multiple domains [42, 74, 84]. Image editors can now replace specific objects or layers in an image with simple queries [1]. Document readers can now extract paper highlights for better understanding [53]. Multi-modal AI models can support even more advanced creative tasks like music composition [39, 54] and video production [13].

For many creative workflows, people perform a range of different ways of "thinking," from reflecting and synthesizing information to framing and reframing the problem, as well as ideating. In the ideation context, researchers have used LLM-generated inspirational text to support users to reframe or explore a design space, and thus generate more creative ideas [22, 37]. Some have argued that LLMs' tendency to hallucinate can propel innovation [27, 45]. On the other hand, research also shows how LLMs can have a homogenizing effect, leading to more similar ideas across groups compared to using paper-based ideation cards [2]. Furthermore, people often attribute ownership to AI-generated content but are reluctant to publicly acknowledge AI's involvement, reflecting a broader conversation about agency in human-AI co-creation [23].

The different ways of situating AI in human-AI co-creation can greatly impact the types of interaction, communication, and collaboration [68] which, in turn, can affect users' cognitive processes and sense of agency. Designers need to think critically about any newly discovered information [50], especially information coming from LLMs that are known to hallucinate or have inaccuracies. Researchers have argued that AI systems should augment, rather than replace, human expertise by maintaining user control while boosting productivity [36, 69]. Given the many AI roles in human-AI collaboration, such as "creative assistant" [31, 93], "creator of content" [16], "friend" [89], or a prompt for critical thinking [49], it is pivotal to understand how the positioning of AI affects cognition and motivation during creative work. Critical thinking in this paper largely refers to the ability to judge, evaluate, and adopt AI-generated content for one's creative work.

Recent research into human-AI collaboration has often framed this as implementing a "human-in-the-loop" technique for machine learning [62, 87], or designing a "co-pilot" [71] or "steerable AI" [29], where both the human and AI agents take actions which, in turn, influence subsequent actions by both parties. Systems can implement human-AI collaboration in a wide range of ways, from letting the AI primarily steer work to allowing humans to drive it while only offloading low-level tasks to AI. For example, consider the differences in the positioning of AI within systems intended to support reading and comprehension. The Paper Plain system [4] can summarize passages and then read the automated summary. The Scim system [25] can intelligently highlight papers for faster skimming. In yet another approach, the Papeo system [53] can locate paper content from talk videos, so readers can better understand both content. While human-AI co-creation take on many forms, less is know about how it affects the user's perception of cognitive load and sense of agency.

In this study, we aim to understand how different ways of integrating AI into a creative workflow impact an individual's ability to reflect and iterate. Our central question is: how does the positioning of AI within a creative workflow impact creative outcomes, the allocation of cognitive resources, and user perceptions of agency and overall value of AI? Heavy-handed AI support may introduce information that extends the individual's existing knowledge, but it might also mislead the user or leave them feeling less in control creatively. In contrast, lighter, more targeted AI support might require users to put in more effort but could strengthen their sense of agency.

To investigate this question, we developed an LLM-enhanced plugin for FigJam (an online collaborative whiteboard). The plugin augments the workflow for creative exercises using one of three ways to integrate LLM support: a No-AI condition, a Co-led condition, and an AI-led condition. For the Co-led and AI-led conditions, we integrated LLMs on the backend to help users gather external information for two common design templates: a "Five Whys" exercise and a "Competitive Analysis" matrix, following prior work [90]. We selected these templates not only because they require users to find and reflect on information (e.g., competitors), but also because their complexity allows us to explore how the different ways of positioning AI impact deeper forms of cognition, such as self-reflection and sense-making.

We ran a between-subjects experiment with N=47 participants randomly assigned to one of three conditions. In the **AI-led** condition, the system fills out the entire template with LLM-generated text, leaving participants with the responsibility of reviewing and editing this information. The **Co-led** condition similarly integrates LLM-generated text, but on a more piecemeal basis, allowing the user to read and provide input box-by-box in response to ideas within the template. As a baseline, in the **No-AI** condition, participants completed the template with the step-by-step instruction provided by the plugin, and no LLM-generated information was available at all. Across all conditions, participants completed two templates: first, they completed the "Five Whys" exercise, iterated on a problem statement, and brainstormed a potential solution; and second, they completed a "Competitive Analysis" around their solution and then iterated on their idea.

The research team analyzed data from several sources, focusing on understanding the impact of AI positioning on creative outcomes, cognitive processing, and user perspectives. To analyze creative outcomes, we coded "idea units" in participants' iterated problem statements, initial solutions, and iterated solutions. To investigate the impact on cognitive processes, the research team analyzed video recordings to code time allocation across various cognitive activities during the workflow. Survey responses provided insights into users' perspectives on their creative processes and the role of AI.

The study found that the presence and positioning of AI 1) Affected the creative outcomes of the design templates: The AI assistance for both interaction styles led to an expanded view of problem framing after completing the Five Whys template, compared to those with no AI support. However, after the Competitive Analysis template, we found that AI-led participants wrote final solutions that exhibited more topic diversity than the those in other conditions. 2) Shifted the focus of cognitive activity: Video analysis revealed differences in how participants allocated attention to different aspects of the design templates. AI-led participants devoted more time to comprehension and synthesis, while Co-led and No-AI participants spent more time writing and revising information in the template. 3) Shaped user perceptions: Co-led and No-AI participants reported higher confidence in their process and creativity of their initial solutions, and Co-led participants self-reported deeper concern for the underlying problems and focused on practical final solutions.

In summary, this paper makes the following contributions:

- We developed an interactive plugin with varying interaction mechanisms for integrating large language models (LLMs) into design templates. This system enhances the cognitive walkthrough process, enabling support for complex cognitive tasks such as self-reflection and iterative design exploration.
- We conducted a between-subjects experiment (N=47) with three conditions to investigate how different human-AI collaboration patterns impact user perspectives, the time spent on different cognitive activities, and the overall creative outcomes.
- The study results align with previous research that finds trade-offs between performance and agency, providing clues on how to integrate AI to optimize both.

2 Related Work

2.1 Human-AI Co-Creativity

Various systems have explored human-AI collaboration to enhance design processes. AI plays a role in inspiring, generating, exploring, combining, and transforming ideas for later expansion [8, 9], impacting the diversity, quantity, and novelty of ideas [49]. O'Toole and Horvát [64] identify five key AI applications in creativity: creative support, ideation support, personalization, co-creativity, and novelty. Guo et al. [30] and Zheng et al. [93] discuss AI's role as a "creative assistant," allowing users to offload repetitive tasks and focus on higher-level thinking. Tools like the Creative Sketching Partner (CSP) [47] and FashionQ [42] identify the assistive role of AI in supporting idea exploration and overcoming design fixation. These tools show how AI can extend human creativity during divergent and convergent thinking processes necessary for idea discovery [42, 72, 84]. ID.8 [3], an AI tool for co-creating visual narratives, identifies lowered barriers and increased access to creative expression as another effect of AI integration in creative processes.

AI's role extends beyond improving creative thought to becoming an active collaborator or "creator of content" [16], shaping human-AI relationships in idea generation. The extent of AI's involvement depends on the human collaborator. Wang et al. [85] found that the need for AI support varies by task . Wordcraft's AI agent [92] exemplifies this by serving as an idea generator, interpolator, and copy editor throughout the design process. However, Partlan et al. [65] argue that dependable AI results alone aren't enough for co-creativity; users need control over outputs to explore diverse ideas [2], manipulation of results, and providing of feedback [81], to keep them engaged in the process.

User control is crucial for affirming their roles as creative collaborators. Rajcic et al. [67] found that users feel more confident in their creative identity when they control AI tools. Jonsson and Tholander [46] highlight that the "friction" between inconsistent AI outputs and human input encourages reflection and rethinking. This interaction leads human collaborators to take on a curatorial role [18], using AI feedback to challenge initial ideas and decide what to keep, remove, or expand upon [26]. As a facilitator in discussions [28] or provocateur in workflows [70], or as a "friend" [89], AI alters human creative roles as they work to integrate it as a collaborator in their creative processes [77].

Current research highlights AI's role in idea generation but often focuses on how users receive information rather than how they collaborate with AI to create new ideas [40]. There is a need to explore "interaction dynamics" [68], like turn-taking, to foster communication between human and AI creators rather than competition [82]. Designing effective co-creative systems requires trustworthiness, adaptiveness [57], and alignment with user values [26] to build confidence in AI collaborators. This study aims to explore how AI can assist in early-stage creative processes, particularly in co-creative brainstorming.

2.2 AI as an Educational Thinking Toolkit

Although formal curricula for AI have been around for decades [15], in recent years, presence of AI in academic settings [63]—regardless of the field of study—is rapidly approaching ubiquity. With the growing accessibility of powerful LLMs to the average students and educators, the widespread adoption of AI in the field of education poses unique opportunities and challenges, such as access to information about a wide variety of topics [79] but limited quality control over the information [59], as well as plagiarism and academic integrity issues [24]. Generative AI chatbots have proved to be useful tools in learning new content [6, 56] and have received positive feedback on their impact [17, 21, 38], but there still remains a lot of terrain to cover when it comes to assessing long-term impacts.

From a theoretical standpoint, generative AI could improve critical thinking by offering fundamental information, creating more opportunities to navigate complex scenarios [83] or learn more in-depth in other modalities [88], but it could also impede critical thinking by generating completed results without necessitating human involvement [33]. The proven positive impacts of AI in academic settings include better computational thinking skills [91], stronger learner self-efficacy [43], more opportunities for reflection [52], and higher confidence in analysis and comprehension of complex concepts [32]. Furthermore, AI has been proven to help users think more comprehensively and thoroughly about problems posed in design thinking exercises [90]. However, there have also been scenarios in which utilizing AI can exacerbate existing weak points and challenges in critical thinking [66].

Exercising effective critical thinking skills is largely tied to the imperfection of AI, and the awareness of this imperfection by its users. Styve et al. [76] finds that students' "practices of critical thinking in programming increased" with "many students… acknowledg(ing) that responses from ChatGPT should not be taken at face value and should be critiqued further", and Shoufan [73] finds that "most (students) believe that (ChatGPT) requires good background knowledge to work with since it does not replace human intelligence". Research suggests that working in tandem with AI yields the best results [86], using it intentionally as a tool while also exercising one's uniquely human intelligence [58, 75]. Brondani et al. [12] finds that instructors could differentiate reflections written by students versus ChatGPT 85% of the time, but that thematic analyses of these reflections completed by qualitative researchers did not substantially differ from ChatGPT-generated analyses. Thus, the differences from engaging with human- or AI-produced work remain ambiguous. We sought to investigate these differences, as well as the impact of AI on critical thinking, in our study.

2.3 Agency in Human-AI Co-Work

The integration of AI in creativity-support tools has significantly expanded the creative possibilities for users, driving research on agency and control in human-AI collaboration. Agency in co-creation involves task delegation to AI systems and the human capacity to guide, critique, and collaborate with AI [19, 61, 78]. These frameworks emphasize the preservation of human agency to ensure meaningful interaction that shape engagement, learning, and reflective thinking.

Agency in AI systems can range from goal-directed assistance to autonomous actions, either assisting with set tasks or dynamically responding to changing contexts [55]. Hwang and Won [41] show that user perceptions of AI agency can be influenced by the system's embodiment, with users attributing more emotional experience to humanoid AIs and greater functional agency to cloud-based systems. Miller [60] and Buçinca et al. [14] both advocate reducing AI overreliance, with Miller [60] proposing "Evaluative AI" to enhance user control and human agency , and Buçinca et al. [14] emphasizing cognitive forcing interventions for more deliberate human thinking. These findings point to a core tension in co-creative systems: while AI can enhance productivity, it risks undermining user agency and critical engagement.

A key consideration for balancing human and AI agency relies on how agency distribution affects users' capacity for reflection and critical thinking. Heyman et al. [37] showed that structured prompts in human-AI interactions enhance creativity by promoting divergent thinking and iterative reflection, helping users refine AI-generated ideas and avoid fixation. Similarly, Lawton et al. [51] found that mixed-initiative systems, alternating between user control and AI suggestions, effectively balance creativity and structure, particularly in open-ended tasks like scene drawing, though they struggle in more structured tasks where user control is essential. Jiang et al. [44] stressed the need for human control in inductive analysis to prevent oversimplification of complex insights. Biermann et al. [7] found that writers who maintained control over creative tasks like character and dialogue generation had a stronger sense of ownership and resisted AI interventions. These results emphasize the importance of designing AI systems that foster not only productivity but also critical engagement, reflection, and contextual adaptability. We designed our own prototype with three versions as a research probe to rigorously examine the different balances between human- and AI-agency in a controlled experiment.

3 System Design

3.1 Design Workflow Integration

To facilitate AI integration in design thinking, we created an interactive plugin on an online collaborative whiteboard to guide the participants through their creative process. Across conditions, the plugin provides instructions for each step of the template to support participant interaction with the workspace. However, each condition provides different ways of engaging with the thinking process, affecting how participants think, write, generate, and reflect. We adopted templates for two common design activities: (1) reflective thinking for problem scoping through the "Five Whys" template and (2) sensemaking for ideation through the "Competitive Analysis" template [20, 90].

The "Five Whys" template (5Y) is a method for scaffolding a root cause analysis around a particular problem. By starting with the initial given problem statement and repeatedly asking the question "Why" (typically five times), the nature of the problem as well as its solution becomes clearer. This iterative technique is designed to interrogate the cause-and-effect relationships underlying a particular problem.

At the beginning of the exercise, participants are given an initial problem statement that is generic and relatable. Their goal is to analyze the problem and its key drivers through the Five Whys exercise. Then, the participants come up with an Iterated Problem Statement and brainstorm an appropriate Initial Solution.

After the participant finishes drafting their Initial Solution, their goal is to gain a better understanding and make sense of the solution space, particularly how their idea compares to existing solutions. This is done through the Competitive Analysis (CA) template provided by the digital whiteboard, a strategic method used to evaluate a new concept against the strengths and weaknesses of existing competitors within the market landscape. This analysis explores opportunities and threats, alongside insights on potential competitors. The CA templates compare the participant's original solution to a few competitors along comparative dimensions (e.g. the "Unique Value Proposition", "Advantages", and "Disadvantages").

Because the competitive analysis exercise has no particular order, rather than a "press-to-continue" flow like the Five Whys, the plugin offers different screens and intended actions for each column, row, and sticky notes within. It supports clicking the column header to edit comparative dimensions, clicking on a row header to edit competitors, and clicking the sticky note within a dimension to edit insights.

Once participants finish this exercise, they distill their insights into an Iterated Solution for the problem statement.

3.2 Condition Design

3.2.1 No-AI. The No-AI condition serves as the control group, representing how users typically approach a problem/solution without

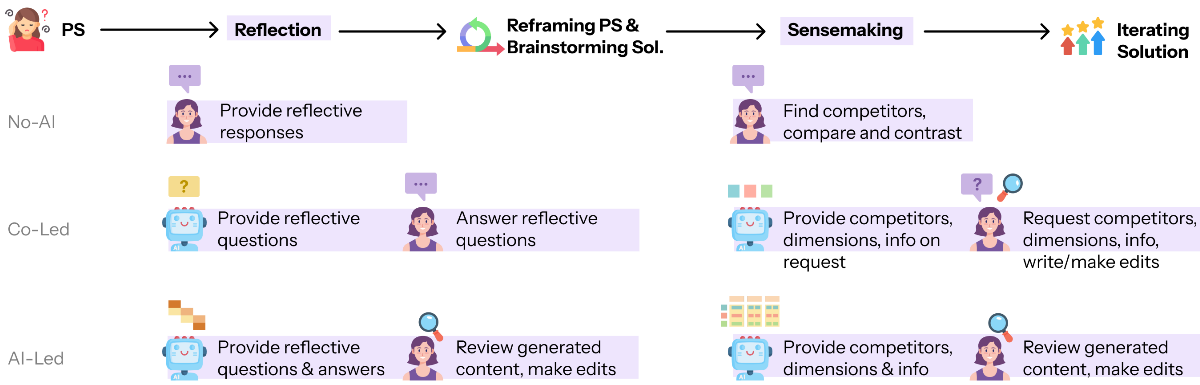

Figure 2: Study workflow broken down by human and AI roles across conditions. In the No-AI condition, participants completed design exercises without AI assistance. In the Co-led condition, participants responded to AI-generated reflective questions and received AI-gathered competitive analysis on demand. In the AI-led condition, participants responded to AI-generated reflective conversations (both questions and answers) and AI-gathered competitive analysis all at once.

LLM assistance. Participants use the default empty templates, filling them out manually based only on their current knowledge. They self-ask questions, respond, and reflect on their answers throughout the template. The plugin in this condition mainly serves to outline the purpose of each step in the template and provide cues that users can click on to move through the template. The main functions for the plugin in this condition are interactive.

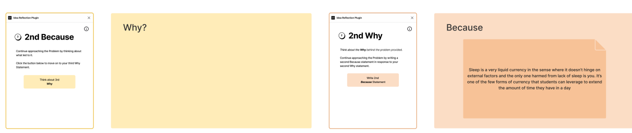

Figure 3: For the No-AI condition, every blank 'Why' box prompts users to reflect and write answers in the respective 'Because' box, after which they will continue to the next 'Why'.

Reflection and Problem Scoping. The plugin in the No-AI condition provides instructions for the participants to push forward their thinking. First, it prompts them to think of a reason 'Why' the problem occurs. The participants then click on a button to move onto writing a 'Because' statement, a response to that 'Why'. Eventually, they write the 'Root Cause' of the problem.

Brainstorming, Sensemaking and Iteration. For the No-AI condition, the competitors and sticky notes inside the columns start out empty. When completing this template, the participant must come up with competitors from their own knowledge or from using web search, filling in table properties manually. If the user wants to add more competitors or dimensions, they can manually add their own.

3.2.2 Co-Led. The Co-led condition introduces LLM generation features in various parts of the design templates to help users reflect during their thought process, as well as provide alternative ideas that could enrich users' understanding of the problem/solution space. Unlike the No-AI condition, participants are not required to

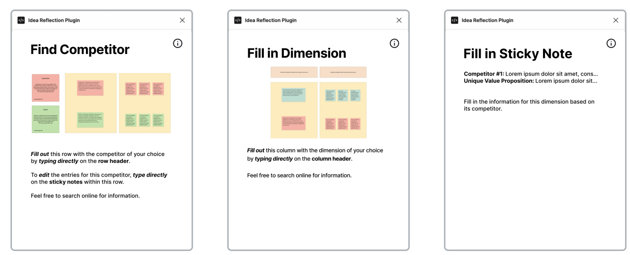

Figure 4: For the No-AI condition, clicking on competitor and dimension headings prompts users to fill out the rows and columns by typing, and clicking on an individual sticky note prompts the user to fill that cell out by typing.

initiate all of their reflective 'Why' thinking or manually write all of their responses. Reducing user initiative allows for more time to reflect on their answers, making the workflow both interactive and reflective.

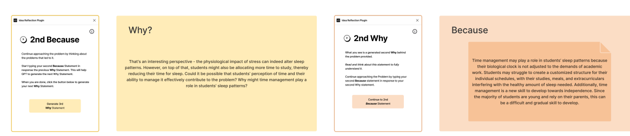

Figure 5: For the Co-led condition, the LLM generates reflective questions in the 'Why' box that the participant must manually answer in the corresponding 'Because' box. Then, they can go on to generate the next 'Why' question.

Reflection and Problem Scoping. In the Co-led condition, to enhance the typical workflow of the Five Whys exercise, the plugin generates thought-provoking questions based on the user's responses. After the participants type a response in the 'Because' box, they can click on the 'Generate Why' button. The plugin proceeds to adjust the user's viewpoint to center the next 'Why' box.

After a few seconds, the next 'Why' question will appear, which proposes a guiding question about a relevant aspect of the problem the user outlined in their reasoning.

Instead of writing the Root Cause manually, the user can generate the root cause of the initial problem by clicking on 'Generate Root Cause' in the plugin after answering all the 'Why' questions. The plugin will generate a concise takeaway from the template by synthesizing all 'Because' responses and 'Why' prompts together.

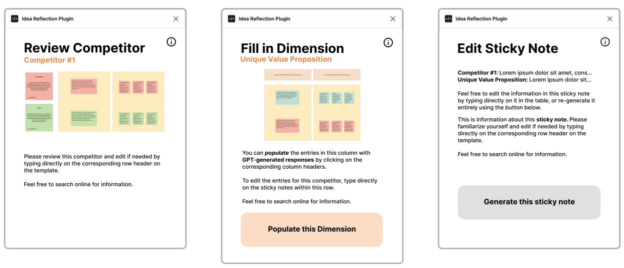

Figure 6: For the Co-led condition, once the competitors and dimensions are generated, participants can review competitors by clicking on the row headers, and populate the dimensions with LLM-generated information by clicking on the column headers. Participants can also generate information for individual cells by clicking directly on a sticky note.

Brainstorming, Sensemaking and Iteration. When the user is ready to move onto the CA exercise, the experimenter selects their initial solution. After a few seconds, the plugin loads their idea into the table, as well as generating three competitors and respective descriptions authored by the LLM. While the user can continue to fill in and edit the template manually, they also have different methods of generating content authored by the plugin.

The user can select an individual sticky note and click on 'Generate this Sticky Note' in the plugin. After a few seconds, the plugin produces insights based on the text in the competitor and dimension headers. The user can also generate an entire dimension. In this case, the user selects the dimension header and clicks on 'Populate this Dimension.'

If the user wants to add another competitor or dimension, the plugin will load three other competitors or dimensions generated by the LLM that are related to the user's initial solution. If the user chooses one of the suggested items, the table will expand accordingly, then add the header and empty sticky notes. Alternatively, the user can manually create their own by clicking the secondary button 'Add your own competitor/dimension' and proceed to fill it out manually similar to the No-AI group.

3.2.3 AI-Led. When users in the AI-led condition begin their tasks, the templates are already populated with content authored by the LLM. Unlike the previous two conditions, the participants do not initiate any writing, using all their time to read and mentally process information. The plugin in this condition guides the participants to review what the LLM generates and forms their understanding around what is presented, leading to a purely reflective workflow.

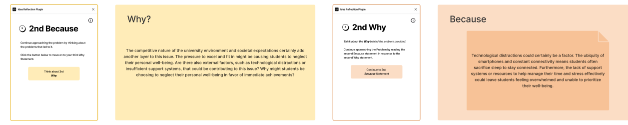

Figure 7: For the AI-led condition, both the 'Why' and 'Because' boxes are populated with LLM-generated text following the line of reasoning set by the problem statement.

Reflection and Problem Scoping. For the Five Whys exercise, all Why boxes have generated guiding questions, all Because boxes have generated responses, and the Root Cause is generated as well. The plugin tasks the participants with reviewing the 'Because' boxes, the 'Why' boxes, and the 'Root Cause' (making any edits if desired) to form a line of reasoning around the problem. Then, they proceed to write an Iterated Problem Statement and Initial Solution similar to the other conditions.

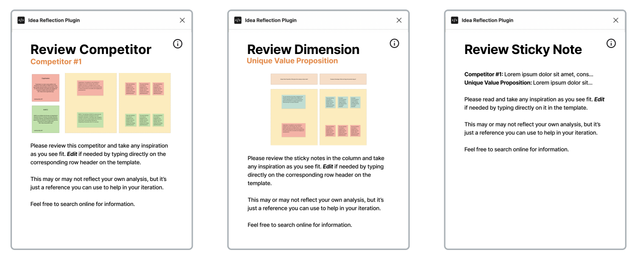

Figure 8: For the AI-led condition, participants can review the LLM-generated information for competitors, dimensions, and individual cells by clicking on the row and column headers and sticky notes respectively.

Brainstorming, Sensemaking and Iteration. When the user is ready to move onto the CA exercise, the experimenter selects their Initial Solution. This time, the plugin will generate three competitors and respective descriptions authored by the LLM, as well as populate all the sticky notes under each dimension.

Similar to the Five Whys exercise in the AI-led condition, the plugin instructs the participants to review the competitors, dimensions, and cells (making any edits if desired). However, they are not able to add more competitors or dimensions. Afterwards, the user writes an Iterated Solution based on the insights they gathered.

3.3 Tech Infrastructure

The system was built using TypeScript and the FigJam Plugin API[1] for front-end interface design, Firebase[2] for data storage, Python Flask[3] for a back-end server, and OpenAI's API to access a large language model (GPT-4)[4]. The system was loaded in the FigJam environment and used as a plugin, sending requests from the front-end interface to the Flask server. The server then prompted the

[1]https://www.figma.com/plugin-docs/api/api-reference/
[2]https://firebase.google.com/
[3]https://flask.palletsprojects.com/en/stable/
[4]https://platform.openai.com/docs/overview

OpenAI API, processed the responses, and returned them to the system's front end. The prompt questions were inspired by a previous study integrating LLMs into the design templates [90].

4 Study

4.1 Participants

A total of 53 individuals were recruited through the university student research participation system. Participants were compensated with extra course credits for their participation. Ultimately, the video recordings of 47 participants' study sessions were successfully captured and analyzed for this study. All participants were college students from a Western public university, with ages ranging from 18 to 32 years old. None of the participants were professional designers. The majority of participants were novices in engaging in design brainstorming activities – 53.3% of the participants had never engaged in any form of design thinking, 30% had only minimal experience, and 16.7% felt comfortable engaging in design thinking.

4.2 Procedure

The user study sessions were conducted in person, with one researcher working one-on-one with each participant. Zoom was used to view and record participants' activities on the computer screen, as well as to capture their responses during post-study interviews. Participants were randomly assigned to one of three experimental conditions, with each condition varying in the extent to which it incorporated the large language model (LLM) into the design thinking process. Of the 47 video recordings captured and analyzed, the breakdown among all three conditions (No-AI:Co-Led:AI-Led) was 16:16:15, indicating a near-even split of participants among the three conditions.

After participants filled out the consent form and pre-study survey, the experimenter introduced them to the problem statement and the generated Five Whys template, which was designed to take no more than 30 minutes to complete. Once participants completed the Root Cause, they were directed to iterate on the original problem statement based on their reflections and write their Iterated Problem Statement (IPS) on a designated sticky note. Finally, they brainstormed a solution to address the problem in their IPS, noting it on a sticky note labeled Initial Solution.

After writing their initial solution, participants transitioned to the sensemaking portion of the study, which took another 30 minutes. Participants were shown their idea in the topmost row of the template, followed by competitors' ideas addressing the same problem. They were instructed to consider and fill in columns for Unique Value Proposition, Advantages, and Disadvantages for each solution idea.

For all three conditions, the experimenter opened a web browser window and informed participants that they could use the internet for real-time online research. After processing the competitive analysis, they were asked to iterate on their Initial Solution and write their Iterated Solution on the sticky note labeled accordingly. Participants then filled out a post-study survey. Upon survey submission, the experimenter conducted an interview with each participant to delve into their thought process and experience.

4.3 Measurements and Analyses

The researchers analyzed the Initial Solution and Iterated Solution to uncover sources of inspiration and understand how participants developed their creativity from the initial problem framing to the final iteration. This involved revisiting video recordings to identify corresponding moments and interview responses, as well as examining original user work to pinpoint specific parts of the templates where participants' cursor movements indicated critical thinking or decision-making moments that contributed to their solution iteration. Inspirational sources were then documented for each solution iteration, detailing specific parts of the Five Whys and Competitive Analysis templates. These sources were also categorized and labeled to summarize the patterns and themes that emerged, providing an overall understanding of how participants drew inspiration and formed their creative solutions. All quantitative analyses were conducted in computational notebooks, and the code was reproducible.

4.3.1 Creative Outcomes. To understand how participants introduced new concepts throughout the process, four researchers coded idea units in participants' framed problem statements and solutions. A defined legend was used to clarify what constituted an "idea unit," with an idea unit being any new concept or added idea that was not simply a rewording or rearrangement of the original Problem Statement. Idea units were then grouped into broader topic categories [34] by related phrases, verbs, or prepositions, and represented different perspectives and considerations mentioned in participants' root causes, initial, and iterated solutions.

To ensure the validity and reliability of coding, four researchers collaborated to cross-check the coding process, discussing discrepancies in weighting or recontextualization of ideas and determining thresholds for similarity to the original wording. Each researcher individually coded approximately fifteen root causes and solutions across the three conditions with multiple rounds of team cross-checking to maintain accuracy. These idea unit categories were later used to calculate averages and standard deviations of idea units across conditions for each problem statement, contributing to an understanding of how participants' creative processes evolved during the iterations. Analysis of covariance (ANCOVA) was conducted to examine the conditional differences on the quantity and diversity of idea units coded by the experimenters, while controlling for the version of the initial problem statement.

4.3.2 Cognitive effort seen in video recordings. Video recordings from Zoom were reviewed to observe participant interactions and behaviors during the tasks, with key moments and responses noted for further qualitative analysis. The research team manually coded the start and end of key activities in the video recordings of each study session. The team inferred the activity at each moment based on the participant's cursor location within the design template and whether they were typing in the moment. For example, 'fill in template' was coded when participants typed in blank boxes in the templates; 'review template' was coded when participants moved cursor or canvas to prior responses in the templates without making edits; and 'edit template' was coded when participants moved back to and edited prior responses. The other categories coded in this analysis included 'web activity', 'scope/update problem', and 'brainstorm/revise/iterate solution' (see Figure 9 and 10).

Furthermore, when participants were on a certain template activity but not typing, we coded that as "Reflection time." The analysis excluded distracted activities such as experimenter instructions, interruptions, and system loading.

To ensure low data discrepancy and high reliability, a group of four researchers conducted four to six rounds of group coding on the same video recordings, rigorously comparing results to resolve discrepancies in both action categorization and time metrics calculation. After compiling a standard legend for consistent use and practicing to resolve all disagreements, each video was coded by two experimenters. The inter-rater reliability between coders was calculated and has reached over 89% agreement.

4.3.3 Survey Responses. The post-study survey asked participants to rate facets of their experience such as workload [35] and critical thinking activities [48] on five-point Likert scales ranging from "strongly agree" to "strongly disagree". Critical thinking activities were assessed through self-report questionnaires adapted from previous research on measuring critical thinking in learning contexts [48]. Example questions included: "I evaluated the claims, inferences, arguments, and explanations of the 'Why' questions for answering 'Because' statements," and "I constructed clear and coherent arguments in my root causes." Additionally, participants rated the usability [5] of the tasks with statements like: "I found the tasks easy to perform," "I felt confident using the templates to complete the tasks," "I needed to learn a lot of things to use the templates effectively," and "I think I would need the support of a technical person to be able to use the templates on my own in the future." Two surveys' data were excluded from this analysis due to a saving error.

4.3.4 Interview Scripts Thematic Analysis. Qualitative coding was employed using thematic analysis [10, 11] to identify common patterns and themes. To ensure high reliability and low discrepancy, four researchers paired up in groups of two and rotated responsibilities to analyze the interviews. Initially, all interview transcripts were transcribed onto a spreadsheet, documenting each participant's information in the following order: assigned experimenter, participant ID, and responses to individual interview questions.

After collecting key direct quotes, a detailed analysis was conducted to identify similar themes and patterns in participants' findings, experiences, opinions, and key takeaways. This process involved both quantitative and qualitative approaches. Quantitative analysis included collecting data about which parts of the templates were most helpful for solution iteration and comparing the confidence levels of participants regarding their filled-out templates. Additionally, data was gathered on how many participants expressed positive, neutral, or negative views on the integration of LLM concerning content accuracy and generation efficiency.

To further understand the interview responses, the research team revisited video recordings to identify corresponding moments, as well as examining original FigJam files to pinpoint specific parts of the templates where participants' cursor movements indicated critical thinking or decision-making moments that contributed to their solution iteration. Inspirational sources were then documented for each solution iteration, detailing specific parts of the Five Whys and Competitive Analysis templates. These sources were also categorized and labeled to summarize the patterns and themes that emerged, providing an overall understanding of how participants drew inspiration and formed their creative solutions.

5 Results

5.1 RQ1: How does the positioning of AI scaffolds within a design template affect creative outcomes?

All participants authored an iterated problem statement and an initial solution (after the Five Whys exercise), and an iterated solution (after the Competitive Analysis exercise). To understand differences in these outcomes across conditions, we measured "idea units" and number of topic categories covered within those key bits of content written by participants. On average, the initial brainstormed ideas covered the same amount of ideas across conditions (~3 idea units) (No-AI: M = 3.05, SD = 1.54; Co-led: M = 3.11, SD = 1.37; AI-led: M = 3.68, SD = 1.97) and the idea units increased by the time of the final solution (No-AI: M = 3.7, SD = 1.52; Co-led: M = 3.74, SD = 1.28; AI-led: M = 5.11, SD = 2.66). However, the analysis revealed no significant difference in the number of idea units in the iterated problem statement between groups (p = 0.11). Idea units in this paper are new ideas and concepts added to the responses and they are further combined to assess the topic convergence in participants' responses (see Section 4.3.1).

5.1.1 AI support helped expand individuals' view of the problem framing. Looking into the participants' iterated problem statements, we observed a significant difference in the number of topic categories per idea across conditions ($F(2,43)$ = 4.23, p = 0.02). Post-hoc analyses using Tukey's HSD revealed that No-AI participants generated significantly fewer topic categories compared to the Co-led condition (MD = -1.38, p = 0.02*, 95% CI: [-2.54, -0.21]) and the AI-led condition (MD = -1.48, p = 0.01, 95% CI: [-2.67, -0.29]). These results indicated that information from AI assistance led to more diverse problem frames; the No-AI participants were not as expansive in their thinking about the problem.

In the qualitative data, participants indicated that interacting with AI-generated responses, especially in the Co-led condition, felt like talking to somebody or oneself. This interaction led to more expansive thinking on different perspectives. P159 [Co-led] mentioned that the interaction with the LLM was like

> "...talking to somebody. We expanded on those thoughts (to) clarify what you were trying to think without trying to say too much, and so it would generate a couple of things that I wasn't really thinking, but it also sparked additional questions."

P123 [Co-led] also treated the AI-generated reflection questions as an equal dialogue partner, continuously building off each other's ideas so that the participant felt compelled to simply respond rather than revise: "...what I found easy was just answering the Why questions because it's just…talking to myself…And it gives a lot of thought because talking with yourself is pretty easy to do…you can just bounce ideas back and forth." Conversely, AI-led participants talked about struggling with writing and adding in their own ideas, "(My only struggle) is reiterating (the) problem in my own way because it was already really clear (P108 [AI-led])."

AI-led participants saw little room to build on already well-formed, completed responses. On the other hand, Co-led participants talked about how the Q&A format helped break down the thought process. As P158 [Co-led] described, the AI support "basically answers the 'Because' and then giving the 'Why' it helps better understand what it breaks it down for you. It gives you different perspectives." Although the AI-led condition instantly produced satisfactory answers, the Co-led interaction segmented the information, gave participants opportunities to reflect on each segment, and potentially added new information that could change the direction of the next AI response.

5.1.2 AI-led participants synthesized information into more diverse solutions. There was no significant difference in the total number of topics generated during the initial brainstorming phase across the three conditions ($p = 0.18$). However, participants in the AI-led condition covered significantly more topics in their iterated solution after working through the competitive analysis compared to those in the No-AI and Co-led conditions ($F(2,43) = 5.33$, $p = 0.009$). The AI-led input during the CA exercise led to more divergent solutions, even though Co-led participants were able to receive about the same external stimuli. The No-AI participants–limited to their own imagination–did not diverge as much, as expected.

The qualitative data suggests that reviewing AI-generated competitive analysis becomes less of an exercise in imagination, and more of an exercise in digesting and then remixing information. P133 [AI-led] identified the competitive analysis as a way for them to include new ideas not present before: "after reading the competitive analysis, I got a clearer idea of…what unique things my solution was missing. So I included that in my solution while…also giving a unique twist to it and making sure that it works for…my problem statement." P144 [AI-led] detailed their process of taking the information gained from competitive analysis to expand on their initial ideas: "My initial solution was a bit broad and less feasible. With my iterated solution, I thought more about what could be feasible, and I also took one of the generated solutions. And I basically just mix them in a way." They highlighted the competitive analysis and generation as an aid to their thinking: "It helped me think about all aspects of what to look for…and I didn't think more about some certain disadvantages, so it helped me definitely think about again what was more feasible and not." Both P133 [AI-led] and P144 [AI-led] felt that their initial solutions were "very broad" but took in a competitor or dimension from the AI-generated responses to help them think more concretely about their ideas.

5.2 RQ2: How does the positioning of AI scaffolds affect the way people allocate attention to different cognitive activities?

Following the coding procedure described above, the research team categorized each segment of participant screen video into the following cognitive activities: 'fill in/review/edit template', 'web activity', 'scope/update problem', and 'brainstorm/revise/iterate solution'. This time spent per activity is the sum of all time segments with that code. Analysis of covariance (ANCOVA) was conducted to examine the conditional differences on the time spent on each activity,

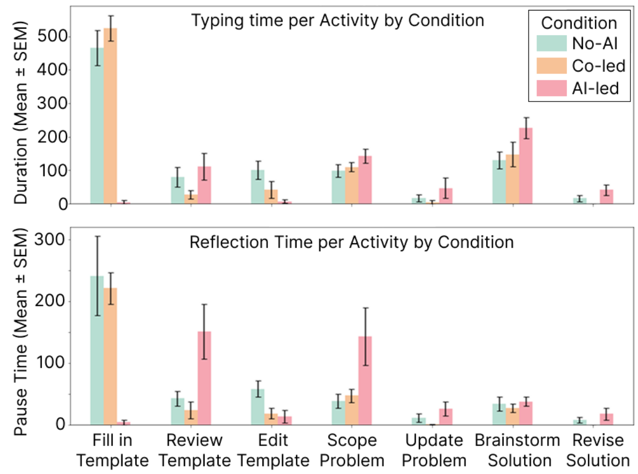

Figure 9: Typing Time per Activity by Condition (top) and Reflection Time per Activity by Condition (bottom) for the Five Whys exercise.

while controlling for the assigned version of the initial problem statement.

No-AI and Co-led participants spent significantly more time overall filling out the templates ($F(2,43) = 54.05$, $p < 0.001$). AI-led participants did not fill out the templates on their own, but comparably spent significantly more time reviewing the content ($F(2,43) = 3.37$, $p = 0.044$).

5.2.1 Increasing AI support shifts the focus of cognitive activity. Compared with No-AI and AI-led participants, Co-led participants were less likely to return to and review their finished responses in the template ($F(2,43) = 3.37$, $p = 0.044$) but still edited the finished responses more than AI-led participants ($F(2,43) = 5.17$, $p = 0.01$). Participants in the No-AI condition were significantly more likely to go back and revise previous text to improve clarity and expression compared to those in the AI-assisted conditions ($F(2,43) = 5.17$, $p = 0.0098$). Through the lens of reviewing and editing behavior, we saw that AI-led participants made comparatively fewer edits. In the absence of AI support, No-AI participants spent the most time on revising previously typed text, which often mean self-revising half-baked ideas or fixing grammar.

According to the qualitative data, these self-corrections may arise from moments of heightened awareness or deliberate selective attention. No-AI participants reported that they needed to take time to activate deeper thinking. P126 [No-AI] said that "the first three Whys was getting through what my initial natural reactions were. So you will see these last two contain more deeper thoughts, as something (that) doesn't come up to you immediately." No-AI participants appeared to selectively focus on a subset of issues. P102 [No-AI] mentioned that "I had to go back to the original problem statement because…the more detailed it got, the more difficult it was to remember what the big problem / big picture was."

Qualitative data suggest that AI-led participants checked back and forth between responses, not to ensure information integrity, but rather to check for information completeness. P110 [AI-led] said they "did a lot of going back and making sure (the generated root cause) was gathering all of the points. So (they) feel the root cause covered everything." This focus on reading and comprehending the AI generated text put emphasis on understanding the content. In contrast, No-AI participants spent a non-zero amount of time focused on the writing process (i.e., grammar, phrasing, or sentence structure), leaving less time to explore different content areas.

5.2.2 AI-led participants devoted more time to comprehension and synthesis. No significant difference was observed among the conditions in the time spent typing a revised problem statement ($p = 0.17$). However, No-AI and Co-led participants demonstrated greater fluency in updating the problem statement following the Five Whys exercise, as indicated by significantly less pause time after completing the templates and before typing the problem statement ($F(2,43) = 4.14$, $p = 0.024$).

After identifying the root cause of the problems, participants across all three conditions spent comparable amounts of time on thinking ($p = 0.68$), as evidenced by the pause time after previous tasks and before typing their initial solutions. Even though not significant, AI-led participants exhibited a marginal increase in the time spent typing their initial solutions in comparison to No-AI and Co-led participants ($F(2,43) = 2.62$, $p = 0.084$).

Qualitative data shows how even the baseline templates helped scaffold the No-AI participants, easing their cognitive processes and helping them approach the problem in different ways. P109 [No-AI] said: "I don't think I've ever thought about this issue this deeply before. But this template did help me think about not just one problem as to why, or one cause as to why people have this problem, but helped me approach it in different ways." It is important to note that the structure of the template and activities themselves provide scaffolding even without AI support.

When AI support was available, participants talked about how it challenged them to reflect and think in ways they would not have on their own. P123 [Co-led] brought up that, "When I think of something I just immediately think, wow! This is great, there's nothing else I can add onto that. And then when you ask a question Why you get these, you have to think more about how you can improve upon your initial statement."

AI-led participants viewed AI-generated responses as polished and logical. P133 [AI-led] mentioned that the AI-generated answers were "a tool to confirm the answer. I already knew the answer and it was a very fleshed out, clean, very well-written answer." P128 [AI-led] also described the responses as logical: "It's very logical. There's no really big leaps of assumptions." P136 [AI-led] mentioned, "I liked how it kept continuing once you get more in-depth into it." But some participants also found the AI-generated responses repetitive. P138 [AI-led] felt "some of (the Five Whys) were a little bit similar to each other… (and) could have gone further into more of an institutional issue than it did."

Although AI-led participants benefited from detailed and clear background information generated by AI, they exhibited indecision when selecting which issues to explore further. P111 [AI-led] expressed the difficulty in creating a single solution in response to the many identified reasons for the problem:

> "Just coming up with a solution that could tackle all the different ones was really hard, considering just for one single problem…The root cause came up with even more problems that I was thinking…how could I come up with different solutions for those programs? I think (what was difficult was) just trying to come up with a clear and good sized answer."

P105 [AI-led] also found it difficult to strike a balance among the given ideas they had to choose from: "(I) incorporated (these ideas) because I lingered longer on these two. I think I incorporated more of these…I also incorporated a little bit of that as well. But I think solutions are really hard to find (like) a balance, so I don't think my solution is perfect." The shift in cognitive focus freed up time for the AI-led participants to reflect more on information, and they also needed extra time due to the abundance of possible angles.

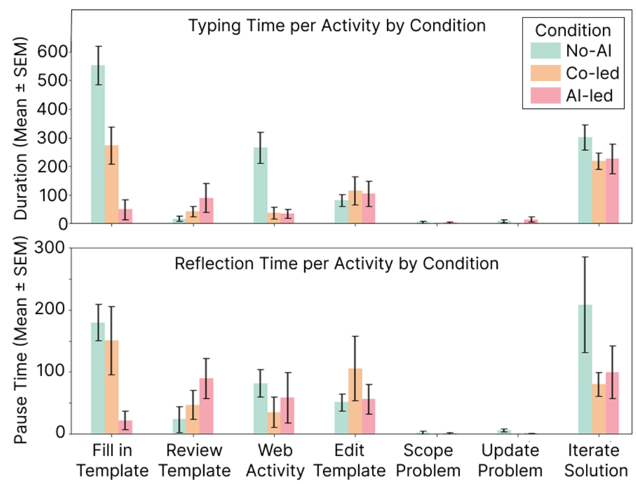

Figure 10: Typing Time per Activity by Condition (top) and Reflection Time per Activity by Condition (bottom) for the Competitive Analysis exercise.

5.3 RQ3: How does AI positioning affect user perceptions about the value of design templates?

5.3.1 No-AI and Co-led participants reported higher confidence in their process and creativity of initial solutions. Participants in the No-AI and Co-led conditions reported significantly higher levels of cognitive engagement and decision-making during the activity compared to those in the AI-led condition in the post-study survey ($F(2,41) = 4.90$, $p = 0.012$). Qualitative data from AI-led participants showed that they tended to summarize the AI responses but did not engage in critical thinking on those responses. P128 [AI-led] stated: "I was trying to summarize what they're saying and what their argument is. I'm taking it as their argument…I don't think I did much independent thinking."

Co-led and No-AI participants also rated their initial brainstorming solutions significantly more creative than AI-led participants did ($F(2,41) = 4.93$, $p = 0.012$). Qualitative data suggest that Co-led

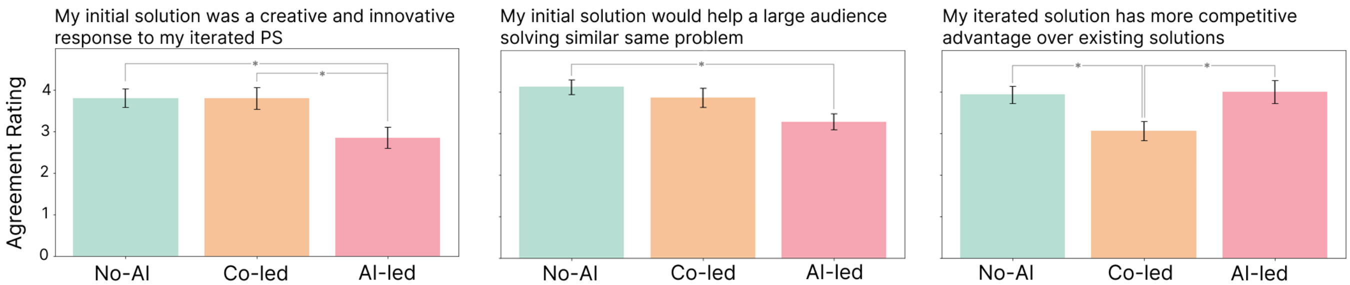

Figure 11: From left to right: (1) No-AI and Co-led participants showed higher self ratings in initial solution creativity. (2) No-AI participants provided higher ratings on whether their solution catered to a larger audience than AI-led participants. (3) Co-led participants rated their final solutions as less novel compared to the No-AI and AI-led participants.

and No-AI participants thought deeply about the problem beyond simply digesting external information about it, giving them more confidence in their initial solution. P102 [No-AI] mentioned that: "…If I just saw the problem statement and it was 'Students constantly have a hard time keeping up with their coursework'...if that's the only part I saw, my solution would just be 'become more effective at time management,' which doesn't say much…But because you had to think deeper about it, my solution was a lot more detailed in (like) the actual ways to be effective with time management."

No-AI participants also believed that their initial solution better targeted a more diverse audience in comparison with AI-led participants from the post-study survey (MD=0.84, p = 0.016). For example, as P146 [No-AI] analyzed the problem, they considered other perspectives: "But then I also considered a general population…I did consider other schools, not just universities, but also community (and) especially high school students too."

5.3.2 Co-led participants showed deeper concern for the underlying problems and focused on practical final solutions. Similar to the reflective exercise (5 Whys), No-AI and Co-led participants reported significantly more on thinking, decision-making, and time pressure for the sensemaking of existing design space (Competitive Analysis) in the post-study survey. However, Co-led participants self-rated their iterated solutions as less competitive than existing solutions compared with No-AI and AI-led participants ($F(2,43)$ = 5.58, p = 0.0071).

Qualitative data suggest that Co-led participants focused more on practicality than novelty. This may also be associated with the lower self-rating on the competitive advantages of their solutions.

> As P113 [Co-led] said: "My competitive analysis statement: it was much less detailed than my previous answers. A lot of the ideas that were generated over here give the company unique value. (But) I was not thinking…from a company perspective. I was more thinking of a generalization, or a cultural change that would be required? The solving of a root cause cannot be done by something commercial."

Individual Co-led participants indicated a more cynical view and deep consideration of the problem. For example, P142 [Co-led] brought up that "I came with terms of reality, and I know that my solution is a little harsh and a little boring to a lot of people, but I had to make it very clear…that it's gonna take time. It's not just like a get rich, quick scheme. It's more like a mindset." The Co-led interaction created a sort of dialogue or feedback loop that helped this participant align their thinking with real-world constraints.

6 Discussion

Our study results provide insights into how different approaches to positioning AI within a design thinking exercise influence creative outcomes, cognitive processes, and user perceptions. Findings revealed that AI assistance, whether interactive (Co-led) or passive (AI-led), offered diverse perspectives and fostered more divergent thinking during problem scoping compared to the absence of AI assistance (No-AI), as indicated by the greater diversity of topics covered in participants' idea descriptions. There was no statistically significant difference in topic diversity between initial solutions across conditions. Additionally, completing the competitive analysis, whether with AI support (Co-led) or without (No-AI), had minimal impact on how users incorporated competitive analysis information into their revised solutions.

During the design thinking exercises, participants in the No-AI and AI-led conditions spent more time reviewing and referencing responses from previous template blocks than Co-led participants. This was reportedly due to their attempts to recall information previously filled in the templates, though the volume of information was overwhelming. In post-study interviews, AI-led participants explained that they reviewed templates to verify whether information had been accurately carried forward within the workflow. In contrast, study video recordings showed that No-AI participants spent time re-editing previous responses as their thoughts evolved through rumination. There was no statistical difference in the typing time of initial solutions, but AI-led participants, who did not contribute to the templates, spent the most time brainstorming before typing compared to other groups.

In summary, AI-led participants spent more time reviewing LLM responses, giving them more time to adopt information into their ideas. However, they also provided low self-reported ratings on their solution creativity compared to Co-led participants, suggesting they did not have the same level of confidence and ownership of their ideas. In contrast, Co-led participants benefited from interacting with the AI assistance for divergent thinking but also had more autonomy and opportunities to express and ruminate on their thinking in the process. Human-led participants still benefited from the interactivity of templates that broke down an otherwise complex activity into procedures. While they did not have AI assistance, they went through the reflections and information gathering, which took more time and effort than reviewing or critiquing.

6.1 Positioning AI to Optimize Creativity

6.1.1 Interaction Dynamics with AI for Creativity. Our study builds upon the concept of "interaction dynamics" to highlight the role of turn-taking and user control in building an effective human-AI partnership by providing varying AI involvement in different conditions. We expand the frameworks by explicitly testing three agency models: (1) In the No-AI condition, users retain full control to own autonomy. (2) In the Co-led condition, we use a mixed-initiative approach that allows for shared agency to balance human input and AI suggestions. (3) In the AI-led condition, users gain agency through reflective engagement rather than active control. The current work systematically compares how varying levels of user and AI agency impact creativity, reflection, and productivity to contribute new insights into designing AI systems that foster balanced collaboration. This expands frameworks that often focus on just a single agency dynamic.

Unlike some of the existing systems that provide scaffolding reactively based on user behavior, our tool incorporates proactive scaffolding, especially in the AI-led condition, where prompts are pre-generated, encouraging users to reflect on pre-existing questions without initiating their own. This differs from tools like Wordcraft [92] that require user-driven initiation. The Co-led condition provides a mix of both proactive and user-initiated scaffolding, allowing participants to engage with AI-generated prompts while still actively contributing their ideas. However, iterative interaction with AI often leads to repetitiveness and errors, which can cause skepticism or aversion. Despite this, the active engagement with reflective thinking prompted by AI led participants to place more value on the feasibility and practicality of ideas, rather than their novelty. As a result, some of them maintained their core ideas with minimal influence from AI as a takeaway. These findings suggest that future AI interventions in human-AI co-creativity could integrate reflective questions with specific focus areas, priming users to develop downstream ideas in targeted directions. By aligning scaffolding approaches with user needs, AI systems can better foster meaningful creativity and balanced collaboration.

6.1.2 AI as Creative Assistant. Previous frameworks often position AI as either an "assistant" or a "collaborator," highlighting varying degrees of user control. Buçinca et al. [14] describe cognitive forcing interventions, where AI prompts users to deliberate more deeply. Heyman et al. [37] demonstrate the use of AI for divergent thinking by providing alternative suggestions to expand users' idea space. Implicitly, this paper explores different forms of AI assistance in a creative context. Building on previous work [90], this paper presents three paradigms of interaction: no AI, interactive AI scaffolding that leaves room for user inquiries, and exemplified text generated by AI without interactivity. Our tool incorporates an iterative scaffold that encourages users to explore, critique, and revise their ideas, going beyond idea generation to broaden the idea space through reflective questions.

Studies such as Partlan et al. [65] emphasize the importance of narrowing down ideas to a solution through AI support, which aligns with our tool's approach. Our tool allows users to evaluate their solutions against AI-generated competitive analysis to understand strengths and weaknesses. Our results show that without AI assistance, people tend to review content back and forth—either by editing existing content or conducting web searches—to select the most valuable information to carry forward to the next stage. The Co-led condition scaffolds depth through iterative back-and-forth interactions, enabling users to refine their ideas using AI-generated competitive insights. LLMs provide impromptu guidance, prompting users to reflect on their ideas based on the current status of their work. When workflows involve reviewing AI-generated information, users aim to gather comprehensive or complementary insights to proceed. They tend to perceive idea quality as a combination of as many ideas as possible rather than focusing on refining a single idea. Future studies could explore offering AI scaffolds as creative assistants at different intervals or through inquiries tailored to varying information needs.

Building on existing expert schema, AI interactivity introduces new opportunities to combine scaffolding methods for creativity. Our tool integrates design templates as a foundation for developing AI scaffolds, enhancing user engagement with complex cognitive tasks such as self-reflection and iteration. More specifically, the templates inspired the LLM prompts as well as the placement and presentation of AI responses. Beyond the expert cognitive scaffolds embedded in the templates, future studies could explore other media for expanding scaffolding approaches in supporting human-AI co-creativity.

6.2 From Accelerating Creativity to Supporting Reflection

In prior works, frameworks for critical thinking position humans as the facilitators. Spector and Ma [75] identify three stages in developing human intelligence: experience, inquiry, and the 3Rs (re-examine, reasoning, reflection). They argue that critical thinking is achieved only through curiosity. Wang et al. [86] explores an AI-supported language learning tool that employs various text and speech modalities. Sarkar [70] discusses the need to design AI as a provocateur in human-AI collaboration. In addition to fostering deeper thinking, users also applied organizational approaches and found motivation from the AI. Our paper expands upon these frameworks by integrating organizational features and feed-forward AI generation to provide users with opportunities to reflect and synthesize insights.

Our observations revealed that AI-led participants felt productive because the AI provided logical and well-written responses, aligning with their expectations. In contrast, Co-led participants effectively expanded their critical reflection, despite being challenged in their thinking. AI-led participants reported the highest average ratings for evaluating information in the template among the three groups. However, post-study interviews indicated that participants were more likely to agree with the provided information than engage in independent thinking. This is noteworthy because prior research has shown that people tend to under-rely on AI if they do not trust their performance [80].

In human-AI co-creativity, where ground truth may be non-essential but practicality still matters, people may lack critical thinking about the information provided to them. Conversely, using AI to generate reflective questions proved effective in encouraging participants to consider issues they might not have otherwise addressed. AI-generated reflective questions helped participants focus on core issues rather than superficial details. For instance, some No-AI participants, without AI assistance, reported that their initial responses were reactionary rather than focused on solving the problem.

However, these reflective questions were sometimes repetitive or emphasized details unrelated to participants' priorities, potentially hindering their natural thought processes. To design future human-AI systems for critical and reflective thinking, AI-generated responses should be logical and meet user expectations, while reflective questions should be diverse and novel to stimulate deeper engagement.

6.3 Limitations and Future Studies

Our study was conducted using common and representative design exercises, such as reflective thinking and sensemaking templates in a design brainstorming context. The scope of our study allowed us to select two exercises, but there is significant potential for training people's general critical thinking skills through interactions with AI. Future human-AI collaboration for creativity research can continue to explore this area. Another limitation arises from the pre-set prompts based on the templates. For example, the competitive analysis prompt only offered tech competitors on the backend. Therefore, participants who started with systemic or social solutions later switched to apps and companies after seeing those were generated as competitors. Future tools may enable users to customize prompts to better fit their information and creative needs. In addition, future studies can also recruit participants with varying levels of design and AI experience. Our participant pool primarily consisted of novices and AI laypeople, as they use novel tools based on intuition. A balanced pool of users for these exposures to design and AI may result in a more comprehensive observation of user behaviors.

7 Conclusion

We integrated an LLM into existing design workflows in three ways (No-AI, Co-led, and AI-led) and used it to conduct an in-person lab experiment with 47 participants to investigate how different AI positioning affected creative outcomes, cognitive processes, and user perceptions. Our results show that the positioning of AI scaffolds shifts the underlying cognition: AI-led participants devoted more time to comprehension and synthesis, while the other conditions spent more time writing and revising information in the template. As a result, Co-led and No-AI participants reported higher confidence in their process and creativity of initial solutions, but AI-led participants wrote problem statements and final solutions that exhibited more topic diversity than other conditions. The study results align with previous research that finds tradeoffs between performance and agency, and provides clues on how to integrate AI to optimize both.

Acknowledgments

We would like to thank everyone in ProtoLab who has provided constructive feedback on this work, especially Jenny Mar and Sydney Zhang for developing early versions of the prototype. This work is supported by the UCSD Cognitive Science Department through teaching positions and NSF funding (#2009003).

References

[1] Adobe. [n. d.]. Generative Fill - AI Fill in Image - Adobe Photoshop. https://www.adobe.com/products/photoshop/generative-fill.html

[2] Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. 2024. Homogenization effects of large language models on human creative ideation. In *Proceedings of the 16th Conference on Creativity & Cognition*. 413–425.

[3] Victor Nikhil Antony and Chien-Ming Huang. 2024. ID. 8: Co-Creating visual stories with Generative AI. *ACM Transactions on Interactive Intelligent Systems* 14, 3 (2024), 1–29.

[4] Tal August, Lucy Lu Wang, Jonathan Bragg, Marti A Hearst, Andrew Head, and Kyle Lo. 2023. Paper plain: Making medical research papers approachable to healthcare consumers with natural language processing. *ACM Transactions on Computer-Human Interaction* 30, 5 (2023), 1–38.

[5] Aaron Bangor, Philip T Kortum, and James T Miller. 2008. An empirical evaluation of the system usability scale. *Intl. Journal of Human–Computer Interaction* 24, 6 (2008), 574–594.

[6] Jose Belda-Medina and Vendula Kokošková. 2023. Integrating chatbots in education: insights from the Chatbot-Human Interaction Satisfaction Model (CHISM). *International Journal of Educational Technology in Higher Education* 20, 1 (2023), 62.

[7] Oloff C Biermann, Ning F Ma, and Dongwook Yoon. 2022. From tool to companion: Storywriters want AI writers to respect their personal values and writing strategies. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference*. 1209–1227.

[8] Volker Bilgram and Felix Laarmann. 2023. Accelerating innovation with generative AI: AI-augmented digital prototyping and innovation methods. *IEEE Engineering Management Review* 51, 2 (2023), 18–25.

[9] Margaret A. Boden. 1998. Creativity and artificial intelligence. *Artificial Intelligence* 103, 1 (1998), 347–356. https://doi.org/10.1016/S0004-3702(98)00055-1 Artificial Intelligence 40 years later.

[10] Virginia Braun and Victoria Clarke. 2012. *Thematic analysis.* American Psychological Association.

[11] Virginia Braun and Victoria Clarke. 2023. Toward good practice in thematic analysis: Avoiding common problems and be(com)ing a knowing researcher. *International Journal of Transgender Health* 24, 1 (2023), 1–6. https://doi.org/10.1080/26895269.2022.2129597 arXiv:https://doi.org/10.1080/26895269.2022.2129597

[12] Mario Brondani, Claudia Alves, Cecilia Ribeiro, Mariana M Braga, Renata C Mathes Garcia, Thiago Ardenghi, and Komkham Pattanaporn. [n. d.]. Artificial intelligence, ChatGPT, and dental education: Implications for reflective assignments and qualitative research. *Journal of Dental Education* ([n. d.]).

[13] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. 2024. Video generation models as world simulators. (2024). https://openai.com/research/video-generation-models-as-world-simulators

[14] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 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 (2021), 1–21.

[15] Francisco J Cantú-Ortiz, Nathalíe Galeano Sánchez, Leonardo Garrido, Hugo Terashima-Marin, and Ramón F Brena. 2020. An artificial intelligence educational strategy for the digital transformation. *International Journal on Interactive Design and Manufacturing (IJIDeM)* 14 (2020), 1195–1209.

[16] Eva Cetinic and James She. 2022. Understanding and creating art with AI: Review and outlook. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 18, 2 (2022), 1–22.

[17] Yu Chen, Scott Jensen, Leslie J Albert, Sambhav Gupta, and Terri Lee. 2023. Artificial intelligence (AI) student assistants in the classroom: Designing chatbots to support student success. *Information Systems Frontiers* 25, 1 (2023), 161–182.

[18] Li-Yuan Chiou, Peng-Kai Hung, Rung-Huei Liang, and Chun-Teng Wang. 2023. Designing with AI: an exploration of co-ideation with image generators. In *Proceedings of the 2023 ACM designing interactive systems conference*. 1941–1954.

[19] Nazli Cila. 2022. Designing Human-Agent Collaborations: Commitment, responsiveness, and support. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 420, 18 pages. https://doi.org/10.1145/3491102.3517500
[20] Design Council. [n. d.]. The Double Diamond. ([n. d.]). https://www.designcouncil.org.uk/our-resources/the-double-diamond/
[21] Yun Dai, Ching-Sing Chai, Pei-Yi Lin, Morris Siu-Yung Jong, Yanmei Guo, and Jianjun Qin. 2020. Promoting students' well-being by developing their readiness for the artificial intelligence age. *Sustainability* 12, 16 (2020), 6597.
[22] Giulia Di Fede, Davide Rocchesso, Steven P Dow, and Salvatore Andolina. 2022. The idea machine: LLM-based expansion, rewriting, combination, and suggestion of ideas. In *Proceedings of the 14th Conference on Creativity and Cognition*. 623–627.
[23] Fiona Draxler, Anna Werner, Florian Lehmann, Matthias Hoppe, Albrecht Schmidt, Daniel Buschek, and Robin Welsch. 2024. The AI Ghostwriter Effect: When Users do not Perceive Ownership of AI-Generated Text but Self-Declare as Authors. *ACM Trans. Comput.-Hum. Interact.* 31, 2, Article 25 (feb 2024), 40 pages. https://doi.org/10.1145/3637875
[24] Yogesh K Dwivedi, Nir Kshetri, Laurie Hughes, Emma Louise Slade, Anand Jeyaraj, Arpan Kumar Kar, Abdullah M Baabdullah, Alex Koohang, Vishnupriya Raghavan, Manju Ahuja, et al. 2023. Opinion Paper:"So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management* 71 (2023), 102642.
[25] Raymond Fok, Hita Kambhamettu, Luca Soldaini, Jonathan Bragg, Kyle Lo, Marti Hearst, Andrew Head, and Daniel S Weld. 2023. Scim: Intelligent skimming support for scientific papers. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 476–490.
[26] Katy Ilonka Gero, Tao Long, and Lydia B Chilton. 2023. Social dynamics of AI support in creative writing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–15.
[27] Bijit Ghosh. 2023. How LLM Hallucinations Propel Scientific Innovation. *Medium* (December 28 2023). https://medium.com/@bijit211987/how-llm-hallucinations-propel-scientific-innovation-0076ea7ac33d Accessed: 2025-02-11.
[28] Gabriel Enrique Gonzalez, Dario Andres Silva Moran, Stephanie Houde, Jessica He, Steven I Ross, Michael J Muller, Siya Kunde, and Justin D Weisz. 2024. Collaborative Canvas: A Tool for Exploring LLM Use in Group Ideation Tasks.. In *IUI Workshops*.
[29] Alberto González Martínez, Billy Troy Wooton, Nurit Kirshenbaum, Dylan Kobayashi, and Jason Leigh. 2020. Exploring collections of research publications with human steerable AI. In *Practice and Experience in Advanced Research Computing*. 339–348.
[30] Alicia Guo, Pat Pataranutaporn, and Pattie Maes. 2024. Exploring the Impact of AI Value Alignment in Collaborative Ideation: Effects on Perception, Ownership, and Output. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 152, 11 pages. https://doi.org/10.1145/3613905.3650892
[31] Xinyu Guo, Yi Xiao, Jiaqi Wang, and Tie Ji. 2023. Rethinking designer agency: A case study of co-creation between designers and AI. (2023).
[32] Ying Guo and Daniel Lee. 2023. Leveraging chatgpt for enhancing critical thinking skills. *Journal of Chemical Education* 100, 12 (2023), 4876–4883.
[33] Ariel Han and Zhenyao Cai. 2023. Design implications of generative AI systems for visual storytelling for young learners. In *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference*. 470–474.
[34] Gunnar Harboe and Elaine M. Huang. 2015. Real-World Affinity Diagramming Practices: Bridging the Paper-Digital Gap. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 95–104. https://doi.org/10.1145/2702123.2702561
[35] Sandra G Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. Sage publications Sage CA: Los Angeles, CA, 904–908.
[36] Jeffrey Heer. 2019. Agency plus automation: Designing artificial intelligence into interactive systems. *Proceedings of the National Academy of Sciences* 116, 6 (2019), 1844–1850.
[37] Jennifer L Heyman, Steven R Rick, Gianni Giacomelli, Haoran Wen, Robert Laubacher, Nancy Taubenslag, Max Knicker, Younes Jeddi, Pranav Ragupathy, Jared Curhan, et al. 2024. Supermind Ideator: How Scaffolding Human-AI Collaboration Can Increase Creativity. In *Proceedings of the ACM Collective Intelligence Conference*. 18–28.
[38] Mohammad Hmoud, Hadeel Swaity, Nardin Hamad, Omar Karram, and Wajeeh Daher. 2024. Higher education students' task motivation in the generative artificial intelligence context: the case of chatgpt. *Information* 15, 1 (2024), 33.
[39] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M Dai, Matthew D Hoffman, Monica Dinculescu, and Douglas Eck. 2018. Music transformer. *arXiv preprint arXiv:1809.04281* (2018).
[40] Angel Hsing-Chi Hwang. 2022. Too late to be creative? AI-empowered tools in creative processes. In *CHI conference on human factors in computing systems extended abstracts*. 1–9.
[41] Angel Hsing-Chi Hwang and Andrea Stevenson Won. 2022. AI in Your Mind: Counterbalancing Perceived Agency and Experience in Human-AI Interaction. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 349, 10 pages. https://doi.org/10.1145/3491101.3519833
[42] Youngseung Jeon, Seungwan Jin, Patrick C Shih, and Kyungsik Han. 2021. FashionQ: an ai-driven creativity support tool for facilitating ideation in fashion design. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–18.
[43] Xi-Hui Jia and Jui-Che Tu. 2024. Towards a New Conceptual Model of AI-Enhanced Learning for College Students: The Roles of Artificial Intelligence Capabilities, General Self-Efficacy, Learning Motivation, and Critical Thinking Awareness. *Systems* 12, 3 (2024), 74.
[44] Jialun Aaron Jiang, Kandrea Wade, Casey Fiesler, and Jed R. Brubaker. 2021. Supporting Serendipity: Opportunities and Challenges for Human-AI Collaboration in Qualitative Analysis. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 94 (apr 2021), 23 pages. https://doi.org/10.1145/3449168
[45] Xuhui Jiang, Yuxing Tian, Fengrui Hua, Chengjin Xu, Yuanzhuo Wang, and Jian Guo. 2024. A Survey on Large Language Model Hallucination via a Creativity Perspective. *ArXiv* abs/2402.06647 (2024). https://api.semanticscholar.org/CorpusID:267627389
[46] Martin Jonsson and Jakob Tholander. 2022. Cracking the code: Co-coding with AI in creative programming education. In *Proceedings of the 14th Conference on Creativity and Cognition*. 5–14.
[47] Pegah Karimi, Jeba Rezwana, Safat Siddiqui, Mary Lou Maher, and Nasrin Dehbozorgi. 2020. Creative sketching partner: an analysis of human-AI co-creativity. In *Proceedings of the 25th international conference on intelligent user interfaces*. 221–230.
[48] David Kember, Doris YP Leung, Alice Jones, Alice Yuen Loke, Jan McKay, Kit Sinclair, Harrison Tse, Celia Webb, Frances Kam Yuet Wong, Marian Wong, et al. 2000. Development of a questionnaire to measure the level of reflective thinking. *Assessment & evaluation in higher education* 25, 4 (2000), 381–395.
[49] Jingoog Kim and Mary Lou Maher. 2023. The effect of AI-based inspiration on human design ideation. *International Journal of Design Creativity and Innovation* 11, 2 (2023), 81–98.
[50] Lucy Kimbell. 2011. Rethinking design thinking: Part I. *Design and culture* 3, 3 (2011), 285–306.
[51] Tomas Lawton, Francisco J Ibarrola, Dan Ventura, and Kazjon Grace. 2023. Drawing with Reframer: Emergence and Control in Co-Creative & AI. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 264–277. https://doi.org/10.1145/3581641.3584095
[52] Yiren Liu, Si Chen, Haocong Cheng, Mengxia Yu, Xiao Ran, Andrew Mo, Yiliu Tang, and Yun Huang. 2024. How ai processing delays foster creativity: Exploring research question co-creation with an llm-based agent. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–25.
[53] Kyle Lo, Joseph Chee Chang, Andrew Head, Jonathan Bragg, Amy X Zhang, Cassidy Trier, Chloe Anastasiades, Tal August, Russell Authur, Danielle Bragg, et al. 2023. The semantic reader project: Augmenting scholarly documents through ai-powered interactive reading interfaces. *arXiv preprint arXiv:2303.14334* (2023).
[54] Ryan Louie, Andy Coenen, Cheng Zhi Huang, Michael Terry, and Carrie J Cai. 2020. Novice-AI music co-creation via AI-steering tools for deep generative models. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.
[55] Michael Luck, Mark d'Inverno, et al. 1995. A Formal Framework for Agency and Autonomy.. In *Icmas*, Vol. 95. 254–260.
[56] Kleopatra Mageira, Dimitra Pittou, Andreas Papasalouros, Konstantinos Kotis, Paraskevi Zangogianni, and Athanasios Daradoumis. 2022. Educational AI chatbots for content and language integrated learning. *Applied Sciences* 12, 7 (2022), 3239.
[57] Bahar Mahmud, Guan Hong, and Bernard Fong. 2023. A Study of Human–AI Symbiosis for Creative Work: Recent Developments and Future Directions in Deep Learning. *ACM Transactions on Multimedia Computing, Communications and Applications* 20, 2 (2023), 1–21.
[58] Lina Markauskaite, Rebecca Marrone, Oleksandra Poquet, Simon Knight, Roberto Martinez-Maldonado, Sarah Howard, Jo Tondeur, Maarten De Laat, Simon Buckingham Shum, Dragan Gašević, et al. 2022. Rethinking the entwinement between artificial intelligence and human learning: What capabilities do learners need for a world with AI? *Computers and Education: Artificial Intelligence* 3 (2022), 100056.
[59] Rosario Michel-Villarreal, Eliseo Vilalta-Perdomo, David Ernesto Salinas-Navarro, Ricardo Thierry-Aguilera, and Flor Silvestre Gerardou. 2023. Challenges and opportunities of generative AI for higher education as explained by ChatGPT.

Education Sciences 13, 9 (2023), 856.

[60] Tim Miller. 2023. Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven Decision Support using Evaluative AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 333–342. https://doi.org/10.1145/3593013.3594001

[61] Caterina Moruzzi and Solange Margarido. 2024. Customizing the balance between user and system agency in human-AI co-creative processes. In *Proceedings of the 15th International Conference on Computational Creativity (ICCC'24). Association for Computational Creativity*. 1–10.

[62] Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. 2023. Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review* 56, 4 (2023), 3005–3054.

[63] Mir Murtaza, Yamna Ahmed, Jawwad Ahmed Shamsi, Fahad Sherwani, and Mariam Usman. 2022. AI-based personalized e-learning systems: Issues, challenges, and solutions. *IEEE access* 10 (2022), 81323–81342.

[64] Katherine O'Toole and Emőke-Ágnes Horvát. 2024. Extending human creativity with AI. *Journal of Creativity* 34, 2 (2024), 100080.

[65] Nathan Partlan, Erica Kleinman, Jim Howe, Sabbir Ahmad, Stacy Marsella, and Magy Seif El-Nasr. 2021. Design-driven requirements for computationally co-creative game AI design tools. In *Proceedings of the 16th International Conference on the Foundations of Digital Games*. 1–12.

[66] James Prather, Brent N Reeves, Juho Leinonen, Stephen MacNeil, Arisoa S Randrianasolo, Brett A Becker, Bailey Kimmel, Jared Wright, and Ben Briggs. 2024. The Widening Gap: The Benefits and Harms of Generative AI for Novice Programmers. In *Proceedings of the 2024 ACM Conference on International Computing Education Research-Volume 1*. 469–486.

[67] Nina Rajcic, Maria Teresa Llano Rodriguez, and Jon McCormack. 2024. Towards a Diffractive Analysis of Prompt-Based Generative AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–15.

[68] Jeba Rezwana and Mary Lou Maher. 2023. Designing creative AI partners with COFI: A framework for modeling interaction in human-AI co-creative systems. *ACM Transactions on Computer-Human Interaction* 30, 5 (2023), 1–28.

[69] Shadan Sadeghian, Alarith Uhde, and Marc Hassenzahl. 2024. The Soul of Work: Evaluation of Job Meaningfulness and Accountability in Human-AI Collaboration. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 130 (apr 2024), 26 pages. https://doi.org/10.1145/3637407

[70] Advait Sarkar. 2024. AI Should Challenge, Not Obey. *Commun. ACM* 67, 10 (2024), 18–21.

[71] Abigail Sellen and Eric Horvitz. 2024. The rise of the AI Co-Pilot: Lessons for design from aviation and beyond. *Commun. ACM* 67, 7 (2024), 18–23.

[72] Orit Shaer, Angelora Cooper, Osnat Mokryn, Andrew L Kun, and Hagit Ben Shoshan. 2024. AI-Augmented Brainwriting: Investigating the use of LLMs in group ideation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17.

[73] Abdulhadi Shoufan. 2023. Exploring students' perceptions of ChatGPT: Thematic analysis and follow-up survey. *IEEE Access* 11 (2023), 38805–38818.

[74] Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2024. Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers. *arXiv preprint arXiv:2409.04109* (2024).

[75] Jonathan Michael Spector and Shanshan Ma. 2019. Inquiry and critical thinking skills for the next generation: from artificial intelligence back to human intelligence. *Smart Learning Environments* 6, 1 (2019), 1–11.

[76] Arne Styve, Outi T Virkki, and Usman Naeem. 2024. Developing Critical Thinking Practices Interwoven with Generative AI Usage in an Introductory Programming Course. In *2024 IEEE Global Engineering Education Conference (EDUCON)*. IEEE, 01–08.

[77] Minhyang Suh, Emily Youngblom, Michael Terry, and Carrie J Cai. 2021. AI as social glue: uncovering the roles of deep generative AI during social music composition. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–11.

[78] Lev Tankelevitch, Viktor Kewenig, Auste Simkute, Ava Elizabeth Scott, Advait Sarkar, Abigail Sellen, and Sean Rintel. 2024. The metacognitive demands and opportunities of generative AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–24.

[79] Ahmed Tlili, Boulus Shehata, Michael Agyemang Adarkwah, Aras Bozkurt, Daniel T Hickey, Ronghuai Huang, and Brighter Agyemang. 2023. What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart learning environments* 10, 1 (2023), 15.

[80] Suzanne Tolmeijer, Markus Christen, Serhiy Kandul, Markus Kneer, and Abraham Bernstein. 2022. Capable but Amoral? Comparing AI and Human Expert Collaboration in Ethical Decision Making. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 160, 17 pages. https://doi.org/10.1145/3491102.3517732

[81] Josh Urban Davis, Fraser Anderson, Merten Stroetzel, Tovi Grossman, and George Fitzmaurice. 2021. Designing co-creative ai for virtual environments. In *Proceedings of the 13th Conference on Creativity and Cognition*. 1–11.

[82] Florent Vinchon, Todd Lubart, Sabrina Bartolotta, Valentin Gironnay, Marion Botella, Samira Bourgeois-Bougrine, Jean-Marie Burkhardt, Nathalie Bonnardel, Giovanni Emanuele Corazza, Vlad Glăveanu, et al. 2023. Artificial Intelligence & Creativity: A manifesto for collaboration. *The Journal of Creative Behavior* 57, 4 (2023), 472–484.

[83] Yoshija Walter. 2024. Embracing the future of Artificial Intelligence in the classroom: the relevance of AI literacy, prompt engineering, and critical thinking in modern education. *International Journal of Educational Technology in Higher Education* 21, 1 (2024), 15.

[84] Qian Wan, Siying Hu, Yu Zhang, Piaohong Wang, Bo Wen, and Zhicong Lu. 2024. "It Felt Like Having a Second Mind": Investigating Human-AI Co-creativity in Prewriting with Large Language Models. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–26.

[85] Bryan Wang, Yuliang Li, Zhaoyang Lv, Haijun Xia, Yan Xu, and Raj Sodhi. 2024. LAVE: LLM-Powered Agent Assistance and Language Augmentation for Video Editing. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*. 699–714.

[86] Xinghua Wang, Qian Liu, Hui Pang, Seng Chee Tan, Jun Lei, Matthew P Wallace, and Linlin Li. 2023. What matters in AI-supported learning: A study of human-AI interactions in language learning using cluster analysis and epistemic network analysis. *Computers & Education* 194 (2023), 104703.

[87] Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. 2022. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems* 135 (2022), 364–381.

[88] Yi Wu. 2024. Critical Thinking Pedagogics Design in an Era of ChatGPT and Other AI Tools—Shifting From Teaching "What" to Teaching "Why" and "How". *Journal of Education and Development* 8, 1 (2024), 1.

[89] Zhuohao Wu, Danwen Ji, Kaiwen Yu, Xianxu Zeng, Dingming Wu, and Mohammad Shidujaman. 2021. AI creativity and the human-AI co-creation model. In *Human-Computer Interaction. Theory, Methods and Tools: Thematic Area, HCI 2021, Held as Part of the 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings, Part I 23*. Springer, 171–190.

[90] Xiaotong (Tone) Xu, Jiayu Yin, Catherine Gu, Jenny Mar, Sydney Zhang, Jane L. E, and Steven P. Dow. 2024. Jamplate: Exploring LLM-Enhanced Templates for Idea Reflection. In *Proceedings of the 29th International Conference on Intelligent User Interfaces* (Greenville, SC, USA) *(IUI '24)*. Association for Computing Machinery, New York, NY, USA, 907–921. https://doi.org/10.1145/3640543.3645196

[91] Ramazan Yilmaz and Fatma Gizem Karaoglan Yilmaz. 2023. The effect of generative artificial intelligence (AI)-based tool use on students' computational thinking skills, programming self-efficacy and motivation. *Computers and Education: Artificial Intelligence* 4 (2023), 100147.

[92] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: story writing with large language models. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*. 841–852.

[93] Qingxiao Zheng, Zhongwei Xu, Abhinav Choudhary, Yuting Chen, Yongming Li, and Yun Huang. 2023. Synergizing human-AI agency: a guide of 23 heuristics for service co-creation with LLM-based agents. *arXiv preprint arXiv:2310.15065* (2023).