Productive vs Reflective: How different ways of integrating AI into design workflows affects cognition and motivation

ANONYMOUS AUTHOR(S)

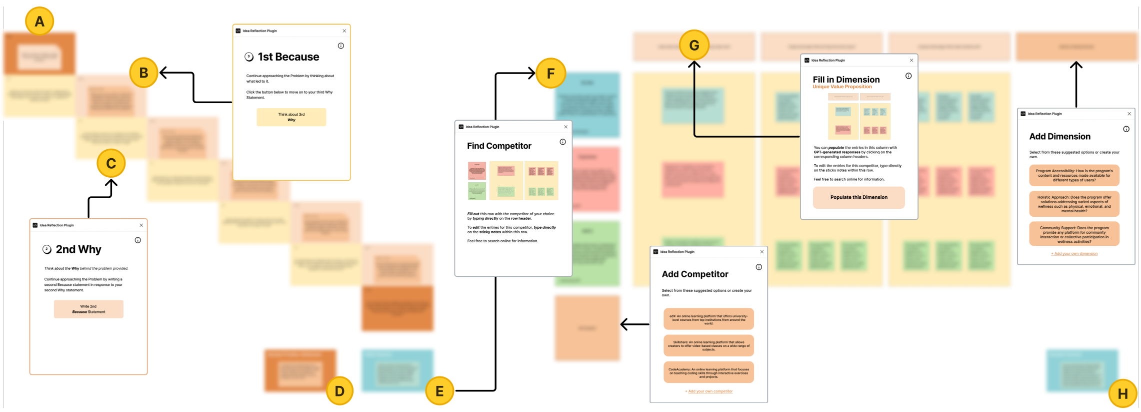

Fig. 1. Overview of the design workflow in co-led condition. A. Participants were given a problem statement at the start of the study and saw a AI-generated reflective question on it in the box below. B. They need to answer in the 'Because' box to address reflective question of 'Why' on the left. C. The 'Why' box prompts participant to reflect on cause of the prior 'Because' answer. D. Participant generated an Iterated Problem Statement (IPS) after providing the reflective answers. E. Then they propose Initial Solution address IPS. F. Series of AI-proposed competing solutions to the Initial Solution. Participant can request more. G. Series of dimensions across which the different solutions will be analyzed. Participant can request more. H. Participant proposed an Iterated Solution after analyzing the competitive ideas.

An increasing number of creativity-support tools now integrate AI-empowered features, extending the ability of users—especially novices—to produce creative work. While AI could assume various roles within such tools, less is known about how the positioning of AI support affects an individual's cognitive processes and sense of agency. To examine this relationship, we built a collaborative whiteboard plugin that integrates a large language model into design templates that facilitate reflective brainstorming activities. We conducted a between-subjects experiment with three different techniques (human-led vs. co-led vs. AI-led) with n=47 design novices to compare how the balance between human and machine effort impacts the allocation of cognitive resources. Our results showed that participants who viewed AI responses had produced more written content. AI-led participants had more divergent thinking but co-led participants had more in-depth details. Those who engaged with providing reflective responses also reported higher self-efficacy on the creativity of solutions.

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

Additional Key Words and Phrases: Creativity, Critical Thinking, Self-reflection, Learning, Brainstorming, Human-AI Collaboration, Agency, Co-piloting, Steerable AI, LLMs

ACM Reference Format:
Anonymous Author(s). 2024. Productive vs Reflective: How different ways of integrating AI into design workflows affects cognition and motivation. In *Proceedings of The ACM (Association of Computing Machinery) CHI conference on Human Factors in Computing Systems (CHI '25).* ACM, New York, NY, USA, 22 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

The recent generative AI evolution has led to many innovative features for creativity-support tools across multiple domains [37, 67, 76]. For example, image editors can now replace parts of images with simple queries [1] and document readers can now extract paper highlights [46]. AI models have been trained to support specific creative tasks (e.g., music composition [34, 47], image generation [63], etc.). AI's specific capabilities and human-AI co-creativity varies in the types of interaction, communication, and collaboration [61], which may have different effects on a user's cognitive processes and sense of agency. Heer [2019] and Sadeghian et al. [2024] identified that AI systems designed to augment rather than replace human expertise enhance decision-making and creativity by maintaining user control while boosting productivity [31, 62]. Prior work in human-AI co-creativity leverages this technology to expand users' knowledge, to deepen their understanding of problems, and to cultivate associative thinking.

For creative workflows, in particular, people perform a range of different ways of "thinking", from reflecting, to gathering information, to framing and reframing the problem, and then to ideating, to name but a few. Recent studies have shown a mixed influence of AI-generated text on ideation. In the ideation context, researchers have used LLM-generated inspirational text to support users to reframe or explore the design space, and thus generate more creative ideas [32]. However, LLMs may also led participants to produce more similar ideas and made them feel less responsible for the generated ideas [2]. Adopting AI-generated information requires that people can think critically about it. Nonetheless, people often attribute ownership to AI-generated content but are reluctant to publicly acknowledge AI's involvement, reflecting a broader distinction in understanding agency [20]. Given the many AI roles in human-AI collaboration, such as "creative assistant" [27, 85], "creator of content" [14], and a "friend" [81], it is pivotal to understand how the positioning of AI affects critical thinking and reflection.

Recent research into human-AI collaboration has often framed this as implementing "human-in-the-loop" technique for machine learning [55, 79], or designing a "co-pilot" [64] or "steerable AI" [25], 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 and only offloading low-level tasks to AI. For example, consider the differences between the positioning of AI within systems intended to support reading and comprehension [46]. The Paper Plain system [4] can summarize passages and then read the automated summary. The Scim system [22] can intelligently highlight papers for faster skimming. In yet another approach, the Papeo system [46] can locate paper content from talk videos. It is unclear how the positioning of an AI and the agency that occurs during human-AI collaboration influences people's learning and critical thinking. In this study, we aim to understand how different ways of integrating AIs into creative tasks shape people's reflection and idea iteration. Specifically, we asked the following research question: How does the positioning of AI within a creative workflow affect creative outcomes, the allocation of a user's cognitive resources, and user perceptions of the value of AI?

To investigate this, we developed a FigJam (an online collaborative whiteboard) plugin that integrates LLMs into a design workflow for common creative exercises in three ways (Human-only vs. Co-led vs. AI-led). Two templates (Five Whys and Competitive Analysis) were selected because they often require a designer to gather external information, tasks that LLMs generally perform well. The LLM is integrated through prompting based on the data structure of the template [82]. The templates also invoke a range of cognitive processes, allowing us to explore how the AI positioning impacts cognition. Participants engaged in a reflective activity, thinking deeply about the root causes of a problem, and a sensemaking activity through conducting competitive analysis.

We ran a between-subjects experiment where N=47 participants were randomly assigned to one of three conditions. In the AI-led condition, the system fills out the entire template with LLM-generated text leaving the participants to review and edit this information. The co-led condition similarly integrates LLM-generated text, but on a more piece-meal basis, allowing the user to read and provide input box by box withing the template. As a baseline, in the Human-only condition, participants completed the activity with the plugin instructing them on what to do at each step on the template but no LLM-generated information was available at all. In all conditions, participants iterated on a problem statement after the reflective activity, brainstormed an idea, and finally iterated on the idea after the competitive analysis.

The research team analyzed the information from workflows to uncover inspirational sources for user-generated problem statements and solutions. We counted and categorized any new concept or added idea that was not simply a rewording or rearrangement of the original Problem Statement. For short, we will use the term “idea unit” to capture this coding. We used participants’ iterated problem statements (IPS), initial solutions, and iterated solutions to measure creative outcome, coded the study videos to investigate cognitive processes, and survey responses to understand self-efficacy.

The study finds that AI-led participants focused on ensuring information comprehensiveness and exhibited expansive thinking influenced by AI-generated content. Co-led participants, on the other hand, engaged more in reflective thinking, developed deeper problem statements, and self-rated their initial solutions as more creative compared to AI-led participants. They also intentionally focused less on the novelty of their ideas after reviewing and analyzing AI-gathered solutions.

In summary, this paper makes the following contributions:

- We developed an interactive plugin that integrates LLMs and the existing cognitive walkthroughs during the design process to facilitate the sample complex cognitive tasks (e.g., self-reflection and iteration).
- We conducted a three-condition between-subject in-person experiment (N=47) to investigate how collaboration with LLMs influences critical reflection, sensemaking and idea iteration process.
- The study insights shed light on how co-piloting with LLMs affects critical thinking and self-efficacy, and can help future researchers implement better practices that enable learners to maintain skill gains while keeping their work processes efficient and productive.

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 [43]. O’Toole and Horvát [2024] identify five key AI applications in creativity: creative support, ideation support, personalization, co-creativity, and novelty [57]. Guo et al. [2024] and Zheng et al. [2023] discuss AI’s role as a “creative assistant,” allowing users to offload repetitive tasks and focus on higher-level thinking [26, 85]. Tools like the Creative Sketching Partner (CSP) [41] and FashionQ [37] 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 [37, 65, 76]. ID.8, an AI tool for co-creating visual narratives [3], identifies lowered barriers and increased access [81] 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" [14], shaping human-AI relationships in idea generation. The extent of AI's involvement depends on the human collaborator. Wang et al. [2024] found that the need for AI support varies by task [77]. Wordcraft's AI agent [84] exemplifies this by serving as an idea generator, interpolator, and copy editor throughout the design process. However, Partlan et al. [2021] argue that dependable AI results alone aren't enough for co-creativity [58]; users need control over outputs to explore diverse ideas [2], manipulation of results, and providing of feedback [73], to keep them engaged in the process.

User control is crucial for affirming their roles as creative collaborators. Rajcic et al. [2024] found that users feel more confident in their creative identity when they control AI tools [60]. Jonsson and Tholander [2022] highlight that the "friction" between inconsistent AI outputs and human input encourages reflection and rethinking [40]. This interaction leads human collaborators to take on a curatorial role [16], using AI feedback to challenge initial ideas and decide what to keep, remove, or expand upon [23]. As a facilitator in discussions [24] or as a "friend" [81], AI alters human creative roles as they work to integrate it as a collaborator in their creative processes [70].

Current research highlights AI's role in idea generation but often focuses on how users receive information rather than the collaboration with AI to create new ideas [35]. There is a need for "interaction dynamics" [61], like turn-taking, to foster communication between human and AI creators rather than competition [74]. Designing effective co-creative systems requires trustworthiness, adaptiveness [50], and alignment with user values [23] 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 Educational Thinking Toolkit

Although formal curricula for AI have been around for decades [13], in recent years the presence of AI in academic settings [56] – regardless of field of study – is rapidly approaching ubiquity. With the growing accessibility of powerful LLMs to the average student and educator, 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 [71] but limited quality control over the information [52], as well as plagiarism and academic integrity issues [21]. Generative AI chatbots have proved to be useful tools in learning new content [6, 49] and have received positive feedback on their impact [15, 19, 33], 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 [75] or learn more in-depth in other modalities [80], but it could also impede critical thinking by generating completed results without necessitating human involvement [29]. The proven positive impacts of AI in academic settings include better computational thinking skills [83], stronger learner self-efficacy [38], more opportunities for reflection [45], and higher confidence in analysis and comprehension of complex concepts [28]. Furthermore, AI has been proven to help users think more comprehensively and thoroughly about problems posed in design thinking exercises [82]. However, there have also been scenarios in which utilizing AI can exacerbate existing weak points and challenges in critical thinking [59].

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. [2024] 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" [69], and Shoufan [2023] finds that "most [students] believe that [ChatGPT] requires good background knowledge to work with since it does not replace human intelligence" [66]. Research suggests that working in tandem with AI yields the best results [78], using it intentionally as a tool while also exercising one's uniquely human intelligence [51, 68]. Brondani et al. [2024] 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 [11]. 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 [17, 54]. These frameworks emphasize the preservation of human agency to ensure meaningful interaction that shapes engagement, learning, and reflective thinking.

Agency in AI systems can range from goal-directed assistance to autonomous actions, either assisting set tasks or dynamically responding to changing contexts [48]. Hwang and Won [2022] 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 [36]. Miller [2023] and Buçinca et al. [2021] both advocate reducing AI over-reliance, with Miller proposing "Evaluative AI" to enhance user control and human agency [53], and Buçinca et al. emphasizing cognitive forcing interventions for more deliberate human thinking [12]. 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. [2024] 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 [32]. Similarly, Lawton et al. [2023] 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 [44]. Jiang et al. [2021] stressed the need for human control in inductive analysis to prevent oversimplification of complex insights [39]. Biermann et al. [2022] found that writers who maintained control over creative tasks like character and dialogue generation had a stronger sense of ownership and resisted AI interventions [7]. 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.

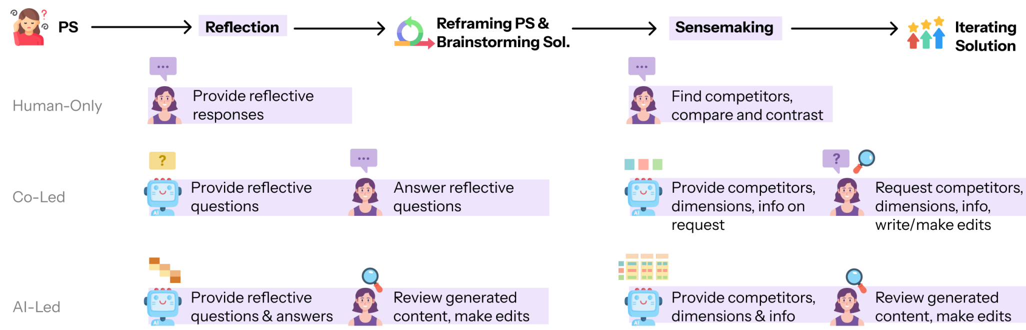

Fig. 2. Study workflow broken down by human and AI roles across conditions. In Human-Only condition, participants completed design exercises without AI assistance. In Co-led condition, participants responded to AI-generated reflective questions and received AI-gathered competitive analysis on demand. In AI-led condition, participants responded to AI-generated reflective conversations (both questions and answers) and AI-gathered competitive analysis all at once.

3 SYSTEM DESIGN

3.1 Design Workflow Integration

To facilitate AI integration in design thinking, we created an interactive plugin on a 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. But 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: reflective thinking for problem scoping and sensemaking for ideation [18, 82].

This reflective process is facilitated with the Five Whys (5Y) template provided by the digital whiteboard, 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 interrogative technique is designed to explore the cause-and-effect relationships underlying a particular problem.

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

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 with 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 potential, 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 written at the end of the 5Ys exercise to a small number of competitors along a fixed set of 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 5Ys, 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 Human-only. The Human-only condition serves as the control group, representing how users typically approach a problem/solution without LLM assistance. Participants use the default empty templates, filling them out manually based 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.

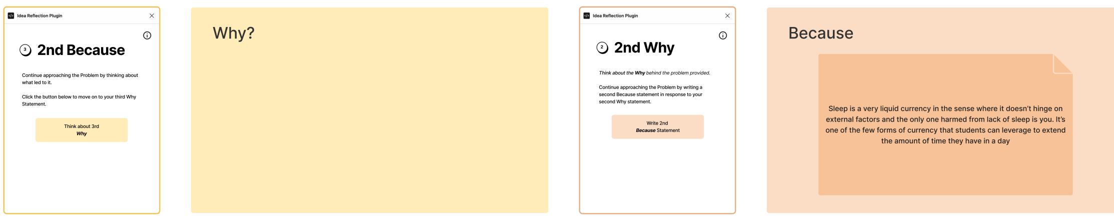

Fig. 3. For the Human-only condition, every blank 'Why' box prompts users to reflect and write answers in the respective 'Because' box, after which they will continue to the next 'Why'.

Reflection and Problem Scoping. The plugin in the Human-only 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.

In the Human-only condition, participants first see a blank 5 Whys template in the first task. To push forward, the default template question simply asks 'Why?' in boxes, without any context-specific information. The participants then provide an answer in the adjacent 'Because' boxes that largely depends on the user's existing knowledge and interpretation.

When participants complete a section, buttons on the interface prompt them to move onto the next, writing the next Because statement or thinking about the next Why. Regardless of previous experience with the exercise or platform, participants are able to complete the entire template with little friction in workflow.

At the end of the exercise, participants combine their findings into a manually written Root Cause. From there, participants write an Iterated Problem Statement in order to reframe the problem based on what they learned. Lastly, participants draft an Initial Solution to the Iterated Problem Statement.

Brainstorming, Sensemaking and Iteration. For the Human-only 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

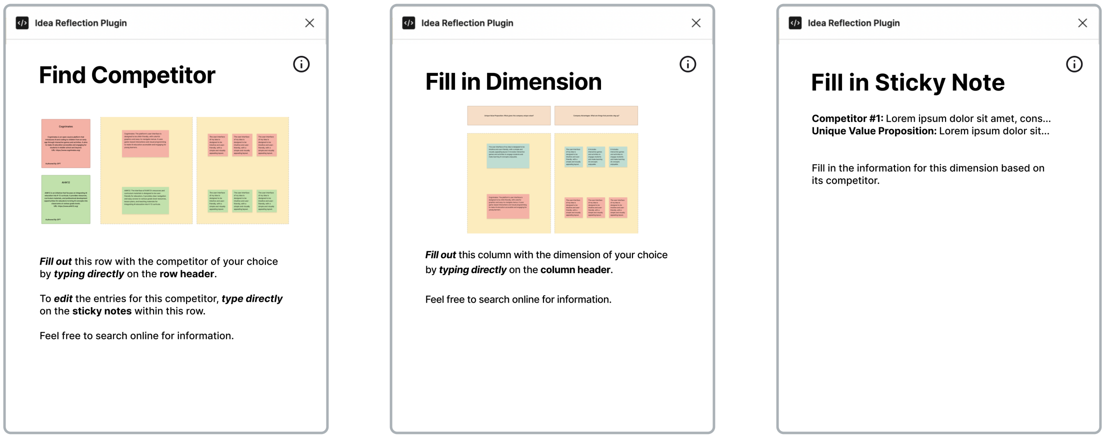

Fig. 4. For the Human-only condition, clicking on competitor and dimension headings prompts users to fill out the rows and columns by typing, and clicking on an individual sticky note prompts the user to fill that cell out by typing.

of the problem/solution space. Unlike the Human-only condition, participants are not required to 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.

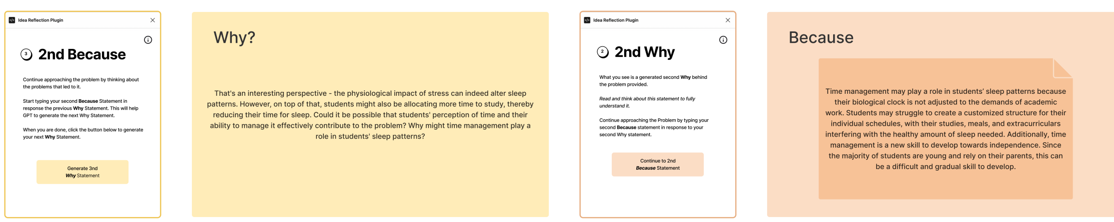

Fig. 5. For the Co-led condition, the LLM generates reflective questions in the 'Why' box that the participant must manually answer in the corresponding 'Because' box. Then, they can go on to generate the next 'Why' question.

Reflection and Problem Scoping. In the Co-led condition, to enhance the typical workflow of the 5Ys 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. And after a few seconds, the next 'Why' question will appear, which proposes a guided 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.

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

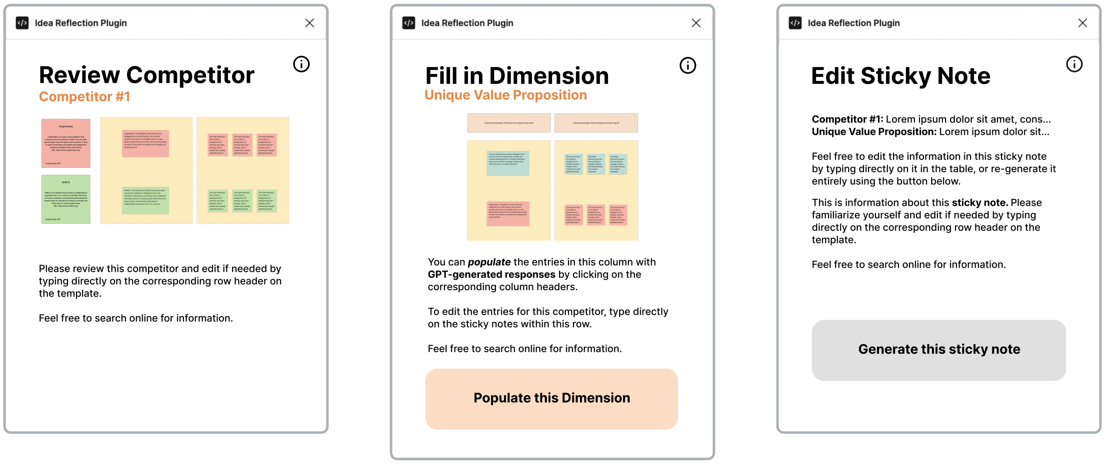

Fig. 6. For the Co-led condition, once the competitors and dimensions are generated, participants can review competitors by clicking on the row headers, and populate the dimensions with LLM-generated information by clicking on the column headers. Participants can also generate information for individual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

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 Human-only 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.

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.

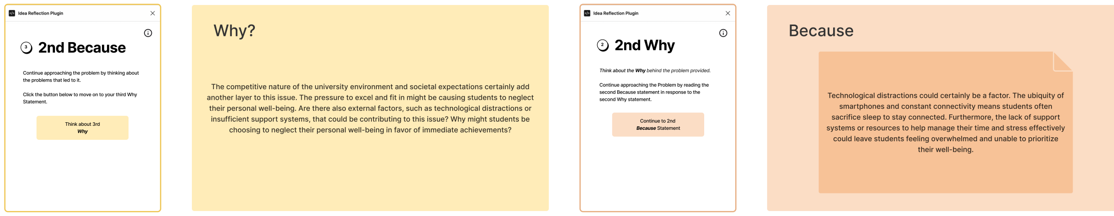

Fig. 7. For the AI-led condition, both the 'Why' and 'Because' boxes are populated with LLM-generated text following the line of reasoning set by the problem statement.

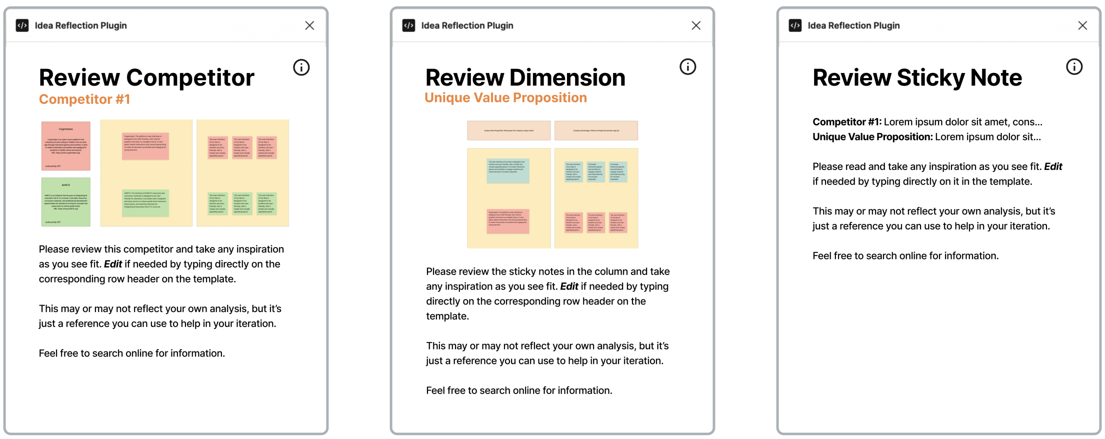

Fig. 8. For the AI-led condition, participants can review the LLM-generated information for competitors, dimensions, and individual cells by clicking on the row and column headers and sticky notes respectively.

Brainstorming, Sensemaking and Iteration. When the user is ready to move onto the CA exercise, the experimenter selects their Initial Solution. This time, the plug-in 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 5Ys 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. Afterward, the user writes an Iterated Solution based on the insights they gathered.

3.3 Tech Infrastructure

The system was built using TypeScript the FigJam Plugin API[1] for front-end interface design, Firebase[2] for storing data, 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. Then the server prompts the OpenAI API and receives and processes responses before returning them to the

[1] https://www.figma.com/plugin-docs/api/api-reference/
[2]
[3]
[4]

system front-end. The prompt questions were inspired by a previous study integrating LLMs into the design templates [82].

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. 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 Study Sessions

The user study sessions were conducted in-person, with one researcher working one-on-one with each participant. Additionally, Zoom was used to share screens and record participants' activities on the computer screen, as well as to capture their responses during post-study interviews. Prior to their arrival at the study site, 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 (Human-only:Co-Led:AI-Led) is 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 participants to the problem statement and the generated Five Whys template, which should take no more than 30 minutes to complete. After writing the Root Cause, participants 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 participants completed writing their initial solution, they transitioned to the sensemaking portion of the study, which took another 30 minutes. Participants were shown their idea in the topmost row in 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 they could use the internet for real-time online research for checking information online. 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 inspirational sources 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.

4.3.1 Idea Quality. 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 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. ANCOVA were 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 Video Coding. 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. To analyze the video recordings, the researchers employed a detailed video coding legend that categorized specific actions to measure and understand participant engagement and activity accurately.

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, all of the researchers calculated inter-rater reliability.

4.3.3 Survey Responses. The post-study survey questions asked about self-rating on five-point Likert scales ranging from "Strongly Disagree" to "Strongly Agree" to rate workload [30] and critical thinking activities [42] involved in the process. 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 survey 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] 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. A detailed analysis was conducted to identify similar themes and patterns in participants' findings, experiences, opinions, and key takeaways.

5 RESULTS

5.1 How does the positioning of an AI within design templates affect creative outcomes?

All participants had a similar quantity of idea units in the iterated problem statement (F(2,43)=2.34, p=0.11), but the convergence of topics varied.

5.1.1 Co-led participants wrote deeper problem statements. Significantly fewer topic categories of idea units were generated in the Co-led participants than those were in the Human-only and AI-led condition (F(2,43)=4.23, p=0.02*), indicating a more concentrated and detailed delve into a perspective in the problem statement. P159 [Co-led] mentioned that the interaction with the LLM was like "talking to somebody. We expanded on those thoughts like 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." P158 [Co-led] felt the 'Why' questions helped them break down the issues more deeply: "It 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."

5.1.2 AI-led participants developed broad solutions initially and diverged in iteration. Much like the number of ideas in the solutions, the diversity of ideas in the solutions emerged comparable patterns across conditions. Participants across all three conditions generated a similar number of topics in the initial brainstorming solution (F(2,43)=1.78, p=0.18). Those who viewed LLM-generated competitive analysis (AI-led) spanned more topics in their idea iteration compared to those who conducted a Human- and Co-led competitive analysis, indicating more divergent thinking (F(2,43)=5.33, p<0.009**). 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.

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."

5.2 How does the positioning of an AI within design templates affect the cognitive process?

Timestamps were coded when activities start and end based on the video recordings of study sessions. Time duration when participants explicitly typed on canvas components and the thinking time before they started typing were calculated. ANCOVA were conducted to examine the conditional differences on the time spent on each activity, while controlling for the version of the initial problem statement. The analyses excluded distracted activities, such as experimenter instructions, interruptions, and system loading.

5.2.1 Human-only participants edited for accuracy, while AI-led participants ensured information remained intact. As per design, Human-only and Co-led participants overall spent significantly more time on filling out the template than AI-led participants (F(2,43)=54.05, p<0.001**), who spent significantly more time on reviewing the content (F(2,43)=3.36, p=0.044*). Human-only participants were the most likely to go back and forth editing their previous responses to

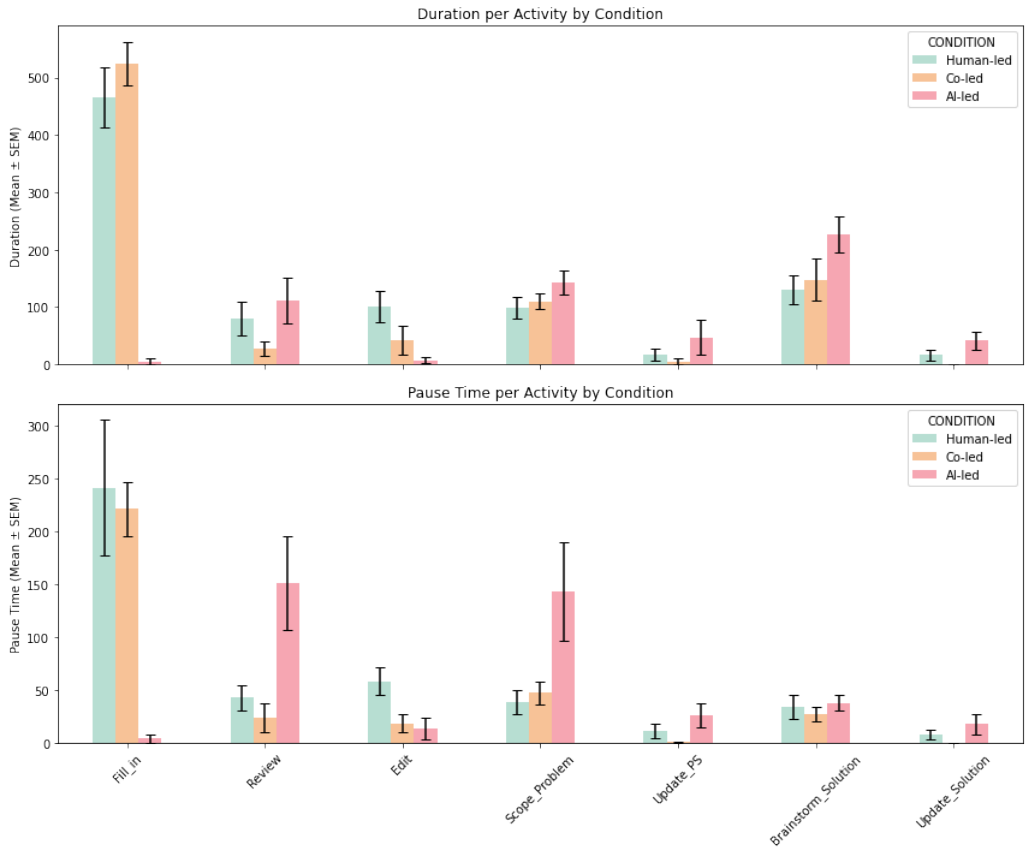

Fig. 9. Duration per Activity by Condition (top) and Pause Time per Activity by Condition (bottom) for the Five Whys exercise.

correctly express themselves compared to the other conditions ($F(2,43)=5.17$, $p=0.0098^{**}$). They took longer time to settle in the reflections than Co-led participants. P126 [Human-only] said that "the first 3 Whys was getting through what my initial nature reactions were. So you will see these last 2 contain more deeper thoughts, as something doesn't come up to you immediately." Human-only participants might also have been overwhelmed by the information at the end and selectively focused on a subset of issues came across. P102 [Human-only] 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."

Compared with Human-only and AI-led participants, Co-led participants were less likely to change their existing answers in the template. AI-led participants also checked back and forth but did not edit the template. They checked the previous responses not due to information integrity but more so 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 I feel the root cause covered everything."

5.2.2 Human-only and Co-led reflections prepared participants' thoughts about problems whereas AI-led responses were more confirmatory. Human-only and Co-led participants were also more fluent in updating the problem statement after the Five Whys exercise, spending less time on preparing and planning before typing ($F(2,43)=4.14$, $p=0.024^*$). They spent about the same amount of time typing out the problem statement ($F(2,43)=1.84$, $p=0.17$).

Human-only participants indicated that the interactive template eased the otherwise stressful cognitive process and helped them approach the problem in different way. P109 [Human-only] said: "I don't think I've ever thought about this issue this deeply before. I was just probably like, oh. I'm just stressed, you know? 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".

Co-led participants also indicated that the reflective questions challenged them to think differently than otherwise they would even know possible. 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 about more how you can improve upon your initial statement."

AI-led participants viewed AI-generated responses as polished and expected. 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 flushed out, clean, very well-written answer." P128 [AI-led] also described the responses as logical: "And then [the questions] are expected. It's very logical. There's no really big leaps of assumptions." P136 [AI-led] claimed that the AI-generated responses were continuous and in-depth: "I liked how it kept continuing once you get more in-depth into it." But they sometimes may also be repetitive. P138 [AI-led] felt "some of (the 5Ys) were a little bit similar to each other…The (Whys) could have gone farther into more of an institutional issue than it did."

5.2.3 AI-led participants spent more time writing an initial solution and Co-led participants edited template contents less. When brainstorming their first solutions, participants across three conditions spent similar amounts of time thinking ($F(2,43)=0.39$, $p=0.68$), but AI-led participants marginally spent more time on writing ($F(2,43)=2.62$, $p=0.84$). P111 [AI-led] expressed the difficulty in creating a single solution in response to the many identified reasons for the problem: "I think 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 like trying to come up with a clear and good sized answer." P136 [AI-led] described the level of detail and clarity already provided by the AI-generated responses: "...[It] just basically [gives] me ideas and answers to my questions about these. And [it] gives me an idea of what I know and this is a problem, that maybe we should fix that." P108 [AI-led] added to this, struggling with writing because they are building on well-formed, completed responses that are not their own: "[My only struggle] is reiterating [the] problem in my own way because it was already really clear."

Almost no Co-led participants went back and forth editing the contents in the templates, compared to the other two groups. P115 [Co-led] mentioned the repetition of questioning and responding in allowing them to identify a final solution: "I think there's a constant asking why it addresses a problem that I did not think was a problem…eventually it asked me a question where I came to the solution that, oh, students just don't prioritize sleep." P113 [Co-led] found this repetition can also work to prevent further consideration and iteration of ideas: "...I think the generation of the Why questions were a little confusing in the sense that they were very repetitive. So it was actually clogging my thought process instead of helping it, even though…I had to look over the same points again and again, which was nice in some sense. But in another sense it makes it very difficult to think because it doesn't spread out my thought process."

Alternatively, P123 [Co-led] 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. . . sort of like. . . talking to myself. . . And it gives a lot of thought because talking with yourself is pretty easy to do. . . you can just (like) bounce back ideas back and forth."

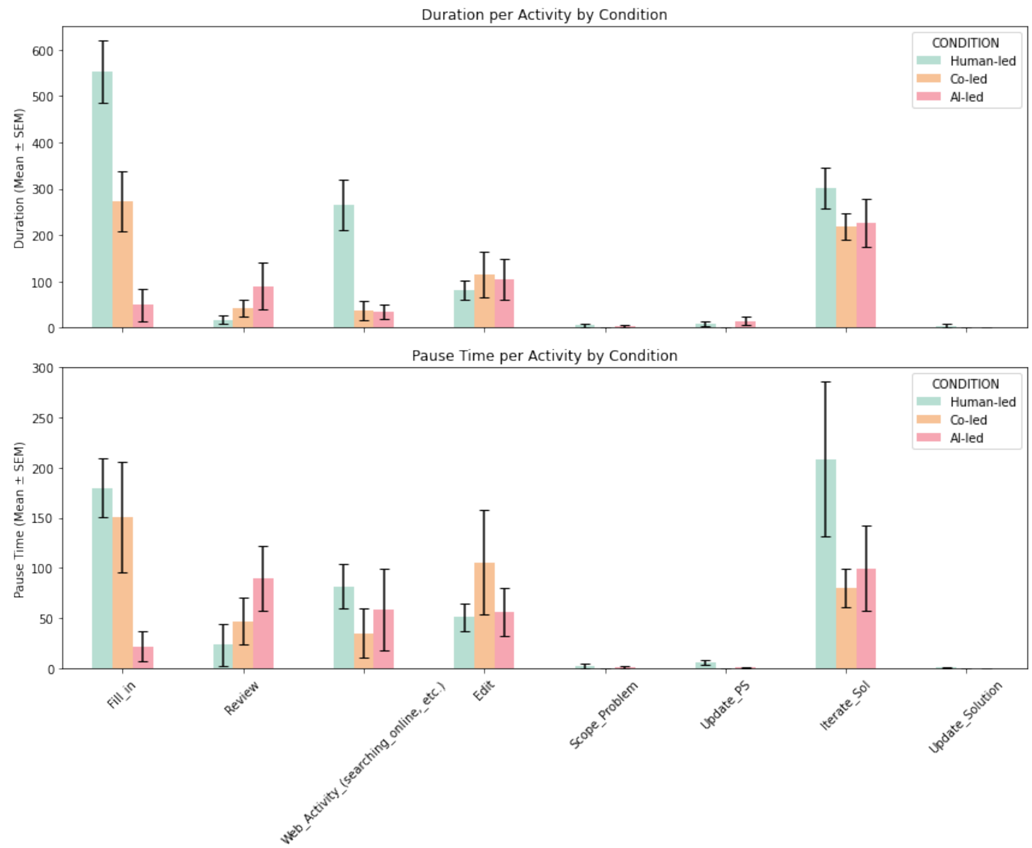

Fig. 10. Duration per Activity by Condition (top) and Pause Time per Activity by Condition (bottom) for the Competitive Analysis exercise.

5.3 How does the positioning of an AI within design templates affect self efficacy?

5.3.1 Human-only and Co-led participants reported more engagement and higher self-efficacy in solution creativity. Participants who engaged with providing reflective answers (Human-only and Co-led) also reported that they were engaged with more thinking and decision making when completing this activity ($F(2,41)=4.90$, $p=0.012^*$) than those who just reviewed AI-generated reflective answers (AI-led). AI-led participants tended to summarize the AI responses but lacked critical thinking on those responses. P128 [AI-led] mentioned said : "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 Human-only participants also rated their initial brainstorming solutions more creative than AI-led participants ($F(2,41)=4.93$, $p=0.012^*$). P102 [Human-only] mentioned that because of the need to think deeper beyond merely reading the problem statement, they were able to craft a better solution: ". . . If I just saw the problem statement

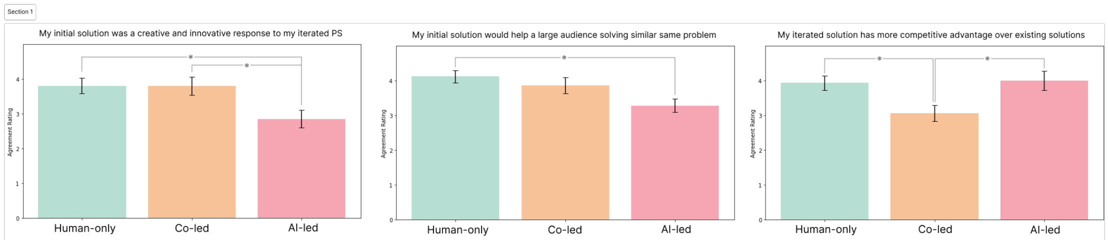

Fig. 11. From left to right: (1) Human-only and Co-led participants showed higher self-efficacy in initial solution creativity. (2) Human-only participants rated their solution targeting more audience than AI-led participants. (3) Co-led participants perceived their ideas less novelty-focused compared to the existing solutions that AI-gathered.

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”. In contrast, P105 [AI-led] found that it was difficult to strike balance among the given ideas they had to choose from: “I guess [I] incorporated [these ideas] because I lingered longer on these 2. 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.”

Human-only participants also believed that their initial solution better targeted a more diverse audience in comparison with AI-led participants (MD=0.84, p=0.016*). As P146 [Human-only] 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 comparing AI-provided competitors showed deeper concern of problems and focused more on practicality of solutions. Similar to the reflective exercise, Human-only and Co-led participants reported more on thinking, decision-making, and time pressure for the sensemaking of existing design space. However, Co-led participants self-rated their iterated solutions as less competitive than existing solutions compared with Human-only and AI-led participants (F(2,43)=5.58, p=0.0071**). It might be that they focused more on the practicality than novelty, as P113 [Co-led] said: ”My competitive analysis statement: I think it was much less detailed than my previous answers. I think a lot of the ideas that were generated over here gives the company unique value. (But) I was not thinking…from a company perspective. I was more thinking of a generalization, or like a cultural change that would be required? I think the solving of a root cause cannot be done by something commercial.” Co-led participants indicated a more cynical view and deep consideration of the problem. 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.”

6 DISCUSSION

In summary, our study results shed light into how different degrees of agency in design thinking exercises influence creative outcome, cognitive processes, and self-efficacy. Interacting with AI for reflective thinking guided people to think more deeply, while passively consuming AI-generated content expanded people’s divergent thinking. In design activities that were led or co-led by humans, people tended to focus on precisely expressing their own viewpoints. When working in the AI-led design activity, people tended to maintain comprehensive information rather than intervening with their own opinions. Co-led participants responded to reflective questions and also had a higher self-reported ratings on the novelty of their initial solutions. They also valued the practicality of their iterated solutions after comparing with existing AI-gathered solutions.

6.1 Finding the right balance for AI-assisted reflective thinking

We observed that AI-led participants felt productive because the AI provided logical and well-written thoughts, as they had expected. In contrast, Co-led participants effectively expanded their critical reflection despite being challenged in their thinking. AI-led participants self-reported the highest average ratings for evaluating the information in the template among the three groups. However, based on post-study interviews, participants showed more agreement with the provided information than with independent thinking. This is noteworthy because prior research has indicated that people are likely to under-rely on AIs if they do not trust their performance [72]. However, 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. On the other hand, using AI to generate reflective questions was effective in engaging participants to consider issues they might not have thought about otherwise. AI-generated reflective questions helped participants focus on core issues rather than superficial details, as some Human-only participants without AI assistance reported that their initial responses were more about reactions than immediate problem-solving. However, these reflective questions were sometimes be repetitive or focus on details that differ from what participants want to address, potentially hindering their normal thought processes. When designing future human-AI systems for critical and reflective thinking, AI-generated responses should be logical and meet expectations, while AI-generated reflective questions should be diverse and novel.

6.2 Enhancing creative cognition with sense of user agency

Our results across different behavioral patterns in various conditions show that without AI assistance, people tend to review content back and forth (either by editing existing content or through web searching) to select the most valuable information to carry forward to the next stage. When the workflow is supplemented with AI-generated information, people aim to gather comprehensive or complementary information to proceed. They perceive the quality of ideas as a combination of as many ideas as possible, rather than honing in on one specific idea. 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. Based on this, future AI interventions in human-AI co-creativity can implement reflective questions with specific focus, priming people to develop downstream ideas in targeted directions.

6.3 Future study and limitations

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 AIs. Future human-AI collaboration for creativity research can continue to explore this area.

7 CONCLUSION

We integrated an LLM into existing design workflows in three ways (human-only, co-led, and AI-led) and used it to conduct an in-person lab experiment with 47 design novices to investigate how different AI positioning affected creative outcomes, cognitive processes, and self-efficacy. Our results showed that the co-led condition provoked deeper thinking about the problem and led participants to focus more on the practicality of their solutions, whereas the AI-led condition helped in expanding ideas that people can incorporate into their own.

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] 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.]).

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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

[18] Design Council. [n. d.]. The Double Diamond. ([n. d.]). https://www.designcouncil.org.uk/our-resources/the-double-diamond/

[19] 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.

[20] 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

[21] 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.

[22] 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.

[23] 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.

[24] 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*.

[25] 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.

[26] 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

[27] Xinyu Guo, Yi Xiao, Jiaqi Wang, and Tie Ji. 2023. Rethinking designer agency: A case study of co-creation between designers and AI. (2023).

[28] Ying Guo and Daniel Lee. 2023. Leveraging chatgpt for enhancing critical thinking skills. *Journal of Chemical Education* 100, 12 (2023), 4876–4883.

[29] 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.

[30] 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.

[31] Jeffrey Heer. 2019. Agency plus automation: Designing artificial intelligence into interactive systems. *Proceedings of the National Academy of Sciences* 116, 6 (2019), 1844–1850.

[32] 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.

[33] 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.

[34] 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).

[35] 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.

[36] 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

[37] 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.

[38] 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.

[39] 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

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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

[45] 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.

[46] 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).

[47] 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.

[48] Michael Luck, Mark d'Inverno, et al. 1995. A Formal Framework for Agency and Autonomy.. In *Icmas*, Vol. 95. 254–260.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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

[54] 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.

[55] 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.

[56] 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.

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

[58] 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.

[59] 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.

[60] 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.

[61] 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.

[62] 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

[63] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems* 35 (2022), 36479–36494.

[64] 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.

[65] 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.

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

[67] 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).

[68] 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.

[69] 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.

[70] 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.

[71] 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.

[72] 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

[73] 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.

[74] 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.

[75] 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.

[76] 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.

[77] 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.

[78] 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.

[79] 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.

[80] 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.

[81] 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.

[82] 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

[83] 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.

[84] 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.

[85] 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).