An Evidence-based Workflow for Studying and Designing Learning Supports for Human–AI Co-creation

Frederic Gmeiner
HCI Institute
Carnegie Mellon University
Pittsburgh, PA, USA
fog@andrew.cmu.edu

Jamie Conlin
Arcadia University
Glenside, PA, USA
jconlin@arcadia.edu

Eric Tang
HCI Institute
Carnegie Mellon University
Pittsburgh, PA, USA
ehtang@andrew.cmu.edu

Nikolas Martelaro*
HCI Institute
Carnegie Mellon University
Pittsburgh, PA, USA
nikmart@cmu.edu

Kenneth Holstein*
HCI Institute
Carnegie Mellon University
Pittsburgh, PA, USA
kjholste@cs.cmu.edu

ABSTRACT

Generative artificial intelligence (GenAI) systems introduce new possibilities for enhancing professionals' workflows, enabling novel forms of human–AI co-creation. However, professionals often struggle to learn to work with GenAI systems effectively. While research has begun to explore the design of interfaces that support users in learning to co-create with GenAI, we lack systematic approaches to investigate the effectiveness of these supports. In this paper, we present a systematic approach for studying how to support learning to co-create with GenAI systems, informed by methods and concepts from the learning sciences. Through an experimental case study, we demonstrate how our approach can be used to study and compare the impacts of different types of learning supports in the context of text-to-image GenAI models. Reflecting on these results, we discuss directions for future work aimed at improving interfaces for human–AI co-creation.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**.

KEYWORDS

Human–AI Interaction, Generative AI, Support Interfaces, Study Method, Learning, Human–AI Co-creation, Case Study

ACM Reference Format:

Frederic Gmeiner, Jamie Conlin, Eric Tang, Nikolas Martelaro, and Kenneth Holstein. 2024. An Evidence-based Workflow for Studying and Designing Learning Supports for Human–AI Co-creation. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3613905.3650763

*Co-senior authors contributed equally.

CHI EA '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0331-7/24/05
https://doi.org/10.1145/3613905.3650763

1 INTRODUCTION

Generative AI systems are rapidly improving, enabling new forms of human–AI "co-creation" [12]. Such systems have the potential to enhance professionals' creative practices, for example, by generating media artifacts like text, images, or video from user descriptions. Various studies have shown that GenAI systems can enhance creative work, especially at early stages such as ideation or sketching [22, 23].

However, research also shows that professionals struggle to work with GenAI systems effectively—especially in more goal-directed tasks where professionals aim to produce a specific outcome with the help of GenAI. In recent years, the HCI community has documented various challenges that professionals face in working with GenAI systems across a range of domains such as coding, illustration design, or engineering [17, 27, 29]. For example, in the context of text prompt-based interactions, users often struggle to craft input prompts in ways that will achieve desired outcomes, and they face difficulties in interpreting and repairing erroneous outputs [49].

Given these challenges, research has begun to explore mechanisms and interfaces to better support professionals in working co-creatively with GenAI. For example, Zamfirescu-Pereira *et al.* [49] propose interfaces that help users interactively explore AI model capabilities and limitations by labeling and comparing LLM-generated utterances in the context of a chatbot dialogue design system. Following a different approach, several works have proposed support interfaces that automatically suggest prompt variations that are meant to reflect best practices for prompting [6, 41].

While such systems explore promising directions, recent work has called for a more systematic approach to studying how to support human–AI co-creation [39, 43]. We build on these calls, drawing particular attention to the dearth of knowledge about how to support people in *learning* how to work co-creatively. Given the rapidly changing capabilities of GenAI systems, there is a need for new support interfaces that help professionals adapt to emerging modes of AI-augmented work.

In this paper, we take a first step to address this gap: **We present a methodological approach for studying how to support learning to co-create with AI systems.** Our approach is inspired by prior evidence-based methods from the field of learning sciences

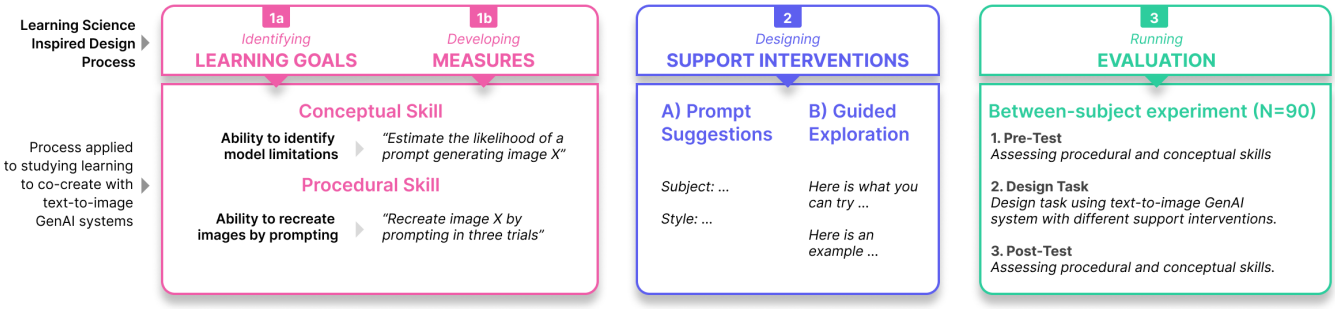


Figure 1: A process for studying and designing learning supports for human–AI co-creation. Top row: Learning science-inspired design process. Bottom row: Process as applied in our experimental case study.

for systematically studying the impacts of particular *learning interventions*, with respect to particular *learning goals* [11, 14, 25]. While these methods have been used to study and support learning in well-defined, closed-ended tasks in domains such as geometry or stoichiometry [15, 20], it is less clear how they can be adapted for more open-ended, co-creative contexts.

In this work, we explore an adaptation of this approach to study and design learning supports for working with GenAI systems. We propose following a learning science inspired pipeline consisting of the following steps: **(1) identifying learning goals**, **(2) developing measures**, **(3) designing support mechanisms**, and **(4) evaluating the resulting learning effects** (see Figure 1 top). Through an experimental case study with professional illustrators who worked with a text-to-image diffusion GenAI model for the first time, we demonstrate how this approach can be used to study and compare the impacts of different types of learning supports for human–AI co-creation with GenAI. Reflecting on these results, we highlight directions for future work aimed at studying and designing learning supports in the context of human–AI co-creation.

Overall, this paper makes two contributions: (1) we introduce an evidence-based workflow for studying and designing learning supports for human–AI co-creation, and (2) we provide an end-to-end demonstration of this workflow through an experimental case study.

2 TOWARD A RIGOROUS APPROACH FOR STUDYING LEARNING TO CO-CREATE WITH AI

Outside of human–AI (HAI) interaction research, the field of learning sciences has developed evidence-based methods for studying the impacts of learning interventions with respect to specific goals for human learning [11, 14, 25]. Core to these approaches is a "backward" approach to instructional design. Rather than starting with the design of the learning interventions themselves, researchers and designers first identify a set of fine-grained *learning goals*: specific skills and knowledge that learners should be able to demonstrate if a learning intervention is successful. Next, these approaches involve identifying or developing respective *measures*: instruments that can assess learning with respect to the identified learning goals. *Learning interventions* are then designed to align with these learning goals. Finally, the learning interventions' effectiveness, with respect to the learning goals, is evaluated through experimental studies using the learning measures.

Following this approach, studies have shown that measuring learning with respect to finer-grained learning goals that can be tied to specific observable abilities, rather than vague notions of "understanding," allows one to gain more informative insights for evaluating and refining learning interventions [4, 42]. We believe a similar methodological approach is a promising starting point for rigorously studying how to support learning to co-create with AI systems.

In this work, we explore to what extent this approach can be used to study how to support learning to co-create with AI systems. We propose a pipeline consisting of the following steps: (1) *identifying learning goals*, (2) *developing measures*, (3) *designing support mechanisms*, and (4) *evaluating the resulting learning effects* (see Figure 1 top). In the next section, we demonstrate this process through an experimental case study.

3 AN INITIAL END-TO-END CASE STUDY

To explore the value of this approach in the context of HAI co-creation, we set up and ran an initial experiment. Below, we describe the process of designing the experimental study, following the four steps outlined in the previous section.

3.1 Context: Supporting illustrators in learning to co-create with text-to-image GenAI

Our goal was to study learning to co-create with GenAI in an open-ended task domain. In this case study, we focused on illustration design tasks given that these involve complex conceptualization and communication skills [16]. Illustration design often balances goal-directed and open-ended requirements, making it a rich domain for studying human–AI co-creation.

Our study focused on supporting illustrators in learning to create images for a children's book using Stable Diffusion—a prompt-based open-source text-to-image diffusion model [38]. Such GenAI models take an input text prompt and try to return an image that best matches the input description. To run our experiment, we

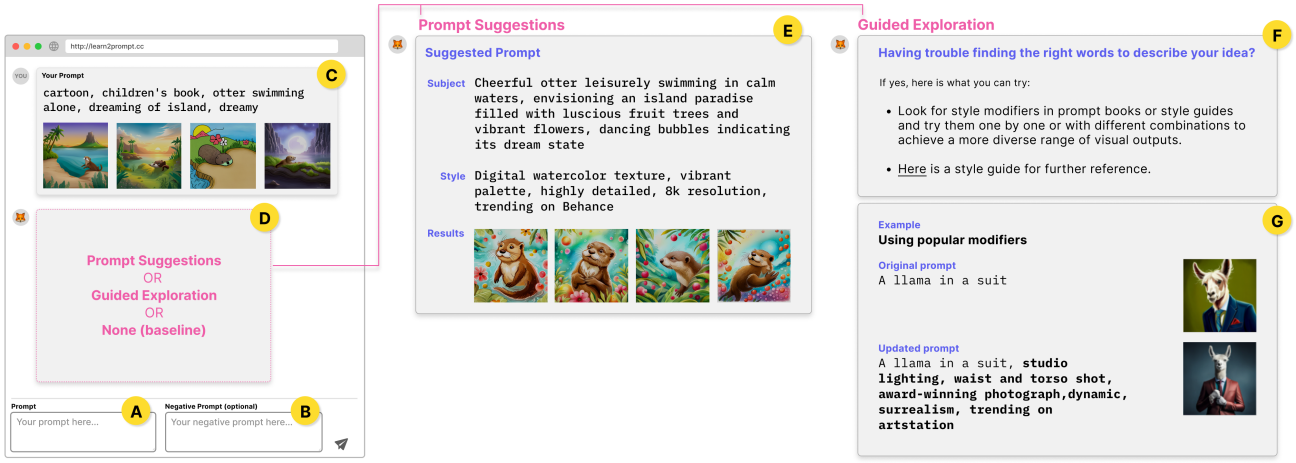


Figure 2: Interface elements of the chat-based GenAI design tool developed for our study. (A) First, participants submit a text prompt describing the desired image and (B) optionally, a negative text prompt describing concepts or elements that should be excluded from the image. (C) Next, the generated image and the original prompt appear in the chat history view. (D) Depending on the study condition, additional system-generated messages are displayed: (E) *Prompt Suggestions* offer alternative prompt variations by expanding the user's prompt with subject and style modifiers. (F) *Guided Exploration* suggests context-dependent trial-and-error exploration strategies along with (G) related worked examples.

developed a React-based web application that integrates Stable Diffusion into a simple chat-based interface (Figure 2).

Given our focus on illustration, we chose to study support mechanisms in the context of learning to co-create with text-to-image GenAI. Effectively working with current text prompt-based systems relies on a user's prompt engineering abilities—the skill of iteratively crafting effective input text prompts to achieve desired model outcomes. Beyond today's models' limitations, we expect that prompt engineering in its current form will substantially change, for example, due to models' increasing capabilities in interpreting user intents [48] or by enabling multi-modal input modalities, including sketching [32]. Nonetheless, we anticipate that the need will remain for users to learn and adapt their communication to the capabilities of specific co-creative AI agents, particularly when working in open-ended task domains. Thus, we use current text-to-image GenAI systems as a context through which to explore our broader approach.

3.2 Identifying Learning Goals

For the purpose of testing our approach, we selected two examples of skills that we speculate to be relevant for effective human–AI co-creation in prompt-based text-to-image AI models. However, we stress that there are many other learning objectives that are likely equally or more relevant to HAI co-creation, which should be explored by future research. In the following, we describe our process of identifying two goals for studying learning effects.

To select our two learning goals, we took inspiration from prior research that aims to improve human–AI collaboration by leveraging mechanisms known to support effective human-human collaborations, such as *group cognition*, *shared mental models* and *theory of mind* [1, 2, 8, 19, 24, 46, 47, 51]. While it remains an open question to what extent principles from human-human collaboration apply to human–AI teams, such constructs may provide useful starting points for identifying valuable goals and measures for human–AI co-creation. For example, in our study context, to work successfully with a given text-to-image AI model, a user must first learn about the model's limitations (for example, that achieving certain complex image compositions is challenging for the model). Furthermore, after learning about existing model limitations, the user then has to learn which prompting techniques can help to overcome these limitations (for example, emphasizing certain keywords in the prompt). This forms the foundation for our two selected learning goals, described below.

3.2.1 Conceptual Skill. For the first goal, we selected **knowledge of a specific model's limitations** (classified as a *conceptual skill*). Past research on HAI collaboration in decision-making contexts has shown that users' ability to recognize model limitations can enhance the overall quality of decision-making. For example, Bansal et al. [1] demonstrated that when humans rely on AI-generated output for decision-making, their understanding of the AI's error boundary, i.e., the areas where the AI is accurate versus inaccurate, helps them anticipate possible errors and decide when to override the automated inference. In the context of human-human collaboration, knowledge of a teammate's capabilities and limitations is known to enable adaptation and thereby support more effective collaborative work [40].

3.2.2 Procedural Skill. For the second goal, we selected **the ability to overcome model limitations by employing model-specific prompting techniques or strategies**, which we classified as a *procedural skill*. This skill was informed by previous studies on text prompt-based GenAI systems, which identified that effective

prompting requires specific techniques and strategies to overcome model limitations. For example, following instruction-like template structures, including specific keywords (prompt modifiers) or repeating phrases to emphasize concepts, can drastically improve model output [9, 28, 33]. Furthermore, previous studies indicate that effective prompting techniques can be learned through hands-on trial-and-error practice and that recreating images through prompting can serve as a proxy task to measure this skill [34, 45]. Through the lens of human-human collaboration, a crucial skill for effective collaboration is the ability to (partially) adapt one's own behavior to other group members' abilities and limitations [7, 40]. Translated to the context of learning to co-create with prompt-based text-to-image AI models, we speculate that this skill relates to a user's ability to adapt their prompting strategy according to the models' limitations to generate desired image outputs.

3.3 Developing Measures

We next developed measures to assess learning toward each of these two learning goals. To assess the *conceptual skill* (the ability to identify model limitations), we constructed a **survey instrument that asks users to estimate the likelihood that a text-to-image model would generate a given image based on a shown input prompt** (see Appendix Figure 3). The instrument consists of a prompt-image pair and several six-point Likert-like items. Each item asks participants to estimate the likelihood of the shown prompt resulting in an image that would match the depicted image in terms of *style*, *composition*, and *meaning*.

Furthermore, in line with prior literature, we constructed different assessment items (prompt-image pairs) to differentiate between *near* and *far* transfer assessment items to measure how well the application of a learned skill generalizes across different contexts. In particular, "near transfer" assessment items were more similar to those that participants would practice during the design task itself. "Far transfer" assessment items were intended to test generalization by introducing aspects participants would not encounter during the design task.

To assess the *procedural skill* (the ability to overcome model limitations), we constructed an **interactive survey instrument that asks participants to recreate a challenging image as closely as possible by prompting** (see Appendix Figure 4). The instrument asks users to provide positive and negative prompts to recreate a depicted target image. After submission, a new image is generated from the provided prompts and shown next to the target image for visual comparison. Then, users can refine their prompts and regenerate images two more times. After generating the third image, users select one image that best matches the target. This task design was inspired by previous studies of text-to-image AI models that used similar approaches to study prompting practices and model steerability [34, 45].

3.4 Designing Support Interfaces

After defining goals and measures, we implemented two support interfaces: *Prompt Suggestions* and *Guided Exploration* (Figure 2). We designed these based on interactions proposed in prior literature for supporting text prompt-based GenAI tasks. Although these support interfaces from prior literature were designed to support GenAI workflows, and were not explicitly designed to support *learning*, our aim was to understand the impacts these interactions might have on human learning.

We designed the first support interface, **Prompt Suggestions**, based on prior research prototypes [6] and commercial systems such as DALL-E 3 [5] that support users in working with text prompt-based GenAI systems by automatically suggesting prompt variations. In our implementation, after a user submits a prompt, the system automatically suggests an alternative prompt that follows common prompting best practices by extending and editing the user's original prompt. To implement this mechanism, we utilized an LLM pipeline similar to recent support systems such as [6] (see Appendix Figure 5 for further implementation details).

The second support interface, **Guided Exploration**, was inspired by prior work that proposes to support users in working with GenAI through a trial-and-error exploration of model capabilities, such as systematically testing different input-output combinations [49]. In our implementation, the system frequently provides support messages with suggestions for systematic trial and error strategies, along with worked examples of successful prompt and image pairs. We identified these prompt image-pair examples by conducting a literature review [9, 28, 33], screening online prompt support resources [21, 26, 44], and running formative pilot sessions prior to the actual case study experiment. The aim of the developed support interface is to guide users in systematically testing out different prompting techniques to overcome specific model limitations. We implemented this mechanism as a rule-based chatbot that would provide context-aware support messages (see Appendix Figure 6 for further implementation details).

3.5 Evaluation Study: Procedure

To investigate the learning impacts of each support interface with respect to our two specific learning goals, we conducted a remote between-subject study with 90 illustrators *(age in years M=34.3, SD=11.7)*. We recruited participants via the online platform Prolific who had at least two years of professional illustration experience *(years M=10.5, SD=9.6)* and little or no prior experience in working with prompt-based generative AI tools (such as ChatGPT or DALL-E). All participants were native or fluent in writing English and were paid 20 USD per hour. The study underwent approval by our university's IRB *(#2023_00000192)*. Each study session took 120 minutes and was split into four phases: (1) Onboarding, (2) Pre-test, (3) Design Task, and (4) Post-test.

Onboarding: At the beginning of the session, participants were presented with a prompt guide that described common prompting

Condition	**Number of Participants**	**Professional Experience** (years)
PROMPT SUGGESTIONS	*30*	*M=9.1, SD=7.2*
GUIDED EXPLORATION	*30*	*M=10.7, SD=9.1*
BASELINE	*30*	*M=11.6, SD=12.0*

Table 1: Participant counts and professional experience, by experimental condition.

techniques, basic functionalities, and general limitations of the Stable Diffusion model (see Appendix Figure 9 for further details).

Pre-test: In the 30-minute long pre-test, participants completed several activities that tested their abilities to recreate images through prompting and identify model limitations (the measures for assessing conceptual and procedural skills described in section 3.3).

Design Task: After the pre-test, participants started with the design task. First, they watched a short video explaining the design tool's interface functionality. Then, they were presented with a design brief that asked them to create two illustrations for a children's book using our generative AI web application. After that, participants had 20 minutes to work on each illustration task, during which they could generate as many images as desired. Participants were randomly assigned to one of the three support conditions that controlled the type of learning intervention they would be supported with during the design task: (1) *Prompt Suggestions*, (2) *Guided Exploration*, and (3) *Baseline – no interactive support* (see Table 1 for details). After task completion, they submitted one final image for each illustration task.

Post-test: Lastly, after the design task, participants completed a 30-minute post-test identical to the pre-test. To mitigate possible order effects, we randomized the sequence of the individual activities across all pre- and post-tests.

3.6 Analysis

To assess participants' learning, we compared their individual scores from the pre- and post-tests. To measure the impacts on participants' ability to identify model limitations, we compared the pre- and post-test scores of the Likert-like item responses for the style and composition likelihood estimations of the near and far transfer tasks. To measure participants' ability to overcome model limitations, we compared whether a participant's images from the pre- or post-test were visually closer to the given target image (in terms of style and composition). To assess the visual similarity of each image with the target image, we computed the cosine similarity between the image's CLIP embeddings and the target image's text prompt [36]. To analyze learning effects for each skill and transfer distance, we fitted linear models—estimated using ordinary least squares (OLS)—to predict participants' post-test scores with the support mechanism conditioned on participants' pre-test scores (formula: *post-score ~ mechanism * pre-score*).

3.7 Highlighted Findings

Overall, our results show that designers who received support through any of the two interactive interfaces performed slightly better than those designers without interactive support (Table 2). Among participants with lower prior ability (as measured by the pre-test), *Prompt Suggestions* had a positive effect on their ability to *overcome* model limitations in the image recreation task, compared with the *Baseline* condition, but had no significant effect on their ability to explicitly *identify* these limitations (see Appendix Figure 7 for further details). Meanwhile, *Guided Exploration* had the opposite effect. This intervention had a positive effect on participants' ability to *identify* model limitations (particularly in identifying *style similarity* but not *compositional similarity*) compared with the *Baseline* condition. However, *Guided Exploration* had no significant effect on participants' ability to *overcome* these limitations (see Appendix Figure 8 for further details).

Condition	**Conceptual Skill** *The ability to identify model limitations*	**Procedural Skill** *The ability to recreate an image through prompting*
PROMPT SUGGESTIONS	No effect	**Positive effect** *
GUIDED EXPLORATION	**Positive effect** *	No effect
BASELINE	No effect	No effect

**statistically significant*

Table 2: Overview of measured learning effects of conceptual and procedural skills after receiving different support messages (conditions). Participants in the *Prompt Suggestions* condition improved in the *procedural skill*, while those in the *Guided Exploration* condition improved in the *conceptual skill*. There were no learning effects observed for participants in the *Baseline* condition (see Appendix Figures 7 and 8 for further details).

4 DISCUSSION

In the previous sections, we proposed an evidence-based workflow for studying and designing learning supports for human–AI co-creation, drawing upon established approaches from the field of learning sciences. Through an experimental case study, we demonstrated that this approach can yield insight into the learning impacts of different interventions with respect to specific learning goals for human–AI co-creation. More broadly, we believe that the demonstrated approach provides a valuable and practical methodological foundation for future research aimed at improving HAI co-creation across task domains and AI models. Below, we reflect on our takeaways from this initial experiment and discuss possible opportunities to refine and apply this methodological approach in future research on HAI co-creation.

4.1 Reflections on the Case Study

Overall, although the observed effects were small, the results of our experiment indicate that both interactive support interfaces can support learning to work with generative AI systems. However, **each support mechanism promotes human learning toward different skills.** This points to the importance of designing learning supports for human–AI co-creation with particular learning goals in mind, and running evaluations to ensure that the targeted goals are truly supported.

Our *Guided Exploration* interface was intended to guide users in systematically testing out different ways to overcome a specific model's limitations. However, to do so, this interface often explicitly highlighted the model limitations themselves. This may be why this intervention improved participants' ability to explicitly identify such limitations, even though it did not improve their ability to overcome them. By contrast, the *Prompt Suggestions* interface appears to have been more effective in helping participants learn effective prompting strategies by example. However, this intervention did

not explicitly highlight and name specific model limitations, which may be why participants in this condition improved in their ability to overcome model limitations, but not to explicitly identify these limitations.

4.2 Toward a Taxonomy of Learning Goals for Human–AI Co-creation.

The learning goals we selected in the case study primarily served to test and demonstrate our approach to measuring fine-grained learning effects. While these goals for HAI co-creation via text-based prompting were inspired by factors identified in prior literature, we could have chosen many other reasonable learning goals that might exist. Here, we deliberately focused on these learning goals for the purpose of an initial exploration via an end-to-end case study. However, by building upon our process, future work should continue to identify more fine-grained learning goals for effectively working with GenAI systems. This includes skills for working with text-based prompt systems (as in our demonstration) but also for other non-text input modalities and domains, such as 2D sketching [50] or 3D geometries [30]. Comparing the results of different studies would also allow identifying more generalizable HAI co-creation skills across different GenAI models and task domains. **Ultimately, the goal would be to create a taxonomy of learning goals for human–AI co-creation**. As a starting point, future work could seek further inspiration from prior literature that has identified learning goals in open-ended tasks across various domains, such as debugging in CS education, project-based learning in maker spaces, or team learning within design teams [3, 35, 37].

4.3 Refining Measures and Analytics.

We constructed the measures in our case study to assess the extent to which participants would master a specific prompt-related skill. In the case of the *conceptual* skill (the ability to identify model limitations from given prompt-image pairs), the assessment relied on a likelihood estimation through Likert-like survey items, which directly reflects users' estimations. However, to assess the *procedural* skill (the ability to recreate images through prompting), we relied on measuring the visual similarity of the user-generated images with the target image. While we carefully calibrated our similarity comparison mechanisms based on comparable approaches found in prior literature [31], this analysis does not reveal any insights into specific prompting techniques and strategies employed in creating the images. Therefore, future work should, in addition to assessing the generated outcomes, also develop methods to analyze the prompting strategies of users. For example, in the case of text-based prompting, some prior work has started to utilize NLP methods to better assess users' prompting patterns and structures [13, 34].

4.4 Designing Support Mechanisms Tailored to Learning Goals.

Our primary goal was to provide an end-to-end demonstration of our method's ability to generate insights about different learning interventions' impacts with respect to specific, fine-grained learning goals for human–AI co-creation. For the purpose of testing our study approach, we chose to design and implement two support interfaces that follow different approaches for supporting working with GenAI systems as proposed by recent HCI work. Since the results show different learning effects for both interfaces, we are confident that the method generally allows us to detect nuanced learning effects in the context of HAI co-creation tasks. In future work, our workflow can be used twofold: (1) To evaluate the effectiveness of *existing* support strategies and interfaces to inform their improvement and (2) to design *new* support interventions *specifically tailored* toward fine-grain learning goals.

4.5 Exploring Continuous Evaluation Mechanisms

The case study's evaluation approach utilized pre- and post-tests to gauge learning effects within an experimental study setting. This approach allowed us to control the number of opportunities for participants to demonstrate their knowledge and to directly compare learning effects between support conditions. However, future work should also explore **approaches to continuously evaluate skills inside the task itself to allow more opportunities for skill demonstration**. Such knowledge tracing-inspired mechanisms [10] could eventually be directly integrated into a co-creative tool itself as a way to interactively adjust and offer support interventions depending on a user's skill level.

5 CONCLUSION

While emerging GenAI systems have the potential to augment professional work and enable new forms of human–AI co-creation, current systems pose many challenges for professional users to adopt GenAI systems into their workflows. Recent research has started to suggest mechanisms to better support users in learning to work with GenAI systems. However, we currently lack a systematic approach to evaluate the impact of support interventions on humans' learning to co-create with AI. In this paper, we presented an evidence-based workflow for studying and designing learning supports for human–AI co-creation by taking inspiration from prior studies in the field of learning sciences. Furthermore, we demonstrated that by following this approach, we were able to gain insights into the impacts of different support interventions on fine-grained learning goals within the context of working with text-to-image GenAI models. While this paper represents a first attempt to study the learning effects of support interfaces within text-to-image GenAI tasks, we believe that the demonstrated approach provides a valuable and practical methodological foundation for future research aiming at improving human–AI co-creation across task domains and AI models. We hope our work inspires future research to build upon and collectively enhance support systems for more effective and complementary human–AI co-creation.

ACKNOWLEDGMENTS

We want to thank all study participants for supporting this work and Minjune Song for supporting the development of the design tool. This material is based upon work supported by *Prolific* and the *National Science Foundation* under Grant No. 2118924 *Supporting Designers in Learning to Co-create with AI for Complex Computational Design Tasks*. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s)

and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S. Lasecki, Daniel S. Weld, and Eric Horvitz. 2019. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7, 1 (Oct. 2019), 2–11.

[2] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S. Weld, Walter S. Lasecki, and Eric Horvitz. 2019. Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff. *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (July 2019), 2429–2437. https://doi.org/10.1609/aaai.v33i01.33012429

[3] Yoav Bergner, Samuel Abramovich, Marcelo Worsley, and Ofer Chen. 2019. What Are the Learning and Assessment Objectives in Educational Fab Labs and Makerspaces?. In *Proceedings of FabLearn 2019*. ACM, New York NY USA, 42–49. https://doi.org/10.1145/3311890.3311896

[4] Dawn Berk and James Hiebert. 2009. Improving the Mathematics Preparation of Elementary Teachers, One Lesson at a Time. Teachers and Teaching: Theory and Practice, 15(3), 337-356. *Teachers and Teaching* 15 (June 2009), 337–356. https://doi.org/10.1080/13540600903056692

[5] James Betker, Gabriel Goh, Li Jing, TimBrooks, Jianfeng Wang, Linjie Li, LongOuyang, JuntangZhuang, JoyceLee, YufeiGuo, WesamManassra, PrafullaDhariwal, CaseyChu, YunxinJiao, and Aditya Ramesh. 2023. Improving Image Generation with Better Captions.

[6] Stephen Brade, Bryan Wang, Mauricio Sousa, Sageev Oore, and Tovi Grossman. 2023. Promptify: Text-to-Image Generation through Interactive Prompt Exploration with Large Language Models. arXiv:2304.09337 [cs]

[7] C. Shawn Burke, Kevin C. Stagl, Eduardo Salas, Linda Pierce, and Dana Kendall. 2006. Understanding Team Adaptation: A Conceptual Analysis and Model. *Journal of Applied Psychology* 91, 6 (Nov. 2006), 1189–1207. https://doi.org/10.1037/0021-9010.91.6.1189

[8] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 1–24. https://doi.org/10.1145/3359206

[9] Minsuk Chang, Stefania Druga, Alexander J. Fiannaca, Pedro Vergani, Chinmay Kulkarni, Carrie J Cai, and Michael Terry. 2023. The Prompt Artists. In *Creativity and Cognition*. ACM, Virtual Event USA, 75–87. https://doi.org/10.1145/3591196.3593515

[10] Albert T. Corbett and John R. Anderson. 1995. Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. *User Modelling and User-Adapted Interaction* 4, 4 (1995), 253–278. https://doi.org/10.1007/BF01099821

[11] Alejandro Mauricio Dávila Rubio. 2017. Wiggins, G., & McTighe, J. (2005) Understanding by Design (2nd Ed.). Alexandria, VA: Association for Supervision and Curriculum Development ASCD. *Colombian Applied Linguistics Journal* 19, 1 (Feb. 2017), 140. https://doi.org/10.14483/calj.v19n1.11490

[12] Nicholas Davis, Chih-Pin Hsiao, Yanna Popova, and Brian Magerko. 2015. An Enactive Model of Creativity for Computational Collaboration and Co-creation. In *Creativity in the Digital Age*, Nelson Zagalo and Pedro Branco (Eds.). Springer London, London, 109–133. https://doi.org/10.1007/978-1-4471-6681-8_7

[13] Michael Desmond and Michelle Brachman. 2024. Exploring Prompt Engineering Practices in the Enterprise. https://doi.org/10.48550/arXiv.2403.08950 arXiv:2403.08950 [cs]

[14] Walter Dick, Lou Carey, and James O. Carey. 2005. *The Systematic Design of Instruction* (6th ed ed.). Pearson/Allyn and Bacon Boston, Boston.

[15] Karen L. Evans, David Yaron, and Gaea Leinhardt. 2008. Learning Stoichiometry: A Comparison of Text and Multimedia Formats. *Chem. Educ. Res. Pract.* 9, 3 (2008), 208–218. https://doi.org/10.1039/B812409B

[16] Michael Fleishman. 2007. *How to Grow as an Illustrator*. Allworth Press, New York, NY.

[17] Frederic Gmeiner, Humphrey Yang, Lining Yao, Kenneth Holstein, and Nikolas Martelaro. 2023. Exploring Challenges and Opportunities to Support Designers in Learning to Co-create with AI-based Manufacturing Design Tools. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–20. https://doi.org/10.1145/3544548.3580999

[18] Gustavosta. 2023. MagicPrompt-Stable-Diffusion. https://huggingface.co/Gustavosta/MagicPrompt-Stable-Diffusion.

[19] Kenneth Holstein, Vincent Aleven, and Nikol Rummel. 2020. A Conceptual Framework for Human–AI Hybrid Adaptivity in Education. In *Artificial Intelligence in Education*, Ig Ibert Bittencourt, Mutlu Cukurova, Kasia Muldner, Rose Luckin, and Eva Millán (Eds.). Vol. 12163. Springer International Publishing, Cham, 240–254. https://doi.org/10.1007/978-3-030-52237-7_20

[20] Amanda Jansen, Tonya Bartell, and Dawn Berk. 2009. The Role of Learning Goals in Building a Knowledge Base for Elementary Mathematics Teacher Education. *The Elementary School Journal* 109, 5 (May 2009), 525–536. https://doi.org/10.1086/597000

[21] Groove Jones. 2023. AI Prompts for Generative Art – Midjourney, DALL-E, and Stable Diffusion. https://groovejones.com/explore-prompting-for-midjourney-generative-ai-artificial-intelligence-art/.

[22] Hyeonsu B. Kang, David Chuan-En Lin, Nikolas Martelaro, Aniket Kittur, Yan-Ying Chen, and Matthew K. Hong. 2023. BioSpark: An End-to-End Generative System for Biological-Analogical Inspirations and Ideation. https://doi.org/10.48550/arXiv.2312.11388 arXiv:2312.11388 [cs]

[23] 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 (IUI '20)*. Association for Computing Machinery, New York, NY, USA, 221–230. https://doi.org/10.1145/3377325.3377522

[24] Janin Koch and Antti Oulasvirta. 2018. Group Cognition and Collaborative AI. In *Human and Machine Learning*, Jianlong Zhou and Fang Chen (Eds.). Springer International Publishing, Cham, 293–312. https://doi.org/10.1007/978-3-319-90403-0_15

[25] Kenneth R. Koedinger, Albert T. Corbett, and Charles Perfetti. 2012. The Knowledge-Learning-Instruction Framework: Bridging the Science-Practice Chasm to Enhance Robust Student Learning. *Cognitive Science* 36, 5 (2012), 757–798. https://doi.org/10.1111/j.1551-6709.2012.01245.x

[26] Lexica. 2023. Lexica. https://lexica.art/.

[27] Michael Xieyang Liu, Advait Sarkar, Carina Negreanu, Benjamin Zorn, Jack Williams, Neil Toronto, and Andrew D. Gordon. 2023. "What It Wants Me To Say": Bridging the Abstraction Gap Between End-User Programmers and Code-Generating Large Language Models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–31. https://doi.org/10.1145/3544548.3580817

[28] Vivian Liu and Lydia B Chilton. 2022. Design Guidelines for Prompt Engineering Text-to-Image Generative Models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–23. https://doi.org/10.1145/3491102.3501825

[29] Vivian Liu, Han Qiao, and Lydia Chilton. 2022. Opal: Multimodal Image Generation for News Illustration. arXiv:2204.09007 [cs]

[30] Justin Matejka, Michael Glueck, Erin Bradner, Ali Hashemi, Tovi Grossman, and George Fitzmaurice. 2018. Dream Lens: Exploration and Visualization of Large-Scale Generative Design Datasets. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–12. https://doi.org/10.1145/3173574.3173943

[31] Dylan Moore, Tobias Dahl, Paula Varela, Wendy Ju, Tormod Næs, and Ingunn Berget. 2019. Unintended Consonances: Methods to Understand Robot Motor Sound Perception. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300730

[32] Changhoon Oh, Jungwoo Song, Jinhan Choi, Seonghyeon Kim, Sungwoo Lee, and Bongwon Suh. 2018. I Lead, You Help but Only with Enough Details: Understanding User Experience of Co-Creation with Artificial Intelligence. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3174223

[33] Jonas Oppenlaender. 2022. A Taxonomy of Prompt Modifiers for Text-To-Image Generation. https://doi.org/10.48550/arXiv.2204.13988 arXiv:2204.13988 [cs]

[34] Jonas Oppenlaender, Rhema Linder, and Johanna Silvennoinen. 2023. Prompting AI Art: An Investigation into the Creative Skill of Prompt Engineering. arXiv:2303.13534 [cs]

[35] Cyril Picard, Cécile Hardebolle, Roland Tormey, and Jürg Schiffmann. 2022. Which Professional Skills Do Students Learn in Engineering Team-Based Projects? *European Journal of Engineering Education* 47, 2 (March 2022), 314–332. https://doi.org/10.1080/03043797.2021.1920890

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs]

[37] Kathryn M. Rich, Carla Strickland, T. Andrew Binkowski, and Diana Franklin. 2019. A K-8 Debugging Learning Trajectory Derived from Research Literature. In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*. ACM, Minneapolis MN USA, 745–751. https://doi.org/10.1145/3287324.3287396

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. https://doi.org/10.48550/arXiv.2112.10752 arXiv:2112.10752 [cs]

[39] Wout Schellaert, Fernando Martínez-Plumed, Karina Vold, John Burden, Pablo A. M. Casares, Bao Sheng Loe, Roi Reichart, Sean Ó hÉigeartaigh, Anna Korhonen, and José Hernández-Orallo. 2023. Your Prompt Is My Command: On Assessing the Human-Centred Generality of Multimodal Models. *Journal of Artificial Intelligence Research* 77 (June 2023), 377–394. https://doi.org/10.1613/jair.1.14157

[40] Valerie Sessa and Manuel London (Eds.). 2007. *Work Group Learning: Understanding, Improving and Assessing How Groups Learn in Organizations* (1 ed.). Psychology Press. https://doi.org/10.4324/9780203809747

[41] Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 4222–4235. https://doi.org/10.18653/v1/2020.emnlp-main.346

[42] John C. Stamper and Kenneth R. Koedinger. 2011. Human-Machine Student Model Discovery and Improvement Using DataShop. In *Artificial Intelligence in Education (Lecture Notes in Computer Science)*, Gautam Biswas, Susan Bull, Judy Kay, and Antonija Mitrovic (Eds.). Springer, Berlin, Heidelberg, 353–360. https://doi.org/10.1007/978-3-642-21869-9_46

[43] Lev Tankelevitch, Viktor Kewenig, Auste Simkute, Ava Elizabeth Scott, Advait Sarkar, Abigail Sellen, and Sean Rintel. 2023. The Metacognitive Demands and Opportunities of Generative AI. https://doi.org/10.48550/arXiv.2312.10893 arXiv:2312.10893 [cs]

[44] thesephist. 2023. Collection of Useful Stable Diffusion Prompt Modifiers. https://gist.github.com/thesephist/376afed2cbfce35d4b37d985abe6d0a1.

[45] Kailas Vodrahalli and James Zou. 2023. ArtWhisperer: A Dataset for Characterizing Human-AI Interactions in Artistic Creations. arXiv:2306.08141 [cs]

[46] Dakuo Wang, Elizabeth Churchill, Pattie Maes, Xiangmin Fan, Ben Shneiderman, Yuanchun Shi, and Qianying Wang. 2020. From Human-Human Collaboration to Human-AI Collaboration: Designing AI Systems That Can Work Together with People. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–6. https://doi.org/10.1145/3334480.3381069

[47] Qiaosi Wang, Koustuv Saha, Eric Gregori, David Joyner, and Ashok Goel. 2021. Towards Mutual Theory of Mind in Human-AI Interaction: How Language Reflects What Students Perceive About a Virtual Teaching Assistant. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–14. https://doi.org/10.1145/3411764.3445645

[48] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. 2023. Human Preference Score: Better Aligning Text-to-Image Models with Human Preference. arXiv:2303.14420 [cs]

[49] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–21. https://doi.org/10.1145/3544548.3581388

[50] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. https://doi.org/10.48550/arXiv.2302.05543 arXiv:2302.05543 [cs]

[51] Rui Zhang, Nathan J. McNeese, Guo Freeman, and Geoff Musick. 2021. "An Ideal Human": Expectations of AI Teammates in Human-AI Teaming. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (Jan. 2021), 1–25. https://doi.org/10.1145/3432945

A APPENDIX

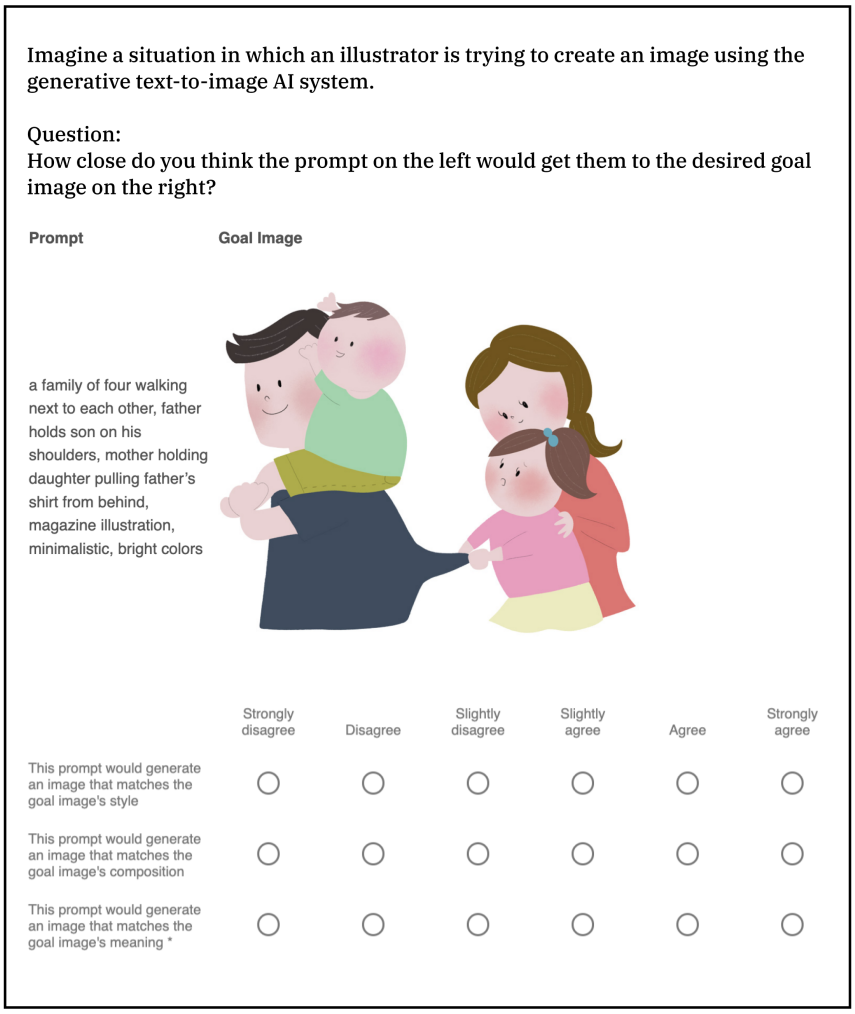


Figure 3: Example of the survey instrument to assess participants' ability to identify GenAI model limitations (conceptual skill). The instrument consists of a prompt-image pair (middle) and several six-point Likert-like items (bottom). Each item asks participants to estimate the likelihood of the shown prompt resulting in an image that would match the depicted image in terms of *style*, *composition*, and *meaning*. In this example, the shown prompt is relatively simple and would probably not generate an image that matches the goal image's composition because it depicts different characters with complex interactions, which is challenging to achieve with this model.

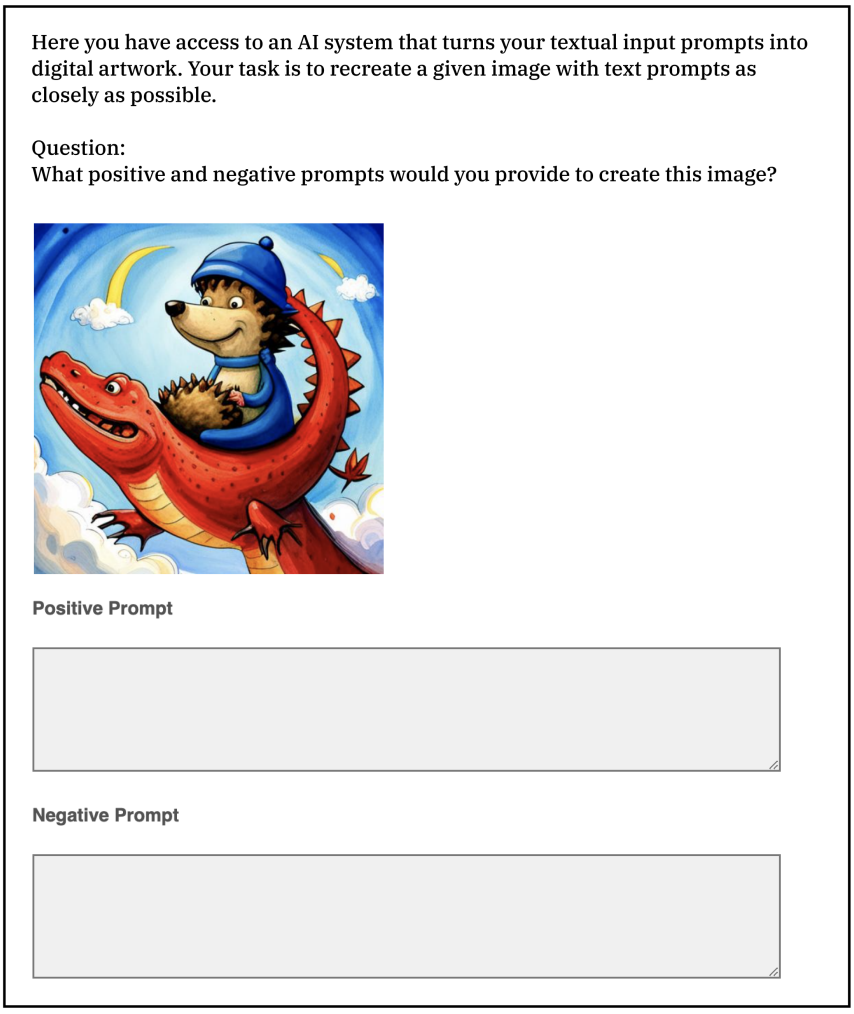


Figure 4: Example of the interactive survey instrument to assess participants' ability to recreate a given image through prompting (procedural skill). The instrument asks users to provide positive and negative prompts to recreate a depicted target image. Not shown in this figure: After submission, a new image is generated from the provided prompts and shown next to the target image for visual comparison. Then, users can refine their prompts and regenerate images two more times. After generating the third image, users select one image that best matches the target.

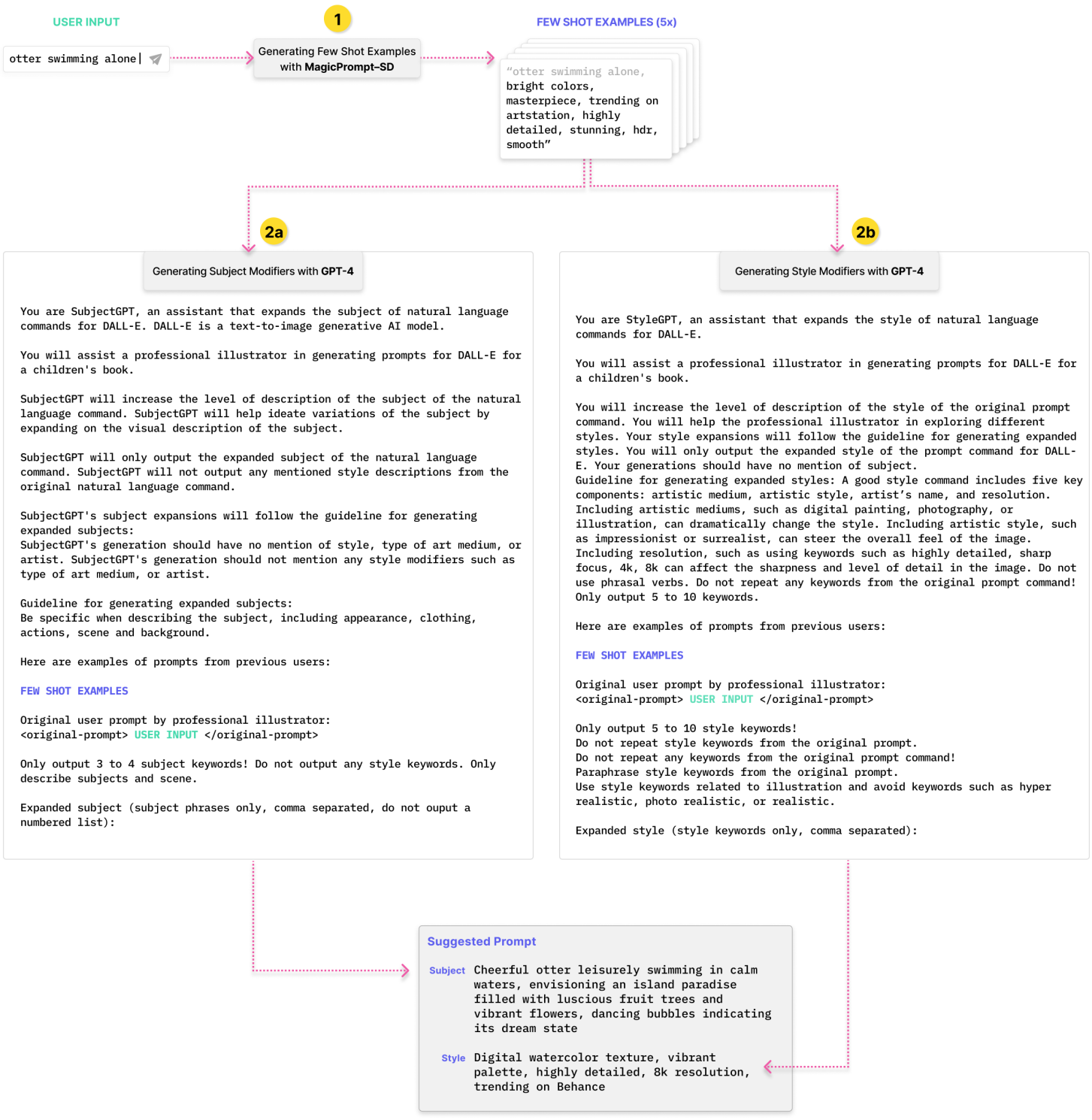


Figure 5: Process diagram of LLM pipeline for generating Prompt Suggestions: (1) First, five different prompt variations are generated from the original user prompt using MagicPrompt-SD [18], a GPT-2-based large language model fine-tuned on 80,000 stable diffusion prompts from the online community platform lexica.ai. This is done to generate domain context-specific few-shot prompt examples for GPT-4 at a later stage. Then, subject and style modifiers are generated separately (2a and 2b) by prompting GPT-4. These prompts consist of static instructions along with the dynamically generated few shot examples and the original user input prompt.

*Not really sure if the **style modifiers** in your prompt are actually having an effect on the image?*

If yes, here is what you can try:

- Experiment with **adding** or **removing** modifiers one at a time.
- Generate a new image to **compare the difference.**

Example

Adding a style modifier

Original prompt
an ice cream cone

Updated prompt
a **cartoon** ice cream cone

Are you not seeing drastic enough changes between iterations?

If yes, here is what you can try:

- Introduce **volatility** by looking for **style modifiers** online and introducing them one by one into your prompt.
- Adding **specific types** of style modifiers like **medium, artist name,** or **time period** may help.

Example

Adding specific types of modifiers

Original prompt
a shark lying on the sand at the beach

Updated prompt
a **digital painting** of a shark lying on the sand at the beach, **by Paul Cezanne, Gothic Period**...

*Is there a **mood, atmosphere, object** or **interaction** you want to **emphasize** more in your image?*

If yes, here is what you can try:

- Experiment with **repeating** those elements to reinforce their importance.
- Keep exploring **variations** of the prompt to see how the AI responds.

Example

Repeating Keywords

Original prompt
...an illustration of a koala drinking tea and eating cookies...

Updated prompt
an illustration of a koala drinking tea and eating cookies, ... **cookies, cookies, cookies, cookies**...

*Are there subjects that get **merged together** in the images?*

If yes, here is what you can try:

- Experiment with **describing interactions** of your subjects.
- **Explore** variations of the prompt to see how the AI interprets it differently.

Example

Describing interactions

Original prompt
... a bear and a rabbit ....

Updated prompt
... a bear **dancing with** a rabbit .....

*Are there any elements you would like to **remove** from an image?*

If yes, here is what you can try:

- **Add** them to the **negative prompt** or **modify** your **positive prompt** accordingly.

Example

Using Negative Prompt

Original prompt
...a photo of an astronaut in space...

Positive prompt
...a photo of an astronaut in space.....

Negative prompt
moon, planet

*Do all of your latest images **look the same?***

If yes, here is what you can try:

- Set the Seed to "**Auto**" and try generating the prompt multiple times to observe if there are variations in the results.
- Sometimes, **generating the prompt repeatedly** can yield different outputs that might align better with your expectations.

Example

Generating with the same prompt

First Attempt
...a cat sleeping under a tree...

Second Attempt
...a cat sleeping under a tree.....

*Has the system generated an image **close** to what you envision, but you want to **make it more similar**?*

If yes, here is what you can try:

- **Reuse the seed** of that image and only **slightly modify** the prompt in each iteration to continue working off of it with more acute changes.
- Click the ***Reuse Prompt and Parameters*** button below an image to reuse its seed and prompts.

Example

Reusing an image seed

Original prompt
... a butterfly, flowers, ...

[seed: 8633]

Second Attempt
... an **organge** butterfly, flowers, ...

[seed: 8633]

*Does the AI system understand your prompt **like a person would?***

If not, here is what you can try:

- Find **better descriptions** of your design goal for the AI by experimenting with **synonyms** or **alternative words** to your original description.
- Try using **less obvious** or **specific terminology**, which might lead to outcomes closer to your intended goal.

Example

Synonyms

Original prompt
a Greek goddess...drinking from a crystal cup

Updated prompt
a Greek goddess ...drinking from a crystal **wine glass**...

*Having trouble **finding the right words** to describe your idea?*

If yes, here is what you can try:

- Look for **style modifiers** in **prompt books** or **style guides** and try them one by one or with different combinations to achieve a more diverse range of visual outputs.

→ Here is a **style guide** for further reference

Example

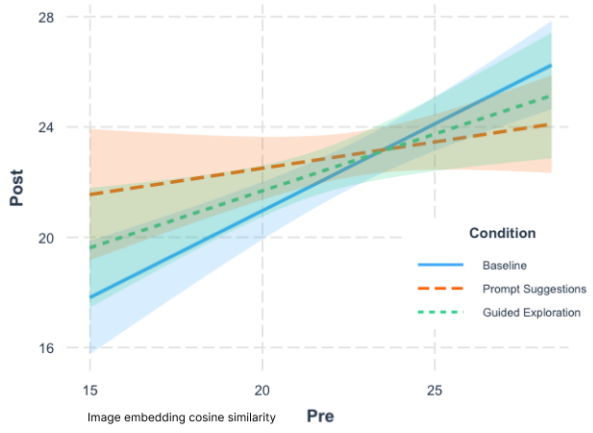
Using popular modifiers

Original prompt
a llama in a suit

Updated prompt
a llama in a suit, **studio lighting, waist and torso shot, award-winning photograph,dynamic, surrealism,trending on artstation**

Figure 6: Examples of messages designed for the Guided Exploration support bot mechanism. Each message contains suggestions for systematic trial and error exploration (upper part) and worked examples of successful prompts and image pairs (lower part) following best practices identified in prior literature.

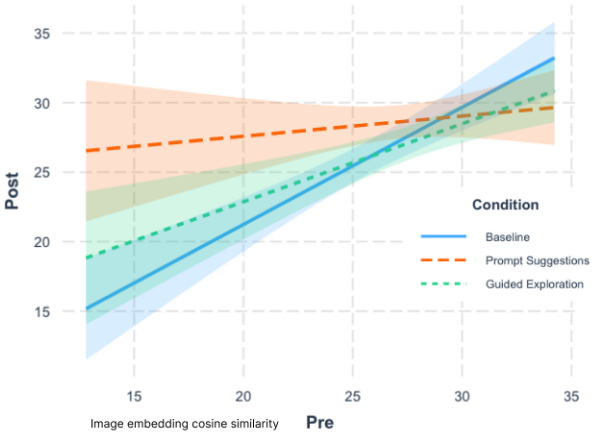
Procedural Skill – Near Transfer Task



	Post		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	8.37	2.77 – 13.98	**0.004**
Pre	0.63	0.39 – 0.87	**<0.001**
Condition [Prompt Suggestions]	10.31	1.69 – 18.94	**0.020**
Condition [Guided Exploration]	5.06	-3.68 – 13.81	0.253
Pre × Condition [Prompt Suggestions]	-0.44	-0.81 – -0.06	**0.022**
Pre × Condition [Guided Exploration]	-0.22	-0.61 – 0.18	0.277
Observations	90		
R^2 / R^2 adjusted	0.304 / 0.263		

We fitted a linear model (estimated using OLS) to predict Post with Pre and Condition (formula: **Post ~ Pre * Condition**). Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95% Confidence Intervals (CIs) and p-values were computed using a Wald t-distribution approximation.

Procedural Skill – Far Transfer Task

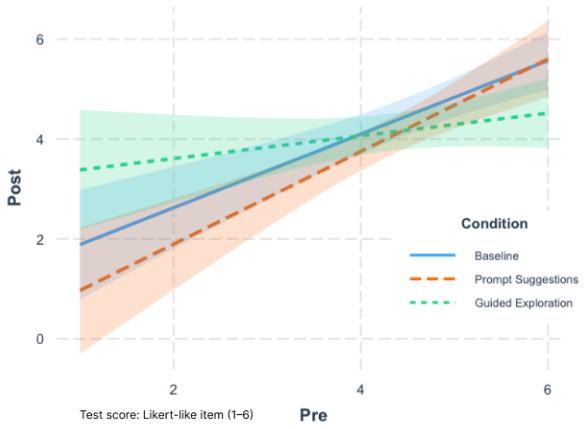


	Post		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	4.38	-2.62 – 11.38	0.217
Pre	0.84	0.57 – 1.11	**<0.001**
Condition [Prompt Suggestions]	20.30	8.58 – 32.02	**0.001**
Condition [Guided Exploration]	7.27	-3.83 – 18.36	0.197
Pre × Condition [Prompt Suggestions]	-0.70	-1.13 – -0.26	**0.002**
Pre × Condition [Guided Exploration]	-0.28	-0.69 – 0.13	0.172
Observations	90		
R^2 / R^2 adjusted	0.426 / 0.392		

We fitted a linear model (estimated using OLS) to predict Post with Pre and Condition (formula: **Post ~ Pre * Condition**). Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95% Confidence Intervals (CIs) and p-values were computed using a Wald t-distribution approximation.

Figure 7: Linear regression results for assessment items targeting participants' ability to overcome model limitations (procedural skill).

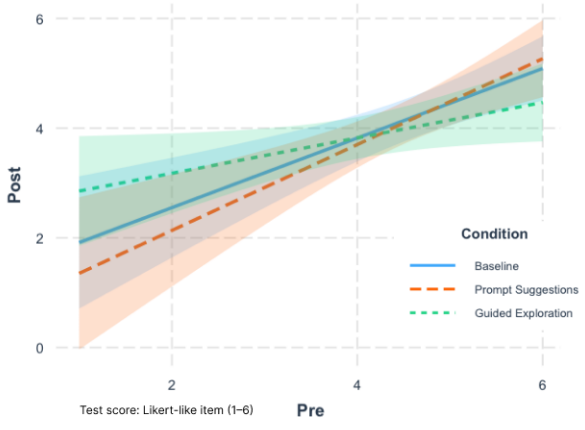
Conceptual Skill – Far Transfer Task (Style Similarity)



		Post	
Predictors	*Estimates*	*CI*	*p*
(Intercept)	1.15	-0.22 – 2.52	0.099
Pre	0.74	0.44 – 1.03	**<0.001**
Condition [Prompt Suggestions]	-1.11	-3.24 – 1.02	0.302
Condition [Guided Exploration]	2.00	-0.05 – 4.06	0.056
Pre × Condition [Prompt Suggestions]	0.19	-0.28 – 0.67	0.428
Pre × Condition [Guided Exploration]	-0.51	-0.96 – -0.06	**0.028**
Observations	90		
R^2 / R^2 adjusted	0.394 / 0.358		

We fitted a linear model (estimated using OLS) to predict Post with Pre and Condition (formula: **Post ~ Pre * Condition**). Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95% Confidence Intervals (CIs) and p-values were computed using a Wald t-distribution approximation.

Conceptual Skill – Far Transfer Task (Compositional Similarity)



		Post	
Predictors	*Estimates*	*CI*	*p*
(Intercept)	1.28	-0.23 – 2.79	0.095
Pre	0.63	0.32 – 0.95	**<0.001**
Condition [Prompt Suggestions]	-0.71	-3.03 – 1.61	0.543
Condition [Guided Exploration]	1.25	-0.73 – 3.23	0.214
Pre × Condition [Prompt Suggestions]	0.15	-0.35 – 0.65	0.555
Pre × Condition [Guided Exploration]	-0.31	-0.75 – 0.13	0.163
Observations	90		
R^2 / R^2 adjusted	0.314 / 0.273		

We fitted a linear model (estimated using OLS) to predict Post with Pre and Condition (formula: **Post ~ Pre * Condition**). Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95% Confidence Intervals (CIs) and p-values were computed using a Wald t-distribution approximation.

Figure 8: Linear regression results for assessment items targeting participants' ability to identify model limitations (conceptual skill).

🎨 Guide to creating images with generative text-to-image AI systems

Generative **text-to-image AI systems** can turn **textual descriptions into images** via **prompting**.
The AI model you work with during this study is called **"Stable Diffusion."**

1. Prompt Example 🖼️

Here is an **example prompt** and **corresponding image**:

Prompt	**Generated Image**
a beautiful landscape of a tiny futuristic village in the french countryside during spring season, painting by studio ghibli backgrounds and frederic edwin church hd and louis remy mignot hd, nice spring afternoon lighting, smooth tiny details, soft and clear shadows, low contrast, perfect	

2. Negative Prompts ❌

In addition to the prompts, you can specify negative prompts, which can be used to say **what should NOT appear in the image**.

The example negative prompt *"Bright colors, people, realism"* will result in images **avoiding bright colors, people as a subject, and realism as a style**.

3. Seed 🌱

The seed parameter is a **number that guides the creation of the image** by defining a **starting point**. The same seed with the same prompt produces the same image every time. **Keeping the same seed** enables **reproducibility** while **changing the seed** enables the **exploration of different variations**.

4. Common Best Practices 👍

- For image generation, sticking to a **standard prompt structure** like
  [Subject] [Medium] [Artist(s)] [Details] [Style Modifiers]
  may help make it an easier format for the AI to understand your design goal.
- Prompts can be composed of so called ***style modifiers***, which are **descriptive words or short phrases** that are **separated by commas** that are used to dictate the aesthetic of the image. Modifiers can include **historical or artistic periods, artists, mediums, or adjectives.** These can heavily influence the overall look and quality of an image, so do not be afraid to include as many as you can to communicate your design goal with the AI system.
- **Negative Prompt:** Usually, it is better to simply input the elements or styles that you don't want (**without "no" or "remove"**) in the negative prompt. E.g. "horse" instead of "no horse" if you want to exclude a horse from your image.
- If you find part of your prompt is missing from the generated images, you can try **adding synonyms**, using **repetition**, or **rephrasing**.
- Remember that the **AI "understands" words differently from people** and may respond better to specific keywords, styles, or internet terms.

5. How does this work? 🤖

The underlying generative AI system is **trained with images and image captions from the internet**.

By detecting patterns and structures within these image and text pairs, the models can generate new images that match a given text description (prompt).

However, generative models, like all AI models, are impacted by the data used during their training. When biases exist within the training data, these biases can manifest in the images produced by the model in various ways. In particular, **due to its training data sourced from the internet, the model could exhibit a pronounced inclination towards specific keywords/aesthetics that prevailed as popular trends on social media platforms and applications.**

6. Known Limitations of Generative Models ☝️

- **Hands** and **feet** are famously **difficult** for AI **to generate**.
- The systems **require specific and detailed textual descriptions**. If the prompt is vague or lacks clarity, the resulting image may not accurately depict the intended concept.
- Producing **realistic and precise images** poses a **problem for generative models**, as they often introduce extra elements or fail to capture essential details.
- The systems can face **difficulty in comprehending the contextual connections among objects** within an image, leading to the generation of image outputs that may appear odd or unrealistic.

7. Safety Filter ❌

- Like many other generative AI models, this text-to-image model has a **content filter** to **prevent** generating **images that contain potentially harmful** or **illegal content**.
- Even if your prompt does not explicitly request harmful or illegal images, terms such as *"children," "kids,"* or similar terms might trigger the content filter mechanism in certain cases.
- If you encounter such issues, **rephrase your prompt** and **avoid potentially problematic terms**.

8. Online Help Resources 📚

There are many helpful **online resources** available for learning to work with generative AI systems. For example, a useful resource for prompting is **lists of popular style modifiers** or **collections of generated images along with their prompts**.

Here are some websites you can use to get started:
Please open these links and keep them open in separate tabs

- A collection of popular Style Modifiers grouped by categories
- A list of Art Mediums with image examples
- A collection of Aesthetics

Figure 9: The prompt guide that all participants received at the beginning of the study session before the pre-test.