Exploring the Potential of Metacognitive Support Agents for Human-AI Co-Creation

Frederic Gmeiner
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, PA, USA
gmeiner@cmu.edu

Kaitao Luo
Carnegie Mellon University
Pittsburgh, PA, USA
kaitaol@andrew.cmu.edu

Ye Wang
Autodesk Research
San Francisco, CA, USA
whyzcandy@gmail.com

Kenneth Holstein
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, PA, USA
kjholste@cs.cmu.edu

Nikolas Martelaro
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, PA, USA
nikmart@cmu.edu

Abstract

Despite the potential of generative AI (GenAI) design tools to enhance design processes, professionals often struggle to integrate AI into their workflows. Fundamental cognitive challenges include the need to specify all design criteria as distinct parameters upfront (intent formulation) and designers' reduced cognitive involvement in the design process due to cognitive offloading, which can lead to insufficient problem exploration, underspecification, and limited ability to evaluate outcomes. Motivated by these challenges, we envision novel *metacognitive support agents* that assist designers in working more reflectively with GenAI. To explore this vision, we conducted exploratory prototyping through a Wizard of Oz elicitation study with 20 mechanical designers probing multiple metacognitive support strategies. We found that agent-supported users created more feasible designs than non-supported users, with differing impacts between support strategies. Based on these findings, we discuss opportunities and tradeoffs of metacognitive support agents and considerations for future AI-based design tools.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → *Artificial intelligence.*

Keywords

Human-AI Interaction, Support Interfaces, Metacognition, Wizard of Oz

ACM Reference Format:
Frederic Gmeiner, Kaitao Luo, Ye Wang, Kenneth Holstein, and Nikolas Martelaro. 2025. Exploring the Potential of Metacognitive Support Agents for Human-AI Co-Creation. In *Designing Interactive Systems Conference (DIS '25), July 05–09, 2025, Funchal, Portugal.* ACM, New York, NY, USA, 26 pages. https://doi.org/10.1145/3715336.3735785

DIS '25, Funchal, Portugal

ACM ISBN 979-8-4007-1485-6/25/07
https://doi.org/10.1145/3715336.3735785

1 Introduction

Generative AI (GenAI) models offer increasing capabilities in supporting design workflows by generating images [84], videos [48], or complex mechanical parts [41, 102]. In mechanical design, working with AI allows designers to co-create designs that would be extremely tedious or even infeasible without AI support, such as reducing the weight of an electric wheelchair component [41] or generating parts using emerging manufacturing processes [102] with 3D geometric GenAI solvers. However, despite the growing promise of AI design tools to augment design processes, professionals often struggle to effectively integrate AI into their workflows [42, 107]. GenAI demands new (computational) workflows that require designers to work differently than they are used to or trained in [42, 69, 91, 101]. Current GenAI-supported workflows pose a set of unique cognitive challenges, including:

(1) **Intent Formulation**: Designers have to specify all design criteria necessary for generating feasible parts as distinct parameters upfront instead of iteratively modeling, testing, and visualizing a part's 3D geometry [42, 89]. This is a particular challenge for GenAI systems with lengthy and expensive inference times, such as complex 3D geometric solvers (e.g., [5]).
(2) **Problem Exploration**: To tackle design problems sufficiently, designers must thoroughly think through and consider many aspects, but GenAI workflow automation can reduce cognitive engagement and foster overreliance due to "cognitive offloading," making problem exploration more challenging [64, 106, 107].
(3) **Outcome Evaluation**: Designers are also required to evaluate generated designs according to the problem, but when their problem understanding is limited due to cognitive offloading, they won't be able to effectively evaluate and refine generated designs [91, 107].

Motivated by such cognitive challenges, we explore interaction patterns and interfaces to support professionals in more effectively working with AI-driven design tools. In this work, we follow an exploratory prototyping approach [105] to explore the potential

of voice-based agents that support designers' *metacognition* [40][1] while working on a manufacturing design task in a 3D GenAI CAD tool, where the designer specifies their goals as parameters and geometry within a graphical CAD interface. Broadly, we ask: *What interfaces and interaction patterns can support designers in better thinking through and formulating design problems, and evaluating generated outcomes, when working with GenAI-based design tools?*

Informed by theories and findings from human-AI interaction, learning sciences, and the study of design processes, we engaged in exploratory prototyping [105] to explore a design space for metacognitive support agents through three different design probes [14] and observe how each influences designers' processes and outcomes in a GenAI-based design task. In this prototyping process, we used the "Wizard of Oz" (WoZ) technique [29, 105], in which a human operator controlled the agent probes in a flexible manner but within certain probe-dependent constraints. Each probe followed a different support approach: (1) *SocratAIs* asks reflective questions to prompt deeper reflection-in-action and (2) *HephAIstus* prompts task planning and diagramming supported by suggestions for design strategies and software operation. While the first author enacted these two agents, we also included (3) external experts in mechanical and generative design from Autodesk to act as wizards in some sessions, who we invited to provide *their own* interpretations of metacognitive support strategies in a freeform manner.

Since CAD-based work is highly visual-spatial, we explore all support agents through voice modality to reduce cognitive load. Inspired by the concept of *"think-aloud computing"* [59], we prompt designers to verbalize their thoughts while working on the design task to foster deeper reflection-in-action and to elicit their knowledge and situational intentions for the support agents.

We conducted a formative study with 20 trained mechanical engineers new to working with generative AI systems. The designers were supported by one agent probe (or received no support in a control condition) while working on a realistic mechanical design task in the "Generative Design" extension of the commercial CAD software Autodesk Fusion 360 [5].

By comparing the design processes, outcome quality, and participant post-task interviews through video interaction and thematic analysis, we investigate the following research questions:

RQ1 *How do different agent support strategies impact the design process?*

RQ2 *What are the perceived benefits and challenges of metacognitive support agents?*

Overall, we found that agent-supported users created more feasible designs than unsupported users. Most users actively engaged with and appreciated the agent's support in helping them think through the design task and operate the software. We also identified that different agent strategies had different impacts. For example, question-asking strategies that prompted mental simulations or visualizations through sketching helped designers with intent formulation and problem exploration regarding the part's mechanical loads. However, we also observed that asking questions alone was less impactful when users had solidified incorrect assumptions, and that sometimes agent support could lead to additional overreliance. Finally, our findings provide insight into design trade-offs and differences in user preferences for metacognitive support agents.

We conclude by discussing design implications for future metacognitive support agents for GenAI-based design tools. While our paper explores support for mechanical design tasks, we discuss how our findings may generalize to other GenAI design activities. In sum, this paper makes three main contributions:

(1) Opening a design space for metacognitive think-aloud support agent interfaces for computational design tasks;
(2) Sharing empirical insights into how designers interact with metacognitive think-aloud support interfaces in the context of GenAI-based manufacturing design workflows;
(3) Proposing design considerations for future metacognitive support interfaces for GenAI-assisted design tasks.

2 Related Work

2.1 Challenges of AI-Assisted Design Workflows

Many GenAI design tools operate as black boxes—designers specify objectives and then examine one or more generated designs. This poses key barriers to iterative trial-and-error design workflows, especially for GenAI systems with lengthy and expensive inference times, such as geometric solvers (e.g., [5]). Therefore, research has explored the design of systems that facilitate faster, more interactive design exploration paired with computational design techniques [23, 31, 54, 56, 66, 104]. However, recent research has also identified several unique cognitive challenges professionals face when using AI-based design tools:

Intent formulation: GenAI tools demand designers to specify all design criteria required for generating feasible parts upfront, shifting focus to careful upfront planning of design requirements and formulating design intents in distinct parameters instead of iteratively modeling, testing, and visualizing a part's 3D geometry [42, 89]. This design process demands a shift in designers' attitudes, skills, and mental processes compared with traditional (CAD modeling) practices [69, 101].

Problem exploration: Design typically requires designers to think carefully through many different facets (explicit and implicit) to tackle design problems sufficiently. However, GenAI workflow automation can foster reduced cognitive involvement in the design process and overreliance due to "cognitive offloading" [81, 91], making it more challenging to explore and define design problems adequately [64, 106, 107]. For example, empirical research found that designers in geometric "traditional" modeling environments engaged more in semantic-level actions, leading to unexpected discoveries and diverse design solutions. In contrast, those in parametric environments followed a top-down process with fewer exploration and goal changes [18].

Outcome evaluation: Designers need to assess generated designs in relation to the design problem at hand. However, if their understanding of the problem is limited due to AI-imposed cognitive offloading, their capacity to effectively evaluate and refine the generated designs will also be constrained [91, 107].

Motivated by such challenges, recent research highlights the need to rethink parametric design tools and develop systems that

[1] *Metacognition* refers to mental processes of thinking about one's own thinking, enabling individuals to regulate and improve their cognitive strategies by reflecting on their decisions and problem-solving approaches.

better support designers in parametric modeling and computational thinking [98]. Similarly, other recent work has emphasized better support for the metacognitive challenges imposed by GenAI-based workflows [91]. *Metacognition* [40] involves reflecting on and regulating one's own thinking to improve decision-making and problem-solving strategies. For GenAI workflows, Tankelevitch et al. [91] highlight three critical phases that demand more explicit metacognitive support: (1) "prompting" GenAI (formulating inputs), (2) evaluating GenAI outputs, and (3) deciding on if and how to incorporate GenAI into one's workflow best.

Motivated and building atop prior work identifying (meta)cognitive challenges of GenAI-assisted design workflows, we explore novel support interfaces to help users work better with GenAI-assisted workflows. In the next sections, we will review metacognitive support strategies from learning science and design, and then highlight the role of asking questions in design and problem-solving as a distinct metacognitive support strategy.

2.2 Metacognitive (Design) Support Strategies

In the cognitive and learning sciences, metacognitive support has been shown to play a crucial role in enhancing problem-solving abilities by enabling individuals to reflect upon and actively regulate their own cognitive processes [45, 60]. *Self-regulated learning (SRL)* is closely tied to metacognition and involves studying and supporting learners' ability to manage and direct their learning processes through metacognitive skills like planning, monitoring, and evaluating their actions, typically occurring in distinct cyclical phases [10, 47, 73, 97]. Prior research has identified effective metacognitive support strategies such as "self-explanation", where prompting individuals to articulate their reasoning and underlying assumptions to themselves supports them in clarifying and organizing their own understanding [96]. This process, often in combination with think-aloud-style verbalizations of thoughts, promotes the integration of new information with prior knowledge, fostering critical thinking and cognitive engagement in a task [45, 46]. Research has studied metacognitive support strategies and interactive systems to promote reflection and problem-solving in various contexts, including software debugging [32, 57, 61, 75, 90], data analysis [36], learning computational skills [22] and exploratory learning [21].

Design research increasingly emphasizes the central role of metacognitive monitoring and control processes for design activities. For example, Ball and Christensen [9, p. 49] explicitly draw parallels between metacognitive processes and prior design theories, such as the role of *"reflection in and on action"* in design practice [35, 85, 86]. Furthermore, research has found evidence for the importance of metacognition for learning and mastering design skills [53, 62, 76, 78]. Building on this understanding of metacognition in design, the following section explores how questioning strategies can foster metacognitive engagement, critical thinking, and deeper cognitive exploration while supporting designers in tackling complex design challenges.

2.3 The Role of Asking Questions in Design and Problem-Solving

Questioning can support thinking and foster deeper cognitive engagement during problem-solving. In educational contexts, deep-level reasoning questions (e.g., questions probing underlying principles or causal relationships) or inquiry-based prompts (e.g., prompts encouraging student-led questioning and investigation) have been shown to enhance learning outcomes by stimulating critical thinking and deeper exploration of complex concepts [12, 27, 33, 44, 92]. A specific strategy is the *Socratic Method*, which employs guided open-ended questioning to stimulate critical thinking and reflection [38]. This approach has been used effectively across various fields, such as in healthcare education to develop critical thinking skills among students [49], programming to aid novice debuggers in identifying and resolving code issues [2, 3, 57, 100], supporting academic career development [74], and in creativity research to foster co-creativity between humans [88].

Similarly, asking questions plays a central role in guiding designers' thinking in exploring and refining ideas by challenging their assumptions as they work through complex, evolving problems [39]. When tackling ill-defined problems, designers must navigate the solution and problem spaces simultaneously, often using abductive reasoning to reframe problems and synthesize new insights and possibilities [28, 34, 58]. Research analyzing design team communication has shown the crucial role of question asking for problem-solving [4], idea generation [26], design reviews [19], and design studio education [51]. Other work has explored ways of supporting designers through targeted questioning to enhance the design process by helping them articulate product requirements [99], stimulate idea generation [83], and highlight awareness of bias in designerly thinking [79].

Eris developed a taxonomy of questions asked during design teamwork [39], building on prior taxonomies by Lehnert [65] and Graesser [44]. Eris's taxonomy outlines three types of questions for design: 1. *Low-level questions* for clarification, 2. *Deep reasoning questions* for causal explanations, and 3. *Generative design questions* for exploring alternative solutions. Research showed that student teams asking more Deep Reasoning and Generative Design questions achieved more innovative design outcomes [39].

2.4 Multimodal and Collaborative CAD Systems

Current computer-aided design (CAD) tools primarily rely on WIMP (Windows, Icons, Menus, Pointer) interfaces, using pointer movements and keystrokes as input [70] to modify geometry in visual-spatial interfaces. However, multimodal inputs for CAD work, such as gestures or speech in combination with WIMP [55], can offer advantages, as experimentally demonstrated by Ren et al. [80]. This combination—leveraging the split-attention [6] and the modality effect [20]—allows users to process visual and auditory information simultaneously, which can reduce cognitive load and enhance performance in complex tasks.

Similarly, user research studies often utilize the concurrent think-aloud protocol [63] to understand participants' thoughts and actions by encouraging them to speak about their thoughts as they perform a task. Concurrent verbalization can offer rich insights into users' knowledge and intents while only slightly increasing cognitive

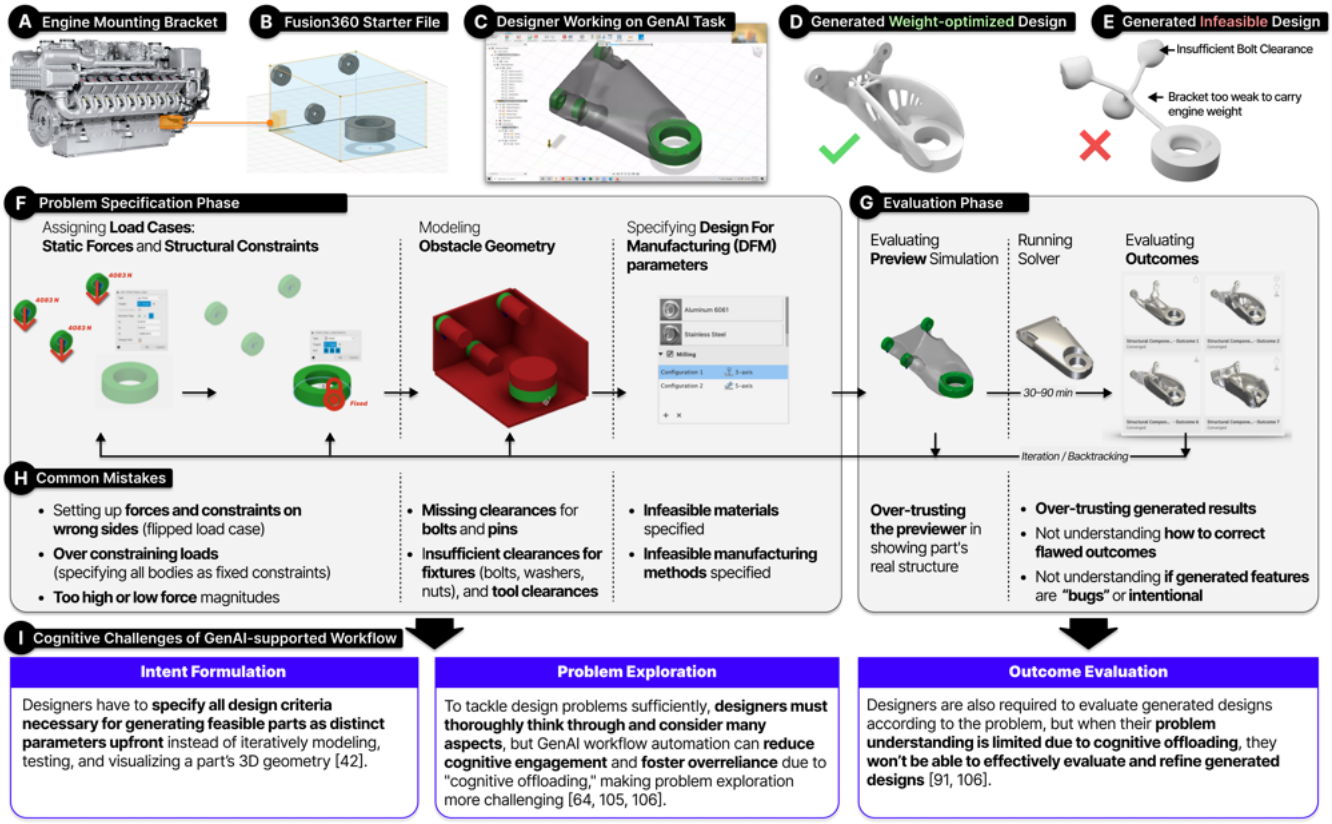

Figure 1: Overview of the Fusion 360 design task (A-E), workflow (F-G), common user mistakes (H), and cognitive challenges (I). The task involves (A) designing an engine bracket that connects the engine to a damper. (B) A starter file containing connection holes and bounding dimensions is provided to the users to initiate the design in (C) Fusion 360. The user is prompted to create (D) a viable design while minimizing weight and avoiding (E) infeasible features. (F–G) The workflow involves six steps, and based on the AI system's solutions, the user may iterate the design by adjusting the design constraints and criteria to produce new solutions. Task taken from [42]. (Image A: © Rolls-Royce Power Systems)

load during complex tasks [95], such as annotating existing CAD models through speech [77]. Inspired by the think-aloud protocol, Krosnick et al. [59] propose the interaction paradigm of *think-aloud computing* where computer users are encouraged to speak while working on a CAD design task to transcribe and capture rich knowledge with relatively low effort in real-time.

Other empirical research shows how conversation and real-time support during CAD sessions can improve problem-solving. For example, revealing unique communication patterns that make multi-disciplinary engineering design work more effective [82] or systems for supporting CAD users by connecting them with human CAD experts in real-time [25, 52] or automatically provide context-sensitive learning resources [67].

In this study, we aim to explore voice-based agent support interfaces and think-aloud user interactions for augmenting GenAI-driven CAD workflows by enabling low-effort continuous speech-based user intent and context-elicitation.

3 Case Study: Challenges of AI-Assisted Mechanical Design Tasks in Fusion360 Generative Design

Our work aims to develop support agents that help designers overcome the cognitive challenges they face when working with AI-assisted design tools. We replicate our previous study of mechanical designers working with AI assistance [42], which found that designers working with AI for the first time often failed to create feasible mechanical parts despite being familiar with the design tasks and CAD tools in general. In that study, designers worked with the "Generative Design" feature of Autodesk Fusion360 [5], which helps designers create lightweight and strong parts through topology optimization and genetic algorithms [66]. In the task (Figure 1A–E), the designer is asked to design a material-efficient and structurally sound engine mounting bracket by considering the optimal manufacturing and material combination from a large pool of possibilities. While designing mounting brackets is common for mechanical engineers, optimizing designs for different manufacturing methods and materials is difficult without simulation and AI

Table 1: Overview of agent probes' support strategies and behaviors.

	SocratAIs	HephAIstus	Expert Freeform
Main Strategies	**Asking Questions**	**Planning, Sketching, with Suggestions**	**Freeform: Determined by Invited Experts**
Support Principles	Asking questions to **prompt self-explanation and reflection** [45, 96].	Providing **planning and sketching support** [7, 42] while supporting these activities by **suggesting design strategies** and **workflows** [25, 52].	Determined by invited expert acting as wizard
Response to User Queries	Responding with questions only	Answering factual user questions	Determined by wizard
Modalities	Voice	Voice, link sharing, screen annotations	Voice, screen annotations
Timing and Frequency of Messages	Determined by wizard	Determined by wizard	Determined by wizard
Wizard Enacted by	Researcher (first author)	Researcher (first author)	Generative Design experts from Autodesk
Example	`"What might be reasons why the GenAI system generated the shapes this way?"`	`"Can you walk me through your load cases and constraints by sketching a free-body diagram? I shared a link to a board for you to sketch on in the chat."`	

support. Traditionally, engineers build a part and then gradually remove or add material based on structural analysis to derive a weight-optimized part. Exploring different manufacturing options is necessary for every material and manufacturing process—a time-consuming and tedious task. In contrast, Generative Design can automatically generate many options based on specified requirements, which the designer can explore and choose from.

Concretely, in Autodesk Generative Design, designers specify the *structural loads* a part has to hold, the GenAI solver's *obstacle geometry* (part areas that must remain free of material, such as clearances for bolt holes), and the *material and manufacturing properties* (Figure 1 F). Designers then optionally request a preview simulation before running the solver and then evaluate many AI-generated solutions to identify viable designs (Figure 1 G). If no outcomes are deemed satisfactory, designers might iterate the design by adjusting the input criteria.

In our prior study [42], we observed that **while most designers learned to specify some of these input parameters successfully over time, many failed to correctly specify structural loads and obstacle geometry for sufficient part clearances.** As a result, most of the submitted designs were unfeasible because they were either too heavy or weak, larger than the allowable, or had insufficient clearances around bolt holes, preventing the bracket from being mounted.

Figure 1 H lists common errors that occur during the task of designing a ship engine mounting bracket using Generative Design related to insufficiently **specified loads**, **obstacle geometry**, **materials and manufacturing options (DFM)**, and during **outcome evaluation**. These observed common mistakes relate to key cognitive GenAI workflow challenges of *intent formulation*, *problem exploration*, and *outcome evaluation* (see Figure 1 I and Section 2.1). In the following section, we describe how we probe different support strategies aimed at helping designers overcome these challenges.

4 Constructing Support Agent Design Probes

Our motivation was to prototype, study and compare different metacognitive support strategies in the context of GenAI-supported design tasks. Previous studies have shown that existing support resources and strategies (such as help menus, online forums, or video tutorials) seem to be ineffective in helping designers overcome the cognitive challenges involved in working with GenAI [42]. Thus, we speculated that metacognitive support strategies, such as asking reflective questions to prompt self-explanation or planning and sketching activities, might be more effective. Inspired by previous work and findings on metacognitive support, we therefore asked:

1) *What if a support agent simply asked questions? Could this in itself be enough to promote productive reflection-in-action and improve human-AI co-creation?*
2) *What if an agent prompted designers to plan and sketch while supporting these activities through suggestions?*

To explore these metacognitive support strategies prior to developing functional AI-based agent systems, we engaged in exploratory prototyping [105] and constructed two different agents as design probes [14]: (1) *SocratAIs*, a Socratic agent that asks reflective questions to prompt deeper reflection-in-action, and (2) *HephAIstus* that prompts task planning and diagramming supported by suggestions for design strategies and software operation (see Table 1). These two agent probes were enacted by the first author using an exploratory *"Wizard of Oz"* [29, 105] approach. The wizard followed guidelines to adhere to the general rules for each agent while also having flexibility over when to send messages and the

exact phrasing of messages given the in-the-moment context of each participant's session (see 5.4 for details).

We hypothesized that each of these support strategies could be effective in supporting designers on their own, but also likely in combination. However, as a start, we wanted to investigate how certain strategies would work in isolation to better understand their impact, benefits, and tradeoffs on designers' metacognition and design process.

In addition to these two probes, to move beyond our research team's assumptions, we also asked:

3) *How would human experts in generative design support designers new to working with GenAI in this task?*

To answer this question, we also invited (3) external experts from Autodesk to act as wizards during some sessions, to observe and compare their natural strategies of supporting other designers in this task (freeform). The following sections describe our three agent probes *SocratAIs*, *HephAIstus*, and *Expert-Freeform* in more detail.

4.1 *SocratAIs* Probe – Asking Questions

Inspired by previous research on prompting self-explanation [45, 96], *SocratAIs* proactively asks users questions as they complete a design task to **prompt self-explanation and reflection on their design decisions**. Specifically, this agent follows a **Socratic questioning approach** to support designers' metacognition by constructing questions relevant to the phase of a design task that a designer is currently working on, such as specifying the part's loads, obstacle geometry, manufacturing considerations, or evaluating outcomes (see Appendix for agent guide). In line with a Socratic questioning approach, *SocratAIs* **only responds to user requests with further questions** and refuses to provide direct answers.

4.2 *HephAIstus* Probe – Planning, Sketching, with Suggestions

To explore our second question (*what if an agent prompted designers to plan and sketch while supporting these activities through suggestions?*), we constructed *HephAIstus* (*referencing Hephaestus, the Greek god of craftsmanship).* This agent provides metacognitive support in the form of **planning and sketching** and by supporting these activities with suggestions around design strategies and tool operation. Inspired by prior research on the benefits of externalization activities in design [7, 42], the agent offers deliberate planning and sketching activities parallel to the CAD workspace to help users think through the design problem more strategically and visually. For planning, the agent suggests the user engage in a **project planning activity** by sharing a pre-generated text document outlining critical project-relevant aspects with the user (see Appendix A.1.4 for an example document). This strategy aims to encourage users to think through the design task more deeply before switching to the CAD interface.

For the sketching activity, the agent suggests that the designer **sketch out load case-relevant forces and constraints as a *free-body diagram***[2] by sharing a link to a 2D drawing canvas containing the side and top view of the bracket as a starting point (see Appendix Figure 8 for an example).

To support these planning and sketching activities, the agent proactively offers suggestions for the design task and software operation, inspired by work on supporting software learning and work processes [25, 52]. This entails providing alternative design options, highlighting overlooked software features, notifying about unintentional execution errors, or recommending tools and techniques to improve the overall design process.

In contrast to *SocratAIs'* question-asking approach, *HephAIstus* responds to user queries with direct answers, similar to chat assistants such as ChatGPT. Lastly, the agent can visually highlight areas on the user's screen to direct their attention to what the agent is talking about.

4.3 *Expert-Freeform* – Support from an External Generative Design Expert

Lastly, we explored how human experts attempt to support designers' metacognition in this task, when provided agency over how to do so. We invited experts in mechanical and generative design from Autodesk (the maker of Fusion360)—who were not a part of our research team—to serve as wizards, allowing them to provide their own interpretation of metacognitive support (freeform). For these sessions, we recruited wizards from Autodesk's employee pool via internal mailing lists and snowball sampling (see Appendix Table 5). Participants ranged between 27 and 47 years of age, with mechanical design experience between 3 and 15+ years. All experts had high self-rated proficiency in Fusion360 Generative Design, were closely involved with its development, and had substantial experience training others to use the tool.

The expert wizards were instructed to support the other designer in working with Generative Design and the design task by controlling the voice agent. We refrained from explicitly telling them to follow a specific support strategy, and instead, they were asked to provide real-time support in their preferred way, so long as they did not directly instruct the other designer on what to do.

4.4 Common Agent Capabilities

Besides the differing support strategies outlined above, all support agent probes shared the common capabilities:

- the agent possesses (non-exhaustive) knowledge of additive manufacturing and generative design tasks
- the agent has access to the users' screen and think-aloud speech in real-time;
- the agent can identify inconsistencies between the requirements stated in the design brief and the GenAI parameters specified by the designer by comparing the design brief and screen activities (e.g., detecting over/under-constrained load cases, infeasible material combinations, or wrong force setup)[3];
- the agent can send voice messages to the user and (in *HephAIstus* and *Expert-Freeform* cases) annotate the user's screen and share links via chat.

[2]Free body diagrams are common mechanical visual representations to illustrate the forces acting on physical objects in a given situation, helping to simplify complex mechanical problems and reason about its structure.

[3]Such requirements could be explicit nature (e.g., the force the bracket needs to hold) or implicit features, such as bolt clearances, which were not explicitly mentioned in the design brief

5 Study Design

To elicit the impacts, potential benefits, and drawbacks of the support agent probes, we conducted a formative between-subjects study with trained mechanical designers new to working with generative AI. Each designer was supported by a different agent probe (facilitated by a human operator in the background) while working with the Autodesk Generative Design tool to design a ship engine mount [4].

We used an exploratory "Wizard of Oz" (WoZ) prototyping approach [105], where a human operator controlled the voice agent probes in the background to simulate different support strategies. Instead of only following strict predefined rules, the wizards had certain degrees of freedom in enacting the agent probes' support strategies to explore broader design possibilities and implications in response to emerging situations during user sessions [105] (see 5.4).

While working on the task, designers were asked to think aloud to elicit their cognitive processes (e.g., mental models [24], learning [103]) and knowledge and intents so that they could be used by the (WoZ-controlled) voice-based support agents. Participants worked between 31 and 99 minutes, then submitted their designs and completed a semi-structured interview to reflect on their experience working with the support agent.

We collected the following data:

- Video, screen, and audio recordings with machine-generated transcripts of the agent-supported think-aloud design sessions
- Audio recordings and machine-generated transcripts of the post-task interviews
- 3D designs created during the think-aloud sessions
- Log files with timestamps of all human-facilitated agent messages

5.1 Participants

We recruited 20 designers (aged 20 to 42 *(M = 26.1, SD = 5.9))* with mechanical engineering backgrounds from engineering departments of North American universities and through the Upwork freelance hiring platform[5] (see Appendix Table 4). Participants had between one and ten years of Mechanical Design experience and between zero to ten years of industry experience, as determined via a screening questionnaire. All participants had at least two years of experience using CAD and Autodesk Fusion360 but no experience working with the Generative Design extension. We recruited participants familiar with Fusion360 so that they could focus on learning to work with the AI-driven Generative Design feature rather than learning the CAD tool's user interface. Participants included a mix of undergraduates, graduate students, and professional engineers. Before the study, all participants signed a consent form approved by our institution's IRB. Participants were compensated 20 USD per hour.

[4]See Section 3 for a description of this task, previously used in [42].
[5]http://www.upwork.com

5.2 Design Task and System

Participants were instructed to design a light and strong engine mounting bracket with Autodesk Fusion360's [5] "Generative Design" feature (see Section 3 and Figure 1 for a detailed description). Since we adopted the task from an existing study [42], we verified the suitability of the task for our study by first piloting it with mechanical engineers from our institution and an external user proficient with Generative Design, all of whom successfully completed the task without receiving any support.

5.3 Procedure

The study was structured into four phases, split into two sessions:

1) Onboarding (30 minutes): After an introduction to the study, participants received a hands-on tutorial demonstrating Fusion360 Generative Design's core functionalities through a step-by-step example design task.

2) Design Task - Part 1 (up to 70 minutes): After onboarding, participants were introduced to the design brief, task, and starter file containing predefined geometric constraints. They were also told that a virtual AI agent would support them during the design task (except for the members of the No Support group). Sessions were conducted over video conference (Zoom) with audio, screen, and video recording.

Participants worked while sharing their screens and thinking aloud, with research team members following the video call remotely and operating the support agent. Participants were allowed to use any available support resources, such as internal Autodesk help files, external video tutorials, or online user forums.

Participants worked until they completed specifying the generative design inputs. They then started the Generative Design extension's solver, which completed the first session. Since the solver required 30 minutes of runtime, participants took a 30-minute break and then returned for the second session.

3) Design Task - Part 2 (up to 30 minutes): After the solver finished, participants evaluated the generated designs. If satisfied with the results, they could directly select three designs. Otherwise, they could re-adjust the design criteria and restart the solver, in which case they would return to evaluating the generated designs after the exit interview and select their final designs.

4) Exit Interview and Debriefing (20 min): After task completion, participants participated in a semi-structured remote interview with a research team member. Participants were asked to reflect on their experience working with the Generative Design extension, the think-aloud activity, and the agent support (see the Appendix for interview protocol). Additionally, after the expert-facilitated sessions, we interviewed the experts to gain further insight into their support strategies and challenges they perceived. The interview was audio and video recorded and the interviewer took notes. At the end of the inteview, participants were debriefed about how humans had actually controlled the AI support agents.

5.4 Wizard of Oz Setup

Overall, we followed an exploratory Wizard of Oz prototyping approach [105] as a design space exploration where wizards would have some flexibility in enacting the agent probes. This allowed

Figure 2: Process diagram of Wizard of Oz setup. The remotely located wizard (right) followed the designer's actions (left) by listening to their verbalizations and observing their screen and webcam stream. Using a web interface, the wizard could (1) type messages and send these as (2) synthesized voice messages to the user as agent messages. (3) All agent messages were logged with timestamps.

us to make meaningful comparisons between the support strategies (*asking questions* vs. *planning and sketching support* vs. *expert freeform*) while also giving wizards flexibility on how to enact the different agent probes in detail (such as the exact message timing and phrasing). Below, we detail the instructions given to wizards and the study setting.

5.4.1 Wizard Details. The *SocratAIs*, and *HephAIstus* agents were facilitated by the first author with experience in mechanical engineering, Fusion360, and generative gesign [6].

This wizard followed these general guidelines:

1) Follow the designer's verbalizations and screen actions and pay close attention to the task-specific design steps and challenges as outlined in Section 3.
2) Pay close attention to inconsistencies between the requirements stated in the design brief and the input parameters set by the designer[7].
3) Never directly tell the participant what to do, but rather provide supportive questions, hints, or suggestions (depending on the enacted agent type).
4) You are free to send messages whenever and how often you consider it helpful to the designer. However, pay special attention to moments in which designers transition between design sub-tasks (such as from specifying obstacle geometry to specifying loads), as well as when designers show hesitation or use hedging expressions.
5) You are free to formulate the messages in a way you consider to be most helpful, while adhering to the agent's support strategy (e.g., only asking questions).

For the *Expert-Freeform* agent wizards, we did not provide specific guidelines, but only instructed them to provide support in their preferred way, so long as they did not directly instruct the supported designer on what to do (see Section 4.3).

5.4.2 Setting. For all sessions, participants and wizards were in separate locations during the design task, and communication between the wizards and participants was established via Zoom video conferencing software (see Figure 2). Although most sessions were co-located, with participants and wizards in separate but neighboring rooms of our research lab, eight sessions were conducted remotely. In the lab sessions, participants completed the task on a computer workstation running Fusion360 with the Generative Design extension. Remote participants were provided access to a web-based computer[8] with the same setup for remote sessions.

Participants shared their screens via Zoom and wore an audio headset during the task to capture their verbalizations and ensure they could hear the agent's voice. The wizard joined the same video call using a generic name ('Agent') with a deactivated webcam to follow the participant's screen actions and verbalizations. In addition, the wizard could generate and send agent voice messages using a self-developed web control interface (see Figure 2 right). For the *Expert-Freeform* sessions, in addition to the researcher and the designer, the external task expert from Autodesk anonymously attended the conference call in the background and remote-controlled the voice agent via the web interface.

[6] In some sessions, a second research team member with experience in mechanical engineering and generative design was co-present, verbally supporting the wizard.
[7] Such requirements could be explicit (e.g., the force the bracket needs to hold) or implicit features, such as bolt clearances, which were not explicitly mentioned in the design brief.
[8] using Paperspace

The agent control interface was developed in React.js and uses Google's text-to-speech API to generate the agent voice from the wizard-typed text (see Figure 2). The interface features a button to toggle the playback of an idle sound cue to sonically indicate an 'agent is processing' state to the participant. Additionally, the tool logs all generated messages with a timestamp and session ID exportable in JSON format. We used audio-routing software[9] on the wizard's computer to inject the generated agent speech audio into the video call. To mitigate possible gender bias effects, we deliberately selected a gender-ambiguous voice option for the agent, following suggested best practices from prior research [93].

5.5 Measures and Analysis

To gain insight into our research questions, **RQ1** *How do different support strategies impact the design process?* and **RQ2** *What are the perceived benefits and challenges of metacognitive support agents?*, we evaluated the design outcomes and analyzed ~19 hours of think-aloud videos and ~6 hours of interview recordings using a combination of video interaction analysis and reflexive thematic analysis.

5.5.1 ***Design outcome evaluation.*** We evaluated the design outcome feasibility by checking the submitted engine brackets against the requirements in the design brief, rating across five criteria, each yielding one point:

(1) The structural soundness was validated using finite element analysis (FEA).
(2) The feasible load case setup was checked in their Fusion360 project file.
(3) The optimized mass was not extremely light or heavy.
(4) The part had feasible fastener clearance (i.e., clear bolt holes)
(5) The part's mass and volume fit within the acceptable bounding dimensions.

The GenAI solver generated around 20 designs, and to compensate for possible variability in the generated outcomes, we asked participants to choose three feasible parts from which we then selected the highest-scoring part as their final design.

5.5.2 ***Video interaction analysis to determine agent impact.*** We used *video interaction analysis* [11] of the think-aloud recordings to understand how agent support impacted participants' design process. Specifically, to determine the impact of the agents' messages on the design process, we analyzed whether participants considered new design aspects after receiving a message based on their verbal reflections or concrete actions. The think-aloud video and transcript data were equally distributed among three researchers who applied the following coding procedure:

1) Tracking GenAI input specifications: First, the coders tracked participants' interactions with the Generative Design features relevant to the design task and documented whether the actions would produce satisfactory outcomes. Specifically, they tracked how participants specified *(1) structural loads* (forces), *(2) mechanical constraints*, and the obstacle geometry feature to control the bracket's *(3) bolt* and *(4) dampener pin clearances*, and *(5) overall size*.

[9]https://existential.audio/blackhole/

2) Coding message impact: Second, the coders tracked the impact of the agent's messages on the design process: For each agent message, they coded if the message had an observable impact on the participant considering a new aspect related to the design task, which needed to be apparent from the designer's verbalizations or actions *(coded with 'none,' 'weak,' or 'strong')*. For the design assistant agent probe (HephAIstus), they also tracked users' direct messages to the agent and if agent messages were generally observably helpful to the user *(yes/no)*.

3) Coding planning and sketching interactions: Third, the coders tracked when the agent sent the planning sheet or the free-body diagram sketching board and when the designers interacted with these.

Between these coding sessions, the researchers met frequently to discuss edge cases and ensure consistency in their coding practices. From this data, we then created time-series event plots for each session with `R` and `ggplot2` to visually identify patterns (see Appendix Figure 7). In addition, we created *summary videos* for each participant, highlighting all situations featuring agent messages or other interesting designer-agent interactions (please see the video figure in the supplementary material for an example).

5.5.3 ***Reflexive thematic analysis.*** To understand participants' attitudes toward the agent support, we performed a *reflexive thematic analysis* [16] of the interview data (transcripts). We followed an iterative inductive coding process and generated themes through affinity diagramming. We used ATLAS.ti to analyze transcripts, audio, video, and Miro for affinity diagramming.

First, the first author coded the interview transcript data utilizing both a *semantic* (what people said) and *latent* (our interpretations of the data) coding strategy. Next, the research team collectively identified initial codes and themes. Based on the time series plots from the video interaction analysis and the summary videos, we then associated the participant statements from the interviews with specific situations in the design sessions to cross-validate the impact of agent messages and identify additional qualitative themes and interaction patterns. We iteratively reviewed and revised codes and themes until we identified a stable network of coherent and rich themes.

6 Findings

All participants completed the task without abandoning it. Overall, most agent-supported designers overcame more GenAI workflow challenges and produced more feasible designs than unsupported designers (Table 2). However, different agent strategies impacted the design process in different ways. Most designers saw benefits in agent support, but we also elicited various trade-offs and differing preferences for support interactions. In the following sections, we present our findings on the impact of different support agent strategies on the design process, along with the perceived benefits and challenges associated with these.

6.1 Impact of different support strategies on design process (RQ1)

6.1.1 Design Outcome Comparison. Overall, participants with support produced notably higher-quality parts (see Table 2 and Figure 3 and 4), with an average outcome score of $M = 3.5$ ($SD = 1.4$),

Table 2: Table summarizing participants' outcome design scores across five criteria (checkmarks) and process statistics by support group. Normalized Message Frequency represents the number of agent-initiated messages divided by the session duration. Note that the number of agent-initiated messages is lower than the total agent messages for the HephAIstus and Expert-Freeform groups since these exclude agent messages in response to user-initiated queries.

Support Condition	**SocratAIs**						**HephAIstus**						**Expert-Freeform**						**No Support**					
Participant	**S1**	**S2**	**S3**	**S4**	**S5**		**H1**	**H2**	**H3**	**H4**	**H5**		**E1**	**E2**	**E3**	**E4**	**E5**		**B1**	**B2**	**B3**	**B4**	**B5**	
Passing Structural Analysis (FEA)		✓	✓	✓	✓			✓	✓	✓			✓	✓		✓	✓								
Correct Load Setup		✓		✓	✓				✓	✓			✓			✓	✓							
Mass Optimized		✓	✓	✓	✓				✓	✓			✓			✓	✓							
Feasible Fastener Clearances	✓	✓		✓	✓		✓	✓		✓	✓		✓	✓	✓	✓	✓			✓	✓			
Feasible Part Size	✓	✓	✓	✓	✓	*AVG*	✓	✓	✓	✓	✓	*AVG*	✓	✓				*AVG*	✓			✓	✓	*AVG*
Outcome Quality Score	2	5	3	5	5	**4.0**	2	3	4	5	2	**3.2**	5	3	1	4	4	**3.4**	1	1	1	1	1	**1**
# Agent Messages	38	27	21	17	15	**23.6**	38	64	21	40	32	**39**	27	29	21	36	52	**33**						
# User Messages							15	39	8	8	12	**14.4**	2	8	8	7	12	**7.4**						
# Triggered New Considerations	8	13	2	2	5	**6.0**	4	5	0	4	3	**3.2**	4	8	1	2	10	**5.0**						
Duration (min)	74	74	32	42	56	**55.5**	92	89	59	99	70	**81.7**	37	47	31	43	55	**42.5**	53	37	63	83	59	**59.0**
Normalized Message Frequency	0.5	0.4	0.7	0.4	0.3	**0.4**	0.2	0.2	0.2	0.2	0.2	**0.2**	0.6	0.3	0.5	0.5	0.6	**0.5**						

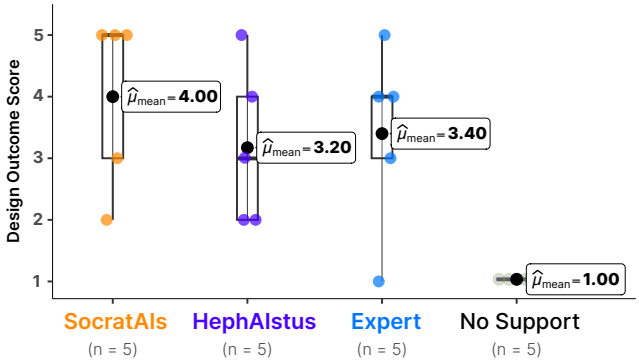

Figure 3: Plot showing the design outcome scores between agent-supported groups and no support.

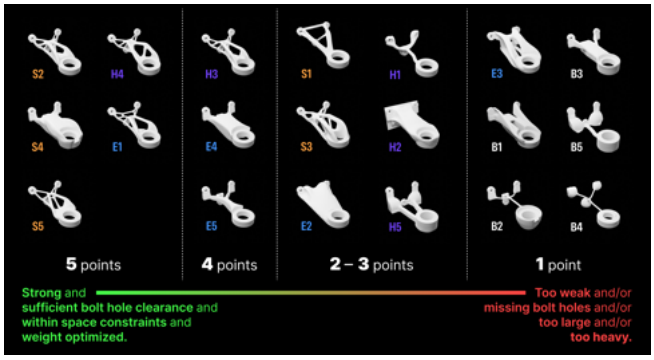

Figure 4: Overview of engine bracket designs created by participants grouped by quality score (1–5). White IDs indicate participants from the *No Support* group (all one point).

compared to participants with no support who had an average outcomes score of $M = 1.0$ ($SD = 0.0$). All participants in the *No Support* group incorrectly specified the bracket's load case, and consequently, no final design passed the structural analysis (see designs with white labels in Figure 4). Additionally, brackets created in the *No Support* group had inaccurately specified obstacle geometry, resulting in infeasible fastener clearances or material exceeding the required package size. These low-quality outcomes match our prior study's results in the same unsupported task [42].

In contrast, while the outcome quality varied within and across the agent-supported groups, the majority of supported designers created brackets that passed the structural analysis and had feasible fastener clearances while staying within the required space limitations. Between supported groups, the average design score varied slightly, with *SocratAIs*-supported users having the highest number of designs fulfilling the load and spatial requirements. While the small sample size per group prevents us from drawing conclusions about statistically significant differences between the support conditions, the consistent gap between supported and unsupported users points to clear benefits of having agent support.

6.1.2 Comparison of Agent Message Frequency and Impact (across conditions). In terms of **number of messages**, agents sent between 15 and 64 messages per session, with the highest group average being $M = 39 (SD = 15.8)$ in the *HephAIstus* group, $M = 33 (SD = 11.9)$ in the *Expert-Freeform* group, and $M = 23.6 (SD = 9.3)$ in the *SocratAIs* group (see Table 2). However, these counts of the number of messages sent by *HephAIstus* and *Expert-Freeform* also include responses to user-initiated queries and therefore are naturally higher than for the *SocratAIs* group. To gain a **normalized comparison of message frequency between conditions**, we calculated the number of messages initiated by the agents divided by the session duration (agent-initiated messages/session duration), which revealed a similar message frequency per minute across agent

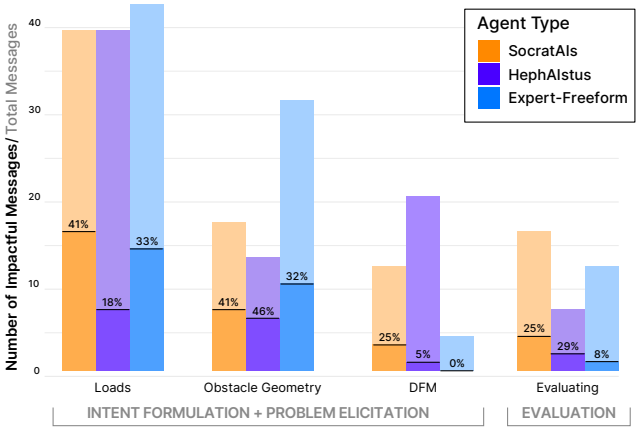

Figure 5: Plot illustrating the total number of messages per support topic category and agent type (unsaturated colored bars) and the percentage of observable impactful messages that triggered observable new considerations (saturated areas).

groups of $M = 0.4(SD = 0.15)$ for *SocratAIs*, $M = 0.2(SD = 0.03)$ for *HephAIstus*, and $M = 0.5(SD = 0.14)$ for *Expert-Freeform*.

Regarding the messages' impact on the design process, the number of messages that triggered observable new considerations (i.e., impactful messages, as defined in 5.5.2) ranged from zero to 13, with means ranging between $M = 6.0(SD = 4.6)$ *SocratAIs*, $M = 5.0(SD = 3.9)$ *Expert-Freeform*, and $M = 3.2(SD = 1.9)$ *HephAIstus*. Interestingly, although the *HephAIstus* group had fewer messages that triggered observable new considerations compared to the other two groups, their final design outcomes were comparable. This suggests that the planning and sketching activities prompted by the *HephAIstus* agent may have supported productive design reasoning, even when fewer individual messages were coded as impactful.

Analyzing the messages' topics across all agent groups, most agent messages concerned *intent formulation* and *problem exploration* in the *Loads*, *Obstacle Geometry* and *DFM* categories followed by messages supporting *Evaluation* (see Figure 5). Further analyzing the messages' topics in terms of their impact on considering new design-relevant aspects, the highest number of impacts had messages supporting intent formulation and problem exploration: *Obstacle Geometry* (between 32% to 46% across groups) and *Loads* (41% *SocratAIs* and 33% *Expert-Freeform*).

In terms of the **contrast between the number of messages and impact on design considerations**, agents varied drastically between the support aim topics: *HephAIstus' Loads* and *DFM* categories had only half or a third of the impact (18% and 5%) as *SocratAIs* while having a similar or larger number of total messages. Similarly, the *Expert-Freeform's* messages supporting *Evaluating* had only half of the impact (8%) as the other support agents, while having a similar number of total messages. In contrast, for the *Obstacle Geometry* and *Evaluating* categories, messages in the *HephAIstus* group had the highest consideration impact while having the lowest number of messages compared to the other groups (46% and 29%).

6.1.3 *SocratAIs' Effects on the Design Process.*

A) User-Agent Interaction Dynamics: Overall, participants paused their think-aloud verbalizations when listening to agent messages (118/118 questions). However, depending on the user's situation, they responded differently: immediately responding to the agent's questions by giving an answer (61/118 questions); finishing their line of thought and sub-task before replying (28/118); pausing to think and reflect silently before verbally replying (13/118); or directly responded with simple acknowledgments after thinking for a while in silence, such as *"Yeah, you are right,"* or *"That's a good point,"* even if the message was not a direct suggestion but an open-ended question (11/118); or providing no response (5/118). Some participants (2/5) asked the agent a question at the beginning of the session, but they stopped asking the agent more questions afterward, recognizing that it would not provide an answer but reply to user requests only with questions.

B) Agent Impact on Overcoming GenAI-Related Challenges: *SocratAIs* had mixed impacts on helping designers overcome design challenges across participant sessions (see example timelines in Figure 6 and Appendix Figure 7 for all sessions). In some sessions (2/5), agent messages had an observable strong impact on helping designers overcome design challenges (S2, S5). Meanwhile, in other sessions (2/5), we could only observe weak evidence of impact, where some agent messages had a positive impact, but overall, designers were unable to overcome most major challenges (S1, S3). In one session (S4), the designer created feasible outcomes without facing major challenges or showing observable agent impact, but *SocratAIs* helped them to consider additional design-task-related factors.

C) *SocratAIs*' Positive Effects:
In most sessions (S2, S3, S4, S5), **reflective agent questions helped users with more precise intent formulation and problem exploration.** For example, an agent question probing reflection on potential additional shape requirements of the part helped S5 think through the part's clearance requirements and consider and specify additional important details, leading to a feasible bracket design:

> SocratAIs: When specifying the bolt clearances, how do these impact the assembly and servicing of the bracket?
> S5: [Looks at the preview simulation] So this part won't be serviceable [...] you'd need enough clearance for the socket. So I will go back to edit model..." [user adds more obstacle geometry].

(see also timeline S5 in Figure 6). Later, S5 reflected on the helpfulness of the agent's message: *"It was asking something about how the obstacle geometry affects the serviceability of the part. That was basically telling me that I needed to leave some clearance for tools and for maintenance. That was very helpful"* (S5).

We also noticed in several cases that **questions were especially effective** in helping designers formulate intent and specify

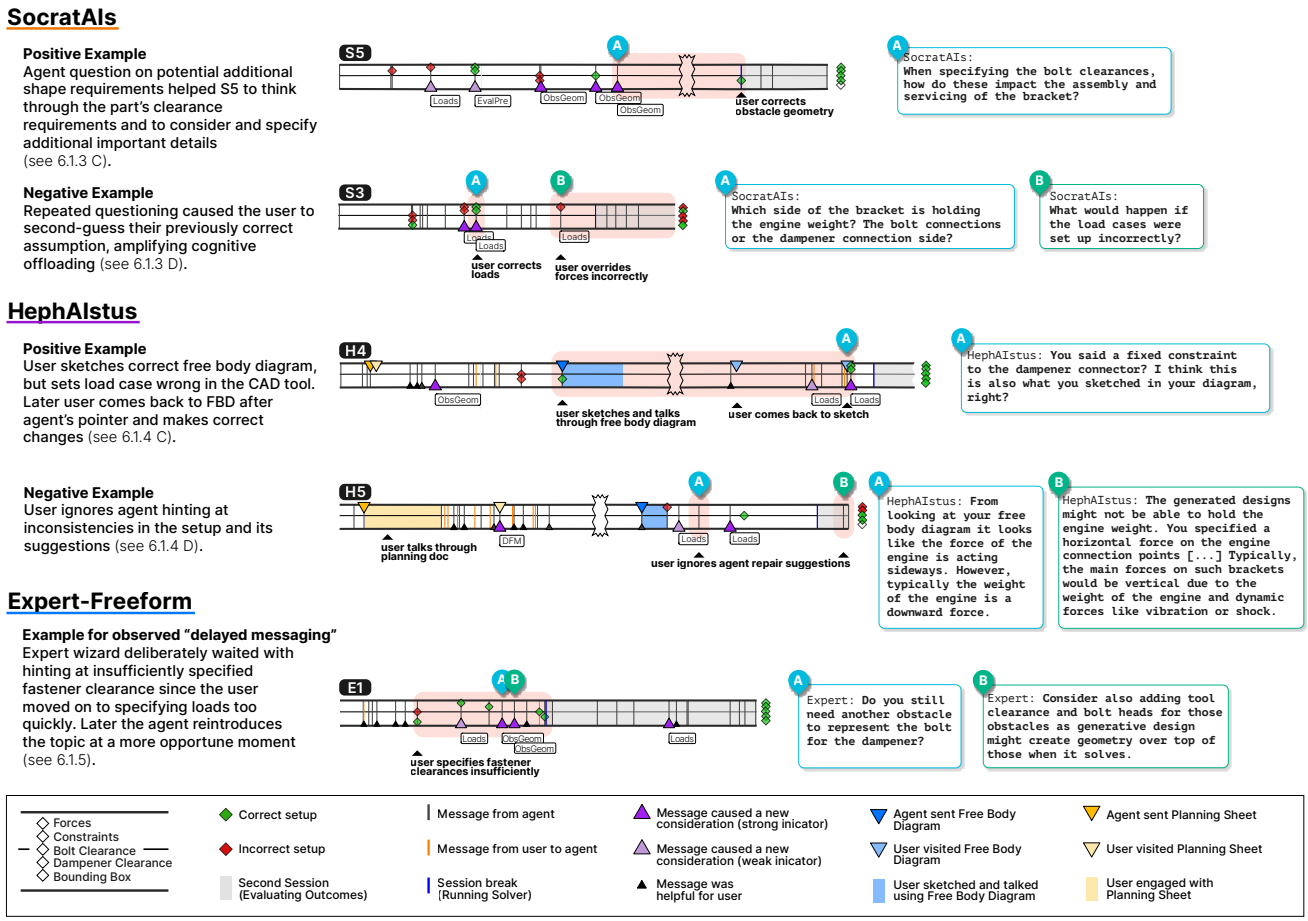

Figure 6: Timeline excerpts visualizing participant and agent interactions throughout the design task; timelines are divided into lanes, each showing (in)correct GenAI input specifications (diamond shapes) for (1) forces, (2) constraints, (3) bolt and, (4) dampener clearances, (5) bounding box (from top to bottom); black and orange vertical lines represent exchanged agent and user messages with purple and black triangles indicating an observable impact on the design process.

the design problem more accurately **when prompting users to mentally simulate the real-world aspects of the bracket**. For example, S3 had mistakenly modeled the load case reversed to the real-world situation ("flipped" load case), causing the solver to build a structurally unstable bracket. An agent question prompting them to reflect on the part's function from a real-world perspective (`Which side of the bracket is holding the engine weight? The bolt connections or the dampener connection side?`) helped the user update their mental model and load case.

In two sessions (S2, S5), **reflective questions during the preview and outcome evaluation phases helped designers better evaluate and correct faulty designs.** For example, S2 generated brackets based on incorrectly specified loads (forces and constraints assigned to the wrong sides). While evaluating the design, the user noticed the structurally weak parts but was unsure about the cause. An agent question probing deeper thought about the GenAI solver's mechanism caused the designer to realize their flawed mental model and correct the load specification:

`SocratAIs: What might be the reasons why the solver generated the shapes this way?`
`S2: [user thinks] Yeah, because it did not consider building material between these three [bolt connections] ... maybe because of the forces... [user checks forces again] It should be acting downward based on the weight of the engine [user corrects load case].`

For confident users, questions helped in problem exploration by considering additional design factors.

In one session, S4 did not encounter major challenges, and while the agents also asked similar questions as in the other sessions, these also had no observable impact on supporting critical cognitive challenges. However, agent questions (such as "`Considering the engine's environment, what alternative or additional assumptions might we make about the load cases?`") helped

the user to consider additional relevant factors, such as suitable materials for the bracket's maritime environment or additional forces resulting from ship movements. Later, S4 reflected in the interview: *"I think it was pretty useful [...] it prompted me about the assumptions about the loading. And I was like: 'oh yeah, this is on a ship, so it has to survive lateral loads and not just ground loads'." (S4)*

D) *SocratAIs*' Negative Effects:
In one session (S3), **repeated questioning amplified cognitive offloading, leading to flawed results.** At first, *SocratAIs*' questions during the initial setup phase helped S3 correct their overconstrained load case. However, a later question (`What would happen if the load cases were set up incorrectly?`) made them doubt their initially correctly specified load case setting and change it for the worse (see S3 in Figure 6).

In another case, **questions were less impactful in correcting a user's solidified wrong assumptions**. S1 had incorrectly set up a load case with fixed constraints and forces assigned to the same geometry, which would cancel out the impact of the force on the bracket's structure in the solver. An agent question (`Can you walk me through your intention of assigning a force and a fixed constraint to the same geometry?`) caused S1 to provide an explanation of their reasoning, eliciting their incorrect assumptions. But, instead of realizing misconceptions, their explanations reinforced their assumptions, preventing them from correcting issues.

6.1.4 *HephAIstus' Effects on the Design Process.*

A) User-Agent Interaction Dynamics:
Besides proactive support, *HephAIstus* also responded to user requests like a voice-based chat assistant such as Alexa or Siri, and users addressed the agent between 8 to 39 (M=14.4) times per session (see Table 2 and orange vertical lines in Figures 6 and 7). These user-initiated requests included asking the agent about manufacturing-related facts, such as material properties, or asking for help with load case-related tasks, such as calculating forces. Some users requested confirmation or feedback from the agent on the design process (*"Am I missing anything, agent?"*) or requested guidance (*"Ok, what's next?"*). Overall, these sessions were characterized by phases of active back-and-forth conversations between the user and agent, as apparent from the clusters of dense orange and black vertical lines in the event timelines in Figures 6 and 7. The proactive messages initiated by the agent included suggestions regarding Fusion360 operation (such as specific tools within the Generative Design extension) and overcoming design challenges (such as reminding users to use the preview or pointing out mismatches between the design brief and their setup).

B) Agent Impact on Overcoming GenAI-Related Challenges:
Overall, we found that ***HephAIstus* had mixed impacts on helping designers overcome design challenges across participant sessions.** In 1/5 sessions, the agent messages had an observable strong impact on helping designers overcome design challenges (H4), while in 3/5 sessions, weak impacts on overcoming design challenges were observable (H1, H2, H5). In 1/5 sessions, a designer created almost feasible outcomes without facing major challenges or observable agent impact (H3), but it was clearly observable that the agent helped the user operate the software more effectively.

C) *HephAIstus*' Positive Effects:
The agent-provided sketching board helped some designers in loads-related problem specification and intent formulation. When the agent prompted designers to explain the bracket's load case by sketching a free-body diagram (FBD) and sharing a link to a prepared drawing board, all users followed the link and used the board to sketch out diagrams while verbalizing their thoughts (see blue triangles and highlighted passages in Figure 6 and 7 and example board in Appendix). In several cases, the sketched free-body diagram served as a conversational anchor and reference point between the user and the agent. For example, while H4 had first sketched an FBD with feasible load cases, they then incorrectly specified the load case in Fusion360. Later, the agent pointed out the inconsistency between the sketched FBD and the load case setup in the CAD tool, which led the designer to correct the input specification (see H4 in Figure 6).

All designers explored the agent's project planning sheet, but few revisited it during the session. While 2/5 users quickly went through the document at the beginning, 3/5 users (H2, H3, H5) spent between 3–10 minutes in the document, utilizing its provided structure to talk through, reflect on and plan the design process step by step before starting to work in Fusion360 (see yellow triangles and highlighted passages in Figure 6 and 7). For example, the planning sheet supported H5 in reflecting on and exploring suitable material options while using the document to add notes about different material characteristics (see H5 in Figure 6). However, only two users revisited the document later in the session (H4, H5).

We also observed that in many cases, **proactive agent suggestions reminded designers about overlooked steps, unintentional execution errors (slips), or software features.** For example, the agent reminded H2 to run a preview simulation to better assess bolt clearances before starting the solver (`You might run a preview simulation at a later point to evaluate the clearances before starting the solver`). This message led H2 to run a preview and realize insufficient bolt clearance.

D) *HephAIstus*' Negative Effects:
While 52% of agent messages had an observable impact on helping the user work on the task and operating the software (101/196 messages, see black triangles in Figure 6), only a fraction (0.06%) directly helped overcome cognitive design challenges by considering new design-task relevant aspects (16/196 messages, purple triangles in Figure 7). We also observed that **directly pointing out inconsistencies in the users' setup only helped some users correct existing issues.** For example, in two cases (H1, H5), designers repeatedly failed to correct ill-defined load case setups despite the agent directly pointing these out and providing concrete suggestions for correcting them (see H5 in Figure 6). In those situations, participants decided not to follow the agent's suggestions and instead followed their own (partially incorrect) intuition.

6.1.5 *Expert-Freefrom Observed Support Strategies.*

From analyzing the session videos and post-task interviews with the expert facilitators, we identified several support strategies that helped designers similar to our support agent probes.

We observed that **experts frequently supported designerly thinking and metacognition while also highlighting overlooked design issues to guide users**. Similar to *HephAIstus*, experts proactively highlighted potential issues, such as missing or misrepresented GenAI parameters, to ensure critical considerations were addressed early.

Additionally, we observed that **some experts used a question-asking strategy** to support users in intent formulation, problem exploration, and outcome evaluation similar to *SocratAIs*.

Some **experts also deliberately delayed messages** when the user moved on to a different sub-task too quickly and waited to reintroduce the topic later at a more opportune moment (see E1 in Figure 6).

Lastly, we observed that **experts also frequently supported users in navigating CAD software features**, helping them by offering guidance on Generative Design functions, recommending workflow optimizations, helping with calculations such as load distribution, and providing help to locate tools and options as needed.

6.2 Perceived benefits and challenges of metacognitive support agents (RQ2)

Overall, participants appreciated the support from the agents. S2 stated that *"it's doing a good job [...] by assisting you throughout the whole design work"* and H1 noted a perceived efficiency gain: *"I feel like without [the agent], [...] it would have definitely taken a longer amount of time."* Besides positive aspects, participants highlighted trade-offs and challenges, which we present in the following sections, organized around the different support strategies and general agent interactions.

6.2.1 *User Feedback on SocratAIs (Question-Asking).*

Participants consistently found the agents' questions valuable, particularly those that prompted reflection on key design aspects, such as load cases and clearances. These questions helped refine GenAI input specifications by encouraging them to reconsider functional details and correct initial assumptions, as one participant noted, *"The questions [...] helped me reflect and go back over my train of thought and see, 'Am I missing something? Does this look like I'm doing what I'm supposed to do?'"* (S5).

Participants also reported that questions encouraged them to slow down and critically evaluate their thinking, much like a professor would in a one-on-one setting, as S1 explained: *"when [..] you're just sitting down designing by yourself, you don't often run through those things. So having someone to stop you and say, 'Why do you think that works?' is a good check every now and then"* (S1).

Some users found agent questions redundant but preferred them over missing important steps. For example, one participant noted that while *"25% of the questions actually helped,"* the rest pulled their attention away from the current task (S1). Others mentioned that the agents sometimes asked questions about actions they were already performing, which felt unnecessary as they were already thinking through those steps. However, participants acknowledged that redundancy was preferable to missing something important and found the frequency of questions appropriate.

Some found the questions more useful for problem exploration and intent formulation: *"It did a good job [during] the initial part of the setting up of the design"* (S3). **Others saw more benefit in supporting outcome evaluation**, particularly when analyzing and comparing designs: *"[Questions during outcome evaluation phase were more] valuable because when you're comparing this many designs, it's good to be reminded of what's most important to compare and prioritize"* (S2).

6.2.2 *User Feedback on HephAIstus (Planning, Sketching, with Suggestions).*

The agent-provided project planning sheet was generally perceived as helpful, giving users a structured way to approach their tasks and a document to guide their process, as this participant stated: *"To show you an actual work plan from the design to the actual fabrication and production of the piece is good. [It helps you to] separate [the design process] into different steps and how we're going to work from here"* (H2).

Designers found that sketching helped them visually think through the design problem: *"Sketching a free body diagram was definitely helpful. I mean, it was just good to see before I had set it up in Fusion, sort of my plan for where the loads and constraints were gonna go"* (H3). Several participants suggested that the sketching feature could be improved by making it more interactive—for example, by providing real-time calculations, augmenting sketches with force vectors, and offering a library of example diagrams—to support thinking through a part's design requirements.

Many participants valued the agent's proactive suggestions, and when the agent pointed out possible inconsistencies in their load specifications, as H1 described: *"I wasn't sure why [the solver] was generating so thin [parts] and having the agent explain to me, 'hey, it's probably because of the constraints that you set, you have canceling loads, you should not do that.' That was good feedback to modify the constraints"* (H1).

Designers also liked when *HephAIstus* helped them catch slips and correct mismatches in real-time. For example, one user appreciated a hint that pointed out a mismatch between the force requirements in the design brief and the load setup: *"I thought it was extremely helpful. I [mis]read the instructions with the weight capacity [...] I'm glad I received a prompt to make sure that the weight distribution was accurate because it knew there was a difference between the weight the bracket was supposed to hold and the engine's total weight. I was impressed that the agent provided me that prompt to check and make that design change"* (H1).

Additionally, **participants also found the agent more efficient than searching online or sifting through video tutorials for answers.** Being able to ask the agent questions directly about specific software functions or design issues saved time and allowed them to stay focused on the task without interrupting their workflow: *"I didn't have to stop what I was doing [...] to go to Google and find information"* (H3).

The fact that the agent would also annotate the screen and highlight relevant interface elements was widely appreciated for reducing the need to search for tool functions manually,

as one participant noted: *"when I had doubts about specific functions in Fusion, I asked the agent and it was very helpful in that. And the fact that it would highlight the word to click, that was very useful"* (H2).

6.2.3 Reflections and Suggestions of the Expert-Freeform Wizards. In the post-task wizard interview (after supporting a designer), the task experts reflected on their support strategies and challenges. Some **experts emphasized the difference between operating Fusion360 and thinking through design problems, underscoring that while both of these tasks are equally important, they often require two distinct "mindsets":** *"I was thinking of it from a 'how do I use Fusion' standpoint. [But it's] kind of like two parts of the brain: One is like, 'I know where the buttons are, I know the workflow,' and there is like, really creative problem-solving"* (Expert 4).

Furthermore, some also suggested that **agent-initiated "design reviews" during outcome evaluation could help users in critical evaluation of GenAI outcomes**. Similarly, others suggested to **introducing deliberate "checkpoints" or reflection phases between design task steps** (e.g., when the user transitions from specifying loads to obstacle geometry), which could allow for better scaffolding of (metacognitive) support throughout the process.

Several **experts also saw the potential to pre-structure workflows by offering early guidance on problem exploration and setting up generative design inputs.** Some highlighted the value of "preemptive" planning activities, similar to *HephAIstus'* planning sheet—suggesting that preparing a clear design plan before switching to the GenAI tool could improve the overall design process.

Some experts deliberately focused on supporting users in outcome evaluation and also suggested that designers run a solver preview early to obtain visual feedback, helping users quickly assess if their setup was correct. Another observed recurring strategy was that during outcome evaluation, *Expert-Freeform* agents suggested looking back to realize flawed load specifications, as this expert described in the interview: *"If we get really hefty results like blocky stuff, then as the agent, I can say, 'Do these parts seem over-designed? Let's look back at our load cases!' And then we can recognize, 'okay, we applied that load to the full load to every entity'"* (Expert 1).

Lastly, experts also suggested instead of directly speaking agent messages, to **annotate screen elements to signal available feedback from the agent**, for example, by circling critical parts, to provide users opportunities to initiate a conversation with the agent when desired by clicking on these highlighted regions.

6.2.4 Feedback on General Interactions With Agent Probes . Here we summarize participants' feedback on general aspects of interacting with the support agents. **Participants generally appreciated the voice modality of the agent, finding it faster and more efficient than typing**, as H2 described, *"talking is more time efficient because in chat I'll first have to explain the issue and it will take longer and be less clear"* (H2).

Others highlighted the benefit of voice-based interactions for complex tasks that required creative or visual exploration: *"In a software like this, I could definitely see use in [voice]. Text can be convenient, [but] you might miss it. Having a voice is helpful [for] anything that requires an exploratory or creative process"* (H1). However, some users noted that **voice interaction could be impractical in shared or public workspaces**, such as offices or labs, and would prefer an additional text-based alternative in such environments.

Users also shared differing opinions about the overall benefits of the question-asking and support suggestion strategies. Designers in the *SocratAIs* group generally valued open-ended questions rather than providing direct answers, as it encouraged them to think critically about their design decisions and allowed for flexibility in their approach, as S5 shared: *"I don't think it should have pulled me straight up the answer. I think it was better to tell me, 'Hey, you should think about this,' because not every part is going to be maintained the same way or serviced the same way [...] By asking you a more open-ended question, it pointed you in a direction, but it also left open the possibility to ignore it"* (S5). However, a few users also suggested that **more direct educational scaffolding with explanations could be more helpful for less experienced users**, while other, more experienced users could prefer shorter prompts to ensure that their workflow stayed on track. Some participants in the *HephAIstus* group also emphasized that to fully rely on the agent, they would need to trust its understanding of complex design contexts and ensure that its recommendations are accurate, especially for critical engineering tasks.

7 Discussion

In the following sections, we discuss our findings and their implications for metacognitive design support systems and agent-based CAD support while highlighting key learnings, design considerations, and open questions for future GenAI design support systems (see Table 3).

7.1 Toward Metacognitive Design Support Systems

Our findings indicate that agent-facilitated metacognitive support can play a positive role in helping designers overcome the cognitive challenges of GenAI workflows: Designers receiving some form of support often had improved design outcomes compared to those without assistance[10]. In our exploratory prototyping study, we categorized support strategies into distinct agent probes to reveal nuanced benefits and tradeoffs, but also saw that none of our agents served as a one-size-fits-all solution. This suggests that *combining multiple strategies may ultimately prove more effective in practice*, and future work should explore systems with blended approaches.

Below we reflect on the findings of this study and highlight design considerations for future metacognitive GenAI support systems (Table 3). Regarding specific metacognitive support strategies, our findings indicate that **(A1) cueing users with thought-provoking open-ended questions can help with intent formulation, problem exploration, and outcome evaluation, leading to improved AI-generated outcomes** (see 6.1.2—6.1.3, 6.1.5). These findings align with prior evidence on the crucial role of questions within design processes [19, 39, 79]. Similar to other recent work [30, 74], our findings also emphasize AI agents' possible role as facilitators that can stimulate users' critical thinking, which

[10]Our result in the unsupported group aligns with previous findings using the same task with a similar population [42], and statistical tests also showed no significant differences in population characteristics between supported and unsupported groups in our study.

Table 3: Overview of design considerations and key learnings.

	Design Considerations / Key Learnings	Seen in
*Opportunities for agent-based **metacognitive support***		
A1	Cueing users with **thought-provoking open-ended questions** can help with *intent formulation*, *problem exploration*, and *outcome evaluation* in GenAI-assisted design tasks.	6.1.2, 6.1.3, 6.1.5
A2	Prompting **mental simulations through questions and sketching** can assist designers in thinking through design problems and more accurately formulating intents and specifying GenAI model inputs (supporting *intent formulation* and *problem exploration*).	6.1.3, 6.1.4
A3	Offering metacognitive support in **key moments** of GenAI-based design processes can enhance cognitive engagement, for example, by offering users **agent-driven "design review sessions" during part evaluation** or introducing **dedicated "reflection checkpoints"** when transitioning between subtasks.	6.1.3, 6.2.3
A4	Giving users **control over the type of metacognitive support** depending on their needs and experience level.	6.2.4
A5	Providing designers **custom-generated user-editable design checklists** to support planning and reflection of design decisions.	6.1.4, 6.2.2
*Opportunities for agent-based **CAD support***		
B1	Offering **suggestions for design decision and tool operation** in combination with **metacognitive support** to help **improve users' tool fluency** and **overcome cognitive GenAI workflow challenges**.	6.1.5, 6.2.3
B2	Enabling users to **verbally request support from agents** can help to maintain focus and reduce context-switching in complex and visual-heavy CAD tasks.	6.1.4, 6.2.2
B3	In addition to **voice agent feedback**, utilizing visual **screen annotations and text** can reduce cognitive load.	6.2.2
B4	Agents that follow user behavior over time offer the potential for **proactively providing reminders, hinting at inconsistencies, and suggestions for metacognitive support, tool operation, and design task considerations**.	6.2.2, 6.1.5
B5	**Visually signaling available agent feedback** for users to optionally engage in can reduce task interruptions.	6.2.3

challenges the common notion of GenAI systems as "oracles" that only provide definitive (but possibly inaccurate) solutions or answers. However, going further, a challenge will lie in determining when "asking" versus "telling" the user would be most appropriate. Further investigations could draw on principles from learning science, suggesting metacognitive processing may only be effective if preceded by adequate knowledge or initial instruction [13, 45].

Also, we saw that asking questions alone can have limitations: In our study, questions were less effective at challenging ingrained incorrect assumptions, indicating that guidance beyond questioning may sometimes be required, especially when users hold deep misconceptions. Similarly, repeated questioning also presented a dual effect: while it often helped users to repair flawed inputs, it sometimes led to over-reliance, with designers thinking the AI might know something more than them or that the AI is right, rather than engaging in deeper reflection. This risk of dependency aligns with other findings on in-action feedback during design tasks, where excessive guidance was observed to diminish self-reflection and critical evaluation [37, 107]. Future work should, therefore, explore *when* and *how* metacognitive support systems could provide assistance without increasing automation reliance.

From a technical perspective, recent advancements in natural language processing (NLP) have enabled automated generation of Socratic questions for teaching math [87] or debugging [2] and generating domain-specific educational questions by pre-training LLMs [17]. Building atop such technical foundations, future research should explore design task-specific question generation to prompt designers' self-reflection and critical thinking aligned to specific design domains. Furthermore, inspired by emerging process mining techniques focused on metacognition and self-regulated learning phases [1, 15, 108], future work could investigate ways to further optimize systems, for example by tailoring prompts to designers' specific situational metacognitive needs, such as *intent formulation*, *problem exploration* or *outcome evaluation* phases.

Regarding specific metacognitive strategies, our analysis indicates that **(A2) prompting mental simulations through questions and sketching can assist designers to think through and more accurately formulate intents and specify GenAI model inputs** (see 6.1.3, 6.1.4). These findings align with previous research on design cognition, suggesting that mental simulation presents a vital metacognitive process in design activities [8, 9]. In addition, we saw that providing distinct support for thought externalization and visualization through sketching (a known cognitive amplification strategy in design [7]) helped designers more carefully think through input specifications for the GenAI solver,

thus improving designers' intent formulation and problem specification. Overall, metacognitive agent support might potentially be helpful for many design processes, whether GenAI-supported or not. Based on these findings, future work should explore further metacognitive support mechanisms relevant to design within and outside of GenAI-assisted tasks, such as prompting mental simulations through questions or guiding users in gradually sketching and eliciting relevant input criteria.

While many questions may not have directly helped designers[11], we saw that the *right* reflective question at the *right* time can have a significant impact on the design process. However, anticipating and catching the right moment can be tricky, but some situations seemed to be more opportune than others. For example, during the GenAI preview and outcome evaluation phases, reflective questions helped designers in assessing and correcting generated parts by linking back components' structural errors to insufficiently specified model inputs (see 6.1.3). Likewise, some external expert wizards and users also emphasized that support during evaluation phases would be especially useful (see 6.2.3 and 6.2.1). Building atop this learning, future systems could **(A3) offer users agent-driven "design review sessions" during part evaluation** (similar to reviews in design education or professional collaborations [43, 72]) **and introducing dedicated "reflection checkpoints" between GenAI setup steps** (e.g., when the user transitions from specifying loads to obstacle geometry) to better scaffold (metacognitive) support throughout GenAI-based design processes (see 6.2.3).

Our findings also revealed differences in user preferences between more confident and inexperienced designers regarding questions versus suggestions (see 6.2.4), highlighting the desire for systems to give users **(A4) control over the type of metacognitive support depending on their needs and experience level.**

Lastly, future metacognitive support systems could provide **(A5) custom-generated user-editable design checklists to support planning and scaffolding of design decisions** (see 6.1.4, 6.2.2).

7.2 Opportunities for Agent-based CAD Support

While our study primarily focused on supporting designers working with GenAI within CAD environments, the findings also revealed interesting insights and opportunities for designing agent-based CAD support systems that can complement metacognitive strategies. For example, in our study, *HephAIstus*—in addition to its metacognitive planning and sketching support—had a positive impact on helping users work on the design task and software operation. However, *HephAIstus*' suggestions helped designers less to overcome GenAI-related cognitive challenges than *SocratAIs*' questions. Especially for supporting *intent formulation* and *problem exploration* related to load cases, *SocratAIs* was twice as effective as *HephAIstus* (see Section 6.1.2 and Figure 5), indicating that for intent formulation and problem exploration, questions paired with planning and sketching support might be more effective than suggestions. Consequently, we conclude that **(B1) metacognitive support through reflective questions, planning, and sketching is equally crucial for effectively supporting designers in GenAI tasks as providing suggestions for design decision and tool operation** [12] (see 6.1.5, 6.2.3).

Furthermore, based on the insights derived from *HephAIstus* and *Expert-Freeform*, we see various opportunities for systems providing real-time support for design tasks and software operation, along with metacognitive support in CAD and GenAI design workflows. Notably, we see opportunities for **(B2) enabling users to directly request information and metacognitive support from agents verbally while working on a task**, which seemed to have helped users maintain focus while reducing context-switching (see 6.1.4, 6.2.2).

Similarly, users also appraised the agent's **(B3) visual tool guidance by directly highlighting relevant GUI elements** (see 6.2.2). Recent NLP advancements in speech processing and synthesis [68, 94], as well as the increasing ability of multimodal AI models to visually understand and operate software GUIs [50, 71], provide promising foundations for future research to explore such multimodal conversational support agents further.

Additionally, by capturing and responding to user behavior, verbalizations, and (screen) context over time, **(B4) agents can proactively provide reminders, hinting at inconsistencies and suggestions for metacognitive support, tool operation, and design task considerations** (see 6.2.2, 6.1.5). With multimodal LLM's increasing context windows, such contextual longitudinal support seems to become increasingly feasible.

Lastly, instead of support agents always directly verbalizing their messages, future support systems could instead **(B5) visually signal available agent feedback for users to engage in if and when desired**, which could reduce task interruptions (see 6.2.3).

7.3 Limitations

We highlight the following limitations: The study followed an exploratory prototyping approach [105] that enabled us to compare the different agents' metacognitive strategies while allowing flexibility in how support was delivered (e.g., message timing and phrasing). As a result, our design insights are partially shaped by the individual wizards (the first author and four external experts), and repeating the study with different wizards may yield slightly different outcomes. Furthermore, to analyze the impact of agents on the design process, we used video interaction analysis to identify moments when participants visibly considered new, relevant aspects in response to agent messages. While this yielded valuable findings, future work could incorporate additional user-agent interaction dynamics to further surface complementary insights. In terms of population, our participants represent only a subset of engineering designers. While all participants had relevant training in design and experience with 3D CAD software, many had limited industry exposure. To address this imbalance, we included five professionals with more extensive industry experience. Most of these professional users were part of the expert-facilitated agents, which might have biased the results. However, we disregarded this potential bias since the observed behaviors were similar across all supported groups. Furthermore, as the participants in our study were self-selected, they were likely interested in or receptive to

[11] On average only 6 out of 23 questions of *SocratAIs*' sessions had an observable positive impact on the design process where the designer considered a new relevant aspect after receiving the message (see Table 2).

[12] This is also indicated by the comparable quality of outcomes across the different agent-supported groups.

GenAI systems. This openness to AI-supported work may have influenced some of our findings. Additionally, although we aimed to ensure the design tasks felt realistic, participants knew they were part of a research study and that their designs wouldn't be produced. They might have invested more time learning the tool and thinking through the problem to create practical designs in a real-world setting.

8 Conclusion

While GenAI tools promise to enhance design processes, many professionals struggle to work effectively with AI. Key challenges include specifying all design criteria upfront (intent formulation) and reduced cognitive engagement due to cognitive offloading, which can limit problem exploration and outcome evaluation. To address this, we explored metacognitive support agents in a Wizard of Oz user study. Our findings show that users with agent support developed more viable designs, though outcomes varied depending on support strategy. While designers recognized the benefits of such assistance, we also uncovered trade-offs and differing user preferences. Based on these results, we highlight opportunities and trade-offs of metacognitive support agents and implications for AI-based design tools. While this work explores metacognitive support agents for GenAI-assisted mechanical part creation, the findings and design considerations offer promising avenues for research in other AI-assisted workflows and insights for developing new support techniques for AI-based design applications.

Acknowledgments

We thank all study participants, Autodesk experts, and the research assistants Anna Xu and Claire Malella for supporting this work. We would also like to thank Aniket Kittur, Christopher McComb, and Nur Yildirim for providing early feedback on a draft of this manuscript and the reviewers of this submission for their constructive suggestions. This material is based upon work supported by the National Science Foundation under Grant No. #2118924 Supporting Designers in Learning to Co-create with AI for Complex Computational Design Tasks.

References

[1] Mark Abdelshiheed, John Wesley Hostetter, Tiffany Barnes, and Min Chi. 2023. Leveraging Deep Reinforcement Learning for Metacognitive Interventions Across Intelligent Tutoring Systems. In *Artificial Intelligence in Education*, Ning Wang, Genaro Rebolledo-Mendez, Noboru Matsuda, Olga C. Santos, and Vania Dimitrova (Eds.). Vol. 13916. Springer Nature Switzerland, Cham, 291–303. doi:10.1007/978-3-031-36272-9_24

[2] Erfan Al-Hossami, Razvan Bunescu, Ryan Teehan, Laurel Powell, Khyati Mahajan, and Mohsen Dorodchi. 2023. Socratic Questioning of Novice Debuggers: A Benchmark Dataset and Preliminary Evaluations. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*. Association for Computational Linguistics, Toronto, Canada, 709–726. doi:10.18653/v1/2023.bea-1.57

[3] Zeyad Alshaikh, L. Tamang, and V. Rus. 2020. Experiments with a Socratic Intelligent Tutoring System for Source Code Understanding. In *The Florida AI Research Society*.

[4] Marco Aurisicchio, Rob H Bracewell, Ken M Wallace, et al. 2007. Characterising Design Questions That Involve Reasoning. In *DS 42: Proceedings of ICED 2007, the 16th International Conference on Engineering Design, Paris, France, 28.-31.07. 2007.*

[5] Autodesk. 2020. Fusion 360 Generative Design. https://www.autodesk.com/solutions/generative-design/manufacturing.

[6] Paul Ayres and John Sweller. 2005. The Split-Attention Principle in Multimedia Learning. In *The Cambridge Handbook of Multimedia Learning* (1 ed.), Richard Mayer (Ed.). Cambridge University Press, 135–146. doi:10.1017/CBO9780511816819.009

[7] Maral Babapour. 2015. Roles of Externalisation Activities in the Design Process. *Swedish Design Research Journal* 2014 (May 2015). doi:10.3384/svid.2000-964x.14134

[8] Linden J. Ball and Bo T. Christensen. 2009. Analogical Reasoning and Mental Simulation in Design: Two Strategies Linked to Uncertainty Resolution. *Design Studies* 30, 2 (March 2009), 169–186. doi:10.1016/j.destud.2008.12.005

[9] Linden J. Ball and Bo T. Christensen. 2019. Advancing an Understanding of Design Cognition and Design Metacognition: Progress and Prospects. *Design Studies* 65 (Nov. 2019), 35–59. doi:10.1016/j.destud.2019.10.003

[10] Maria Bannert, Peter Reimann, and Christoph Sonnenberg. 2014. Process Mining Techniques for Analysing Patterns and Strategies in Students' Self-Regulated Learning. *Metacognition and Learning* 9, 2 (Aug. 2014), 161–185. doi:10.1007/s11409-013-9107-6

[11] Eric P.S. Baumer and Bill Tomlinson. 2011. Comparing Activity Theory with Distributed Cognition for Video Analysis: Beyond "Kicking the Tires". In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. Association for Computing Machinery, New York, NY, USA, 133–142. doi:10.1145/1978942.1978962

[12] Warren Berger. 2014. *A More Beautiful Question: The Power of Inquiry to Spark Breakthrough Ideas*. Bloomsbury USA, New York, NY.

[13] Dianne C. Berry and Donald E. Broadbent. 1984. On the Relationship between Task Performance and Associated Verbalizable Knowledge. *The Quarterly Journal of Experimental Psychology Section A* 36, 2 (May 1984), 209–231. doi:10.1080/14640748408402156

[14] Kirsten Boehner, Janet Vertesi, Phoebe Sengers, and Paul Dourish. 2007. How HCI Interprets the Probes. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. Association for Computing Machinery, New York, NY, USA, 1077–1086. doi:10.1145/1240624.1240789

[15] Conrad Borchers, Jiayi Zhang, Ryan S. Baker, and Vincent Aleven. 2024. Using Think-Aloud Data to Understand Relations between Self-Regulation Cycle Characteristics and Student Performance in Intelligent Tutoring Systems. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*. 529–539. doi:10.1145/3636555.3636911 arXiv:2312.05675 [cs]

[16] Virginia Braun and Victoria Clarke. 2019. Reflecting on Reflexive Thematic Analysis. *Qualitative Research in Sport, Exercise and Health* 11, 4 (Aug. 2019), 589–597. doi:10.1080/2159676X.2019.1628806

[17] Sahan Bulathwela, Hamze Muse, and Emine Yilmaz. 2023. Scalable Educational Question Generation with Pre-trained Language Models. In *Artificial Intelligence in Education*, Ning Wang, Genaro Rebolledo-Mendez, Noboru Matsuda, Olga C. Santos, and Vania Dimitrova (Eds.). Vol. 13916. Springer Nature Switzerland, Cham, 327–339. doi:10.1007/978-3-031-36272-9_27

[18] Çetin Tünger, Çetin Tünger, Şule Taşlı Pektaş, and Sule Tasli Pektas. 2020. A Comparison of the Cognitive Actions of Designers in Geometry-Based and Parametric Design Environments. *Open House International* 45 (June 2020), 87–101. doi:10.1108/ohi-04-2020-0008

[19] Carlos Cardoso, Ozgur Eris, Petra Badke-Schaub, and Marco Aurisicchio. 2014. Question Asking in Design Reviews: How Does Inquiry Facilitate the Learning Interaction?. In *Proceedings of the 10th Design Thinking Research Symposium (DTRS)*. Purdue University, 18.

[20] Juan C. Castro-Alonso and John Sweller. 2020. The Modality Effect of Cognitive Load Theory. In *Advances in Human Factors in Training, Education, and Learning Sciences*, Waldemar Karwowski, Tareq Ahram, and Salman Nazir (Eds.). Springer International Publishing, Cham, 75–84.

[21] Catherine C. Chase, Jenna Marks, Deena Bernett, Melissa Bradley, and Vincent Aleven. 2015. Towards the Development of the Invention Coach: A Naturalistic Study of Teacher Guidance for an Exploratory Learning Task. In *Artificial Intelligence in Education*, Cristina Conati, Neil Heffernan, Antonija Mitrovic, and M. Felisa Verdejo (Eds.). Vol. 9112. Springer International Publishing, Cham, 558–561. doi:10.1007/978-3-319-19773-9_61

[22] Rimika Chaudhury, Taha Liaqat, and Parmit K. Chilana. 2023. Exploring the Needs of Informal Learners of Computational Skills: Probe-Based Elicitation for the Design of Self-Monitoring Interventions.

[23] Xiang 'Anthony' Chen, Ye Tao, Guanyun Wang, Runchang Kang, Tovi Grossman, Stelian Coros, and Scott E. Hudson. 2018. Forte: User-Driven Generative Design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–12. doi:10.1145/3173574.3174070

[24] Michelene T. H. Chi. 2006. Laboratory Methods for Assessing Experts' and Novices' Knowledge. In *The Cambridge Handbook of Expertise and Expert Performance*, K. Anders Ericsson, Neil Charness, Paul J. Feltovich, and Robert R. Hoffman (Eds.). Cambridge University Press, Cambridge, 167–184. doi:10.1017/CBO9780511816796.010

[25] Parmit K. Chilana, Nathaniel Hudson, Srinjita Bhaduri, Prashant Shashikumar, and Shaun K. Kane. 2018. Supporting Remote Real-Time Expert Help: Opportunities and Challenges for Novice 3D Modelers. (Oct. 2018), 157–166. doi:10.1109/vlhcc.2018.8506568

[26] Carlos Coimbra Cardoso, Petra Badke-Schaub, and Ozgur Eris. 2016. Inflection Moments in Design Discourse: How Questions Drive Problem Framing during Idea Generation. *Design Studies* 46 (Sept. 2016), 59–78. doi:10.1016/j.destud.2016.07.002

[27] Scotty D. Craig, Jeremiah Sullins, Amy Witherspoon, and Barry Gholson. 2006. The Deep-Level-Reasoning-Question Effect: The Role of Dialogue and Deep-Level-Reasoning Questions During Vicarious Learning. *Cognition and Instruction* 24, 4 (2006), 565–591. jstor:27739846

[28] Nigel Cross. 2006. *Designerly Ways of Knowing*. Springer, Berlin London.

[29] Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. 1993. Wizard of Oz Studies — Why and How. *Knowledge-Based Systems* 6, 4 (Dec. 1993), 258–266. doi:10.1016/0950-7051(93)90017-N

[30] Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. 2023. Don't Just Tell Me, Ask Me: AI Systems That Intelligently Frame Explanations as Questions Improve Human Logical Discernment Accuracy over Causal AI Explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–13. doi:10.1145/3544548.3580672

[31] Nicholas Davis, Chih-PIn Hsiao, Kunwar Yashraj Singh, Lisa Li, Sanat Moningi, and Brian Magerko. 2015. Drawing Apprentice: An Enactive Co-Creative Agent for Artistic Collaboration. In *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition*. ACM, Glasgow United Kingdom, 185–186. doi:10.1145/2757226.2764555

[32] David DeLiema, Maggie Dahn, Virginia J. Flood, Ana Asuncion, Dor Abrahamson, Noel Enyedy, and Francis Steen. 2019. *Debugging as a Context for Fostering Reflection on Critical Thinking and Emotion* (1 ed.). Routledge, London, 209–228. doi:10.4324/9780429323058-13

[33] J. T. Dillon. 1984. The Classification of Research Questions. *Review of Educational Research* 54, 3 (Sept. 1984), 327–361. doi:10.3102/00346543054003327

[34] Kees Dorst and Nigel Cross. 2001. Creativity in the Design Process: Co-Evolution of Problem–Solution. *Design Studies* 22, 5 (Sept. 2001), 425–437. doi:10.1016/S0142-694X(01)00009-6

[35] Graham Dove, Nicolai Brodersen Hansen, and Kim Halskov. 2016. An Argument For Design Space Reflection. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction (NordiCHI '16)*. Association for Computing Machinery, New York, NY, USA, 1–10. doi:10.1145/2971485.2971528

[36] Ian Drosos, Advait Sarkar, Xiaotong Xu, Carina Negreanu, Sean Rintel, and Lev Tankelevitch. 2024. "It's like a Rubber Duck That Talks Back": Understanding Generative AI-Assisted Data Analysis Workflows through a Participatory Prompting Study. In *Proceedings of the 3rd Annual Meeting of the Symposium on Human-Computer Interaction for Work (CHIWORK '24)*. Association for Computing Machinery, New York, NY, USA, 1–21. doi:10.1145/3663384.3663389

[37] Jane L. E, Yu-Chun Grace Yen, Isabelle Yan Pan, Grace Lin, Mingyi Li, Hyoungwook Jin, Mengyi Chen, Haijun Xia, and Steven P. Dow. 2024. When to Give Feedback: Exploring Tradeoffs in the Timing of Design Feedback. In *Proceedings of the 16th Conference on Creativity & Cognition (C&C '24)*. Association for Computing Machinery, New York, NY, USA, 292–310. doi:10.1145/3635636.3656183

[38] Linda Elder and Richard Paul. 2016. *The Thinker's Guide to The Art of Socratic Questioning*. Foundation for Critical Thinking Press.

[39] Ozgur Eris. 2004. *Effective Inquiry for Innovative Engineering Design*. Springer US, Boston, MA. doi:10.1007/978-1-4419-8943-7

[40] John H. Flavell. 1979. Metacognition and Cognitive Monitoring: A New Area of Cognitive–Developmental Inquiry. *American Psychologist* 34, 10 (Oct. 1979), 906–911. doi:10.1037/0003-066X.34.10.906

[41] formlabs. 2020. Generative Design 101. https://formlabs.com/blog/generative-design/.

[42] 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. doi:10.1145/3544548.3580999

[43] Gabriela Goldschmidt, Hagay Hochman, and Itay Dafni. 2010. The Design Studio "Crit": Teacher–Student Communication. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing* 24, 3 (Aug. 2010), 285–302. doi:10.1017/S089006041000020X

[44] Arthur C. Graesser and Natalie K. Person. 1994. Question Asking During Tutoring. *American Educational Research Journal* 31, 1 (March 1994), 104–137. doi:10.3102/00028312031001104

[45] Douglas J. Hacker (Ed.). 1998. *Metacognition in Educational Theory and Practice*. Erlbaum, Mahwah, NJ.

[46] Robert Hausmann and Kurt Vanlehn. 2010. The Effect of Self-Explaining on Robust Learning. *I. J. Artificial Intelligence in Education* 20 (Jan. 2010), 303–332. doi:10.3233/JAI-2010-010

[47] Sofie Heirweg, Mona De Smul, Emmelien Merchie, Geert Devos, and Hilde Keer. 2020. Mine the Process: Investigating the Cyclical Nature of Upper Primary School Students' Self-Regulated Learning. *Instructional Science* 48 (Aug. 2020). doi:10.1007/s11251-020-09519-0

[48] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. 2022. Imagen Video: High Definition Video Generation with Diffusion Models. https://arxiv.org/abs/2210.02303v1.

[49] Yueh-Ren Ho, Bao-Yu Chen, and Chien-Ming Li. 2023. Thinking More Wisely: Using the Socratic Method to Develop Critical Thinking Skills amongst Healthcare Students. *BMC Medical Education* 23, 1 (March 2023), 173. doi:10.1186/s12909-023-04134-2

[50] Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxuan Zhang, Juanzi Li, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. 2023. CogAgent: A Visual Language Model for GUI Agents. arXiv:2312.08914

[51] Ada Hurst, Shirley Lin, Claire Treacy, Oscar G. Nespoli, and John S. Gero. 2023. Comparing Academics and Practitioners Q&A Tutoring in the Engineering Design Studio. *Proceedings of the Design Society* 3 (July 2023), 997–1006. doi:10.1017/pds.2023.100

[52] Nikhita Joshi, Justin Matejka, Fraser Anderson, Tovi Grossman, and George Fitzmaurice. 2020. MicroMentor: Peer-to-Peer Software Help Sessions in Three Minutes or Less. (April 2020), 1–13. doi:10.1145/3313831.3376230

[53] Shabnam Kavousi, Patrick A. Miller, and Patricia A. Alexander. 2020. The Role of Metacognition in the First-Year Design Lab. *Educational Technology Research and Development* 68, 6 (Dec. 2020), 3471–3494. doi:10.1007/s11423-020-09848-4

[54] Rubaiat Habib Kazi, Tovi Grossman, Hyunmin Cheong, Ali Hashemi, and George Fitzmaurice. 2017. DreamSketch: Early Stage 3D Design Explorations with Sketching and Generative Design. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*. ACM, Québec City QC Canada, 401–414. doi:10.1145/3126594.3126662

[55] Sumbul Khan and Bige Tunçer. 2019. Gesture and Speech Elicitation for 3D CAD Modeling in Conceptual Design. *Automation in Construction* 106 (Oct. 2019), 102847. doi:10.1016/j.autcon.2019.102847

[56] Hannah Kim, Jaegul Choo, Haesun Park, and Alex Endert. 2016. InterAxis: Steering Scatterplot Axes via Observation-Level Interaction. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (Jan. 2016), 131–140. doi:10.1109/TVCG.2015.2467615

[57] Amy J. Ko and Brad A. Myers. 2004. Designing the Whyline: A Debugging Interface for Asking Questions about Program Behavior. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Vienna Austria, 151–158. doi:10.1145/985692.985712

[58] Jon Kolko. 2010. Abductive Thinking and Sensemaking: The Drivers of Design Synthesis. *Design Issues* 26, 1 (Jan. 2010), 15–28. doi:10.1162/desi.2010.26.1.15

[59] Rebecca Krosnick, Fraser Anderson, Justin Matejka, Steve Oney, Walter S. Lasecki, Tovi Grossman, and George Fitzmaurice. 2021. Think-Aloud Computing: Supporting Rich and Low-Effort Knowledge Capture. *International Conference on Human Factors in Computing Systems* (May 2021). doi:10.1145/3411764.3445066

[60] Kelly Y. L. Ku and Irene T. Ho. 2010. Metacognitive Strategies That Enhance Critical Thinking. *Metacognition and Learning* 5, 3 (Dec. 2010), 251–267. doi:10.1007/s11409-010-9060-6

[61] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*. ACM, Atlanta Georgia USA, 126–137. doi:10.1145/2678025.2701399

[62] Mustafa Kurt and Sevinc Kurt. 2017. Improving Design Understandings and Skills through Enhanced Metacognition: Reflective Design Journals. *International Journal of Art & Design Education* 36, 2 (2017), 226–238. doi:10.1111/jade.12094

[63] Hannu Kuusela and Pallab Paul. 2000. A Comparison of Concurrent and Retrospective Verbal Protocol Analysis. *The American Journal of Psychology* 113, 3 (2000), 387. doi:10.2307/1423365 jstor:1423365

[64] J. D. Lee and K. A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46, 1 (Jan. 2004), 50–80. doi:10.1518/hfes.46.1.50_30392

[65] Wendy G. Lehnert. 1978. *The Process of Question Answering: A Computer Simulation of Cognition* (1 ed.). Routledge, London. doi:10.4324/9781003316817

[66] 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. doi:10.1145/3173574.3173943

[67] Justin Matejka, Tovi Grossman, and George Fitzmaurice. 2011. Ambient Help. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. Association for Computing Machinery, New York, NY, USA, 2751–2760. doi:10.1145/1978942.1979349

[68] Ambuj Mehrish, Navonil Majumder, Rishabh Bhardwaj, Rada Mihalcea, and Soujanya Poria. 2023. A Review of Deep Learning Techniques for Speech Processing. arXiv:2305.00359

[69] Achim Menges and Sean Ahlquist (Eds.). 2011. *Computational Design Thinking* (1. publ ed.). Wiley, Chichester.

[70] Hongwei Niu, Cees Van Leeuwen, Jia Hao, Guoxin Wang, and Thomas Lachmann. 2022. Multimodal Natural Human–Computer Interfaces for Computer-Aided Design: A Review Paper. *Applied Sciences* 12, 13 (June 2022), 6510.

doi:10.3390/app12136510
[71] Runliang Niu, Jindong Li, Shiqi Wang, Yali Fu, Xiyu Hu, Xueyuan Leng, He Kong, Yi Chang, and Qi Wang. 2024. ScreenAgent: A Vision Language Model-driven Computer Control Agent. (2024). doi:10.48550/ARXIV.2402.07945
[72] Yeonjoo Oh, Suguru Ishizaki, Mark D. Gross, and Ellen Yi-Luen Do. 2013. A Theoretical Framework of Design Critiquing in Architecture Studios. *Design Studies* 34, 3 (May 2013), 302–325. doi:10.1016/j.destud.2012.08.004
[73] Ernesto Panadero. 2017. A Review of Self-regulated Learning: Six Models and Four Directions for Research. *Frontiers in Psychology* 8 (April 2017), 422. doi:10.3389/fpsyg.2017.00422
[74] Soya Park, Hari Subramonyam, and Chinmay Kulkarni. 2024. Thinking Assistants: LLM-Based Conversational Assistants That Help Users Think By Asking Rather than Answering. doi:10.48550/arXiv.2312.06024 arXiv:2312.06024 [cs]
[75] Maria Teresa Parreira, Sarah Gillet, and Iolanda Leite. 2023. Robot Duck Debugging: Can Attentive Listening Improve Problem Solving?. In *Proceedings of the 25th International Conference on Multimodal Interaction (ICMI '23)*. Association for Computing Machinery, New York, NY, USA, 527–536. doi:10.1145/3577190.3614160
[76] Carolyn Plumb, Rose Marra, Douglas Hacker, and John Dunlosky. 2018. Measuring Engineering Students' Metacognition with a Think-Aloud Protocol. In *2018 ASEE Annual Conference & Exposition Proceedings*. ASEE Conferences, Salt Lake City, Utah, 30796. doi:10.18260/1-2--30796
[77] Raquel Plumed, Carmen González-Lluch, David Pérez-López, Manuel Contero, and Jorge D Camba. 2021. A Voice-Based Annotation System for Collaborative Computer-Aided Design. *Journal of Computational Design and Engineering* 8, 2 (April 2021), 536–546. doi:10.1093/jcde/qwaa092
[78] Pontificia Universidad Javeriana, Juanita Tobón, Fabio Tellez, and Oscar Alzate. 2019. Metacognition in the Wild: Metacognitive Studies in Design Education. In *Insider Knowledge - Proceedings of the Design Research Society Learn X Design Conference, 2019*. Design Research Society. doi:10.21606/learnxdesign.2019.09128
[79] Rebecca Anne Price and Peter Lloyd. 2022. Asking Effective Questions: Awareness of Bias in Designerly Thinking. In *Handbook of Engineering Systems Design*, Anja Maier, Josef Oehmen, and Pieter E. Vermaas (Eds.). Springer International Publishing, Cham, 1–16. doi:10.1007/978-3-030-46054-9_24-3
[80] Xiangshi Ren, Gao Zhang, and Guozhong Dai. 2000. An Experimental Study of Input Modes for Multimodal Human-Computer Interaction. In *Advances in Multimodal Interfaces — ICMI 2000 (Lecture Notes in Computer Science)*, Tieniu Tan, Yuanchun Shi, and Wen Gao (Eds.). Springer, Berlin, Heidelberg, 49–56. doi:10.1007/3-540-40063-X_7
[81] Evan F. Risko and Sam J. Gilbert. 2016. Cognitive Offloading. *Trends in Cognitive Sciences* 20, 9 (Sept. 2016), 676–688. doi:10.1016/j.tics.2016.07.002
[82] Debrina Roy, Nicole Calpin, Kathy Cheng, Alison Olechowski, Andrea P. Argüelles, Nicolás F. Soria Zurita, and Jessica Menold. 2024. Designing Together: Exploring Collaborative Dynamics of Multi-Objective Design Problems in Virtual Environments. *Journal of Mechanical Design* 146, 3 (March 2024), 031702. doi:10.1115/1.4063658
[83] Marta Royo, Elena Mulet, Vicente Chulvi, and Francisco Felip. 2021. Guiding Questions for Increasing the Generation of Product Ideas to Meet Changing Needs (QuChaNe). *Research in Engineering Design* 32, 3 (July 2021), 411–430. doi:10.1007/s00163-021-00364-x
[84] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. https://arxiv.org/abs/2205.11487v1.
[85] Donald A. Schön. 1983. *The Reflective Practitioner: How Professionals Think in Action*. Basic Books, New York.
[86] Moushumi Sharmin and Brian P. Bailey. 2011. "I Reflect to Improve My Design": Investigating the Role and Process of Reflection in Creative Design. In *Proceedings of the 8th ACM Conference on Creativity and Cognition (C&C '11)*. Association for Computing Machinery, New York, NY, USA, 389–390. doi:10.1145/2069618.2069710
[87] Kumar Shridhar, Jakub Macina, Mennatallah El-Assady, Tanmay Sinha, Manu Kapur, and Mrinmaya Sachan. 2022. Automatic Generation of Socratic Subquestions for Teaching Math Word Problems. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 4136–4149. doi:10.18653/v1/2022.emnlp-main.277
[88] Keith Stenning, Alexander Schmoelz, Heather Wren, Elias Stouraitis, Theodore Scaltsas, Constantine Alexopoulos, and Amelie Aichhorn. 2016. Socratic Dialogue as a Teaching and Research Method for Co-Creativity? *Digital Culture and Education* 8, 2 (2016), 13.
[89] Hari Subramonyam, Roy Pea, Christopher Pondoc, Maneesh Agrawala, and Colleen Seifert. 2024. Bridging the Gulf of Envisioning: Cognitive Challenges in Prompt Based Interactions with LLMs. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–19. doi:10.1145/3613904.3642754
[90] Lasang Jimba Tamang, Zeyad Alshaikh, Nisrine Ait Khayi, Priti Oli, and Vasile Rus. 2021. A Comparative Study of Free Self-Explanations and Socratic Tutoring Explanations for Source Code Comprehension. *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education* (March 2021), 219–225. doi:10.1145/3408877.3432423
[91] 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. doi:10.48550/arXiv.2312.10893 arXiv:2312.10893 [cs]
[92] Andrew A. Tawfik, Arthur Graesser, Jessica Gatewood, and Jaclyn Gishbaugher. 2020. Role of Questions in Inquiry-Based Instruction: Towards a Design Taxonomy for Question-Asking and Implications for Design. *Educational Technology Research and Development* 68, 2 (April 2020), 653–678. doi:10.1007/s11423-020-09738-9
[93] Suzanne Tolmeijer, Naim Zierau, Andreas Janson, Jalil Sebastian Wahdatehagh, Jan Marco Marco Leimeister, and Abraham Bernstein. 2021. Female by Default? – Exploring the Effect of Voice Assistant Gender and Pitch on Trait and Trust Attribution. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, 1–7. doi:10.1145/3411763.3451623
[94] Andreas Triantafyllopoulos, Björn W. Schuller, Gökçe İymen, Metin Sezgin, Xiangheng He, Zijiang Yang, Panagiotis Tzirakis, Shuo Liu, Silvan Mertes, Elisabeth André, Ruibo Fu, and Jianhua Tao. 2023. An Overview of Affective Speech Synthesis and Conversion in the Deep Learning Era. *Proc. IEEE* 111, 10 (Oct. 2023), 1355–1381. doi:10.1109/JPROC.2023.3250266
[95] Maaike Van Den Haak, Menno De Jong, and Peter Jan Schellens. 2003. Retrospective vs. Concurrent Think-Aloud Protocols: Testing the Usability of an Online Library Catalogue. *Behaviour & Information Technology* 22, 5 (Sept. 2003), 339–351. doi:10.1080/0044929031000
[96] Kurt VanLehn, Randolph M. Jones, and Michelene T.H. Chi. 1992. A Model of the Self-Explanation Effect. *Journal of the Learning Sciences* 2, 1 (Jan. 1992), 1–59. doi:10.1207/s15327809jls0201_1
[97] Geoffrey Vaughan. 2022. Metacognition and Self-Regulated Learning: Recent Perspectives for an International Context. *Opus et Educatio* 9, 2 (Aug. 2022). doi:10.3311/ope.501
[98] Tom Veuskens, Danny Leen, and Raf Ramakers. 2022. Identifying Opportunities to Reimagine Parametric Modeling for Makers. (2022).
[99] Min Wang and Yong Zeng. 2009. Asking the Right Questions to Elicit Product Requirements. *International Journal of Computer Integrated Manufacturing* 22, 4 (April 2009), 283–298. doi:10.1080/09511920802232902
[100] Judith D. Wilson. 1987. A Socratic Approach to Helping Novice Programmers Debug Programs. In *Proceedings of the Eighteenth SIGCSE Technical Symposium on Computer Science Education - SIGCSE '87*. ACM Press, St. Louis, Missouri, United States, 179–182. doi:10.1145/31820.31755
[101] Robert Woodbury. 2010. *Elements of Parametric Design*. Routledge, London.
[102] Humphrey Yang, Kuanren Qian, Haolin Liu, Yuxuan Yu, Jianzhe Gu, Matthew McGehee, Yongjie Jessica Zhang, and Lining Yao. 2020. SimuLearn: Fast and Accurate Simulator to Support Morphing Materials Design and Workflows. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. ACM, Virtual Event USA, 71–84. doi:10.1145/3379337.3415867
[103] Kirsty Young. 2009. Direct from the Source: The Value of 'think-Aloud' Data in Understanding Learning. *The Journal of Educational Enquiry* 6 (2009).
[104] Loutfouz Zaman, Wolfgang Stuerzlinger, Christian Neugebauer, Rob Woodbury, Maher Elkhaldi, Naghmi Shireen, and Michael Terry. 2015. *GEM-NI*: A System for Creating and Managing Alternatives In Generative Design. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, Seoul Republic of Korea, 1201–1210. doi:10.1145/2702123.2702398
[105] J.D. Zamfirescu-Pereira, David Sirkin, David Goedicke, Ray LC, Natalie Friedman, Ilan Mandel, Nikolas Martelaro, and Wendy Ju. 2021. Fake It to Make It: Exploratory Prototyping in HRI. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI '21 Companion)*. Association for Computing Machinery, New York, NY, USA, 19–28. doi:10.1145/3434074.3446909
[106] 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. doi:10.1145/3544548.3581388
[107] Guanglu Zhang, Ayush Raina, Jonathan Cagan, and Christopher McComb. 2021. A Cautionary Tale about the Impact of AI on Human Design Teams. *Design Studies* 72 (Jan. 2021), 100990. doi:10.1016/j.destud.2021.100990
[108] Jiayi Zhang, Conrad Borchers, Vincent Aleven, and Ryan Shaun Baker. 2024. Using Large Language Models to Detect Self-Regulated Learning in Think-Aloud Protocols. doi:10.35542/osf.io/hrtz6

A Additional Materials

Table 4: Overview of study participants.

ID	Agent Group	Age	Role	MechDes Exp. Years	Indus. Exp. Years	CAD Exp. Years	FEA Prof.	DFM Prof.
B1	No Support	22	Student, MA Mechanical Engineering	3–5	0	2–4	5	2
B2	No Support	28	Student, PhD Mechanical Engineering	3–5	1–2	5+	6	2
B3	No Support	27	Researcher, Mechanical Engineering	6–10	0	5+	7	5
B4	No Support	23	Student, MA Mechanical Engineering	3–5	3–5	2–4	7	5
B5	No Support	39	Researcher, Mechanical Engineering	3–5	0	5+	1	1
S1	SocratAIs	26	Student, BS Mechanical Engineering	6–10	0	5+	4	1
S2	SocratAIs	23	Student, MS Mechanical Engineering	3–5	0	2–4	5	5
S3	SocratAIs	22	Student, MS Mechanical Engineering	1–2	1–2	2–4	4	1
S4	SocratAIs	26	Student, PhD Mechanical Engineering	6–10	1–2	5+	1	5
S5	SocratAIs	20	Student, BA Mechanical Engineering	3–5	0	2–4	4	6
H1	Hephaistus	30	Student, PhD Mechanical Engineering	3–5	3–5	5+	4	4
H2	Hephaistus	42	Student, PhD Mechanical Engineering	6–10	6–10	5+	4	1
H3	Hephaistus	22	Student, BA Mechanical Engineering	1–2	1–2	2–4	3	2
H4	Hephaistus	26	Mechanical Engineer	3–5	3–5	5+	2	2
H5	Hephaistus	21	Student, BA Mechanical Engineering	1–2	1–2	2–4	3	5
E1	Expert-Freeform	38	Mechanical Engineer	10+	10+	5+	7	7
E2	Expert-Freeform	26	Mechanical Engineer	1–2	3–5	5+	4	6
E3	Expert-Freeform	29	Mechanical Designer	10+	3–5	5+	5	7
E4	Expert-Freeform	29	Mechanical Engineer	6–10	6–10	5+	5	5
E5	Expert-Freeform	23	Student, MS Mechanical Engineering	3–5	0	5+	2	5

Table 5: Overview of demographics of Autodesk Fusion360 Generative Design experts who acted as wizards in the Expert-Freeform condition. Fusion360 Generative Design software proficiency was self-rated on a 1–7 scale.

ID	Age	Role	MechDes Exp. Years	F360 GenDes Prof.	F360 GenDes Training Exp.	Paired with
Expert 1	31	Senior Research Engineer	3 – 5	6/7	Trained customers, students, colleagues	E2, E5
Expert 2	35	Sr. Research & Design Engineer	6 – 10	7/7	Taught lectures, trained colleagues	E1
Expert 3	47	Principal Research Engineer	15+	7/7	Trained customer support teams	E4
Expert 4	27	Research and Design Engineer	3 – 5	6/7	Trained customers and colleagues	E3

Figure 7: Timeline plots visualizing participant and agent interactions throughout the design task; timelines are divided into lanes, each showing (in)correct GenAI input specifications (diamond shapes) for (1) forces, (2) constraints, (3) bolt and, (4) dampener clearances, (5) bounding box (from top to bottom); black and orange vertical lines represent exchanged agent and user messages with purple and black triangles indicating an observable impact on the design process.

A.1 Wizard Guidelines

The agents *SocratAIs* and *HephAIstus* were facilitated by the first author, with experience in mechanical engineering, Fusion360, and Generative Design. In some sessions, a second research team member with experience in mechanical engineering and Generative Design was co-present to provide additional verbal support for the main wizard.

A.1.1 ***Guidelines for the SocratAIs and HephAIstus wizard.***

The wizard of *SocratAIs* and *HephAIstus* followed these general guidelines:

1) Follow the designer's verbalizations and screen actions and pay close attention to the task-specific design steps and challenges as outlined in Section 3, such as specifying the bracket's load cases (forces and structural constraints), modeling appropriate geometry

for keeping bolts and dampeners free of material (obstacle geometry), defining DFM parameters such as materials and manufacturing options, and also to support users in evaluating the design previews and generated outcomes.
2) Pay close attention to inconsistencies between the requirements stated in the design brief and the input parameters set by the designer. Such requirements could be explicit (e.g., the force the bracket needs to hold) or implicit features, such as bolt clearances, which were not explicitly mentioned in the design brief.
3) Never directly tell the participant what to do, but rather provide supportive questions, hints, or suggestions (depending on the enacted agent type).
4) You are free to send messages whenever and how often you consider it helpful to the designer. However, pay special attention to moments in which designers transition between design sub-tasks (such as from specifying obstacle geometry to specifying loads), as well as when designers show hesitation or use hedging expressions (e.g., *"I am unsure if..."*).
5) You are free to formulate the messages in a way you consider to be most helpful, while adhering to the agent's support strategy (e.g., only asking questions).

A.1.2 ***Guidelines for the Expert-Freeform wizards.***

The *Expert-Freeform* wizards (external experts not part of the research team) received fewer instructions since we wanted to observe their natural support behavior. However, experts were told not to directly tell the participant what to do, but rather to help them work on the design task and with the GenAI system.

A.1.3 ***SocratAIs Agent Introduction*** .

SocratAIs: Hey! I am a voice agent here to support you during the design task. I can hear what you are saying, and I can see your screen and follow along with you while you work on the task. From time to time, I will ask you questions that are supposed to help you think through the design task. You can also ask me questions at any time.

A.1.4 ***HephAIstus.***

Agent introduction:

HephAIstus : Hey! I am a voice agent here to support you during the design task. I can hear what you are saying, and I can see your screen and follow along with you while you work on the task. Feel free to ask me any questions, and I'll do my best to provide you with answers. From time to time, I'll also chime in with tips and guidance to help you along the way.

Project Planning Activity:

HephAIstus provided metacognitive support by suggesting the user engage in a **project planning activity** by sharing a pre-generated text document outlining critical project-relevant aspects with the user:

HephAIstus : Hey! I noticed you are working on designing a ship engine mounting bracket and that you are planning to use Autodesk Fusion 360 Generative Design to explore different materials and manufacturing options. Before you start working on the task, I suggest having a planning session together to help you walk through the steps and design considerations in designing the bracket. What do you think?
(waiting for confirmation from user)
HephAIstus : I created a shared doc with a high-level breakdown of the steps and design considerations to be discussed. You can find the link in the chat.

Planning Sheet Document:

1) Project Scope and Requirements

- Define the objectives of the bracket design.
- Understand the operational conditions of the ship engine.
- Identify load types (static, dynamic, thermal) and magnitudes.
- Clarify space constraints and installation considerations.

2) Material Selection

- Discuss different material options (metal alloys, composites, etc.).
- Consider material properties such as strength, weight, corrosion resistance, and cost.
- Review the material performance under extreme marine conditions.

3) Manufacturing Method

- Determine feasible manufacturing methods (casting, machining, additive manufacturing, etc.).
- Evaluate the pros and cons of each method concerning the design objectives.
- Discuss generative design constraints for each manufacturing process.

4) Generative Design Parameters

- Set up load cases and boundary conditions in Fusion 360.
- Define the design space and apply necessary constraints and conditions.
- Choose the resolution of the generative design mesh.

5) Design Constraints and Criteria

- Set criteria for minimum safety factors.
- Define requirements for vibration dampening.
- Consider access for maintenance and installation.

6) Simulation and Analysis

- Plan for simulations to predict performance under various loads.
- Analyze stress distribution, deformation, and fatigue life.
- Review thermal and fluid flow analysis if necessary.

7) Optimization Objectives

- Establish the optimization goals, such as weight reduction, strength optimization, cost efficiency, etc.
- Discuss trade-offs between different optimization objectives.

8) Compliance and Standards

- Ensure the design meets marine industry standards and regulatory compliance.
- Consider classification society requirements and certifications.

9) Integration with Existing Systems

- Discuss how the bracket will integrate with the ship's engine and surrounding structures.
- Plan for interfaces with other systems and parts.

10) Lifecycle Considerations

- Consider the lifecycle impacts, such as ease of manufacture, sustainability, recyclability, and end-of-life disposal.
- Maintenance.

Free-body Diagram Sketching Activity:
The agent can suggest that the designer sketch out load case-relevant forces and constraints as a free-body diagram by sharing a link to a 2D drawing canvas containing the side and top view of the bracket as a starting point:

HephAIstus : Can you walk me through your load cases and constraints by sketching a free-body diagram? I shared a link to a board for you to sketch on in the chat (see Fig. 8).

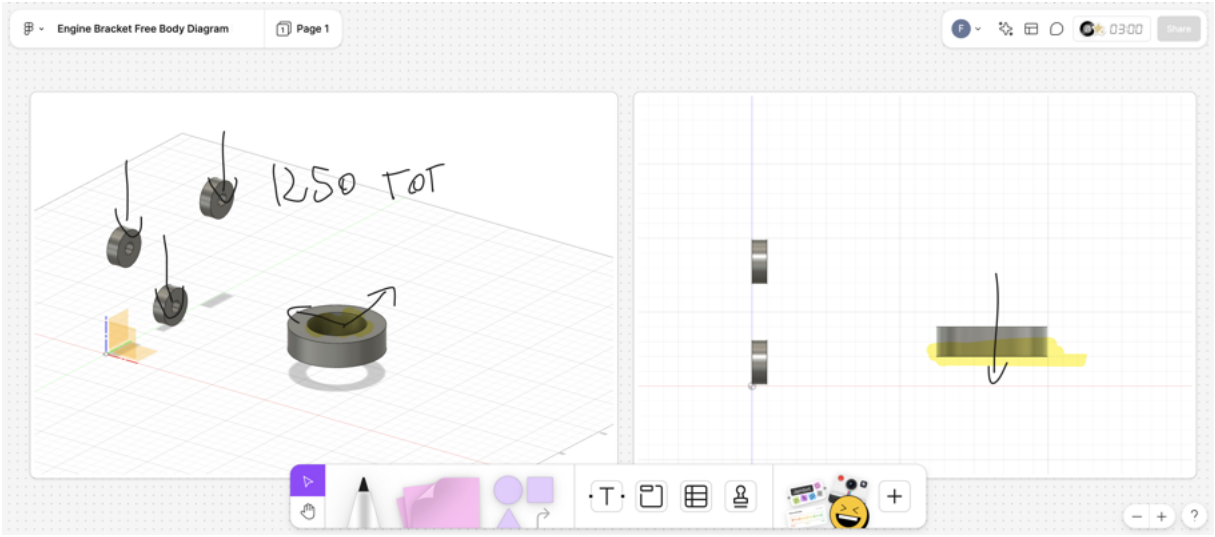

Figure 8: Screenshot of the sketching board *HephAIstus* sent to users (with scribbles from H3 on it).

A.1.5 ***Expert-Freeform Agent Introduction.***

Expert Agent: Hey! I am a voice agent here to support you during the design task. I can hear what you are saying, and I can see your screen and follow along with you while you work on the task. Feel free to ask me any questions, and I'll do my best to provide you with answers. From time to time, I'll also chime in with tips and guidance to help you along the way.

Table 6: Interview protocol with questions of the semi-structured post-task interview.

Nr	Question
General Feedback on task and thinking aloud	
Q1	Did you encounter any technical difficulties during the design session(s) that limited your ability to work on the task?
Q2	How challenging was the design task of designing an engine bracket for you in general?
Q3	How did it feel to think aloud during the task? Do you think that thinking aloud impacted your ability to complete the task in any way?
Feedback on working with the Generative Design feature	
Q4	Could you tell me what it was like to work with the Fusion 360 Generative Design feature in general?
Q5	Are you satisfied with the final design in general? How closely does it match the design brief?
Q6	Did you encounter any challenges in designing the engine bracket using the generative design feature?
Q7	Could you imagine using this tool in the future?
Q8	How much do you trust the results from the design tool?
Feedback on Support Agent	
Q9	Could you tell me what it was like to work with the design support agent in general?
Q10	Do you remember situations in which you found the agent's support helpful? In which not?
Q11	What would you want the design support agent to do more of?
Q12	What would you want the design support agent to do differently?
Q13	Would you use a tool like the design support agent in your work? Why or why not?
Q14	How did you like the frequency of messages?
Q15	How did you like the planning doc and sketching board?
Q16	In which phases did you find the support more or less helpful?
Q17	How useful did you find the questions that the agent asked you? Do you remember specific questions that you found helpful or unhelpful? Please explain.
Q18	Is there anything else you would like to share with us, or do you think we should know about?