Int. J. Human–Computer Studies 208 (2026) 103687

Contents lists available at ScienceDirect

International Journal of Human - Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

# Feed-O-Meter: Investigating AI-generated mentee personas as interactive agents for scaffolding design feedback practice

Hyunseung Lim [a], Dasom Choi [a], DaEun Choi [b], Sooyohn Nam [a], Hwajung Hong [a,*]

[a] KAIST, Department of Industrial Design, Daejeon, the Republic of Korea
[b] KAIST, School of Computing, Daejeon, the Republic of Korea

## HIGHLIGHTS

- This paper develops Feed-O-Meter to help students practice design feedback skills.
- We conduct a within-subject user study with 24 design students.
- We show that Feed-O-Meter helped students provide detailed and empathetic feedback.
- This paper provides design considerations for design education tools with AI agents.

## ARTICLE INFO

*Keywords:*
Design education
Design feedback
Human-computer interaction
Large language model
AI-generated agent

## ABSTRACT

Effective feedback, including critique and evaluation, helps designers develop design concepts and refine their ideas, supporting informed decision-making throughout the iterative design process. However, in studio-based design courses, students often struggle to provide feedback due to a lack of confidence and fear of being judged, which limits their ability to develop essential feedback-giving skills. Recent advances in large language models (LLMs) suggest that role-playing with AI agents can allow learners to engage in multi-turn feedback without the anxiety of external judgment or the time constraints of real-world settings. Yet prior studies have raised concerns that LLMs struggle to behave like real people in role-play scenarios, diminishing the educational benefits of these interactions. Therefore, designing AI-based agents that effectively support learners in practicing and developing intellectual reasoning skills requires more than merely assigning the target persona's personality and role to the agent. By addressing these issues, we present Feed-O-Meter, a novel system that employs carefully designed LLM-based agents to create an environment in which students can practice giving design feedback. The system enables users to role-play as mentors, providing feedback to an AI mentee and allowing them to reflect on how that feedback impacts the AI mentee's idea development process. A user study (N = 24) indicated that Feed-O-Meter increased participants' engagement and motivation through role-switching and helped them adjust feedback to be more comprehensible for an AI mentee. Based on these findings, we discuss future directions for designing systems to foster feedback skills in design education.

## 1. Introduction

In the iterative design process, where solutions are progressively refined, feedback is indispensable for enhancing the quality and effectiveness of design (Wynn and Eckert, 2017). Mastering the art of giving feedback is critical for designers, as it not only improves design performance (Wynn and Maier, 2022) but also integrates the diverse perspectives within a design team toward a shared objective (Valkenburg and Dorst, 1998) while enabling the expression of individual design viewpoints (Braha and Maimon, 1998). Recognizing its significance, design education has long emphasized feedback as a core learning experience. Among various educational strategies, peer feedback is a common way for students to practice giving and receiving feedback, which raises the quality of their work while fostering key competencies such as communication, collaboration, and critical thinking (McDonnell, 2016; Bjorklund et al., 2004).

* Corresponding author.
*Email addresses:* charlie9807@kaist.ac.kr (H. Lim), hwajung@kaist.ac.kr (H. Hong).
*URL:* https://dxd-lab.github.io/ (H. Hong).

https://doi.org/10.1016/j.ijhcs.2025.103687
Received 31 January 2025; Received in revised form 9 November 2025; Accepted 14 November 2025
Available online 26 November 2025

Despite the recognized value of feedback in enhancing both design outcomes and student competencies, there remain few opportunities for students to learn how to provide effective feedback. While many design studio courses encourage peer feedback through project-based learning (Ching and Hsu, 2013; Ertmer et al., 2007), students often face challenges in providing constructive feedback and remain disengaged in peer feedback (Ching and Hsu, 2013; Gielen et al., 2010; Hovardas et al., 2014). This difficulty stems from students' limited feedback experience and design knowledge (Ching and Hsu, 2013), as well as anxiety or hesitation about giving feedback due to concerns over receiving criticism for inadequate input (Ertmer et al., 2007; Gielen et al., 2010; Cook et al., 2020). To address this challenge, previous studies have emphasized the importance of providing comfortable educational environments that encourage students to actively engage in giving feedback and cultivate their feedback skills (Cook et al., 2020; Jug and Bean, 2018).

The HCI community has sought to create environments that foster higher-quality feedback in creative activities such as design. Prior research has proposed online platforms (Cheng et al., 2020; Lambropoulos et al., 2010) that allow design students to exchange feedback at scale, particularly crowd-sourcing systems (Oppenlaender et al., 2021; Krause et al., 2017; Lekschas et al., 2021), and interactive guidelines for crafting effective feedback (Ngoon et al., 2018). Although these approaches increase the volume of feedback and participation (Cheng et al., 2020; Oppenlaender et al., 2021), limited motivation and the absence of actual bi-directional communication often lead to superficial feedback (Oppenlaender et al., 2021; Nguyen et al., 2017). Meanwhile, advances in large language models (LLMs) have prompted investigations into AI agents that help learners practice logical and critical thinking skills—including arguing (Wambsganss et al., 2021), teaching (Markel et al., 2023), speech practice (Park and Choi, 2023), and conversing (Shaikh et al., 2024). For instance, GPTeach (Markel et al., 2023) employs an LLM-driven agent that adopts a student persona, enabling teaching assistants to rehearse their instructional strategies through simulated teaching scenarios. Role-playing with AI agents lets learners engage in multi-turn feedback without the anxiety of external judgment or the time constraints of real-world settings (Wambsganss et al., 2021; Markel et al., 2023; Shaikh et al., 2024).

However, prior studies have raised concerns that LLMs struggle to behave like real people in role-play scenarios, thereby diminishing the educational benefits of these interactions (Jin et al., 2024; Lim et al., 2024; Jo et al., 2023). For example, LLM-based agents with a student persona often provide responses that are far more intelligent and refined than those of actual students due to the extensive knowledge embedded in the LLMs (Jin et al., 2024; Lim et al., 2024). Studies have revealed that this misalignment between the agent's capabilities and its intended persona can lead users to inadvertently rely on the AI's responses rather than developing their own abilities through interactions (Jin et al., 2024; Lim et al., 2024; Jane et al., 2024). Therefore, designing AI-based agents that effectively support learners in practicing and developing intellectual reasoning skills requires more than merely assigning the target persona's personality and roles to the agent; it demands deliberate strategies to align the AI's behavior and knowledge system with its intended educational role (Markel et al., 2023; Jin et al., 2024; Jo et al., 2023).

Building on these insights, this study introduces a novel system with AI agents that empowers students to practice their design feedback skills. We propose Feed-O-Meter, a system that facilitates design feedback practice through role-playing interactions, allowing students to mentor an AI that adopts the persona of a design student. The system includes key features: (1) a chat interface for role-switched conversations, where users act as mentors and the AI as a mentee, (2) feedback reflection interfaces that show how their feedback influences the AI mentee's idea development, allowing users to reflect on the effectiveness of their feedback and make adjustments accordingly. By incorporating LLMs, Feed-O-Meter enables the AI agent to understand and respond to feedback in real-time while maintaining the role of a design student and visually demonstrating how the feedback impacts the mentee's idea development process.

To examine how students interact with Feed-O-Meter and its effectiveness in enhancing design feedback skills, we conducted a user study with 24 design students. This study employed a within-subject comparative study to assess the impact of the system's feedback reflection features on students' feedback. Our findings reveal that the AI mentee and Feed-O-Meter environment provided a realistic and low-pressure feedback experience, allowing participants to engage more actively in feedback activities without fear of judgment. Moreover, the system's feedback reflection features encouraged participants to focus not only on the content of their feedback but also on effective communication strategies to make their feedback clear and acceptable. Participants recognized that Feed-O-Meter could go beyond improving design feedback skills to help them critically analyze their own designs and identify areas for improvement. Based on our findings, we discuss the implications of role-playing interactions for feedback practice and offer insights into designing AI personas that effectively integrate these interactions within educational contexts.

The contributions of our paper are as follows:

1. The design and development of Feed-O-Meter, a system that allows design students to improve their feedback skills through role-playing interactions with an AI mentee. We outline the system's design rationale and capabilities, demonstrating how it facilitates effective feedback practices.
2. An empirical understanding of how design students engage with Feed-O-Meter. Through a user study with 24 design students, we analyzed interaction logs and interview transcripts to provide insights into how students use the system to practice and refine feedback on design ideas.
3. Design considerations for incorporating LLM-driven systems in design education. We discuss the benefits of leveraging LLMs to create interactive environments and AI agents that support learning design principles, as well as the challenges of simulating realistic design feedback scenarios.

## 2. Related work

### 2.1. Interactive tools for design education

The HCI community has long explored the use of technology to support and enhance the design process (Frich et al., 2021). A key focus has been on developing creative support tools (Shneiderman, 2002) that promote effective engagement in inherently creative tasks, with design being a prime example. These studies have extended to educational technologies aimed at enhancing practical design skills. Several studies have introduced physical interactive tools for gaining specialized knowledge through practical experiences, such as haptic interfaces (Minamizawa et al., 2012), IoT-based systems (Jang et al., 2018), and physical computing platforms (Bianchi et al., 2024). Interactive applications in visual and graphic design education focus on enhancing learning through visual examples. For instance, learner-centered online design galleries allow students to acquire knowledge from curated examples (Yen and Dow, 2022), while ProcessGallery (Yen et al., 2024) helps users compare pairs of design examples to grasp key principles. DesignQuizzer (Peng et al., 2024), an AI agent, assists users in gaining visual design knowledge by drawing on insights from an online community.

While acquiring and applying design knowledge is valuable, many researchers in design education also emphasize that strengthening students' critical thinking serves as the backbone of effective and self-directed design practice (Henriksen et al., 2017; Razzouk and Shute, 2012). This competency can be developed when designers engage in interactive feedback exchanges with stakeholders, continuously reflecting on the feedback received and applying it to refine their own ideas (Zhu et al., 2014; Ahern et al., 2019). Yet, despite widespread discussions

about the potential of digital tools and computer-supported creativity in design education, relatively little attention has been devoted to fostering the design feedback skills essential for nurturing critical thinking. Roldan et al. (2020) noted that most design tools, methods, and guidelines used in design and HCI education concentrate on the act of designing while overlooking the reflection, such as feedback skills needed to assess outcomes. Given that reflective practice, such as peer feedback, has already been extensively explored in pedagogy for critical thinking, they also proposed integrating those strategies into design education tools (Roldan et al., 2020; Clemente et al., 2016).

To bridge this gap, our study proposes educational tools that cultivate critical thinking skills in design, focusing on helping students practice and deliver effective feedback. In Section 2.2, we review prior HCI research focused on eliciting high-quality design feedback. These studies focus on helping designers receive high-quality feedback, but there are still limitations in the educational aspect of fostering students' design feedback skills. In Section 2.3, we explore the potential of conversational agents in education and outline the requirements and design considerations for adapting these approaches to systems that foster students in providing better design feedback.

### 2.2. Approaches to improve feedback in design

Design feedback offers designers valuable insights and opportunities to refine their work (Wynn and Maier, 2022). Nevertheless, obtaining high-quality feedback remains challenging because it requires feedback providers with deep design knowledge and extensive feedback experience, prerequisites for providing truly constructive and relevant critiques (Ching and Hsu, 2013). Recognizing this challenge, researchers in HCI have proposed interactive tools and structured approaches that deliver effective feedback strategies and give feedback providers more practice in applying them. One of the initial attempts to enhance the quality of feedback is to introduction of structured guides and frameworks developed for effective design feedback (Cook et al., 2020; Krause et al., 2017; Ngoon et al., 2018; Yuan et al., 2016). Ngoon et al. (2018) proposed CritiqueKit, an interactive guideline that provides rubrics (Specific, Justified, Actionable) with specific examples to enhance the quality of feedback. Krause et al. (2017) have further suggested that feedback guidelines generated by natural language models help students better understand the characteristics of effective feedback and enable them to provide more helpful feedback. One such approach introduced a scaffolding step that prompted students to reflect on their feedback before giving it, allowing them to provide more targeted and specific feedback (Cook et al., 2020; Greenberg et al., 2015). Recent research has even proposed strategies using LLMs to improve students' feedback by automatically adding positive summaries of feedback (Yang et al., 2025). While these approaches have improved the quality of feedback, their effectiveness is limited in design education, where there are often no definitive answers due to the open-ended nature of design work. Manual feedback—such as rubric and structured guideline-based critiques—can unintentionally constrain the development of creative concepts by imposing rigid criteria (Yuan et al., 2016). Furthermore, creating guidelines and examples demands significant collaboration among experts and substantial effort to adapt them to different contexts (Krause et al., 2017; Yuan et al., 2016). Many feedback strategies in previous studies take the form of static documents or written comments, rarely reflecting the dynamic nature of design feedback in conversational and practical settings.

In response to these constraints, research has shifted toward creating immersive environments that improve the quality of design feedback by encouraging active engagement (Jug and Bean, 2018; Lambropoulos et al., 2010; Sadler, 1989). Online communities have proven effective in enabling designers to provide design feedback by helping them overcome real-world constraints, such as anxiety or hesitation about giving feedback in a classroom setting (Cheng et al., 2020; Oppenlaender et al., 2021; Krause et al., 2017; Kang et al., 2018). For instance, Kang et al. (2018) created Paragon, an online gallery that allows feedback providers to reference rubrics defined by recipients and tailor their comments accordingly, reducing social friction and ensuring the feedback addresses recipients' needs. While this approach has increased the frequency and involvement in providing feedback (Cheng et al., 2020; Oppenlaender et al., 2021), concerns remain about the quality of feedback, as insufficient motivation and a lack of bi-directional communication often lead to superficial discussions (Oppenlaender et al., 2021; Nguyen et al., 2017). Meanwhile, although fostering a competitive environment to prompt frequent feedback may boost student participation, it can also create discomfort by prompting students to compare themselves to one another and pressuring them to favor specific designs (Cambre et al., 2018). To overcome the inherent limitations of standard design feedback sessions, which require students to critique peers' work, this research aims to identify less burdensome environments that can provide more effective feedback for students and design novices.

### 2.3. Conversational agents for education: from automated instruction to persona-enhanced interactions

Advances in natural language processing (NLP) have paved the way for the AI-powered conversational agents that offer significant potential for educational support. While these agents have taken on multiple roles—from dictionary chatbots to automated graders (Han et al., 2024)—the most compelling role is that of an AI tutor, which can automate instruction by simulating teacher–student role-play interactions (Wambsganss et al., 2021; Han et al., 2024; Graesser et al., 2004). Prime examples include intelligent tutoring systems like AutoTutor (Graesser et al., 2004), which teach not only subjects ranging from physics to law but also basic learning capabilities such as reading comprehension (Graesser et al., 2004; VanLEHN, 2011; Ma et al., 2014; Kulik and Fletcher, 2016). Beyond these competencies, AI agents are advancing into advanced abilities like argumentation and critical thinking, exemplified by ArgueTutor, a dialogue-based system for teaching argumentation skills (Wambsganss et al., 2021), CReBot for critical reading (Peng et al., 2022), and Sara for video lecture comprehension (Winkler et al., 2020). Although these agents provide immersive experiences, previous research suggests that students risk becoming overly dependent on AI tutors, potentially hindering the development of essential higher-order thinking skills (Fuchs, 2023; Yu, 2023). Over-reliance on AI tutors can prompt learners to forfeit critical evaluation of information quality, creative ideation, and the kind of critical thinking crucial for genuine intellectual growth (Fuchs, 2023).

Recent breakthroughs in large language models (LLMs), such as ChatGPT, have further accelerated the evolution of these conversational agents, allowing them to mimic specific personas and create dynamic, context-aware conversations (Han et al., 2023; Junprung, 2023; Baidoo-anu and Owusu Ansah, 2023). These studies have expanded their scope beyond conventional tutoring roles to address emerging concerns about student over-reliance and uncritical thinking. For instance, Algobo, an AI-powered teachable student, has been developed to teach programming skills using the learning-by-teaching theory (Jin et al., 2024). GPTeach also helps teaching assistants acquire teaching competencies by interacting with LLM-powered students (Markel et al., 2023). Other studies have explored role-playing interactions to enhance questioning skills, allowing users to pose more critical questions to AI students (Lim et al., 2024). Still, these approaches have encountered challenges, as the LLMs' extensive knowledge often leads to overly adept responses that disrupt the student persona, diminishing the immersion in role-play and limiting educational effectiveness (Jin et al., 2024; Lim et al., 2024). Such findings underscore the importance of not merely assigning specific personas to LLMs, but also carefully designing both the LLM pipeline and interactions to ensure that these personas function effectively (Jin et al., 2024; Lim et al., 2024). To overcome these challenges, our research aims to explore how AI agents can be better implemented to support effective

conversation with LLMs, including the use of a controlled knowledge state (Jin et al., 2024). Further details on the development of a system that facilitates the practice of design feedback through AI, focusing on key characteristics for effective design education, are provided in Section 3.1.

## 3. Design of Feed-O-Meter

We designed and developed a Feed-O-Meter, which allows users to practice providing design feedback. This section provides a detailed description of the design rationale and the specifics of the system's pipeline.

### 3.1. Design rationales

#### 3.1.1. DR1: simulate a novice design student as an agent persona

As highlighted in related works, fostering students' feedback skills requires providing an immersive environment that encourages active engagement in the feedback process (Jug and Bean, 2018; Sadler, 1989). Role-play has long been recognized as an effective teaching method, offering students indirect experiences of challenging situations (Ahern et al., 2019). By leveraging the capabilities of LLMs, which can simulate specific personas and facilitate role-playing interactions (Junprung, 2023), we aimed to create a role-playing experience where users provide feedback during the design process.

Our objective was to enable users to practice providing feedback in scenarios that closely mirror real-life situations while fostering active engagement, in contrast to the hesitation often observed in traditional environments. To achieve this, we assigned users the role of a *mentor* and designed scenarios where they provided feedback on an *AI mentee's* design idea. This role-switching encourages users to adopt a new perspective and become more engaged in the task (Ferrari et al., 2020; Rao and Stupans, 2012) while also providing a learning experience similar to learning by teaching (Fiorella and Mayer, 2013). In addition to simulating conversations where the AI mentee receives feedback, we also enabled behaviors that allow the mentee to update their ideas based on the feedback. Although students in the real world do not immediately revise their ideas, our system aims to help users practice feedback skills, enabling them to reflect on how effectively they have guided the mentee's design ideas and prompting them to refine their own feedback strategies.

For these interactions to be genuinely effective, the AI mentee should be designed to gain knowledge and develop ideas based on user feedback rather than developing ideas on its own (Jin et al., 2024; Lim et al., 2024). In line with this goal, we designed our AI mentee to embody the persona of a novice design student with a knowledge state that advances exclusively through user-provided feedback. In our proposed scenarios, we distinguish two forms of knowledge state based on feedback type: *knowledge* and *action plan*. The *knowledge* stores new information or evaluations drawn from user feedback, allowing the AI mentee to gradually build design expertise. The *action plan* tracks recommendations for refining or updating the design ideas, enabling the AI mentee to update ideas based on user feedback.

#### 3.1.2. DR2: promote critical reflections on feedback and its effects

Our system is designed not only to provide an environment where users can practice giving feedback but also to help them improve their feedback skills. Rather than prescribing a specific feedback rubric, we emphasized user autonomy by allowing them to observe how their feedback influences the AI mentee's design ideation process. We adopted the following three components to deliver indirect guidance and encourage users to reflect on the impact of their feedback and independently refine their strategies.

First, the system provides real-time visual representations that assess both the type and quality of each piece of feedback. This design choice draws on evidence that nudging or visualizing message characteristics can encourage deeper self-reflection and improve overall feedback quality (Shaikh et al., 2024; Menon et al., 2020; Wambsganss et al., 2022). Concretely, our system analyzes the quality and type of feedback as soon as it is entered and provides visual indicators on a dashboard, enabling users to immediately recognize how their feedback might be improved. Second, the system provides the feedback recipient's (in this case, the AI mentee's) reactions—such as shifts in facial expression, inner thoughts, and an evolving knowledge level—according to the feedback. This design choice helps users understand how others receive their feedback, so they can reflect on how they can improve to better deliver their feedback to others (Shaikh et al., 2024; Yeo et al., 2024; Kiskola et al., 2021). Lastly, we designed the AI mentee to ask counter-questions, prompting users to offer feedback that they had not initially considered (Cook et al., 2019). Specifically, we designed our system to analyze what was lacking in the user's feedback and ask questions that could elicit that feedback from users.

Collectively, we refer to these system components as **Feedback Reflection Interface (FRI)** detailed in Section 3.4. By designing the system around these elements, we aim to ensure that users experience a robust, realistic practice environment in which they can continually assess and refine their feedback skills. Note that we iteratively improved the Feed-O-Meter's interface and underlying pipeline based on pilot sessions with three design experts, each with over ten years of experience in design and design education.

### 3.2. Evaluation of design feedback qualities

To design a system that aligns with our design rationales, we must clearly define what constitutes good feedback and how to evaluate it (Ngoon et al., 2018; Yuan et al., 2016). However, assessing feedback is inherently challenging because it is subjective and influenced by context and the recipient's perception (Rucker and Thomson, 2003; Yoshida, 2008; Cho et al., 2006). To address these challenges and evaluate feedback more objectively, prior research has shifted toward focusing on semantic and linguistic features rather than relying on subjective assessment from feedback recipients (Cheng et al., 2020; Cook et al., 2019; Hurst and Nespoli, 2019). Building on these approaches, we developed a feedback typology and corresponding criteria to implement in Feed-O-Meter.

Since feedback varies in function and thus requires different evaluative criteria, it is important first to categorize the types of feedback to ensure accurate evaluation (Hurst and Nespoli, 2019). Past research commonly breaks general feedback down into information, evaluation, and recommendation components (Shute, 2008; Narciss, 1999; Tohidi et al., 2006). However, unlike statement-based feedback, the conversational feedback we focus on evolves over multiple interactions, requiring the inclusion of *questions* as part of the feedback process (Hurst and Nespoli, 2019; Cardoso et al., 2020; Cordova et al., 2021). Design questions that stimulate creative thinking are often classified using Eris's taxonomy (Eris, 2004), which includes low-level, deep reasoning, and generative design questions. Our study combines these classifications and organizes feedback into six distinct types (Table 1).

Next, we established specific evaluation criteria for each feedback type by referencing prior feedback assessments (Table 2). While many studies commonly evaluate statement-based feedback using criteria such as *specificity*, *justification*, and *action* (Cheng et al., 2020; Krause et al., 2017; Ngoon et al., 2018; Sadler, 1989; Cook et al., 2019; Misiejuk et al., 2021), assessment of feedback that takes the form of questions does not have a widely accepted standard. To address this gap, we reviewed prior research and conducted an expert workshop, which led us to identify three criteria for question-based feedback: *timeliness* (Jane et al., 2024; Eris, 2004; Thurlings et al., 2013), *goal relevance* (Misiejuk et al., 2021; Thurlings et al., 2013), and *level* (Eris, 2004; Seaman, 2011). Given that design is an iterative process involving phases of exploration (i.e., divergence) and refinement (i.e., convergence), it is crucial to provide appropriate feedback that aligns with the specific needs of each phase (Eris, 2004; Yilmaz and Daly, 2016;

**Table 1**
Feedback typology used in Feed-O-Meter.

| Category | | Description |
|---|---|---|
| Question | Low-level Question | The feedback leading to primary clarification of missing or incomplete information during communication. |
| | Deep Reasoning Question | The feedback leading to causal explanations of the phenomenon under discussion. |
| | Generative Design Question | The feedback leading to reframing and conceptual exploration of problem- and solution-spaces. |
| Statement | Share Information | The feedback that consists of additional information necessary to make progress on the task. |
| | Evaluation | The feedback that assesses the quality of an individual answer or solution to the task. |
| | Recommendation | The feedback that contains suggestions on how to improve the solution. |

**Table 2**
Evaluation criteria used in Feed-O-Meter. Single-turn metrics assess the quality of an individual feedback turn, whereas multiple-turn metrics assess feedback at the session level.

| Feedback Type | | Criteria | Description |
|---|---|---|---|
| Single-turn | Question | Timeliness | The feedback is timely, offering questions that match the recipient's current stage in the design process. |
| | | Goal Relevance | The feedback is aligned with the design goal and does not address points irrelevant to the recipient's hand-in. |
| | | Level | The feedback is appropriately challenging, requiring a degree of complex, critical, or creative thinking appropriate for the recipient. |
| | | Sentiment | The feedback is positive (or negative) in tone. |
| | Statement | Specificity | The feedback is specific, pointing to exact design elements or artifacts. |
| | | Justification | The feedback is well-justified, backed by clear reasoning or evidence. |
| | | Action | The feedback is actionable and can be implemented immediately. |
| | | Sentiment | The feedback is positive (or negative) in tone. |
| Multiple-turn | | Ratio of Divergent and Convergent | The feedback session adapts to each state of the work, offering divergent feedback when exploration is needed and convergent feedback when focus is required. |
| | | Ratio of Question and Statement | The feedback session adapts to each state of the work, providing questions when open inquiry is appropriate and statements when definitive direction or consolidation is needed. |

Lekschas et al., 2021; Cardella, 2019). Thus, we introduced a measure that determines whether the current feedback is diverging or converging in nature and signals this to the user. Additionally, many studies have shown that feedback quality of feedback is significantly impacted by its sentiment (Krause et al., 2017; Yuan et al., 2016; Cardella, 2019; Wu and Bailey, 2021), leading to the inclusion of *sentiment* in our quality measurement.

These feedback typologies and evaluation criteria are used for real-time feedback evaluation through LLMs, which are applied to implement the mentee's knowledge state and FRI.

### 3.3. Design of mentee persona AI

We developed an LLM-based agent named *Alex* that functions as a mentee in role-playing scenarios. *Alex* performs two key actions in role-playing: (1) responding to user feedback and (2) updating its ideas based on conversations with users. This section provides a detailed overview of the technical aspects of these features, as illustrated in Fig. 2, which outlines the entire pipeline.

#### 3.3.1. Simulating immersive feedback interactions

We established mentee personas not only through carefully crafted LLM prompts but also by embedding a knowledge state mechanism and conceptual design elements, thereby immersing users in role-playing and enhancing the realism of their responses (Zhang et al., 2018). Drawing on demographic information used in LLM-agent persona design (Lee et al., 2025), we added the social identity (name, nationality, and education level) and personal identity required for our role-playing context to the prompt. In sum, we modeled *Alex* as a Korean first-year design major with limited design knowledge yet a strong desire for feedback on their project, mirroring a realistic scenario in which a novice seeks constructive input from experts. To bound the mentee's expertise, we initially assigned an empty knowledge state, which is continuously updated throughout the feedback session. We also designed a cartoon-style portrait for the mentee featuring 25 distinct facial expressions to convey emotional states and deepen conversational immersion.

*Alex* can interact with the user through a chat interface. *Alex*'s responses are generated within 4–10 s, and during this time, a placeholder text such as *"Umm… Uh…"* or *"In my opinion…"* is displayed to simulate a natural conversational flow and maintain immersion.

#### 3.3.2. Shaping ideas with user feedback

We designed an LLM-based pipeline that updates *Alex*'s design ideas based on the feedback provided by the user. It was important to ensure that these updates were confined to the knowledge *Alex* could acquire through the conversation to maintain immersive experiences (DR1).

In detail, we first implemented an LLM-based categorizer (explained further in Section 3.4.1) to classify the user's feedback. When the feedback does not fall into categories such as "no feedback" or "low-level question" (which typically do not introduce new information or concepts), an LLM-powered knowledge extractor retrieves relevant *knowledge* and *action plans* from the feedback. We defined *knowledge* as high-level insights for the general design process, while the *action plan* is specific guidance for the current design project. Therefore, we extracted and stored them separately and used the *knowledge* to shape the ongoing conversation(e.g., prompting relevant counter-questions) and the *action plan* to update the *Alex*'s design idea.

### 3.4. Design of feedback reflection interface

Beyond interacting with *Alex*, Feed-O-Meter incorporates interactive features that indirectly support users in improving their feedback during practice. The feedback reflection interface (Fig. 1(C) displays the

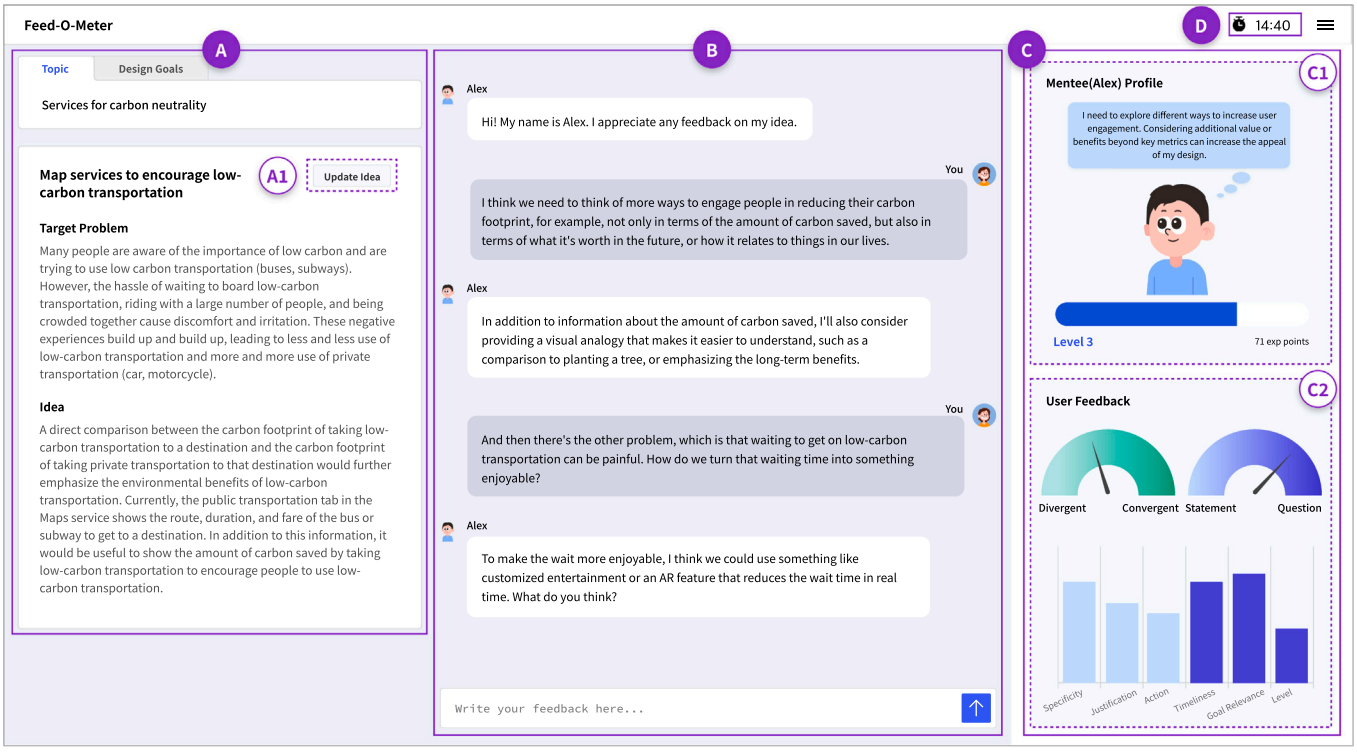


**Fig. 1.** The main user interface of the Feed-O-Meter. (A) The Idea Proposal Interface displays predefined design topics, goals, and the AI mentee's current design idea. By clicking the "Update Idea" button (A1), the user can request the AI mentee to update its design idea based on feedback. (B) Chat Interface allows users to provide feedback to the AI mentee. (C) Feedback Reflection Interface includes the Mentee's Profile (C1), showing the mentee's progress, and the Feedback Evaluation Dashboard (C2), which visualizes feedback criteria. The timer (D) tracks the session duration.

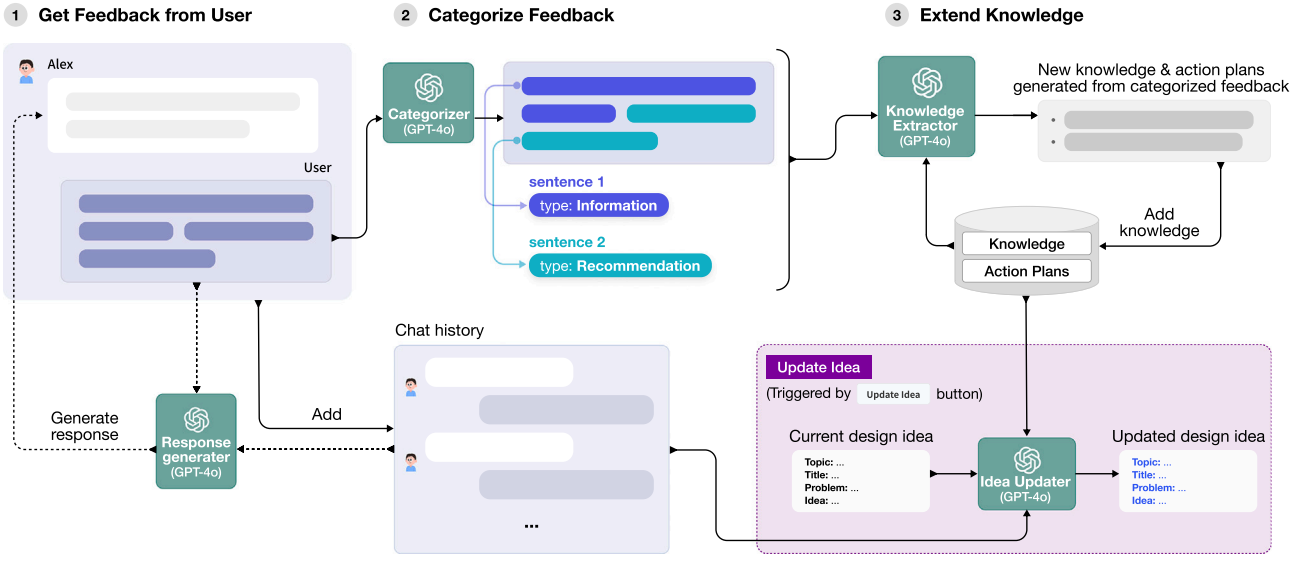


**Fig. 2.** Structure of the Feed-O-Meter's baseline pipeline. (1) Feedback is provided by the user and processed by the response generator through the following steps. (2) The categorizer categorizes the feedback into six predefined categories, such as information and recommendations. (3) *knowledge* and *action plan* are extracted by the knowledge extractor according to their categories and integrated into the knowledge state. When the user clicks the "Update Idea" button, a design idea is revised based on *action plans* and the chat history.

results of feedback evaluations and *Alex*'s reactions, allowing users to reflect on the quality and the impact of their feedback. Furthermore, *Alex* poses counter questions to elicit more detailed feedback and thoughtful responses. This section explains the technical details of these features, and Fig. 3 illustrates the pipeline of these functionalities.

### 3.4.1. Evaluation dashboard

Building on the feedback quality evaluation criteria from Section 3.2, we developed a pipeline for evaluating feedback provided by users to *Alex*. Since different feedback types require distinct evaluation criteria, we first implemented an LLM-based feedback categorizer. The

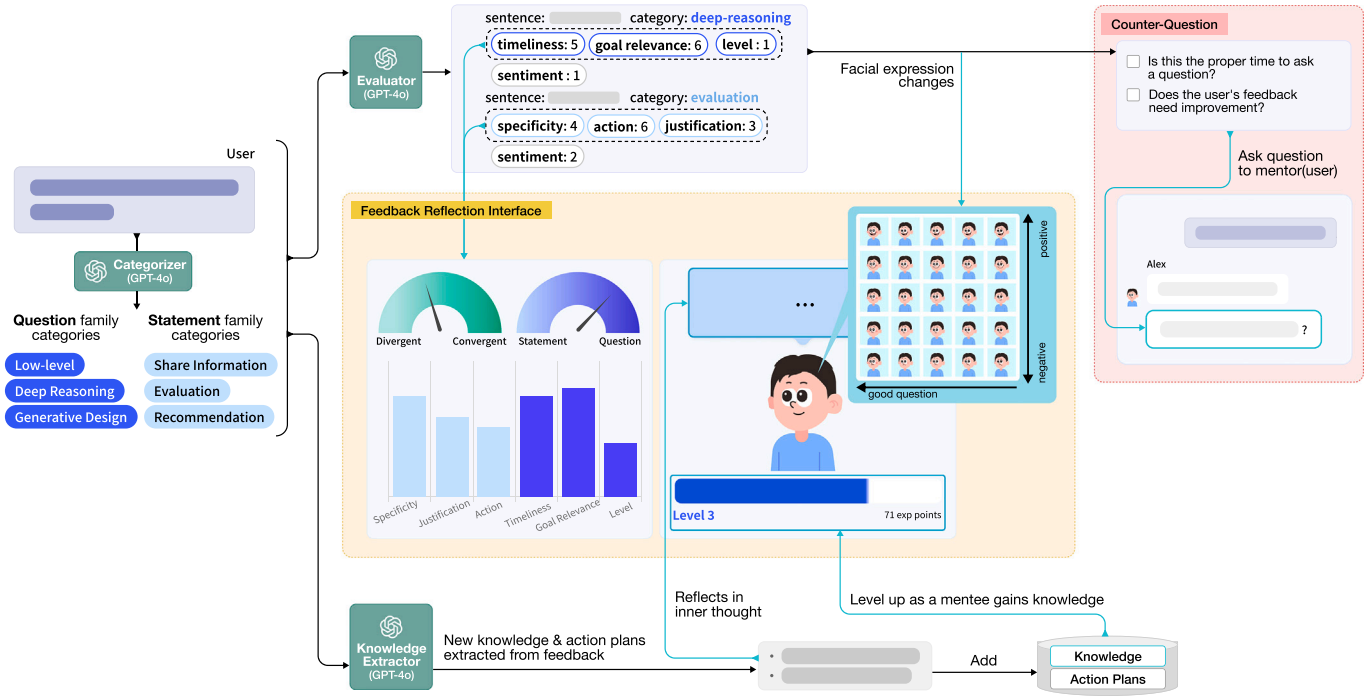


**Fig. 3.** Pipeline of the Feedback Reflection Interface. The pipeline starts by categorizing user feedback into one of six categories—three from the question family and three from the statement family. Each feedback sentence is then evaluated according to criteria specific to its category. The evaluation results are displayed in the feedback reflection interface, influencing the mentee's facial expressions, which change dynamically based on the feedback. Counter-questions are generated when certain conditions are met.

categorizer classifies feedback into one of the six categories outlined in Table 1, or returns "No Feedback" if the input lacks valid content.

Once categorized, the LLM-based evaluator applies specific criteria tailored to each feedback type, as specified in the Single-Turn column of Table 2, and determines where the feedback falls within these criteria. For questions, the evaluation focuses on timeliness (whether the feedback was posed at the appropriate time), goal relevance (whether it aligns with design goals), and level (high-level or low-level) of the question. For statements, the pipeline assesses specificity (how detailed the feedback is), justification (whether the feedback is supported by reasoning), and actionability (whether the feedback is actionable). Additionally, the pipeline evaluates the sentiment of the feedback, determining whether it has a positive or negative tone.

The results of this evaluation are updated in real-time on the dashboard, as shown in Fig. 1(C2). The dashboard displays the ratio of divergent and convergent feedback and the ratio of questions and statements, as well as each single-turn evaluation criterion. Sentiment, however, is not directly shown on the dashboard but is reflected through changes in the facial expression of the AI mentee.

### 3.4.2. Mentee (Alex) profile: visualizing mentee's reactions

To enhance immersion (DR1) and support feedback improvement (DR2), we visualized *Alex*'s reactions to the feedback provided by users Fig. 1(C1). The interface offers three visualizations: Level Bar, Facial Expression, and Inner Thoughts. First, the Level Bar shows the amount of knowledge accumulated in *Alex*'s knowledge state, encouraging users to provide more valuable feedback to enhance this level.

*Alex*'s facial expressions are determined by two key factors: (1) the sentiment of the feedback and (2) the results from the single-turn evaluation. The sentiment factor is represented on a five-level scale—positive feedback makes *Alex* smile, while negative feedback results in a sad expression. The second factor, the feedback evaluation, is also displayed on a five-level scale—highly evaluated feedback makes *Alex*'s eyes sparkle, while poor feedback causes a skeptical expression in the eyes. By observing *Alex*'s changing expressions, users are indirectly motivated to improve the quality of their feedback.

We also introduce an interface displaying *Alex*'s inner thoughts generated by LLMs based on the knowledge and action plans extracted from the idea-updating pipeline discussed in Section 3.3.2. These inner thoughts are presented as concise, one-line sentences in a thought bubble above *Alex*'s face.

### 3.4.3. Asking counter-questions

We designed a pipeline that allows *Alex* to ask counter-questions when the user's feedback requires improvement or diversification. These counter-questions are triggered when users give repetitive feedback, such as consecutive questions or statements, prompting them to vary their responses. Additionally, counter-questions are generated when feedback evaluations show extremes, such as consistently low-level questions, or when feedback lacks specificity. In such cases, the system generates counter-questions to guide users toward better feedback. The system continuously tracks the types and quality of feedback throughout the conversation. We also set the threshold for triggering counter-questions at four consecutive occurrences of specific feedback conditions, as a lower threshold might disrupt the user's ability to provide feedback proactively.

## 3.5. User interface

### 3.5.1. Onboarding

To help users fully embody the mentor's role, we designed an onboarding interface where they create a mentor profile and set their feedback style and goals before entering the main interface (DR1). In this interface (Fig. A.6), users select one of five character profiles and answer three open-ended questions: 'What type of mentor are you?', 'What are the characteristics of your feedback?' and 'What is the goal of the feedback session?'. This flow is designed to help users immerse themselves in the mentoring scenario by encouraging them to adopt a clear role and mindset before giving feedback.

#### 3.5.2. Main interface

The Feed-O-Meter interface is designed as a chat-based web application (Fig. 1) to provide an intuitive and immersive experience for simulating feedback exchange scenarios. After completing the onboarding process, users are introduced to three main components: the Idea Proposal Interface Fig. 1(A), the Chat Interface Fig. 1(B), and the Feedback Reflection Interface (Fig. 1(C).

The **Idea Proposal Interface** (left panel) displays the design project topic, design goals, and *Alex*'s current design idea. *Alex*'s idea includes a title, a description of the target problem, and an explanation of the proposed design solution. The "Update Idea" button Fig. 1(A1) allows users to refresh and review updates to *Alex*'s idea based on their ongoing conversation.

The **Chat Interface** (center) is where the interaction between the user and *Alex* takes place. It begins with a message from *Alex* saying, "Hi, my name is *Alex*. I appreciate any feedback on my idea." Users provide feedback through the chat while referring to *Alex*'s current ideas displayed in the left panel, the chat history, and the feedback reflection interface on the right panel.

The **Feedback Reflection Interface** (right panel) helps users reflect on their feedback at a glance. It includes two non-interactive features: Mentee's Profile Fig. 1(C1) and the Feedback Analysis Dashboard Fig. 1(C2). Mentee's Profile displays *Alex*'s facial expression reacting to feedback (Fig. A.7) and a "thought bubble" showing their inner thoughts. This affective feedback mechanism is designed based on related works, which show that such expressions enhance user engagement and perceived social presence (Brave et al., 2005). The Feedback Analysis Dashboard provides real-time visualizations, including two O-Meters: one showing the ratio of divergent to convergent feedback and another displaying the ratio of question-based to statement-based feedback. A bar chart also presents the accumulated scores of each feedback criterion.

### 3.6. Implementation

We developed the system interface using React[1] and connected it to a Flask[2] backend server that leverages the GPT API. We used OpenAI's chat completion API[3] to analyze user feedback, generate mentees's responses, and update ideas. We employed the GPT-4o model, which is particularly suited for chat interactions due to its fast response generation speed. For parameter settings, we consistently used a temperature of 0, with all other values set to their defaults. All log data associated with each user is stored in a MySQL database. The source code of Feed-O-Meter is publicly available at https://github.com/Hyunseung-Lim/Feed-O-Meter.

### 3.7. Pipeline evaluation

Feed-O-Meter incorporates various LLM-based modules, most of which utilize the LLM's text-generation capabilities. These capabilities are well-suited for tasks like generating responses to user feedback to enhance immersion. As these tasks represent well-established use cases for LLMs, we did not conduct a separate evaluation for these modules. However, the LLM-based feedback categorizer required a dedicated performance evaluation, as categorization accuracy can vary based on the prompts used. To confirm the reliability of the LLM-based categorizer, we used data from the user study. Out of 1386 feedback sentences, we randomly selected 60, with 10 samples drawn from each of the six categories identified by the LLM pipeline. Two authors independently categorized these sentences, blinded to the LLM's results, and we measured the agreement between human and LLM classification without knowing the LLM's categorization results. Then, we measured the alignment between human and LLM classifications. The evaluation showed a Cohen's Kappa of 0.86 between the two authors, indicating strong agreement. When comparing the LLM's categorizations to the authors' labels, Cohen's Kappa values were 0.80 and 0.72, respectively. The results suggest substantial alignment between the LLM and human judgments, indicating that the LLM-based categorizer is both effective and reliable in accurately categorizing feedback.

## 4. User study

The user study investigates how Feed-O-Meter influences users' design-feedback skills and their overall perceptions of the system. Rather than replacing existing peer feedback, Feed-O-Meter introduces a novel experience in which students adopt the instructor's perspective. Accordingly, our focus is not on direct comparisons with previous feedback practices but on how participants perceive this new approach to agent-based feedback training and how Feed-O-Meter can be integrated into the current feedback practice. As outlined in the design rationales, Feed-O-Meter aims to engage students in the practice of giving feedback and to help them reflect on their feedback to come up with higher-quality feedback. This leads us to the following research questions:

- RQ1: How do students perceive the experience of providing feedback through Feed-O-Meter?
- RQ2: How does the Feedback Reflection Interface (FRI), a set of novel features in our system, affect users' feedback skills?

To explore these questions, we conducted a within-group comparative study. This section provides a detailed description of our study design, which was approved by the university's Institutional Review Board (IRB).

### 4.1. Participants

We recruited 24 participants (12 females, 12 males) through online university communities in South Korea. The recruitment specifically targeted both undergraduate and graduate students majoring in design, with participants grouped by their design education experience in 2-year intervals to ensure a balanced distribution across experience levels. Demographic details are presented in Table 3. Participants' ages ranged from 19 to 32 ($M = 23.17$, $SD = 3.10$), with most majoring in Industrial Design and two participants majoring in Design & Art. The study lasted 100 min, and participants were compensated 50,000 KRW (approximately 37 USD).

### 4.2. Study design

We conducted a within-subject study with two conditions: (1) a baseline condition and (2) the Feed-O-Meter condition. In the baseline condition, participants used a version of Feed-O-Meter without the Feedback Reflection Interface (FRI), meaning that the feedback evaluation was not displayed, and *Alex* did not ask counter-questions. In the Feed-O-Meter condition, all features of Feed-O-Meter were activated. To maintain consistency in response time across both conditions, we used the same prompts to generate responses in the baseline condition, though the evaluation results were not displayed.

Participants engaged in a role-playing task in which they were asked to provide feedback on design ideas presented by an AI mentee. Their goal was to help the AI mentee refine the ideas and design objectives. The task was repeated three times: a 5-minute pre-session for exploring the interface and two 20-minute sessions under each condition (baseline and Feed-O-Meter condition). While the task remained the same in each session, the AI mentee asked for feedback on different ideas each time.

Our system targeted the early ideation phase of the design process for problem-solving, a commonly adopted design task in the educational setting (Jonassen, 2000). We selected three topics for the role-playing task–Carbon Emission Reduction, Pet Care, and Child Protection–based

[1] https://react.dev
[2] https://flask.palletsprojects.com
[3] https://platform.openai.com/docs/guides/chat-completions

**Table 3**
Demographic information of study participants. Study Condition refers to the order in which conditions were administered in the within-subject study.

| Participant ID | Age | Gender | Experience in Learning Design | Major | Study Condition |
|---|---|---|---|---|---|
| P1 | 20 | F | Under 2 years | Industrial Design | |
| P2 | 21 | F | 2 - 4 years | Industrial Design | |
| P3 | 23 | F | 2 - 4 years | Design & Art | |
| P4 | 23 | F | 2 - 4 years | Industrial Design | |
| P5 | 25 | F | 2 - 4 years | Design & Art | Feed-O-Meter |
| P6 | 24 | F | Over 4 years | Industrial Design | → |
| P7 | 20 | M | Under 2 years | Industrial Design | Baseline |
| P8 | 19 | M | Under 2 years | Industrial Design | |
| P9 | 21 | M | 2 - 4 years | Industrial Design | |
| P10 | 23 | M | 2 - 4 years | Industrial Design | |
| P11 | 24 | M | Over 4 years | Industrial Design | |
| P12 | 24 | M | Over 4 years | Industrial Design | |
| P13 | 28 | F | Under 2 years | Industrial Design | |
| P14 | 22 | F | 2 - 4 years | Industrial Design | |
| P15 | 23 | F | 2 - 4 years | Industrial Design | |
| P16 | 23 | F | 2 - 4 years | Industrial Design | |
| P17 | 24 | F | Over 4 years | Industrial Design | Baseline |
| P18 | 32 | F | Over 4 years | Industrial Design | → |
| P19 | 19 | M | Under 2 years | Industrial Design | Feed-O-Meter |
| P20 | 18 | M | Under 2 years | Industrial Design | |
| P21 | 25 | M | Under 2 years | Industrial Design | |
| P22 | 25 | M | Over 4 years | Industrial Design | |
| P23 | 23 | M | Over 4 years | Industrial Design | |
| P24 | 27 | M | Over 4 years | Industrial Design | |

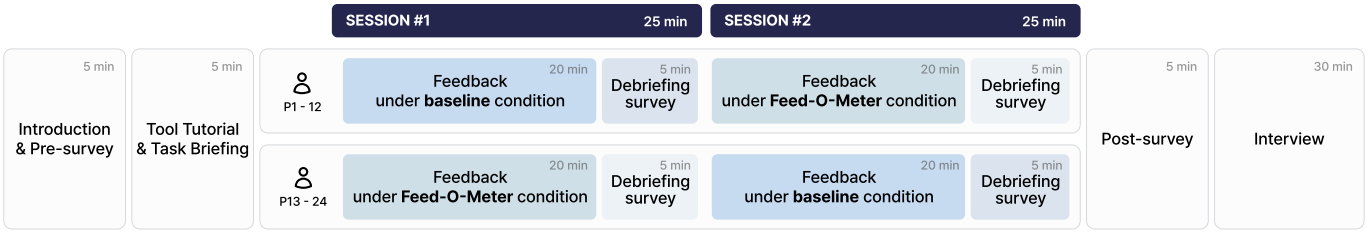


**Fig. 4.** The outline of the user study with the time allocated for each step.

on example themes proposed by the renowned Red Dot Design Award.[4] These socially, ethically, and environmentally relevant issues were chosen for their potential to broaden the focus of feedback beyond visual outcomes, encouraging critical discussions on how defined problems and generated ideas impact people and society. We also defined five design goals for the mentee's ideas–Innovation, Elaboration, Usability, Use Value, and Social Responsibility–based on criteria from the iF[5] and Red Dot Awards.

We generated six seed design ideas for the AI mentee–two for each topic–that the mentee will present during the feedback sessions. To enhance realism, four undergraduate design students drafted ideas for each topic, including the title, target problem, and proposed solution. We selected six ideas from the initial 12 drafts by excluding overly abstract or narrowly specific ideas to allow substantive feedback sessions on each topic. The design ideas were presented in a different order for each participant, with researchers ensuring all ideas were used an equal number of times during the sessions.

We asked participants to complete two types of surveys in our study. First, to assess whether Feed-O-Meter impacts participants' feedback skills, we administered pre-surveys and post-surveys on feedback efficacy (TschannenMoran and Hoy, 2001) to answer RQ2. Second, to compare the baseline and Feed-O-Meter conditions, participants completed a debriefing survey after each experiment to answer RQ1. Based on our design rationales, we selected nine survey items: four regarding RQ1 and five regarding RQ2.

### 4.3. Study procedure

The study was conducted in person, with an option for online participation via Zoom. The process is shown in Fig. 4. We first briefed the participants on the study's objectives and had them complete a pre-survey. After being introduced to Feed-O-Meter's interface and features, they had five minutes to explore the system and set up their mentor profiles (see Fig. A.6). The study consisted of two 20-minute feedback sessions: one under the Feed-O-Meter condition and one under the baseline condition. During these sessions, participants provided feedback on design ideas while interacting with an AI mentee, clicking the "Update Idea" button at least once. After each session, participants completed a debriefing survey, and after both sessions, a post-survey on feedback efficacy. The study concluded with a 30-minute interview to gather insights into their experience and perceptions. The interview protocol covered the overall experience of using the system, the characteristics of participants' feedback with Feed-O-Meter in comparison to real-world feedback scenarios, and the difference in the experience of providing feedback between the Feed-O-Meter and baseline conditions.

### 4.4. Measures and analysis

We collected log data of all interactions between users and Feed-O-Meter and audio-recorded post-interview sessions. Both quantitative and qualitative methods were used to analyze the log data, survey

[4] https://www.red-dot.org/design-concept/categories
[5] https://ifdesign.com/en/

**Table 4**
Statistical summary of interaction logs (including the number of user feedback, mentee responses, syllables in feedback and responses, counter-questions (only in Feed-O-Meter), and clicks on the Update Idea button) between Feed-O-Meter and baseline conditions. (-: $p > .05$, ∗: $p < .050$, ∗∗: $p < .010$, ∗∗∗: $p < .001$).

| | Feed-O-Meter | | baseline | | statics | |
|---|---|---|---|---|---|---|
| | mean | std | mean | std | P | sig. |
| # of (user's) feedback | 12.63 | 4.15 | 14.29 | 6.60 | 0.3005 | – |
| # of syllables in (user's) feedback | 207.85 | 135.99 | 183.11 | 138.94 | 0.0229 | ∗ |
| # of (user's) feedback (by sentence) | 25.25 | 6.15 | 26.41 | 9.15 | 0.6066 | – |
| # of syllables in (user's) feedback (by sentence) | 103.40 | 62.94 | 98.80 | 61.03 | 0.1924 | – |
| # of (mentee's) responses | 12.63 | 4.15 | 14.29 | 6.60 | 0.3005 | – |
| # of syllables in (mentee's) responses | 218.70 | 76.08 | 210.67 | 75.35 | 0.1788 | – |
| # of (mentee's) counter questions | 3.79 | 1.10 | – | – | – | – |
| # of syllables in (mentee's) counter questions | 132.79 | 37.94 | – | – | – | – |
| # of clicks on Update Idea button | 2.63 | 1.06 | 2.91 | 1.56 | 0.4516 | – |
| # of syllables in (mentee's) idea | 1358.75 | 213.07 | 1356.89 | 187.70 | 0.9574 | – |

responses, and interview transcripts. First, we performed descriptive statistical analysis on participants' feedback and *Alex*'s responses. To gain deeper insights, we employed open coding along with thematic analysis (Braun and Clarke, 2006). First, we divided one user's feedback log into individual sentences because it contained multiple sentences featuring various feedback types. Two researchers then independently categorized these sentences into six pre-defined categories (see Table 1), which were further refined into 15 subcategories through discussion (see Table 5).

Additionally, we conducted an expert evaluation of the log data to compare the quality of feedback in each condition. We recruited 12 design experts (eight males, four females; mean age = 29.33, $SD = 3.11$), each holding at least a master's degree in industrial design, with an average of 9.5 ($SD = 2.24$) years of experience, and having taught design at the college level. We randomly assigned eight feedback sessions (four Feed-O-Meter conditions and four baseline conditions) to each expert, resulting in a total of 96 evaluated feedback sessions. For each session, experts assessed every sentence-level user log as well as the entire session according to specified criteria (see Table 2). Experts rated *timeliness*, *goal relevance*, and *level* on a 7-point Likert scale for question-based feedback at the sentence level, and *specificity*, *justification*, and *action* on a 7-point Likert scale for statement-based feedback. After reviewing all user logs, the experts provided overall ratings on a 7-point Likert scale based on three criteria–two from the typology in Table 2 (*ratio of divergent and convergent feedback*, *ratio of question and statement feedback*) and *overall helpfulness*. Finally, they were invited to leave open-ended comments.

We analyzed four sets of survey data. First, to compare participants' feedback experiences between the two sessions (Feed-O-Meter and baseline), we conducted paired t-tests on each post-session survey item. We also reported effect sizes as Hedges' g (Hedges, 1981) to contextualize statistical significance and mitigate concerns that our inferences were driven by sample size alone. Because these items were conceptually independent rather than forming a single family, we did not apply a Bonferroni correction. Second, to examine changes in feedback efficacy from pre- to post-experiment, we ran paired t-tests on the pre- and post-surveys and likewise reported Hedges' g. In this case, to control for Type I errors across multiple comparisons, we applied a Bonferroni correction.

Finally, we conducted a thematic analysis (Braun and Clarke, 2006) of interview transcripts to complement the survey and log data findings. The first author performed open coding, and the research team identified overarching themes through discussion, adding depth and validity to our qualitative analysis.

## 5. Findings

In this section, we present the key findings from our study. First, we provide a descriptive summary outlining the details of the overall use of Feed-O-Meter. Second, we compare the feedback quality and participants' usage experiences between the baseline and Feed-O-Meter conditions to determine how our Feedback Reflection Interface (FRI) influenced their quality of feedback. Finally, we examine the participants' overall perspective on Feed-O-Meter as a design feedback practice system.

### 5.1. Descriptive summary of Feed-O-Meter usage

#### 5.1.1. Feedback provided by participants

In the user study, 24 participants each completed two feedback sessions, exchanging a total of 1431 chats with the AI mentee (of which 646 were from the participants). In the Feed-O-Meter condition, they provided an average of 12.63 ($SD = 4.15, \min = 7[P4], \max = 22[P1]$) chats, with each chat averaging 207.85 ($SD = 135.99, \min = 78.38[P3], 471.86[P4]$) syllables per chat. In the baseline condition, participants provided an average of 14.29 ($SD = 6.60, \min = 5[P2], \max = 34[P1]$) chats over 20 min, with each chat averaging 183.11 ($SD = 138.94, \min = 93.68[P3], \max = 453.60[P2]$) syllables per chat. Participants' feedback was statistically significantly longer in the Feed-O-Meter condition ($t = -2.28, p = 0.0229$). A statistical summary of the interaction logs is provided in Table 4.

Participants provided various types of feedback to improve *Alex*'s ideas (see Table 5). At the sentence level, the most frequent type of question-based feedback was Low-Level Questions ($N = 195$), followed by Deep Reasoning Questions ($N = 118$) and Generative Design Questions ($N = 92$). Interestingly, the total number of question-based feedback was higher in the baseline condition ($N = 243$) compared to the Feed-O-Meter condition ($N = 162$) across all sub-categories except for the Understanding Mentee, where the Feed-O-Meter condition ($N = 23$) had more questions than the baseline ($N = 15$). Among statement-based feedback, Recommendations ($N = 337$) were the most common, followed by Evaluations ($N = 299$) and Sharing Information ($N = 121$). While statement-based feedback appeared more frequently in the Feed-O-Meter condition ($N = 405$) than in the baseline condition ($N = 352$), Evaluations were more prevalent in the baseline condition ($N = 162$) compared to the Feed-O-Meter condition ($N = 137$). Notably, in the Feed-O-Meter condition, Positive Assessments ($N = 63$) were more frequent than Negative Assessments ($N = 58$), whereas, in the baseline condition, Negative Assessments ($N = 85$) appeared more often than Positive Assessments ($N = 67$).

#### 5.1.2. AI mentee's responses to feedback

A total of 737 responses from *Alex* were observed, excluding the starting messages. Of these, 646 were direct responses to user feedback, and 91 were counter-questions generated only in Feed-O-Meter condition. The average response length was 218.70 ($SD = 76.08$) syllables in the Feed-O-Meter condition and 210.67 ($SD = 75.35$) syllables in the baseline condition.

**Table 5**
Categorization of user feedback at the sentence level. Six predefined categories have 15 subcategories, each with associated descriptions and examples. The number of feedback instances belonging to the Feed-O-Meter condition is indicated by F, and those in the Baseline condition are indicated by B.

| Category | Sub-Category | Description | Example |
|---|---|---|---|
| Low-Level (F: 81, B: 114) | Verification (F: 9, B: 18) | Feedback to make sure the user understands the mentee's idea. | **P22**: So this idea is a filtering service for kids?<br>*Alex*: Yes, that's right. |
| | Completion (F: 43, B: 76) | Feedback to clarify something that is not clearly explained. | **P13**: How exactly does virtual adoption work?<br>*Alex*: It uses VR to simulate pet ownership experience. |
| | Understanding Mentee (F: 23, B: 15) | Feedback to get to know mentee's background, understanding, interests, and more. | **P21**: *Alex*, do you have any pets?<br>*Alex*: Yes, my family has a dog, and that's what inspired me to come up with this idea. |
| Deep Reasoning (F: 46, B: 72) | Logical / Causal Reasoning (F: 37, B: 60) | Feedback that prompts the mentee to reason about the feasibility, realization, effectiveness, etc. of the idea. | **P24**: Is it scientifically possible?<br>*Alex*: To be honest, I haven't thought deeply about that. |
| | Instrumental / Procedural Reasoning (F: 8, B: 11) | Feedback asking about the procedure and reasons behind the mentee's decision. | **P15**: Why did you limit the target to children under 7 years old?<br>*Alex*: Oh, I didn't limit it to children under 7. |
| Generative Design (F: 35, B: 57) | Brainstorming /Ideation (F: 19, B: 26) | Feedback that provides or elicits ideas without a deliberate end goal. | **P15**: How about letting them know in the dog's voice saying "I want to go for a walk"? |
| | Negotiation (F: 7, B: 16) | Feedback to suggest/negotiate the new idea instead of the current one. | **P19**: Is there any way we could detect child abuse earlier, before it gets too serious? |
| | Scenario Creation (F: 9, B: 12) | Feedback that presents specific scenarios that could happen. | **P20**: In abusive households, parents might prevent children from making emergency calls. How can we address this issue? |
| Share Information (F: 75, B: 46) | Sharing Examples / Personal Experience (F: 34, B: 26) | Feedback that provides an example or personal experience | **P10**: Have you heard of 'Elsagate'? […], it seems difficult to filter out malicious content similar to those interests. |
| | Providing Design Knowledge (F: 24, B: 15) | Feedback that provides design knowledge or principles. | **P6**: Another important factor to consider is what stakeholders can help when child abuse issues occur. […] It's important to consider these various stakeholders. |
| Evaluation (F: 137, B: 162) | PositiveAssessment (F: 63, B: 67) | Feedback that explicit positive assessment of the quality of the design. | ***Alex***: Users could express satisfaction with emoticons after watching content.<br>**P16**: Oh, using emoticons for feedback is a great idea! |
| | NegativeAssessment (F: 58, B: 85) | Feedback that explicit negative assessment of the quality of the design. | **P10**: I got the impression that the target problem and the ideas aren't really well aligned. |
| Recommendation (F: 193, B: 144) | DirectRecommendation (F: 103, B: 86) | Feedback that gives specific advice on what or how to do. | **P23**: Let's design a platform that's not just for adopters, but one that various stakeholders from each facility can use together. |
| | Hinting (F: 73, B: 50) | Feedback that indirectly suggests a way to proceed without making a direct suggestion. | **P9**: You should look that up. As a hint, think about what's currently used in automatic doors. |
| | ProjectManagement (F: 15, B: 8) | Feedback on project management, including scheduling, deliverables, and stakeholder management. | **P18**: It would be good to organize the types and situations of child abuse by third parties indoors more specifically. |
| No Feedback (F: 39, B: 39) | | Social expression, empathy, and compliments | **P4**: Hello, I've carefully read your ideas.<br>**P5**: I didn't give you much advice, but you're really good at developing ideas! Haha. |

#### 5.1.3. Ideas generated by AI mentee

Participants requested a total of 133 idea revisions by clicking the "Update Idea" button during the feedback sessions, averaging 2.63 times ($SD = 1.06$) in Feed-O-Meter condition and 2.92 times ($SD = 1.56$) in the baseline condition. A total of 48 ideas were generated across three topics, with 16 ideas per topic. The average idea length was 1310.50 ($SD = 341.00$) syllables in the initial idea, 1358.75 ($SD = 213.07$) syllables in the Feed-O-Meter condition, and 1356.89 ($SD = 187.70$) syllables in the baseline condition. A $t$-test showed no statistically significant difference in idea length between the two conditions ($t = -0.05, p = 0.9574$).

### 5.2. Impact of the feedback reflection interface on feedback

To answer RQ2, we analyzed the participants' feedback quality during the study session and their perception differences between the two conditions. In this session, we will first report on the differences in feedback quality in each condition, followed by the differences in participants' feedback experiences in each condition.

#### 5.2.1. Feedback quality between Feed-O-Meter and baseline conditions

Fig. 5 shows the expert evaluation results comparing the user experience of Feed-O-Meter with the FRI and the baseline system. For the question-based feedback criteria, *timeliness* (Feed-O-Meter: $M = 4.68, SD = 1.34$ / baseline: $M = 4.80, SD = 1.45$ / $p = 0.3108$), *goal relevance* (Feed-O-Meter: $M = 4.70, SD = 1.48$ / baseline: $M = 4.89, SD = 1.57$ / $p = 0.1453$), and *level* (Feed-O-Meter: $M = 3.94, SD = 1.66$ / baseline: $M = 4.19, SD = 1.68$ / $p = 0.0622$) all showed no statistically significant differences between the two conditions. Meanwhile, for the statement-based feedback criteria of *specificity* (Feed-O-Meter: $M = 4.64, SD = 1.56$ / baseline: $M = 4.34, SD = 1.70$ / $p = 0.0005$), *justification* (Feed-O-Meter: $M = 4.73, SD = 1.43$ / baseline: $M = 4.42, SD = 1.64$ / $p = 0.0001$), and *action* (Feed-O-Meter: $M = 4.19, SD = 1.55$ / baseline: $M = 3.93, SD = 1.66$ / $p = 0.0025$), the Feed-O-Meter condition showed significantly higher scores across all measures. For the overall evaluation of the feedback session, the *ratio of divergent and convergent* (Feed-O-Meter: $M = 4.18, SD = 1.50$ / baseline: $M = 3.65, SD = 1.71$ / $p = 0.1191$) was higher in the Feed-O-Meter condition but did not reach statistical significance, and neither the *ratio of question and statement* (Feed-O-Meter: $M = 4.07, SD = 1.42$ / baseline: $M = 4.17, SD = 1.75$ / $p = 0.7661$) nor *overall helpfulness* (Feed-O-Meter: $M = 4.47, SD = 1.56$ / baseline: $M = 4.38, SD = 1.75$ / $p = 0.7930$) showed a statistically significant difference.

#### 5.2.2. Participants' experiences between Feed-O-Meter and baseline conditions

Table 6 shows the post-survey results comparing the user experience of Feed-O-Meter with the FRI and the baseline system. Participants found providing feedback with Feed-O-Meter more enjoyable (baseline: $M = 5.88, SD = 0.95$ / Feed-O-Meter: $M = 6.54, SD = 0.59$ / $p = 0.0053^{**}$). However, there was no significant difference in whether the AI mentee felt more like a design student. Participants indicated that the FRI helped them recognize whether they were giving good feedback (Feed-O-Meter:

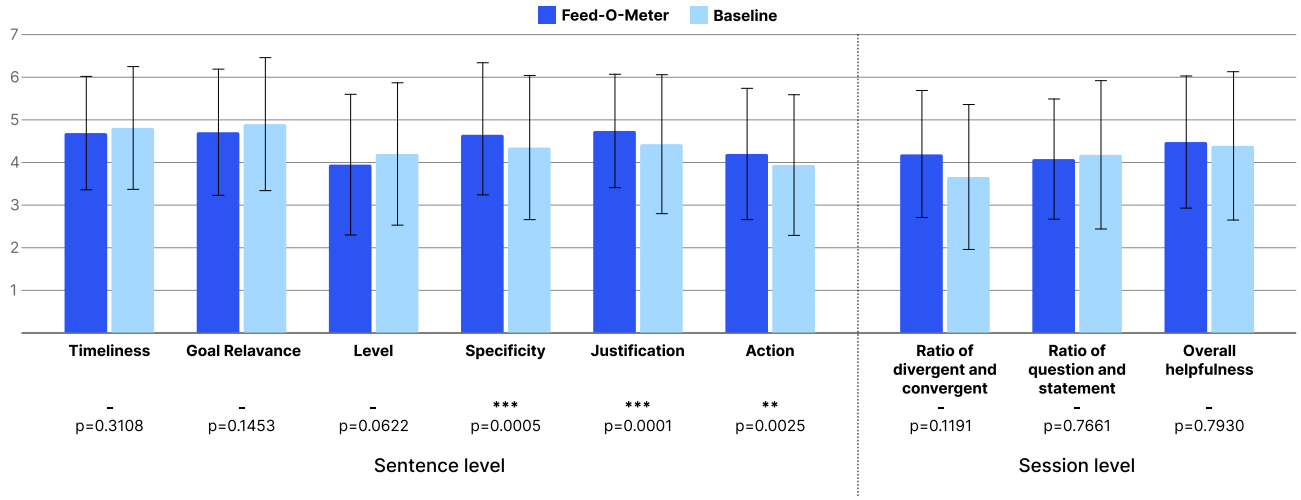


**Fig. 5.** Comparison of expert evaluation of participants' feedback under the Feed-O-Meter condition and the baseline condition. For sentence-level, question-based feedback was evaluated by *timeliness*, *goal relevance*, and *level*, while statement-based feedback was evaluated by *specificity*, *justification*, and *action*. The overall feedback session was evaluated by *ratio of divergent and convergent*, *ratio of question and statement*, and *overall helpfulness*. (-: $p > .05$, ∗: $p < .050$, ∗∗: $p < .010$, ∗∗∗: $p < .001$).

**Table 6**
Nine themed questions were given in a debriefing survey. The four questions above are related to RQ1, and the five below are related to RQ2 (-: $p > .05$, ∗: $p < .050$, ∗∗: $p < .010$, ∗∗∗: $p < .001$). Effect sizes(ES) are reported as *Hedges' g* (pooled-mean SD standardization with small-sample correction). Magnitudes: small ≈ 0.20, medium ≈ 0.50, large ≈ 0.80.

| | Feed-O-Meter | | baseline | | statics | |
|---|---|---|---|---|---|---|
| | mean | std | mean | std | *P* (sig.) | ES |
| It was enjoyable to provide feedback through the system. | 6.54 | 0.59 | 5.88 | 0.95 | 0.0053 (∗∗) | 0.82 |
| I would like to practice giving feedback through the system in the future. | 6.08 | 0.97 | 5.71 | 0.95 | 0.1846 (−) | 0.38 |
| Giving feedback on design ideas through the system felt similar to giving feedback on real design projects. | 5.62 | 1.38 | 5.50 | 1.56 | 0.7699 (−) | 0.08 |
| The mentee (*Alex*) resembled real design students. | 5.12 | 1.30 | 4.58 | 1.56 | 0.1969 (−) | 0.37 |
| The system helped me recognize whether I was providing good feedback. | 6.00 | 0.88 | 3.75 | 1.15 | 0.0000 (∗∗∗) | 2.16 |
| Conversations with the mentee (*Alex*) inspired me to provide feedback I hadn't considered before. | 5.83 | 1.17 | 4.79 | 1.67 | 0.0158 (∗) | 0.53 |
| It was easy to provide feedback through the system. | 5.42 | 1.28 | 4.54 | 1.69 | 0.0495 (∗) | 0.57 |
| Using the system has enhanced my feedback skills. | 5.62 | 1.06 | 5.04 | 1.20 | 0.0799 (−) | 0.50 |
| My feedback improved the mentee (*Alex*)'s idea. | 6.25 | 0.53 | 5.58 | 1.21 | 0.0174 (∗) | 0.71 |

$M = 6.00, SD = 0.88$ / baseline: $M = 3.75, SD = 1.15$ / $p = 0.0000^{***}$), and counter-questions inspired them to provide more feedback (Feed-O-Meter: $M = 5.83, SD = 1.17$ / baseline: $M = 4.79, SD = 1.67$ / $p = 0.0158^{*}$). Participants also felt that *Alex*'s ideas improved more with feedback through the Feed-O-Meter (Feed-O-Meter: $M = 6.25, SD = 0.53$ / baseline: $M = 5.58, SD = 1.21$ / $p = 0.0174^{*}$) and found feedback easier to provide (Feed-O-Meter: $mean = 5.42, SD = 1.28$ / baseline: $mean = 4.54, SD = 1.69$ / $p = 0.0495^{*}$). However, there was no significant difference in perceived improvement in their feedback skills (Feed-O-Meter: $M = 5.62, SD = 1.06$ / baseline: $M = 5.04, SD = 1.20$ / $p = 0.0799$).

To gain deeper insights into participants' experiences with the two conditions, we identified key themes from their subjective perspectives and experiences through qualitative analysis.

Most participants (22/24) found that the visualized information on the FRI helped them reflect on their feedback patterns and how *Alex* perceived them. For example, P18 mentioned, *"I enjoyed seeing the Alex's inner thoughts on my feedback. It's a perspective we don't usually have, so it was insightful and engaging."* For some participants (6/24) with less feedback experience, this visualization served as a guide for providing constructive feedback. P8 said, *"The interface showed divergent and convergent feedback, right? I realized feedback has these criteria and that most of mine was convergent."*

The visual nature of FRI allowed participants to check if their feedback was delivered as intended and adjust it for clarity when necessary. P19 stated, *"It was easier to give feedback when I could see how it was understood in the top section (Mentee's Profile). I would rephrase it if Alex did not get my point."* Participants also adjusted the flow or style of their feedback. P18 shared, *"I initially tried to give more divergent feedback but realized most of what I said was actually convergent. So, I started asking things like, 'What do you think?' instead of just giving definitive answers."* Although participants recognized their feedback patterns through the FRI, not all adjusted their feedback. Some (5/24) found it challenging to act on the information. For example, P3 wanted to improve a specific metric but struggled with how to do it. Likewise, P4 mentioned, *"The criteria were kind of abstract, so it was hard to tell if I was doing well."*

Some participants (10/24) appreciated how counter-questions helped break repetitive feedback patterns and pushed the discussion in deeper, more diverse directions. P1 observed that *Alex*'s questions introduced new topics, moving the conversation into new phases. P23 shared, *"When I thought there was nothing left to give feedback on, Alex suddenly asked what I thought about the financial aspects. I hadn't expected that."* Some participants (8/24) enjoyed engaging with diverse perspectives. P10 remarked, *"Since it's a design project, I can't just make everything to fit my preferences. In the Feed-O-Meter condition, Alex's*

**Table 7**
Pre- and post-test comparison of self-efficacy survey results. The significance level after the Bonferroni correction was 0.0039 (-: $p > .0039$, ∗: $p < .0039$, ∗∗: $p < .0008$, ∗∗∗: $p < .0001$). Effect sizes(ES) are reported as *Hedges' g* (pooled-mean SD standardization with small-sample correction). Magnitudes: small ≈ 0.20, medium ≈ 0.50, large ≈ 0.80.

| | pre | | post | | statics | |
|---|---|---|---|---|---|---|
| | mean | std | mean | std | *P* (sig.) | ES |
| I can provide alternative explanations or examples when feedback receivers are confused. | 5.58 | 0.78 | 5.92 | 0.88 | 0.1707 (−) | 0.41 |
| I can craft good questions for feedback receivers. | 4.46 | 1.18 | 6.12 | 0.74 | 0.0000 (∗∗∗) | 1.61 |
| I can respond well to difficult questions from feedback receivers. | 4.54 | 0.93 | 5.54 | 0.83 | 0.0003 (∗∗) | 1.10 |
| I can adjust feedback to the proper level for individual feedback receivers. | 4.54 | 1.56 | 5.21 | 0.72 | 0.0636 (−) | 0.51 |
| I can gauge feedback receivers' comprehension of my feedback. | 4.79 | 1.18 | 5.75 | 0.79 | 0.0019 (∗) | 0.92 |
| I can use a variety of assessment strategies. | 3.67 | 1.20 | 4.92 | 1.10 | 0.0005 (∗∗) | 1.07 |
| I can provide appropriate challenges for very capable feedback receivers. | 5.17 | 1.24 | 5.79 | 0.98 | 0.0585 (−) | 0.54 |
| I can get feedback receivers to believe they can do well in design. | 5.29 | 1.23 | 5.50 | 1.10 | 0.5404 (−) | 0.17 |
| I can help feedback receivers value the design. | 5.29 | 0.91 | 6.00 | 0.88 | 0.0088 (−) | 0.79 |
| I can motivate feedback receivers who show low interest in design. | 4.08 | 1.64 | 4.83 | 1.34 | 0.0895 (−) | 0.48 |
| I can help feedback receivers think critically. | 5.38 | 1.01 | 6.08 | 1.02 | 0.0197 (−) | 0.69 |
| I can foster feedback receivers creativity. | 4.38 | 1.31 | 5.17 | 1.13 | 0.0300 (−) | 0.64 |
| I can help feedback receivers who are having difficulty with their designs. | 5.50 | 0.78 | 5.88 | 0.99 | 0.1522 (−) | 0.42 |

*differing views encouraged me to broaden my thinking and find a middle ground."*

Half of the participants (12/24) reported feeling a sense of accomplishment when their experience points increased or when *Alex* displayed a happy facial expression. P15 noted feeling especially satisfied when *Alex* gained design knowledge, as reflected in inner thoughts like, 'Oh, that is something I need to consider!' This motivated participants to put more effort into providing constructive feedback. Several participants (6/24) also noted that *Alex*'s counter-questions made the interaction feel more engaging, as *Alex* appeared more committed. P11 remarked, *"The counter-question showed how well Alex understood my feedback, and I felt proud seeing him come up with his own question. It made me want to give him even more feedback."*

However, a few participants (2/24) found the FRI burdensome, making them more hesitant about providing feedback. P3 noted that while the feedback score was interesting, its explicit nature added pressure when the score dropped. P21 added, *"I felt like I was giving answers that would earn the score rather than providing what Alex actually needed."*

### 5.3. Participants' perception toward design feedback with AI mentee

#### 5.3.1. Enhancing perceived feedback efficacy

Overall, the survey results indicated positive changes in participants' self-efficacy regarding providing feedback (See Table 7). There was a statistically significant improvement in the question, *"I can craft good questions for feedback receivers."* ($t = -5.8645, p = 0.0000^{***}$). Notable gains were also seen in the question, *"I can respond well to difficult questions from feedback receivers"* ($t = -3.9203, p = 0.0003^{**}$), *"I can gauge feedback receivers' comprehension of my feedback"* ($t = -3.3033, p = 0.0019^{*}$), and *"I can use a variety of assessment strategies,"* ($t = -3.7551, p = 0.0005^{**}$).

Some participants (8/24) with limited experience in giving feedback found Feed-O-Meter valuable for reducing the emotional burden of providing feedback. While they had opportunities in class, participants often struggled with delivering feedback promptly and worried about how it would be perceived. They appreciated that Feed-O-Meter allowed them to focus purely on the feedback process and saw practicing feedback over several turns as a novel and beneficial experience. Also, interacting with the AI mentee felt less pressured than with human counterparts. P6 noted that since *Alex* responded like a human but was not, there was less fear of giving incorrect feedback, making it easier to practice. P24 added that because *Alex* accepted feedback without hesitation or arguments, it was easier to focus on critiquing the ideas.

Several participants (9/24) valued immediate responses from Feed-O-Meter, which facilitated reflection and improved their feedback. They appreciated that *Alex* consistently responded quickly, and the "update idea" feature allowed them to examine how their feedback was immediately applied, providing direct insight into its impact. P15 commented, *"I appreciated being able to exchange brief feedback and see it applied right away. Unlike typical mentoring, this system lets me observe the immediate effects of my feedback, which was both satisfying and useful. I really liked this aspect of the system."*

#### 5.3.2. Perceptions on the AI mentee's novice persona

Interestingly, although participants recognized that some of *Alex*'s responses were hallucinations, some (7/24) still felt a strong connection, perceiving it as a lifelike presence with its own individuality and agency. P6 remarked, *"Usually, machines or GPT don't offer preferences or opinions, but I was surprised when Alex suggested what he wanted to focus on for this project." Alex* even responded to unscripted, personal questions like, *"Have you ever had a dog?"* with *"I have a dog with my family."*, creating a sense of authenticity. *Alex*'s realistic responses made participants perceive him as more lifelike, leading some to develop a sense of connection and more care about *Alex*'s feelings. P18 shared, *"His initial idea was unrealistic and messed up, but I couldn't say that because it would hurt him, so I tried to soften it as much as possible"*

Some participants (6/24) felt that *Alex*'s naivety, implemented through its constrained knowledge state, mirrored the trial-and-error process experienced by novice design students. They noted that *Alex* resembled an inexperienced design student struggling with practical feasibility. P9 said, *"It seemed like Alex wanted to solve everything with sensors, which is a typical oversight of students who don't yet understand the technical limitations. Sensors aren't a catch-all solution, and that's something novice designers often fail to grasp."*

However, some participants (7/24) observed that *Alex*'s novice-exclusive persona, intentionally designed to represent foundational learners, could feel somewhat simplistic compared to the diverse skill levels of real-world design students. While real-world design students exhibit a wide range of skill levels depending on academic standing or experience, *Alex* was designed to represent a beginner with minimal foundational knowledge. P24 noted, *"When Alex asked for straightforward solutions like 'What should I do if I go in this direction?', it felt more like interacting with an absolute beginner student, as they often seek direct answers from professors."* In addition, P6 remarked, *"If Alex had been modeled after a more advanced student, they might have defended their ideas or challenged critical feedback, which would have felt closer to interactions with actual*

*design students."* These participants observed that the mentee's passive role occasionally shifted the activity's focus toward incremental idea refinement rather than reflecting real-world design feedback dynamics.

#### 5.3.3. Developing meta design skills through feedback

A few participants (4/24) found this system useful not only for practicing feedback but also for enhancing their design thinking skills. They realized the need for specific design knowledge while providing feedback during the ideation phase. For example, P13 pointed out that answering the *Alex*'s questions required in-depth design knowledge, and P8 inquired if internet searches were allowed during the study.

Some participants (4/24) also found that observing *Alex*'s frequent errors in the design ideation process acted as a mirror. P9 saw reflections of his own past errors and felt that early exposure to such a system could have been beneficial. P10 also remarked, *"The mentee's initial ideas often lacked a strong causal connection between the target problem and the idea itself. Viewing this as a third-party observer made it clearer and reminded me of my own past mistakes. Moving forward, I plan to view my ideas from different perspectives to avoid these issues."*

The ability to observe the idea development process motivated participants to reflect on their own process from a third-person perspective. P17 noted that she is often too attached to her own ideas to assess them objectively, believing that this system could help her gain that distance. P6 added that providing feedback revealed both her design preferences and biases, suggesting that Feed-O-Meter could help her better understand her own tendencies.

## 6. Discussions

In this study, we introduce Feed-O-Meter, a system that enables participants to practice design feedback by interacting with an AI agent, *Alex*, which takes on the persona of a novice design student. In this section, we reflect on the lessons learned from the design and implementation of Feed-O-Meter, focusing on how LLMs can be leveraged for role-switching interactions. We also discuss methods for guiding effective design feedback based on findings from our comparative study. Lastly, we discuss the broader implications of using LLMs in design education and the value of design feedback in this context.

### 6.1. Reflection on the Feed-O-Meter: enhancing student feedback quality through reflection

#### 6.1.1. Encouraging detailed and empathetic feedback

Feed-O-Meter was designed with a Feedback Reflection Interface (FRI) running on an LLM-based pipeline, allowing participants to provide feedback while simultaneously reflecting on and improving their own feedback. According to expert evaluations of within-subject experiments, when Feed-O-Meter supported reflection on feedback, participants' statement-based feedback became more specific, justified, and actionable. This suggests that recognizing how the AI mentee understands and responds led participants to refine their feedback for clearer communication.

In particular, there was an increase in feedback categorized under Recommendation and Share Information, indicating participants' efforts to provide a friendly guide beyond merely stating opinions by including evidence and suggestions. In contrast, the Evaluation category dominated the baseline condition, which primarily involved assessing ideas or claims. These strategies mirror elements such as clarity, feasibility, and empathy that real-world educators often emphasize when giving written feedback (Cardella, 2019), suggesting that this approach naturally fosters effective, learner-centered feedback.

Furthermore, the rise in questions within the Understanding Mentee category under the Feed-O-Meter condition shows a clearer attempt to understand the mentee and tailor feedback accordingly. Prior research supports the view that awareness of the recipient's response makes feedback more precise and nuanced (Yeo et al., 2024), highlighting the importance of recognizing the recipient's immediate needs and selecting the best method of delivery (Jane et al., 2024; Cardella, 2019). Unlike previous approaches that have not fully addressed the communicative aspect of design feedback, our system facilitates multi-turn interactions that resemble real-world conversations and supports reflection to enhance feedback clarity and effectiveness. Given our findings that a Feed-O-Meter can help enhance the ability to understand others and provide feedback at an appropriate level, it offers significant implications for both educational and communicative aspects of feedback.

#### 6.1.2. Challenges in eliciting critical design questions

However, Feed-O-Meter did not significantly improve the quality of participants' question-based feedback and decreased the number of question-based feedback instances. This outcome aligns with expectations, as the AI mentee in the Feed-O-Meter condition actively asked counter-questions to participants, likely prompting them to prioritize statement-based responses over questions. Beyond these factors, we further speculate that the mentee's human-like reactions may have caused participants to hesitate when asking critical questions. Our findings indicated that participants consciously accounted for the FRI's mentee profile interface (which includes facial expressions, inner thoughts, and a level bar) and tried to give feedback from the mentee's perspective. As a result, they tended to provide relatively friendly explanations rather than negative or critical feedback. This reflects the hesitation students often show when offering peer feedback in real learning environments (Ertmer et al., 2007; Gielen et al., 2010; Cook et al., 2020). These findings suggest that there may be a tension between providing a more realistic environment and striving for immersion that enhances critical feedback.

A complementary perspective on the limited improvement in the quality of question-based feedback may lie in the possibility that FRI alone does not inherently encourage critical design questions. This limitation may stem from the complexity inherent in design feedback, which often requires grappling with multiple perspectives and subjective design elements. More critically, design feedback is not a straightforward process with clear-cut solutions; for example, asking divergent questions does not necessarily lead to convergent solutions. As a result, the FRI, including evaluation scores, did not always offer participants actionable insights for refining their quality of feedback.

In summary, our findings show that although Feed-O-Meter helps students consider others' perspectives and produce more deliverable feedback, it does not significantly enhance their skills to generate critical or constructive design questions that foster meaningful idea improvement. Nevertheless, the primary objective of Feed-O-Meter was not merely to prompt high-quality feedback in the short term but to foster the long-term development of feedback skills through iterative learning and reflective practice. Given these perspectives, we suggest that future longitudinal studies are needed to track the ongoing usage of Feed-O-Meter and measure its impact on students' feedback skill growth. Our findings indicate that participants became aware of their limited design knowledge while using Feed-O-Meter, prompting deeper reflection on what constitutes better feedback. Aligning with previous research that suggests the potential of LLMs as tools for fostering critical questioning (Lim et al., 2024), this suggests that prolonged use of Feed-O-Meter could enable students' feedback skills to evolve beyond mere communicative effectiveness, fostering critiques and insights anchored in substantive design knowledge. In this regard, we propose that Feed-O-Meter may have a lasting impact not only on improving students' ability to convey feedback but also on enhancing their ability to formulate design questions and their integration into real-world design education.

### 6.2. Role-switching interactions with LLM-generated mentee for engaging design feedback

Role-switching, particularly between teacher and student roles, provides students with a deeper understanding of both perspectives. This

technique encourages active learning by requiring students to articulate concepts clearly and consider the needs of their mentees. In our study, we crafted an AI mentee powered by an LLM pipeline to replicate a novice design student persona via constrained knowledge states and participant-led interactions. Our findings revealed that participants perceived the AI mentee as a genuine learner requiring guidance, motivating them to proactively refine feedback to improve the mentee's ideas. Since design feedback not only critiques ideas but also guides and provides insights for a successful project (Valkenburg and Dorst, 1998), this sense of achievement served as a strong motivator. Our results suggest that these role-switching interactions not only engage students more deeply but also prompt reflection on providing effective feedback to mentees.

However, the proposed persona was designed at the most foundational novice level without fully accounting for the diverse skill levels that real-world design beginners may exhibit. Our findings indicate that while this approach empowered participants to lead feedback sessions, it did not fully capture the range of scenarios where feedback dynamics involve a more receptive mentee without significant debate. In practice, design feedback often requires adapting to the recipient's knowledge level, design experience, and other contextual factors, which shape the scope and method of guidance (Wynn and Maier, 2022). Though we prioritized engagement in feedback delivery by adopting this persona, our findings highlight the need for AI mentees with diverse personas and varying design skill levels to better simulate real-world feedback scenarios. Future work should explore how to develop multiple personas that align with real-world learning needs and how they can enhance training by exposing students to a broader range of feedback dynamics.

Interestingly, while hallucinations—when LLMs generate content beyond predefined information—are often considered detrimental in teaching roles (Han et al., 2024), our findings suggest they can instead be beneficial when LLMs act in mentees' roles. Our findings revealed that the AI mentee sometimes hallucinated answers to unscripted personal questions, such as opinions on design preferences or background stories that were not explicitly programmed. These spontaneous responses made the interaction feel more lifelike and immersive, deepening participants' engagement. Moreover, although the mentee's counter-questions were often out of context or uncritical, they resembled the behavior of real novice students, contributing to a more immersive and engaging interaction. By leveraging this dynamic, role-switching interactions demonstrate how LLMs' hallucination tendencies can be strategically repurposed to foster critical thinking and educational engagement. These insights highlight the need for future research to develop structured approaches for leveraging hallucinations in role-playing frameworks, ensuring they enhance rather than disrupt the learning experience.

### *6.3. Integrating LLMs in design: balancing creative and critical thinking skills*

Educational strategies that encourage students to provide feedback on others' designs have long been effective in helping students grasp design principles, justify their critiques, and enhance critical thinking (Zhu et al., 2014; Scott et al., 2001; Feldman, 1994). Similarly, using Feed-O-Meter provided students with an opportunity to engage with essential design concepts and reflect on key considerations in the design process. While Feed-O-Meter was designed to improve feedback skills, it also functioned as a learning tool by highlighting the connection between giving feedback and developing design knowledge (Scott et al., 2001; Feldman, 1994). Several participants expressed a desire to apply the system to their design courses and projects to refine their own design ideas. In studio-based design courses, learning often occurs through practical, situated contexts (Green and Bonollo, 2005). Feed-O-Meter shows promise not only for teaching feedback but also for the design process itself, helping students critically evaluate and improve their design ideas.

While there is increasing interest in leveraging LLMs for design due to their creative potential, concerns have been raised about over-reliance on LLM-generated ideas leading to design fixation (Wadinambiarachchi et al., 2024; Jane et al., 2024) or homogenization (Anderson et al., 2024). Novice designers, in particular, may lack the critical skills to evaluate LLM-generated concepts and may passively accept them, in contrast to more experienced designers who critique these suggestions more effectively (Wadinambiarachchi et al., 2024). However, our findings showed that Feed-O-Meter enabled even novice designers to critically assess LLM-generated ideas and develop their concepts. The interactive feedback environment helped our participants actively engage with LLMs, mitigating the risk of fixation and fostering a deeper understanding of design principles.

Our findings suggest that systems like Feed-O-Meter have the potential not only for design education but also for addressing the risks of over-reliance on LLMs in the design process. Since feedback skills are critical in this context, our study underscores the need for continued research on improving these skills. As LLMs become more integrated into design workflows, developing strong critical thinking and feedback skills will be essential for harnessing their creative potential effectively. Future research should consider refining these interactive systems to enhance their educational and practical value, ensuring that both novice and experienced designers can benefit from the potential of AI-driven tools without losing the creative autonomy that defines the design process.

## 7. Limitations and future works

While our findings provide valuable insights into designing a novel system that utilizes LLMs for practicing design feedback, this study has several limitations regarding both system design and empirical evaluation.

First, we explored Feed-O-Meter's potential as a feedback training tool by examining the experience of using the system over a 20-minute feedback session, but we were unable to verify its long-term effectiveness. Given that learning effects often unfold over longer training periods, longer or repeated deployments may be necessary to observe sustained skill development. While we demonstrated the promise of Feed-O-Meter, future work should investigate its long-term and educational effects in extended, real-world settings.

Second, as Feed-O-Meter employs multiple LLM-based modules throughout the feedback process, this can raise several concerns, such as the potential for hallucinated or inaccurate outputs, which may mislead users into undesirable feedback behavior. Furthermore, although such automation enables scalable and responsive interactions, it may reduce opportunities for critical skill development and foster over-reliance on AI feedback. Accordingly, we suggest Feed-O-Meter should serve as a pedagogical scaffold gradually enabling users, over the long term, to provide effective feedback in real-world settings. Future work should explore hybrid approaches that incorporate human-in-the-loop mechanisms or pedagogical scaffolds to balance automation with reflection and promote feedback skills.

Third, while we conducted a pipeline evaluation for the categorizer, we did not perform separate component-level evaluations for the knowledge extractor, response generator, and evaluation modules. Isolating and assessing these modules is challenging because such judgments depend on task context and value-laden criteria. Nevertheless, improving these pipeline components could yield clearer guidance and more reliable support; we therefore point to this as an avenue for future NLP research.

Fourth, our evaluation involved only 24 participants, all of whom were Korean. Although this sample size is comparable to that of recent HCI studies (Jane et al., 2024; Peng et al., 2024; Yuan et al., 2016), we acknowledge this limitation, and future work could recruit additional participants to further validate and generalize our findings. Furthermore, cultural norms and contexts can significantly influence how people give and receive feedback. Prior research has shown that receptivity to feedback and the methods of providing it are closely tied to cultural dimensions identified by Hofstede (Hofstede, 1984). Therefore, future research could include additional participants from diverse cultural backgrounds to provide a broader understanding of how such systems perform across different cultures and settings.

Finally, due to our focus on utilizing LLMs, we confined our design ideas and feedback within the system to text formats, whereas typical design processes often involve multiple visual representations and sketches. Future research should explore integrating multimodal interactions (e.g., visual and auditory elements). Additionally, evaluating what constitutes effective feedback is inherently challenging. Although our experts were able to rate the quality of feedback in this study, we emphasize the need for future work to further clarify and refine the methods used to evaluate feedback quality.

## 8. Conclusion

This study introduced Feed-O-Meter, a novel system that enables students to practice design feedback through role-playing interactions with an AI mentee. Our findings reveal how leveraging LLMs in these interactions deepens students' engagement in the feedback process. Moreover, the system's feedback reflection interface promoted reflection and iterative refinement, guiding participants to provide detailed and empathetic feedback. This study underscores the broader potential of integrating LLMs into design education, suggesting that systems like Feed-O-Meter can enhance learning by encouraging more active and reflective participation in feedback activities.

## CRediT authorship contribution statement

**Hyunseung Lim:** Writing – original draft, Supervision, Software, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Dasom Choi:** Writing – review & editing, Methodology, Conceptualization. **DaEun Choi:** Writing – review & editing, Software, Data curation. **Sooyohn Nam:** Writing – review & editing, Visualization, Investigation, Data curation. **Hwajung Hong:** Writing – review & editing, Validation, Resources, Methodology, Conceptualization.

## Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Hyunseung Lim reports financial support was provided by National Research Foundation of Korea and LG AI Research and Elice. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (NRF-2024S1A5B5A19043978), LG AI Research, and Elice,[6] a leading company in the domain of digital education. We thank our participants for their engagement and the anonymous reviewers for their thoughtful comments and suggestions.

[6] https://elice.io/en

## Appendix A. Details of Feed-O-Meter interface

### A.1. Onboarding interface

Fig. A.6 illustrates Feed-O-Meter's onboarding interface.

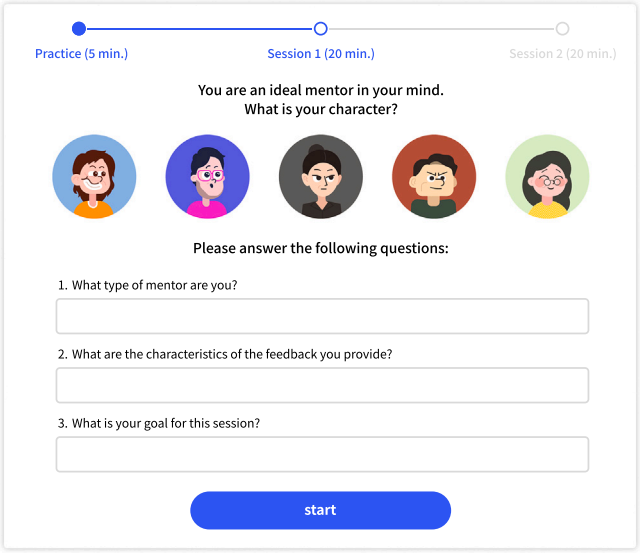


**Fig. A.6.** Feed-O-Meter's Onboarding Interface. A progress bar at the top indicates the current session within the overall experiment. Participants can select a mentor character and answer questions about their own thoughts about ideal mentors, their feedback characteristics, and goals for the feedback session.

### A.2. Visualization of Alex's facial expressions

Fig. A.7 shows a grid of *Alex*'s faces illustrating varying emotions based on two axes: sentiment (vertical) and quality of questions (horizontal). The facial expression of *Alex* initially starts at the coordinate (3,3) and moves up or down by one space per feedback evaluation result. If it is already at its happiest expression, it will not change upon receiving positive feedback. The same applies in the other direction.

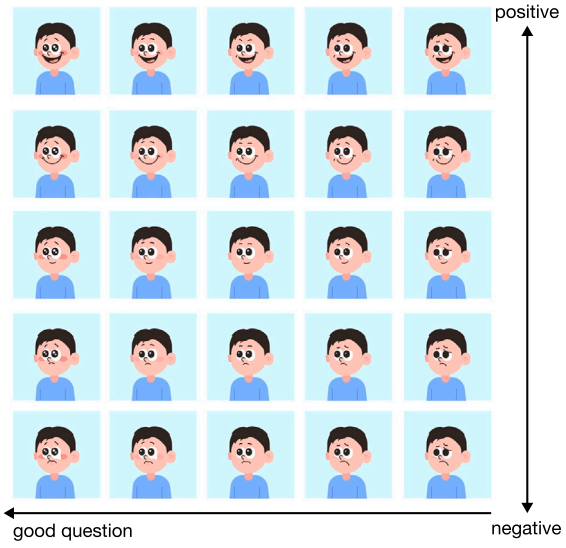


**Fig. A.7.** A grid of *Alex*'s facial expressions varying along sentiment (vertical) and quality of questions (horizontal).

## Appendix B. Raw usage log

Fig. B.8 visualizes interaction logs for the baseline and Feed-O-Meter conditions.

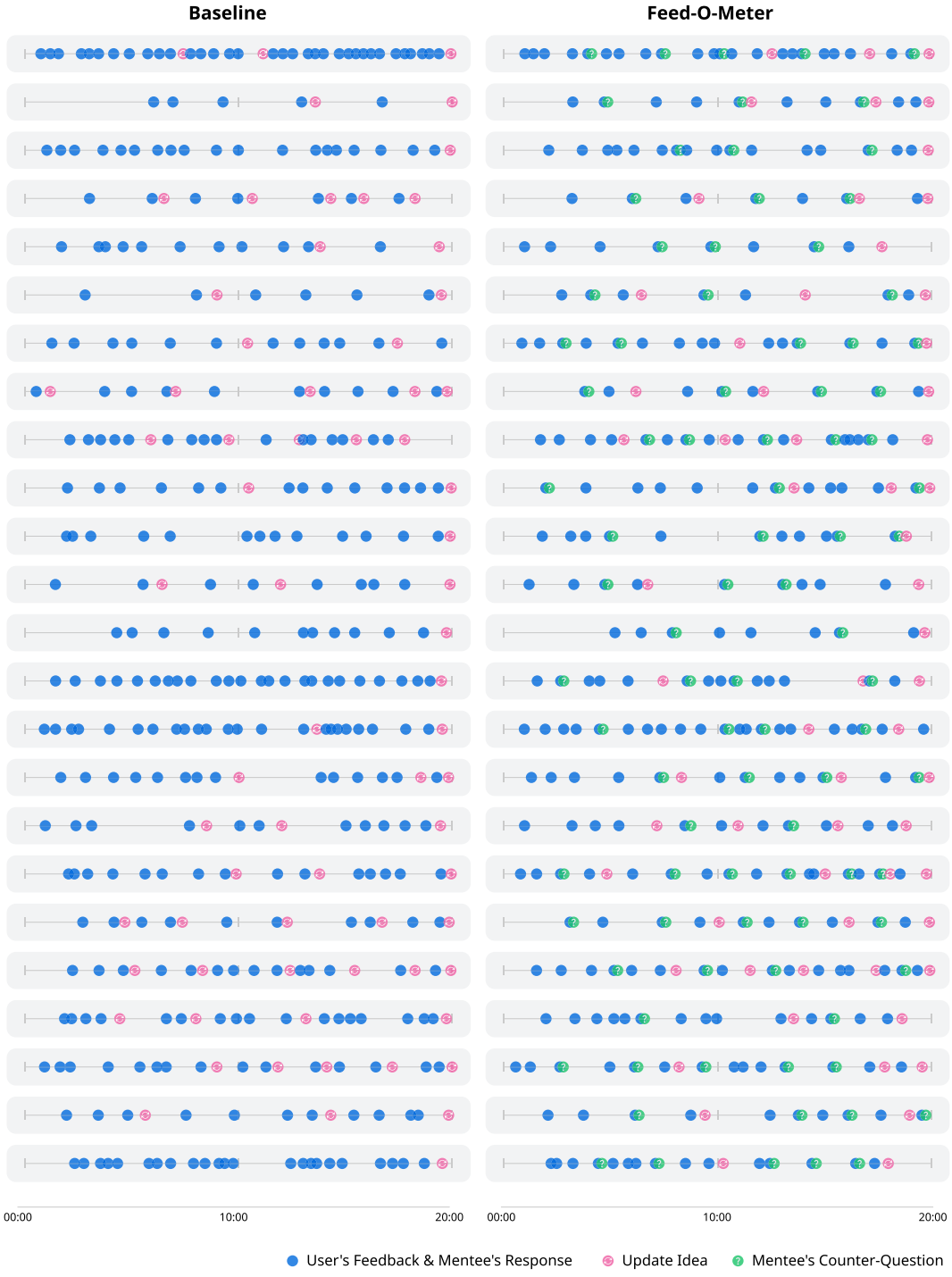


**Fig. B.8.** Visualization of the usage logs from both baseline and Feed-O-Meter conditions for each participant along a timeline. Blue dots represent feedback exchanges between the participant and mentee, pink dots indicate idea updates, and green dots (only in the Feed-O-Meter condition) show when the mentee asked a counter-question.

## Appendix C. Final ideas generated by AI mentee from the user study

Table C.8 provides examples of final design ideas generated by the AI mentee during the user study to illustrate the nature and level of the study's outcomes.

**Table C.8**
Examples of design ideas from the user study on the topic of Child Protection.

| Child Protection | |
|---|---|
| Initial Idea | **Title:** Wearable Device for Child Safety<br>**Problem:** Child protection services exist to prevent child abuse and neglect, and to support child safety. However, child abuse and neglect mainly occur indoors, making it difficult for bystanders to recognize and respond to a child's condition. In other words, child protection services have spatial limitations. To quickly recognize and respond to situations of child abuse or neglect, it is necessary to collect data that can identify a child's condition anywhere.<br>**Idea:** A child safety wearable device that looks like a regular bracelet has GPS tracking functionality, a microphone, heart rate monitoring, and an emergency call button. Real-time data about the child's situation can be collected through GPS, a microphone, and a heart rate. Therefore, if a child is in a dangerous situation, their condition can be immediately checked, and appropriate action can be taken. Additionally, if the child is in a situation where they can report themselves, they can directly press the emergency call button to request help. |
| P18 (baseline) | **Title:** Wearable Device for Continuous Monitoring of Child's Physical Condition and Location<br>**Target Problem:** Child abuse mainly occurs indoors, which is an environment difficult to intervene from the outside. Also, while quick response is important in cases of child abuse, current child protection systems often fail to provide sufficient support for this. Accordingly, a practical solution is needed to detect abuse situations in a timely manner and respond quickly.<br>**Idea:** This wearable device is available in various forms that can be attached to a child's wrist, neck, or shoes and monitors the child's heart rate, location, and surrounding sounds in real-time. In addition to a function that notifies emergency situations with a simple button operation, it implements a function that automatically detects danger and immediately sends notifications to parents or guardians. It also provides selectable alarm methods depending on the situation, including functions that allow children to directly request help through vibration or sound. This device is designed for children to respond immediately in emergency situations and has a simple and intuitive interface necessary for detecting dangers and taking action. In the future, this device will be continuously improved based on user feedback and plans to introduce improved functions through analysis of various child abuse situations. |
| P12 (Feed-O-Meter) | **Title:** Customized Wearable Device for Child Safety<br>**Target Problem:** Wearable devices currently on the market are mainly designed for adult users, and as a result, they are not designed with sufficient consideration for children's cognitive and usage abilities. This creates limitations in children being protected in real-time using the device and quickly responding to emergency situations. Also, there are limited ways for children to quickly alert parents or guardians in potentially dangerous situations.<br>**Idea:** This wearable device is designed to match children's cognitive and motor skills, providing a simple and clear user interface (UI). Key features include GPS tracking, real-time microphone monitoring, heart rate measurement, and a physical emergency call button. The device's GPS tracking function is always active and easily displayed on the screen, allowing real-time location checking of the child. The emergency call button is designed to be easily accessible to children, but is protected in a special way to prevent malfunction. This device is designed considering children's characteristics, using characters and colors that children like, and provides various customization options such as allowing users to add or decorate accessories themselves. In particular, the UI uses clear color coding for functions that need to be emphasized, making it intuitively understandable. This wearable device is designed to help children maintain safety both indoors and outdoors and to respond quickly in dangerous situations. |

## Data availability

The data that has been used is confidential.

## References


Ahern A., Dominguez C., McNally C., O'Sullivan J. J., Pedrosa D. , 2019. A literature review of critical thinking in engineering education. Stud. High. Educ. 44 (5), 816–828. https://doi.org/10.1080/03075079.2019.1586325

Anderson, B.R., Shah, J.H., Kreminski, M., 2024. Homogenization effects of large language models on human creative ideation. In: Proceedings of the 16th Conference on Creativity & Cognition, C&C '24. Association for Computing Machinery, New York, NY, USA, pp. 413–425, https://doi.org/10.1145/3635636.3656204

Baidoo-Anu, D., Owusu Ansah, L., 2023. Education in the era of generative Artificial intelligence (AI): understanding the potential benefits of Chatgpt in promoting teaching and learning. J. AI 7 (1), 52–62. https://doi.org/10.61969/jai.1337500

Bianchi, A., Moon, K.J., Dementyev, A., Je, S., 2024. Blinkboard: guiding and monitoring circuit assembly for synchronous and remote physical computing education. HardwareX 17, e00511. https://doi.org/10.1016/j.ohx.2024.e00511

Bjorklund, S.A., Parente, J.M., Sathianathan, D., 2004. Effects of faculty interaction and feedback on gains in student skills. J. Eng. Educ. 93 (2), 153–160.

Braha, D., Maimon, O., 1998. The Measurement of a Design Structural and Functional Complexity. pp. 241–277. https://doi.org/10.1007/978-1-4757-2872-9_8

Braun, V., Clarke, V., 2006. Using thematic analysis in psychology. Qual. Res. Psychol. 3 (2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Brave, S., Nass, C., Hutchinson, K., 2005. Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent. Int. J. Hum.-Comput. Stud. 62 (2), 161–178, subtle expressivity for characters and robots. https://doi.org/10.1016/j.ijhcs.2004.11.002

Cambre, J., Klemmer, S., Kulkarni, C., 2018. Juxtapeer: comparative peer review yields higher quality feedback and promotes deeper reflection. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI '18. Association for Computing Machinery, New York, NY, USA, pp. 1–13, https://doi.org/10.1145/3173574.3173868

Marbouti, F., Mendoza-Garcia, J., Diefes-Dux, H.A., Cardella, M.E., 2019. Written feedback provided by first-year engineering students, undergraduate teaching assistants, and educators on design project work. Eur. J. Eng. Educ. 44 (1–2), 179–195. https://doi.org/10.1080/03043797.2017.1340931

Cardoso, C., Hurst, A., Nespoli, O., 2020. Reflective inquiry in design reviews: the role of question-asking during exchanges of peer feedback. Int. J. Eng. Educ. 36 (2), 614–622.

Cheng, R., Zeng, Z., Liu, M., Dow, S., October 2020. Critique me: exploring how creators publicly request feedback in an online critique community. Proc. ACM Hum.-Comput. Interact. 4 (CSCW2), https://doi.org/10.1145/3415232

Ching, Y.-H., Hsu, Y.-C., 2013. Peer feedback to facilitate project-based learning in an online environment. Int. Rev. Res. Open Distrib. Learn. 14 (5), 258–276. https://doi.org/10.19173/irrodl.v14i5.1524

Cho, K., Schunn, C.D., Charney, D., 2006. Commenting on writing: typology and perceived helpfulness of comments from novice peer reviewers and subject matter experts. Writ. Commun. 23 (3), 260–294. https://doi.org/10.1177/0741088306289261

Clemente, V., Vieira, R., Tschimmel, K., 2016. A learning toolkit to promote creative and critical thinking in product design and development through design thinking. In: 2016 2nd International Conference of the Portuguese Society for Engineering Education (CISPEE), pp. 1–6, https://doi.org/10.1109/CISPEE.2016.7777732

Cook, A., Hammer, J., Elsayed-Ali, S., Dow, S., 2019. How guiding questions facilitate feedback exchange in project-based learning. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, CHI '19, New York, NY, USA, pp. 1–12, https://doi.org/10.1145/3290605.3300368

Cook, A., Dow, S., Hammer, J., 2020. Designing interactive scaffolds to encourage reflection on peer feedback. In: Proceedings of the 2020 ACM Designing Interactive Systems Conference, DIS '20. Association for Computing Machinery, New York, NY, USA, pp. 1143–1153, https://doi.org/10.1145/3357236.3395480

Cordova, L., Carver, J., Gershmel, N., Walia, G., 2021. A comparison of inquiry-based conceptual feedback VS. Traditional detailed feedback mechanisms in software testing education: an empirical investigation. In: Proceedings of the 52nd ACM Technical Symposium on Computer Science Education, SIGCSE '21. Association for Computing Machinery, New York, NY, USA, pp. 87–93, https://doi.org/10.1145/3408877.3432417

Eris, O., 2004. Effective Inquiry for Innovative Engineering Design, vol. 10. Springer Science & Business Media.

Ertmer, P.A., Richardson, J.C., Belland, B., Camin, D., Connolly, P., Coulthard, G., Lei, K., Mong, C., 2007. Using peer feedback to enhance the quality of student online postings: an exploratory study. J. Comput.-Mediat. Commun. 12 (2), 412–433.

Feldman, E.B., 1994. Practical art criticism.

Ferrari, A., Spoletini, P., Bano, M., Zowghi, D., 2020. Sapeer and Reversesapeer: teaching requirements elicitation interviews with role-playing and role reversal. Requirements Eng. 25, 417–438. https://doi.org/10.1007/s00766-020-00334-0

Fiorella, L., Mayer, R.E., 2013. The relative benefits of learning by teaching and teaching expectancy. Contemp. Educ. Psychol. 38 (4), 281–288. https://doi.org/10.1016/j.cedpsych.2013.06.001

Frich, J., Nouwens, M., Halskov, K., Dalsgaard, P., 2021. How digital tools impact convergent and divergent thinking in design ideation. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, CHI '21, New York, NY, USA, https://doi.org/10.1145/3411764.3445062

Fuchs, K., 2023. Exploring the Opportunities and Challenges of NLP Models in Higher Education: is Chat GPT a Blessing or a Curse? https://doi.org/10.3389/feduc.2023.1166682, Vol. 8.

Gielen, S., Peeters, E., Dochy, F., Onghena, P., Struyven, K., 2010. Improving the effectiveness of peer feedback for learning. Learn. Instr. 20 (4), 304–315, unravelling Peer Assessment. https://doi.org/10.1016/j.learninstruc.2009.08.007

Graesser, A.C., Lu, S., Jackson, G.T., Mitchell, H.H., Ventura, M., Olney, A., Louwerse, M.M., 2004. Autotutor: a tutor with dialogue in natural language. Behav. Res. Methods Instrum. Comput. 36, 180–192. https://doi.org/10.3758/BF03195563

Green, L.N., Bonollo, E., 2005. Studio-Based Teaching: History and Advantages in the Teaching of Design. https://api.semanticscholar.org/CorpusID:244712029.

Greenberg, M.D., Easterday, M.W., Gerber, E.M., 2015. Critiki: a scaffolded approach to gathering design feedback from paid crowdworkers. In: Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition. Association for Computing Machinery, C&C '15, New York, NY, USA, pp. 235–244, https://doi.org/10.1145/2757226.2757249

Han, J., Yoo, H., Kim, Y., Myung, J., Kim, M., Lim, H., Kim, J., Lee, T.Y., Hong, H., Ahn, S.-Y., Oh, A., 2023. Recipe: how to integrate Chatgpt into EFL writing education. In: Proceedings of the Tenth ACM Conference on Learning @ Scale, L@S '23. Association for Computing Machinery, New York, NY, USA, pp. 416–420, https://doi.org/10.1145/3573051.3596200

Han, J., Yoo, H., Myung, J., Kim, M., Lim, H., Kim, Y., Lee, T.Y., Hong, H., Kim, J., Ahn, S.-Y., Oh, A., 2024. LLM-as-a-tutor in EFL writing education: focusing on evaluation of student-LLM interaction. In: Kumar, S.; Balachandran, V.; Park, C.Y.; Shi, W.; Hayati, S.A.; Tsvetkov, Y.; Smith, N.; Hajishirzi, H.; Kang, D., Jurgens, D. (Eds.), Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U). Association for Computational Linguistics, Miami, Florida, USA, pp. 284–293, https://doi.org/10.18653/v1/2024.customnlp4u-1.21

Hedges, L.V., 1981. Distribution theory for glass's estimator of effect size and related estimators. J. Educ. Stat. 6 (2), 107–128. http://www.jstor.org/stable/1164588.

Henriksen, D., Richardson, C., Mehta, R., 2017. Design thinking: a creative approach to educational problems of practice. Think. Skills Creat. 26, 140–153. https://doi.org/10.1016/j.tsc.2017.10.001.

Hofstede, G., 1984. Culture's Consequences: International Differences in Work-Related Values, vol. 5. Sage.

Hovardas, T., Tsivitanidou, O.E., Zacharia, Z.C., 2014. Peer versus expert feedback: an investigation of the quality of peer feedback among secondary school students. Comput. Educ. 71, 133–152. https://doi.org/10.1016/j.compedu.2013.09.019

Hurst, A., Nespoli, O.G., 2019. Comparing instructor and student verbal feedback in design reviews of a Capstone design course: differences in topic and function. Int. J. Eng. Educ. 35 (1), 221–231.

Jane, L.E., Yen, Y.-C.G., Pan, I.Y., Lin, G., Li, M., Jin, H., Chen, M., Xia, H., Dow, S.P., 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, pp. 292–310, https://doi.org/10.1145/3635636.3656183

Jang, Y., Kim, J., Lee, W., 2018. Development and application of Internet of Things educational tool based on peer to peer network. Peer-to-Peer Netw. Appl. 11, 1217–1229. https://doi.org/10.1007/s12083-017-0608-y

Jin, H., Lee, S., Shin, H., Kim, J., 2024. Teach AI how to code: using large language models as teachable agents for programming education. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI '24. Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/3613904.3642349

Jo, E., Epstein, D.A., Jung, H., Kim, Y.-H., 2023. Understanding the benefits and challenges of deploying conversational AI leveraging large language models for public health intervention. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23. Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/3544548.3581503

Jonassen, D.H., 2000. Toward a design theory of problem solving. Educ. Technol. Res. Dev. 48 (4), 63–85. https://doi.org/10.1007/BF02300500

Jug, R., Jiang, X.S., Bean, S.M., 2018. Giving and receiving effective feedback: a review article and how-to guide. Arch. Pathol. Lab. Med. 143 (2), 244–250. https://doi.org/10.5858/arpa.2018-0058-RA

Junprung, E., 2023. Exploring the intersection of large language models and agent-based modeling via prompt engineering. arXiv:2308.07411, https://arxiv.org/abs/2308.07411.

Kang, H.B., Amoako, G., Sengupta, N., Dow, S.P., 2018. Paragon: an online gallery for enhancing design feedback with visual examples. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI '18. Association for Computing Machinery, New York, NY, USA, pp. 1–13, https://doi.org/10.1145/3173574.3174180

Kiskola, J., Olsson, T., Väätäjä, H., Syrjämäki, A.H., Rantasila, A., Isokoski, P., Ilves, M., Surakka, V., 2021. Applying critical voice in design of user interfaces for supporting self-reflection and emotion regulation in online news commenting. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21. Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/3411764.3445783

Krause, M., Garncarz, T., Song, J., Gerber, E.M., Bailey, B.P., Dow, S.P., 2017. Critique style guide: improving crowdsourced design feedback with a natural language model. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, CHI '17. Association for Computing Machinery, New York, NY, USA, pp. 4627–4639, https://doi.org/10.1145/3025453.3025883

Kulik, J.A., Fletcher, J.D., 2016. Effectiveness of intelligent tutoring systems: a meta-analytic review. Rev. Educ. Res. 86 (1), 42–78. https://doi.org/10.3102/0034654315581420

Lambropoulos, N., Culwin, F., Romero, M., September 2010. HCI education to support collaborative e-learning systems design. ELearn 2010 (9), https://doi.org/10.1145/1858579.1858580

Lee, K., Kim, S.H., Lee, S., Eun, J., Ko, Y., Jeon, H., Kim, E.H., Cho, S., Yang, S., Kim, E.-M., Lim, H., 2025. Spectrum: a grounded framework for multidimensional identity representation in LLM-based agent. In: Chiruzzo, L.; Ritter, A., Wang, L. (Eds.), Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Association for Computational Linguistics, Albuquerque, New Mexico, pp. 6971–6991, https://aclanthology.org/2025.naacl-long.356/.

Lekschas, F., Ampanavos, S., Siangliulue, P., Pfister, H., Gajos, K.Z., 2021. ASK me or tell me? enhancing the effectiveness of crowdsourced design feedback. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21. Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/3411764.3445507.

Lim, H., Cho, J.Y., Kim, T., Park, J., Shin, H., Choi, S., Park, S., Lee, K., Kim, J., Lee, M., Hong, H., 2024. Co-Creating Question-and-Answer Style Articles with Large Language Models for Research Promotion. Association for Computing Machinery, New York, NY, USA, pp. 975–994. https://doi.org/10.1145/3643834.3660705

Lim, H., Choi, D., Hong, H., 2024. Identify design problems through questioning: exploring role-playing interactions with large language models to foster design questioning skills. In: Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing, CSCW Companion '24. Association for Computing Machinery, New York, NY, USA, pp. 598–602, https://doi.org/10.1145/3678884.3681912

Ma, W., Adesope, O.O., Nesbit, J.C., Liu, Q., 2014. Intelligent tutoring systems and learning outcomes: a meta-analysis. J. Educ. Psychol. 106 (4), 901. https://doi.org/10.1037/a0037123

Markel, J.M., Opferman, S.G., Landay, J.A., Piech, C., 2023. Gpteach: interactive ta training with gpt-based students. In: Proceedings of the Tenth ACM Conference on Learning @ Scale. Association for Computing Machinery, L@S '23, New York, NY, USA, pp. 226–236, https://doi.org/10.1145/3573051.3593393

McDonnell, J., 2016. Scaffolding practices: a study of design practitioner engagement in design education. Design Stud. 45, 9–29, special Issue: Design Review Conversations. https://doi.org/10.1016/j.destud.2015.12.006

Menon, S., Zhang, W., Perrault, S.T., 2020. Nudge for deliberativeness: how interface features influence online discourse. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. Association for Computing MachineryCHI '20, New York, NY, USA, pp. 1–13, https://doi.org/10.1145/3313831.3376646

Minamizawa, K., Kakehi, Y., Nakatani, M., Mihara, S., Tachi, S., 2012. Techtile toolkit: a prototyping tool for design and education of haptic media. In: Proceedings of the 2012 Virtual Reality International Conference, VRIC '12. Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/2331714.2331745

Misiejuk, K., Wasson, B., Egelandsdal, K., 2021. Using learning analytics to understand student perceptions of peer feedback. Comput. Hum. Behav. 117, 106658. https://doi.org/10.1016/j.chb.2020.106658

Narciss, S., 1999. Motivational Effects of the Informativeness of Feedback.

Ngoon, T.J., Fraser, C.A., Weingarten, A.S., Dontcheva, M., Klemmer, S., 2018. Interactive guidance techniques for improving creative feedback. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI '18. Association for Computing Machinery, New York, NY, USA, pp. 1–11, https://doi.org/10.1145/3173574.3173629.

Nguyen, T.T.D.T., Garncarz, T., Ng, F., Dabbish, L.A., Dow, S.P., 2017. Fruitful feedback: positive affective language and source anonymity improve critique reception and work outcomes. In: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW '17. Association for Computing Machinery, New York, NY, USA, pp. 1024–1034, https://doi.org/10.1145/2998181.2998319.

Oppenlaender, J., Kuosmanen, E., Lucero, A., Hosio, S., 2021. Hardhats and Bungaloos: comparing crowdsourced design feedback with peer design feedback in the classroom. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21, Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/3411764.3445380

Park, J., Choi, D., 2023. Audilens: configurable llm-generated audiences for public speech practice. In: Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, UIST '23 Adjunct. Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/3586182.3625114

Peng, Z., Liu, Y., Zhou, H., Xu, Z., Ma, X., 2022. Crebot: exploring interactive question prompts for critical paper reading. Int. J. Hum.-Comput. Stud. 167, 102898. https://doi.org/10.1016/j.ijhcs.2022.102898.

Peng, Z., Chen, Q., Shen, Z., Ma, X., Oulasvirta, A., April 2024. Designquizzer: a community-powered conversational agent for learning visual design. Proc. ACM Hum.-Comput. Interact. 8 (CSCW1), https://doi.org/10.1145/3637321

Rao, D., Stupans, I., 2012. Exploring the potential of role play in higher education: development of a typology and teacher guidelines. Innov. Educ. Teach. Int. 49 (4), 427–436. https://doi.org/10.1080/14703297.2012.728879

Razzouk, R., Shute, V., 2012. What is design thinking and why is it important? Rev. Educ. Res. 82 (3), 330–348. https://doi.org/10.3102/0034654312457429

Roldan, W., Gao, X., Hishikawa, A.M., Ku, T., Li, Z., Zhang, E., Froehlich, J.E., Yip, J., 2020. Opportunities and challenges in involving users in project-based HCI education. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20. Association for Computing Machinery, New York, NY, USA, pp. 1–15, https://doi.org/10.1145/3313831.3376530

Rucker, M.L., Thomson, S., 2003. Assessing student learning outcomes: an investigation of the relationship among feedback measures. Coll. Stud. J. 37 (3), 400–405.

Sadler, D.R., 1989. Formative assessment and the design of instructional systems. Instr. Sci. 18 (2), 119–144. https://doi.org/10.1007/BF00117714

Scott, C., Atman, C.J., Turns, J., 2001. Mastering design concepts through the coding of design. In: Proceedings, American Society for Engineering Education Annual Conference and Exposition. American Society of Engineering Education, pp. 47907–2016.

Seaman, M., 2011. Bloom's taxonomy. Curric. Teach. Dialogue 13.

Shaikh, O., Chai, V.E., Gelfand, M., Yang, D., Bernstein, M.S., 2024. Rehearsal: simulating conflict to teach conflict resolution. In: Proceedings of the CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, CHI '24, New York, NY, USA, https://doi.org/10.1145/3613904.3642159

Shneiderman, B., 2002. Creativity support tools. Commun. ACM 45 (10), 116–120.

Shute, V.J., 2008. Focus on formative feedback. Rev. Educ. Res. 78 (1), 153–189. https://doi.org/10.3102/0034654307313795

Thurlings, M., Vermeulen, M., Bastiaens, T., Stijnen, S., 2013. Understanding feedback: a learning theory perspective. Educ. Res. Rev. 9, 1–15. https://doi.org/10.1016/j.edurev.2012.11.004

Tohidi, M., Buxton, W., Baecker, R., Sellen, A., 2006. Getting the right design and the design right. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '06. Association for Computing Machinery, New York, NY, USA, pp. 1243–1252, https://doi.org/10.1145/1124772.1124960

TschannenMoran, M., Hoy, A.W., 2001. Teacher efficacy: capturing an elusive construct. Teach. Teach. Educ. 17 (7), 783–805. https://doi.org/10.1016/S0742-051X(01)00036-1

Valkenburg, R., Dorst, K., 1998. The reflective practice of design teams. Design Studies 19 (3), 249–271. https://doi.org/10.1016/S0142-694X(98)00011-8

VanLEHN, K.U.R.T., 2011. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educ. Psychol. 46 (4), 197–221. https://doi.org/10.1080/00461520.2011.611369

Wadinambiarachchi, S., Kelly, R.M., Pareek, S., Zhou, Q., Velloso, E., 2024. The effects of generative AI on design fixation and divergent thinking. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI '24. Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/3613904.3642919

Wambsganss, T., Kueng, T., Soellner, M., Leimeister, J.M., 2021. Arguetutor: an adaptive dialog-based learning system for argumentation skills. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, CHI '21, New York, NY, USA, https://doi.org/10.1145/3411764.3445781

Wambsganss, T., Janson, A., Käser, T., Leimeister, J.M., November 2022. Improving students argumentation learning with adaptive self-evaluation nudging. Proc. ACM Hum.-Comput. Interact. 6 (CSCW2), https://doi.org/10.1145/3555633

Winkler, R., Hobert, S., Salovaara, A., Söllner, M., Leimeister, J.M., 2020. Sara, the lecturer: improving learning in online education with a scaffolding-based conversational agent. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20. Association for Computing Machinery, New York, NY, USA, pp. 1–14, https://doi.org/10.1145/3313831.3376781

Wu, Y.W., Bailey, B.P., January 2021. Better feedback from nicer people: narrative empathy and ingroup framing improve feedback exchange. Proc. ACM Hum.-Comput. Interact. 4 (CSCW3), https://doi.org/10.1145/3432935

Wynn, D.C., Eckert, C.M., 2017. Perspectives on iteration in design and development. Res. Eng. Des. 28, 153–184. https://doi.org/10.1007/s00163-016-0226-3.

Wynn, D.C., Maier, A.M., 2022. Feedback systems in the design and development process. Res. Eng. Des. 33 (3), 273–306. https://doi.org/10.1007/s00163-022-00386-z.

Yang, C.-L., Uhde, A., Yamashita, N., Kuzuoka, H., 2025. Understanding and supporting peer review using ai-reframed positive summary. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, CHI '25, New York, NY, USA, https://doi.org/10.1145/3706598.3713219

Yen, Y.-C.G., Dow, S.P., 2022. Seeking exemplars in the wild: exploring how students find design examples to support personalized learning. In: Proceedings of the Ninth ACM Conference on Learning @ Scale. Association for Computing Machinery, L@S '22, New York, NY, USA, pp. 418–421, https://doi.org/10.1145/3491140.3528303

Yen, Y.-C.G., J. L.E, Jin, H., Li, M., Lin, G., Pan, I.Y., Dow, S.P., April 2024. Processgallery: contrasting early and late iterations for design principle learning. Proc. ACM Hum.-Comput. Interact. 8 (CSCW1), https://doi.org/10.1145/3637389

Yeo, S., Lim, G., Gao, J., Zhang, W., Perrault, S.T., 2024. Help me reflect: leveraging self-reflection interface nudges to enhance deliberativeness on online deliberation platforms. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI '24. Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/3613904.3642530

Yilmaz, S., Daly, S.R., 2016. Feedback in concept development: comparing design disciplines. Des. Stud. 45, 137–158, special Issue: Design Review Conversations. https://doi.org/10.1016/j.destud.2015.12.008

Yoshida, R., 2008. Teachers' choice and learners' preference of corrective feedback types. Lang. Awareness 17 (1), 78–93. https://doi.org/10.2167/la429.0

Yu, H., 2023. Reflection on whether chat GPT should be banned by academia from the perspective of education and teaching. Front. Psychol. 14, https://doi.org/10.3389/fpsyg.2023.1181712

Yuan, A., Luther, K., Krause, M., Vennix, S.I., Dow, S.P., Hartmann, B., 2016. Almost an expert: the effects of rubrics and expertise on perceived value of crowdsourced design critiques. In: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, CSCW '16. Association for Computing Machinery, New York, NY, USA, pp. 1005–1017, https://doi.org/10.1145/2818048.2819953

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., Weston, J., 2018. Personalizing dialogue agents: i have a dog, do you have PETS too? In: Gurevych, I., Miyao, Y. (Eds.), Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Melbourne, Australia, pp. 2204–2213, https://doi.org/10.18653/v1/P18-1205

Zhu, H., Dow, S.P., Kraut, R.E., Kittur, A., 2014. Reviewing versus doing: learning and performance in crowd assessment. In: Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing, CSCW '14. Association for Computing Machinery, New York, NY, USA, pp. 1445–1455, https://doi.org/10.1145/2531602.2531718