Design Activity Simulation: Opportunities and Challenges in Using Multiple Communicative AI Agents to Tackle Design Problems

Boyin Yang
Department of Engineering
University of Cambridge
Cambridge, United Kingdom
by266@cam.ac.uk

John J Dudley
Department of Engineering
University of Cambridge
Cambridge, United Kingdom
jjd50@cam.ac.uk

Per Ola Kristensson
Department of Engineering
University of Cambridge
Cambridge, United Kingdom
pok21@cam.ac.uk

Abstract

Large Language Models (LLMs) can enhance structured design thinking, yet existing copilot approaches integrate them into human workflows rather than exploring their autonomous potential. This paper investigates how LLM-based communicative AI agents can independently tackle open-ended design problems and how their strengths and limitations inform human-AI collaboration. We iteratively design a system where AI agents play different roles and simulate human design activity through conversational turns. The agents investigate user needs, identify design constraints, and explore the design space, with useful insights emerging from their interactions. To assess reasoning quality, we conducted a human jury evaluation with five HCI researchers and explored potential applications through a contextual inquiry with seven professionals. Our findings demonstrate that integrating human design thinking techniques enhances AI reasoning. AI agents effectively tackle design problems, generating low-novelty yet well-grounded and practical solutions that meet key design requirements.

CCS Concepts

• **Human-centered computing** → **HCI design and evaluation methods**; **Interaction design process and methods**.

Keywords

Generative AI; End-user interaction with LLMs and Multimodal models

ACM Reference Format:
Boyin Yang, John J Dudley, and Per Ola Kristensson. 2025. Design Activity Simulation: Opportunities and Challenges in Using Multiple Communicative AI Agents to Tackle Design Problems. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces (CUI '25), July 08–10, 2025, Waterloo, ON, Canada.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3719160.3736609

CUI '25, Waterloo, ON, Canada

ACM ISBN 979-8-4007-1527-3/25/07
https://doi.org/10.1145/3719160.3736609

1 Introduction

Just as past industrial revolutions saw machine automation greatly improving labor efficiency, the rise of generative artificial intelligence (GenAI) heralds a promising future for automating knowledge work, with the potential for dramatic productivity gains. Specifically, recent advancements in Large Language Models (LLMs) have brought elements of automation into the Human-Centered Design (HCD) process. Existing applications include: understanding user needs [76]; informing brainstorming, designing, and coding of a project at the early stage [99]; augmenting human creativity and inspiring divergent thinking [91]; facilitating nontechnical communication and system understanding between technical and nontechnical team members [72]; providing enjoyable design experiences [51]; assessing the feasibility of designs during the early design stage [51, 65]; informing the strategic decision-making process by analyzing user trends, preferences and sentiments [65]; and many more.

However, despite these promising advances, current copilot approaches leave critical gaps unaddressed. In practice, HCD projects face persistent challenges due to the need for multidisciplinary expertise, sustained user engagement, and complex coordination, all of which require substantial human effort. While LLM-based copilots can assist with specific tasks, they do not relieve human teams of the broader burdens of integration, management, and oversight. Moreover, the design process is deeply experience-driven, which can lead to blind spots or unproductive directions when teams lack expertise or carry biases. This raises a critical question: can we move beyond task-level augmentation toward fully autonomous systems capable of simulating the end-to-end HCD process? Such systems could not only test the boundaries of design automation but also offer scalable, accessible, and potentially transformative approaches to human-AI collaboration. Therefore, this paper investigates how a fully autonomous AI system can directly tackle design problems. It serves as a foundational study that guides future research on human-AI collaboration in design automation.

Essentially, HCD prioritizes users' needs, preferences, and experiences throughout the design process, aiming to create intuitive, effective, and satisfying products, services, and systems [81]. Achieving this requires a deep understanding of user behavior, emotions, aesthetics, and the interplay between technology and users. Further, given that this knowledge spans business, technical, manufacturing, and marketing domains, mastering HCD is beyond the capacity of any single individual. Consequently, HCD is typically carried out by multidisciplinary teams, often involving

direct participation from end-users [33]. This process is inherently time-consuming and resource-intensive [11, 19, 82], with discovery activities lasting four to eight weeks and iterative testing cycles extending the overall design timeline [87]. Additionally, the necessity for sustained user engagement and collaboration across diverse expertise further adds to the complexity and investment required. To address this problem, Schmidt et al. [76] point out that LLMs can be used to play user roles based on personas in the design problem scenario. The rationale behind this is that LLMs encode a broad array of experiences that people have recorded, which can offer a vast amount of information mimicking human feedback.

Building on the work of Schmidt et al. [76], we conjecture that LLMs can go even further. Beyond simulating end-users, LLMs can simulate the roles of members within a design team, including users and professionals. This presents significant potential benefits for human design teams in terms of availability, scalability, and efficiency:

Availability It is challenging to assemble multidisciplinary teams as such expertise may be difficult to source or might be unavailable at the time. AI agents can potentially fulfill such roles, either fully or partially.

Scalability Automated or semi-automated design processes provide resource-efficient ways of exploring larger design spaces by increasing the pace of ideation and critique. This enables design teams to potentially consider more opportunities for design than otherwise.

Efficiency Projects are resource-constrained in terms of time, money, personnel, and so on. By allowing AI to manage some tasks in the early design stage, human designers can focus on other tasks. For example, by bootstrapping a design project with a partial AI-generated design, human designers can more rapidly focus on other aspects, such as stakeholder engagement or detailed design.

We emphasize that humans can play a supervisory role in guiding and monitoring AI agents' work where humans do not directly engage in the design for frontline tasks but stand behind the AI agents to monitor their work and intercept and steer the work when necessary. We hypothesize that this approach can increase the level of automation of using AI in HCD tasks so that humans only need to engage in the design process when necessary. However, this requires a deeper understanding of how to enable AI agents to automatically work with each other and produce usable outcomes in tackling design problems.

To explore the automation boundary of AI agents in HCD tasks, we iteratively design and develop a group of communicative AI agents, each assigned distinct roles and responsibilities, to autonomously tackle design problems through divergent and convergent design processes. Additionally, the system includes a chatbot-like assistant agent that helps users navigate and analyze the extensive text-based outcomes generated throughout the AI agents' design activity (see Figure 1).

We iteratively evaluate this system with five HCI experts and seven professionals from different domains who take the supervisory role of these AI agents, and seek to answer the following four research questions:

(1) How can we enable AI agents to autonomously tackle open-ended design problems and improve the quality of their outputs?
(2) How do humans perceive the quality of AI agents' work?
(3) How can we present AI-agent-based design activity in a way that is meaningful to humans?
(4) What insights can we gain from this approach?

The contributions of this work are the following:

- We present a system of multiple LLM-based communicative AI agents, each assuming a distinct role within a design team. These roles include end-users, a user researcher, a product manager, a business analyst, an ethical advisor, an interaction designer, and a developer. Working collaboratively, these agents autonomously tackle open-ended HCD problems by simulating human design activities. Their interactions follow a high-level structured framework, enabling them to systematically investigate user needs, identify design constraints, and explore the design space.
- We develop an interactive web application that presents the generated design activity and enables post-hoc analysis of insights produced during AI agents' interactions.
- The results demonstrate that AI agents can effectively collaborate to generate informative and feasible design outcomes with speed. Both a human jury evaluation with five HCI researchers and a contextual inquiry with seven professionals find that while AI-generated solutions tend to lack novelty, they are well-grounded, practical, and comprehensive—addressing key aspects such as user needs, business requirements, ethics, user experience, and technical feasibility. Additionally, the seven professionals identified potential applications for design students, user researchers, designers, consultants, product managers, and project managers.

2 Related work

2.1 Large Language Models in Human-Centered Design

Recent work has found that LLMs can be a powerful tool for augmenting human creativity and inspiring divergent thinking [101]. These AI tools are increasingly integrated into the HCD process to assist designers in ideation [8, 25, 35], rapid prototyping [54, 76], and decision-making [4, 32]. Typically, LLM-based AI tools generate a wide range of design alternatives or enhance existing ones, enabling designers to explore novel possibilities and refine their work more efficiently [71].

In addition to functioning as design tools, LLMs are increasingly being utilized to simulate end users [76, 102]. This allows for more informed decision-making early in the design process, potentially minimizing the need for extensive user testing and iteration. The underlying logic is that if the model can provide insights comparable to those from human users, it may be more efficient to rely on the model. However, when human feedback offers deeper or more valuable insights, bypassing it in favor of models could lead to suboptimal outcomes.

Schmidt et al. [76] suggest that, besides simulating users, if we see LLMs conceptually as a massive database of experiences, it

should be possible to use their input to replace human feedback in certain parts of the design process. Schmidt et al. [76] propose four activities that LLMs can assist with: (1) understanding and specifying user needs; (2) prototyping and implementing interfaces and systems; (3) evaluating systems; and (4) discussing systems, including teasing out ethical implications.

In this paper, we extend this idea by simulating the entire design team's cooperation and collaboration. While much research has focused on how AI supports individual designers, less attention has been given to how AI could simulate or augment team interactions during the design process. Exploring how AI can transform collaborative practices is essential to fully realizing its potential in HCD. In this vein, we seek to fully automate the entire design process. Different from existing AI design tools serving as copilot systems where humans play the role of co-workers of AI, we highlight that humans can play a supervisory role in inspecting AI agents' work. This requires a system-level automation perspective for AI agents.

2.2 Traditional Multi-Agent Systems and Their Opportunities in the AI Era

Multi-agent systems (MAS) research has a rich history encompassing various fields and several evolutions. A typical MAS is usually an automated system that consists of multiple interacting computing elements known as agents. These agents possess two key capabilities: (1) they are at least to some extent capable of autonomous actions of deciding for themselves what to do to achieve their goals; and (2) they can interact with other agents through human-like social activities such as cooperation, coordination, and negotiation [96].

These agents can observe their environment and interact in a shared environment to achieve common or conflicting goals. Each agent has specified goals, and they can take specific actions to modify the environment or send information to other agents, which means they can impact each other's decisions and actions directly and indirectly [3].

This theoretical foundation suggests that MAS could be leveraged to address design problems, where each agent can be a member with different roles in a design team. In a MAS, individual agents may have different prior knowledge about the environment while often only observing some partial information about the state of the environment. Different agents may receive different observations about the environment. All these agents in the MAS are goal-directed, and their actions are driven by their goals.

Traditional MASs work well with quantitative-data-driven tasks in relatively restricted environments, such as autonomous driving [2, 78], multi-robot factories [38, 64], robotic rescue teams [27, 55], automated trading [43, 56], and commercial games [29, 40]. These tasks have two common characteristics. First, the environment is relatively static. Although the data may change, most features for achieving the goal are known in advance. Second, the data are quantitative, involving measurable variables such as distance, energy consumption, weight, time, or money.

These conditions enable agents to calculate payoffs for their actions using Shapley value [80] (or other values, such as Myerson value [62]) so that they can adjust their strategies for finding an optimal solution (i.e., finding the strategy for the highest overall payoff). These ideas, rooted in economics and game theory, are widely applied in traditional MAS research.

However, in knowledge-based tasks, such as those found in HCD, calculating payoffs becomes exceedingly difficult. The data in these tasks are qualitative (e.g., natural language conveying semantic meaning and logic), making it challenging for traditional MAS approaches to handle such complex problems. Thus, to address knowledge work challenges, we need alternatives beyond traditional MAS.

The rapid advancements in LLMs have demonstrated their capability to effectively process semantic data, revealing great potential for utilizing multiple LLM-based AI agents to process complex knowledge work. For example, Zhang et al. [100] use LLM-based MAS to facilitate ontology alignment, which typically necessitates the involvement of cross-domain human experts. However, Abdelnabi et al. [1] introduce a benchmark to systematically evaluate the arithmetic, inference, exploration, and planning capabilities of LLM-based MASs via text-based, multi-agent, multi-issue, semantically rich negotiation games. Their findings reveal that even state-of-the-art models, including GPT-4 and Llama-3 70B, still underperform in these tasks. Similarly, tackling design problems can be framed as a text-based, multi-agent, multi-issue, semantically rich negotiation process, where AI agents must reason about conflicting business requirements, user needs, ethical considerations, and technical constraints. This suggests that simply combining existing LLM-based AI agents is insufficient for reliably handling such complex challenges. Accordingly, there is a need for a dedicated system—one that is flexible enough to navigate open-ended design tasks while maintaining a structured approach to ensure high-quality performance.

2.3 Large Language Model-based Human Activity Simulation and Human-Centered Design

LLMs can be used to serve as agents to simulate human activities. This is usually achieved by assigning personas to AI agents for role-playing [1, 67]. The key observation is that LLMs encode a wide range of human behavior from their training data [12, 14, 67, 79]. For specific use cases, researchers either fine-tune the LLMs [79] or use a prompt chain [67, 97] to generate agent behaviors in context by crafting an agent architecture that handles retrieval where past experience is dynamically updated at each time step and mixed with agents' current contexts and plans. These agents are described with personas, defining their identity, occupation, and relationship with other agents [67].

Personas are also widely adopted in HCD [44, 58, 59], providing a comprehensive and realistic representation of the target users to aid in the design of more user-centered products and services. However, there is little research on simulating the entire design process, expanding the focus from simulating individual-level agent performance to simulating the performance and dynamics of a group of co-working agents, which, besides persona, requires dedicated group-level agent interaction mechanisms.

Traditional HCD theories, frameworks, and approaches provide rich resources as foundations for simulating both individual-level and group-level agent activities in design problem-tackling tasks

from at least four perspectives: (1) contextual understanding [15, 16, 37, 74, 77, 89]; (2) ideation and creative engagement [21, 24, 30, 46, 48, 66, 90]; (3) design team and user engagement [5, 6, 9, 18, 26, 39, 45, 47, 61, 69, 75, 77, 84]; and (4) overall design process and strategies [10, 13, 20, 60]. In this paper, we mainly focus on the overall design process and strategies, which can build a foundation for various follow-up studies from all these perspectives, ultimately improving the quality of work produced by AI agents tackling real-world design problems.

2.4 Interaction Between Multiple AI Agents

The rapid development of chat-based LLMs has substantially advanced the ability to tackle complex tasks. Recent studies have explored using multiple AI agents to improve reasoning abilities, evaluated in standard testing scenarios, such as mathematical and strategic reasoning tasks [28, 52, 93]. However, these works intend to improve the quality of responses in a single output rather than maintaining a continuous discussion. This may be the best fit for single-issue tasks in which one final answer can solve a problem.

However, a real-world problem can consist of a series of subquestions, some of which are initially uncertain and can only be identified during the process. Li et al. [50] address this by proposing a communicative agent framework in which two agents autonomously collaborate to complete instruction-following tasks with minimal initial human intervention. This work demonstrates that, through prompt engineering and feeding the content generated by one agent to another, two LLM-based AI agents can continuously ask and answer questions to tackle problems. Thereby it introduces a way for two AI agents to iteratively tackle a problem that fundamentally aligns with the cooperative work process of humans. However, despite its promise, this approach still requires substantial improvements to generate more usable outputs for real-world design problems. For example, instead of two agents, tackling design problems usually requires more agents in a team to cover various aspects of a design project. In order to produce human-centered outcomes, these agents need to work together to carry out typical divergent and convergent design thinking processes using HCD approaches to clarify design goals, propose various design options, evaluate these ideas, and form an overall design solution [81].

Building on the work of Li et al.[50], we investigate how multiple LLM-based communicative AI agents can simulate human design activities and generate valuable outputs for tackling design problems.

3 Developing a System that Allows Multiple AI Agents to Tackle Design Problems

Building upon the concept of MAS, we develop an automated system where multiple LLM-based conversational AI agents play different roles in a design team. These agents collaboratively tackle open-ended design problems through interactions in a shared environment, operating entirely without human involvement. Figure 1 shows the system user interface where human designers can interactively review and analyze the simulated design activity. From left to right of the web application, human designers can gain an overview of the simulated design activity outcome from the product requirements document (PRD), which provides a structured summary of the AI-generated design activity, including design goals, stakeholders, system features, and milestones. The center of the interface displays the original activity details, allowing designers to explore the original reasoning history of AI agents. On the right, human designers can interact with the AI project assistant agent, enabling further inquiries, refinements, and deeper engagement with the design process.

3.1 Establishing Lines of Communication and Knowledge Sharing

Traditional MAS research offers a formal way of describing multiple agents cooperatively working to solve a shared problem. We use these concepts to guide the design of our system of multiple LLM-based communicative AI agents cooperatively tackling design problems. Formally, we define Equation 1 to describe the cooperation modality [7]:

$$\langle\langle C\rangle\rangle P \tag{1}$$

where C is a collection of agents and P is a property (or condition) of interest. $\langle\langle C\rangle\rangle P$ means that there exists a collection of strategies for the agents C such that, if they follow these strategies, then P will be guaranteed to hold, regardless of the actions of other agents outside C. More generally, $\langle\langle N\rangle\rangle P$ (where N is any agent set) denotes that the agents in N can cooperatively enforce P under some strategy. In this context, $\langle\langle\rangle\rangle$ expresses the notion of enforceability—either over all possible paths (universally) or over some paths (existentially), depending on the surrounding formalism. Van der Hoek and Wooldrige [88] added knowledge modalities $K_{\mathcal{A}_i}$ for each agent $\mathcal{A}_i$ (we amend the notation and expression to accommodate our study):

$$\langle\langle \mathcal{A}_i\rangle\rangle K_{\mathcal{A}_j} P \tag{2}$$

Equation 2 denotes the *communication* between two agents that $\mathcal{A}_i$ can ensure $\mathcal{A}_j$, who has knowledge $K_{\mathcal{A}_j}$, comes to know P. In addition, Equation 3 defines the *task processing* of an agent $\mathcal{A}_i$:

$$\langle\langle \mathcal{A}_i\rangle\rangle P \leftarrow K_{\mathcal{A}_i} Q \tag{3}$$

It denotes that for the agent $\mathcal{A}_i$ who has knowledge $K_{\mathcal{A}_i}$, knowing Q is a necessary condition for being able to achieve P. Accordingly, we distill six essential rules from the MAS for creating the AI agent design team to tackle design problems:

(1) Unlike human communication, which can naturally be one-to-one or one-to-many, one-to-many communication among AI agents is composed of multiple one-to-one communications (see Section 3.2).
(2) An AI agent must be equipped with practical knowledge and provided with appropriate information to generate solutions (see Section 3.3).
(3) AI agents can generate knowledge for other AI agents but cannot generate knowledge for themselves (see Section 3.5).
(4) Therefore, the initial piece of knowledge (i.e., the design problem) must be provided by humans as a prompt (see Section 3.5).
(5) New information can be generated through communication between two AI agents or by an individual AI agent processing the provided information (see Section 3.2 and Section 3.5).

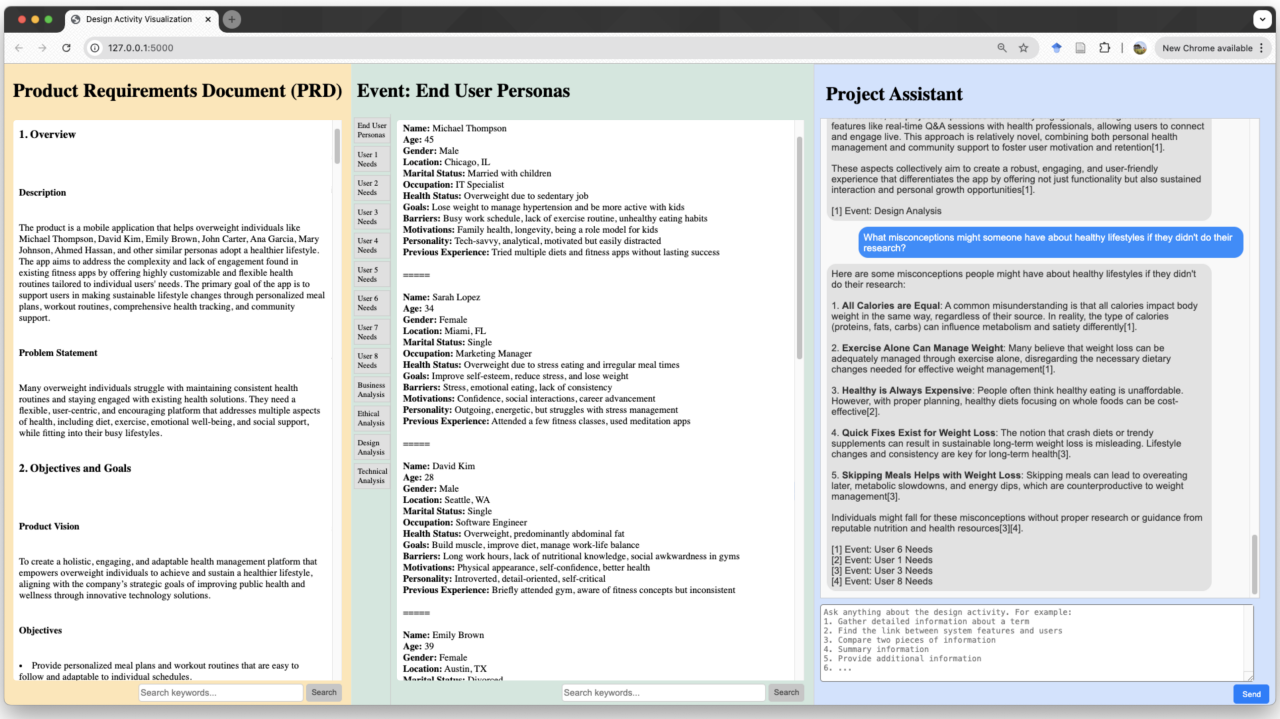


Figure 1: The user interface (UI) of the interactive system. This screenshot is taken when a human expert (E7) uses the system during the user study. The UI is split into three columns. The left column displays the overview of the design activity outcomes of the AI agents, named the product requirements document (PRD). The middle column illustrates the event's details in the entire design activity simulation. Human users can click the tabs on the left to switch between events. The right column is an LLM-based project assistant agent that can interactively carry out analyses and provide information based on the design activity shown in the left and the middle columns.

(6) By default, each AI agent can communicate with any other agents in the team. A communication network must be designed if a specific system pipeline is required (see Section 3.5).

3.2 Implementing Bilateral AI-Agent Communication

To create a system where multiple AI agents can communicate and contribute to the design process, we must first establish the fundamental one-to-one communication unit between two AI agents. In a recent study, Li et al. [50] introduce *inception prompting*, a prompt engineering technique that allows two AI agents to automatically prompt each other in a loop, starting with only a seed prompt provided by humans. The termination condition is typically predetermined by the number of iterations.

In our study, this technique enables bilateral AI agent communication, functioning as a Q&A process with context memory, where one agent asks questions and the other provides answers. Through this iterative exchange, AI agents collaboratively explore the problem by generating, analyzing, evaluating, and selecting ideas from their professional perspectives. The outcome of each conversation serves as new context knowledge passed to the next pair of AI agents with different expertise, ultimately tackling the given problem. This pipeline is detailed in Section 3.5.

We reproduce the mathematical expression of bilateral AI-agent communication [50] to accommodate the purpose of this study. The inquiry AI agent message obtained at time t is denoted by $\mathcal{I}_t$, and the reply AI agent message is denoted by $\mathcal{R}_t$. The set of messages obtained up until time t is denoted by Equation 4:

$$\mathcal{M}_t = \{(\mathcal{I}_0, \mathcal{R}_0), \ldots, (\mathcal{I}_t, \mathcal{R}_t)\} = \{(\mathcal{I}_i, \mathcal{R}_i)\}\big|_{i=0}^{t} \tag{4}$$

At the next time step, $t+1$, the inquiry AI agent $\mathcal{A}_I$ takes the historical conversation message set $\mathcal{M}_t$ and provides a new inquiry $\mathcal{I}_{t+1}$, as shown in Equation 5:

$$\mathcal{I}_{t+1} = \mathcal{A}_I(\mathcal{M}_t) \tag{5}$$

The new inquiry, along with the message set $\mathcal{M}_t$, is then passed to the reply AI agent for generating a new reply $\mathcal{R}_{t+1}$, as shown in Equation 6:

$$\mathcal{R}_{t+1} = \mathcal{A}_R(\mathcal{M}_t, \mathcal{I}_{t+1}) \tag{6}$$

After gathering the reply $\mathcal{R}_{t+1}$ to the inquiry $\mathcal{I}_{t+1}$, the updated message set $\mathcal{M}_{t+1}$ is defined by Equation 7:

$$\mathcal{M}_{t+1} \leftarrow \mathcal{M}_t \cup (\mathcal{I}_{t+1}, \mathcal{R}_{t+1}) \tag{7}$$

3.3 Imbuing Agents with Individual Intelligence

Individual-level intelligence is the fundamental component that defines the capabilities and attributes of an AI agent. Figure 2 (a) illustrates a neutral AI agent, which is the basis of different roles in the design team.

This individual-level intelligence can be realized using an LLM, such as GPT-4o. The LLM enables the AI agent to process and generate human-like text, facilitating effective interaction across diverse task contexts. The persona, goal, knowledge, and code of conduct are established through prompt engineering.

Given the variability of design problems, each AI agent assumes a distinct role within the design team. Consequently, while the content of the tasks may differ, the underlying problem-tackling strategy remains relatively consistent. Thus, prompt engineering should focus on guiding the strategy level rather than specific content, ensuring that the AI agent can adapt to varying contexts while maintaining a coherent approach to problem-tackling.

Traditional HCD theories, frameworks, and approaches are design thinking strategies that provide valuable insights to prompt engineering, as mentioned in Section 2.3. We later report in Section 4 how embedding these design thinking strategies in AI agents improves the outcome of the simulated design activities regarding user simulation realism and the quality of design task analysis.

3.4 Assigning Roles to Agents

Edmondson and Nembhard [31] highlight the value of bringing professionals from different functions to work together on development projects to create the highest quality product in the shortest time. This business objective aligns with the principle of co-design, which states that stakeholders should work together to form design solutions [73]. In the context of industry-focused HCD, problem-tackling refers to systematically identifying user needs and challenges and employing empathetic and collaborative approaches to develop reliable solutions that align with user requirements and business objectives to enhance product usability and market relevance. Therefore, we identify six essential professional roles for typical information technology (IT) industry project designs. We do not consider project management as this is a subsequent stage of product design, which highly relies on the actual development team. We list the duties of these roles in the design stage in Table 1.

3.5 Forming Agents into a Team

We create a group of AI agents simulating the seven roles within the design team (see Table 1) by adhering to the rules listed in Section 3.1. Figure 2 (b) illustrates the network of the AI agent design team: After the AI user researcher agent gathers the user needs from the AI user agent, the AI product manager agent coordinates the remaining design activity, communicating with other AI professional agents to address the design problem from their respective expert perspectives.

We streamline this design process, as shown in Figure 3. The entire design problem-solving task is composed of seven design events. Specifically, there are three types of event compositions consisting of multiple threads: a pair of AI agents communicating with each other, a human assigning tasks to an AI agent, and a single AI agent processing an assigned task. Each thread processes a sub-task of the event, which involves one or several runs. A run is a conversation turn between two participants. In this case, we assign two chronological threads: one for understanding and tackling the problem, and another for summarizing. The former focuses on identifying and addressing the design challenge, while the latter extracts information from the conversation history. This summarized conversation history is used for internal access, feeding knowledge to other AI agents. The original conversation is displayed in the web application, providing richer details about the design activity.

As shown in Figure 3, Events 1 and 7 do not involve inter-AI-agent discussions, as only one AI agent processes the information in each run. Events 2–6, however, are multi-turn discussions between AI agents. Based on their functions, the events can be categorized into two types: *design development events* and *design support events*. Design development events, such as "gather business analysis" (the first design event after understanding user needs), involve AI agents working to form a loose or preliminary design solution by exploring various aspects in depth. This is accomplished through the agents' use of the expansive exploration strategy, as discussed in Section 4. Design support events, which include tasks like "gather user needs", "gather ethical analysis", "gather technical analysis", and "gather design analysis", can occur either before or after design development events. In these events, AI agents focus on providing detailed information or exploring essential elements to support or refine the design. This is accomplished through the iterative deepening strategy adopted by the AI agents, as described in Section 4.

3.6 Iterative Refinement

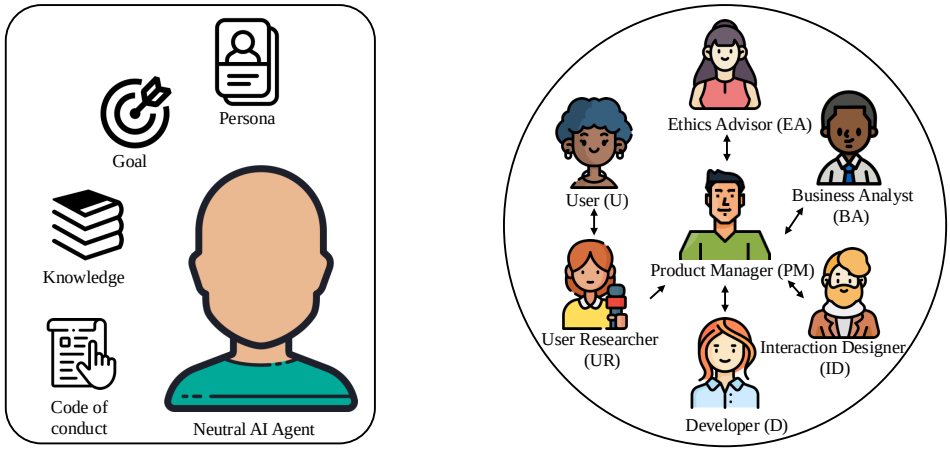
To understand how to improve the quality of AI agents' outputs and how humans benefit from this system, we evolved the design of this system over three main iterations. Each of these iterations focused on addressing a particular concern as summarized below:

- **Iteration 1: Strategy Concerns:** We enable the fully automated design activity simulation using multiple AI agents. The evaluation focuses on understanding the strategies of AI agents tackling design problems.
- **Iteration 2: Quality Concerns:** We investigate methods to improve the quality of simulated design activity, focusing on how human designers perceive its effectiveness. To evaluate this, we conducted a human jury study with five HCI professionals, assessing both the first and second versions of the system.
- **Iteration 3: Presentation Concerns:** We seek ways to interactively present AI agents' work to human users. The evaluation engages seven professionals, focusing on the utility value of this approach.

For conciseness and clarity, this paper mainly introduces and evaluates the current system rather than the previous versions of the system. The Supplementary Material contains the complete quantitative and qualitative evaluations that serve as a foundation of the current system design. For brevity, we summarize the key insights and findings from the previous systems here.

Through these three design iterations, we develop a system consisting of three main parts:

(1) Multiple AI agents working as a design team in the back end to tackle the given design problem (see Section 3.1–3.5);



(a) The construction of individual level intelligence (b) The network of group level intelligence for the AI agent design team

Figure 2: (a) The individual-level intelligence composition. Each AI agent is composed of four key elements: persona, goal, knowledge, and code of conduct. The persona defines the agent's role and background, the goal specifies the objectives the agent seeks to achieve during the design task, knowledge refers to the external information available to complete the task, and the code of conduct regulates the agent's behavior in asking and answering questions to sustain the conversation. (b) The group-level intelligence network for the design team. All seven AI agents work together as a team. The circle represents the shared environment in which all agents collaborate on the same problem. A two-way arrow indicates a conversation between two AI agents, while a one-way arrow represents a data-sharing process.

Role	Duties
User (U)	People who directly use the designed product. They have unique personas and experiences associated with the design problem. Their needs serve as the core value of the design.
User Researcher (UR)	Focuses on gathering insights about user behaviors, needs, and pain points. They use research methods, such as interviews, to inform the design and development process and ensure the product is user-centered.
Business Analyst (BA)	Responsible for identifying business needs and market gaps and translating them into design requirements for other team members. They aim to ensure the product aligns with the company's strategic goals and solves business problems.
Ethics Advisor (EA)	Ensures that the product complies with ethical standards and regulations, particularly concerning data privacy, user rights, and broader societal impacts. They help navigate ethical dilemmas and provide recommendations to avoid harmful outcomes for users or the business.
Interaction Designer (ID)	Focusing on creating an effective and user-friendly product design. They design product features by considering user needs, business requirements, and ethical considerations.
Developer (D)	Responsible for providing technical implementation options for product features, which may lead to different interaction experiences. They also analyze the technical feasibility and implementation to ensure the product is functional, scalable, and secure.
Product Manager (PM)	Responsible for defining the product vision, strategy, and roadmap. They act as a bridge between various stakeholders, including the user researcher, business analyst, ethics advisor, interaction designer, and developer, to ensure the product meets all requirements from various perspectives.

Table 1: Roles and duties of AI agents for design problem tackling.

(2) A web application for visualizing the design activity simulation (see Figure 1); and
(3) A project assistant agent integrated into the web application that interactively responds to human queries (see Section 6).

4 Addressing Strategy Concerns

We identified four main findings from the previous iterations.

First, **AI agents can adaptively switch between two conversational strategies.** The inductive thematic analysis reveals that when there is limited information about system features, AI

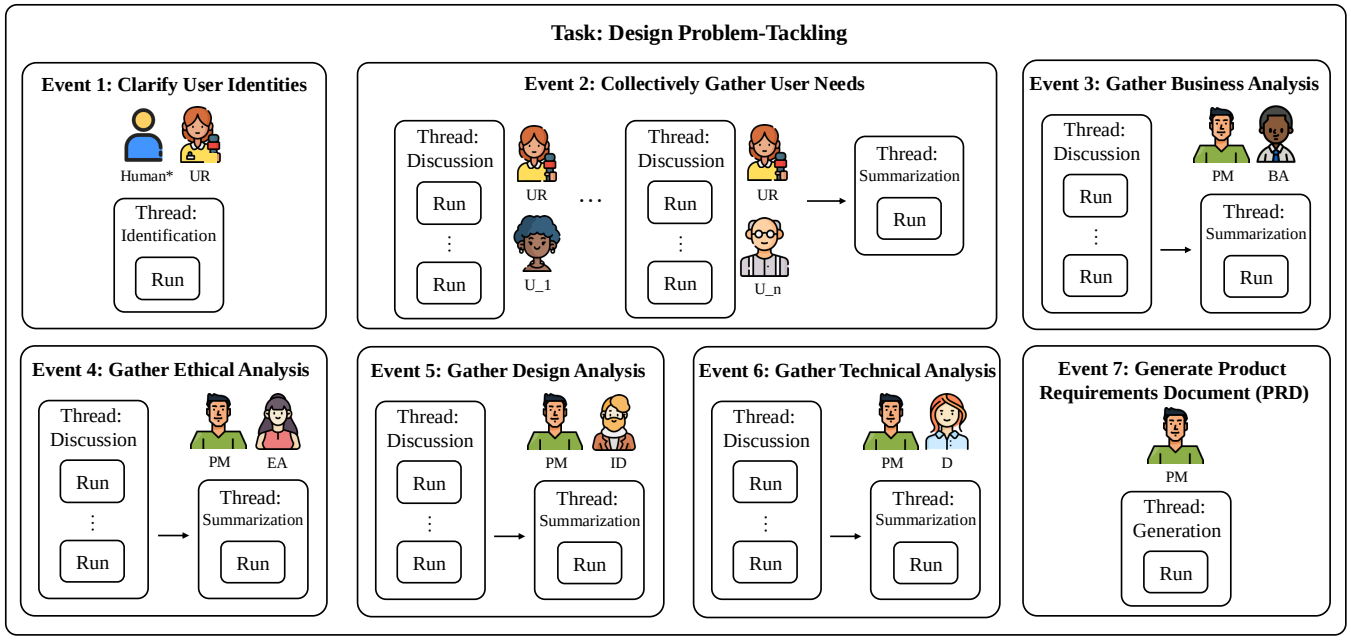


Figure 3: The pipeline of the AI agent design team tackling design problems (see AI agent roles and abbreviations in Figure 2 (b)). The *task* denotes the whole design problem-tackling process, which consists of multiple events. The *event* is a sub-task of the overall task, with specific predefined focuses. Each event involves a pair of players, either a human and an AI agent or two AI agents, except Event 7. An event can contain one or multiple *threads*. A thread is a mini-task of the sub-task, such as identification, discussion, summarization, and generation. Each thread consists of one or multiple *runs*. A run denotes a conversation turn. In Event 1, the human inputs a design problem, and an AI user researcher agent (UR) identifies multiple end-user personas, which are passed to Event 2. In Event 2, the AI user researcher agent conducts interviews with each AI user agent and provides an overall user interview summary of the interviews. This summary is passed to Events 3–6, where AI agents carry out design problem-tackling discussions, focusing on business requirements, ethical requirements, interaction design, and technical feasibility, respectively. Each event of 2–6 has two outputs: discussion history and discussion summary. Finally, in Event 7, the AI product manager agent (PM) generates a product requirements document (PRD) based on all the discussions. Note: *The human participant provides the design problem as an input and is not involved in the subsequent design discussion.

agents tend to explore more information by adopting an *expansive exploration* strategy. Conversely, when more information is available, they tend to dig deep into certain topics by adopting an *iterative deepening* strategy. This shift happens even without deliberate prompt engineering. That is, humans do not directly control these two strategies by default.

Second, **AI agents adopt a top-down thinking strategy.** Unlike humans, who usually are good at producing details and examples, AI agents tend to produce concepts and summaries. To address this, we implemented a follow-up questioning mechanism that prompts agents to go into more detail. However, this approach can sometimes push the discussion into irrelevant or unproductive directions. Incorporating an interactive human-AI conversation process remains an important avenue for future work, enabling human designers to help steer the depth and relevance of AI-generated content.

Third, **refining the design process and prompts can address missing design thinking steps in AI agents.** Based on the deductive thematic analysis driven by human-based design thinking theories [85], the first version of the system tended to focus on divergent (generating options) while ignoring convergent (evaluating options) design processes, leading to the design focusing on secondary requirements while overlooking primary goals. This is improved in the second system design iteration by embedding design thinking theories, such as the concept of the double diamond design process [20] (i.e., to incorporate divergent and convergent phases) and user interview techniques [95] (e.g., to avoid asking leading questions), to the prompt and adding an extra conversation turn for evaluating proposed ideas. Figure 4 illustrates an improvement in AI agents' performance in tackling design problems, motivating us to incorporate more user study and design thinking techniques in the final version.

Fourth, **AI agents can swiftly generate comprehensive ideas.** AI agents have the ability to facilitate parallel communication, which allows them to ask and answer multiple questions simultaneously in one conversation turn—substantially enhancing the communication rate compared to human conversations. As a result, they can swiftly generate a large number of requirements and

potential solutions. In the first design iteration, 92 questions were asked and answered across 20 conversation turns throughout six events in an eight-minute span, summarized into 17 requirements, and yielded 100 potential solutions. In terms of efficiency, each run in the first version of the system incurred a cost of approximately two dollars and took eight minutes. In the second system version, each run cost around two dollars and took 20–25 minutes. Figure 5 illustrates the comprehensiveness of the generated design space of an assigned design problem.

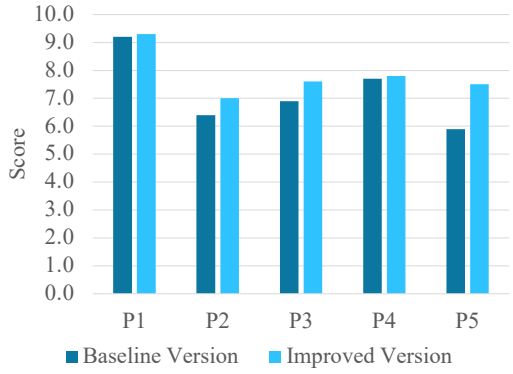


Figure 4: Bar plot of the overall score for baseline and improved versions from five human jurors after the second design iteration. The scores indicate jurors' perception of AI performance (1–2: very poor, 3–4: poor, 5–6: neutral, 7–8: good, 9–10: very good). The overall score is calculated as the average of 50 sub-category scores, grouped into 22 categories (see a detailed analysis in the Supplementary Material). Jurors P1 and P4 did not provide differentiated scores between the two versions (<0.1), while P2, P3, and P5 assigned higher overall scores to the refined version (>0.5). This finding demonstrates that more than half of the human jurors perceived a clear improvement from adopting human design thinking methods and user study techniques. This suggests that further integration of these approaches could lead to even more noticeable performance gains.

5 Addressing Quality Concerns

We carry out a human jury evaluation with five jurors with HCI backgrounds to understand the quality of the content generated by the AI agents. Their feedback on both the baseline and improved versions of the system demonstrates the three potential benefits for human designers in terms of availability, scalability, and efficiency, as mentioned in Section 1. However, there are three key challenges: (1) creativity dilution; (2) in-depth discussion avoidance; and (3) information presentation.

5.1 Creativity Dilution

To simulate design activities, AI agents tackle design problems by iteratively asking and answering questions. Although AI agents can produce vivid details at the beginning of the design activity, they tend to discuss explicit requirements while overlooking the implicit information and capture common ideas while ignoring unique points. As a result, the design activity simulation opens up a large design space that is insightful while ending up giving safe but standard solutions. This limitation highlights the value of incorporating human designers into the system loop to help steer the discussion between AI agents, guide exploration, and surface less obvious insights. For example, human designers could identify promising themes emerging from the automated process and actively direct the agents' focus toward novel or underexplored directions. This represents a concrete reason why human designers remain essential and should not be fully replaced by AI. We see this as an important and promising line of future work.

5.2 In-depth Discussion Avoidance

One of the key reasons for creativity dilution is that AI agents avoid in-depth discussion by rephrasing, rather than answering, questions posed by their paired AI agent. This is improved by embedding design thinking methods into the prompt in the second design iteration, as mentioned in Section 4. For example, prompts are used to enable AI agents to capture detailed information that can serve as potential design points and to ask relevant follow-up questions. However, AI agents may still fall short in capturing these nuances compared to experienced and creative human designers. More powerful LLMs may also improve this, which, however, is out of the scope of this paper.

5.3 Information Presentation

AI agents generate a large amount of text-based information that is hard for humans to review, which includes around 20,000 to 25,000 words per design activity for tackling a given design problem. We recognize this as the primary challenge among all challenges. The third design iteration addresses this by providing an interactive interface with an advanced chatbot, which we refer to as the *Project Assistant* (see Figure 1).

Knowing the benefits and challenges of the design activity simulation using multiple AI agents, we focus on addressing the primary challenge of providing an efficient information presentation interface for this multiple AI agents-based system. The rest of the paper focuses on introducing and evaluating the third version of the system, which is designed and developed based on the previous two iterations.

6 Addressing Presentational Concerns

To enable humans to interact with the content generated by the AI agents, we designed and developed an AI agent role as a project assistant. The project assistant can help human users process the large amount of text-based information generated by the rest of the AI agents. Retrieval-Augmented Generation (RAG) [49] is adopted to empower the AI project assistant.

RAG is a hybrid approach that combines the strengths of information retrieval and natural language generation to enhance human interaction with large text datasets. RAG works by retrieving relevant information from an external corpus based on a user query and then using a language model to generate a coherent, contextually relevant response. This technique allows for more

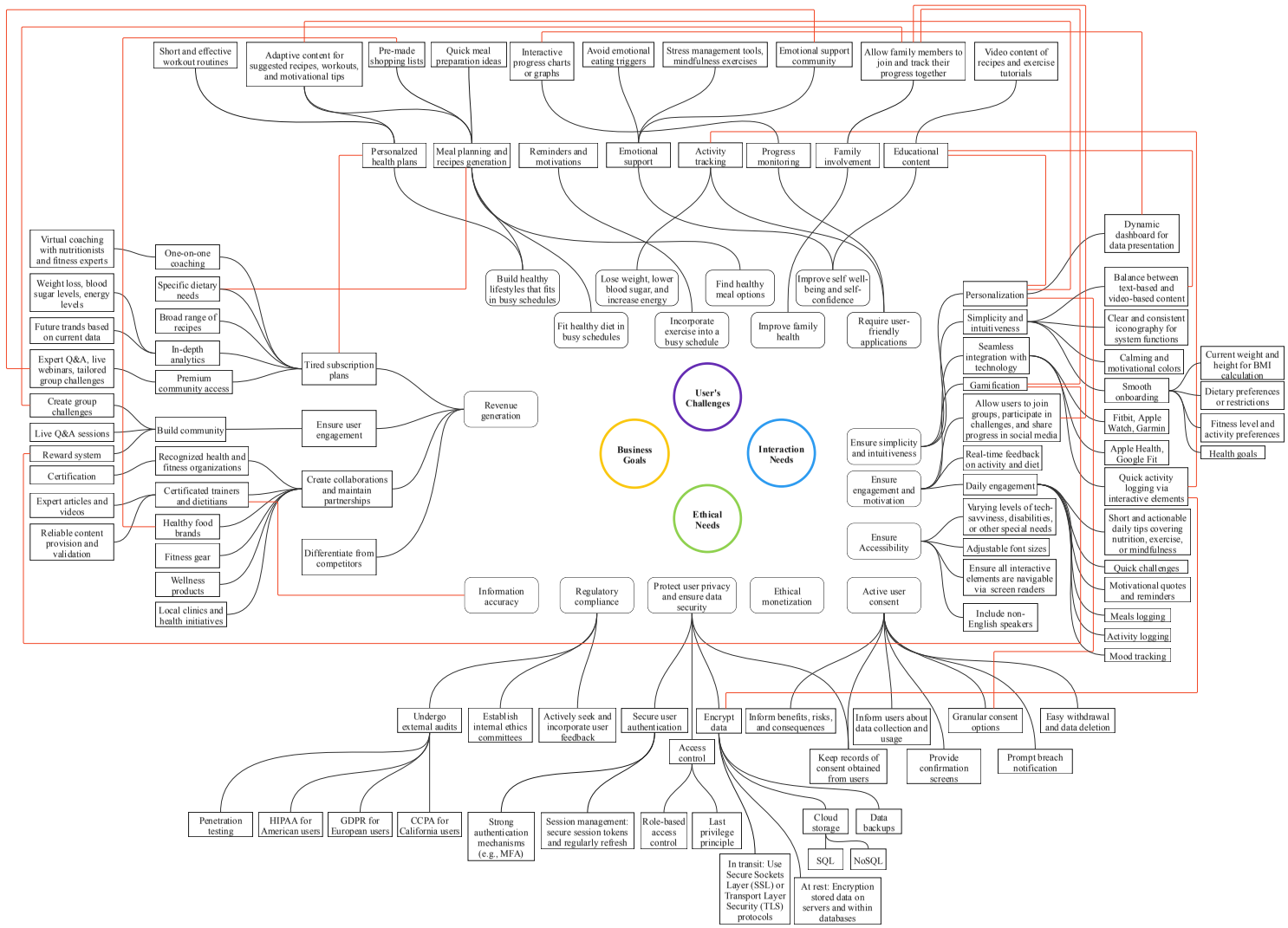


Figure 5: The design space visualization consists of four main aspects (circles), 17 requirements (rounded rectangles), and 100 solutions (rectangles), all extracted from the AI agents' conversation history. Black curves represent connections between requirements, solutions, and sub-solutions, while red polylines highlight shared solutions across different aspects. This figure demonstrates the AI agents' ability to comprehensively understand potential user needs, identify design constraints, and explore the design space based on a simple task description (*"Design an app that helps overweight people adopt a healthy lifestyle."*) within a relatively short time (20 – 25 minutes in this case) — a process that typically takes a human design team several weeks.

dynamic and accurate responses than standalone language models by grounding the generated text in specific, retrieved data.

In the context of our system, RAG facilitates human-AI interaction by enabling human users to query large, text-based design activity datasets generated by the AI agents and obtain meaningful summaries or detailed insights from the data. Figure 6 illustrates this mechanism. The retrieval component ensures that the generated responses are based on specific, relevant sections of the dataset, making it easier for users to explore and engage with a large amount of complex information.

7 Contextual Inquiry

In this user study, we conduct a contextual inquiry with seven experts from different backgrounds to understand how to design an interactive application that is useful for humans. We present each expert with a web application that displays the original design activity details along with an interactive AI project assistant agent (see the web application in Figure 1).

7.1 Study Process

Each participant will engage in a 1.5-hour contextual inquiry session using the provided web application, with the total session time controlled to stay within two hours. Participants are offered a £15 Amazon voucher as compensation for their time.

Considering the nature of generative AI, as emphasized by Weisz et al. [94], where outputs may vary in character or quality even when the input remains unchanged (referred to as generative variability), we assign the same design task to the AI agents for all participants: *Design an app that helps overweight people adopt a healthy lifestyle*. New design activities are generated for each participant to account for this variability.

Before the user study, the researcher explains the study's process and goals. The web application and its key functions are demonstrated, including navigating text materials and interacting with the AI project assistant.

Each participant takes on the role of a supervisor overseeing the AI agent team and completes three practice tasks at the beginning of the session, such as finding the designed feature in the PRD section, finding the motivation of a particular user in the Event section, and asking the AI project assistant agent to compare the different needs between two users (see the user interface in Figure 1). During the test, the researcher conducts a contextual inquiry by observing the participant's interactions and asking questions. After the session, participants complete a questionnaire evaluating the realism of the AI agents' ability to simulate human behavior and their understanding of events, as well as the quality of the task analysis. The user study concludes with an interview to gather feedback on the participant's experience with the web application.

Specifically, we categorize the seven events in the design problem-tackling task as follows:

User simulation and understanding events This includes clarifying user identities and gathering user needs. Participants evaluate how closely the AI user agents reflect reality, considering the realism of the generated personas and how well the AI agents' responses align with those personas.

Task analysis events This includes gathering business, ethical, technical, and design analyses, as well as generating the product requirements document (PRD). Participants evaluate the quality of ideas in terms of novelty, feasibility, relevance, and specificity.

Based on the following detailed evaluation criteria, the participant is asked to score the material from 1 to 10 (1–2: very poor, 3–4: poor, 5–6: neutral, 7–8: good, 9–10: very good).

For user simulation and understanding events, we propose two main dimensions and their sub-dimensions:

Persona Realism This dimension evaluates how accurately the AI-generated personas represent real-world users based on the given design task.
- Authenticity: The degree to which the generated persona accurately represents the given HCD task.
- Consistency: The consistency with which the AI user agent remains aligned with the persona throughout the interaction without deviating from the defined traits or motivations.
- Context Appropriateness: The appropriateness of the agent's responses based on the specific context of the HCD task or scenario.

User Role Realism This measures how well the AI agents adhere to and perform their assigned roles, simulating real user behavior during interactions.
- Role Alignment: How well the AI user agent adheres to its assigned role during interactions, such as answering questions like a real user in that role would answer.
- Engagement Realism: The extent to which the AI user agent demonstrates realistic engagement, including nuanced and contextually relevant responses typical of a human user in the same situation.

For task analysis events, we adopt the idea quality evaluation metrics [22] for this study, focusing on four main dimensions with several sub-dimensions: novelty (originality and paradigm relatedness), feasibility (acceptability and implementability), relevance (applicability and effectiveness), and specificity (explainability, completeness, and clarity).

7.2 Recruiting Participants

We recruited seven experts from different backgrounds to understand how different people might use this system. All participants' professional backgrounds are listed below:

E1 From the education technology field with four years of research experience in information science.
E2 One year of experience in consulting and three years of experience in political research.
E3 Eight years of experience in strategy consulting.
E4 Five years of experience as a business owner in electronics manufacturing with four years of research experience in material science.
E5 Four years of research experience in design engineering.
E6 Six years of experience in human-computer interaction research, user experience design, and mobile game design.
E7 20+ years of research experience in user-centered design, creativity, and drawing.

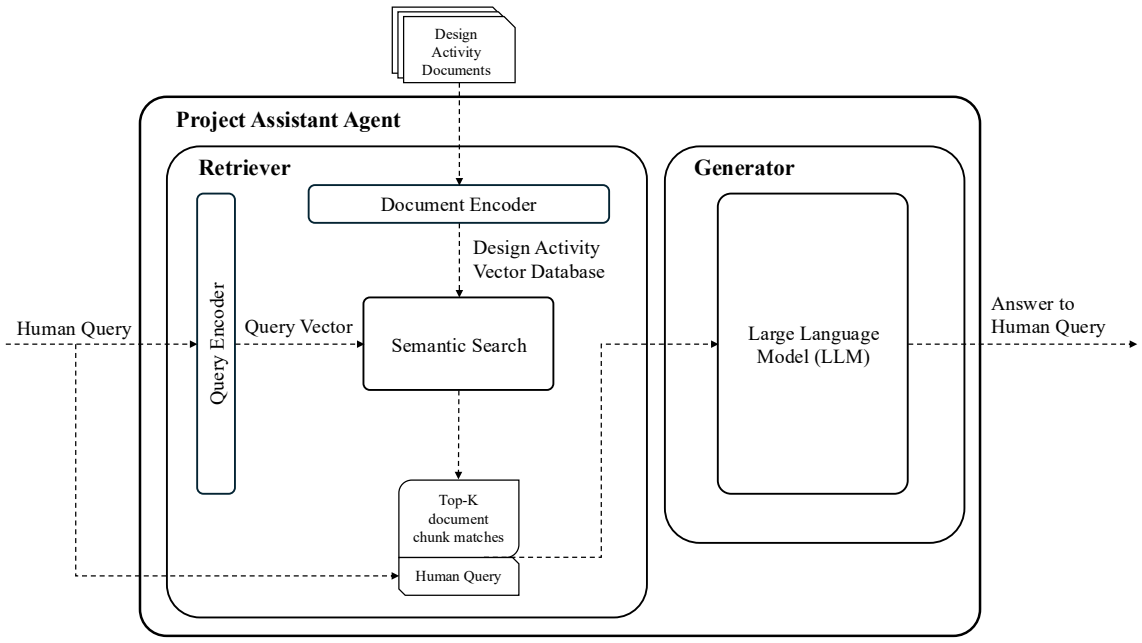


Figure 6: The retrieval-augmented generation (RAG) model for the AI project assistant agent. It consists of two parts: a retriever and a generator. The retriever encodes the human query and the text-based design activity documents into vectors, respectively. A semantic search is conducted to find top-K document chunk matches. These text chunks, along with the original human query, are sent to a LLM-based generator to produce the answer to the human query.

8 Results

8.1 Performance Ratings

Table 2 shows six experts rate the AI agents' average performance as good (E1–E6 scores are all above 7.0), and one expert (E7) thinks the overall performance is better than neutral. Interestingly, four non-designers (E1–E4) hold a similar view that the realism of personas and user roles are all above average, while three designers (E5–E7) hold an opposite view that categories of feasibility, relevance, and specificity are above average. This difference roughly reflects that designers and non-designers have different values in different aspects. Notably, all seven participants agreed that novelty is one of the lowest-performing categories of AI agents, with below-average scores.

Two designers (E6 and E7) think the persona realism is neutral or above neutral, and the rest of the participants believe it is good or very good. For each category – user role realism (E6), relevance (E3), and specificity (E4) – a different participant gave a rating above neutral. For novelty, all participants except E4 believe the AI agents perform poorly or neutrally. On the contrary, every participant rates feasibility as good or very good.

This result reflects that people with different backgrounds may have different standards and expectations for each category of AI agents' performance. Designers (E5, E6, E7) are more critical of user simulation and understanding events as they are the foundation of human-centered design. E6 and E7 acknowledge that the system can identify typical end-users who meet expectations. However, it does not include people who are not end-users but may still provide value to the design. For example, E7 mentioned that he may recruit people who live a healthy lifestyle, as they can be the goal for end-users, which can further inspire design. In addition, E5 and E6 notice that the AI user agents occasionally fail to follow their personas. However, the business owner (E4) values the implementation plan of the design so that it can be used directly for work arrangements. Consultants expect (E2, E3) more reliable data sources with richer details for business analysis so that they can better understand the actual market sizes, market gaps, and business innovation points.

Expert	Persona Realism	User Role Realism	Novelty	Feasibility	Relevance	Specificity	*Average*
E1	8.3	8.0	6.0	8.0	8.0	7.0	7.6
E2	9.3	10.0	5.0	8.0	9.0	9.0	8.4
E3	9.0	9.0	5.0	8.0	6.5	8.3	7.6
E4	8.3	9.0	7.5	9.0	7.0	6.7	7.9
E5	7.0	7.0	6.5	8.5	8.5	9.0	7.8
E6	6.7	6.5	5.0	9.5	8.0	7.7	7.2
E7	6.0	7.5	3.0	7.5	7.5	7.3	6.5

Table 2: Scores from all seven experts. Each score represents the average of subcategory scores within a given category. Scores above or equal to seven, indicating favorable evaluations, are shown in green text, while those below seven are shown in red text. Scores five to six are neutral, indicating the result is not necessarily bad. Additionally, the background is shaded green for scores above the participant's average, and red for those below.

In summary, the result reveals that the system can identify the typical end-users, but is not able to adopt advanced user study skills in recruiting participants out of the normal scope. The user simulation is accurate in general but occasionally shows hallucinations. Almost all experts think AI agents lack novelty but perform well in other aspects, except they fail to provide accurate data sources and merely generate a general project plan that can not be used to assign work to the human product team.

8.2 Interview Findings

8.2.1 AI agents ask high-quality questions and generate safe but not innovative ideas. When asked about the quality of AI agent-generated ideas, E1, E2, and E4 believe it is very high overall. E2 and E3 highlighted the quality of follow-up questions to the previous conversation turns. E3, E5, E6, and E7 think these ideas are good in terms of coverage (E3, E5, E6, E7), user needs understanding (E7), ethical analysis (E6), technical analysis (E6), and can self-correct (E5). However, the more prior experience participants have with tackling design problems, the less they feel the design generated by AI agents is innovative. Regarding innovation, consultants and designers hold different views. As mentioned by the consultant participant (E3),

> *"I've been in the industry for many years, and these solutions all feel very familiar. They can solve the problems, but they don't push beyond my current thinking or bring any innovative value. However, it is unfair to judge AI as not being creative in these projects. Actually, in real work, most projects only aim to play safe. Due to the realistic constraints of every project, we often sacrifice novelty and creativity to play safe. It performs at a similar creativity level as the projects I am involved in."*

However, the designer participant (E6) comments on the AI agents' outcome for the healthy lifestyle application design that,

> *"I would prioritize gamification because changing habits is not an easy task. It requires innovative elements to engage users enough to keep them using the product."*

8.2.2 AI agents efficiently generate comprehensive and effective information that helps humans inspect the design problem. All participants believe AI agents generate a lot of structured and comprehensive information that can effectively help humans systematically inspect the design problem. Particularly, E7 highlighted that,

> *"I actually think it's pretty good, like it gives lots of good lists. There are sorts of things that even a very insight-driven project might neglect, such as accessibility or data privacy. So that's good at making sure it's covered all the bases."*

E1, E2, E3, and E6 highlighted that they could use around 70% of the material generated by AI if assigned this design problem-tackling task, which they believe can save tremendous time in collecting information so they can spend more time on analyzing and innovating. E2 and E3 mentioned that it usually takes the team four weeks to conduct user interviews in an eight-week consulting project, while this system only costs two dollars and takes less than 25 minutes. E5 expressed that, even if the system can not replace a conventional user study with simulated users, it can still be used as a supplementary tool and replace certain portions of what real humans contribute to a user study. Interestingly, E4 expressed that, as a business owner, he would give the report generated by the AI agents to the product team to create a prototype so that they can test it rapidly using an agile development strategy.

8.2.3 Informative details raised in the AI agents' discussions are not distilled into clear value propositions . Regarding user needs understanding, E3, E5, E6, and E7 all mentioned that, sometimes, what users say may not fully reflect what they actually want. This implies that there are implicit value propositions hidden behind the conversation details. However, as E3 and E6 pointed out, junior human user researchers also find it challenging to distill fundamental design directions or value propositions that motivate and support the design from these implicit details.

8.2.4 AI agents use different communication styles than humans. E3, E5, and E6 indicated that the AI agents adopt a parallel questioning and answering strategy, where multiple questions or answers are asked or answered at a time, making it efficient for AI agents' communication while inefficient for humans to read and process. E3 also pointed out that,

> *"AI agents adopt a top-down strategy by providing concepts and abstract ideas or guidance first while offering more detailed information if asked. This is the opposite of humans' natural conversation habit, as we are good at giving examples and providing details but may find it challenging to summarize concepts quickly and correctly."*

The rationale behind this could be that the training data of current LLMs are recorded human experiences that can be regarded as a massive database [76]. Only when we clearly specify the question can LLMs retrieve the specific information in this massive database. This can be achieved by enabling AI agents to iteratively ask and answer follow-up questions, as introduced in Section 3.2.

8.2.5 Ten ways of using the AI project assistant agent. Through the user study, at least ten ways of using the AI project assistant were observed:

(1) Gathering detailed information on topics or concepts. - E1, E2, E3, E4, E5, E6, E7
(2) Finding the link between two pieces of information, such as system features and user needs. - E1, E2, E3, E4, E6, E7
(3) Helping the human participant understand the task or content, although it is not always successful. -E2, E3, E4, E5, E6, E7
(4) Retrieving information. - E1, E2, E4, E5, E6
(5) Summarizing information. - E2, E3, E4, E5, E6
(6) Filtering information. - E2, E4, E5, E6, E7
(7) Providing additional information. - E1, E2, E3, E5, E7
(8) Comparing different pieces of information to gain a better understanding of the design activity. - E1, E3, E7
(9) Running step-by-step role-playing simulation. - E5, E6
(10) Providing external information. - E5

All participants find it easy and efficient to interact with the AI project assistant agent. Particularly, E7 points out that misconceptions can trigger human inspiration. For example, E7 asks the agent *"What misconceptions might someone have about healthy lifestyles if they didn't do their research?"* and finds the reply inspirational (see Figure 1). However, asking high-quality questions requires a deep understanding of innovation and design theory in order to distill an inspirational design space from the design activity. Similarly, unique questioning skills, such as running step-by-step role-playing simulations, also require extensive design backgrounds. These observations from design experts can inform the future design iterations of the system.

8.2.6 Building trust with the system by cross-checking. Non-design experts, such as E1, E3, E4, and E6, carry out different extents of cross-checking at the beginning of the user study. After confirming that the content in the PRD, original event logs, and AI project assistant agent feedback are consistent, they are happy to trust the AI agents' output about the domains with which they are unfamiliar. E4 notes that,

> *"The process of cross-checking is pretty intuitive for humans to build trust when I am in my work. I naturally apply this technique to inspect AI agents' work."*

Similarly, E6 expresses that,

> *"I didn't trust the AI-based user research at the beginning, but I built trust after I checked the conversation."*

This finding reveals the mindset of people attempting to develop trust in such multiple AI agent systems.

8.2.7 The missing parts of the current system. All participants recognized the value of the system's current output but suggested that it could be further improved by addressing certain gaps to make it more applicable for real-world tasks. For instance, E1, E2, and E3 expressed the need for AI agents to provide more precise sources for the numbers mentioned in the PRD or conversations. They noted that the sources currently provided are too vague, making it difficult to directly locate the original information, which reduces their confidence in using the data. E4 suggested that the system offer a more detailed project plan, such as a comprehensive Gantt chart. The current plan is too basic and lacks the level of instruction needed for practical use. E3, E5, E6, and E7 pointed out that the text-based interaction with the system can be exhausting and unintuitive. They recommended incorporating more diagrams, tables, charts, figures, storyboards, and various data visualization techniques to help them quickly understand the core concepts of the generated design activity.

8.2.8 Increasing the level of control can improve system output and increase trust. The experts noted that the AI agents relieved them from tedious tasks within the design process, a benefit they appreciated in their role as "supervisors". However, they all agreed that increasing their level of control and allowing them to guide the direction of the AI agents' conversations would not only help build trust but also improve the quality of the output.

On the one hand, E5 and E7 mentioned that it is natural for users to pay less attention to things they are not involved with. Seeing the output is impacted by one's input can build a sense of ownership and increase trust. An important follow-up question is when and how people could provide input to the system.

In addition, E6 mentioned that the perceived issues with system output quality may due to the AI agents and humans not having consistently aligned goals. This misalignment can result in the AI agents focusing on aspects of the design problem that a human might consider to be of less value. This misalignment at an early stage of the design process can have an amplified impact on the result. E3 and E6 highlighted that allowing people to supervise AI agents' work at essential decision-making points can enable a higher goal and value alignment between humans and AI agents. They point out that the essential decision-making points are typically at the end of each event, and human users can supervise by commenting on or amending interim results. In this vein, people can save their cognitive resources for more important and higher-level decisions in the task. For example, deciding whether "voice input for meal logging" (proposed by AI agents) is a necessary system feature for the given design problem, rather than writing a detailed interview script.

9 Discussion

9.1 Paying Attention to the Early Design Stage

Existing LLM-based productivity and design tools [4, 8, 25, 35] are copilot systems providing fine-grained control and leveraging the power of AI to augment experienced designers in producing specific design artifacts.

Producing beautiful designs is important, and finding the correct design problem is equally important, if not more so. This is because we do not want to waste resources in designing a product to solve a problem that does not exist.

Therefore, we suggest paying more attention to the early design stage, which can potentially save a great amount of resources in terms of time and money. Our system focuses on the early stage, helping human users quickly understand user needs, market gaps, ethical risks, potential system features, user experience, and technical dependencies. It can serve as a consulting tool for individuals or organizations to better understand the design problem before they carry out any design and development operations.

9.2 Generalizability of the System

To assess the generalizability of our system, we tested it on additional design tasks beyond the initial evaluation domains. The outcomes demonstrated a similar level of performance, suggesting that the system's core mechanisms and prompting strategies are robust across contexts. This generalizability arises because the system's role-based agent structure and communication framework are designed to be domain-agnostic and task-independent, focusing on the process of divergent and convergent thinking rather than specific content knowledge. While the current system operates effectively without embedded domain expertise, future extensions could integrate specialized knowledge bases or expert models, enabling it to handle tasks that require deep domain-specific reasoning and further expanding its practical applicability. We found that the system generalizes well across different design contexts, maintaining stable performance due to its domain-agnostic mechanisms. In addition to the primary design case used in this study, we assigned

the system several other example design tasks to further assess its versatility: (1) *Design a virtual career coach system for MBA students.* (2) *Design a mobile health app to help elderly users manage chronic conditions.* (3) *Design a digital toolkit for small business to reduce carbon footprint.* (4) *Design an AI-powered personal finance assistant for Gen Z users.* Example outcomes for these tasks can be found in the Supplementary Material. Moreover, we provide open-sourced code so that readers can explore the system's performance on their own design tasks.

9.3 Strengths and Opportunities of Automated Design Activity

Existing LLM-based HCD tools and methods [76, 99, 101] leverage the power of AI to assist designers with information collection, idea generation, and design analysis at different stages of the design process. For example, a recent study [23] reveals that individuals working with AI performed at a quality level similar to two-person human teams without AI and that design teams can produce more balanced, cross-functional solutions with AI. In these approaches, AI plays the role of a tool to augment the designer's ability, which still requires designers to have a clear understanding of the design process and techniques, such as asking the right questions at the right time or providing appropriate prompts to these LLM-based tools.

Our work extends this by providing a system that covers a wide range of design processes with a holistic view of co-design, automatically carrying out the design process by identifying end users, gathering user needs, analyzing market opportunities, understanding ethical requirements, providing design ideas, and evaluating technical feasibility. AI agents are the design team members carrying out the actual design activity, while the humans are supervisors inspecting the AI agents' work. This new human-AI cooperative relationship introduces several major opportunities.

9.3.1 Reducing HCD Barriers. Human users are not required to have HCD knowledge to use this system, as design activities are carried out automatically by assigning a brief design problem description. This dramatically reduces the barrier of carrying out HCD for non-experts or even junior practitioners. As supervisors, humans can evaluate the outcomes rather than actively push the design process.

9.3.2 Improving Project Progress. Automated design problem-tackling activity dramatically increases efficiency. The high quality of the output from the AI agents allows the user to start a project with detailed materials rather than collect and analyze information from scratch. In turn, this gives the user more time and financial resources to focus on essential aspects based on the work of the AI agents.

9.3.3 Scaling ideation and exploration. The system can rapidly generate and evaluate a wide range of design directions, enabling exploration of larger design spaces than what human teams could reasonably cover on their own. This supports broader innovation opportunities and helps surface alternative solutions that may otherwise be overlooked.

9.3.4 Enhancing decision-making and focus. By automating large-scale information collection, synthesis, and preliminary design exploration, the system supports human designers in focusing their attention on higher-level, value-driven decisions. Rather than being consumed by routine or mechanical tasks, human designers can dedicate more effort to critical areas such as aligning the design with stakeholder needs, refining nuanced interactions, and ensuring long-term ethical and social impacts. Additionally, because the system explicitly surfaces ethical considerations as part of its design activity, it encourages human supervisors to engage with these issues proactively, rather than treating them as ad hoc or after-the-fact fixes.

9.3.5 Providing a testbed for human-AI collaboration. Beyond immediate practical benefits, this system offers a valuable platform for studying the dynamics of human-AI collaboration in creative tasks. It allows researchers to investigate when, where, and how human supervision adds value, how trust and control evolve, and how automation boundaries can be effectively balanced.

9.4 Limitations of the System

While the system demonstrates promising capabilities for automated design activity, several limitations must be acknowledged.

9.4.1 Creativity dilution. It is common to experience goal misalignment in a team, which can reduce the efficiency and the quality of teamwork [68]. AI agents in a fully automated workflow can also experience goal misalignment with human designers, often missing subtle or insightful design points that human creativity and intuition can easily capture. While the system explores a broad design space, it tends to converge on safe, conventional solutions, limiting the novelty and divergence of outcomes.

9.4.2 Modality constraints. The current system operates through text-based interactions and struggles with design tasks that require visual [42], spatial [98], or embodied reasoning [53], such as creating physical prototypes [63], interpreting gestures [57, 83], or accounting for environmental context [17]. Although recent studies have individually explored these design modalities [70, 86, 92], expanding the existing system's capabilities to effectively handle multimodal design tasks remains an open challenge.

9.4.3 Generative biases. Although the AI agents cover a wide range of design roles, their outputs are ultimately constrained by the underlying LLMs [36]. This can lead to biases inherited from training data, lack of cultural or contextual nuance, and a tendency to produce familiar or widely available solutions, potentially limiting the novelty and originality of the outcomes.

9.5 Risks of Automated Design Activity

While automated design offer great opportunities, the deployment also introduces important risks that need to be critically considered.

9.5.1 Reduced situation awareness. Automated systems can lower the user's ability to monitor and stay conscious of all relevant aspects of the task, a capacity known as *situation awareness* [34]. As AI agents autonomously handle complex design activities, human designers may lose track of how decisions are being made, which options are being considered or excluded, and when critical

turning points arise. This can lead to designers operating with an incorrect or incomplete understanding of the system's current state, increasing the chance of overlooking errors, biases, or emerging issues.

9.5.2 Complacency and skill degradation. There is a risk that designers may excessively delegate tasks to automation, reducing meaningful human involvement to passive oversight [41]. Over time, this can erode human agency and degrade essential design skills, as designers and researchers become less engaged in critical tasks such as problem framing, evaluation, and creative exploration. Additionally, as users grow accustomed to the system's high-speed, high-volume outputs, they may develop a false sense of confidence in the AI's results, overlooking the need for scrutiny and missing the nuanced understanding, empathy, and contextual sensitivity that only human designers can provide.

9.5.3 Automation reliability and error costs. The system's reliability depends on the underlying models, prompts, and mechanisms [36, 50], which can produce flawed or biased outputs, especially when encountering edge cases or poorly defined problems. Errors introduced by the system can propagate through the design process, potentially incurring significant downstream costs in terms of rework, misaligned products, or user dissatisfaction.

Overall, these risks we have raised here not only highlight practical challenges but also open important avenues for future exploration of human-AI cooperation in design. Specifically, they raise questions about how to effectively balance automation and human oversight, how to design interaction mechanisms that preserve human skills and situational awareness, and how to ensure accountability and trust calibration in collaborative AI systems, offering a rich agenda for future research.

9.6 Potential Users of the System

As part of the contextual inquiry, the first author discussed the potential users of the system with the experts.

9.6.1 Design students. As suggested by E5, the system's comprehensive and structured approach makes it a promising tool for teaching design students how to tackle complex design problems. The recorded conversations between AI agents can serve as rich case study materials, helping students understand various aspects of the design process, including problem framing, qualitative data analysis, and interdisciplinary collaboration.

9.6.2 User researchers and designers. E3 and E7 appreciate the high speed and low cost of generating simulated users and note how the system could potentially be used by user researchers and designers as a supplementary approach to traditional user studies. We suggest that simulated users could be used either before, during, or after the user study to prepare better for the user interview or to understand the boundary of the end-user group. However, we should not arbitrarily use AI agents as the only approach to understanding user needs. One of the reasons is that LLMs can also be biased, considering the training data may not equally represent each group. Another reason is that simulated users may not fully represent the real users, even when assigning a real user's persona to an AI agent. Moreover, designers can gain extra insights from nonverbal observations.

9.6.3 Consultants, product managers, and project managers. Suggested by E2, E3, and E4, this system can also potentially be used by consultants, product managers, and project managers to swiftly gain an overview of project ideas at an early design stage. This can substantially reduce the time spent collecting fundamental information and clarifying initial directions. They can choose either to accept or reject ideas generated by the AI agents by judging their quality and validity, having established some trust in the fact that the ideas are coherent and meaningful but may lack novelty.

10 Future Work

This work opens up several promising directions for future research, both in evaluation design and system development.

First, future studies should involve a larger and more diverse sample of participants to enhance the generalizability of the findings across different user groups, domains, and contexts. Expanding the participant base would strengthen the robustness and external validity of the system's performance evaluations. Additionally, refining the evaluation framework with a more fine-grained breakdown of assessment criteria, along with more detailed statistical analyses, would provide deeper insights into the system's strengths and limitations. This would help researchers more effectively identify areas for improvement and guide future system iterations.

Beyond evaluation improvements, several system-level enhancements are necessary to improve usability, control, and user experience. The current fully automated design process allows humans only to monitor the system and conduct ad hoc evaluations, which can reduce both the usability of outputs and users' sense of control. Moreover, the system's long, text-based outputs substantially impede user experience when humans attempt to inspect and make sense of the generated materials.

These observations reveal three critical design requirements for future work:

First, **design for efficient and effective system control.** Allowing AI agents to run automatically in detailed tasks while increasing human control at critical decision-making points can ensure efficiency and effectiveness at the same time. Moreover, providing semi-structured interaction guidance for human users can help with on-boarding and reduce mental load.

Second, **design for ownership and user trust.** In addition to allowing humans to steer the design activity, enabling easy cross-checking of the information presented in the system can also build trust. This can be achieved by, for example, providing source links to specific data or allowing users to rapidly find information generated by AI agents in the design activity when it is referenced by other AI agents.

Third, **use a variety of information presentation techniques.** We wish to ensure generated design solutions are correctly and efficiently conveyed to humans. This can be achieved by presenting information through tables, figures, and charts that improve the readability of the system output for humans, and by providing storyboards, wireframes, and sketches that make the design solutions more understandable.

11 Conclusion

This paper has demonstrated the feasibility of automating cooperative knowledge work through the use of multiple LLM-based communicative AI agents. We designed, developed, refined, and evaluated a system of LLM-based AI agents for tackling design problems. The system is capable of exploring design spaces, identifying constraints, and generating effective and feasible design ideas based on a HCD problem given by the user. The system can also interactively present this information via a web application with a chatbot-like assistant agent. This approach enables users to quickly gain a comprehensive understanding of a design problem at an early stage, incorporating multiple perspectives such as business requirements, ethical considerations, interaction design, and technical feasibility. This established understanding provides a rich foundation for further design focusing on the aspects that AI agents are not good at, such as novelty.

This approach is both low-cost and efficient, with each run costing approximately two dollars and taking less than half an hour—substantially faster and more affordable than traditional design activities. Conventional methods typically span at least one to two months and require significant financial investment to involve experts across various domains and organize user studies. However, we emphasize that we should not replace a conventional HCD process with an approach based on AI agents. Instead, the current method is a supplementary means for undertaking HCD.

Having demonstrated the potential for AI agents to simulate the way we carry out design activities, our future work will focus on improving the usability of the system and exploring how to better integrate human input into the AI agents' workflow. This is likely to involve improving data visualization, finding optimal levels of automation to ensure efficiency, and increasing user control to enhance the quality of the design activity outputs.

12 Open Science

To stimulate further research on this topic, we make our code available here: https://github.com/boyiny/design_activity_simulation.

References

[1] Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. 2023. LLM-Deliberation: Evaluating LLMs with Interactive Multi-Agent Negotiation Games. https://doi.org/10.60882/cispa.25233028.v1

[2] Stefano V. Albrecht, Cillian Brewitt, John Wilhelm, Balint Gyevnar, Francisco Eiras, Mihai Dobre, and Subramanian Ramamoorthy. 2021. Interpretable Goal-based Prediction and Planning for Autonomous Driving. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, Xi'an, China, 1043–1049. https://doi.org/10.1109/ICRA48506.2021.9560849

[3] Stefano V Albrecht, Filippos Christianos, and Lukas Schäfer. 2024. *Multi-Agent Reinforcement Learning: Foundations and Modern Approaches*. MIT Press, Cambridge, Massachusetts.

[4] Shm Garanganao Almeda, J. D. Zamfirescu-Pereira, Kyu Won Kim, Pradeep Mani Rathnam, and Bjoern Hartmann. 2024. Prompting for Discovery: Flexible Sense-Making for AI Art-Making with Dreamsheets. arXiv:2310.09985 [cs.HC] https://arxiv.org/abs/2310.09985

[5] Esteve Almirall and Jonathan Wareham. 2008. Living Labs and Open Innovation: Roles and Applicability. *eJOV: The Electronic Journal for Virtual Organization & Networks* 10 (2008), 21–46.

[6] Esteve Almirall and Jonathan Wareham. 2011. Living Labs: Arbiters of Mid- and Ground-Level Innovation. *Technology Analysis & Strategic Management* 23, 1 (2011), 87–102.

[7] Rajeev Alur, Thomas A Henzinger, and Orna Kupferman. 2002. Alternating-Time Temporal Logic. *Journal of the ACM (JACM)* 49, 5 (2002), 672–713.

[8] Tyler Angert, Miroslav Suzara, Jenny Han, Christopher Pondoc, and Hariharan Subramonyam. 2023. Spellburst: A Node-based Interface for Exploratory Creative Coding with Natural Language Prompts. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 100, 22 pages. https://doi.org/10.1145/3586183.3606719

[9] Paul Bate and Glenn Robert. 2006. Experience-based Design: from Redesigning the System Around the Patient to Co-designing Services with the Patient. *BMJ quality & safety* 15, 5 (2006), 307–310.

[10] Kent Beck, Mike Beedle, Arie Van Bennekum, Alistair Cockburn, Ward Cunningham, Martin Fowler, James Grenning, Jim Highsmith, Andrew Hunt, and Ron Jeffries. 2001. Manifesto for Agile Software Development.

[11] M Bird, M McGillion, EM Chambers, J Dix, CJ Fajardo, M Gilmour, K Levesque, A Lim, S Mierdel, C Ouellette, et al. 2021. A Generative Co-design Framework for Healthcare Innovation: Development and Application of an End-user Engagement Framework. *Research Involvement and Engagement* 7 (2021), 1–12.

[12] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2022. On the Opportunities and Risks of Foundation Models. arXiv:2108.07258 [cs.LG] https://arxiv.org/abs/2108.07258

[13] Tim Brown. 2008. Design Thinking. *Harvard business review* 86, 6 (2008), 84.

[14] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models are Few-shot Learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[15] John M Bryson. 2004. What to Do When Stakeholders Matter: Stakeholder Identification and Analysis Techniques. *Public management review* 6, 1 (2004), 21–53.

[16] Tony Buzan and Barry Buzan. 2006. *The Mind Map Book*. Pearson Education, London, UK.

[17] Junlong Chen, Jens Grubert, and Per Ola Kristensson. 2025. Analyzing Multimodal Interaction Strategies for LLM-Assisted Manipulation of 3D Scenes. In *2025 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*. IEEE, Saint-Malo, France, 206–216.

[18] Henry William Chesbrough. 2003. *Open Innovation: The New Imperative for Creating and Profiting from Technology*. Harvard Business Press, USA.

[19] R. S. Contreras-Espinosa, A. Frisiello, J. L. Eguia-Gomez, and A. Blanco. 2023. Co-creation, Co-design, and Co-production: Enablers and Barriers for Implementation and Use of Digital Technologies. In *Communication and Applied Technologies*, Paulo Carlos López-López, Daniel Barredo, Ángel Torres-Toukoumidis, Andrea De-Santis, and Óscar Avilés (Eds.). Springer Nature Singapore, Singapore, 81–90.

[20] Design Council. 2003. *The Double Diamond*. Design Council. Retrieved July 7, 2024 from https://www.designcouncil.org.uk/our-resources/the-double-diamond/

[21] Edward De Bono. 2017. *Six Thinking Hats: The Multi-million Bestselling Guide to Running Better Meetings and Making Faster Decisions*. Penguin, UK.

[22] Douglas L Dean, Jill Hender, Tom Rodgers, and Eric Santanen. 2006. Identifying Good Ideas: Constructs and Scales for Idea Evaluation. *Journal of Association for Information Systems* 7, 10 (2006), 646–699.

[23] Fabrizio Dell'Acqua, Charles Ayoubi, Hila Lifshitz, Raffaella Sadun, Ethan Mollick, Lilach Mollick, Yi Han, Jeff Goldman, Hari Nair, Stewart Taub, et al. 2025. *The Cybernetic Teammate: A Field Experiment on Generative AI Reshaping Teamwork and Expertise*. Technical Report. National Bureau of Economic Research.

[24] Sebastian Deterding, Dan Dixon, Rilla Khaled, and Lennart Nacke. 2011. From Game Design Elements to Gamefulness: Defining "Gamification". In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments* (Tampere, Finland) *(MindTrek '11)*. Association for Computing Machinery, New York, NY, USA, 9–15. https://doi.org/10.1145/2181037.2181040

[25] Giulia Di Fede, Davide Rocchesso, Steven P. Dow, and Salvatore Andolina. 2022. The Idea Machine: LLM-based Expansion, Rewriting, Combination, and Suggestion of Ideas. In *Proceedings of the 14th Conference on Creativity and Cognition* (Venice, Italy) *(C&C '22)*. Association for Computing Machinery, New York, NY, USA, 623–627. https://doi.org/10.1145/3527927.3535197

[26] Sara Donetto, Vicki Tsianakas, Glenn Robert, et al. 2014. Using Experience-based Co-design (EBCD) to Improve the Quality of Healthcare: Mapping Where We are Now and Establishing Future Directions. , 5–7 pages.

[27] Daniel S Drew. 2021. Multi-Agent Systems for Search and Rescue Applications. *Current Robotics Reports* 2 (2021), 189–200.

[28] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving Factuality and Reasoning in Language Models through Multiagent Debate. arXiv:2305.14325 [cs.CL] https://arxiv.org/abs/2305.14325

[29] Joel Dyer, Arnau Quera-Bofarull, Ayush Chopra, J. Doyne Farmer, Anisoara Calinescu, and Michael Wooldridge. 2023. Gradient-Assisted Calibration for Financial Agent-Based Models. In *Proceedings of the Fourth ACM International Conference on AI in Finance* (Brooklyn, NY, USA) *(ICAIF '23)*. Association for Computing Machinery, New York, NY, USA, 288–296. https://doi.org/10.1145/3604237.3626857

[30] Bob Eberle. 1996. *Scamper on: Games for Imagination Development*. Prufrock Press Inc., Austin, TX.

[31] Amy C Edmondson and Ingrid M Nembhard. 2009. Product Development and Learning in Project Teams: The Challenges are the Benefits. *Journal of product innovation management* 26, 2 (2009), 123–138.

[32] Eva Eigner and Thorsten Händler. 2024. Determinants of LLM-assisted Decision-Making. arXiv:2402.17385 [cs.AI] https://arxiv.org/abs/2402.17385

[33] B-N Sanders Elizabeth and Uday Dandavate. 1999. Design for Experiencing: New Tools.

[34] Mica R Endsley. 2021. Situation Awareness. , 434–455 pages.

[35] Ziv Epstein, Hope Schroeder, and Dava Newman. 2022. When Happy Accidents Spark Creativity: Bringing Collaborative Speculation to Life with Generative AI. arXiv:2206.00533 [cs.HC] https://arxiv.org/abs/2206.00533

[36] Mohamed Amine Ferrag, Norbert Tihanyi, and Merouane Debbah. 2025. From LLM Reasoning to Autonomous AI Agents: A Comprehensive Review.

[37] Bill Gaver, Tony Dunne, and Elena Pacenti. 1999. Design: Cultural Probes. *interactions* 6, 1 (1999), 21–29.

[38] Ivan Gochev, Gorjan Nadzinski, and Mile Stankovski. 2017. Path Planning and Collision Avoidance Regime for A Multi-Agent System in Industrial Robotics. *Machines. Technologies. Materials.* 11, 11 (2017), 519–522.

[39] Theresa Green, Ann Bonner, Laisa Teleni, Natalie Bradford, Louise Purtell, Clint Douglas, Patsy Yates, Margaret MacAndrew, Hai Yen Dao, and Raymond Javan Chan. 2020. Use and Reporting of Experience-based Codesign Studies in the Healthcare Setting: A Systematic Review. *BMJ quality & safety* 29, 1 (2020), 64–76.

[40] Mais Haj Qasem, Mohammad Aljaidi, Ghassan Samara, Raed Alazaidah, Ayoub Alsarhan, and Mohammed Alshammari. 2023. An Intelligent Decision Support System Based on Multi Agent Systems for Business Classification Problem. *Sustainability* 15, 14 (2023), 10977.

[41] Lydia Harbarth, Eva Gößwein, Daniel Bodemer, and Lenka Schnaubert. 2025. (Over) Trusting AI Recommendations: How System and Person Variables Affect Dimensions of Complacency. *International Journal of Human–Computer Interaction* 41, 1 (2025), 391–410.

[42] Lihua Huang and Peng Zheng. 2023. Human-Computer Collaborative Visual Design Creation Assisted by Artificial Intelligence. *ACM transactions on Asian and low-resource language information processing* 22, 9 (2023), 1–21.

[43] Yuling Huang, Chujin Zhou, Kai Cui, and Xiaoping Lu. 2024. A Multi-Agent Reinforcement Learning Framework for Optimizing Financial Trading Strategies Based on TimesNet. *Expert Systems with Applications* 237 (2024), 121502.

[44] Tra Huynh, Adrian Madsen, Sarah McKagan, and Eleanor Sayre. 2021. Building Personas from Phenomenography: A Method for User-Centered Design in Education. *Information and Learning Sciences* 122, 11/12 (2021), 689–708.

[45] Barbara A Israel, Amy J Schulz, Edith A Parker, and Adam B Becker. 1998. Review of Community-based Research: Assessing Partnership Approaches to Improve Public Health. *Annual review of public health* 19, 1 (1998), 173–202.

[46] Robert Jungk and Norbert Müllert. 1987. *Future Workshops: How to Create Desirable Futures*. Inst. for Social Inventions, London, UK.

[47] Finn Kensing and Jeanette Blomberg. 1998. Participatory Design: Issues and Concerns. *Computer supported cooperative work (CSCW)* 7 (1998), 167–185.

[48] Finn Kensing and Halskov Madsen. 2020. Generating Visions: Future Workshops and Metaphorical. In *Design at work*. CRC Press, Boca Raton, Florida, 155–168.

[49] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.

[50] Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative Agents for " Mind" Exploration of Large Language Model Society. *Advances in Neural Information Processing Systems* 36 (2023), 51991–52008.

[51] Jie Li, Hancheng Cao, Laura Lin, Youyang Hou, Ruihao Zhu, and Abdallah El Ali. 2024. User Experience Design Professionals' Perceptions of Generative Artificial Intelligence. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 381, 18 pages. https://doi.org/10.1145/3613904.3642114

[52] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate. arXiv:2305.19118 [cs.CL] https://arxiv.org/abs/2305.19118

[53] Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. 2024. Aligning Cyber Space with Physical World: A Comprehensive Survey on Embodied AI.

[54] Qiuyu Lu, Jiawei Fang, Zhihao Yao, Yue Yang, Shiqing Lyu, Haipeng Mi, and Lining Yao. 2024. Enabling Generative Design Tools with LLM Agents for Building Novel Devices: A Case Study on Fluidic Computation Interfaces. arXiv:2405.17837 [cs.HC] https://arxiv.org/abs/2405.17837

[55] Cai Luo, Andre Possani Espinosa, Danu Pranantha, and Alessandro De Gloria. 2011. Multi-Robot Search and Rescue Team. In *2011 IEEE International Symposium on Safety, Security, and Rescue Robotics*. IEEE, Kyoto, Japan, 296–301.

[56] Yuan Luo, Kecheng Liu, and Darryl N Davis. 2002. A Multi-agent Decision Support System for Stock Trading. *IEEE network* 16, 1 (2002), 20–27.

[57] Zhuoyue Lyu and Per Ola Kristensson. 2025. Objestures: Bimanual Interactions with Everyday Objects and Mid-Air Gestures in Mixed Reality (Video Showcase). In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 917, 2 pages. https://doi.org/10.1145/3706599.3721347

[58] Tomasz Miaskiewicz, Susan Jung Grant, and Kenneth A Kozar. 2009. A Preliminary Examination of Using Personas to Enhance User-Centered Design. , 697 pages.

[59] Tomasz Miaskiewicz and Kenneth A Kozar. 2011. Personas and User-Centered Design: How Can Personas Benefit Product Design Processes? *Design studies* 32, 5 (2011), 417–430.

[60] Pietro Micheli, Sarah JS Wilner, Sabeen Hussain Bhatti, Matteo Mura, and Michael B Beverland. 2019. Doing Design Thinking: Conceptual Review, Synthesis, and Research Agenda. *Journal of Product innovation management* 36, 2 (2019), 124–148.

[61] Michael J. Muller. 2002. *Participatory Design: the Third Space in HCI*. L. Erlbaum Associates Inc., USA, 1051–1068.

[62] Roger B Myerson. 1977. Graphs and Cooperation in Games. *Mathematics of operations research* 2, 3 (1977), 225–229.

[63] Jim Philippe Neussl, Rebecca Zheng, George Hanson, and Boyin Yang. 2019. TechBuddies: Engaging Students to Teach Retirees about Technology. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3309694

[64] Zixiang Nie, Kwang-Cheng Chen, and Kyeong Jin Kim. 2024. Social-Learning Coordination of Collaborative Multi-Robot Systems Achieves Resilient Production in a Smart Factory. *IEEE Transactions on Automation Science and Engineering* 22 (2024), 1–15. https://doi.org/10.1109/TASE.2024.3435443

[65] Damilola Oluwaseun Ogundipe, Sodiq Odetunde Babatunde, and Emmanuel Adeyemi Abaku. 2024. AI and Product Management: A Theoretical Overview from Idea to Market. *International Journal of Management & Entrepreneurship Research* 6, 3 (2024), 950–969.

[66] Alex F Osborn. 1953. *Applied Imagination*. Scribner's, New York, USA.

[67] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 2, 22 pages. https://doi.org/10.1145/3586183.3606763

[68] Christopher OLH Porter, Brittney Amber, and Ernie Wang. 2019. Team Motivation and Goal (Mis) Alignment: The Missing Link in Human Capital Resources Research. In *Handbook of research on strategic human capital resources*. Edward Elgar Publishing, Cheltenham, 315–337.

[69] Coimbatore K Prahalad and Venkat Ramaswamy. 2004. Co-creating Unique Value with Customers. *Strategy & leadership* 32, 3 (2004), 4–9.

[70] Nathalie Riche, Anna Offenwanger, Frederic Gmeiner, David Brown, Hugo Romat, Michel Pahud, Nicolai Marquardt, Kori Inkpen, and Ken Hinckley. 2025. AI-Instruments: Embodying Prompts as Instruments to Abstract & Reflect Graphical Interface Commands as General-Purpose Tools.

[71] Steven R. Rick, Gianni Giacomelli, Haoran Wen, Robert J. Laubacher, Nancy Taubenslag, Jennifer L. Heyman, Max Sina Knicker, Younes Jeddi, Hendrik Maier, Stephen Dwyer, Pranav Ragupathy, and Thomas W. Malone. 2023. Supermind Ideator: Exploring Generative AI to Support Creative Problem-Solving. arXiv:2311.01937 [cs.AI] https://arxiv.org/abs/2311.01937

[72] Danissa V Rodriguez, Katharine Lawrence, Javier Gonzalez, Beatrix Brandfield-Harvey, Lynn Xu, Sumaiya Tasneem, Defne L Levine, and Devin Mann. 2024. Leveraging Generative AI Tools to Support the Development of Digital Solutions in Health Care Research: Case Study. *JMIR Human Factors* 11, 1 (2024), e52885.

[73] Elizabeth B-N Sanders and Pieter Jan Stappers. 2008. Co-creation and the New Landscapes of Design. *Co-design* 4, 1 (2008), 5–18.

[74] EB-N Sanders. 2000. Generative Tools for Co-designing. In *Collaborative design: proceedings of codesigning 2000*. Springer, London, United Kingdon, 3–12.

[75] David Satcher. 2005. *Methods in Community-based Participatory Research for Health*. John Wiley & Sons, San Francisco, CA, USA.

[76] Albrecht Schmidt, Passant Elagroudy, Fiona Draxler, Frauke Kreuter, and Robin Welsch. 2024. Simulating the Human in HCD with ChatGPT: Redesigning Interaction Design with AI. *Interactions* 31, 1 (2024), 24–31.

[77] Douglas Schuler and Aki Namioka. 1993. *Participatory Design: Principles and Practices*. CRC press, New Jersey, USA.

[78] Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. 2016. Safe, Multi-Agent, Reinforcement Learning for Autonomous Driving. arXiv:1610.03295 [cs.AI] https://arxiv.org/abs/1610.03295

[79] Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-LLM: A Trainable Agent for Role-Playing. arXiv:2310.10158 [cs.CL] https://arxiv.org/abs/2310.10158

[80] Lloyd S Shapley. 1953. *A Value for N-Person Games*. Princeton University Press Princeton, Princeton, New Jersey, USA.

[81] Helen Sharp, Jenny Preece, and Yvonne Rogers. 2023. Interaction Design: Beyond Human-Computer Interaction.

[82] Peter Slattery, Alexander K Saeri, and Peter Bragge. 2020. Research Co-design in Health: A Rapid Overview of Reviews. *Health research policy and systems* 18 (2020), 1–13.

[83] Zhaomou Song, John J Dudley, and Per Ola Kristensson. 2025. Investigating Visualization and Control for Assisting Gesture Interaction in Virtual Reality. In *2025 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, Saint-Malo, France, 1218–1219.

[84] Clay Spinuzzi. 2005. The Methodology of Participatory Design. *Technical communication* 52, 2 (2005), 163–174.

[85] Joachim Stempfle and Petra Badke-Schaub. 2002. Thinking in Design Teams - An Analysis of Team Communication. *Design studies* 23, 5 (2002), 473–496.

[86] Sirui Tao, Ivan Liang, Cindy Peng, Zhiqing Wang, Srishti Palani, and Steven P. Dow. 2025. DesignWeaver: Dimensional Scaffolding for Text-to-Image Product Design. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 425, 26 pages. https://doi.org/10.1145/3706598.3714211

[87] U.S. Government. 2025. Human-Centered Design - Performance.gov. https://www.performance.gov/cx/hcd/ Accessed: 2025-02-22.

[88] Wiebe van der Hoek and Michael Wooldridge. 2002. Tractable Multiagent Planning for Epistemic Goals. In *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems: Part 3* (Bologna, Italy) *(AAMAS '02)*. Association for Computing Machinery, New York, NY, USA, 1167–1174. https://doi.org/10.1145/545056.545095

[89] Froukje Sleeswijk Visser, Pieter Jan Stappers, Remko Van der Lugt, and Elizabeth BN Sanders. 2005. Contextmapping: Experiences from Practice. *CoDesign* 1, 2 (2005), 119–149.

[90] Eric Von Hippel. 1986. Lead users: a source of novel product concepts. *Management science* 32, 7 (1986), 791–805.

[91] Samangi Wadinambiarachchi, Ryan M. Kelly, Saumya Pareek, Qiushi Zhou, and Eduardo Velloso. 2024. The Effects of Generative AI on Design Fixation and Divergent Thinking. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 380, 18 pages. https://doi.org/10.1145/3613904.3642919

[92] Zijun Wan, Jiawei Tang, Linghang Cai, Xin Tong, and Can Liu. 2024. Breaking the Midas Spell: Understanding Progressive Novice-AI Collaboration in Spatial Design.

[93] Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2024. Unleashing the Emergent Cognitive Synergy in Large Language Models: A Task-Solving Agent through Multi-Persona Self-Collaboration. arXiv:2307.05300 [cs.AI] https://arxiv.org/abs/2307.05300

[94] Justin D. Weisz, Jessica He, Michael Muller, Gabriela Hoefer, Rachel Miles, and Werner Geyer. 2024. Design Principles for Generative AI Applications. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 378, 22 pages. https://doi.org/10.1145/3613904.3642466

[95] Chauncey Wilson. 2013. *Interview Techniques for UX Practitioners: A User-Centered Design Method*. Newnes, Oxford, UK.

[96] Michael Wooldridge. 2009. *An Introduction to Multiagent Systems*. John wiley & sons, Chichester, West Sussex, United Kingdom.

[97] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 385, 22 pages. https://doi.org/10.1145/3491102.3517582

[98] Yue Yang, Fan-Yun Sun, Luca Weihs, Eli VanderBilt, Alvaro Herrasti, Winson Han, Jiajun Wu, Nick Haber, Ranjay Krishna, Lingjie Liu, et al. 2024. Holodeck: Language Guided Generation of 3D Embodied AI Environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Seattle, USA, 16227–16237.

[99] Eric York. 2023. Evaluating ChatGPT: Generative AI in UX Design and Web Development Pedagogy. In *Proceedings of the 41st ACM International Conference on Design of Communication* (Orlando, FL, USA) *(SIGDOC '23)*. Association for Computing Machinery, New York, NY, USA, 197–201. https://doi.org/10.1145/3615335.3623035

[100] Shiyao Zhang, Yuji Dong, Yichuan Zhang, Terry R. Payne, and Jie Zhang. 2024. Large Language Model Assissted Multi-Agent Dialogue for Ontology Alignment. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems* (Auckland, New Zealand) *(AAMAS '24)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2594–2596.

[101] Ruican Zhong, Donghoon Shin, Rosemary Meza, Predrag Klasnja, Lucas Colusso, and Gary Hsieh. 2024. AI-Assisted Causal Pathway Diagram for Human-Centered Design. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 2, 19 pages. https://doi.org/10.1145/3613904.3642179

[102] Lixi Zhu, Xiaowen Huang, and Jitao Sang. 2024. How Reliable is Your Simulator? Analysis on the Limitations of Current LLM-based User Simulators for Conversational Recommendation. In *Companion Proceedings of the ACM Web Conference 2024* (Singapore, Singapore) *(WWW '24)*. Association for Computing Machinery, New York, NY, USA, 1726–1732. https://doi.org/10.1145/3589335.3651955