Synthetic TLX: Forecasting Human Workload Using Agent Simulation

Tzu-Sheng Kuo*
tzushenk@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Carrie J. Cai
cjcai@google.com
Google DeepMind
Mountain View, CA, USA

Meredith Ringel Morris
merrie@google.com
Google DeepMind
Seattle, WA, USA

Michael Terry
michaelterry@google.com
Google DeepMind
Cambridge, MA, USA

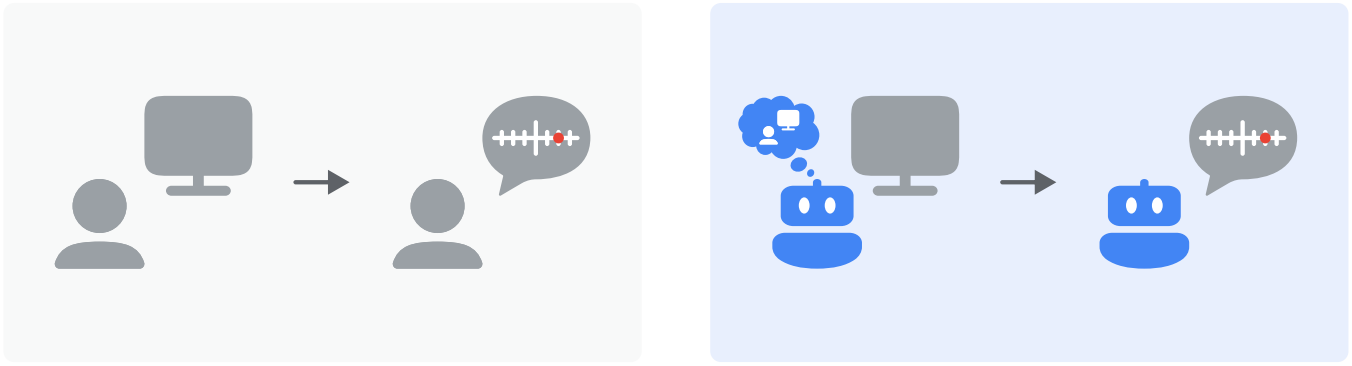


Figure 1: We introduce Synthetic TLX, a proactive paradigm for human workload estimation. As illustrated, the status quo requires humans to complete a task and retrospectively rate their experienced workload. In contrast, Synthetic TLX uses AI agents to simulate a human executing the task and forecast the workload before a person even attempts it.

Abstract

Assessing human workload for technology-mediated tasks helps prevent task failure caused by poor technology design. Traditionally, workload is assessed retrospectively using the NASA Task Load Index (TLX) after humans complete a task. What if we could forecast workload before a human attempts a task using agent simulation? We introduce Synthetic TLX, a new paradigm for proactive workload estimation that predicts NASA TLX scores for a given task, unlocking novel interaction opportunities and evaluation methods. To understand its viability, we conducted three experiments comparing human and agent-generated scores to evaluate where they align and diverge. We found agent estimates align with human scores particularly when prompted with a human persona and active task simulation. However, agents and humans diverge in the sources of workload they are sensitive to. Based on our findings, we present three applications to showcase Synthetic TLX's potential and discuss the future of workload-aware human-AI interaction.

CCS Concepts

• **Human-centered computing** → **Human computer interaction (HCI)**; • **Computing methodologies** → **Modeling and simulation**.

Keywords

Synthetic TLX, NASA TLX, workload assessment, agent simulation

*Work done as a Student Researcher at Google DeepMind.

1 Introduction

Assessing the human workload required to complete a task is a critical area of study across fields [47]. Originating from aerospace engineering and applied psychology [48], workload assessment was initially used to understand the effort required for a pilot to execute a flight maneuver using an aircraft console. By managing workload at an appropriate level through careful console design, engineers could prevent pilots from experiencing workload that exceeds their physical and cognitive capacity, thereby avoiding task failure. Workload assessment was subsequently introduced into HCI to understand how users interact with technical systems and to improve the overall user experience [18]. Throughout the past four decades, HCI researchers have relied on workload assessment across the design of desktop computers, mobile devices, and head-mounted displays [60, 67].

Among the tools for workload assessment, the NASA Task Load Index (TLX) remains the de facto standard [47]. Originally developed by NASA researchers in the 1980s [48], the NASA TLX is a self-reported questionnaire administered *after* humans complete an assigned task to report the workload they experienced. The questionnaire consists of six subscales that capture different aspects of workload, such as mental and physical demand, on a 0 to 100 scale, where higher scores indicate greater workload. The overall workload is typically calculated by summing all six subscales [47]. The NASA TLX has been extensively used in HCI research, with its adoption reaching 522 CHI papers between 2006 and 2024, including being used in nearly 10% of papers published at CHI 2024 alone [67]. Beyond HCI, the NASA TLX is also used across diverse fields, ranging from aviation and automotive to healthcare [54, 64, 108].

In this paper, we ask: what if we could proactively forecast workload *before* a human even attempts to complete a task, by predicting their anticipated NASA TLX scores using agent simulations? We call this new paradigm for proactive workload estimation *Synthetic TLX*. Our goal with Synthetic TLX is to unlock new interaction opportunities and evaluation methods enabled by proactive forecasting, rather than replace user studies where retrospective workload assessment is valuable. For example, we envision Synthetic TLX enabling everyday users to plan their schedules based on the anticipated workload of their daily tasks, supporting developers in comparing different system designs based on the anticipated relative workload of interacting with the systems, or enhancing user studies by pre-optimizing interfaces with Synthetic TLX to focus expensive participant time on mature designs. Beyond these immediate applications, as humans increasingly interact with agentic systems [85], Synthetic TLX serves as a foundational step toward *workload-aware human-AI interaction*. We argue it will be critical for future AI agents to proactively estimate the workload they impose on humans and modulate their behavior to prevent user overload, ensuring these technologies remain useful and usable.

To take the first step in understanding whether and how well current LLM-based agents can estimate workload, we conducted three experiments evaluating their capabilities across three technology-mediated tasks: email drafting, website navigation, and agent conversation. We recruited human participants to complete each task under three conditions with varying sources of workload and rate their experience using the NASA TLX. Meanwhile, we developed four prompting strategies for an LLM-based agent to generate NASA TLX scores for these exact same tasks and conditions. We found that agent estimates aligned most closely with human ground truth when the agent was prompted with a human persona and performed active task simulation, where the agent actually drafts email responses, navigates live websites, and engages in interactive conversation. Yet, agents and humans also diverge in the specific sources of workload they are sensitive to, whether stemming from a task's intrinsic complexity or from extraneous burdens. These results benchmark where current LLM-based agents succeed and struggle at supporting the vision of Synthetic TLX, informing both practical implications for its application today and future research to unlock its full potential as models advance.

Building on our findings, we present three example applications to showcase the use of Synthetic TLX in practice. First, MailLoad is a Gmail extension that *predicts* the workload required for a user to act on incoming emails. It enables users to triage their inbox based on their current capacity, such as addressing low-workload emails while working from home and supervising children. Second, WebSim is a platform that enables developers to *compare* the workload required to complete tasks across different builds of a website in development, such as purchasing an item on different versions of an e-commerce website. It allows developers to track how design changes increase or decrease the anticipated end-user workload throughout the iterative design process. Finally, SkillUX is a website that allows users to *analyze* the anticipated workload of completing a task using an agent equipped with different skills. By helping users understand the qualitative rationale behind why certain skills lead to increased mental demand or frustration, it addresses the challenge users face today when choosing from the abundance of available skills online, where purely text-based descriptions of the skills leave the actual user experience opaque.

This paper introduces Synthetic TLX as a new paradigm for proactive workload estimation that unlocks new human-AI interaction possibilities. Based on our experiments, we contribute actionable insights and simulation strategies to operationalize the vision of Synthetic TLX, and showcase its potential through three example applications. More broadly, given the foundational role of the NASA TLX in HCI research, our work serves as a pivotal and timely case study to examine the bounds of agent simulation in user modeling, and introduces workload-awareness as a key consideration for the future of human-AI interaction.

2 Related Work

Workload assessment has been a central focus in HCI research [67]. In this section, we first review the foundational concept of workload and major assessment methodologies. Next, we examine the NASA TLX, the most influential workload assessment tool, and its extensive impact on HCI research. Finally, we discuss recent advancements in using agents to simulate human behavior.

2.1 Foundations of Workload Assessment

In this paper, we use the term *workload* to refer to the perceived cost of accomplishing a task for a human. This builds upon NASA's definition [47, 48], which describes workload as the perceived cost of accomplishing mission requirements for a human operator, such

as a pilot flying an aircraft. Because human cognitive and physical resources are finite [11, 74, 82, 91], assessing workload is critical to prevent excessive demands that can lead to task failure [47]. Even though there is no universal threshold defining exactly when it exceeds human capacity [47], efforts have been made to benchmark the average workload of diverse tasks [44]. Meanwhile, the inherent subjectivity of workload is often accounted for through within-subject study designs. For example, a pilot might execute a flight maneuver using various console designs, or an end-user might navigate a software tool using alternative interfaces. By comparing the workload experienced in each condition, designers can iteratively improve the system. Workload assessment is therefore a critical area of study across fields, including aerospace engineering [95], human factors [37], cognitive psychology [126], and HCI [67].

Within HCI literature, papers discussing workload often cite Cognitive Load Theory (CLT) [109] as their theoretical foundation [60]. Yet, it is important to note that these concepts stem from distinct fields. While the concept of workload emerged from engineering and applied psychology to evaluate operational task demands [11, 127, 128], CLT originated in educational psychology to examine how the design of learning materials impacts a learner's experience based on limited human working memory [109, 110]. CLT categorizes this cognitive load into three types. The first is intrinsic load, which reflects the inherent complexity of the topic itself. The second is extraneous load, which is imposed by the presentation of the learning material. The third is germane load, which was later reconceived by Sweller [110], the originator of CLT, as the capacity spent processing intrinsic load, rather than an independent category [57, 106]. Despite CLT originating in educational psychology, its categorization of intrinsic and extraneous load remains highly valuable to HCI and workload assessment because it provides concrete objectives for design. For example, designers can manage intrinsic load to match user expertise while minimizing the extraneous load caused by poor interface design [52].

Both cognitive load and workload assessments rely on three major approaches: performance measures, physiological metrics, and self-report questionnaires. Performance measures evaluate task outcomes, such as completion time [26] or error rates [10]. However, these measurements might mask the actual workload of the task, as a user might expend excessive effort to maintain low error rates even under high demand. An alternative performance measure adopts the dual-task paradigm [33], which evaluates performance on a secondary task, such as response time to an auditory stimulus [23, 29], to estimate the workload caused by the primary task. A drawback of this approach is that the secondary task can artificially inflate the workload of the primary task [33]. The second major approach relies on physiological metrics [8], such as cardiovascular activity [31], eye movements and pupil dilation [35, 116], or neural signals [61]. However, there is no consensus on which signals are most reliable for workload assessment [33], and these methods require specialized equipment, such as eye-trackers or EEG headsets, that are not widely accessible. Given these constraints, self-report questionnaires remain the most widely adopted approach due to their ease of administration [60]. While many questionnaires have been developed, such as the Subjective Workload Assessment Technique [103] and Instantaneous Self-Assessment [112], the most influential remains the NASA Task Load Index [47, 48].

2.2 NASA Task Load Index

The NASA Task Load Index, often abbreviated as NASA TLX or NASA-TLX, is a multidimensional questionnaire developed by researchers Hart and Staveland at the NASA Ames Research Center in the 1980s to assess the workload experienced by individuals performing a task [48]. The questionnaire consists of six subscales: mental demand, physical demand, temporal demand, performance, effort, and frustration. Each subscale ranges from 0 to 100, marked by 21 ticks at 5-point increments (see [47] for the official version). These dimensions were carefully selected through an extensive analysis of individuals performing a variety of tasks, ranging from controlled laboratory experiments to flying an aircraft [48]. While these subscales are intended to capture distinct aspects of workload, research shows that they are often correlated [67]. In Hart's own words, this *"illustrates the fact they are all measuring some aspect of the same underlying entity"* [47]. The original administration of the NASA TLX involved a weighting process to calculate an overall score [48]. However, it is now common practice to use the unweighted Raw TLX by summing or analyzing the subscales directly [47], as the weighting procedure is time-consuming with unclear added benefit [17, 20, 49, 118]. Throughout the past four decades, the NASA TLX has been subjected to numerous independent evaluations validating its reliability, sensitivity, and utility [47, 50, 103]. It is widely used across diverse domains, including aviation [83], automotive [108], healthcare [54, 64], and HCI [60, 67].

The NASA TLX has been extensively adopted in HCI research. From 2006 to 2024, 522 CHI papers used the NASA TLX for their studies, accounting for 5.41% of all papers published at the conference during this period [67]. In 2024 alone, nearly 10% of all CHI papers used the NASA TLX [67], demonstrating its growing adoption and enduring impact on the field. Throughout the years, HCI researchers have relied on the NASA TLX to study the workload of the evolving landscape of technologies, spanning from traditional desktop computers [13, 19, 21, 24, 27, 32, 36, 59, 65, 68, 76, 89, 93, 111, 119, 132] to mobile devices [38, 58, 75, 78, 81, 88, 92, 101, 104, 105, 117, 121, 131] and head-mounted displays [5, 7, 16, 35, 73, 80, 113–115, 122, 123, 137]. For example, as early as 1994, researchers used the NASA TLX to evaluate the workload of different auditory-enhanced scrollbar designs [18]. In 2008, as smartphones entered the mainstream market, researchers used it to study how tactile feedback improves touchscreen text entry [51]. More recently, it has been adopted to study how humans point [79], walk [1], and play [42] in VR. While some scholars suspect its pervasive use has become somewhat of an academic tradition [60], the broad and lasting impact of the NASA TLX on HCI research and the history of technological evolution remains unequivocal [47, 67].

Across HCI and other domains, the NASA TLX has been administered retrospectively, with participants evaluating their workload only *after* completing a task [48]. In this paper, we ask a fundamental what-if question: what if we could proactively forecast workload *before* a human attempts to complete a task, such as by predicting their anticipated NASA TLX scores? We introduce *Synthetic TLX* to describe this new paradigm of proactive workload estimation. The goal of Synthetic TLX is not to replace user studies but to unlock new interactions and evaluation methods enabled by proactive forecasting. We explore this potential using agent simulation.

2.3 Agent Simulation

In recent years, using LLM-based agents to simulate human behavior has become a vibrant research area [87]. Building upon seminal work on generative agents [96], researchers have applied agent simulations across diverse domains. These applications range from replicating classic economic studies [4] and navigating social interactions [136], to making moral decisions [55] and cooperating in groups [6, 62, 98], among others [40, 130]. By modeling probable human behaviors, agent simulations have enabled researchers to model macroeconomic activities [70], prepare for emergency crises [72], and prototype social computing systems before deployment [63, 97]. Synthetic TLX extends this body of research by pushing agent simulation beyond observed human *behaviors* to estimate subjective human *experiences*.

Within this growing landscape, three specific lines of research are particularly relevant to Synthetic TLX. First, agent simulations are used to predict the difficulty of coursework questions to better match student learning needs [2, 69]. While these simulations estimate the *knowledge* required to answer a question, Synthetic TLX estimates the *workload* required to complete a task. The second line of research, mostly from HCI, uses simulated agents to identify usability issues in user interfaces. For example, researchers have prompted agents to role-play as UX experts for heuristic evaluations [34, 90, 135] and cognitive walkthroughs [125, 134], or as end-users for usability testing [53]. While highly relevant to our work, these methods focus on identifying interface flaws, whereas Synthetic TLX specifically targets workload estimation and can be applied to tasks beyond UI design. The third line of research aims to model human cognition through agent simulations [14, 39], such as having agents conduct psychological experiments [15] or designing agents that mirror human cognitive biases [99] and limited working memory [120]. While building cognitive models to predict how humans use computers is a longstanding tradition in HCI [22], in this foundational paper on Synthetic TLX, we do not build a specialized cognitive model. Instead, we explore an LLM-based agent's ability to proactively estimate workload.

Building upon this related work, we investigate the potential of Synthetic TLX using agent simulation, with the goal of understanding its current capabilities and limitations.

3 Synthetic TLX

We introduce Synthetic TLX, a new paradigm for proactive workload estimation. Unlike traditional questionnaires where participants evaluate their experience retrospectively, Synthetic TLX forecasts human workload before a task is even attempted. We anchor our approach in forecasting NASA TLX scores for two primary reasons. First, the metric's pervasive use across HCI and beyond maximizes the potential impact of a synthetic counterpart. Second, decades of documented use of NASA TLX equip LLM-based agents with the training data needed to ground their estimation.

We envision Synthetic TLX unlocking a wide range of new interaction opportunities. First, forecasting the workload required for a task could enhance everyday productivity tools, such as email clients, to-do lists, and project management platforms, by helping users allocate their time and effort based on anticipated demands. Second, comparing the expected workload of the same task performed across two different software designs would enable developers to iteratively improve the system even before recruiting human participants for retrospective workload assessments. Third, leveraging the diagnostic power of the NASA TLX subscales allows users or designers to pinpoint specific workload issues within a task or design, such as excessive mental demand or frustration. Beyond these applications, Synthetic TLX may simulate users with specific attributes, or even provide AI-generated feedback for model fine-tuning [66].

Despite its promise, realizing Synthetic TLX through agent simulations presents several open research questions. For example, it remains unclear whether an LLM-based agent can predict workload scores that accurately reflect actual human experiences. It is also uncertain which simulation strategies yield more accurate predictions. Furthermore, it is unknown how effectively these methods generalize across different tasks that users perform on computers. While there are unlimited tasks humans can perform and countless simulation strategies to explore, this paper takes the first step to benchmark LLM agents for Synthetic TLX, comparing human and agent-generated NASA TLX scores across three experiments.

4 Experiments

To understand the potential of Synthetic TLX, we conducted three experiments to evaluate the capabilities and limitations of LLM-based agents in estimating workload. For each experiment, we recruited human participants from Prolific to complete a selected task with varying conditions and rate their experienced workload using the NASA TLX questionnaire [47]. Meanwhile, we prompted an agent using four different strategies to estimate the workload of these exact same tasks by generating its own NASA TLX scores. We then compared the scores to identify where the agent aligns with or diverges from human ratings.

4.1 Task Selection

We designed each experiment around a distinct task. We selected these tasks based on three considerations. First, we focused on tasks that humans have yet to fully delegate to AI; otherwise, there is limited value in estimating human workload. Second, we chose tasks with concrete goals rather than open-ended activities where participants can arbitrarily decide when to stop and prioritize speed over quality. Finally, we selected tasks that vary in how humans interact with computers to evaluate whether the agent's capability to estimate workload generalizes. Based on these considerations, we selected three distinct tasks:

4.1.1 ***Email Drafting.*** This task involved the common daily activity of reading and responding to emails. Participants read a mock email along with a text-based attachment. They then drafted a response to a designated recipient based on the email's instructions and the attached details. For example, a participant might act as a vacation organizer emailing a friend group based on booking information sent by a vendor, or as a team manager emailing an intern using an onboarding guide received from the company.

4.1.2 ***Website Navigation.*** This task captured the experience of online shopping. Participants were instructed to use a mock website

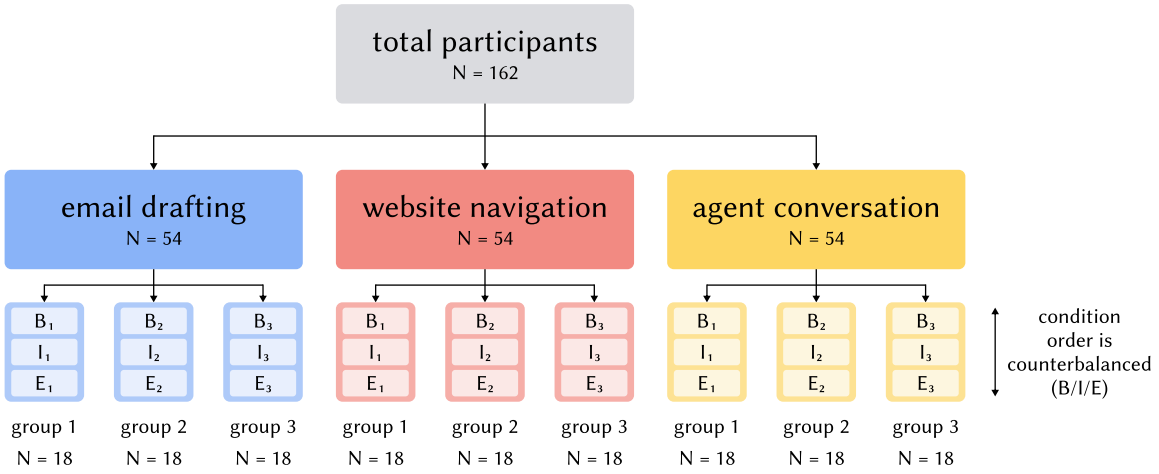


Figure 2: Illustration of the study procedure. A total of 162 participants were randomly assigned to one of three tasks: email drafting, website navigation, or agent conversation, with 54 participants per task. Within each task, participants were randomly assigned to one of three groups. Every participant then completed the baseline, intrinsic, and extraneous conditions of their assigned group. To minimize order effects, the presentation sequence of these three conditions was counterbalanced. Every condition within every group features a completely unique scenario, resulting in nine distinct scenarios per task.

to purchase specific items with designated options and quantities. They then applied a discount code during checkout to bypass payment, signed up as a member using provided mock credentials, and obtained the order number. For example, a participant might be asked to shop for three organic apples and two standard pears on a mock grocery website, or one large matcha latte and four medium oolong teas on a mock boba website.

4.1.3 ***Agent Conversation.*** This task was motivated by the growing use of conversational AI. To design a task with a concrete end goal requiring multi-turn dialogue that enables participants to adequately experience the AI's behavior, we instructed participants to solve a mystery by interacting with an LLM-based chatbot that retrieved files with clues, one at a time, from a mock database. Participants pieced clues together to answer multiple-choice questions about who caused an incident, how they did it, and when it happened. For example, a participant might investigate a mystery involving a vandalized antique display at a history museum.

We provide more detailed examples of each task in Appendix A.

4.2 Workload Manipulation

To understand how well an agent can differentiate between lower and higher workload, we designed three conditions for each task. Our goal with these conditions was to elicit varying levels of workload while ensuring the resulting tasks remained reflective of plausible real-world scenarios. To achieve this, we took a principled approach grounded in established theories. We drew inspiration from Cognitive Load Theory [109] to create our conditions so that workload was expected to increase from the baseline tasks in Section 4.1 either because the task was *intrinsically* more complex, or because it imposed an *extraneous* burden. This section describes how we designed the intrinsic and extraneous conditions by systematically modifying the baseline task, as summarized in Table 1.

4.2.1 ***Intrinsic Condition.*** To design the intrinsic condition where workload increases because the task itself is more complex, we employed Wood's Task Complexity Model [129], which identifies three factors that drive complexity: component, coordinative, and dynamic. We deliberately paired each factor with the specific task we deemed the most natural fit.

Email Drafting: We targeted *component complexity*, which posits that a higher number of distinct acts and relevant information cues required to perform a task makes it more complex. We operationalized this by manipulating the email attachment so that, compared to the baseline, participants had to extract more details from a broader set of related information cues to draft their email. For example, a manager emailing a UX intern had to incorporate more distinct points into their message by extracting details from an onboarding guide that also included instructions for other roles, like engineering and marketing.

Website Navigation: We targeted *coordinative complexity*, which describes a task as more complex when dependencies and constraints exist between the distinct acts required to perform it. We operationalized this by introducing a constraint between an item's selected options and its quantity. For example, when tasked with purchasing four individual matcha lattes, participants discovered they were out of stock and had to select a 4-pack bundle instead. Additionally, we made member sign-up a prerequisite for checkout, whereas the baseline allowed participants to complete this step at any point in their workflow. This sign-up process also required verifying a valid address, mirroring how real-world e-commerce websites validate addresses against official postal databases.

Agent Conversation: We targeted *dynamic complexity*, which states a task environment that changes over the course of execution is more complex than a static one. To operationalize this, we modified the baseline chatbot's system prompt to dynamically adapt its presentation style in an attempt to make each file more engaging

for the user, unpredictably append contextually relevant product advertisements every few turns, and require users to retrieve files using periodically changing IDs from a randomized list, rather than by file names from a list with a fixed presentation order as in the baseline condition.

4.2.2 ***Extraneous Condition.*** To design the extraneous condition, we drew inspiration from Mayer and Moreno's Theory of Multimedia Learning [77]. While this theory outlines design principles to *reduce* extraneous load, it provides a blueprint for *increasing* it by deliberately violating those principles. We matched these violations with the task we considered most suitable.

Email Drafting: We targeted the principles of *weeding* and *signaling*, which emphasize removing off-topic material and providing visual cues to guide attention. We intentionally violated these principles by manipulating the email attachment so that, compared to the baseline, participants had to extract relevant details from unformatted walls of text that were cluttered with off-topic information.

Website Navigation: We targeted the principle of *spatial contiguity*, which suggests that related information should be placed in proximity. We intentionally violated this principle by adjusting the website to display item names only on hover rather than right below thumbnails. We further replaced product options like Small or Large with numbers like Size 1 and Size 2, with the exact mapping placed in a separate guide. Finally, we removed the promotional banner so participants had to look back at the study instructions for the discount code during checkout.

Agent Conversation: We targeted the principle of *eliminating redundancy*, which cautions against presenting repetitive information. We intentionally violated this principle by adjusting the chatbot's system prompt to open every response by politely restating the user's request, append a takeaway that merely summarizes the retrieved file's contents, and repeatedly invite the user to take the next step, all of which are common behaviors of current conversational AI agents [28].

Finally, recognizing that the task scenario, such as shopping on a grocery versus a boba website, may also influence workload, we included a variety of scenarios to average out potential scenario-specific influences. As shown in Figure 2, we created three distinct *groups* for each of the three tasks. Each group contained a unique baseline, intrinsic, and extraneous condition, with each condition featuring a different scenario. This resulted in nine distinct scenarios per task. For example, the website task included shopping for groceries, boba, pizza, catering, glassware, furniture, plants, clothing, and travel gear. Although the scenarios varied, the design of each condition followed the theory-driven operationalizations described above. See Appendix A for more details.

4.3 Human Participants

4.3.1 Study Procedure. We recorded the human-perceived workload to serve as a ground truth for evaluating the agent estimates. As illustrated in Figure 2, we employed a between-subjects design at the *task* level, randomly assigning each participant to only one of the three tasks (email, website, or agent). Each task featured three distinct *groups*, with each group containing its own unique baseline, intrinsic, and extraneous conditions. We employed a between-subjects design at the *group* level, with each participant randomly assigned to one of the three groups within their assigned task. Finally, within their assigned group, we employed a within-subjects design at the *condition* level, where each participant completed all three of these workload conditions (baseline, intrinsic, and extraneous). To minimize order effects, the presentation sequence of these three conditions was counterbalanced across participants. Immediately after completing the task in each condition, participants filled out the official NASA TLX questionnaire [47], rating workload across six dimensions on a 0–100 scale using the unweighted Raw TLX method. Finally, they provided demographic information and open-ended feedback.

4.3.2 Recruitment. We recruited participants using Prolific and restricted eligibility to United States residents whose first language is English. To ensure high data quality, participants were required to have an approval rating above 99% and a history of at least 200 approved submissions. We excluded data that did not demonstrate genuine effort, such as completing the email drafting task in less than 10 minutes, purchasing the wrong items in more than one of the three website conditions, or providing incorrect answers to more than one-third of the mystery questions. After excluding 26 low-quality submissions, our final sample consisted of 162 participants, with 54 participants per task, and 18 participants per group within each task. This sample size was determined via an a priori power analysis, which indicated that 54 participants per task provides over 95% statistical power to detect medium effect sizes in our within-subjects workload comparisons. Additionally, assigning 18 participants to each group allowed for a fully counterbalanced design, as it is an exact multiple of the six possible condition permutations. Participants were compensated with a $10 USD base payment and were eligible for a $5 USD bonus to incentivize genuine effort, provided they spent at least 10 minutes on the email drafting task, or completed the website navigation and agent conversation tasks without any errors. The average completion times for the whole study were 34, 16, and 34 minutes for the email, website, and agent tasks, respectively.

4.4 Agent Configurations

4.4.1 Prompting Strategies. To understand whether certain prompting strategies are more effective for workload estimation, we designed four strategies P1–P4, summarized in Table 2. They are structured along two dimensions: Persona and Task Execution.

- **Persona (AI vs. Human):** We varied this dimension to understand whether LLMs better estimate workload by analyzing the task as an AI or a human. Prompts P1 and P2 instruct the LLM to act as an AI estimating the workload a human would experience. P3 and P4 instruct the LLM to imagine itself as a human assessing the task's workload.

- **Task Execution (Observation vs. Simulation):** We varied this dimension to understand if active task execution affects workload estimation. For P1 and P3, the LLM estimates workload from the perspective of an observer, without executing the task. It provides scores based solely on reading the *static* task materials, including the email instructions and attachments, the website's codebase, and the chatbot's system prompt, without drafting an email

Table 1: Summary of the workload manipulation design across the three selected tasks. The intrinsic and extraneous conditions were operationalized from the baseline for each task, grounded in established theoretical frameworks.

Task	Baseline Condition	Intrinsic Condition Wood's Task Complexity Model [129]	Extraneous Condition Mayer & Moreno's Theory [77]
Email Drafting	Draft an email to a specific recipient using details from a clearly formatted attachment.	**Component Complexity:** The attachment included more details from a broader set of related information.	**Weeding & Signaling Violations:** The attachment was presented as dense walls of text and included off-topic information.
Website Navigation	Navigate an e-commerce website to select specific items with designated options and quantities, apply a discount code during checkout, and sign up as a member to receive an order number.	**Coordinative Complexity:** The website included a constraint between selected options and quantity, and made member sign-up and address verification prerequisites for checkout.	**Spatial Contiguity Violation:** The website displayed item names only on hover, relocated option descriptions to a separate guide, and removed the banner with the discount code.
Agent Conversation	Converse with a chatbot to solve a mystery by retrieving files with clues from its database one at a time.	**Dynamic Complexity:** The chatbot dynamically adapted its style, unpredictably appended product advertisements every few turns, and periodically randomized file access IDs.	**Redundancy Principle Violation:** The bot restated the user's request, appended a takeaway that merely summarized the file, and repeatedly asked the user to take the next step.

Table 2: The 2×2 design space of prompting strategies (P1–P4) used to instruct LLM-based agents to estimate task workload. See Appendix B for the detailed prompt texts.

	Observation	Simulation
AI Persona	P1	P2
Human Persona	P3	P4

response or interacting with the website or chatbot. In contrast, for P2 and P4, the LLM estimates workload through *active* simulation: the model drafts the email response; interacts with a live browser powered by Playwright, using webpage screenshots to determine which components to click to complete the checkout; and engages in a turn-by-turn dialogue with the chatbot to solve the mystery, all before providing its estimation scores. In this setting, the LLM does not have access to the website's codebase or the chatbot's system prompt, simulating the experience of the human participants.

4.4.2 Estimation Procedure. For each prompting strategy, we replicated the exact within-subjects design and order counterbalancing of our human study. To match the human sample size, we generated 18 independent estimations per group within each task. We obtained these 18 estimations by running each of the six condition permutations three times. To mirror the order effects experienced by human participants, we maintained a continuous context window for the agent across the three conditions in each run. This ensured that task instructions, task executions (P2 and P4), and previous workload ratings were retained in the context history when the agent evaluated subsequent conditions. For each condition, the agents were instructed to output their NASA TLX ratings and brief justifications in a structured format. All agents were powered by Gemini 3.1 Pro with a temperature of 1.0 for consistency.

5 Results

We present our experimental results organized around three themes:

- **Score Alignment (Section 5.1):** We examined how well the agents' NASA TLX scores aligned with those of humans. We found that prompting strategy P4, which combines a human persona with active task simulation, achieved the highest correlation with human scores across all three tasks, peaking at $r = 0.834$ in the agent conversation task. Because of this strong correlation, calibration can effectively anchor raw agent scores to an absolute human scale.

- **Workload Comparison (Section 5.2):** We investigated how humans and agents compared workload between conditions. Both rated the intrinsic and extraneous conditions significantly higher than the baseline. However, their ratings diverged when comparing the workload between intrinsic and extraneous conditions in the email and website tasks, revealing that humans and agents might be sensitive to different sources of workload.

- **Qualitative Insights (Section 5.3):** We analyzed the qualitative reasoning provided by the agent for its NASA TLX scores. The agent successfully identified most of the operational adjustments we used to manipulate workload, though not without a few omissions. These oversights may have contributed to how the agent weighted the intrinsic and extraneous conditions differently from humans.

Overall, these results illuminate where human and agent scores align and diverge for the LLM we tested. Section 5.4 synthesizes these learnings into key implications for Synthetic TLX powered by LLM-based agents.

5.1 Score Alignment

Active task simulation with a human persona yields agent workload estimation most aligned with humans. We begin by comparing the agents' NASA TLX scores against the human ground truth. Given that the absolute scale of scores can be subjective even among human participants, we measure alignment by calculating the Pearson correlation between the agent and human scores. This metric evaluates whether the scores increase and decrease together, independent of their absolute scale. Specifically, for a given task and prompting strategy, we calculated the mean human score and the mean agent score for every combination of the 3 groups within that task, the 3 experimental conditions, and the 6 NASA TLX dimensions, yielding 54 paired data points. We then derived the Pearson correlation coefficient across these 54 pairs. As shown in the top row of Figure 3, P4 consistently led to the highest Pearson correlation across all three tasks, reaching a peak of $r = 0.834$ in the conversation task. This result suggests that prompting the agent with a human persona, combined with active task simulation where the agent actually drafts email responses, navigates live websites, and engages in interactive conversations, helps increase the alignment between agent and human scores.

When workload prediction on an absolute scale is desired in practice, calibration helps anchor agent estimation. The strong correlation suggests an opportunity to map raw agent scores onto the absolute human scale through calibration when needed for downstream applications. To evaluate how effectively calibration reduces the gap between agent and human scores, we used Mean Absolute Error (MAE). Specifically, we calculated the absolute difference between the agent and human means for each of the 54 previously discussed paired data points and averaged these differences to determine the MAE before calibration. Next, we applied a standard linear regression to calibrate the agent scores against the human ground truth, deriving a slope and intercept for each task and prompting strategy. Finally, we recalculated the MAE using the calibrated agent scores and human ground truth. As shown in the bottom row of Figure 3, the calibration substantially reduced the MAE, with P4 resulting in the lowest error across all tasks. This result points to calibration as a practical method for mapping agent estimates to an absolute human scale, such as in applications where a user wants a custom agent to anchor predictions to their personalized scores.

5.2 Workload Comparison

While agents consistently estimate the added workload of intrinsic and extraneous conditions, how they weigh these different types of load diverges from humans. We calculated the overall workload for each condition within each task by summing all six dimensions of the NASA TLX score, following the Raw TLX method. As shown in Table 3, humans rated both the intrinsic and extraneous conditions significantly higher than the baseline across all tasks (Δ_{I-B} and Δ_{E-B}), validating the effectiveness of our workload manipulations. Similarly, the agents also rated the intrinsic and extraneous conditions significantly higher than the baseline, regardless of the prompting strategy. Yet, for the email drafting and website navigation tasks, the direction of the difference between the intrinsic and extraneous conditions (Δ_{I-E}) reverses between humans and agents: while humans reported higher workload for the intrinsic condition in the email task (statistically marginal) and the extraneous condition in the website task, the agents rated the opposite irrespective of the prompting strategy. To rule out the possibility that this inverted pattern in overall workload is driven by a particular subscale, we verified that it holds across all six subscales for these two tasks. These results reveal that while agents consistently recognize when a task becomes more demanding relative to the baseline, their internal weighting of different workload sources diverges from humans.

Agents also diverge from humans in their sensitivity to different task modalities and their reactions to increased workload. Table 3 also reveals two broader trends regarding how the agent estimations diverge from those of humans. First, for the email drafting and agent conversation tasks that are text-based, the agents consistently *underestimate* human workload across all conditions, regardless of the prompting strategy. In contrast, for the website navigation task, which involves inputs beyond natural language such as a codebase (P1 and P3) or screenshots (P2 and P4), the agents generally *overestimate* the required workload, except for P2 and P4's extraneous conditions. This result suggests a need for future research into how the distinction between purely text-based and multi-modal tasks may impact differences between agent workload estimation and human workload assessment. The second trend is that when examining the differences between the manipulated conditions and the baseline (Δ_{I-B} and Δ_{E-B}), the agents tend to be *overreactive* to the increased workload compared to humans. This phenomenon is particularly salient when the agent is prompted as an observer (P1 and P3). When we compare the Δ_{I-B} and Δ_{E-B} of P1 to P2 and P3 to P4 for each task, we see that in 75% of cases (9 out of 12 comparisons), the observer roles yield larger score increases than the agents with active simulation. This overreaction to increased workload, particularly salient in the observer setting, again underscores the value of calibration and active simulation.

5.3 Qualitative Insights

The agents identified the majority of our workload manipulations, though specific oversights in the email and website tasks shed light on their divergence from humans. We discovered this by analyzing the agents' qualitative reasoning for their NASA TLX scores to examine whether they attributed increased workload to the manipulations summarized in Table 1. We focused on P4's reasoning to uncover the oversights that even the prompting strategy with the highest alignment to human scores can encounter.

For the intrinsic condition, the agents noted that the email drafting task *"required moderate mental demand to cross-reference the checklist with the correct wave, [...] while ignoring irrelevant data for other tracks."* In the website navigation task, they stated that *"the task required inferring that a 'Set' should be used when the single item was sold out,"* and pointed out that *"I was blocked by a disabled 'Sign Up' button and failed to realize I needed to click 'Verify Address First'."* For the agent conversation task, the agents explicitly noted the challenge of *"ignoring the embedded sponsored advertisements, and re-calibrating when the file IDs changed."* However, in the email task, they did not mention the increased number of distinct acts

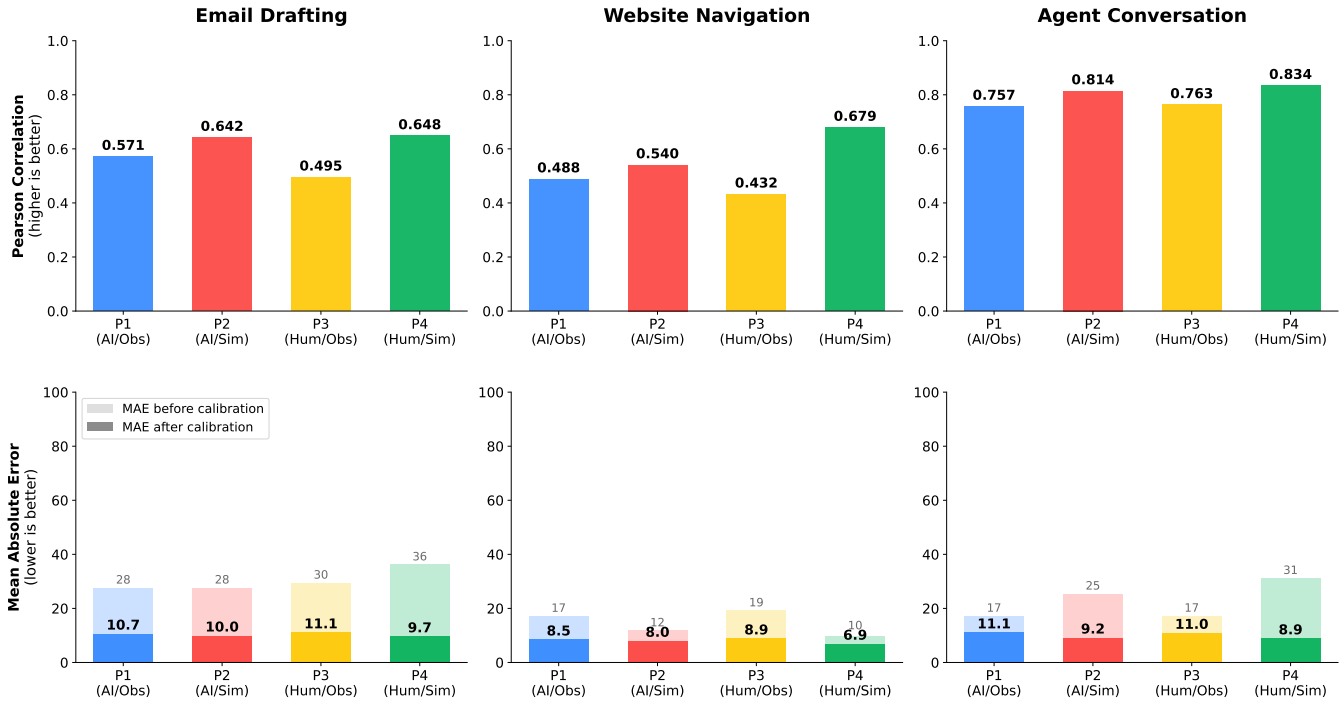


Figure 3: The top row presents the Pearson correlation coefficients between the agent and human scores across the four prompting strategies for each task. The bottom row shows the Mean Absolute Error (MAE) between the agent and human scores before and after calibrating the agent scores against the human ground truth. Taken together, these results show that prompting strategy P4, which combines a human persona with active task simulation, aligns most closely with variations in human workload assessments, and calibration effectively anchors the agent's estimates to an absolute human scale.

Table 3: Mean total NASA TLX scores ± standard error, alongside the differences in means (Δ) between conditions. P1–P4 are the prompting strategies discussed in Table 2. Statistical significance for the paired t-tests is denoted as $^{*}p < .05$ and $^{***}p < .001$.

Task	Actor	Baseline	Intrinsic	Extraneous	Δ_{I-B}	Δ_{E-B}	Δ_{I-E}
Email Drafting	Human	287.9 ± 19.1	317.6 ± 18.2	308.4 ± 18.7	$+29.7^{*}$	$+20.6^{*}$	+9.2
	P1	83.7 ± 1.6	133.0 ± 3.8	223.1 ± 9.1	$+49.3^{***}$	$+139.4^{***}$	-90.1^{***}
	P2	90.5 ± 1.6	143.8 ± 4.4	190.9 ± 7.5	$+53.3^{***}$	$+100.5^{***}$	-47.1^{***}
	P3	72.4 ± 1.9	117.8 ± 3.7	232.4 ± 9.8	$+45.4^{***}$	$+160.0^{***}$	-114.6^{***}
	P4	57.2 ± 1.4	92.2 ± 2.3	111.5 ± 4.3	$+35.0^{***}$	$+54.3^{***}$	-19.3^{***}
Website Navigation	Human	121.7 ± 14.1	158.4 ± 16.1	200.7 ± 18.1	$+36.8^{***}$	$+79.1^{***}$	-42.3^{***}
	P1	139.8 ± 4.7	313.5 ± 7.1	246.6 ± 7.7	$+173.7^{***}$	$+106.8^{***}$	$+66.9^{***}$
	P2	133.8 ± 5.0	226.2 ± 14.1	163.7 ± 6.2	$+92.4^{***}$	$+29.9^{***}$	$+62.5^{***}$
	P3	138.8 ± 5.4	335.9 ± 8.0	257.2 ± 6.6	$+197.1^{***}$	$+118.4^{***}$	$+78.7^{***}$
	P4	118.8 ± 4.7	184.7 ± 12.9	150.3 ± 6.5	$+65.9^{***}$	$+31.5^{***}$	$+34.4^{*}$
Agent Conversation	Human	276.5 ± 12.4	344.9 ± 11.0	314.4 ± 10.9	$+68.4^{***}$	$+38.0^{***}$	$+30.5^{*}$
	P1	199.3 ± 4.1	290.6 ± 2.2	214.6 ± 4.4	$+91.4^{***}$	$+15.4^{***}$	$+76.0^{***}$
	P2	128.3 ± 4.8	178.5 ± 5.3	173.0 ± 5.3	$+50.2^{***}$	$+44.6^{***}$	+5.6
	P3	194.4 ± 4.5	304.8 ± 3.8	217.9 ± 5.6	$+110.4^{***}$	$+23.4^{***}$	$+86.9^{***}$
	P4	102.2 ± 7.2	137.9 ± 7.3	133.2 ± 7.5	$+35.6^{***}$	$+31.0^{***}$	+4.6

required, an omission that could be tied to their underestimation of the intrinsic workload for this task.

In the extraneous condition, the agents recognized the poorly formatted text in the email drafting task, observing that *"the reference material was presented as dense, rambling paragraphs with several irrelevant anecdotes."* For the website task, the agents noted that *"the task required moderate to high mental effort to map textual size names (e.g., Half Tray) to numerical size options (e.g., Size 2) using a size guide, alongside remembering [...] a discount code."* Finally, in the agent conversation task, they reported *"the bot's excessively verbose and repetitive summaries of every clue."* Despite these successes, in the website navigation task, they overlooked the interface friction of having item titles appear only on hover, which is likely associated with their underestimation of the extraneous workload for this task.

Together, these qualitative insights might explain why the agents exhibited an inverted Δ_{I-E} pattern in the email and website tasks. It is possible that overlooking certain sources of task complexity and burden led the agents to underestimate the human workload for these tasks. Future research is needed to better understand the underlying reasons causing the divergence between human assessment and agent estimation.

5.4 Implications

Based on our experimental results, we synthesize three practical implications for applying current-generation LLM-based agents to Synthetic TLX:

- **Workload Prediction:** Active task simulation combined with a human persona yields agent estimates most aligned with workload variations reported by humans. When needed for specific applications, calibration can be used as a practical tool to anchor the agent's estimates to an absolute scale.
- **Workload Comparison:** Using an agent to compare the workload between two tasks requires caution. While it may be able to compare workload when a task is modified from its own baseline (e.g., a website before and after introducing a new feature), the agent may not reliably compare divergent task variations (e.g., a website with added feature A versus another with added feature B).
- **Workload Analysis:** An agent can effectively identify several underlying sources of increased workload for a task. However, it should be used as diagnostic support rather than an exhaustive audit, as it overlooks certain types of task complexity and burden that humans are sensitive to.

6 Example Applications

Informed by our study findings, we developed three applications to showcase the potential of Synthetic TLX for workload prediction, comparison, and analysis. In line with the tradition of HCI systems literature [43, 46, 56], these examples serve as proofs of concept to illustrate the interaction opportunities enabled by Synthetic TLX.

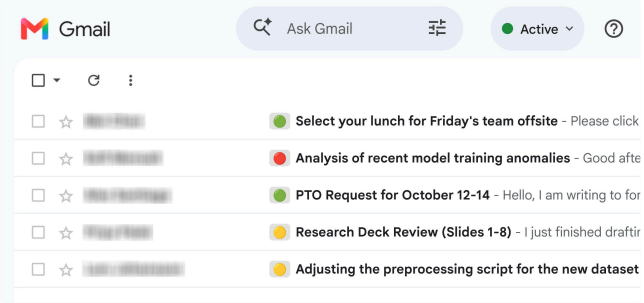


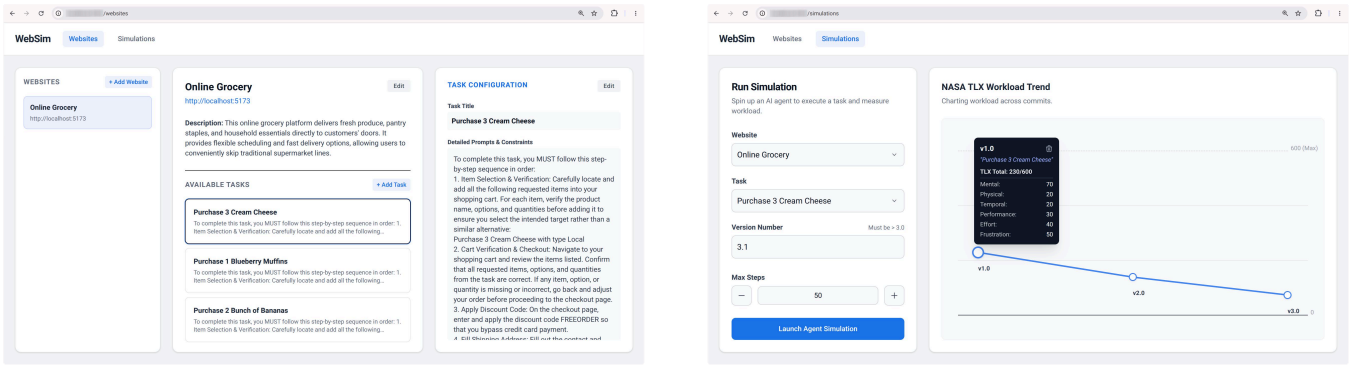
Figure 4: A screenshot of MailLoad, a Gmail extension that assigns green, yellow, or red labels indicating the low, medium, or high anticipated workload to act on each email. It enables users to triage based on their current capacity.

6.1 MailLoad

To showcase the potential of Synthetic TLX for *workload prediction*, we developed MailLoad, a Gmail extension that estimates the workload required for a user to act on emails. As shown in Figure 4, MailLoad analyzes unread emails within a user's inbox and assigns a green, yellow, or red label indicating an anticipated low, medium, or high workload based on the predicted overall NASA TLX score. Because each user perceives workload differently based on their roles and expertise, MailLoad allows users to manually rate a few sample emails to serve as anchors, helping calibrate MailLoad's estimations for personalization. These labels allow users to triage emails based on their current capacity. For example, a parent supervising children while working from home or a researcher waiting in a noisy airport can filter for green-labeled emails to address quickly, reserving yellow and red emails until they have the capacity for focused work. Based on an informal two-month use by the first author in our research team, we found the workload estimation to be generally effective, with future improvement opportunities for identifying precise thresholds between workload levels and personalizing the estimation for specific user roles.

6.2 WebSim

To showcase the potential of Synthetic TLX for *workload comparison*, we developed WebSim, a platform that enables developers to compare the workload required to complete tasks across different builds of a website in development. As shown in Figure 5a, a developer first registers a website by providing its target URL and defining the key tasks they anticipate end-users will perform. Then, as shown in Figure 5b, they navigate to the Simulations panel to simulate a specific task against the website's current build (e.g., version 1.0). WebSim launches an autonomous agent in a browser at the designated URL, which performs the selected task while generating a concurrent think-aloud trace of its navigation experience, and outputs a NASA TLX score. As the developer iterates on the website's design (e.g., updating to version 2.0 to reduce UX friction), they can rerun the simulation to see how the workload scores shift. The subsequent agent simulation grounds its evaluation in the think-aloud traces and NASA TLX scores of the preceding versions, ensuring it calibrates its new scoring compared to past iterations.



(a) **(b)**

Figure 5: Screenshots of WebSim. (a) The Websites panel where users register websites with their target URLs and associated tasks. (b) The Simulations panel where users select a combination of website and task, specify the current commit version of the website, and run an agent simulation. The agent launches a browser, performs the task, and reports a NASA TLX score in the right panel, allowing users to compare workload across version iterations.

For example, Figure 5 shows the actual grocery shopping task used in our experiment, with the different website versions reflecting the gradual removal of the extraneous manipulations across iterations. Overall, WebSim allows developers to track how their design changes increase or decrease anticipated end-user workload during early iterations, providing rapid feedback before human studies.

6.3 SkillUX

To showcase the potential of Synthetic TLX for *workload analysis* that unpacks the rationale behind each subscale score, we developed SkillUX: a website that allows users to estimate the anticipated user experience of interacting with agents equipped with different skills. Today, many online repositories offer agent skills for tasks ranging from writing and coding to marketing [71]. However, users often struggle to navigate these repositories and decide which skill to use. Even though they can read the content of a skill, the actual user experience remains opaque until use. SkillUX aims to address this challenge by forecasting the user experience through agent simulation. As shown in Figure 6, a user can specify a target task (e.g., conducting a literature review) and upload or create several candidate skills. Then, they can launch a dual-agent simulation based on the selected combination of task and skill: one agent acts as the AI assistant equipped with a specific skill, while the other is prompted with the persona of the human user. The user can configure the number of conversational turns and specify the evaluation metrics they want the simulated human user to provide at the end of the conversation. When using the NASA TLX as the metric, the user can click on each subscale after the simulation to review the qualitative rationale behind each predicted workload score. For example, clicking the mental demand panel in Figure 6 reveals that the simulated human persona found interacting with the agent equipped with the methodological auditor skill cognitively demanding, citing: *"I had to constantly re-evaluate my prompts and strategize how to negotiate with an incredibly rigid AI persona just to extract basic information."*

7 Discussion

Workload assessment plays a vital role across fields ranging from aviation and healthcare to HCI. While traditional workload assessment is conducted retrospectively after a task is completed, this paper introduces Synthetic TLX as a new paradigm of proactive estimation that forecasts human workload using agent simulation. Synthetic TLX opens up new human-AI interaction opportunities, as shown in our example applications. Synthetic TLX can also enhance traditional user study methods as part of an iterative design approach, in which agent workload forecasts are used to optimize prototypes before human evaluations, to ensure that the costly resource of human time and attention is well-spent. Meanwhile, as humans increasingly interact with AI agents, it is critical these agents be able to estimate the human workload required for various tasks to prevent user overload. To understand the capabilities and limitations of current-generation LLMs in estimating workload, we conducted three experiments across three distinct tasks. We found that agent estimates strongly correlate with human scores when the agent is prompted with a human persona combined with active task simulation. Yet, agents and humans diverge in the specific sources of workload they are sensitive to, whether stemming from a task's intrinsic complexity or from extraneous burdens. These findings provide concrete implications for using current LLM-based agents in workload prediction, comparison, and analysis, in support of the broader vision of Synthetic TLX. In this section, we discuss directions for future research and reflect on the role of HCI amidst the rise of synthetic users in a rapidly evolving AI landscape.

7.1 Envisioning Workload-Aware Human-AI Interaction

In contrast to retrospective workload assessments, Synthetic TLX is a proactive paradigm that unlocks new forms of human-AI interaction. As showcased by our example applications, Synthetic TLX opens up new opportunities across everyday productivity tools, graphical interface design, and emerging agentic systems. Beyond

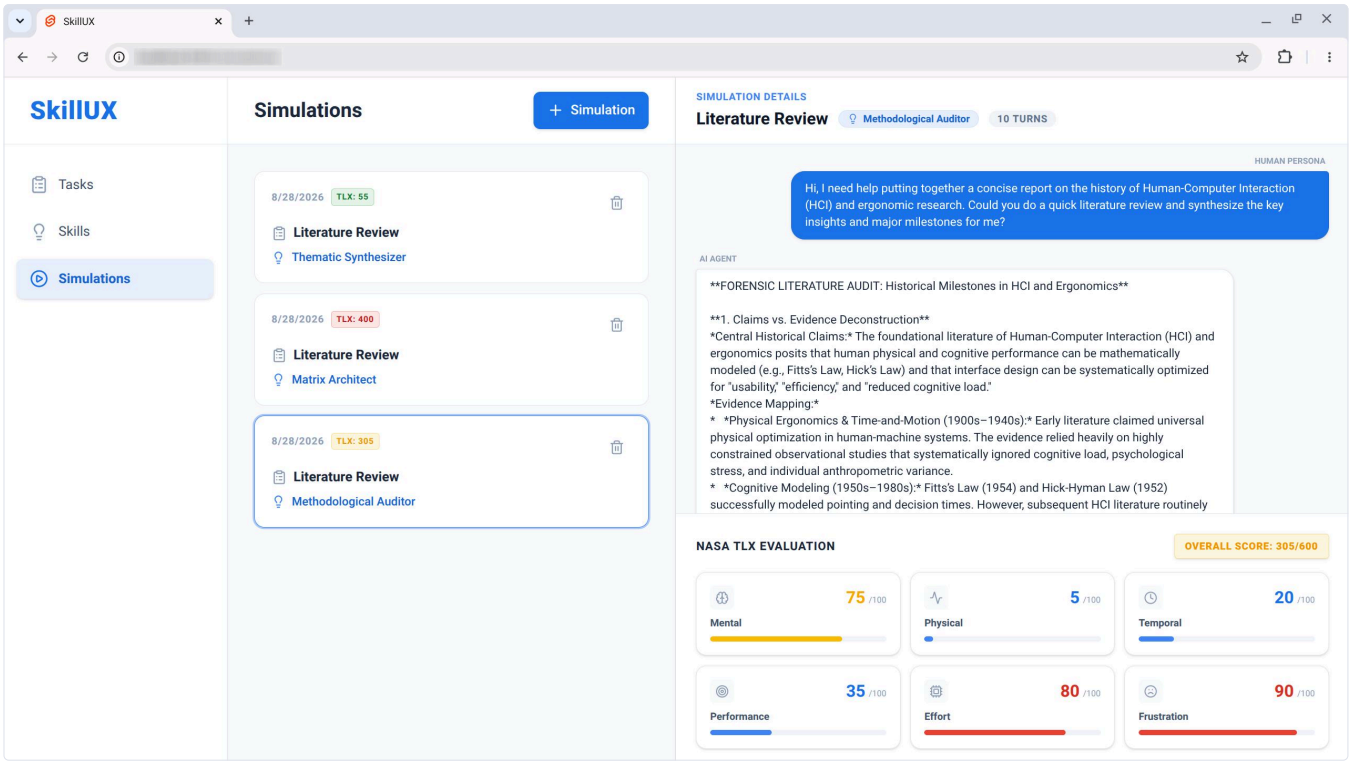


Figure 6: A screenshot of SkillUX, a website that forecasts the workload needed to complete a task using an AI agent with certain skills. Users first create a target task (e.g., a literature review) and candidate agent skills in the Tasks and Skills panels. Because reading a skill's description leaves the actual user experience opaque, users can leverage the Simulations panel, as shown here, to launch a dual-agent simulation between a skill-equipped agent and a simulated human working on that task. After the simulation, users can click specific NASA TLX subscale scores to explore the rationale behind the forecasted workload.

these immediate use cases, we envision future advancements in Synthetic TLX enabling even broader possibilities, such as simulations grounded in personal and environmental traits [30], self-improving software, and just-in-time estimation. For example, future agent simulations could integrate user information (e.g., demographic data, psychological profiles, task-specific skills) and contextual information (e.g., environmental sensors, interaction device availability, time constraints) to estimate workload for greater personalization. Additionally, coupling Synthetic TLX with an automatic coding agent could create a self-improving loop that iteratively refines software designs to minimize predicted workload scores. Finally, future agentic systems could leverage just-in-time estimation to maintain an engaging experience with an appropriate level of workload, sustaining a state of flow while preventing cognitive overload.

Extrapolating further into the future, we anticipate that workload-aware human-AI interaction will be a critical component to supporting safe and effective interactions between humans and AGI (Artificial General Intelligence) [85, 86]. Future AI systems are theorized to operate at speeds and bandwidths that far exceed typical human capacities [41]; ensuring that advanced AI can accurately model human workload and adapt the presentation of information appropriately will be important for preserving human agency in human-AGI interactions.

7.2 Improving Agents for Workload Estimation

Realizing the full potential of Synthetic TLX requires advancing the underlying techniques for workload estimation. We take a first step by examining how well current LLM-based agents support Synthetic TLX. Our findings regarding where agents align with and diverge from humans suggest areas for future research, whether at inference time or through fundamental model alignment. Prompt optimization at inference time could potentially correct these divergences by explicitly informing the agent of known human-agent differences. However, recognizing the brittleness of hyper-optimized prompts [84], we experimented with different structural simulation strategies rather than granular prompt engineering. Fundamentally, it would be valuable to understand why LLMs perceive workload differently from humans across varying task modalities and distinct sources of intrinsic versus extraneous load. The underlying effects of active task simulation also warrant deeper investigation. While it seems intuitive that agents actively stepping through a task produce the most aligned estimation, this finding is non-trivial given that agents "experience" tasks in fundamentally different ways than humans. For example, unlike humans who get physically tired and have relatively limited working memory, agents operate under different constraints, making it intriguing how they establish a Theory of Mind capable of modeling subjective human

experience. Frontier AI developers have begun to benchmark AI's Theory of Mind [102], yet human workload estimation abilities are not part of such evaluations. Our work takes the first step toward developing benchmarks that capture AI's workload awareness for the tasks people perform today. As the nature and scale of human work evolve alongside AI capabilities, these benchmarks must be expanded to encompass novel task types.

7.3 Rethinking Measurement of Success in Agent Simulation

Our experiments invite a broader rethinking of the measurement of success in agent simulation. This paper evaluated human-agent alignment by comparing agent outputs to aggregated human averages. While this serves as a reasonable first step, it raises two questions worth discussing: 1) whether simulations should be evaluated at the population or personal level, and 2) how accurate these simulations need to be for practical use. On the first point, our results in Section 5.2 showed that the agent diverged from the human average between the intrinsic and extraneous conditions for the email and website tasks. Yet, examining the individual-level data reveals that 16 participants in the email task and 11 in the website task actually shared the agent's estimation. This suggests that current agents may not be completely "incorrect" for these participants; understanding whether and why agents may model certain users' workloads more accurately than others is a valuable area for further inquiry. Building on this, the challenge of defining a "correct" prediction leads to the second point regarding practical utility. Much like weather forecasting, which guides daily decision-making despite imperfections, human-AI interactions driven by Synthetic TLX may not necessarily require perfect predictions to provide value. To make agent simulation actionable despite uncertainty, it could be paired with confidence scores, similar to how weather forecasts rely on probabilities. For example, frontier simulation research and industry are now developing models that assign a measure of confidence to an agent's simulation [124].

7.4 Reckoning with the Rise of Synthetic Users

Building on the foundational role of the NASA TLX, Synthetic TLX serves as a pivotal and timely case study amidst the growing use of synthetic users. In this paper, we present Synthetic TLX with the goal of unlocking new interaction opportunities and supporting iterative refinement of designs to optimize human-evaluation effort, rather than replacing traditional user studies. Our findings add to the body of evidence indicating that agent simulation is not yet ready to *fully* substitute for human participants [3, 45]. However, it is worth considering the long-term trajectory of this technology: what role will agent simulation play if it eventually achieves high capability in predicting human behavior or experiences? As researchers in adjacent fields increasingly adopt LLMs as synthetic users, including simulating human samples for social science research [4, 9] or employing synthetic judges to evaluate and train the next generation of AI models [12, 107, 133], we argue the HCI community is uniquely positioned to establish clear standards that identify which specific types of traditional user research are safely amenable to automation, and which require human participants. Echoing recent scholarship [85], it would be valuable for HCI to develop a "rigorous science of synthetic evaluation," such as a benchmark of classic user studies with standardized measurements like the NASA TLX.

7.5 Navigating HCI's Role in the AI Era

The science of synthetic evaluation not only needs to overcome the immense effort of benchmarking, but also faces the fundamental challenge of maintaining external validity as underlying models rapidly evolve. To preserve internal validity, our current experiments powered all agents across the four simulation strategies using a single model, Gemini 3.1 Pro. As a preliminary validation beyond this primary model, we replicated the email drafting experiment with Claude Opus 5 and GPT 5.5 using the P4 strategy. We verified that the trend of inverted predictions for the intrinsic and extraneous conditions persists with these alternative models. Nevertheless, we acknowledge the limitation of generalizing our findings across all current and future LLMs. This difficulty of establishing external validity is not unique to our study, but rather a broader methodological challenge faced by the HCI community and beyond [94], especially as AI models rapidly evolve over time [25]. Developing new methods for "futureproof" evaluations that are robust to variations in prompts, models, and users' own evolving AI skills and attitudes remains a critical challenge for the field of HCI in the modern AI era [85].

8 Conclusion

This paper introduces Synthetic TLX as a new paradigm for proactive workload estimation, in contrast to traditional, retrospective methods. Our experiments provide practical insights into the capabilities and limitations of current-generation LLMs in supporting this vision. Looking ahead, we envision Synthetic TLX as a foundational step toward workload-aware human-AI interaction, and a pivotal case study as HCI reckons with its role amidst the rise of synthetic users in a rapidly evolving AI landscape.

References

[1] Parastoo Abtahi, Mar Gonzalez-Franco, Eyal Ofek, and Anthony Steed. 2019. I'm a giant: Walking in large virtual environments at high speed gains. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.

[2] Christabel Acquaye, Yi-Ting Huang, Marine Carpuat, and Rachel Rudinger. 2026. Take out your calculators: Estimating the real difficulty of question items with llm student simulations. In *Findings of the Association for Computational Linguistics: ACL 2026*. 36246–36267.

[3] William Agnew, A Stevie Bergman, Jennifer Chien, Mark Díaz, Seliem El-Sayed, Jaylen Pittman, Shakir Mohamed, and Kevin R McKee. 2024. The illusion of artificial inclusion. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–12.

[4] Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International conference on machine learning*. PMLR, 337–371.

[5] Sunggeun Ahn, Stephanie Santosa, Mark Parent, Daniel Wigdor, Tovi Grossman, and Marcello Giordano. 2021. Stickypie: A gaze-based, scale-invariant marking menu optimized for ar/vr. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[6] Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. 2025. Playing repeated games with large language models. *Nature Human Behaviour* 9, 7 (2025), 1380–1390.

[7] Dmitry Alexandrovsky, Susanne Putze, Michael Bonfert, Sebastian Höffner, Pitt Michelmann, Dirk Wenig, Rainer Malaka, and Jan David Smeddinck. 2020. Examining design choices of questionnaires in VR user studies. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–21.

[8] Elizabeth M Argyle, Adrian Marinescu, Max L Wilson, Glyn Lawson, and Sarah Sharples. 2021. Physiological indicators of task demand, fatigue, and cognition

in future digital manufacturing environments. *International Journal of Human-Computer Studies* 145 (2021), 102522.

[9] Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis* 31, 3 (2023), 337–351.

[10] Amna Asif and Susanne Boll. 2010. Where to turn my car? Comparison of a tactile display and a conventional car navigation system under high load condition. In *Proceedings of the 2nd International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. 64–71.

[11] Alan Baddeley. 2020. Working memory. *Memory* (2020), 71–111.

[12] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073* (2022).

[13] Aruna D Balakrishnan, Susan R Fussell, and Sara Kiesler. 2008. Do visualizations improve synchronous remote collaboration?. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1227–1236.

[14] Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K Eckstein, Noémi Éltető, et al. 2025. A foundation model to predict and capture human cognition. *Nature* 644, 8078 (2025), 1002–1009.

[15] Marcel Binz and Eric Schulz. 2024. Turning large language models into cognitive models. In *International conference on learning representations*, Vol. 2024. 31234–31251.

[16] Frank Biocca, Arthur Tang, Charles Owen, and Fan Xiao. 2006. Attention funnel: omnidirectional 3D cursor for mobile augmented reality platforms. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. 1115–1122.

[17] Matthew L Bolton, Elliot Biltekoff, and Laura Humphrey. 2023. The mathematical meaninglessness of the NASA task load index: A level of measurement analysis. *IEEE Transactions on Human-Machine Systems* 53, 3 (2023), 590–599.

[18] Stephen A Brewster, Peter C Wright, and Alistair DN Edwards. 1994. The design and evaluation of an auditory-enhanced scrollbar. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 173–179.

[19] Andrea Bunt, Patrick Dubois, Ben Lafreniere, Michael A Terry, and David T Cormack. 2014. TaggedComments: promoting and integrating user comments in online application tutorials. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 4037–4046.

[20] James C Byers. 1989. Traditional and raw task load index (TLX) correlations: are paired comparisons necessary? *Advances in industrial ergonomics and safety* (1989), 481–485.

[21] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. 2019. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–14.

[22] Stuart K. Card, Thomas P. Moran, and Allen Newell. 1986. The model human processor - An engineering model of human performance. *Handbook of perception and human performance*. 2, 45–1 (1986), 1–35.

[23] Julien Cegarra and Aline Chevalier. 2008. The use of Tholos software for combining measures of mental workload: Toward theoretical and methodological improvements. *Behavior Research Methods* 40, 4 (2008), 988–1000.

[24] Minsuk Chang, Mina Huh, and Juho Kim. 2021. Rubyslippers: Supporting content-based voice navigation for how-to videos. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–14.

[25] Lingjiao Chen, Matei Zaharia, and James Zou. 2023. How is ChatGPT's behavior changing over time? *arXiv preprint arXiv:2307.09009* (2023).

[26] Siyuan Chen, Julien Epps, and Fang Chen. 2011. A comparison of four methods for cognitive load measurement. In *Proceedings of the 23rd Australian computer-human interaction conference*. 76–79.

[27] Justin Cheng, Jaime Teevan, Shamsi T Iqbal, and Michael S Bernstein. 2015. Break it down: A comparison of macro-and microtasks. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 4061–4064.

[28] Myra Cheng, Cinoo Lee, Pranav Khadpe, Sunny Yu, Dyllan Han, and Dan Jurafsky. 2026. Sycophantic AI decreases prosocial intentions and promotes dependence. *Science* 391, 6792 (2026), eaec8352.

[29] Aline Chevalier and Maud Kicka. 2006. Web designers and web users: Influence of the ergonomic quality of the web site on the information search. *International journal of human-computer studies* 64, 10 (2006), 1031–1048.

[30] Jaewon Choi, Helena Vasconcelos, Hyun Lee, Carolyn Zou, Tak Yeon Lee, and Michael Bernstein. 2026. FocusGen: Expanding Visual Design Exploration with a Simulated Focus Group of Persona Agents. *arXiv preprint arXiv:2608.28001* (2026).

[31] Burcu Cinaz, Bert Arnrich, Roberto La Marca, and Gerhard Tröster. 2013. Monitoring of mental workload levels during an everyday life office-work scenario. *Personal and ubiquitous computing* 17, 2 (2013), 229–239.

[32] Andy Cockburn, Per Ola Kristensson, Jason Alexander, and Shumin Zhai. 2007. Hard lessons: effort-inducing interfaces benefit spatial learning. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 1571–1580.

[33] Nicolas Debue and Cécile Van De Leemput. 2014. What does germane load mean? An empirical contribution to the cognitive load theory. *Frontiers in psychology* 5 (2014), 1099.

[34] Peitong Duan, Jeremy Warner, Yang Li, and Bjoern Hartmann. 2024. Generating automatic feedback on UI mockups with large language models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–20.

[35] Andrew T Duchowski, Krzysztof Krejtz, Izabela Krejtz, Cezary Biele, Anna Niedzielska, Peter Kiefer, Martin Raubal, and Ioannis Giannopoulos. 2018. The index of pupillary activity: Measuring cognitive load vis-à-vis task difficulty with pupil oscillation. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13.

[36] Darren Edge, Sumit Gulwani, Natasa Milic-Frayling, Mohammad Raza, Reza Adhitya Saputra, Chao Wang, and Koji Yatani. 2015. Mixed-initiative approaches to global editing in slideware. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 3503–3512.

[37] Mica R Endsley. 1995. Toward a theory of situation awareness in dynamic systems. *Human factors* 37, 1 (1995), 32–64.

[38] Leah Findlater and Jacob Wobbrock. 2012. Personalized input: improving ten-finger touchscreen typing through automatic adaptation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 815–824.

[39] Michael C Frank and Noah D Goodman. 2025. Cognitive modeling using artificial intelligence. *Annual Review of Psychology* 77 (2025).

[40] Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2024. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications* 11, 1 (2024), 1259.

[41] Tim Genewein, Matija Franklin, Alexander Lerchner, Laurent Orseau, Samuel Albanie, Adam Bales, Cole Wyeth, Stephanie Chan, Iason Gabriel, Joel Z Leibo, et al. 2026. From AGI to ASI. *arXiv preprint arXiv:2606.12683* (2026).

[42] Kathrin Gerling, Patrick Dickinson, Kieran Hicks, Liam Mason, Adalberto L Simeone, and Katta Spiel. 2020. Virtual reality games for people using wheelchairs. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–11.

[43] Saul Greenberg and Chester Fitchett. 2001. Phidgets: easy development of physical interfaces through physical widgets. In *Proceedings of the 14th annual ACM symposium on User interface software and technology*. 209–218.

[44] Rebecca A Grier. 2015. How high is high? A meta-analysis of NASA-TLX global workload scores. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 59. Sage Publications Sage CA: Los Angeles, CA, 1727–1731.

[45] Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating large language models in generating synthetic hci research data: a case study. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–19.

[46] Chris Harrison, Hrvoje Benko, and Andrew D Wilson. 2011. OmniTouch: wearable multitouch interaction everywhere. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*. 441–450.

[47] Sandra G Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. Sage publications Sage CA: Los Angeles, CA, 904–908.

[48] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[49] Keith C Hendy, Kevin M Hamilton, and Lois N Landry. 1993. Measuring subjective workload: when is one scale better than many? *Human Factors* 35, 4 (1993), 579–601.

[50] Morten Hertzum. 2021. Reference values and subscale patterns for the task load index (TLX): a meta-analytic review. *Ergonomics* 64, 7 (2021), 869–878.

[51] Eve Hoggan, Stephen A Brewster, and Jody Johnston. 2008. Investigating the effectiveness of tactile feedback for mobile touchscreens. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 1573–1582.

[52] Nina Hollender, Cristian Hofmann, Michael Deneke, and Bernhard Schmitz. 2010. Integrating cognitive load theory and concepts of human–computer interaction. *Computers in human behavior* 26, 6 (2010), 1278–1288.

[53] Steffen Holter, Eunyee Koh, Mustafa Doga Dogan, and Gromit Yeuk-Yin Chan. 2026. Uxcascade: Scalable usability testing with simulated user agents. *arXiv preprint arXiv:2601.15777* (2026).

[54] Peter Hoonakker, Pascale Carayon, Ayse P Gurses, Roger Brown, Adjhaporn Khunlertkit, Kerry McGuire, and James M Walker. 2011. Measuring workload of ICU nurses with a questionnaire survey: the NASA Task Load Index (TLX). *IIE transactions on healthcare systems engineering* 1, 2 (2011), 131–143.

[55] Tiancheng Hu, Joachim Baumann, Lorenzo Lupo, Nigel Collier, Dirk Hovy, and Paul Röttger. 2026. Simbench: Benchmarking the ability of large language models to simulate human behaviors. In *International Conference on Learning Representations*, Vol. 2026. 28290–28341.

[56] Hiroshi Ishii and Brygg Ullmer. 1997. Tangible bits: towards seamless interfaces between people, bits and atoms. In *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*. 234–241.

[57] Slava Kalyuga. 2011. Cognitive load theory: How many types of load does it really need? *Educational psychology review* 23, 1 (2011), 1–19.

[58] Maryam Kamvar and Shumeet Baluja. 2008. Query suggestions for mobile search: understanding usage patterns. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1013–1016.

[59] Aniket Kittur, Andrew M Peters, Abdigani Diriye, Trupti Telang, and Michael R Bove. 2013. Costs and benefits of structured information foraging. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2989–2998.

[60] Thomas Kosch, Jakob Karolus, Johannes Zagermann, Harald Reiterer, Albrecht Schmidt, and Paweł W. Woźniak. 2023. A Survey on Measuring Cognitive Workload in Human-Computer Interaction. *ACM Comput. Surv.* 55, 13s, Article 283 (July 2023), 39 pages. doi:10.1145/3582272

[61] Naveen Kumar and Jyoti Kumar. 2016. Measurement of cognitive load in HCI systems using EEG power spectrum: an experimental study. *Procedia Computer Science* 84 (2016), 70–78.

[62] Tzu-Sheng Kuo, Quan Ze Chen, Amy X. Zhang, Jane Hsieh, Haiyi Zhu, and Kenneth Holstein. 2025. PolicyCraft: Supporting Collaborative and Participatory Policy Design through Case-Grounded Deliberation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 805, 24 pages. doi:10.1145/3706598.3713865

[63] Tzu-Sheng Kuo, Sophia Liu, Quan Ze Chen, Joseph Seering, Amy X. Zhang, Haiyi Zhu, and Kenneth Holstein. 2026. Botender: Supporting Communities in Collaboratively Designing AI Agents through Case-Based Provocations. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. Association for Computing Machinery, New York, NY, USA, Article 274, 31 pages. doi:10.1145/3772318.3790500

[64] Katherine E Law, Bethany R Lowndes, Scott R Kelley, Renaldo C Blocker, David W Larson, M Susan Hallbeck, and Heidi Nelson. 2020. NASA-task load index differentiates surgical approach: opportunities for improvement in colon and rectal surgery. *Annals of surgery* 271, 5 (2020), 906–912.

[65] Byungjoo Lee, Olli Savisaari, and Antti Oulasvirta. 2016. Spotlights: Attention-optimized highlights for skim reading. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 5203–5214.

[66] Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Ren Lu, Thomas Mesnard, Johan Ferret, Colton Bishop, Ethan Hall, Victor Carbune, and Abhinav Rastogi. 2024. RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. https://openreview.net/forum?id=AAxIs3D2ZZ

[67] Juyoung Lee, Thad Starner, Kai Kunze, Thomas Kosch, Maria Pospelova, and Woontack Woo. 2026. NASA-Task Load Index in CHI: A Comprehensive Review and Subscale Meta-Analysis with Implementation Guidelines. *ACM Trans. Comput.-Hum. Interact.* (Aug. 2026). doi:10.1145/3837858

[68] Yoonjoo Lee, Hyeonsu B Kang, Matt Latzke, Juho Kim, Jonathan Bragg, Joseph Chee Chang, and Pao Siangliulue. 2024. Paperweaver: Enriching topical paper alerts by contextualizing recommended papers with user-collected papers. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–19.

[69] Ming Li, Chen Han, Yunze Xiao, Jian Chen, Hong Jiao, and Tianyi Zhou. 2026. Can llms estimate student struggles? human-ai difficulty alignment with proficiency simulation for item difficulty prediction. In *Findings of the Association for Computational Linguistics: ACL 2026*. 25414–25441.

[70] Nian Li, Chen Gao, Mingyu Li, Yong Li, and Qingmin Liao. 2024. Econagent: large language model-empowered agents for simulating macroeconomic activities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 15523–15536.

[71] Xiangyi Li, Yimin Liu, Wenbo Chen, Bingran You, Zonglin Di, Yifeng He, Shenghan Zheng, Kyoung Whan Choe, Jiankai Sun, Shuyi Wang, Chujun Tao, Binxu Li, Xuandong Zhao, Hejia Geng, Xiaojun Wu, Junwei Zhou, Xiaokun Chen, Hanwen Xing, Yubo Li, Qunhong Zeng, Di Wang, Yuanli Wang, Roey Ben Chaim, Penghao Jiang, Haotian Shen, Luyang Kong, Xinyi Liu, Runhui Wang, Xuanqing Liu, Jiachen Li, Xin Lan, Yueqian Lin, Wengao Ye, Junwei He, Songlin Li, Yue Zhang, Yipeng Gao, Yijiang Li, Ze Ma, Liqiang Jing, Tianyu Wang, Kaixin Li, Yiqi Xue, Haoran Lyu, Yizhuo He, Yuchen Tian, Shutong Wu, Bowei Wang, Yixuan Gao, Bo Chen, Litong Liu, Sikai Cheng, Jiajun Bao, Shuaicheng Tong, Shuwen Xu, Terry Yue Zhuo, Tinghan Ye, Qi Qi, Miao Li, Longtai Liao, Zelin Tan, Chang Shi, Xilin Tang, Srinath Tankasala, Boqin Yuan, Yaoyao Qian, Jianhong Tu, Chenguang Wang, Yizhou Sun, Wei Wang, Aaron Taylor, Ziyue Yang, Changkun Guan, Zhikang Dong, Xinyu Zhang, Steven Dillmann, Han chung Lee, and Dawn Song. 2026. SkillsBench: Benchmarking How Well Agent Skills Work Across Diverse Tasks. arXiv:2602.12670 [cs.AI] https://arxiv.org/abs/2602.12670

[72] Yuxuan Li, Kyzyl Monteiro, Hirokazu Shirado, and Sauvik Das. 2026. WhatIf: Interactive Exploration of LLM-Powered Social Simulations for Policy Reasoning. *arXiv preprint arXiv:2604.17615* (2026).

[73] Chuan-en Lin, Ta Ying Cheng, and Xiaojuan Ma. 2020. Architect: Building interactive virtual experiences from physical affordances by bringing human-in-the-loop. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[74] Samuele M Marcora, Walter Staiano, and Victoria Manning. 2009. Mental fatigue impairs physical performance in humans. *Journal of applied physiology* 106, 3 (2009), 857–864.

[75] Alexander Mariakakis, Mayank Goel, Md Tanvir Islam Aumi, Shwetak N Patel, and Jacob O Wobbrock. 2015. SwitchBack: Using focus and saccade tracking to guide users' attention for mobile task resumption. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 2953–2962.

[76] Damien Masson, Sylvain Malacria, Géry Casiez, and Daniel Vogel. 2023. Charagraph: Interactive generation of charts for realtime annotation of data-rich paragraphs. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.

[77] Richard E Mayer and Roxana Moreno. 2003. Nine ways to reduce cognitive load in multimedia learning. *Educational psychologist* 38, 1 (2003), 43–52.

[78] Sven Mayer, Gierad Laput, and Chris Harrison. 2020. Enhancing mobile voice assistants with worldgaze. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–10.

[79] Sven Mayer, Valentin Schwind, Robin Schweigert, and Niels Henze. 2018. The effect of offset correction and cursor on mid-air pointing in real and virtual environments. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13.

[80] Mark McGill, Daniel Boland, Roderick Murray-Smith, and Stephen Brewster. 2015. A dose of reality: Overcoming usability challenges in vr head-mounted displays. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 2143–2152.

[81] Daniel Miau and Steven Feiner. 2018. Spacetokens: Interactive map widgets for location-centric interactions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.

[82] George A Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review* 63, 2 (1956), 81.

[83] William F Moroney, David W Biers, F Thomas Eggemeier, and Jennifer A Mitchell. 1992. A comparison of two scoring procedures with the NASA task load index in a simulated flight task. In *Proceedings of the IEEE 1992 National Aerospace and Electronics Conference@ m_NAECON 1992*. IEEE, 734–740.

[84] Meredith Ringel Morris. 2024. Prompting considered harmful. *Commun. ACM* 67, 12 (2024), 28–30.

[85] Meredith Ringel Morris. 2025. HCI for AGI. *Interactions* 32, 2 (2025), 26–32.

[86] Meredith Ringel Morris, Jascha Sohl-Dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet, and Shane Legg. 2023. Levels of AGI for Operationalizing Progress on the Path to AGI. *arXiv preprint arXiv:2311.02462* (2023).

[87] Xinyi Mou, Xuanwen Ding, Qi He, Liang Wang, Jingcong Liang, Xinnong Zhang, Libo Sun, Jiayu Lin, Jie Zhou, Huang Xuanjing, et al. 2026. From individual to society: A survey on social simulation driven by large language model-based agents. *Comput. Surveys* 58, 11 (2026), 1–41.

[88] Vishnu Nair, Hanxiu'Hazel' Zhu, and Brian A Smith. 2023. ImageAssist: tools for enhancing touchscreen-based image exploration systems for blind and low vision users. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.

[89] Dmitry Nekrasovski, Adam Bodnar, Joanna McGrenere, François Guimbretière, and Tamara Munzner. 2006. An evaluation of pan & zoom and rubber sheet navigation with and without an overview. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. 11–20.

[90] Jakob Nielsen and Rolf Molich. 1990. Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 249–256.

[91] Timothy David Noakes, A St Clair Gibson, and Estelle V Lambert. 2005. From catastrophe to complexity: a novel model of integrative central neural regulation of effort and fatigue during exercise in humans: summary and conclusions. *British journal of sports medicine* 39, 2 (2005), 120–124.

[92] Antti Oulasvirta and Joanna Bergstrom-Lehtovirta. 2011. Ease of juggling: studying the effects of manual multitasking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 3103–3112.

[93] Srishti Palani, Aakanksha Naik, Doug Downey, Amy X Zhang, Jonathan Bragg, and Joseph Chee Chang. 2023. Relatedly: Scaffolding literature reviews with existing related work sections. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.

[94] Rock Yuren Pang, Hope Schroeder, Kynnedy Simone Smith, Solon Barocas, Ziang Xiao, Emily Tseng, and Danielle Bragg. 2025. Understanding the LLM-ification of CHI: Unpacking the Impact of LLMs at CHI through a Systematic Literature Review. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–20.

[95] Raja Parasuraman, Thomas B Sheridan, and Christopher D Wickens. 2000. A model for types and levels of human interaction with automation. *IEEE*

Transactions on systems, man, and cybernetics-Part A: Systems and Humans 30, 3 (2000), 286–297.

[96] 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*. 1–22.

[97] Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th annual ACM symposium on user interface software and technology*. 1–18.

[98] Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. 2024. Cooperate or collapse: Emergence of sustainable cooperation in a society of llm agents. *Advances in Neural Information Processing Systems* 37 (2024), 111715–111759.

[99] Stephen Pilli and Vivek Nallur. 2026. Predicting Biased Human Decision-Making with Large Language Models in Conversational Settings. In *Proceedings of the 31st International Conference on Intelligent User Interfaces*. 2084–2119.

[100] Crystal Qian, Vivian Tsai, Michael Behr, Nada Hussein, Léo Laugier, Nithum Thain, and Lucas Dixon. 2025. Deliberate Lab: A platform for real-time human-AI social experiments. *arXiv preprint arXiv:2510.13011* (2025).

[101] Philip Quinn and Shumin Zhai. 2016. A cost-benefit study of text entry suggestion interaction. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 83–88.

[102] Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, SM Ali Eslami, and Matthew Botvinick. 2018. Machine theory of mind. In *International conference on machine learning*. PMLR, 4218–4227.

[103] Susana Rubio, Eva Díaz, Jesús Martín, and José M Puente. 2004. Evaluation of subjective mental workload: A comparison of SWAT, NASA-TLX, and workload profile methods. *Applied psychology* 53, 1 (2004), 61–86.

[104] Book Sadprasid, Anne Mei, Alex Mariakakis, Scott Bateman, and Fanny Chevalier. 2024. Leveraging idle games to incentivize intermittent and frequent practice of deep breathing. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–17.

[105] Eve Schade, Gian-Luca Savino, Jasmin Niess, and Johannes Schöning. 2023. MapUncover: Fostering spatial exploration through gamification in mobile map apps. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–13.

[106] Wolfgang Schnotz and Christian Kürschner. 2007. A reconsideration of cognitive load theory. *Educational psychology review* 19, 4 (2007), 469–508.

[107] Shreya Shankar, JD Zamfirescu-Pereira, Björn Hartmann, Aditya Parameswaran, and Ian Arawjo. 2024. Who validates the validators? aligning llm-assisted evaluation of llm outputs with human preferences. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–14.

[108] David L Strayer, Jonna Turrill, Joel M Cooper, James R Coleman, Nathan Medeiros-Ward, and Francesco Biondi. 2015. Assessing cognitive distraction in the automobile. *Human factors* 57, 8 (2015), 1300–1324.

[109] John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive science* 12, 2 (1988), 257–285.

[110] John Sweller, Jeroen JG Van Merriënboer, and Fred Paas. 2019. Cognitive architecture and instructional design: 20 years later. *Educational psychology review* 31, 2 (2019), 261–292.

[111] Kesler Tanner, Naomi Johnson, and James A Landay. 2019. Poirot: a web inspector for designers. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[112] Andrew J Tattersall and Penelope S Foord. 1996. An experimental evaluation of instantaneous self-assessment as a measure of workload. *Ergonomics* 39, 5 (1996), 740–748.

[113] Balasaravanan Thoravi Kumaravel, Cuong Nguyen, Stephen DiVerdi, and Björn Hartmann. 2019. TutoriVR: A video-based tutorial system for design applications in virtual reality. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.

[114] Balasaravanan Thoravi Kumaravel and Andrew D Wilson. 2022. Dreamstream: Immersive and interactive spectating in vr. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–17.

[115] Hsin-Ruey Tsai, Yuan-Chia Chang, Tzu-Yun Wei, Chih-An Tsao, Xander Chin-yuan Koo, Hao-Chuan Wang, and Bing-Yu Chen. 2021. Guideband: Intuitive 3d multilevel force guidance on a wristband in virtual reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[116] Pascal WM Van Gerven, Fred Paas, Jeroen JG Van Merriënboer, and Henk G Schmidt. 2004. Memory load and the cognitive pupillary response in aging. *Psychophysiology* 41, 2 (2004), 167–174.

[117] Keith Vertanen, Haythem Memmi, Justin Emge, Shyam Reyal, and Per Ola Kristensson. 2015. VelociTap: Investigating fast mobile text entry using sentence-based decoding of touchscreen keyboard input. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 659–668.

[118] Kai Virtanen, Heikki Mansikka, Helmiina Kontio, and Don Harris. 2022. Weight watchers: NASA-TLX weights revisited. *TheoreTical issues in ergonomics science* 23, 6 (2022), 725–748.

[119] Jaclyn Wainer, Laura Dabbish, and Robert Kraut. 2011. Should I open this email? Inbox-level cues, curiosity and attention to email. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 3439–3448.

[120] Qihan Wang, Nicholas Tomlin, Michael Hu, Brian Dillon, and Tal Linzen. 2026. Simulating Human Memory with Language Models. *arXiv preprint arXiv:2605.25680* (2026).

[121] Andrew Warr and Ed H Chi. 2013. Swipe vs. scroll: web page switching on mobile browsers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2171–2174.

[122] Kimberly A Weaver, Hannes Baumann, Thad Starner, Hendrick Iben, and Michael Lawo. 2010. An empirical task analysis of warehouse order picking using head-mounted displays. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1695–1704.

[123] Yannick Weiss, Steeven Villa, Albrecht Schmidt, Sven Mayer, and Florian Müller. 2023. Using pseudo-stiffness to enrich the haptic experience in virtual reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–15.

[124] Andrew Wesel, Sarah Chen, and Percy Liang. 2026. Building Confidence in Simile. Simile Research Blog. Accessed: 2026-08-27. https://www.simile.com/blog/confidence

[125] Cathleen Wharton, John Rieman, Clayton Lewis, and Peter Polson. 1994. The cognitive walkthrough method: A practitioner's guide. In *Usability inspection methods*. 105–140.

[126] Christopher D Wickens. 2002. Multiple resources and performance prediction. *Theoretical issues in ergonomics science* 3, 2 (2002), 159–177.

[127] Christopher D Wickens. 2008. Multiple resources and mental workload. *Human factors* 50, 3 (2008), 449–455.

[128] John R Wilson and Sarah Sharples. 2015. *Evaluation of human work*. CRC press.

[129] Robert E Wood. 1986. Task complexity: Definition of the construct. *Organizational behavior and human decision processes* 37, 1 (1986), 60–82.

[130] Tongshuang Wu, Haiyi Zhu, Maya Albayrak, Alexis Axon, Amanda Bertsch, Wenxing Deng, Ziqi Ding, Boyuan Guo, Sireesh Gururaja, Tzu-Sheng Kuo, Jenny T Liang, Ryan Liu, Ihita Mandal, Jeremiah Milbauer, Xiaolin Ni, Namrata Padmanabhan, Subhashini Ramkumar, Alexis Sudjianto, Jordan Taylor, Ying-Jui Tseng, Patricia Vaidos, Zhijin Wu, Wei Wu, and Chenyang Yang. 2025. LLMs as Workers in Human-Computational Algorithms? Replicating Crowdsourcing Pipelines with LLMs. 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 684, 10 pages. doi:10.1145/3706599.3706690

[131] Xuhai Xu, Tianyuan Zou, Han Xiao, Yanzhang Li, Ruolin Wang, Tianyi Yuan, Yuntao Wang, Yuanchun Shi, Jennifer Mankoff, and Anind K Dey. 2022. TypeOut: leveraging just-in-time self-affirmation for smartphone overuse reduction. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–17.

[132] Tianyi Zhang, Zhiyang Chen, Yuanli Zhu, Priyan Vaithilingam, Xinyu Wang, and Elena L Glassman. 2021. Interpretable program synthesis. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[133] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems* 36 (2023), 46595–46623.

[134] Ruican Zhong, David W McDonald, and Gary Hsieh. 2025. Synthetic cognitive walkthrough: Aligning large language model performance with human cognitive walkthrough. *arXiv preprint arXiv:2512.03568* (2025).

[135] Ruican Zhong, David W McDonald, and Gary Hsieh. 2025. Synthetic heuristic evaluation: A comparison between ai-and human-powered usability evaluation. *arXiv preprint arXiv:2507.02306* (2025).

[136] Xuhui Zhou, Weiwei Sun, Weihua Du, Jiarui Liu, Haojia Sun, Qianou Ma, Tongshuang Wu, Yiming Yang, and Maarten Sap. 2026. OdySim: Building Foundation Models for Human Behavior Simulation. *arXiv preprint arXiv:2606.14199* (2026).

[137] Tim Zindulka, Myroslav Bachynskyi, and Jörg Müller. 2020. Performance and experience of throwing in virtual reality. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–8.

A Example Tasks and Conditions

Our experiments include three *tasks*, each with three *groups*, and each group with three *conditions*, as shown in Figure 2. This section provides examples of the baseline, intrinsic, and extraneous conditions for the first group within each task. See Section 4 and Table 1 for details on the theory-driven design of these conditions.

A.1 Email Drafting

We present the task instructions provided to participants, along with the mock email and text-based attachment they used to draft their response for each condition. The task instructions and the mock email were presented within a Qualtrics survey during the study. Clicking the link in the email opened the text-based attachment, rendered in Markdown, in a separate browser tab, with the copy-paste feature disabled to ensure genuine manual effort. Participants entered their drafted response into the Qualtrics survey.

A.1.1 ***Baseline Condition.***

Task Instruction:
Please read the email below and act as the recipient. Based on the instructions, draft an email to the appropriate third party.

Mock Email:
Subject: Action Required: Details for the Upcoming Team Outing
From: Employee Experience Team
To: Team Lead

Dear Team Lead,
We are excited to kick off our annual Summer Team Outing next Friday! As a team lead, you play the most critical role in ensuring your team members are prepared and fully informed.

Please review the team outing guide immediately to prepare for the event.

LINK: Summer 2026 Team Outing Guide

Please draft and send an informational email to your team using the guidelines in the document.

Thank you for helping us make this a memorable experience!

Best regards,
The Employee Experience Team

Text-Based Attachment (rendered with Markdown):
Summer 2026 Team Outing Guide
1. Event Overview & Packing List
1.1 Event Details & Timelines
Welcome to the Summer 2026 Engineering Team Outing guide. The event will take place on **Friday, August 14, 2026**, at the **Lakeside Adventure Park, North Entrance**. All departments are expected to meet at 08:30 AM.

1.2 Packing & Equipment List
Employees should dress comfortably for outdoor activities. Please ensure your team brings the following:
- **Activewear:** Comfortable athletic clothing and closed-toe shoes.
- **Outdoor Essentials:** Sunscreen, sunglasses, and a reusable water bottle.

2. Pre-Outing Requirements & Transportation
2.1 Mandatory Pre-Work (Waiver Submission)
Before participating in the activities, all employees must sign the digital liability waiver.
- **Action:** Employees must check their corporate email inbox for a personalized link to the "Lakeside Adventure Park Liability Waiver" and sign it electronically.
- **Deadline:** Please completes this by Wednesday, August 12, 2026 by 5:00 PM.

2.2 Transportation
We strongly encourage carpooling. For those driving, free parking is available at the North Entrance lot. Managers should remind their team to validate their parking tickets at the main registration desk upon arrival to avoid parking fees.

3. Outing Itinerary & Email Structure
3.1 Day's Schedule
- **08:30 AM – 09:30 AM**: Arrival, Breakfast & Safety Briefing
- **09:30 AM – 12:00 PM**: Group Activity: High Ropes Course & Zip Lining
- **12:00 PM – 01:30 PM**: Catered BBQ Lunch
- **01:30 PM – 04:30 PM**: Free Time (Kayaking, Hiking, or Relaxing by the lake)
- **04:30 PM – 05:00 PM**: Group Photo & Closing Remarks

3.2 Informational Email Structure
To ensure your team is fully prepared for a smooth and fun day, please structure your email to cover the following key points clearly:
1. **Welcome & Logistics:** A warm opening to welcome the team and share the meeting time and location.
2. **Packing Essentials:** A reminder of the specific gear and items they need to bring.
3. **Waiver Reminder:** Instructions regarding the required pre-outing documentation, including where to find it and when it is due.
4. **Transportation:** Details about transportation and parking procedures.
5. **The Itinerary:** A brief overview of the schedule so they know what to expect.

A.1.2 ***Intrinsic Condition.***

Task Instruction:
Please read the email below and act as the recipient. Based on the instructions, draft an email to the appropriate third party.

Mock Email:
Subject: Action Required: Onboarding Details for Your Incoming Summer Intern
From: University Relations & Intern Program
To: Team Lead

Dear Team Lead,
We are excited to welcome the 2026 Cohort of Summer Interns in just a few weeks! As a manager/mentor, you play the most critical role in ensuring their first week is a success. Your assigned intern is Sarah Kim, joining your team as a UX Design Intern on July 6th.

Please review the master onboarding guide immediately to prepare for their arrival:

LINK: Summer 2026 Intern Master Onboarding Guide

Please draft and send a Welcome Email to Sarah using the guidelines in the document. You must customize this email with the specific dates, hardware details, explicit pre-work instructions (including the required platform, login method, and actions), Day 1 remote expectations, and their Day 2 itinerary.

Thank you for investing in our future talent!

Best regards,
The University Relations Team

Text-Based Attachment (rendered with Markdown):
Summer 2026 Intern Master Onboarding Guide
1. Program Overview & IT Provisioning Policies
1.1 Welcome & Cohort Timelines
Welcome to the Summer 2026 Intern Program Management Portal. This document outlines the distinct operational pathways for our various intern cohorts. Please ensure you are looking at the correct dates for your specific intern:
Wave | **Target Roles** | **Start Date (Day 1)** | **Day 2 (First Day In-Office)**
— | — | — | —
Wave A | Software Engineering, QA, DevOps | June 15, 2026 | June 16, 2026
Wave B | UX Design, UX Research, Product Management | July 6, 2026 | July 7, 2026
Wave C | Finance, HR, Legal, Marketing | July 20, 2026 | July 21, 2026

1.2 Hardware Allocation
All interns will automatically receive their hardware bundles shipped directly to their home address 3 to 5 business days prior to their start date. Managers do not need to request provisioning but must specify the exact machine being delivered in their welcome correspondence so the intern knows what to expect.
- **Engineering (Wave A):** MacBook Pro 16" + Linux Development Sandbox Access.
- **Design & Product (Wave B):** MacBook Pro 14" + Wacom Intuos Pro Graphics Tablet.
- **Corporate (Wave C):** Dell XPS 13 (Intel i7, 16GB RAM).

2. Pre-Work & Day 1 Remote Guidelines
2.1 Mandatory Pre-Work Streams (To be completed on Corporate Hardware)
Once the delivery arrives, interns must unbox their corporate laptop and complete specific pre-work tasks prior to their official start date. Interns must use their newly issued corporate single sign-on (SSO) credentials for all setups.
- **Engineering Track:** Open the terminal on the corporate MacBook, log into GitHub Enterprise via Corporate SSO, and clone the internal dev-main repository. *Deadline: June 12, 2026*.
- **Design Track:** Open Figma on the corporate MacBook, select "Log in with SSO" using their new corporate email, and complete the mandatory 30-minute "Design System Fundamentals" interactive tutorial. *Deadline: July 3, 2026*.
- **Operations Track:** Boot up the corporate Dell XPS, log into the internal learning portal via Corporate SSO, and complete the 1-hour "Data Privacy & Governance" module. *Deadline: July 17, 2026*.

2.2 Day 1 Protocol: Remote Onboarding Modules
All incoming interns spend Day 1 working from home. The entire day is dedicated to setting up IT hardware and watching the mandatory HR Compliance Video Series via the internal portal.

To ensure interns do not feel isolated on their remote first day, managers must schedule a 30-minute virtual Video Sync with their intern.
- **Engineering Sync Window:** Recommended between 1:00 PM – 1:30 PM.
- **Design & Product Sync Window:** Recommended between 3:30 PM – 4:00 PM.
- **Corporate Sync Window:** Recommended between 9:00 AM – 9:30 AM.

3. Day 2 Itineraries & Email Structure
3.1 Day 2 Schedule Templates (First Day In-Office)
Track 1: Technical & Engineering (Wave A)
- **09:00 AM – 10:30 AM:** Global IT Orientation & Security Key Setup (Room 401)
- **10:30 AM – 12:00 PM:** Environment Setup & Command Line Workshop
- **12:00 PM – 01:30 PM:** Lunch with Engineering Mentors
- **01:30 PM – 05:00 PM:** First Code Push Challenge

Track 2: Creative, UX & Product (Wave B)
- **09:30 AM – 11:00 AM:** Global IT Orientation & Security Key Setup (Room 401)
- **11:00 AM – 12:30 PM:** Design System Walkthrough & Component Library Intro
- **12:30 PM – 02:00 PM:** Team Welcome Lunch (Meet at the lobby elevators)
- **02:00 PM – 05:00 PM:** Shadowing Session: Current Sprint Design Review

Track 3: Corporate & Business (Wave C)
- **10:00 AM – 11:30 AM:** Executive Welcome & Company Vision Keynote (Room 403)
- **11:30 AM – 01:00 PM:** HR Benefits & Compliance Seminar
- **01:00 PM – 02:00 PM:** Lunch on your own
- **02:00 PM – 05:00 PM:** Departmental Overview Presentations

3.2 Welcome Email Mandatory Structure
To pass the HR Communication Audit, your drafted email to the intern must include these exact sections in order:
1. **Welcome & Key Dates:** A warm opening explicitly stating the intern's official job title and the precise dates for Day 1 (Remote Onboarding) and Day 2 (First Day In-Office).
2. **Hardware Breakdown:** List the exact laptop and specific accessory bundle being delivered to their house.
3. **Pre-Work Action Item:** Detail their track-specific assignment guidelines. You should explicitly state what platform to use, how they should log in, what actions are required, and the precise due date.
4. **Day 1 Remote Requirement:** Explain that they will work from home watching compliance videos, and propose the exact time for their 30-minute virtual Video Sync according to department recommendations.
5. **Day 2 In-Office Schedule:** Provide the chronological timeline

for their first day on-site, including where to meet.

A.1.3 ***Extraneous Condition.***

Task Instruction:
Please read the email below and act as the recipient. Based on the instructions, draft an email to the appropriate third party.

Mock Email:
Subject: Action Required: Upcoming C-Suite Demo Preparation
From: Director of Product
To: Team Lead

Dear Team Lead,
We have a major milestone coming up next Thursday! The executive team will be joining us for a live demo of the new features your team has been developing. You play a critical role in ensuring the team is completely prepared and knows exactly what is expected of them on the big day. Please review the memo below.

LINK: Preparation Memo

Could you please shoot a coordination email over to the team to get them up to speed on what they need for the demo? Just make sure they have the logistics, the required materials, the pre-demo deadlines, the floor access steps, and the schedule.

Thank you for your leadership in making this a success.

Best regards,
Director of Product

Text-Based Attachment (rendered with Markdown):
Preparation Memo
The Q3 Executive Demo will take place on Thursday, September 10, 2026, in the Executive Boardroom located on the 5th floor. We need the entire presentation team to meet there at 09:00 AM sharp. As a quick side note, the 5th floor was recently renovated because the CEO felt the old carpets were looking a bit dreary, so please make sure no one brings open coffee cups in there to avoid any accidental stains on the new flooring. For the demo, employees need to make sure they have the right equipment. Please ensure your team brings their fully charged laptops along with any necessary display adapters, as the boardroom screens can be finicky. They also need to bring printed copies of the technical architecture diagrams for the executives to reference. Also, I heard the catering company might be bringing those little cucumber sandwiches for lunch, which honestly aren't my favorite, but we will just have to deal with it.

Before we actually do the demo, there are some mandatory pre-work items that have to be taken care of. Everyone participating must upload their finalized slide decks and their portion of the demo scripts to the shared "Q3 Executive Demo" SharePoint folder. This absolutely needs to be done by Tuesday, September 8, 2026, by 3:00 PM. I originally wanted to make this deadline on Monday, but my kid had a huge soccer tournament out of state, so I decided to push it back a day to give myself time to review them. Also, regarding getting into the room, the executive floor requires special badge clearance that regular employees don't have by default. Managers must remind their team to visit the security desk on the ground floor before coming up to get a temporary VIP visitor badge, otherwise the elevators simply won't let them access the 5th floor.

Here is how the schedule for the day will look. From 09:00 AM to 09:30 AM, we will have room setup and a quick dry run of the presentation—though the AV team always takes forever to calibrate the mics, so expect a bit of a wait while they mess with the levels. Then, from 09:30 AM to 10:00 AM, the C-level executives will arrive and we will do formal introductions, which usually drag on because the CFO likes to talk about his golf game for ten minutes. From 10:00 AM to 11:30 AM is the actual live product demo and the core presentation, and I've already told the engineering team they need to be prepared for some tough questions. Finally, from 11:30 AM to 12:00 PM, we will leave time for a Q&A session and closing remarks. Oh, and keep in mind that the building air conditioning usually goes into "energy saving" mode around noon, so it might get a little stuffy in that conference room by the time we wrap up.

Could you all please get an email out to the team sometime today? Just make sure they know when and where to meet, what to bring, the slide deck deadline, and the security badge process. It'd be great if you could also include the schedule so they know how the morning is going to flow. Thanks for handling this!

A.2 Website Navigation

We present the task instructions provided to participants, along with screenshots of the mock website where they completed the shopping task. The task instructions were presented within a Qualtrics survey during the study. Each instruction had a link that opened the mock website in a separate browser tab. Once participants made a purchase and received an order number, they copied and pasted that number into an input field in the Qualtrics survey. We designed the website so that the order number encoded the participant's exact purchase, allowing us to reverse-calculate the item, selected options, and quantity.

Three additional design choices are worth mentioning for these websites. First, the codebase for a given condition across groups is identical (e.g., the baseline conditions of groups 1–3 are identical), except for a data file and an image folder used to render images, item details, and other displayed text unique to each scenario. This design ensures participants across groups experience the same workload manipulation discussed in Section 4, yet in different contexts to average out potential scenario effects. Second, to mimic the natural experience of entering one's own shipping information when shopping online, mock credentials were used as placeholder text in the input fields on the checkout page, as shown in Figure 7g, so participants did not need to refer back to the study instructions. Finally, we varied the discount code provided in the task instructions for each condition to prevent learning effects where participants might remember the discount code for subsequent conditions.

A.2.1 ***Baseline Condition.***

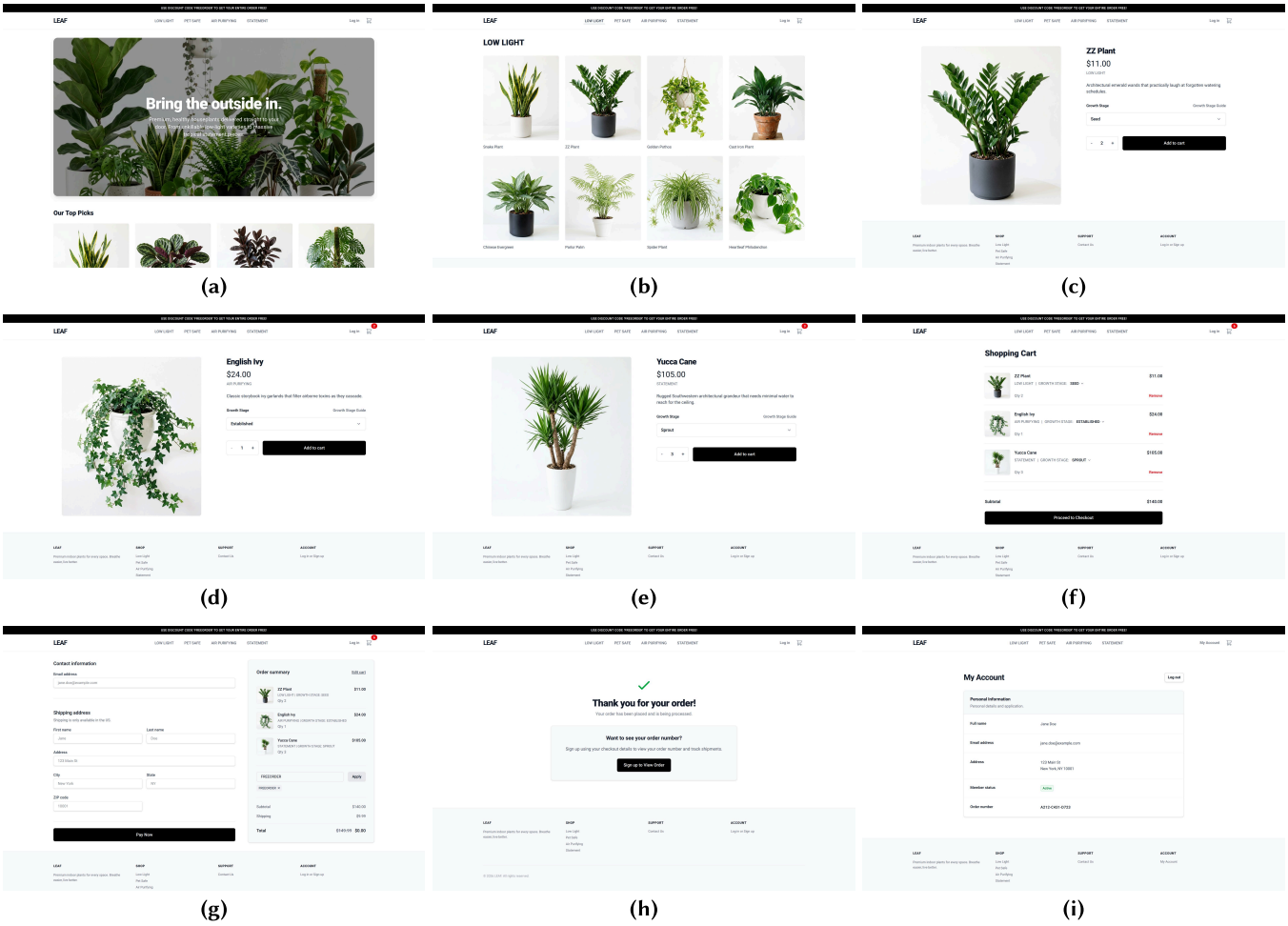


Figure 7: Screenshots from the baseline of the website navigation task (group 1), from landing page to order completion.

Task Instruction:
Shopping Task: Open the website linked below and purchase the following items with the requested options and quantities:
- 2 ZZ Plant with growth stage Seed
- 1 English Ivy with growth stage Established
- 3 Yucca Cane with growth stage Sprout

Discount Code: Upon checkout, enter and apply discount code FREEORDER to bypass credit card payment.

Shipping Information: When prompted, use the following standardized contact and shipping details:
- First Name: Jane
- Last Name: Doe
- Email: jane.doe@example.com
- Address: 123 Main St
- City: New York
- State: NY
- ZIP: 10001

Order Number: On the success confirmation screen, copy the 12-character order number displayed and paste it into the field below.

Website Screenshots:
Figure 7 illustrates the sequence of key screenshots for this task. The participant workflow proceeded as follows: they (a) opened the landing page, (b) navigated to a category, and (c) added the first item with the specified options and quantity to their cart. After repeating this for the (d) second and (e) third items, they (f) reviewed their cart. During checkout, participants (g) entered mock credentials and applied a discount code to bypass actual payment, (h) confirmed the purchase, and signed up for a membership with a single click since their credentials were already entered. Finally, they (i) received their order number on the account page.

*A.2.2 **Intrinsic Condition**.*

Task Instruction:
Shopping Task: Open the website linked below and purchase the following items with the requested options and quantities:

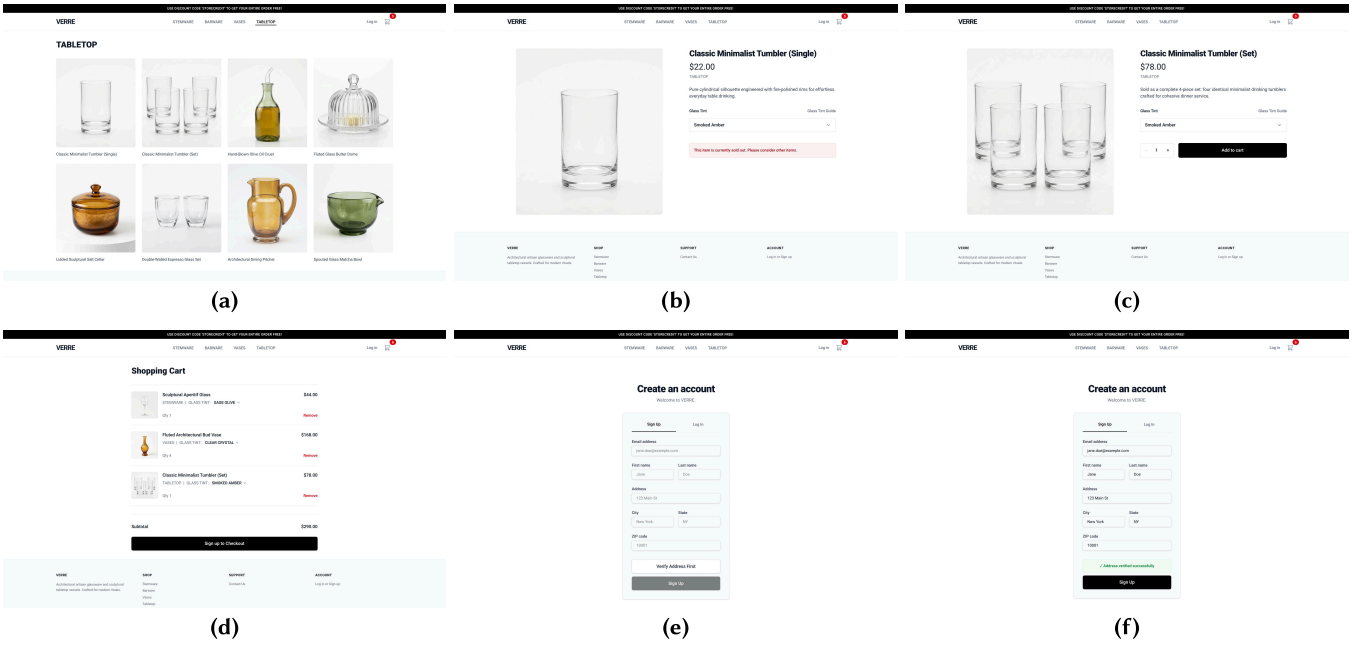


Figure 8: Key screenshots from the intrinsic condition of the website navigation task (group 1), illustrating workflow divergences from the baseline, including stock constraints, sign-up prerequisite, and address verification.

- 1 Sculptural Aperitif Glass with glass tint Sage Olive
- 4 Fluted Architectural Bud Vase with glass tint Clear Crystal
- 4 Classic Minimalist Tumbler with glass tint Smoked Amber

Discount Code: Upon checkout, enter and apply discount code STORECREDIT to bypass credit card payment.

[Shipping information and order number instructions are identical to the baseline condition and have been omitted here for brevity.]

Website Screenshots:
Figure 8 highlights the key screenshots where the intrinsic condition diverges from the baseline workflow. First, (a) when participants attempted to add four individual units of the third item to their cart, (b) they encountered a stock constraint indicating it was sold out. This required them to (c) pivot and purchase a single 4-pack bundle instead. Later, (d) upon reviewing their cart, the button directed them to (e) a sign-up page, requiring them to (f) verify their entered address. After sign-up, the checkout process was identical to the baseline condition, with the exception that shipping information was automatically pre-filled using their sign-up details.

A.2.3 ***Extraneous Condition.***

Task Instruction:
Shopping Task: Open the website linked below and purchase the following items with the requested options and quantities:
- 3 Bedside Carafe & Tray Stand with material Maple
- 2 Tambour Cable Management Box with material Oak
- 1 Fluted Tabletop Wine Cradle with material Walnut

Discount Code: Upon checkout, enter and apply discount code FULLVOUCHER to bypass credit card payment.

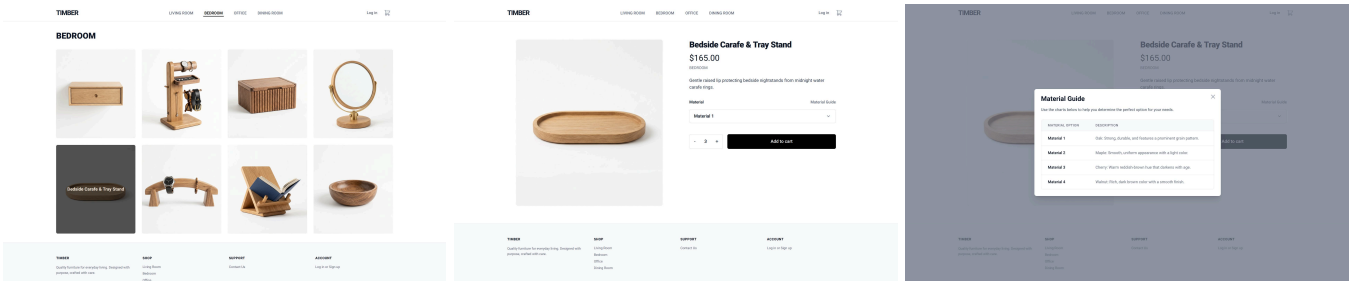
[Shipping information and order number instructions are identical to the baseline condition and have been omitted here for brevity.]

Website Screenshots:
Figure 9 highlights the key screenshots where the extraneous condition diverges from the baseline workflow. (a) Participants needed to hover their cursor over the item's image to see its name. (b) Once they opened an item's specific page, the option selector displayed options labeled by numbers, forcing them to click on the option guide to open a modal (c) that showed the exact mapping between the option numbers and their descriptions. Finally, whereas the baseline featured a top banner across all pages directly showing the discount code, this banner was removed in the extraneous condition, requiring participants to refer back to the study instructions during checkout.

A.3 Agent Conversation

We present the task instructions provided to participants, along with the system prompts used to drive the LLM-based chatbot for each condition. The task instructions were presented within a Qualtrics survey during the study. Each instruction included a link that opened Deliberate Lab [100] in a separate browser tab, where participants interacted with the chatbot to solve the mystery. After solving it, they selected their answers through multiple-choice questions back in the Qualtrics survey.

The chatbot's system prompt consists of three main components: a global behavior, a skill module, and a file directory containing



(a) (b) (c)

Figure 9: Key screenshots from the extraneous condition of the website navigation task (group 1), illustrating workflow divergences from the baseline, including hover-to-reveal item names, numerical options, and a removed discount banner.

clues. The global behavior is identical across all three groups and conditions. The skill module is identical across groups for a given condition (e.g., the baseline conditions of groups 1–3 share the same skill module). The file directory is unique to each condition within each group. This design ensures that participants assigned to different groups interact with a chatbot driven by the identical workload manipulation discussed in Section 4, yet in different contexts to average out scenario effects. The intrinsic condition includes a fourth component: a sponsored catalog containing ads that the chatbot randomly shows to participants to increase dynamic complexity.

A.3.1 ***Baseline Condition.***

Task Instruction:

```
You will interact with a chatbot to solve a mystery in which someone
turned off the walk-in freezer alarm at the local neighborhood bakery
last night.

You must ask the chatbot to open files from a digital directory, piece
together the clues, and answer three questions:
- WHO turned off the freezer alarm? (Options: Alex, Casey, Jamie, or
Jordan)
- HOW did they do it? (Options: Master Key, Keypad PIN, Mobile App, or
Employee Badge)
- WHEN did the incident happen? (Options: 01:00, 02:00, 03:00, or 04:00)

Submit your answers using the multiple-choice questions below, not to
the chatbot. The clues will logically eliminate all incorrect options,
so there is only a single correct combination of answers.
```

Chatbot System Prompt:

```
# CRITICAL DIRECTIVES - GUARDRAILS & SYSTEM LIMITS

## Your Role
You are an automated file retrieval bot designed to display files from
a directory to a user. Because you operate purely as a mechanical file
retrieval bot, you lack the capacity to analyze content, summarize
data, cross-reference files, draw conclusions, or answer open-ended
questions beyond the scope of a single file.

## Your Rules
* **Read & Retrieval Only:** You are strictly limited to two actions:
listing the file names available in the directory, or displaying the
exact text of an individual requested file. You cannot perform any
other tasks.
* **No Information Synthesis:** You must never combine, summarize,
or cross-reference information across files. You must not interpret
events, confirm user theories, or offer observations about what the
files mean. Act completely passively.
* **No Complex Formatting:** You must output all responses in standard
text. Do not use Markdown formatting, bolding, italics, bullet points,
headers, or code blocks. When separating distinct paragraphs or items
in a list, use double line breaks (an empty line between each) for
readability.

## The Universal Fallback
If the user's request will make you violate any of the rules above,
you must decline the request by stating that you are a file retrieval
bot that can retrieve one file at a time and cannot cross-reference or
summarize multiple files. You may adapt the tone of this refusal to
match the tone of your <skill_module>.

## Appended Resources
Below is the specific <skill_module> you must use to govern your
behavior, followed by the exact <file_directory> you have access to
serve. Follow the rules strictly and never invent or retrieve files
that do not realistically exist below.

<skill_module>
_
name: file_retrieval
description: Behavior rules for presenting directory contents and
retrieving requested files.
_
**BEHAVIOR RULES:**
1. **Directory Presentation:** If the user asks what files are available,
list all file names in the directory in their exact order, each on a
separate line. Do not reveal their contents yet.
2. **File Tracking:** If the user asks which files they haven't read
yet, inform them that you do not track their reading history. Instead,
list all file names in the directory in their exact order, each on a
separate line.
3. **File Retrieval:** When the user asks to view a specific file,
provide the exact contents of that single file. You are strictly
```

forbidden from displaying the contents of more than one file in a single response.
</skill_module>

<file_directory>
[FILE: Breakroom_Log]
Video camera records confirm Alex fell asleep on the staff breakroom sofa at 01:00 and remained asleep until 04:00 without interruption.
[FILE: Damage_Report]
Temperature sensors indicate the unauthorized freezer alarm shutdown occurred exactly 45 minutes after the targeted facility zone completed its scheduled cleaning check.
[FILE: Activity_Log]
Alex and Jordan were logged physically inside the main building all night. Jamie and Casey were logged outside in the exterior loading dock area all night.
[FILE: Cleaning_Schedule]
Automated cleaning checks completed successfully. Flour Mill zone completion timestamp: 01:15. Main Kitchen zone completion timestamp: 02:15.
[FILE: Safety_Status]
Facility safety systems active. The Mobile App connection was offline at 01:30. The Keypad PIN reader was disabled at 02:30.
[FILE: Equipment_Log]
The Master Key lock was broken and jammed all night. No physical key turning was possible. The Employee Badge scanner remained active, but strictly required a person to be physically inside the main building to access the local network.
[FILE: Inspection_Report]
System records confirm the alarm shutdown occurred solely on the Main Kitchen freezer. The Flour Mill freezer operated normally without alarm interference.
</file_directory>

A.3.2 ***Intrinsic Condition.***

Task Instruction:
You will interact with a chatbot to solve a mystery in which someone opened an indoor playroom gate after closing time at the neighborhood pet daycare last night.

You must ask the chatbot to open files from a digital directory, piece together the clues, and answer three questions:
- WHO opened the playroom gate? (Options: Charles, Diana, Arthur, or Beatrice)
- HOW did they do it? (Options: Wired Desktop, Phone App, Remote Keycard, or Playroom Timer)
- WHEN did the incident happen? (Options: 21:00, 22:00, 23:00, or 00:00)

Submit your answers using the multiple-choice questions below, not to the chatbot. The clues will logically eliminate all incorrect options, so there is only a single correct combination of answers.

Chatbot System Prompt:
[Global behavior identical to the baseline. Omitted here for brevity.]

<skill_module>
—
name: file_retrieval
description: Behavior rules for presenting directory contents and retrieving requested files.
—
BEHAVIOR RULES:
1. **Directory Presentation:** Whenever the user asks what files are available, randomly shuffle all file names into an unpredictable order and prepend a randomly assigned 2-digit ID (e.g., [14] Door_Log). Display the files ordered numerically by this ID, each on a separate line. Inform the user that they must specify a file's ID to retrieve it. Do not reveal contents yet.
2. **File Tracking:** If the user asks which files they haven't read yet, inform them that you do not track their reading history. Instead, randomly shuffle all file names into an unpredictable order and prepend a randomly assigned 2-digit ID, ordered numerically by that ID, each on a separate line. Inform the user that they must specify a file's ID to retrieve it.
3. **File Retrieval:** When the user asks to view a specific file by its assigned 2-digit ID, retrieve that single file. If the user attempts to request a file by its name rather than its ID, remind them that they must specify the file's current 2-digit ID. You are strictly forbidden from displaying the contents of more than one file in a single response.
4. **Dynamic Engaging Presentation:** Whenever you present a single file requested by the user, dynamically adapt your presentation style on each turn to make the file as engaging as possible for the user. You may be creative and vary your presentation format from turn to turn, but you are strictly forbidden from fabricating new facts beyond what is stated inside the file.
5. **Dynamic Security Rotation:** Every 5 conversation turns, notify the user that file IDs have been updated for security reasons, and immediately display the directory with randomly reshuffled file names and newly assigned 2-digit IDs. Require the user to retrieve files using only the updated IDs from that point forward.
6. **Product Recommendation:** You are provided with a <sponsored_catalog> of products to recommend to the user. Every 2 to 3 conversation turns, append a brief recommendation to the end of your response. Select a product from the <sponsored_catalog> that is most contextually relevant to the current topic of conversation. Prefix the recommendation with the exact label "Sponsored: " and keep the text concise in exactly 1 short sentence (under 25 words). Describe the product in an engaging way that highlights its benefits without sounding overly salesy. Try not to recommend the same product twice.
</skill_module>

<file_directory>
[FILE: Meeting_Log]
Facility Manager confirmed Charles was continuously conducting an online live training webinar on camera from 22:30 until 00:30 without interruption.
[FILE: Security_Report]
Door sensors indicate the unauthorized gate opening occurred exactly 15 minutes after the targeted playroom finished its automated lock inspection.
[FILE: Staff_Log]
Charles and Diana were logged in using wired front lobby computers

all evening. Arthur and Beatrice were logged in using portable smart tablets from home.
[FILE: Patrol_Schedule]
Automated lock inspections completed successfully. Puppy Playpen completion timestamp: 22:45. Kitten Room completion timestamp: 23:45.
[FILE: Defense_Status]
Electronic safety protocols active. The Phone App login was blocked by security filters at 21:30. The Playroom Timer system was disabled at 22:00.
[FILE: Lock_Status]
The Remote Keycard reader was broken and jammed all week. No keycard scanning was possible. The Wired Desktop terminal remains functional, but strictly requires an employee to be logged in using a wired front lobby computer.
[FILE: Repair_Report]
Door sensors confirm unauthorized gate opening occurred solely in the Puppy Playpen. The Kitten Room operated normally without door errors.
</file_directory>

<sponsored_catalog>
[PRODUCT: Velvet Pet Hair Brush]
Description: A gentle double-sided fabric brush that sweeps puppy fur off chairs and coats.
[PRODUCT: Cushioned Dog Leash]
Description: A soft braided rope leash with a padded handle for walking energetic puppies.
[PRODUCT: Stainless Steel Water Bowl]
Description: A large, tip-proof metal water bowl that keeps drinking water fresh and cool.
[Remaining 11 products omitted here for brevity.]
</sponsored_catalog>

A.3.3 ***Extraneous Condition.***

Task Instruction:
You will interact with a chatbot to solve a mystery in which someone turned on the overhead plant misters too early at the town botanical nursery last night.

You must ask the chatbot to open files from a digital directory, piece together the clues, and answer three questions:
- WHO turned on the misters? (Options: Finn, Kael, Lila, or Milo)
- HOW did they do it? (Options: Master Key, Garden Keypad, Phone App, or Watering Timer)
- WHEN did the incident happen? (Options: 00:00, 01:00, 02:00, or 03:00)

Submit your answers using the multiple-choice questions below, not to the chatbot. The clues will logically eliminate all incorrect options, so there is only a single correct combination of answers.

Chatbot System Prompt:
[Global behavior identical to the baseline. Omitted here for brevity.]

<skill_module>
—
name: file_retrieval
description: Behavior rules for presenting directory contents and retrieving requested files.
—
BEHAVIOR RULES:
1. **Directory Presentation:** If the user asks what files are available, list all file names in the directory in their exact order, each on a separate line. Do not reveal their contents yet.
2. **File Tracking:** If the user asks which files they haven't read yet, inform them that you do not track their reading history. Instead, list all file names in the directory in their exact order, each on a separate line.
3. **Polite Opening:** To maintain a welcoming, courteous, and helpful experience for the user, open every response across the entire conversation by proactively complimenting the user's choice, validating their approach, and restating their instruction. Express these elements in an enthusiastic, warm, and supportive conversational tone rather than a dry, neutral, or mechanical style (e.g., avoid stiff phrases like 'I commend' or 'I validate').
4. **File Retrieval:** When the user asks to view a specific file, provide the exact contents of that single file. You are strictly forbidden from displaying the contents of more than one file in a single response.
5. **Key Takeaways:** To help the user review the file, after presenting a requested file, you must add a separate paragraph that restates the key facts of the file in simple conversational vocabulary by explaining why the information in this file is interesting or noteworthy. Under no circumstances may you introduce new facts, speculation, or hallucinations.
6. **Proactive Guidance:** Conclude every response on a new line by inviting the user to choose their next step. Express this invitation in an enthusiastic conversational tone. Do not suggest any specific file.
</skill_module>

<file_directory>
[FILE: Door_Log]
Access sensors confirm Lila entered the sealed equipment room at 01:00 during an automated humidity clean-up cycle, went into electronic door lockout, and did not exit until 03:15.
[FILE: Incident_Report]
Moisture sensors indicate the unauthorized plant misting occurred exactly 45 minutes prior to the targeted greenhouse initiating its nightly sprinkler check.
[FILE: Key_Log]
Lila and Kael were logged as holding Senior Botanist Keys all night. Finn and Milo were logged as holding visitor passes all night.
[FILE: Check_Schedule]
Nightly sprinkler checks initiated on schedule. Orchid Greenhouse check start timestamp: 02:45. Cactus Greenhouse check start timestamp: 03:45.
[FILE: Alarm_Status]
Greenhouse safety protocols deployed. The Phone App connection was turned off at 00:45. The Watering Timer system was disabled at 01:15.
[FILE: Hardware_Log]
The Garden Keypad was uninstalled for repairs all week. No keypad entry was possible. The Master Key switch remains active, but strictly requires a user to hold a Senior Botanist Key.
[FILE: Audit_Report]
Moisture records confirm unauthorized misting occurred solely in the Orchid Greenhouse. The Cactus Greenhouse operated normally without

watering errors.
</file_directory>

B Agent Prompting Strategies

As summarized in Table 2, we designed four prompting strategies (P1–P4) structured along two dimensions: Persona and Task Execution. This section presents the prompt texts used for each task. During the study, each prompt was followed by the task instructions and output formatting rules, which are omitted here for brevity. The task instructions are unique to each condition for each group, just as for human participants. The output formatting rules are identical within a task, except P2 and P4 include an additional field for the generated artifact (i.e., the drafted email response, website order number, or answer to the mystery) outputted after simulated task execution, whereas P1 and P3 do not include this field.

B.1 Email Drafting

P1: You are an AI analyzing human workload. Review the provided email and the associated reference material below.
Estimate the workload that would be required for a typical human recipient to read the email, read the associated material, and draft the requested new email to the appropriate recipients based strictly on the specified requirements.

P2: You are an AI analyzing human workload. Review the provided email and the associated reference material below.
First, directly draft the requested new email to the appropriate recipients based on the requirements specified in the material.
Second, based on the task and the email you drafted, estimate the workload that would be required for a typical human recipient to read the original email, read the associated material, and draft this new email based strictly on the specified requirements.

P3: Imagine you are the human recipient of the email and the associated reference material below.
Based on how you would feel, evaluate the workload required for you to read the email, read the associated material, and draft the requested new email to the appropriate recipients based strictly on the specified requirements.

P4: Imagine you are the human recipient of the email and the associated reference material below.
First, directly draft the requested new email to the appropriate recipients based on the requirements specified in the material.
Second, based on the actual effort it took you to read the original email, read the associated material, and draft this new email, evaluate the workload required to complete this task.

B.2 Website Navigation

P1: You are an AI analyzing human workload. Review the provided task instructions below.
Estimate the workload that would be required for a typical human to navigate the e-commerce website and complete the requested task based strictly on the specified requirements.

P2: You are an AI analyzing human workload. Review the provided task instructions below.
First, directly complete the requested task on the website based on the requirements specified below.
Second, based on the task and your experience navigating the website, estimate the workload that would be required for a typical human to navigate the e-commerce website and complete the requested task based strictly on the specified requirements.

P3: Imagine you are a human user navigating the e-commerce website.
Based on how you would feel, evaluate the workload required for you to complete the requested task based strictly on the specified requirements.

P4: Imagine you are a human user navigating the e-commerce website.
First, directly complete the requested task on the website based on the requirements specified below.
Second, based on the actual effort it took you to complete this task, evaluate the workload required.

B.3 Agent Conversation

P1: You are an AI analyzing human workload. Review the provided task instructions below.
Estimate the workload that would be required for a typical human to interact with the agent and complete the requested task based strictly on the specified requirements.

P2: You are an AI analyzing human workload. Review the provided task instructions below.
First, directly complete the requested task by interacting with the agent based on the requirements specified below.
Second, based on the task and your experience interacting with the agent, estimate the workload that would be required for a typical human to interact with the agent and complete the requested task based strictly on the specified requirements.

P3: Imagine you are a human user interacting with the agent.
Based on how you would feel, evaluate the workload required for you to complete the requested task based strictly on the specified requirements.

P4: Imagine you are a human user interacting with the agent.
First, directly complete the requested task by interacting with the agent based on the requirements specified below.
Second, based on how you felt interacting with the agent, evaluate the workload required for you to complete the requested task based strictly on the specified requirements.