Why Johnny Can't Use Agents: Industry Aspirations vs. User Realities with AI Agents

Pradyumna Shome
pradyumnashome@cmu.edu
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

Sashreek Krishnan
sashreek@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

Sauvik Das
sauvik@cmu.edu
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

Abstract

There is growing imprecision about what "AI agents" are, what they can do, and how effectively they can be used by their intended users. We pose two key research questions: (i) How does the tech industry conceive and market "AI agents"? (ii) What challenges do end-users face when attempting to use commercial AI agents for their advertised uses? We first performed a systematic review of marketed use cases for 102 commercial AI agents, finding that they fall into three umbrella categories: orchestration, creation, and insight. We then evaluated whether end-users could realize these marketed capabilities in practice: we conducted a usability assessment where $N = 31$ participants attempted representative tasks for each of these categories on two popular commercial AI agent tools: Operator and Manus. We found that users were generally impressed with these agents but faced significant usability challenges ranging from agent capabilities that were misaligned with user mental models to agents lacking the meta-cognitive abilities necessary for effective collaboration.

CCS Concepts

• **Human-centered computing** → **Empirical studies in interaction design**; **User studies**; **Usability testing**; **Natural language interfaces**.

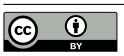
Keywords

AI Agents, Usability Studies, Interface Agents, LLMs

ACM Reference Format:
Pradyumna Shome, Sashreek Krishnan, and Sauvik Das. 2026. Why Johnny Can't Use Agents: Industry Aspirations vs. User Realities with AI Agents. In *ACM Conference on AI and Agentic Systems (CAIS '26), May 26–29, 2026, San Jose, CA, USA*, 25 pages.

1 Introduction

The AI and HCI communities have long been animated by visions of computing systems as intelligent partners for knowledge work: from Vannevar Bush's 1945 dream of the "Memex" to Horvitz's explorations of agent-mediated "mixed-initiative" interfaces [27] to Shneiderman and Maes' canonical debate on direct manipulation versus interface agents [40, 61]. Today, the dream has resurfaced under the banner of large language model (LLM) powered "AI agents". But are these agents, at long last, useful and usable?



CAIS '26, San Jose, CA, USA

ACM ISBN 979-8-4007-2415-2/2026/05
https://doi.org/10.1145/3786335.3813140

Answering that question is complicated because what an "AI agent" is, what it can do, and how those definitions shape user expectations is imprecise and rapidly evolving. Unlike chatbots that use LLMs to generate natural language responses to user prompts, prior work suggests that *operationally agentic* AI agents should have at least some of the following properties: (i) *delegation and initiative allocation*—deciding who between the human user and the AI decides what, when [27]; (ii) *acting in an environment* rather than only producing text—invoking tools, APIs, or GUIs on the user's behalf [37, 40, 61]; (iii) *multi-step execution* with a reason-act loop (e.g., interleaved reasoning and tool use [57, 71]); and (iv) *oversight, intervention, and recovery*—the need for users to supervise and correct ongoing agent activity [45, 59]. But, as with other technologies at the peak of their hype cycle, there are also many *marketed-as-agentic* products that self-label as "AI agents" that may not fully instantiate any of these properties. Because general end-users may not understand the difference, in this paper we consider both scopes as relevant.

To date, much of the conversation about agent performance — in benchmarks, demo videos, venture funding pitches, and media hype — has focused on easy-to-quantify ideals that overestimate true effectiveness [77]. To better design and evaluate "AI agents" that are useful and usable for end-users, we need a systematic understanding of how the tech industry conceives of "AI agents", how those framings shape user expectations, and where end-users' lived experience diverges from marketed aspirations. Against this backdrop, we ask:

RQ1: How does the tech industry conceive of and market "AI agents"? Industry framings matter because they define the menu of software choices that most users can access and set the expectations with which users approach these tools.

RQ2: What challenges do end-users face when using commercial AI agents for their advertised uses? We seek to understand the frictions that emerge when the aspirational capabilities these products imply collide with the realities of use.

To answer RQ1, we constructed a taxonomy of *marketed-as-agentic* products. We started by sourcing advertised uses for AI agents by exploring aggregator websites for AI Agent products, including the AI Agents Directory and Product Hunt, and supplemented our corpus with web searches to increase the diversity of products. We then analyzed themes across verbiage relating to agent functionalities and distilled a taxonomy of three broad advertised use-cases for AI agents: **Orchestration**, **Creation**, and **Insight**.

To answer RQ2, we ran a think-aloud usability study in which $N = 31$ participants attempted representative tasks in each of the

three categories using two popular *operationally agentic* platforms—Operator and Manus. We supplemented these think-aloud studies with semi-structured interviews. Although participants were generally successful at accomplishing the assigned tasks (which were specifically selected to be simple and doable), participants struggled with five critical usability barriers: (i) agent capabilities are misaligned with user mental models, (ii) agents presume trust without establishing credibility, (iii) agents fail to accommodate a diversity of collaboration styles, (iv) agents generate an overwhelming amount of communication overhead, and (v) agents lack the metacognitive abilities that enable productive collaboration.

In summary, this paper makes three core contributions. (1) We introduce a taxonomy of marketed AI agent capabilities, distilling how the tech industry conceives of and frames AI agents. (2) We present an empirical account of five critical usability barriers end-users face when attempting to use these tools on a representative set of tasks. Relative to prior human-AI interaction work, we contribute (a) the expectations-to-breakdowns chain for agentic systems, (b) a mapping of usability barriers to operational agentic properties, and (c) evaluation dimensions for delegation, oversight, and recovery in agentic systems. (3) We distill design and evaluation implications for agentic systems that are usable and effective for end-users.

2 Related Work

Our work is situated at the intersection of three areas of inquiry in the HCI and AI literature: the design and evolution of interface agents, evaluations of AI agents and LLMs, and empirical studies of human-AI interaction.

Design and Evolution of Interface Agents. *Direct manipulation* has long been the dominant interaction paradigm for GUI interfaces [60]. However, *interface agents* that act as intermediaries and invoke commands on users' behalf [40] have been touted as a complementary, and at times oppositional, paradigm. This tension — crystallized in the classic debate over direct manipulation versus interface agents [40, 61] and the broader tradition of mixed-initiative interfaces [27] — has motivated decades of work exploring agent behaviors, interaction styles, and control mechanisms [37, 38]. Researchers have also used Wizard-of-Oz methods to prototype agent experiences beyond the limits of contemporary AI [42], while commercial assistants from Microsoft Clippy to voice assistants have made the promise (and pitfalls) of agent-mediated interaction visible to end-users [9, 10, 19, 68].

Recent transformer-based LLMs (e.g., [15]) have renewed excitement for interface agents by enabling GUI agents that plan and execute multi-step actions in real software environments [28, 74]. Yet, many critical usability questions remain: e.g., what autonomy should agents take, how should tasks be delegated, and how should oversight mechanisms be designed to keep users in control [40, 59]? We contribute an empirical examination of commercially deployed LLM-powered AI agents with real users, identifying the usability barriers that persist in today's renewed instantiation of interface agents and that limit effective delegation, oversight, and collaboration.

Evaluations of LLMs and AI Agents. A broad ecosystem of benchmarks and evaluation frameworks now exists for evaluating AI agents: e.g., capability benchmarks [18, 30, 52, 53], evaluation environments for computer-based knowledge work [21, 50, 69], personalized evaluation [16], new evaluation frameworks [31, 70, 78], and agent cards [8, 62]. However, these evaluations share a common limitation: they narrowly focus on technical competence and elide dimensions that require human-interaction such as usability, interpretability, trust, and user satisfaction. Salaudeen et al. [55] observe that many AI evaluation frameworks suffer from notable validity gaps, and Wallach et al. [64] argue that measurement tasks in generative AI lack the scientific rigor needed for valid conclusions, proposing as redress a framework rooted in social science measurement practices. In short, evaluating only whether an agent *is capable of* completing a task is insufficient — we must also evaluate whether users can effectively direct, oversee, and collaborate with those agents to complete tasks *effectively*. We contribute an evaluation of commercial AI agents with real people, doing tasks representative of how those agents are marketed.

Empirical Studies of Human-AI Interaction. There is a rich tradition of prior empirical work noting usability breakdowns in human-AI interaction. This work emphasizes that effective human-AI teaming hinges on users' mental models, expectations, and trust calibration [32, 36, 47, 76]. Furthermore, prior work suggests that framing and communication shape adoption and reliance on AI systems [33, 34, 72], and that explanations should follow human norms (e.g., being selective and conversational) rather than overwhelming users with undifferentiated detail [44]. Complementary work examines collaboration dynamics and breakdowns: how people interpret competence signals in mixed-initative teams [75], how system updates can disrupt user experience by shifting mental models [11, 22], and how specification burdens and communication mismatches complicate collaboration with language-based systems [12, 35, 73]. Emerging studies of agentic LLM systems further highlight the importance of involving users in planning, avoiding over-delegation assumptions, and aligning procedures with user expectations [24, 29, 58, 65].

Building on this literature, we focus on a distinct empirical gap: how novice end-users experience *operationally agentic* AI agents that *act* on users behalf' by autonomously executing multi-step workflows in real software environments. In these settings, breakdowns can compound across steps and lead to real-world side effects (e.g., committing irreversible actions, producing flawed deliverables, or requiring costly recovery), shifting the core interaction problem from prompting a model to delegating, overseeing, and intervening in ongoing agent activity.

3 Taxonomy of AI Agents

We started by taxonomizing the litany of *marketed-as-agentic* products and services being advertised for immediate or near-term use under the "AI agent" banner (RQ1). This taxonomy, in turn, helps clarify what the technology industry considers an "AI agent" and serves as a *marketing-driven use-case typology*: it captures how products are framed for consumers and what expectations they set. We focus on industry-marketed products because these offerings more clearly define the menu of possibilities for broad consumer use of AI agents — while there is also much academic interest in the topic, the use-cases explored in academia may be less immediately relevant for end-consumers.

Methodology. We assembled a corpus of $N = 102$ *marketed-as-agentic* products by drawing on aggregator lists (e.g., AI Agent Directory [8], Product Hunt [26], Google Cloud use cases [54]) and keyword searches ("AI agents", "LLM agents"). We included products that were marketed or described as "AI agents" (or equivalent), were consumer-facing with concrete marketed use-cases, and targeted general end-users rather than only developers or domain specialists. We did *not* require that products fully instantiate the *operationally agentic* properties discussed prior; our goal for RQ1 is to capture how the "agent" label is applied in industry, so we coded what capabilities are *claimed* in marketing materials. Overlap and broadness across categories are expected because marketing typologies are not mutually exclusive. Two authors performed an inductive qualitative content analysis [56] of marketed capabilities, built a codebook through independent coding and consensus [43], and abstracted codes into higher-level categories following established taxonomy development practices [46]. "AI agents" is a rapidly evolving product category; we do not claim an exhaustive census, but a broad sample sufficient to characterize industry-envisioned use cases. Full search and exclusion criteria, analysis details, and the codebook are in Appendix A.1 and A.3.

Findings. We identified three broad categories of marketed use-cases: **Orchestration**, **Creation**, and **Insight** (Table 1). Categories are not mutually exclusive and, in fact, complex knowledge work often combines insight (problem decomposition), orchestration (executing workflows), and creation (formatting deliverables). The taxonomy thus encodes the *promises* and expectations that *marketed-as-agentic* products set for consumers. We selected one study task per category to test whether those promises are realizable in practice; the usability barriers we report below are, in that sense, expectation failures.

Orchestration agents act in GUIs on the user's behalf: they read interface state (e.g., via vision-language models), generate commands with an LLM, and execute them via a controller, producing actions with real-world side effects and raising delegation and oversight demands. Examples include Salesforce Agentforce [6], LiveX [4], and Expedia's Trip Matching Agent [2].

Creation agents produce structured artifacts (slide decks, websites, apps, documents) with an emphasis on formatting and presentation rather than analytical content; they differ from **Insight** in focusing on deliverable form. Examples include Lovable [5] (apps/websites) and Gamma [3] (slide decks).

Insight agents support research, synthesis, and recommendations by combining web search, APIs, and knowledge bases in multi-turn interactions. They address analysis and decision-making that users would otherwise perform manually. Examples that clearly instantiate the operational properties above include Perplexity Deep Research [7] and Deloitte Care Finder [39]. The full capability map with subcategories and counts, including boundary cases (e.g., recommender-style products marketed as agents), is in Appendix A.2.

4 User Study

We next conducted a usability study to assess whether and how end-users could use *operationally agentic* AI agents for orchestration, insight, and creation tasks in practice (RQ2).

4.1 Methodology

We conducted a think-aloud user study with $N = 31$ participants, with each 1 hour session comprising two AI agent tasks followed by a semi-structured exit interview. During the study, participants would attempt to accomplish each task using one of two commercial AI agents. These tasks were randomly selected from a set of three we had pre-defined, one for each category of our AI agent taxonomy (i.e., insight, orchestration, creation). The two commercial AI agents were Operator [1] and Manus [2].

4.2 The Participants

Participants were recruited via social media (X, LinkedIn) and Prolific. The sample spanned ages 18–65+, included 23 men and 8 women, and was largely composed of frequent users of generative AI tools (e.g., chatbots). However, participants were all novices to the multi-step, task-oriented systems marketed as "AI agents" that we study here. Full demographics are in Appendix B.8 (Table 6).

4.3 The Agents

We studied two commercially deployed *operationally agentic* products: Manus [41] and OpenAI Operator [49] (now integrated into ChatGPT Agent [48]). Both support multi-step task execution via a chat-based prompt interface and can use external tools (including computer use) to act in real software environments, with user takeover and oversight; they thus instantiate the operational agentic properties we discussed earlier (i.e., delegation an initiative allocation, acting in an environment on the user's behalf, multi-step execution, and providing mechanism for oversight, intervention, and recovery). We selected them because they were publicly accessible and covered all three taxonomy categories at the time of study. An extended description of both agents and their interfaces is in Appendix B.3.

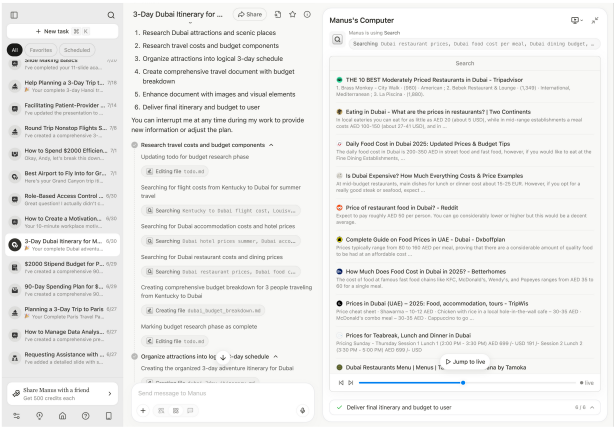


Figure 1: Manus, one of our AI Agents, working on our Holiday Planning task.

[1] https://openai.com/index/introducing-operator/
[2] https://manus.im/

Category	Definition	Examples
ORCHESTRATION	Act on behalf of users to manipulate other software (GUIs, automation).	Agentforce; Copilot
CREATION	Generate structured artifacts: slides, apps, docs, marketing materials.	Lovable; Gamma
INSIGHT	Support analysis, synthesis, recommendations, and information retrieval.	Perplexity Deep Research; Deloitte Care Finder

Table 1: Summary of three marketed use-case categories for AI agents (full capability map in Appendix A.2).

Manus. Manus is a general-purpose AI agent that can perform web searches and control a computer (browser, terminal, filesystem) from a persistent workspace UI (Figure 5). Its interface combines a chat thread, live computer view, and progress indicators, making ongoing work visible. When a user sends a prompt, Manus typically outlines steps and emits a detailed action log with links to intermediate states and outputs. Users can also intervene mid-task via "Take Over," manipulate the computer, and return control.

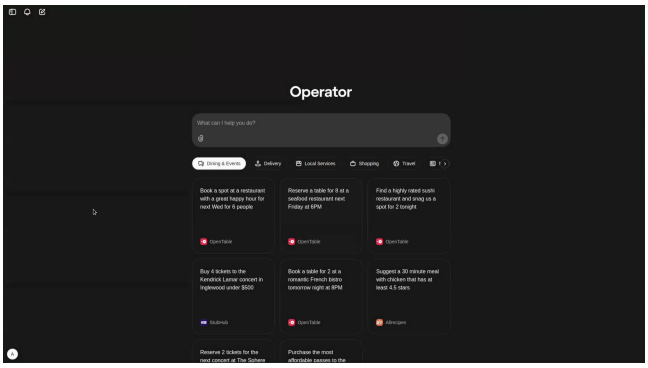


Figure 2: The initial screen on Operator, featuring a text box to enter prompts, and a grid of example tasks in various categories.

Operator. Operator offers comparable computer-using capabilities but with a more minimal, chat-forward interface (Figure 6) and non-streaming responses. It provides task idea cards on entry and then transitions to a standard chat view after the first prompt. When it invokes computer use, it embeds a computer preview in the chat thread (expandable to a larger layout) and supports user takeover for required steps (e.g., authentication) before returning control to the agent.

4.4 The Tasks

We defined three tasks for our participants to attempt during the study — one for each category of our AI agent use-case taxonomy. We designed this study without a non-agent baseline condition because our objective was to explore the novel interaction challenges that arise when end-users delegate tasks to autonomous agents, not to compare agent performance against manual task completion. Tasks were approximately evenly distributed across participants and agents (see Table 2). Specifying tasks is, of course, a very open-ended activity. To pick useful tasks for our study, we selected tasks that were: (i) **familiar** to non-specialists, (ii) **feasible** within the 20-minute think-aloud window, and (iii) **emotionally salient** enough to elicit meaningful preferences and evaluation. Full task-selection criteria and task screenshots are in Appendix B.4.

***ORCHESTRATION** Task: Holiday Planning.* In this task, participants used an AI Agent to produce an itinerary of a 3-day holiday to a location of their choice, including flight tickets, housing arrangements, sightseeing, and other activities of interest. This task involved orchestration (Section 3) in that it requires the agent to complete individual steps by navigating real UIs on an active computer. The task required the agent to navigate flight aggregators, elicit user preferences for activities, and synthesize results into a coherent schedule under complex constraints (budget, seasonal availability, personal preferences). Holiday planning is a familiar, emotionally salient task — the non-trivial costs and personal nature of travel raise the stakes of delegation.

***CREATION** Task: Slide Making.* In this task, participants used an AI agent to create (Section 3) a slide deck for a 10-minute presentation on a topic of their choice. We instructed users that the slides should be something they would be willing to present themselves, raising the stakes and requiring critical evaluation of outputs. Preparing presentations is common in knowledge work, and people have strong variation in both stylistic and content preferences, making this a task that naturally requires collaboration between users and agents.

***INSIGHT** Task: Professional / Personal Growth Stipend Budgeting.* In this task, participants used an AI agent to help them make the best use of USD 2,000 to advance their personal or professional development. This task required insight (Section 3) in that it the agent had to understand user goals, perform web searches, evaluate options against budget constraints, and synthesize results. Budgeting is a familiar, emotionally salient task — what each person considers a good use of disposable money is highly dependent on their values and goals, making a generic output unlikely to satisfy.

4.5 Procedure

We ran 60-minute Zoom sessions with $N = 31$ participants. Our sample size is consistent with qualitative study benchmarks: for example, Caine's analysis of CHI papers reported a modal sample size of 12 [17], and other highly cited studies on sample sizes for qualitative research found that saturation is commonly reached within 12-24 interviews [23, 25]. Each participant completed 2 assigned tasks using Manus or Operator, each of which included a 20 minute think-aloud session and a semi-structured exit interview. We collected

screen and audio recordings, System Usability Scale [14] responses, and interview transcripts. Tasks and agent tools were counterbalanced, with combinations approximately evenly distributed across participants (see Table 2). The interviewer did not answer questions about the agent interface except in limited circumstances where the tool appeared to be broken. Our recruitment flow, compensation, assignment details, and protocol are in Appendix B.2 and our interview script is in Appendix B.5. The study was approved by an institutional review board (IRB).

Consistent with the think-aloud protocol, the interviewer periodically asked participants to articulate what they thought was happening, what they expected to happen next, and how they evaluated the agent's progress. These organic "sync" moments became especially important when progress halted due to apparent platform or tool failures: the interviewer probed participants' interpretations of the blockage and how they were thinking about recovery, but did not provide task instructions. For example, some participants treated failures as technical issues on their own end, some evaluated the agent favorably despite the blockage, and some attempted to pivot to tools or workflows with fewer login requirements. We did not code sessions differently by intervention status; instead, these moments were analyzed as naturally occurring windows into participants' mental models of agent failure modes, intervention possibilities, and recovery strategies.

To analyze the data we collected, we conducted a reflexive thematic analysis [13] to understand what barriers and usability challenges end-users faced while attempting the tasks we had assigned them to complete with the use of an AI agent (RQ2). The first and second authors independently coded three interviews using open inductive coding [56], generating an initial set of codes grounded in concrete, low-level challenges participants encountered. They met weekly and discussed the codes for multiple interviews, merging and de-duplicating codes when realizing they were either the same phenomenon or variants that were different experiences altogether. Through this iterative process, the authors abstracted recurring patterns into higher-level themes that explained general experiences participants went through as they navigated agent software. Consistent with best practices for qualitative research in the HCI community [43], agreement was reached through discussion and consensus rather than inter-rater reliability metrics. Our codebook can be viewed in Appendix B.9.

4.6 Findings

Participants were generally able to complete both assigned tasks with a reasonable degree of success — though we specifically selected tasks that we verified were simple and doable prior to the study. Nevertheless, we identified a number of critical usability barriers that hindered participants' ability to make optimal use of the AI agents we tested. We will start by discussing where the agents succeeded, general impressions towards the agents, and then discuss the usability barriers we identified.

4.6.1 Agent successes. Figure 3 and Table 2 report SUS scores [14] by task and agent. We compare agents within each task using Welch's two-sample t-test, which does not require equal variances between groups. Seven participants completed the same task with both agents, partially violating independence; we treat these tests as descriptive context rather than strict inference. Manus rated significantly lower than Operator on Holiday Planning (−10.8 points, $p = .043$) and significantly higher on Slide Making (+20.8 points, $p = .012$); Stipend Budgeting was not distinguishable (−12.9 points, $p = .184$). With $n = 9$–13 per cell, the study is underpowered to detect modest effects, and contrasts are unadjusted for family-wise error.

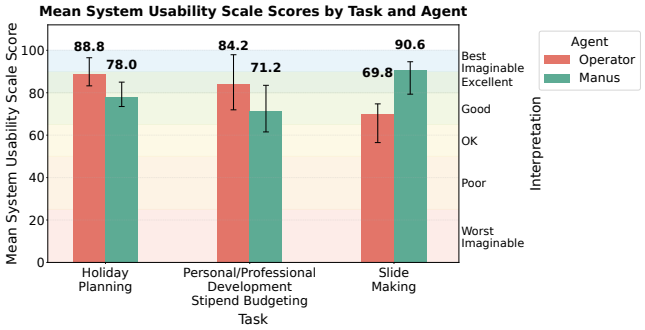


Figure 3: Mean System Usability Scale (SUS) [14] scores by task and agent. Error bars show 95% bootstrap confidence intervals (10,000 resamples); per-cell sample sizes are $n = 9$–13. Per-task standard deviations and Welch's t-test results are reported in Table 2.

Participants were charitable toward both agents even when tasks failed: *"This is worth it in every sense of the word"* (P23); most praised the agents' detail (*"very comprehensive and thorough"*, P26). None had used AI agents before, so this may partly reflect novelty. The favorable SUS scores may appear in tension with the barriers below; we find both genuine. Participants recognized real value, but the barriers—around mental models, trust, collaboration, communication, and metacognition—are frictions that would compound with sustained use, and would likely be magnified in higher-stakes, less structured real-world settings beyond our deliberately simple tasks.

Participants also reported a consistent tradeoff: Manus produced richer first-class outputs (especially for slides) but introduced higher communication overhead and lower editability of intermediate artifacts, while Operator was more concise but stalled more often during computer-use workflows. Both exhibited the barrier classes below.

4.6.2 Barrier 1: Agent behaviors misalign with user mental models. Many participants were unsure what agents could do, how much detail prompts required, and what role they should play during execution. This uncertainty made initial prompting feel high-stakes and encouraged trial-and-error interaction rather than confident delegation. As one participant noted, *"You've got to sit there and make sure that your initial prompt is perfect . . . You ask exactly what you're looking for."* (P26).

Users also misunderstood core interaction mechanics (e.g., when and why takeover was needed), suggesting that current interfaces under-specify agent affordances and execution boundaries. Overall, users built mental models reactively from outcomes rather than proactively from clear system cues, contributing to downstream trust and control breakdowns.

Table 2: Per-task SUS scores by agent. The Manus−Operator estimate, 95% CI, and p value are from Welch's two-sample t-test (which does not assume equal variances between agents). Cells show mean (M) and standard deviation (SD).

Task	Operator		Manus		Manus−Operator	
	n	M (SD)	n	M (SD)	estimate [95% CI]	p
Holiday Planning	10	88.8 (8.6)	10	78.0 (12.8)	-10.8 [-21.1, -0.4]	0.043
Stipend Budgeting	9	84.2 (18.2)	10	71.2 (22.4)	-12.9 [-32.6, +6.8]	0.184
Slide Making	10	69.8 (20.9)	13	90.6 (8.6)	+20.8 [+5.4, +36.2]	0.012

Users were apprehensive and uncertain about prompting agents. End-users generally did not know what to expect when sending an agent a prompt, leading to a behavior that we term *"prompt gambling"*, i.e., users expressing uncertainty and hesitation with prompting agents owing to concerns that it would generate unnecessary and excessive outcomes.

This uncertainty led to a range of different prompting strategies. Some users believed they needed to decompose a task into individual steps, and have an agent perform each step one at a time. These users would craft an initial prompt with only the first step of the task — but the agents they used would assume that this first step was the full task, leaving users feeling trapped. Other users would provide a highly detailed initial prompt to try and prevent agents from going astray. The ultimate effect was that many participants felt that the initial prompt they fed to the agent was high-stakes and essential to get right.

In other cases, users were pleasantly surprised by how much the agent could do with little specification: *"I thought I needed to give it source material for the slides, but it did all the research and organized it in slides that would've taken me weeks to put together."* (P26) These participants gradually updated their model of the agent from a deterministic tool requiring structured hand-holding to something more *"as if it were an administrative assistant"* (P9). This confusion — while not stifling — still indicates that agents' *affordances* were mysterious to users; their conceptions arose only after trial-and-error, and even then remained task-specific and incomplete.

Users had trouble understanding what agents did, and why. Multiple users were confused about what "Computer use" was and how it worked. Attempting to make slides, Operator repeatedly asked users to use *Take Over* to sign in to Google Slides, which felt unexpected and undesirable. One user expected the slide outline and designs before being asked to log in, and expected slides to be prepared within Operator, unaware that slide-making was not native functionality. They also misunderstood *Take Over*, believing it was a request for them to make the slides rather than sign in and return control to the agent. P19 stated: *"it's kind of like I'm giving you a job, and you're throwing the job back at me...You have not even given me anything yet, and you want me to do some stuff for you already."*

A more technically knowledgeable participant questioned the need for Computer Use altogether, preferring web APIs for research tasks over browser control. Computer Use was perceived as far slower than participants' own manual searches.

Some participants developed folk models of how interim outputs affected speed. For example, noting the volume of progress reports, P31 wondered whether hiding workflow logs would make the agent faster.

In short, AI agents are capable enough that simple recall-based chat leaves too much to users' imagination. This creates confusion about what agents can do, what they need from users, how they work, and what role users play in task completion.

4.6.3 Barrier 2: Agents presume trust in sensitive contexts. Several participants (e.g., P22, P26, P30) expressed distrust in working with AI agents to perform either specific parts of a task or the task as a whole. Trust issues with each of the tools spanned concerns around hallucinations, password management, low expectations of agents' capabilities, and disinclination to hand over too much autonomy over tasks users preferred to handle personally.

Not all users wanted to trust the agent with credentials to their accounts: *"I don't particularly want to give it my Google account information"* (P30).

Some users also expressed apprehension about entrusting the agent to do tasks on their behalf, such as planning a trip: *"I would plan manually first, then compare… I still don't trust it completely"* (P26). P22 further explained how these trust issues were compounded by the fact that the agent did not ask them questions about their preferences or constraints: *"it didn't ask me. you know. Do I want one bed or 2? Do I want a a room that looks out on the pool? Or do I want one that looks out on the plaza?"* This example highlights how trust-building is discursive; for tasks that require capturing personal preferences, overly eager agents that run off and do work without confirming user intent can reduce trust.

Several participants also expressed concern about the veracity of the information agents provided, especially when the values conflicted with *a priori* expectations or seemed too good to be true.

4.6.4 Barrier 3: Agents fail to accommodate diverse collaboration styles. Many participants were uncertain how to "collaborate" with agents; several requested pause/stop or finer-grained control (e.g., P23, P26). Users were unsure what they could delegate, how to exert control, and how to recover from errors.

Users lack effective control mechanisms for steering agent behavior during task execution. When agents failed or otherwise produced poor initial outputs, several users believed that it would be faster and more effective to restart from scratch as opposed to asking the agent to iterate on the provided deliverable. They did not believe the agent could effectively make fine-grained edits. P31 captured this feeling when expressing confusion as to why the agent was able to easily create new slides, but had trouble following more detailed instructions to iterate on these slides:

> *"I remember the original one. I mean, it had, like maybe one more block of text on the left hand side, and just for it to remove, you know, that seemed like it took longer than you know, creating all 7 other slides. So that was kind of I don't know why it got so caught up on that."* –P31

Users also expressed wanting better real-time control mechanisms to supervise the agent and ensure it stayed on track. Participant 26 expressed this tritely, *"I'd like a pause button if it's going off the rails."* Similarly, P23 expressed wanting a "stop button" to exert more control over how long the agent worked per prompt.

Others missed or were hesitant to oblige Manus' message inviting additional prompts even while it was responding to a previous prompt. These participants questioned if sending a new prompt would derail the agent or otherwise cause it to lose progress.

Users vary in the level of proactivity they would like out of agents. We observed strong variation in how involved individual users wanted to be with crafting the final output delivered by the agent. For example, for the slide making task, some users were interested in being deeply involved with the design of each slide, whereas others were content with the agent proposing its own outline and building a slide deck around it, with little involvement from them. P9, for example, mentioned "*my typical way of doing this kind of thing is to build all this as independent units on my own.*"

More generally, we found that some users viewed agents as "thought partners" while others viewed agents as execution tools. P16 was emblematic of the former camp: *"I like to do some of the baseline stuff myself… use AI more as… confirmation...making sure that what I'm taking away from something is maybe correct…and then kind of just verifying that I didn't miss anything"* Agents generally did not appear to account for this variation, and tended towards being over-eager execution tools — i.e., assuming users wanted minimal overall involvement.

Agents made incorrect assumptions about users, leading to miscalibrated outputs. Agents also appeared to sometimes make incorrect assumptions about users which could lead to deliverables that were misaligned with user expectations.

For example, for the Slide Making task, P30 requested that the agent make a slide deck about bats. The final slide deck the agent produced was minimal — just a stream of bat images with little text to scaffold the rationale for each image. This lack of scaffolding left P30 feeling frustrated, and feeling like the agent made a faulty assumption: *"It thinks I'm an expert at bats, but I don't know much, and I certainly can't present a talk with slides that just have photos"*. P30 further expressed wanting to explicitly specify roles and expectations in the collaboration: *"Hey, I'm really good at doing this. These are things that you might need. I might need your help on …."*

Similarly, for the Trip Planning task, P13 found many of the travel suggestions the agent made to be fairly basic, and only necessary for people taking a flight or visiting a country for the first time, an assumption that they believe the agent had made incorrectly.

Users struggle with recovering from agent errors due to opaque failure modes. Agents occasionally made errors as they attempted to complete tasks — indeed, 14 out of 31 users encountered at least one operational error during the study, such as being blocked by bot detection workflows (4). Many users struggled with recovering from these errors.

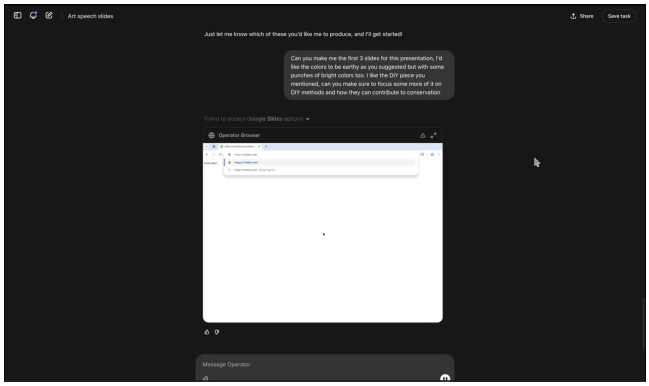


Figure 4: OpenAI Operator, working on the slide making task. Frequently, it was blocked while connecting to online presentation software, potentially due to anti-bot protection.

In particular, for Operator's Computer Use, many users did not even register errors when they did occur. Operator generally froze in these circumstances with no error message or troubleshooting instructions, leaving users uncertain: *"I wasn't quite sure why it was failing."* (P7) As a result, many users attributed Operator's lack of progress to arbitrary causes such as websites being incompatible with Operator, security settings on the user's computer, or because of their own errors. P30, for example, said: *"I didn't realize it was failing... that could just be my own ignorance of the platform it was using."*

These complete blockages prompted interviewer probes about participants' interpretations of the failure and recovery options, but not direct task guidance.

4.6.5 *Barrier 4: Agents create communication overhead for users.*

Participants across the spectrum of progress-visibility preferences (e.g., P16, P21–P24, P31) reported difficulty parsing agent output or articulating intent.

Users have varying expectations around communication of agent progress. We observed a full spectrum of expectations and preferences around how agents should communicate task progress. Some (e.g., P16, P24, P31) preferred final-output-only interaction, while others (e.g., P21, P22, P23) wanted progress updates for oversight and confidence. For example, P16 stated: *"It's just too much to take...Oh, my God! It threw out so much stuff...it's almost an overwhelming amount of information"*. Across both preferences, users reported that the logs were hard to parse because key milestones were buried in dense, uniformly formatted output.

Prompting agents is a technical specification recall task that is cognitively demanding. Many participants expressed that prompting agents was difficult, because they were themselves uncertain about how much they needed to communicate with the agent and how best to communicate the web of interconnected preferences they had in their minds.

> *something that might be obvious to a person, is maybe not always obvious to the AI is needed information. So*

> *it doesn't think to ask, and you don't think to tell it and then realize there's a gap.* –P16

For instance, for the Slide Making task, users expressed a desire to directly modify slides agents had initially improved, rather than asking the agent to iterate on those slides. These users believed it would be too difficult to articulate, in a way agents might be able to understand, their criteria for what they wanted to improve. They simply wanted "better images" or "less text-heavy" slides.

Similarly, on the Trip Planning task, many people wanted to pick flights themselves because of complex, conditionally intertwined criteria that were difficult to express as structured instructions (e.g., frequent flyer miles, number of stops, time of day). The perceived need to do so created intense metacognitive demands.

4.6.6 Barrier 5: Agents lack metacognitive self-assessment. Several users (e.g., P16, P19, P20, P29) observed that agents struggled to assess their own progress, limitations, and output quality in terms users could understand.

Agents don't know what they don't know and can't do. When agents encountered difficulties or made errors, several users observed that agents would fail to recognize these errors and get stuck in a repetitive loop. For example, for the Slide Making task, Operator didn't seem to recognize its inability to control some slide-making websites, which remained frozen for extended periods of time. P16 noted: *"It doesn't have access to certain things, so it couldn't do it. And it just was kind of circling."* P16 then lamented that if they had taken more initiative in troubleshooting, they would have been able to compensate for the agent's failure to recognize its limitations, highlighting meta-cognitive gaps that users must sometimes cover.

Users expressed a desire for agents to express greater reflective abilities. For example, for the Personal Growth Insight task, P20 noticed that Manus hallucinated information about courses that didn't exist at all, even when synthesizing information from Web searches. In this situation, the user (P20) preferred a different response: *"it's seeking to provide an answer rather than to say, I don't know, or to link someone where they should go to find information."* Similarly, for the Slide Making task, P19 expected the agent to take more active role in seeking critique to improve the design: *"It probably covered like 70 to 80% of what I was expecting. I was waiting for it to ask me for some more information."*

Agents use tools with which users are unfamiliar, precluding effective oversight. We observed that agents sometimes used tools with which users were unfamiliar, inhibiting oversight and control. We often saw this in the Slide Making task: agents unexpectedly used programming tools to create slides. P29, for example, was surprised by Manus' decision to use HTML: *"I wasn't expecting it to make the slides all in HTML!"* When Operator did something similar, P25 felt left out of the process and pessimistic about success.

5 Discussion

Our results show a gap between industry aspirations for AI agents and how novice users can reliably use them in practice. In taxonomizing the capabilities of 102 products marketed as "AI agents" and conducting a think-aloud usability study with 31 participants assessing these capabilities across two operationally agentic systems, we found five critical usability barriers: (1) mental-model misalignment, (2) premature trust assumptions, (3) collaboration-style mismatch, (4) communication overload, and (5) weak metacognitive behavior. Table 3 maps each barrier to an implicated agentic property.

Usability barrier	Implicated agentic property
Mental-model misalignment	Hidden plans/boundaries Long-horizon autonomy
Premature trust assumptions	Credential delegation Real-world actions
Collaboration-style mismatch	Mixed-initiative planning Intervention points
Communication overload	Step logs State over time
Weak metacognitive behavior	Stall/loop detection Handoff to users

Table 3: Mapping of the five empirically identified usability barriers to the agentic properties (from prior HCI and ML literatures) that cause or amplify them.

5.1 Relationship to LLM Usability Barriers

We did not include a chatbot baseline, so our claim is not that these barriers are absent from chatbots. Rather, our results show how familiar LLM usability problems are reconfigured when language-based interaction becomes a mechanism for delegating action. In chatbots, users primarily evaluate and revise model outputs. In operationally agentic systems, users delegate work that unfolds over time through plans, tool calls, external interfaces, credentials, and intermediate artifacts. The usability surface expands — prompting remains a challenge, but in addition there are questions around if users can appropriately bound an agent's autonomy, inspect its progress, intervene at the right moment, and recover from failures.

Prior work shows that non-experts prompt LLMs opportunistically and struggle to form stable mental models of language-based systems [12, 35, 73]; generative-AI design work similarly emphasizes the difficulty of specifying, prototyping, and controlling variable model outputs [63, 66]. Our findings suggest that these known challenges become more consequential when ambiguous instructions are enacted rather than merely answered. In a chatbot, a poor prompt may cost another turn. In an agent, the same ambiguity can launch a time- and resource-consuming multi-step execution process that invokes external tools and/or produces irreversible side effects before the user can diagnose the mismatch. Thus, Barrier 1 and Barrier 2 are not only about users knowing what to ask or whether to trust an answer; they are about whether users understand what the agent will do after being asked, where it will act, what assumptions it will make, and what risks it may incur on the user's behalf.

The remaining barriers similarly reflect the shift from response generation to enacted work. *Task diversity* becomes more challenging because agents may manipulate heterogeneous external substrates such as browsers, files, accounts, slide editors, calendars, and software tools in ways that users may not be able to effectively oversee (e.g., when agents create slides programmatically).

Metacognition becomes central because agents can stall, loop, or use unfamiliar tools over many steps; users need to know when the agent is struggling, what it is doing, and when they should intervene. Finally, *collaboration* differs from turn-based chat because agents may take many actions per instruction while users interleave corrections, clarifications, and new goals.

Some manifestations of these barriers may diminish as users gain experience and models improve, but delegation, oversight, intervention, and recovery are structural requirements of agentic systems that will not be addressed simply by expecting more capable users and models in the future. We next translate these barriers into concrete implications for designing and evaluating agentic systems for general end-users.

5.2 Implications for *Designing* AI Agents

Our findings yield six high-priority design implications for end-user AI agents, each directly grounded in the five usability barriers we observed.

5.2.1 "Know your user": Calibrate to user preferences. Users vary in desired proactivity, communication verbosity, and trust requirements (Barriers 2–4). Agents should maintain lightweight user profiles (e.g., preferred collaboration style, verbosity, trusted sources) and ask task-specific preference questions at session start, including acceptable autonomy and when to defer.

5.2.2 "Know yourself": Improve self-assessment. Participants encountered agents that missed blocked progress, presented imperfect drafts as final, or used unexpected tools (Barrier 5). Agents should expose confidence, detect repeated failures, and hand off control when stalled or in need of user input. They should also distinguish completed, blocked, and uncertain states so they neither stop too early nor gratuitously revise work that is already acceptable.

5.2.3 "Be adaptable": Adapt to task and end-user. Users wanted different levels and formats of progress visibility, but interfaces presented uniform, high-volume logs (Barrier 4). Agents should adapt presentation by task type and user preference, using progressive disclosure to emphasize milestones while collapsing routine actions.

5.2.4 "Measure twice, cut once": Support planning and execution control. Users lacked control over plan quality after the first prompt and often felt trapped in poor-quality execution paths (Barriers 1 and 3). Agents should support pre-execution plan review/edit and in-execution controls such as pause, stop, rollback, and checkpointing. This is especially important when agent actions consume time, credentials, or real-world opportunities before inspection.

5.2.5 "Show, don't tell": Support non-textual input. Purely free-text controls imposed a high recall burden and made nuanced intent difficult to express (Barrier 4, leading to Barrier 1). Agents should combine instruction templates, recognition-based controls, and demonstration-based input [20] so users can specify high-impact choices without relying on prompt-engineering know-how, especially for visual edits and conditional constraints.

5.2.6 "Next time's the charm!": Support precise iteration. Participants often trusted agents for first drafts, but not for precise iterations, especially on visual and structural deliverables (Barrier 3). Agents should support artifact-native iteration: inline comments, visual diffs, section reordering, version history, and selective revert. Intermediate outputs should remain inspectable and editable.

5.3 Implications for *Evaluating* AI Agents

Based on our findings, we propose that agentic systems be assessed along at least the following axes to ensure usability: (i) *intervention frequency*—how often users must pause, correct, or redirect the agent (Barriers 1, 3); (ii) *time-to-recovery*—latency and effort to recover from agent errors or stalls (Barriers 3, 5); (iii) *stall/loop rate*—incidents where the agent fails to recognize blocked progress or repeats unproductive actions (Barrier 5); (iv) *plan-edit friction*—ease of reviewing or revising the agent's plan before or during execution (Barriers 1, 3); (v) *credential and handoff points*—clarity and safety of sensitive action delegation (Barrier 2); (vi) *iteration utility*—the ability to make specific, targeted improvements on an interim output (Barrier 3); and, (vii) *rollback/checkpoint use*—availability and usability of undo or checkpointing (Barrier 3). These dimensions complement existing capability-oriented evaluations and support the design of more reliable agentic systems.

5.4 Limitations

As with any empirical study of an emerging technology, our study has limitations related to the tasks we developed, the composition of our participant pool, and the dynamism of the product space. Our user study features three sample tasks, which, while representative of the use cases from our taxonomy, cannot cover the breadth and depth of tasks that AI agents purport to support and that can be found in modern digital knowledge work. Next, our group of participants came exclusively from the US sociocultural context, were relatively young, able, and comprised frequent users of generative AI (but not agentic AI) tools. Finally, we assessed desktop versions of just two commercial AI agents from a large and rapidly expanding space of agent software, whose models, capabilities, interaction paradigms, and resulting usability challenges may differ.

6 Conclusion

In this study, we created a taxonomy of industry-marketed use cases of "AI Agents" and conducted a usability study of two commercial agents, Operator and Manus, to examine how marketed capabilities translate in practice. We identified three broad use cases—orchestration, creation, and insight—and observed five recurring usability barriers: (i) mental-model misalignment, (ii) premature trust assumptions, (iii) collaboration style mismatches, (iv) communication overload, and (v) weak metacognitive behavior. We translate these into implications for designing and evaluating end-user AI agents: calibrate to user preferences, support self-assessment, adapt to task and user, broaden intent input beyond free text, and support precise iteration. Together, these identify central challenges to realizing the longstanding aspirational vision of AI agents.

Acknowledgments

This work was funded, in part, by the National Science Foundation under award #2316768. The authors thank Isadora Krsek, Hao-Ping (Hank) Lee, Yuxuan Li, and Kyzyl Monteiro for insightful discussions, constructive feedback, and helpful suggestions.

References

[1] Anysphere 2025. *Cursor - The AI Code Editor*. Anysphere. https://cursor.com/ Accessed: 2025-09-01.

[2] 2025. Expedia Trip Matching AI Agent. https://www.marketingdive.com/news/expedia-brings-generative-ai-trip-planning-tool-to-instagram/748233/. Accessed: 2025-09-12.

[3] 2025. Gamma Presentation Builder. https://gamma.app/. Accessed: 2025-09-12.

[4] 2025. LiveX AI Agent. https://livex.ai/. Accessed: 2025-09-12.

[5] 2025. *Lovable*. https://lovable.dev/ Accessed: 2025-09-01.

[6] 2025. Salesforce Agentforce. https://www.salesforce.com/agentforce/. Accessed: 2025-09-12.

[7] Perplexity AI. 2024. *Introducing Perplexity Deep Research*. https://www.perplexity.ai/hub/blog/introducing-perplexity-deep-research Accessed: 2025-09-02.

[8] AI Agents Directory. 2025. AI Agent Marketplace & Directory | Find Top AI Agents & AI Agent Solutions. https://aiagentsdirectory.com/ Accessed: 2025-09-01.

[9] Amazon. n.d.. Amazon Alexa. https://alexa.amazon.com/about. Accessed: 2026-02-03.

[10] Apple. n.d.. Siri. https://www.apple.com/siri/. Accessed: 2026-02-03.

[11] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S. Lasecki, Daniel S. Weld, and Eric Horvitz. 2019. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7, 1 (Oct. 2019), 2–11. doi:10.1609/hcomp.v7i1.5285

[12] Gagan Bansal, Jennifer Wortman Vaughan, Saleema Amershi, Eric Horvitz, Adam Fourney, Hussein Mozannar, Victor Dibia, and Daniel S. Weld. 2024. Challenges in Human-Agent Communication. arXiv:2412.10380 [cs.HC] https://arxiv.org/abs/2412.10380

[13] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. doi:10.1191/1478088706qp063oa

[14] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[15] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL] https://arxiv.org/abs/2005.14165

[16] Hongru Cai, Yongqi Li, Wenjie Wang, Fengbin Zhu, Xiaoyu Shen, Wenjie Li, and Tat-Seng Chua. 2025. Large Language Models Empowered Personalized Web Agents. In *Proceedings of the ACM on Web Conference 2025* (Sydney NSW, Australia) *(WWW '25)*. Association for Computing Machinery, New York, NY, USA, 198–215. doi:10.1145/3696410.3714842

[17] Kelly Caine. 2016. Local standards for sample size at CHI. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 981–992.

[18] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*.

[19] Justin Cranshaw, Emad Elwany, Todd Newman, Rafal Kocielnik, Bowen Yu, Sandeep Soni, Jaime Teevan, and Andrés Monroy-Hernández. 2017. Calendar.help: Designing a Workflow-Based Scheduling Agent with Humans in the Loop. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, Denver, CO, USA, 2382–2393. doi:10.1145/3025453.3025780

[20] Allen Cypher, Daniel C. Halbert, David Kurlander, Henry Lieberman, David Maulsby, Brad A. Myers, and Alan Turransky (Eds.). 1993. *Watch what I do: programming by demonstration*. MIT Press, Cambridge, MA, USA.

[21] Alexandre Drouin, Maxime Gasse, Massimo Caccia, Issam H. Laradji, Manuel Del Verme, Tom Marty, Léo Boisvert, Megh Thakkar, Quentin Cappart, David Vazquez, Nicolas Chapados, and Alexandre Lacoste. 2024. WorkArena: How Capable Are Web Agents at Solving Common Knowledge Work Tasks? arXiv:2403.07718 [cs.LG] https://arxiv.org/abs/2403.07718

[22] Dylan Freedman. 2025. The Day ChatGPT Went Cold. *The New York Times* (2025). https://www.nytimes.com/2025/08/19/business/chatgpt-gpt-5-backlash-openai.html Accessed: 2025-09-04.

[23] Greg Guest, Arwen Bunce, and Laura Johnson. 2006. How many interviews are enough? An experiment with data saturation and variability. *Field methods* 18, 1 (2006), 59–82.

[24] Gaole He, Gianluca Demartini, and Ujwal Gadiraju. 2025. Plan-Then-Execute: An Empirical Study of User Trust and Team Performance When Using LLM Agents As A Daily Assistant. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM. doi:10.1145/3706598.3713218

[25] Monique M Hennink, Bonnie N Kaiser, and Vincent C Marconi. 2017. Code saturation versus meaning saturation: how many interviews are enough? *Qualitative health research* 27, 4 (2017), 591–608.

[26] Ryan Hoover. 2013. Product Hunt. https://www.producthunt.com/. Accessed: 2025-09-01.

[27] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces *(CHI '99)*. Association for Computing Machinery, New York, NY, USA, 159–166. doi:10.1145/302979.303030

[28] Xueyu Hu, Tao Xiong, Biao Yi, Zishu Wei, Ruixuan Xiao, Yurun Chen, Jiasheng Ye, Meiling Tao, Xiangxin Zhou, Ziyu Zhao, Yuhuai Li, Shengze Xu, Shenzhi Wang, Xinchen Xu, Shuofei Qiao, Zhaokai Wang, Kun Kuang, Tieyong Zeng, Liang Wang, Jiwei Li, Yuchen Eleanor Jiang, Wangchunshu Zhou, Guoyin Wang, Keting Yin, Zhou Zhao, Hongxia Yang, Fan Wu, Shengyu Zhang, and Fei Wu. 2025. OS Agents: A Survey on MLLM-based Agents for General Computing Devices Use. arXiv:2508.04482 [cs.AI] https://arxiv.org/abs/2508.04482

[29] Song Jiang, Da Ju, Andrew Cohen, Sasha Mitts, Aaron Foss, Justine T Kao, Xian Li, and Yuandong Tian. 2024. Towards Full Delegation: Designing Ideal Agentic Behaviors for Travel Planning. *arXiv preprint arXiv:2411.13904* (2024).

[30] Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. 2024. SWE-bench: Can Language Models Resolve Real-World GitHub Issues? arXiv:2310.06770 [cs.CL] https://arxiv.org/abs/2310.06770

[31] Sayash Kapoor, Benedikt Stroebl, Zachary S. Siegel, Nitya Nadgir, and Arvind Narayanan. 2024. AI Agents That Matter. arXiv:2407.01502 [cs.LG] https://arxiv.org/abs/2407.01502

[32] Harmanpreet Kaur, Alex C. Williams, and Walter S. Lasecki. 2019. Building Shared Mental Models between Humans and AI for Effective Collaboration. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland). Association for Computing Machinery. doi:10.1145/3290605.3300643

[33] Pranav Khadpe, Ranjay Krishna, Li Fei-Fei, Jeffrey T. Hancock, and Michael S. Bernstein. 2020. Conceptual Metaphors Impact Perceptions of Human-AI Collaboration. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 163 (Oct. 2020), 26 pages. doi:10.1145/3415234

[34] Sunnie S. Y. Kim, Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. "I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust. In *Proceedings of the 7th ACM Conference on Fairness, Accountability, and Transparency (FAccT 2024)*.

[35] Andrew J. Ko, Brad A. Myers, and Htet Htet Aung. 2004. Six Learning Barriers in End-User Programming Systems. In *Proceedings of the 2004 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. 199–206. doi:10.1109/VLHCC.2004.47

[36] Hui Li, Jiasheng Zhang, and Kun Huang. 2025. Meta-Analyzing the Trust-Performance Link in Collaboration: Moderating Effects of Conceptual and Contextual Factors. *Public Performance & Management Review* 48, 1 (2025), 1–34. arXiv:https://doi.org/10.1080/15309576.2024.2405839 doi:10.1080/15309576.2024.2405839

[37] Henry Lieberman. 1997. Autonomous Interface Agents. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '97)*. ACM, Atlanta, Georgia, USA, 67–74. doi:10.1145/258549.258592

[38] Henry Lieberman and Ted Selker. 1999. Agents for the User Interface. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. Media Laboratory, Massachusetts Institute of Technology. https://web.media.mit.edu/~lieber/Publications/Agents_for_UI.pdf

[39] Deloitte Consulting LLP. 2025. AWS Marketplace: Care Finder Agent. https://aws.amazon.com/marketplace/pp/prodview-mau5avpo5xog6. https://aws.amazon.com/marketplace/pp/prodview-mau5avpo5xog6 Product listing for Care Finder Agent, an AI-powered healthcare navigation solution by Deloitte..

[40] Pattie Maes. 1994. Agents that reduce work and information overload. *Commun. ACM* 37, 7 (July 1994), 30–40. doi:10.1145/176789.176792

[41] Manus AI. 2025. Manus: General AI agent that bridges mind and action. https://manus.im/?index=1. Accessed: 2025-09-02.

[42] David Maulsby, Saul Greenberg, and Richard Mander. 1993. Prototyping an intelligent agent through Wizard of Oz. In *Proceedings of the INTERACT'93 and CHI'93 Conference on Human Factors in Computing Systems*. 277–284.

[43] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. In *Proceedings of the ACM on Human-Computer Interaction (CSCW, Vol. 3)*. ACM, Article 72. doi:10.1145/3359256

[44] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38. doi:10.1016/j.artint.2018.07.007

[45] Hussein Mozannar, Gagan Bansal, Cheng Tan, Adam Fourney, Victor Dibia, Jingya Chen, Jack Gerrits, Tyler Payne, Matheus Kunzler Maldaner, Madeleine Grunde-McLaughlin, Eric Zhu, Griffin Bassman, Jacob Alber, Peter Chang, Ricky Loynd, Friederike Niedtner, Ece Kamar, Maya Murad, Rafah Hosn, and Saleema Amershi. 2025. Magentic-UI: Towards Human-in-the-loop Agentic Systems. *arXiv preprint arXiv:2507.22358* (2025).

[46] Robert C. Nickerson, Upkar Varshney, and Jan Muntermann. 2013. A method for taxonomy development and its application in information systems. *European Journal of Information Systems* 22, 3 (2013), 336–359. doi:10.1057/ejis.2012.26

[47] Christopher Ong, Kevin McGee, and Teong Leong Chuah. 2012. Closing the human-AI team-mate gap: how changes to displayed information impact player behavior towards computer teammates. In *Proceedings of the 24th Australian Computer-Human Interaction Conference*. 433–439.

[48] OpenAI. 2025. *Introducing ChatGPT agent: bridging research and action.* https://openai.com/index/introducing-chatgpt-agent/ Accessed: 2025-09-02.

[49] OpenAI. 2025. Introducing Operator | OpenAI. https://openai.com/index/introducing-operator/. Accessed: 2025-09-02.

[50] Yichen Pan, Dehan Kong, Sida Zhou, Cheng Cui, Yifei Leng, Bing Jiang, Hangyu Liu, Yanyi Shang, Shuyan Zhou, Tongshuang Wu, and Zhengyang Wu. 2024. WebCanvas: Benchmarking Web Agents in Online Environments. arXiv:2406.12373 [cs.CL] https://arxiv.org/abs/2406.12373

[51] Minjung Park, Jodi Forlizzi, and John Zimmerman. 2025. Exploring the Innovation Opportunities for Pre-trained Models. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference (DIS '25)*. 1973–2005. doi:10.1145/3715336.3735753

[52] Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, Michael Choi, Anish Agrawal, Arnav Chopra, Adam Khoja, Ryan Kim, Richard Ren, Jason Hausenloy, Oliver Zhang, Mantas Mazeika, Dmitry Dodonov, Tung Nguyen, Jaeho Lee, Daron Anderson, Mikhail Doroshenko, Alun Cennyth Stokes, Mobeen Mahmood, Oleksandr Pokutnyi, Oleg Iskra, Jessica P. Wang, John-Clark Levin, Mstyslav Kazakov, Fiona Feng, Steven Y. Feng, Haoran Zhao, Michael Yu, Varun Gangal, Chelsea Zou, Zihan Wang, Serguei Popov, Robert Gerbicz, Geoff Galgon, Johannes Schmitt, Will Yeadon, Yongki Lee, Scott Sauers, Alvaro Sanchez, Fabian Giska, Marc Roth, Søren Riis, Saiteja Utpala, Noah Burns, Gashaw M. Goshu, Mohinder Maheshbhai Naiya, Chidozie Agu, Zachary Giboney, Antrell Cheatom, Francesco Fournier-Facio, Sarah-Jane Crowson, Lennart Finke, Zerui Cheng, Jennifer Zampese, Ryan G. Hoerr, Mark Nandor, Hyunwoo Park, Tim Gehrunger, Jiaqi Cai, Ben McCarty, Alexis C Garretson, Edwin Taylor, Damien Sileo, Qiuyu Ren, Usman Qazi, Lianghui Li, Jungbae Nam, John B. Wydallis, Pavel Arkhipov, Jack Wei Lun Shi, Aras Bacho, Chris G. Willcocks, Hangrui Cao, Sumeet Motwani, Emily de Oliveira Santos, Johannes Veith, Edward Vendrow, Doru Cojoc, Kengo Zenitani, Joshua Robinson, Longke Tang, Yuqi Li, Joshua Vendrow, Natanael Wildner Fraga, Vladyslav Kuchkin, Andrey Pupasov Maksimov, Pierre Marion, Denis Efremov, Jayson Lynch, Kaiqu Liang, Aleksandar Mikov, Andrew Gritsevskiy, Julien Guillod, Gözdenur Demir, Dakotah Martinez, Ben Pageler, Kevin Zhou, Saeed Soori, Ori Press, Henry Tang, Paolo Rissone, Sean R. Green, Lina Brüssel, Moon Twayana, Aymeric Dieuleveut, Joseph Marvin Imperial, Ameya Prabhu, Jinzhou Yang, Nick Crispino, Arun Rao, Dimitri Zvonkine, Gabriel Loiseau, Mikhail Kalinin, Marco Lukas, Ciprian Manolescu, Nate Stambaugh, Subrata Mishra, Tad Hogg, Carlo Bosio, Brian P Coppola, Julian Salazar, Jaehyeok Jin, Rafael Sayous, Stefan Ivanov, Philippe Schwaller, Shaipranesh Senthilkuma, Andres M Bran, Andres Algaba, Kelsey Van den Houte, Lynn Van Der Sypt, Brecht Verbeken, David Noever, Alexei Kopylov, Benjamin Myklebust, Bikun Li, Lisa Schut, Evgenii Zheltonozhskii, Qiaochu Yuan, Derek Lim, Richard Stanley, Tong Yang, John Maar, Julian Wykowski, Martí Oller, Anmol Sahu, Cesare Giulio Ardito, Yuzheng Hu, Ariel Ghislain Kemogne Kamdoum, Alvin Jin, Tobias Garcia Vilchis, Yuexuan Zu, Martin Lackner, James Koppel, Gongbo Sun, Daniil S. Antonenko, Steffi Chern, Bingchen Zhao, Pierrot Arsene, Joseph M Cavanagh, Daofeng Li, Jiawei Shen, Donato Crisostomi, Wenjin Zhang, Ali Dehghan, Sergey Ivanov, David Perrella, Nurdin Kaparov, Allen Zang, Ilia Sucholutsky, Arina Kharlamova, Daniil Orel, Vladislav Poritski, Shalev Ben-David, Zachary Berger, Parker Whitfill, Michael Foster, Daniel Munro, Linh Ho, Shankar Sivarajan, Dan Bar Hava, Aleksey Kuchkin, David Holmes, Alexandra Rodriguez-Romero, Frank Sommerhage, Anji Zhang, Richard Moat, Keith Schneider, Zakayo Kazibwe, Don Clarke, Dae Hyun Kim, Felipe Meneguitti Dias, Sara Fish, Veit Elser, Tobias Kreiman, Victor Efren Guadarrama Vilchis, Immo Klose, Ujjwala Anantheswaran, Adam Zweiger, Kaivalya Rawal, Jeffery Li, Jeremy Nguyen, Nicolas Daans, Haline Heidinger, Maksim Radionov, Václav Rozhoň, Vincent Ginis, Christian Stump, Niv Cohen, Rafał Poświata, Josef Tkadlec, Alan Goldfarb, Chenguang Wang, Piotr Padlewski, Stanislaw Barzowski, Kyle Montgomery, Ryan Stendall, Jamie Tucker-Foltz, Jack Stade, T. Ryan Rogers, Tom Goertzen, Declan Grabb, Abhishek Shukla, Alan Givré, John Arnold Ambay, Archan Sen, Muhammad Fayez Aziz, Mark H Inlow, Hao He, Ling Zhang, Younesse Kaddar, Ivar Ängquist, Yanxu Chen, Harrison K Wang, Kalyan Ramakrishnan, Elliott Thornley, Antonio Terpin, Hailey Schoelkopf, Eric Zheng, Avishy Carmi, Ethan D. L. Brown, Kelin Zhu, Max Bartolo, Richard Wheeler, Martin Stehberger, Peter Bradshaw, JP Heimonen, Kaustubh Sridhar, Ido Akov, Jennifer Sandlin, Yury Makarychev, Joanna Tam, Hieu Hoang, David M. Cunningham, Vladimir Goryachev, Demosthenes Patramanis, Michael Krause, Andrew Redenti, David Aldous, Jesyin Lai, Shannon Coleman, Jiangnan Xu, Sangwon Lee, Ilias Magoulas, Sandy Zhao, Ning Tang, Michael K. Cohen, Orr Paradise, Jan Hendrik Kirchner, Maksym Ovchynnikov, Jason O. Matos, Adithya Shenoy, Michael Wang, Yuzhou Nie, Anna Sztyber-Betley, Paolo Faraboschi, Robin Riblet, Jonathan Crozier, Shiv Halasyamani, Shreyas Verma, Prashant Joshi, Eli Meril, Ziqiao Ma, Jérémy Andréoletti, Raghav Singhal, Jacob Platnick, Volodymyr Nevirkovets, Luke Basler, Alexander Ivanov, Seri Khoury, Nils Gustafsson, Marco Piccardo, Hamid Mostaghimi, Qijia Chen, Virendra Singh, Tran Quoc Khánh, Paul Rosu, Hannah Szlyk, Zachary Brown, Himanshu Narayan, Aline Menezes, Jonathan Roberts, William Alley, Kunyang Sun, Arkil Patel, Max Lamparth, Anka Reuel, Linwei Xin, Hanmeng Xu, Jacob Loader, Freddie Martin, Zixuan Wang, Andrea Achilleos, Thomas Preu, Tomek Korbak, Ida Bosio, Fereshteh Kazemi, Ziye Chen, Biró Bálint, Eve J. Y. Lo, Jiaqi Wang, Maria Inês S. Nunes, Jeremiah Milbauer, M Saiful Bari, Zihao Wang, Behzad Ansarinejad, Yewen Sun, Stephane Durand, Hossam Elgnainy, Guillaume Douville, Daniel Tordera, George Balabanian, Hew Wolff, Lynna Kvistad, Hsiaoyun Milliron, Ahmad Sakor, Murat Eron, Andrew Favre D. O., Shailesh Shah, Xiaoxiang Zhou, Firuz Kamalov, Sherwin Abdoli, Tim Santens, Shaul Barkan, Allison Tee, Robin Zhang, Alessandro Tomasiello, G. Bruno De Luca, Shi-Zhuo Looi, Vinh-Kha Le, Noam Kolt, Jiayi Pan, Emma Rodman, Jacob Drori, Carl J Fossum, Niklas Muennighoff, Milind Jagota, Ronak Pradeep, Honglu Fan, Jonathan Eicher, Michael Chen, Kushal Thaman, William Merrill, Moritz Firsching, Carter Harris, Stefan Ciobâcă, Jason Gross, Rohan Pandey, Ilya Gusev, Adam Jones, Shashank Agnihotri, Pavel Zhelnov, Mohammadreza Mofayezi, Alexander Piperski, David K. Zhang, Kostiantyn Dobarskyi, Roman Leventov, Ignat Soroko, Joshua Duersch, Vage Taamazyan, Andrew Ho, Wenjie Ma, William Held, Ruicheng Xian, Armel Randy Zebaze, Mohanad Mohamed, Julian Noah Leser, Michelle X Yuan, Laila Yacar, Johannes Lengler, Katarzyna Olszewska, Claudio Di Fratta, Edson Oliveira, Joseph W. Jackson, Andy Zou, Muthu Chidambaram, Timothy Manik, Hector Haffenden, Dashiell Stander, Ali Dasouqi, Alexander Shen, Bita Golshani, David Stap, Egor Kretov, Mikalai Uzhou, Alina Borisovna Zhidkovskaya, Nick Winter, Miguel Orbegozo Rodriguez, Robert Lauff, Dustin Wehr, Colin Tang, Zaki Hossain, Shaun Phillips, Fortuna Samuele, Fredrik Ekström, Angela Hammon, Oam Patel, Faraz Farhidi, George Medley, Forough Mohammadzadeh, Madellene Peñaflor, Haile Kassahun, Alena Friedrich, Rayner Hernandez Perez, Daniel Pyda, Taom Sakal, Omkar Dhamane, Ali Khajegili Mirabadi, Eric Hallman, Kenchi Okutsu, Mike Battaglia, Mohammad Maghsoudimehrabani, Alon Amit, Dave Hulbert, Roberto Pereira, Simon Weber, Handoko, Anton Peristyy, Stephen Malina, Mustafa Mehkary, Rami Aly, Frank Reidegeld, Anna-Katharina Dick, Cary Friday, Mukhwinder Singh, Hassan Shapourian, Wanyoung Kim, Mariana Costa, Hubeyb Gurdogan, Harsh Kumar, Chiara Ceconello, Chao Zhuang, Haon Park, Micah Carroll, Andrew R. Tawfeek, Stefan Steinerberger, Daattavya Aggarwal, Michael Kirchhof, Linjie Dai, Evan Kim, Johan Ferret, Jainam Shah, Yuzhou Wang, Minghao Yan, Krzysztof Burdzy, Lixin Zhang, Antonio Franca, Diana T. Pham, Kang Yong Loh, Joshua Robinson, Abram Jackson, Paolo Giordano, Philipp Petersen, Adrian Cosma, Jesus Colino, Colin White, Jacob Votava, Vladimir Vinnikov, Ethan Delaney, Petr Spelda, Vit Stritecky, Syed M. Shahid, Jean-Christophe Mourrat, Lavr Vetoshkin, Koen Sponselee, Renas Bacho, Zheng-Xin Yong, Florencia de la Rosa, Nathan Cho, Xiuyu Li, Guillaume Malod, Orion Weller, Guglielmo Albani, Leon Lang, Julien Laurendeau, Dmitry Kazakov, Fatimah Adesanya, Julien Portier, Lawrence Hollom, Victor Souza, Yuchen Anna Zhou, Julien Degorre, Yiğit Yalın, Gbenga Daniel Obikoya, Rai, Filippo Bigi, M. C. Boscá, Oleg Shumar, Kaniuar Bacho, Gabriel Recchia, Mara Popescu, Nikita Shulga, Ngefor Mildred Tanwie, Thomas C. H. Lux, Ben Rank, Colin Ni, Matthew Brooks, Alesia Yakimchyk, Huanxu, Liu, Stefano Cavalleri, Olle Häggström, Emil Verkama, Joshua Newbould, Hans Gundlach, Leonor Brito-Santana, Brian Amaro, Vivek Vajipey, Rynaa Grover, Ting Wang, Yosi Kratish, Wen-Ding Li, Sivakanth Gopi, Andrea Caciolai, Christian Schroeder de Witt, Pablo Hernández-Cámara, Emanuele Rodolà, Jules Robins, Dominic Williamson, Vincent Cheng, Brad Raynor, Hao Qi, Ben Segev, Jingxuan Fan, Sarah Martinson, Erik Y. Wang, Kaylie Hausknecht, Michael P. Brenner, Mao Mao, Christoph Demian, Peyman Kassani, Xinyu Zhang, David Avagian, Eshawn Jessica Scipio, Alon Ragoler, Justin Tan, Blake Sims, Rebeka Plecnik, Aaron Kirtland, Omer Faruk Bodur, D. P. Shinde, Yan Carlos Leyva Labrador, Zahra Adoul, Mohamed Zekry, Ali Karakoc, Tania C. B. Santos, Samir Shamseldeen, Loukmane Karim, Anna Liakhovitskaia, Nate Resman, Nicholas Farina, Juan Carlos Gonzalez, Gabe Maayan, Earth Anderson, Rodrigo De Oliveira Pena, Elizabeth Kelley, Hodjat Mariji, Rasoul Pouriamanesh, Wentao Wu, Ross Finocchio, Ismail Alarab, Joshua Cole, Danyelle Ferreira, Bryan Johnson, Mohammad Safdari, Liangti Dai, Siriphan Arthornthurasuk, Isaac C. McAlister, Alejandro José Moyano, Alexey Pronin, Jing Fan, Angel Ramirez-Trinidad, Yana Malysheva, Daphiny Pottmaier, Omid Taheri, Stanley Stepanic, Samuel Perry, Luke Askew, Raúl Adrián Huerta Rodríguez, Ali M. R. Minissi, Ricardo Lorena, Krishnamurthy Iyer, Arshad Anil Fasiludeen, Ronald Clark, Josh Ducey, Matheus Piza, Maja Somrak, Eric Vergo, Juehang Qin, Benjámin Borbás, Eric Chu, Jack Lindsey, Antoine Jallon, I. M. J. McInnis, Evan Chen, Avi Semler, Luk Gloor, Tej Shah, Marc Carauleanu, Pascal Lauer, Tran Đuc Huy, Hossein Shahrtash, Emilien Duc, Lukas Lewark, Assaf Brown, Samuel Albanie, Brian Weber, Warren S. Vaz, Pierre Clavier, Yiyang Fan, Gabriel Poesia Reis e Silva, Long, Lian, Marcus Abramovitch, Xi Jiang, Sandra Mendoza, Murat Islam, Juan Gonzalez, Vasilios Mavroudis, Justin Xu, Pawan Kumar, Laxman Prasad Goswami, Daniel Bugas, Nasser Heydari, Ferenc Jeanplong, Thorben Jansen, Antonella Pinto, Archimedes Apronti, Abdallah Galal, Ng Ze-An, Ankit Singh, Tong Jiang, Joan of Arc Xavier, Kanu Priya Agarwal, Mohammed Berkani, Gang Zhang, Zhehang Du, Benedito Alves de Oliveira Junior, Dmitry Malishev, Nicolas Remy, Taylor D. Hartman, Tim Tarver, Stephen Mensah, Gautier Abou Loume, Wiktor Morak, Farzad Habibi, Sarah Hoback, Will Cai, Javier Gimenez, Roselynn Grace Montecillo, Jakub Łucki, Russell Campbell, Asankhaya

Sharma, Khalida Meer, Shreen Gul, Daniel Espinosa Gonzalez, Xavier Alapont, Alex Hoover, Gunjan Chhablani, Freddie Vargus, Arunim Agarwal, Yibo Jiang, Deepakkumar Patil, David Outevsky, Kevin Joseph Scaria, Rajat Maheshwari, Abdelkader Dendane, Priti Shukla, Ashley Cartwright, Sergei Bogdanov, Niels Mündler, Sören Möller, Luca Arnaboldi, Kunvar Thaman, Muhammad Rehan Siddiqi, Prajvi Saxena, Himanshu Gupta, Tony Fruhauff, Glen Sherman, Mátyás Vincze, Siranut Usawasutsakorn, Dylan Ler, Anil Radhakrishnan, Innocent Enyekwe, Sk Md Salauddin, Jiang Muzhen, Aleksandr Maksapetyan, Vivien Rossbach, Chris Harjadi, Mohsen Bahaloohoreh, Claire Sparrow, Jasdeep Sidhu, Sam Ali, Song Bian, John Lai, Eric Singer, Justine Leon Uro, Greg Bateman, Mohamed Sayed, Ahmed Menshawy, Darling Duclosel, Dario Bezzi, Yashaswini Jain, Ashley Aaron, Murat Tiryakioglu, Sheeshram Siddh, Keith Krenek, Imad Ali Shah, Jun Jin, Scott Creighton, Denis Peskoff, Zienab EL-Wasif, Ragavendran P V, Michael Richmond, Joseph McGowan, Tejal Patwardhan, Hao-Yu Sun, Ting Sun, Nikola Zubić, Samuele Sala, Stephen Ebert, Jean Kaddour, Manuel Schottdorf, Dianzhuo Wang, Gerol Petruzella, Alex Meiburg, Tilen Medved, Ali ElSheikh, S Ashwin Hebbar, Lorenzo Vaquero, Xianjun Yang, Jason Poulos, Vilém Zouhar, Sergey Bogdanik, Mingfang Zhang, Jorge Sanz-Ros, David Anugraha, Yinwei Dai, Anh N. Nhu, Xue Wang, Ali Anil Demircali, Zhibai Jia, Yuyin Zhou, Juncheng Wu, Mike He, Nitin Chandok, Aarush Sinha, Gaoxiang Luo, Long Le, Mickaël Noyé, Michał Perełkiewicz, Ioannis Pantidis, Tianbo Qi, Soham Sachin Purohit, Letitia Parcalabescu, Thai-Hoa Nguyen, Genta Indra Winata, Edoardo M. Ponti, Hanchen Li, Kaustubh Dhole, Jongee Park, Dario Abbondanza, Yuanli Wang, Anupam Nayak, Diogo M. Caetano, Antonio A. W. L. Wong, Maria del Rio-Chanona, Dániel Kondor, Pieter Francois, Ed Chalstrey, Jakob Zsambok, Dan Hoyer, Jenny Reddish, Jakob Hauser, Francisco-Javier Rodrigo-Ginés, Suchandra Datta, Maxwell Shepherd, Thom Kamphuis, Qizheng Zhang, Hyunjun Kim, Ruiji Sun, Jianzhu Yao, Franck Dernoncourt, Satyapriya Krishna, Sina Rismanchian, Bonan Pu, Francesco Pinto, Yingheng Wang, Kumar Shridhar, Kalon J. Overholt, Glib Briia, Hieu Nguyen, David, Soler Bartomeu, Tony CY Pang, Adam Wecker, Yifan Xiong, Fanfei Li, Lukas S. Huber, Joshua Jaeger, Romano De Maddalena, Xing Han Lù, Yuhui Zhang, Claas Beger, Patrick Tser Jern Kon, Sean Li, Vivek Sanker, Ming Yin, Yihao Liang, Xinlu Zhang, Ankit Agrawal, Li S. Yifei, Zechen Zhang, Mu Cai, Yasin Sonmez, Costin Cozianu, Changhao Li, Alex Slen, Shoubin Yu, Hyun Kyu Park, Gabriele Sarti, Marcin Briański, Alessandro Stolfo, Truong An Nguyen, Mike Zhang, Yotam Perlitz, Jose Hernandez-Orallo, Runjia Li, Amin Shabani, Felix Juefei-Xu, Shikhar Dhingra, Orr Zohar, My Chiffon Nguyen, Alexander Pondaven, Abdurrahim Yilmaz, Xuandong Zhao, Chuanyang Jin, Muyan Jiang, Stefan Todoran, Xinyao Han, Jules Kreuer, Brian Rabern, Anna Plassart, Martino Maggetti, Luther Yap, Robert Geirhos, Jonathon Kean, Dingsu Wang, Sina Mollaei, Chenkai Sun, Yifan Yin, Shiqi Wang, Rui Li, Yaowen Chang, Anjiang Wei, Alice Bizeul, Xiaohan Wang, Alexandre Oliveira Arrais, Kushin Mukherjee, Jorge Chamorro-Padial, Jiachen Liu, Xingyu Qu, Junyi Guan, Adam Bouyamourn, Shuyu Wu, Martyna Plomecka, Junda Chen, Mengze Tang, Jiaqi Deng, Shreyas Subramanian, Haocheng Xi, Haoxuan Chen, Weizhi Zhang, Yinuo Ren, Haoqin Tu, Sejong Kim, Yushun Chen, Sara Vera Marjanović, Junwoo Ha, Grzegorz Luczyna, Jeff J. Ma, Zewen Shen, Dawn Song, Cedegao E. Zhang, Zhun Wang, Gaël Gendron, Yunze Xiao, Leo Smucker, Erica Weng, Kwok Hao Lee, Zhe Ye, Stefano Ermon, Ignacio D. Lopez-Miguel, Theo Knights, Anthony Gitter, Namkyu Park, Boyi Wei, Hongzheng Chen, Kunal Pai, Ahmed Elkhanany, Han Lin, Philipp D. Siedler, Jichao Fang, Ritwik Mishra, Károly Zsolnai-Fehér, Xilin Jiang, Shadab Khan, Jun Yuan, Rishab Kumar Jain, Xi Lin, Mike Peterson, Zhe Wang, Aditya Malusare, Maosen Tang, Isha Gupta, Ivan Fosin, Timothy Kang, Barbara Dworakowska, Kazuki Matsumoto, Guangyao Zheng, Gerben Sewuster, Jorge Pretel Villanueva, Ivan Rannev, Igor Chernyavsky, Jiale Chen, Deepayan Banik, Ben Racz, Wenchao Dong, Jianxin Wang, Laila Bashmal, Duarte V. Gonçalves, Wei Hu, Kaushik Bar, Ondrej Bohdal, Atharv Singh Patlan, Shehzaad Dhuliawala, Caroline Geirhos, Julien Wist, Yuval Kansal, Bingsen Chen, Kutay Tire, Atak Talay Yücel, Brandon Christof, Veerupaksh Singla, Zijian Song, Sanxing Chen, Jiaxin Ge, Kaustubh Ponkshe, Isaac Park, Tianneng Shi, Martin Q. Ma, Joshua Mak, Sherwin Lai, Antoine Moulin, Zhuo Cheng, Zhanda Zhu, Ziyi Zhang, Vaidehi Patil, Ketan Jha, Qiutong Men, Jiaxuan Wu, Tianchi Zhang, Bruno Hebling Vieira, Alham Fikri Aji, Jae-Won Chung, Mohammed Mahfoud, Ha Thi Hoang, Marc Sperzel, Wei Hao, Kristof Meding, Sihan Xu, Vassilis Kostakos, Davide Manini, Yueying Liu, Christopher Toukmaji, Jay Paek, Eunmi Yu, Arif Engin Demircali, Zhiyi Sun, Ivan Dewerpe, Hongsen Qin, Roman Pflugfelder, James Bailey, Johnathan Morris, Ville Heilala, Sybille Rosset, Zishun Yu, Peter E. Chen, Woongyeong Yeo, Eeshaan Jain, Ryan Yang, Sreekar Chigurupati, Julia Chernyavsky, Sai Prajwal Reddy, Subhashini Venugopalan, Hunar Batra, Core Francisco Park, Hieu Tran, Guilherme Maximiano, Genghan Zhang, Yizhuo Liang, Hu Shiyu, Rongwu Xu, Rui Pan, Siddharth Suresh, Ziqi Liu, Samaksh Gulati, Songyang Zhang, Peter Turchin, Christopher W. Bartlett, Christopher R. Scotese, Phuong M. Cao, Aakaash Nattanmai, Gordon McKellips, Anish Cheraku, Asim Suhail, Ethan Luo, Marvin Deng, Jason Luo, Ashley Zhang, Kavin Jindel, Jay Paek, Kasper Halevy, Allen Baranov, Michael Liu, Advaith Avadhanam, David Zhang, Vincent Cheng, Brad Ma, Evan Fu, Liam Do, Joshua Lass, Hubert Yang, Surya Sunkari, Vishruth Bharath, Violet Ai, James Leung, Rishit Agrawal, Alan Zhou, Kevin Chen, Tejas Kalpathi, Ziqi Xu, Gavin Wang, Tyler Xiao, Erik Maung, Sam Lee, Ryan Yang, Roy Yue, Ben Zhao, Julia Yoon, Sunny Sun, Aryan Singh, Ethan Luo, Clark Peng, Tyler Osbey, Taozhi Wang, Daryl Echeazu, Hubert Yang, Timothy Wu, Spandan Patel, Vidhi Kulkarni, Vijaykaarti Sundarapandiyan, Ashley Zhang, Andrew Le, Zafir Nasim, Srikar Yalam, Ritesh Kasamsetty, Soham Samal, Hubert Yang, David Sun, Nihar Shah, Abhijeet Saha, Alex Zhang, Leon Nguyen, Laasya Nagumalli, Kaixin Wang, Alan Zhou, Aidan Wu, Jason Luo, Anwith Telluri, Summer Yue, Alexandr Wang, and Dan Hendrycks. 2025. Humanity's Last Exam. arXiv:2501.14249 [cs.LG] https://arxiv.org/abs/2501.14249

[53] Christopher Rawles, Sarah Clinckemaillie, Yifan Chang, Jonathan Waltz, Gabrielle Lau, Marybeth Fair, Alice Li, William Bishop, Wei Li, Folawiyo Campbell-Ajala, et al. 2024. Androidworld: A dynamic benchmarking environment for autonomous agents. *arXiv preprint arXiv:2405.14573* (2024).

[54] Matt Renner and Matt A.V. Chaban. 2025. *601 real-world gen AI use cases from the world's leading organizations.* https://cloud.google.com/transform/101-real-world-generative-ai-use-cases-from-industry-leaders Published April 12, 2024; last updated April 9, 2025. Google Cloud..

[55] Olawale Salaudeen, Anka Reuel, Ahmed Ahmed, Suhana Bedi, Zachary Robertson, Sudharsan Sundar, Ben Domingue, Angelina Wang, and Sanmi Koyejo. 2025. Measurement to Meaning: A Validity-Centered Framework for AI Evaluation. arXiv:2505.10573 [cs.CY] https://arxiv.org/abs/2505.10573

[56] Johnny Saldaña. 2021. *The Coding Manual for Qualitative Researchers* (4 ed.). Sage Publications, London.

[57] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 36. arXiv:2302.04761 [cs.CL] https://arxiv.org/abs/2302.04761

[58] Yijia Shao, Humishka Zope, Yucheng Jiang, Jiaxin Pei, David Nguyen, Erik Brynjolfsson, and Diyi Yang. 2025. Future of Work with AI Agents: Auditing Automation and Augmentation Potential across the U.S. Workforce. *arXiv preprint arXiv:2506.06576* (2025).

[59] Thomas B. Sheridan. 1992. *Telerobotics, Automation, and Human Supervisory Control.* MIT Press, Cambridge, MA.

[60] Ben Shneiderman. 1983. Direct manipulation: A step beyond programming languages. *Computer* 16, 08 (1983), 57–69.

[61] Ben Shneiderman and Pattie Maes. 1997. Direct manipulation vs. interface agents. *Interactions* 4, 6 (Nov. 1997), 42–61. doi:10.1145/267505.267514

[62] Leon Staufer, Mick Yang, Anka Reuel, and Stephen Casper. 2025. Audit Cards: Contextualizing AI Evaluations. arXiv:2504.13839 [cs.CY] https://arxiv.org/abs/2504.13839

[63] Hari Subramonyam, Divy Thakkar, Andrew Ku, Juergen Dieber, and Anoop K. Sinha. 2025. Prototyping with Prompts: Emerging Approaches and Challenges in Generative AI Design for Collaborative Software Teams. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3706598.3713166

[64] Hanna Wallach, Meera Desai, A Feder Cooper, Angelina Wang, Chad Atalla, Solon Barocas, Su Lin Blodgett, Alexandra Chouldechova, Emily Corvi, P Alex Dow, et al. 2025. Position: Evaluating generative ai systems is a social science measurement challenge. *arXiv preprint arXiv:2502.00561* (2025).

[65] Zora Zhiruo Wang, Yijia Shao, Omar Shaikh, Daniel Fried, Graham Neubig, and Diyi Yang. 2025. How Do AI Agents Do Human Work? Comparing AI and Human Workflows Across Diverse Occupations. *arXiv preprint arXiv:2510.22780* (2025).

[66] Justin D. Weisz, Jessica He, Michael Muller, Gabriela Hoefer, Rachel Miles, and Werner Geyer. 2024. Design Principles for Generative AI Applications. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–22. doi:10.1145/3613904.3642466

[67] Alma Whitten and J Doug Tygar. 1999. Why Johnny Can't Encrypt: A Usability Evaluation of PGP 5.0.. In *USENIX Security Symposium*, Vol. 348. 169–184.

[68] Wikipedia contributors. 2025. Office Assistant. https://en.wikipedia.org/wiki/Office_Assistant. Accessed: 2025-09-04.

[69] Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. 2025. OSWORLD: benchmarking multimodal agents for open-ended tasks in real computer environments. In *Proceedings of the 38th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '24)*. Curran Associates Inc., Red Hook, NY, USA, Article 1650, 55 pages.

[70] Deshraj Yadav, Rishabh Jain, Harsh Agrawal, Prithvijit Chattopadhyay, Taranjeet Singh, Akash Jain, Shiv Baran Singh, Stefan Lee, and Dhruv Batra. 2019. EvalAI: Towards Better Evaluation Systems for AI Agents. arXiv:1902.03570 [cs.AI] https://arxiv.org/abs/1902.03570

[71] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *International Conference on Learning Representations (ICLR)*. arXiv:2210.03629 [cs.CL] https://arxiv.org/abs/2210.03629

[72] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *CHI 2019*. ACM. https://www.microsoft.com/en-us/research/publication/understanding-

the-effect-of-accuracy-on-trust-in-machine-learning-models/

[73] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 437, 21 pages. doi:10.1145/3544548.3581388

[74] Chaoyun Zhang, Shilin He, Jiaxu Qian, Bowen Li, Liqun Li, Si Qin, Yu Kang, Minghua Ma, Guyue Liu, Qingwei Lin, Saravan Rajmohan, Dongmei Zhang, and Qi Zhang. 2025. Large Language Model-Brained GUI Agents: A Survey. arXiv:2411.18279 [cs.AI] https://arxiv.org/abs/2411.18279

[75] Qiaoning Zhang, Matthew L Lee, and Scott Carter. 2022. You Complete Me: Human-AI Teams and Complementary Expertise. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 114, 28 pages. doi:10.1145/3491102.3517791

[76] Rui Zhang, Nathan J. McNeese, Guo Freeman, and Geoff Musick. 2020. "An Ideal Human": Expectations of AI Teammates in Human-AI Teaming. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (December 2020), 246:1–246:25. doi:10.1145/3432945

[77] Yuxuan Zhu, Tengjun Jin, Yada Pruksachatkun, Andy Zhang, Shu Liu, Sasha Cui, Sayash Kapoor, Shayne Longpre, Kevin Meng, Rebecca Weiss, et al. 2025. Establishing best practices for building rigorous agentic benchmarks. *arXiv preprint arXiv:2507.02825* (2025).

[78] Mingchen Zhuge, Changsheng Zhao, Dylan Ashley, Wenyi Wang, Dmitrii Khizbullin, Yunyang Xiong, Zechun Liu, Ernie Chang, Raghuraman Krishnamoorthi, Yuandong Tian, Yangyang Shi, Vikas Chandra, and Jürgen Schmidhuber. 2024. Agent-as-a-Judge: Evaluate Agents with Agents. arXiv:2410.10934 [cs.AI] https://arxiv.org/abs/2410.10934

A Systematic Review

A.1 Methodology

To construct the taxonomy, we surveyed AI agent product websites and aggregator articles. This is a *marketed-as-agentic* corpus: we included products that are marketed or described as "AI agents" (or equivalent), and the corpus may include boundary cases (e.g., recommender-style or assistant-style products that use the "agent" label) because our goal is to capture how the label is applied in industry. We isolated use cases by deliverable and by verbs describing the automated process, then performed open, inductive coding and clustered use cases until we reached three orthogonal capability types.

Search criteria. We drew on aggregated lists (Google Cloud 601 Real-World AI Use Cases [54], AI Agent Directory [8], Product Hunt [26]) and Google searches for "AI agents" and "LLM agents," assembling $N = 102$ products. We stopped when new searches yielded agents whose use cases overlapped with those already identified.

Exclusion criteria. We excluded academic articles and tools for users with specialized domain knowledge. To capture agents for non-technical end-users (the "Johnny" persona [67, 73]), we excluded developer-focused tools (e.g., Cursor [1]) but included consumer-facing ones (e.g., Lovable [5]).

Limitations. The "AI agent" category is rapidly evolving; we do not claim to have reviewed all products or that our list is representative of all future offerings.

Analysis. Two authors performed an inductive qualitative content analysis [56] of marketed capabilities. Following prior work on pre-trained model use-cases [51], we read customer-journey descriptions, produced initial codes independently, then reached consensus on a codebook [43] and applied it to the full set. We then abstracted codes into higher-level categories using established taxonomy development [46] until three orthogonal categories covered every agent.

A.2 Taxonomy Capability Map

Category	Capability	Count	Example Agents
Orchestration	Automation	36	Salesforce Agentforce; BotPress
	Direct UI / Commands	18	Volkswagen IDA; Copilot
Creation	Writing	25	Ubisoft Ghostwriter
	Audio	2	ElevenLabs
	Websites & Apps	3	Lovable; Den (Workspace)
	Presentations	2	Gamma; Beautiful.AI
	Images	2	L'Oréal AI Agent
Insight	Information Retrieval	98	Perplexity AI; Cohesity Gaia
	Recommendations	44	Netflix; Spotify AI DJ
	Data Analysis	31	Hex; ThoughtSpot Spotter
	Synthesis	17	Fullstory Behavioral Data Agent
	Evaluation	6	Wipro Contract Assistant
	Personalization	2	Mercedes-Benz CLA Agent

Table 4: Full taxonomy of marketed AI agent capabilities (counts by subcategory; products may appear in multiple categories). Entries reflect *marketed* capabilities; products in this map may or may not fully instantiate operationally agentic properties.

A.3 Systematic Review Codebook

Table 5: Codebook: Taxonomy of Marketed Use Cases for AI Agents

Category	Code Name	Definition
Orchestration	**Output Purpose**	**Produces executable actions or commands**.
	Direct UI manipulation	Agent interacts with software interfaces on the user's behalf.
	GUI Automation	Clicking, scrolling, or selecting elements in a graphical interface.
	Terminal Automation	Running shell or command-line instructions.
Creation	**Output Purpose**	**Generates new content shaped by user input.**
	Writing	Producing text-based outputs.
	Emails	Drafting structured, professional or personal communications.
	Articles	Creating longer-form written content.
	Marketing fliers	Designing promotional or advertising copy.
	Letters	Generating formal or informal correspondence.
	Website and Application Development	Writing software, websites, or applications.
	Images	Producing static visual content.
	Music	Composing audio tracks (limited commercial viability).
	Videos	Generating video clips or animations (emerging).
	Sound	Creating audio effects or spoken content.
Insight	**Output Purpose**	**Refines data using cognitive and analytical methods, to guide decision-making.**
	Research	Gathering and synthesizing external knowledge.
	Information Retrieval	Providing factual or domain-specific information.
	Insight	Drawing conclusions or highlighting patterns.
	Data Analysis	Processing and interpreting structured or unstructured data.
	Evaluation	Using the model to judge, score, or refine outputs.
	Synthesis	Combining multiple inputs into a coherent whole.
	Recommendations	Offering next steps, decisions, or actions.

B User Study

B.1 Recruitment, Compensation, & Ethical Review

We recruited participants by advertising on social media platforms such as X and LinkedIn, as well as posting the study on Prolific. 841 prospective participants completed an initial screening questionnaire that we linked to from our social media post, from which we selected 13. We separately recruited 18 participants from Prolific, which resulted in $N = 31$ participants in total. These participants are listed in Table 6. We selected participants to capture a wide range of age groups, occupations, and prior experience with conversational AI tools. Participants were compensated $25 for their interview via Amazon eGift Cards or directly through the approval of Prolific submissions.

B.2 Logistics

For participants recruited through social media outreach, selected participants received a link to sign up for an appointment slot through online scheduling software and a link to a consent form. Before each task, a password manager was used to send temporary links to access each agent. Participants were introduced to the study and asked to open a link to the online survey. They were read the task description and given 20 minutes per task. The interviewer clarified task-related questions and communicated intermittently while the participant worked to capture live reactions. In general, the interviewer did not answer questions about the user interface, except in limited circumstances where the tool appeared to be broken, the participant attempted to solve the task outside of the agent, or attempted to perform tasks beyond the task description. This was done to ensure that the interviews stayed at or close to the expected time and to collect qualitative data on parts of the task that could be completed even when some agent components were not functional, as was the case in several interviews with Operator.

B.3 Agent Interfaces

Manus. Manus is a general-purpose AI agent that can perform Web searches and control a computer with a visible browser, terminal, and filesystem. Its interface (see Figure 5) consists of three panes: a list of past task threads on the left, a chat thread in the center, and a live preview of the agent's computer on the right with a replay dial and task progress checklist.

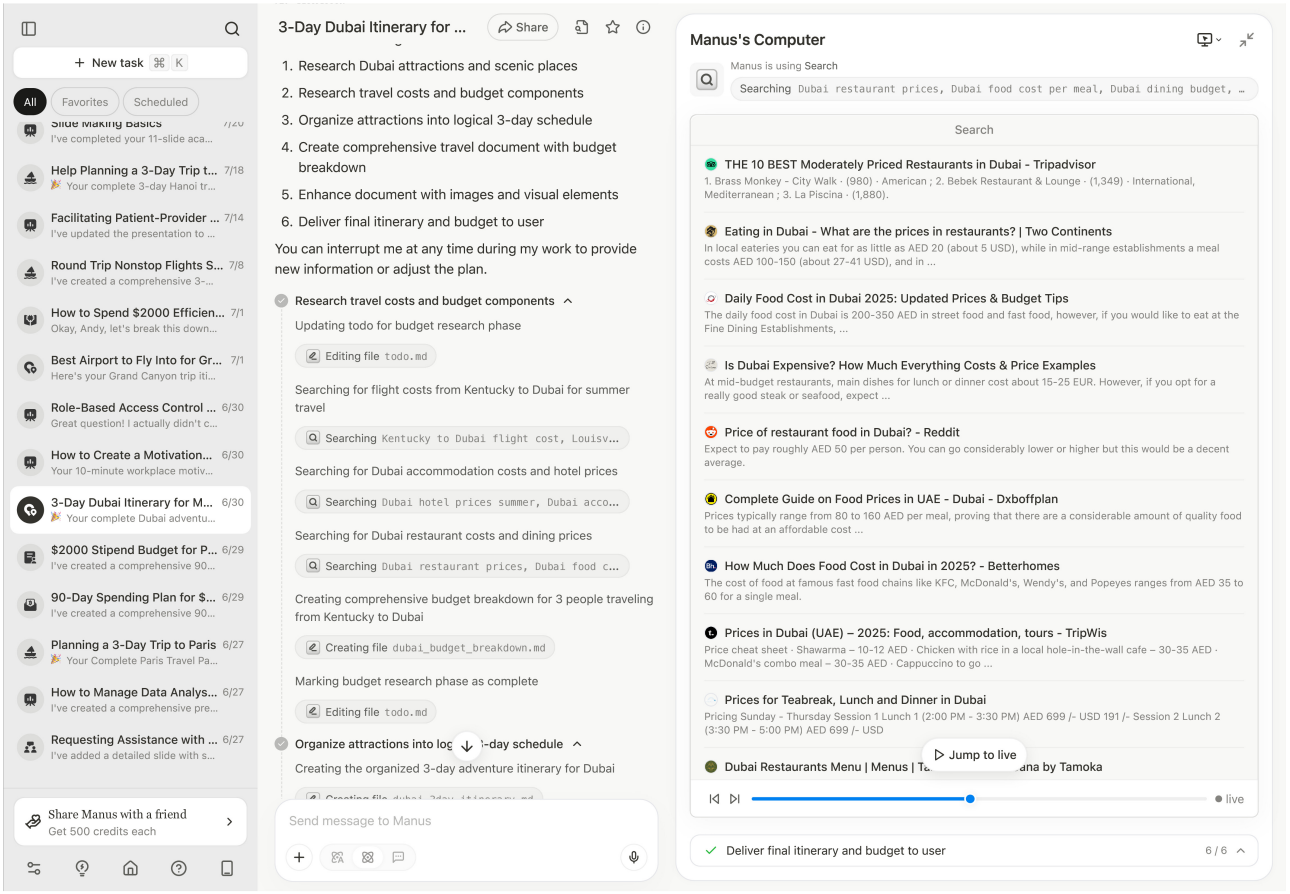


Figure 5: Manus, one of our AI Agents, working on our Holiday Planning task.

When a user sends a prompt, Manus provides a summary of planned steps and shows a detailed log of every action it performs, with interactive chips linking to search results and screen states. It terminates tasks with clickable links to generated files. Users can send additional prompts while the agent is working, and can take direct control of the computer via a "Take Over" button.

Operator. Operator has capabilities comparable to those of Manus, but with a more minimal interface. Its initial screen, as shown in Figure 6, features task idea cards that help users get started with known tasks it can perform. After entering the first prompt, the page transitions to a traditional chat thread view.

After the user provides a prompt, Operator displays a pulsing circle to denote the processing of the user's input. Unlike most generative AI chatbots and Manus, it does not stream responses but displays them all at once, even if the response spans multiple screens. If it determines that computer use is appropriate for a task, it creates a rectangular preview of the computer that is embedded in the chat thread. This can be expanded to another layout where it occupies nearly three-quarters of the screen space from the right. The user can "Take Over" to control the computer, which passes direct manipulation commands such as keystrokes and mouse clicks through to Operator's computer, after which they are requested to record the changes they made before Operator assumes control again.

B.4 Task Selection Criteria and Task Screenshots

We defined three tasks for our participants to attempt during the study–one for each category of our AI agent taxonomy–and selected them using three criteria.

Familiarity. Tasks must be generally familiar to a non-specialized end-user, and ideally ones they have completed in the past without AI assistance. Not only does this ensure tasks are representative of tasks users might want to delegate in the wild, it controls for failure modes unrelated to the use of the AI Agent, the main item of interest in our study.

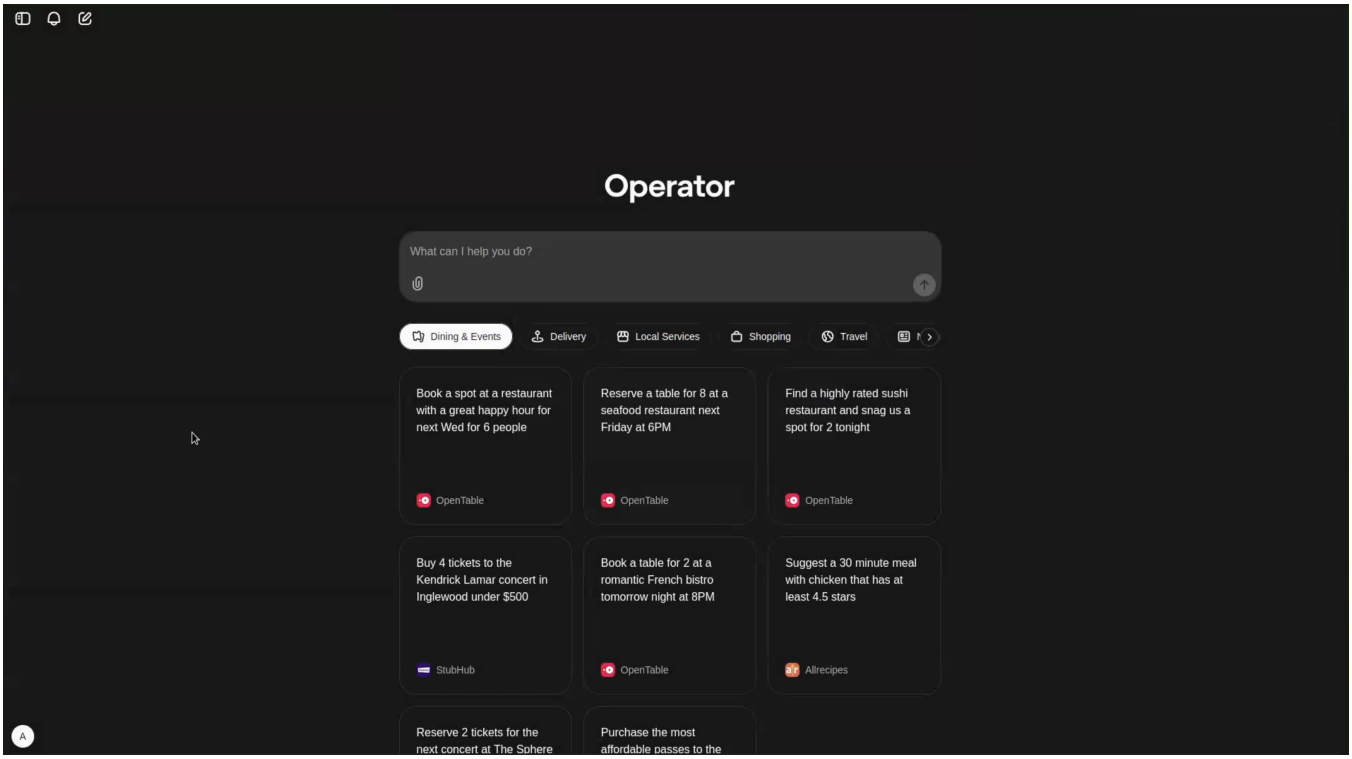


Figure 6: The initial screen on Operator, featuring a text box to enter prompts, and a grid of example tasks in various categories.

Speed. Tasks must be completed relatively quickly. We chose to observe participants working on two tasks in two 20-minute sessions in an hour. This gave users sufficient time to complete each task and allowed the interviewer the opportunity to ask several open-ended questions. Two tasks allowed us to observe differences within-subject when the tasks differed in either the category or the tool used.

Emotional stakes. Tasks must have some emotional stakes so that the user is personally invested in the outcome despite carrying them out within the context of a research study. Most tasks in the real-world have subjective user acceptance criteria, and we would like them to be non-trivial so our results are externally valid. For example, CREATION tasks often incorporate visual design. If the user is not personally invested in the topic, they could trivially declare victory, even if it did not reflect the care they might place on deliverables whose quality was personally significant.

B.5 Session Script

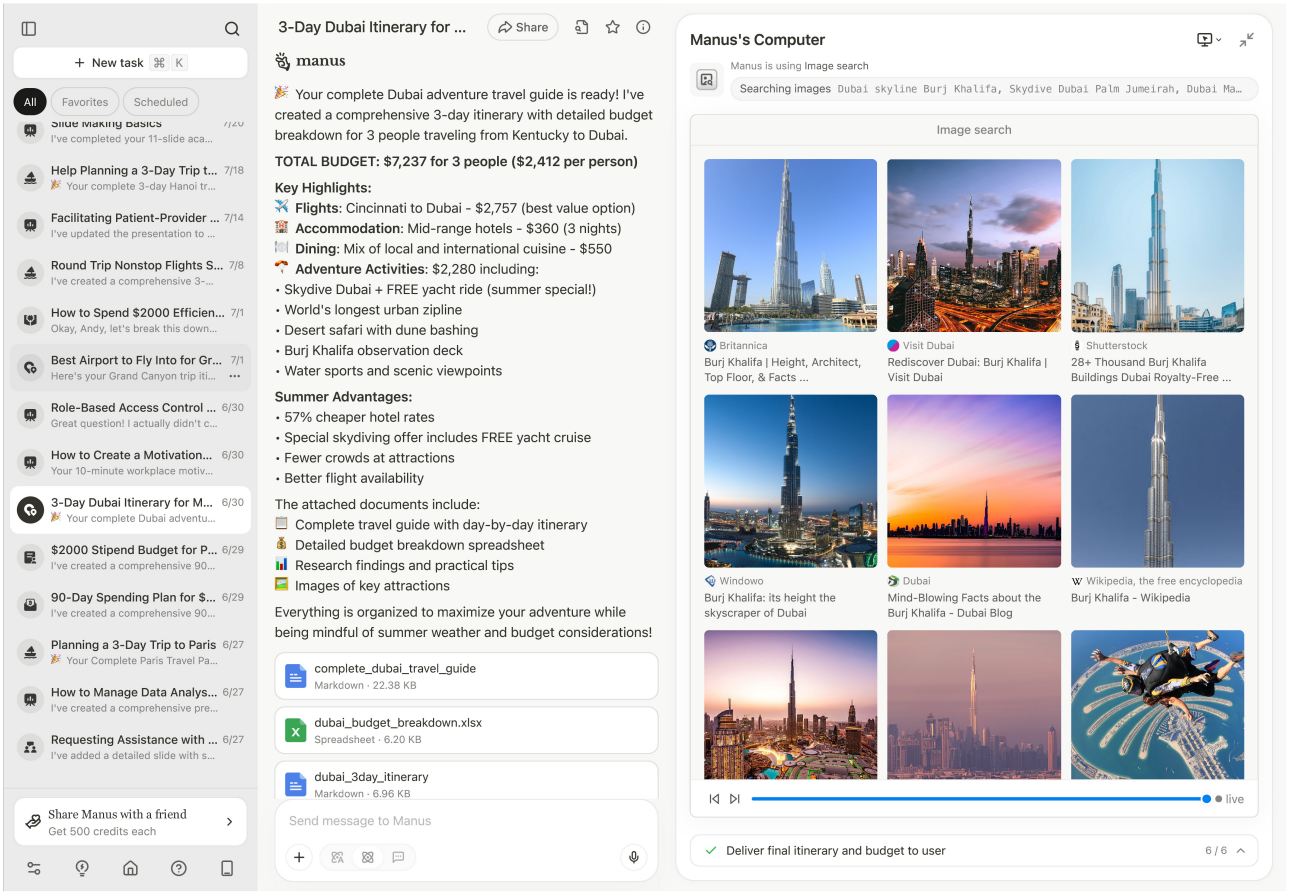


Figure 7: Manus, one of our AI Agents, operating a computer.

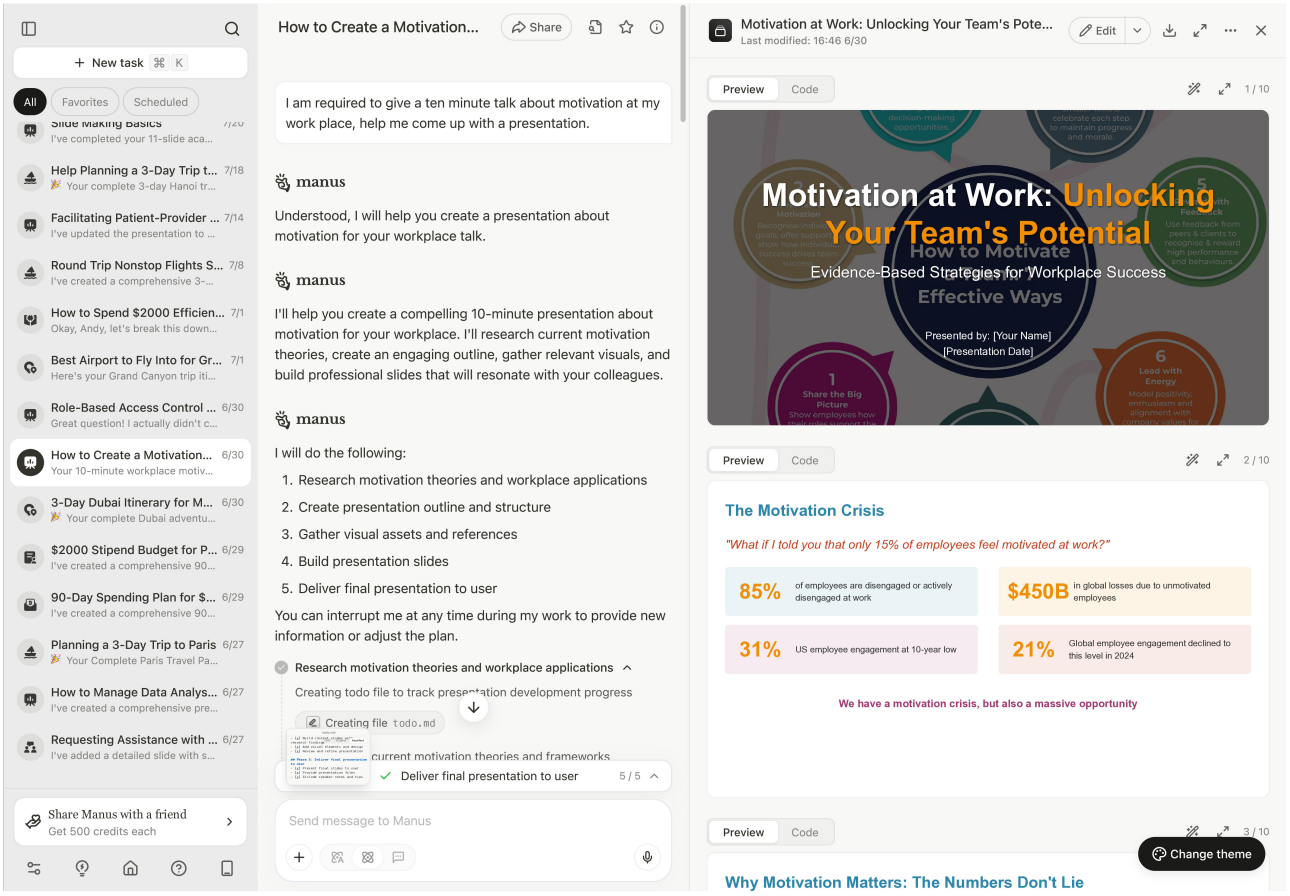


Figure 8: Manus working on the creation task (slide making).

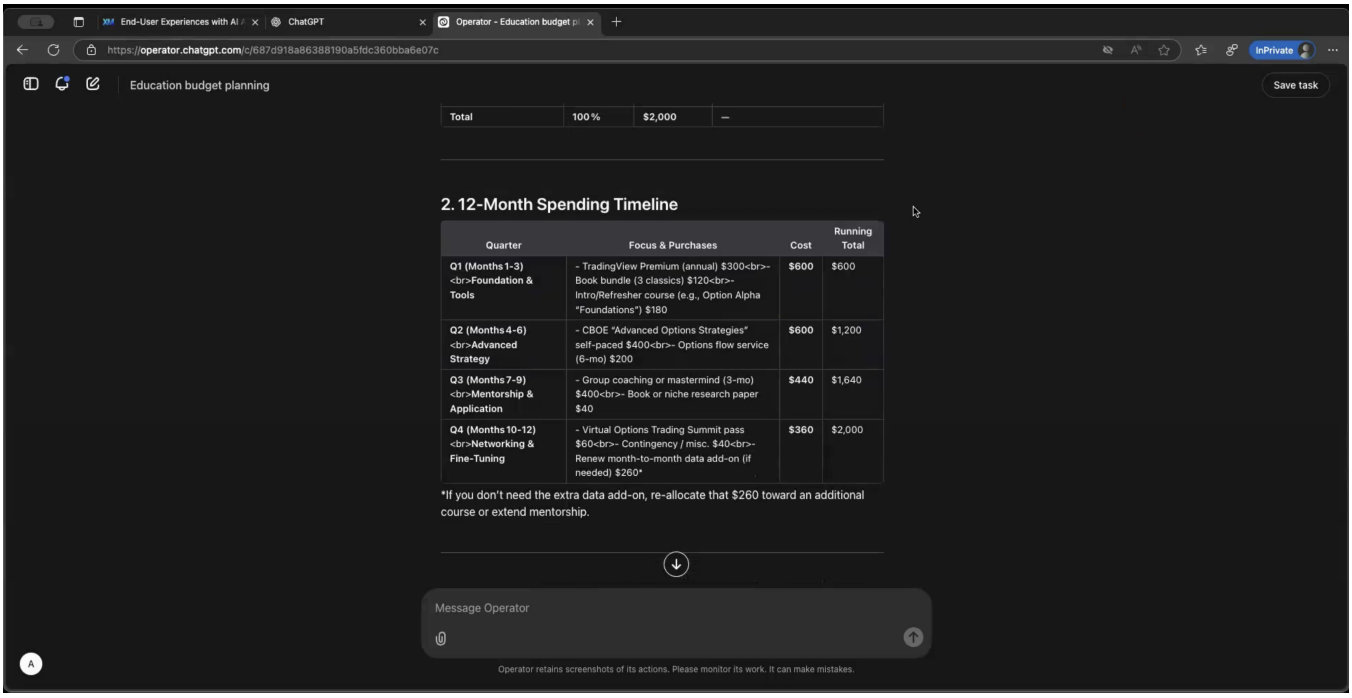


Figure 9: OpenAI Operator, sharing a sample budget with a participant. Participants appreciated its succinct tables.

Interviewer: Hi, thanks for participating in our research study on AI agents! I'm **[First Interviewer Name]**, the Principal Investigator. Before we start, please make sure you're in a private space.

Confirming participant eligibility.

(1) *Interviewer checks if the interviewee filled out the consent form emailed to them, and if not asks them to fill it out.*
(2) *Interviewer asks the interviewee if they are a person, by asking them to come on video, and then turning it off.*
(3) *Interviewer asks the participant to share their screen and open* ***IPInfo****, https://ipinfo.io/country, which displays a page with just a 2 letter country code. US confirms they are connecting to the site from the United States.*
(4) *Interviewer ensures participant is connected via a desktop computer (as opposed to mobile / tablet).*

If the interviewee does not satisfy all the above criteria, they are asked to leave.

Interviewer: We will be recording your audio and screen, so we want to make sure other people aren't accidentally in the recording. Please turn off your video. Please refrain from sharing any identifiable, personal or sensitive information about yourself or others you would not want shared outside the research setting.

Interviewer: Here is a link to the form with survey responses: **[Link to online survey on Qualtrics]**.

Interviewer: Please close tabs that contain personally identifying information, and share just the window with two tabs with me, one for an agent, and the other with the form open.

Interviewer: Let me know once you're ready.

Wait for the participant to load the survey and is ready.

Interviewer: We're recording now. Here's what our session will look like. We're going to have you attempt two tasks, each using one of two agents:

- **Manus**, a general-purpose AI agent.
- **Operator**, a computer-using agent.

Interviewer: After each task, you will complete the questionnaire. We will conclude with general questions and confirm your email address for the gift card.

Interviewer randomly chooses two tasks and tools.

<Task 1 script from Appendix B.6>

Interviewer asks the participant questions including and beyond those from Section B.7.

Participant will fill out their UID as assigned by the PI. They will fill out the task 1 questions on the questionnaire.

<Task 2 script from Appendix B.6>.

Interviewer asks the participant questions including and beyond those from Section B.7

Participant will fill out the task 2 questions on the questionnaire.

Interviewer: Thank you for participating in this study, we're all set.

Participants from Prolific are approved from the study dashboard. For those from social media outreach, confirm their email address before we exist, and purchase a USD 25 Amazon eGift card after the session.

B.6 Task Script

Interviewer: Here are the credentials we'll be using for Manus / Operator.

Interviewer will send the user a temporary 1Password link to credentials for the respective agent, and a link to the agent, that they will open in an incognito window.

Interviewer: Welcome to **[Agent]**. You'll now use a tool that lets you talk to an AI assistant, similar to ChatGPT. In this system, you type your goal or request, and the assistant will respond and try to help. This tool may include memory, workflows, or documents that the assistant can reference as it helps you. Try to complete the task using the assistant however it feels natural to you.

Holiday Planning (Orchestration). Imagine you're traveling to a city you've never visited. You will use the AI Agent to produce an itinerary including flight tickets, housing, activities, and any sightseeing in a new city you're going to on holiday. As you go, you'll be able to revise your preferences or ask the agent to change plans. Any questions?

Slide Making (Creation). You will be using **[agent]** to create a slide deck for a 10 minute talk on a topic you're interested in. You'll have 20 minutes for this task. As you work on the task, I'd like you to talk through your thought process. Once you're done, I'll ask you some questions, and you'll answer some survey questions. As you go, you'll be able to revise your preferences or ask the agent to change plans. Any questions?

Personal / Professional Development Stipend Budgeting (Insight). Imagine you've received a USD 2,000 personal and professional development stipend that expires after 90 days. You will be using Manus / ChatGPT Operator to create a table with the purchases, the cost, and why this is aligned with your life goals. The purchases must be related to education, health, or career. You'll have 20 minutes for this task. As you work on the task, I'd like you to talk through your thought process. Once you're done, I'll ask you some questions, and you'll answer some online survey questions. As you go, you'll be able to revise your preferences or ask the agent to change plans. Any questions?

B.7 Sample Questions for Semi-Structured Interview

- If you could wave a magic wand and change one aspect of the user experience, what would that be?
- Going forward, how would you complete this task?
- Was there anything that surprised you during the experience?
- Were there any moments you weren't sure what to do next?

B.8 Participants

Participant	Gender	Age Group	Generative AI Tool Usage	Education
P01	Man	25-34	Daily or almost daily	Master's
P02	Man	18-24	Daily or almost daily	Bachelor's
P03	Man	25-34	Several times a week	Bachelor's
P04	Man	45-54	About once a week	Bachelor's
P05	Man	25-34	Several times a week	Bachelor's
P06	Man	25-34	Several times a week	Master's
P07	Man	45-54	Daily or almost daily	Master's
P08	Man	25-34	Daily or almost daily	Bachelor's
P09	Man	55-64	Several times a week	Master's
P10	Man	18-24	Daily or almost daily	Some college
P11	Man	25-34	Daily or almost daily	Bachelor's
P12	Man	25-34	Daily or almost daily	Master's
P13	Woman	25-34	Daily or almost daily	Bachelor's
P14	Woman	25-34	Daily or almost daily	High School
P15	Woman	45-54	Several times a week	Some college
P16	Woman	35-44	Several times a week	Bachelor's
P17	Man	35-44	Daily or almost daily	Bachelor's
P18	Man	25-34	Daily or almost daily	Some college
P19	Woman	45-54	Daily or almost daily	Bachelor's
P20	Man	55-64	Daily or almost daily	Bachelor's
P21	Man	35-44	Daily or almost daily	Bachelor's
P22	Man	65+	Daily or almost daily	Some college
P23	Man	45-54	Daily or almost daily	Bachelor's
P24	Woman	18-24	Daily or almost daily	Bachelor's
P25	Man	35-44	Several times a week	Master's
P26	Man	35-44	Daily or almost daily	Master's
P27	Woman	25-34	Several times a week	Trade school
P28	Man	35-44	Daily or almost daily	Bachelor's
P29	Woman	55-64	Several times a week	Bachelor's
P30	Man	25-34	Less than once a month	Bachelor's
P31	Man	25-34	Daily or almost daily	Bachelor's

Table 6: Demographic Data of User Study Participants

B.9 Codebook

Table 7: Codebook: User Experience Codes for AI Agents

Code Name	Definition
AccuracyConcern	User questions whether links/data are trustworthy ("is it hallucinating?")
VerificationRequest	User asks the agent to cite or link to verifiable sources
PaymentTrust	Willingness to let the agent initiate or complete a purchase
CredentialTrust	Willingness to let the agent log in or handle sensitive credentials
CredentialDistrust	Reluctance to let the agent handle credentials or sensitive logins.
OperatorFailure	Operator Computer Use failed enough to need PI intervention
TaskCompletionPerfect	No changes to the deliverable
TaskCompletionHigh	Minimal changes
TaskCompletionPartial	Big picture, but lots of manual editing
TaskCompletionNone	No progress on task
MissedPreferences	Failing to honor user-stated preferences or requirements
OutlineDeviation	Not following the user's requested structure or outline
BudgetOverrun	Exceeding allotted budget or resources
DateError	Using the wrong year (e.g. 2024 vs. 2025)
LinkOmission	Forgetting to provide critical booking or reference links
OverResearch	Spending too much time on background research vs. deliverable
ComponentFailure	A sub-component of the agent did not work as expected
TaskAvoidance	Operator skipping or being unable to complete core steps
SlidesImageQualityIssue	Images in slide deck were unacceptable to user
SlidesTextDensityIssue	Slides or outputs overwhelmed with too much text
AgentTooSlow	User labels the agent response taking more time than acceptable
AgentAcceptableSpeed	User labels agent as taking an acceptable amount of time
AgentFastResponse	User labels agent as being lightening quick in responding and completing the task
NewInfoExposure	Users felt they learned or discovered new information
CommunicationClarity	How clearly the process was explained or narrated
WorkflowTransparency	Visibility into how inputs produce outputs over time
OutputClarity	How understandable the results or deliverables are
OutputHardToFind	User struggles to know if task is complete and where to access deliverable
ErrorCommunicationConfusion	Something goes wrong but user doesn't know what, and how to troubleshoot
HighTextLoad	Too much text output by agent
VisualComplexity	Busy layouts, with many screens and lots happening simultaneously
UndiscoverableFeature	User expresses desire for an action that is possible but cannot be identified
InterruptHesitation	Users hesitate to interrupt when they want to course correct
DesireToTakeOverProcess	Users want to do something themselves
DesireToLeaveAgent	User is frustrated or hits a problem, and wants to switch to a manual process or use an external tool

Code Name	Definition
StrategyMisalignment	Agent follows a different process from user's expectations
ThirdPartyWorkflow	When Operator tries to use 3rd-party tools instead of completing the task in situ
PerceptionAgentMismatchedForTask	Belief that the agent is not suited for the user's task.
DislikedComputerUse	Negative reaction to relying on computer interaction.
UserPrefersDirectManipulation	Preference for manual control over mediated AI use.
PerceptionGoodForAsynchronous Use	Viewed as better fit for tasks done over time, not live.
PerceptionSuggestionsAreKnown	Agent suggestions feel obvious or redundant.
PerceptionTakeControlDefeatsThe PointOfAgents	Too much user control undermines the value of an agent.
PerceptionOperatorAestheticsDelight	User enjoys design/look of the interface.
PreferenceBreadthBeforeDepth	User wants broad options before deep exploration.
ExpectationUseMemoryMore	Expectation that agent should remember context better.
PerceptionUnderwhelmed	Agent output felt unimpressive.
PerceptionSlowerThanHuman	Agent perceived as slower than manual work.
WishDeeperPersonalization	Desire for more tailored outputs.
UserIsSatisfied	Expression of positive outcome with no issues.
ResearcherPromptToConverge	Researcher intervenes to guide the session.
TooltipsAreSuperfluous	User finds on-screen help unnecessary.
PerceptionMarkdownSourceIsConfusing	Markdown/raw text presentation confuses user.
PerceptionComputerUseBadFit	Belief that computer use is poorly matched to task.
DislikedComputerStream	User dislikes continuous text streaming.
LikedComputerStream	User enjoys continuous text streaming.
PromptStrategyBigInitialPrompt	User attempts large all-in-one prompt.
UserIsImpressed	Agent exceeds expectations.
EasyToUse	Agent feels simple and intuitive.
DesireToTweakOutput	User wants fine-grained editing ability.
PromptStrategyStepByStep	User prompts iteratively in smaller steps.
FeltSentient	User perceives the agent as humanlike.
NextStepsUnclear	User unsure what to do after agent's output.
MentalModelChanged	Interaction shifts user's understanding of the system.
AgentIsIntelligent	User describes the agent as smart or capable.
VisualCommunicationDesired	User wants visual aids like diagrams/maps.
ComprehensiveAndDetailed Information	Agent delivers thorough, detailed content.
ProvidedUsefulInformation	Output deemed helpful.
CapturedPreferences	Agent retained and applied user preferences.
StrategyAligned	Agent's approach matched the user's intent.
ErrorCommunicationClarity	Errors explained clearly (or lack thereof noted).
PerceptionComputerUseGoodFit	Belief that computer use suits the task well.
SlidesDesignPraise	Compliments on slide aesthetics.
UserAgentMisalignment	Misfit between agent's behavior and user's needs.
SlidesLayoutOrDesignIssue	Problems with slide formatting or design.
PerceptionPlainText Information IsDifficultToMakeSenseOf	Plain text seen as hard to interpret.
UserWantsSocialProofForInformation	Desire for references or validation of info.
WishOutputMoreInteractiveAndRich	User wants multimedia/interactive results.
UserLikesTheAbilityToInterrupt	Appreciates being able to stop agent mid-task.
WishLessCommunicative	Prefers fewer explanations or chatter.
PerceptionCommunicationSlowsTaskCompletion	Belief that agent's communication delays progress.
UserWantsDirectManipulation	Wants manual control rather than agent mediation.
UserDoesNotWantWorkflowLogs	Prefers not to see step-by-step logs.
CommunicationIsRedundant	Agent repeats unnecessary information.

Code Name	Definition
PerceptionEditsTakeTooLong	Editing with agent is seen as slow.
OutputFileNotPortable	Output format is inconvenient to share.
PerceptionMakesAssumptions	Agent assumes details not provided by user.
PerceptionNotPersonalizedEnough	Response felt generic or untailored.
PerceptionDidNotAskQuestions	Agent failed to clarify user intent.
CommunicationIsNotClear	Agent's output or instructions lack clarity.
OpinionAIReplacesHumans	User reflects on AI taking over human roles.
PerceptionNewUserExperienceUnfamiliar	First-time use felt confusing.
UserDoesNotWantToAuthenticate	Reluctance to log in or provide identity.
UserIsConfused	User expresses confusion.
WishSkillsAndStrengthsPreview	Desire to see what the agent can do upfront.
AgentUsedToolUserIsUnfamiliarWith	Agent invoked tools unknown to user.
WishMoreCommunicative	User wants more explanations or conversation.
DesireDarkMode	Request for dark-theme interface.
DesireSimpleAndAdvancedInteraction Mode	Wish for toggle between basic and advanced use.
DesireForSimplerProcess	User wants fewer steps.
UserIsDissatisfied	Negative reaction to overall experience.
AgentIgnoresUserNeeds	Agent fails to account for stated requirements.
PromptStrategyUsedAmbiguous Keywords	User inputs vague terms in prompts.
AgentIsResponsive	Agent replies quickly and appropriately.
AgentUsesNicheTools	Agent employs specialized or obscure tools.
PerceptionTimeElapsedHeightensExpectations	Longer wait raises user expectations.
AgentSavesALotOfTime	Agent perceived as time-saving.
UserLovesToExploreTheTool	User enjoys experimenting with the agent.
OutcomeQualityExceedsThatOfManualProcess	Results judged better than manual work.
UserConfusedAboutTaskInputs	User unsure what input is required.
PerceptionEffortToProduceIsLow	User feels little effort is needed to create output.
PerceptionUserHasLowInvestmentInAICreatedDeliverables	User less attached to AI-produced results.
PreferenceRestartTaskInsteadOfEditDeliverable	Prefers starting over instead of editing.
DesireToRestartFromScratch	Wants to redo from beginning.
DesirePauseAgentToReviseWork	Wish to halt and revise output mid-flow.
ObservationFirstPromptHasHigh Impact	First prompt shapes overall results strongly.
DesireCompareManualProcess	Wants to contrast AI vs manual process.
PerceptionDoNotTrust	User expresses distrust in the agent.
PerceptionPartialWorkIsValuable	Even incomplete outputs seen as useful.
DesireHideTaskBar	Request to remove or hide task bar.
DesireToUseAgain	User wants to reuse the agent.
DesireCitationsAndSources	Desire for references backing outputs.
InteractionExperienceUsedVoiceMode	User interacted with the agent via voice.