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

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

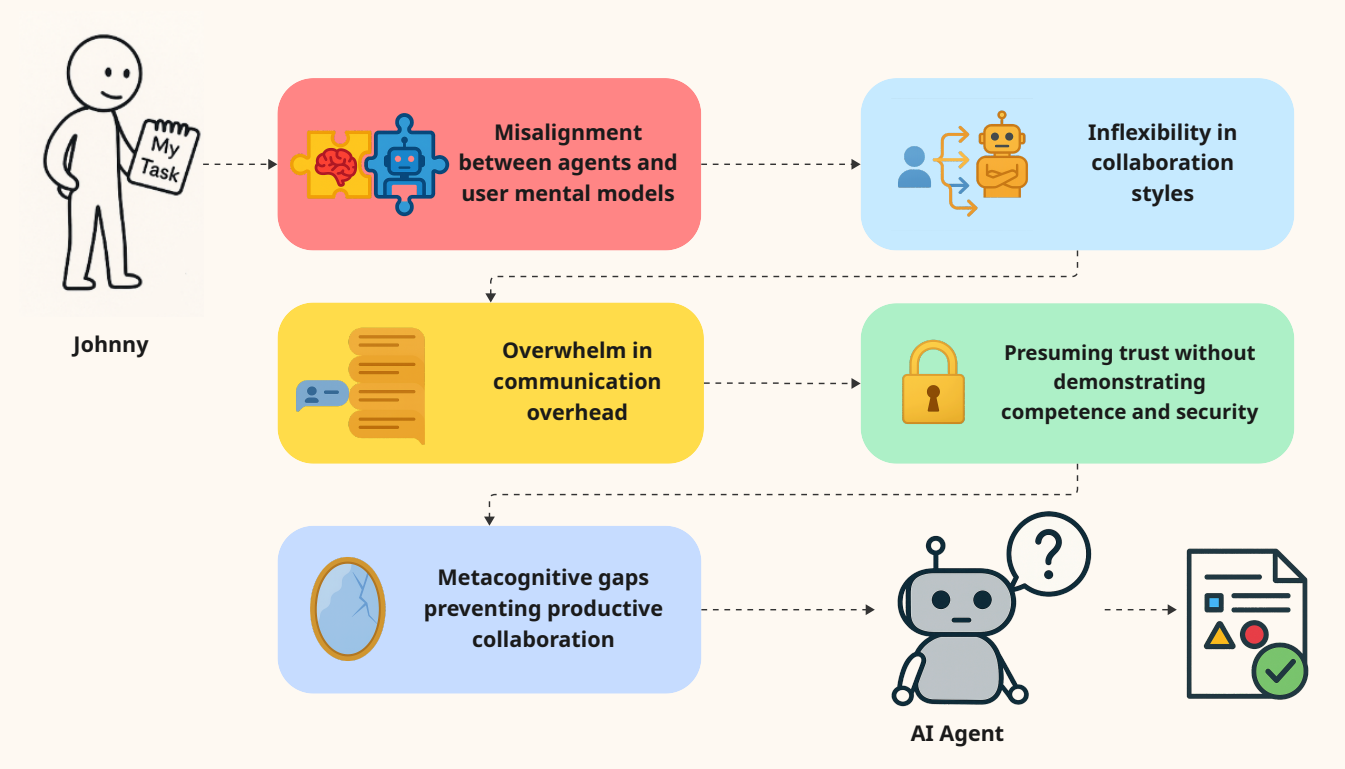

Figure 1: We identify five critical usability barriers novice users face when interacting with AI agents to accomplish a task. As they attempt to delegate their task, they must progressively develop a mental model of the agent's capabilities, navigate its rigid collaboration style, decode its overwhelming communication, all while the agent isn't fully cognizant of its own capabilities, and presumes it has a user's complete trust.

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 of 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. Next, 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 several critical usability challenges ranging from agent capabilities that were misaligned with user mental models to agents lacking the meta-cognitive abilities necessary for effective collaboration.

arXiv '25, Pittsburgh, PA, United States

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

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. 2025. Why Johnny Can't Use Agents: Industry Aspirations vs. User Realities with AI Agent Software. In *Proceedings of arXiv (arXiv '25)*, 27 pages.

1 Introduction

Visions of computing systems as intelligent partners for knowledge work stretch back nearly a century with Vannevar Bush's 1945 accounting of the "Memex," a mechanized desk that could augment human memory and association by linking together trails of knowledge [22]. Two decades later, Licklider's article on "Man–Machine Symbiosis" sharpened this vision into a research program for interactive computing, suggesting that computers might soon "formulate hypotheses" and "make decisions" in concert with their human users [48]. The HCI community has long been animated by these possibilities: from early Wizard-of-Oz experiments simulating what may be possible with intelligent interfaces [34], to Horvitz's explorations of agent-mediated "mixed-initiative" interfaces for scheduling and meeting management [36], to Shneiderman and Maes' canonical debate on direct manipulation versus interface agents [52, 70], to end-user programming systems like if-this-then-that (IFTTT) that help users configure semi-autonomous software assistants [75]. Along the way, ill-fated counter examples such as Microsoft's Clippy [79] highlighted the risks of overzealous agent designs that frustrated users by, as Adar might put it, having UIs that wrote checks the underlying AI couldn't cash [11].

Today, the dream has resurfaced under the banner of "AI agents." In 2025, thousands of start-ups and major technology companies have advertised agentic products that can orchestrate workflows, generate creative outputs, and deliver insights on demand. Commercial products such as OpenAI's ChatGPT Operator and emerging agent platforms like Manus are marketed to non-technical end-users as productivity multipliers. These systems also revive older paradigms "centaur" human–AI teaming [41], in which humans and agents collaborate by sharing complementary strengths.

But, as with many technologies at the peak of their hype cycle, what an "AI agent" is, what it can do, and how well it fits into the messy realities of everyday knowledge work remain underspecified. Much of the conversation about agent performance — in benchmarks, demo videos, venture funding pitches, and media hype — has focused on easy-to-measure ideals. But does an agent scoring 26% versus 24% on "Humanity's Last Exam" matter to a user when a user wants help with booking a flight? More broadly, what we lack is systematic understanding of how the tech industry conceives of "AI agents" as a category of software, and of how end-users actually experience them in practice. Prior Human-AI interaction research cautions that without clear expectations, transparency, and user control [14, 69], agents risk repeating the mistakes of the past.

Against this backdrop, we ask the following two research questions:

1.1 Research Questions

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.

RQ2: What challenges do end-users face when attempting to use commercial AI agents for their advertised uses? Here, we seek to understand the frictions that emerge when aspirational marketing collides with lived practice.

1.2 Summary of Work

To answer these questions, we first conducted a review of the marketed use cases for AI agents. 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 Google searches to increase the diversity of products. Based on our data set, we analyzed themes across verbiage relating to agent functionalities to extract abstract use cases. We distilled a taxonomy of three umbrella categories: **Orchestration**, **Creation**, and **Insight**. Figure 2 provides an overview of our research process.

We then ran a think-aloud usability study in which $N = 31$ participants attempted representative tasks in each of the three categories using two popular commercial agent platforms—Operator and Manus. We supplemented these think-aloud studies with semi-structured interviews. Our study revealed several recurring usability challenges. Although generally successful at accomplishing the assigned tasks (which were designed 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 meta-cognitive abilities that enable productive collaboration. Figure 1 symbolically depicts Johnny, our proverbial end-user, navigating these challenges as they attempt to delegate a task.

These usability breakdowns echo long-standing critiques of agentic systems in HCI: the dangers of ceding too much initiative, the costs of opaque automation, and the challenge of aligning agent actions with human goals.

1.3 Contributions

To summarize, this paper makes three core contributions.

- We reviewed and organized how the tech industry conceives of and markets AI agents, yielding a taxonomy of three core agent "abilities".
- 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.
- We distill a set of design recommendations for designing AI agents that are usable, effective, and collaborative thought partners.

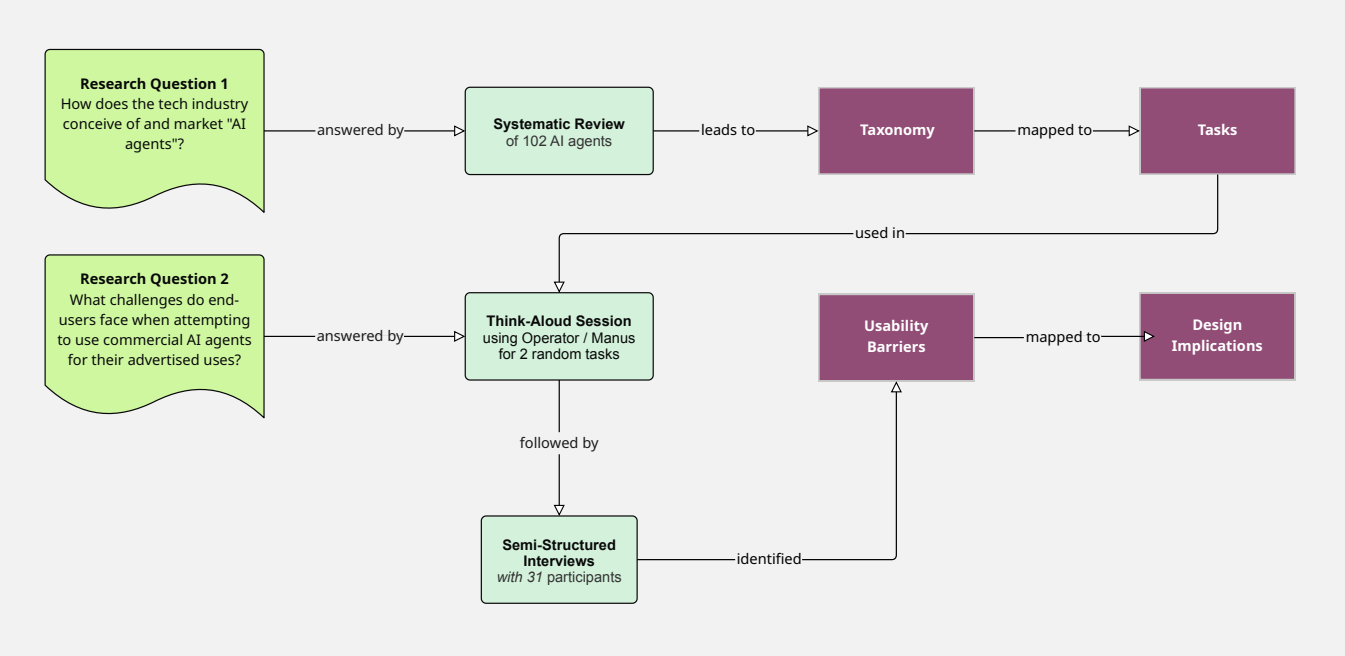

Figure 2: Overview of our research process. We conducted a systematic review to build a taxonomy of marketed use cases of AI agents. Next, we conducted a think-aloud session in which participants completed tasks corresponding to our umbrella use cases, from which we identified usability barriers. These barriers lead to a set of design implications for next-generation AI agent design.

2 Related Work

Our work is situated at the intersection of three areas of inquiry in the HCI and AI literature, spanning work conceptualizing and realizing end-user interface agents, empirical studies of Human-AI collaboration, and evaluations of models and agents.

2.1 Interface Agents

Although direct manipulation [68] has been the dominant interaction paradigm since the invention of graphical user interfaces (GUIs), researchers have long envisioned alternatives. Interface agents, intermediaries between a user and a computing system that invoke system commands on behalf of the user, are one such alternative [52]. Due to the increasing complexity of software, in the 1990s, which had a ripple effect on interface design, HCI researchers conceptualized what a world with interface agents would look like. This included a characterization of agent types, behaviors, interactions, and challenges [36, 49, 50, 52, 70]. To explore the broader design space of intelligent agents, without being bottlenecked by the capabilities of AI at that time, Maulsby et al. [54] introduced Wizard of Oz prototyping for intelligent agents in HCI research.

In the 2000s, agent-oriented programming frameworks [71] such as GOAL [28] and JADE[18] were popularized as a new paradigm for the design of interface agent software by providing explicit constructs to internalize beliefs, decisions, capabilities, and obligations. Such formal software development paradigms enabled the management of multi-agent systems, powering agents that could buy and sell products, discover information, and manage email accounts on behalf of end-users. This period also saw the emergence of virtual assistants such as Microsoft Clippy [79], which attempted to understand the intent of the user to execute direct manipulation commands in a rule-based manner — though these virtual assistants were limited and largely unutilized.

Moving forward, the 2010s represented the era of chatbots and voice assistants, exemplified by Amazon Alexa [78] and Siri [80]. Around this time, AI technology had advanced to enable speech recognition and machine translation, allowing agents to predict needs and initiate interactions. Emerging NLP model capabilities including intent detection, entity extraction, and slot filling were used by researchers to explore designs for scheduling agents such as Calendar.help [26].

Buoyed by strong advances in transformer-based large language models (LLMs) to generate large amounts of coherent text in response to text input [21], the 2020s have witnessed an explosion in the popularity of GUI agents that use commands generated by LLMs to complete goal-directed actions in Web browsers. Several groups of authors have conducted large-scale surveys and enumerations of end-user GUI agents, including ones by Hu et al. on OS Agents [37], and Zhang et al. on so-called "LLM-brained GUI agents" [85]. These surveys examine numerous aspects ranging from the historical evolution of GUI agents, their capabilities, tools and techniques for engineering them, benchmarks and evaluation techniques, security and privacy challenges, and limitations.

Building on this rich tradition of interface agent research, our work examines industry positioning of contemporary "AI agents"

and end-user experiences with commercial implementations, moving beyond technical demonstrations to understand practical usability challenges.

2.2 Human-AI Collaboration

2.2.1 Human Perceptions and Mental Models of AI Teammates. The literature on human-AI teaming underscores the critical role of human perceptions and mental models in fostering successful collaboration [58, 87]. Kaur et al. propose a method to build shared mental models [42], while HACO [31] offers a framework of human-AI teaming concepts, supported by an empirical study of a multi-agent system that enhances teamwork through agent augmentation. Notably, researchers have observed that AI teammates are judged differently than their human counterparts, often receiving more blame for failures [55]. Furthermore, Khadpe et al. [43] reveal that the conceptual metaphors used to describe conversational AI agents significantly influence adoption—paradoxically, emphasizing higher competence can actually reduce users' willingness to adopt.

2.2.2 The Influence of Trust and Explanations in AI Communications. Trust is another critical dimension affecting human-AI collaboration [46]. In an empirical study, researchers discovered that people's trust in models is influenced both by their stated and observed accuracy, showcasing how important not only true capabilities are but also communicated capabilities [83]. In the space of LLMs, another empirical study identified that LLM usage of expressions of uncertainty did reduce overreliance of AI, cementing how communication in human-AI teams materially affects outcomes [44].

The growing field of explainable AI (XAI) [29] seeks to scaffold user trust in model predictions by providing users with rationales for how a model arrived at a prediction. Recent work on "Human-Centered" XAI has sought to broaden perspectives on XAI beyond simple interpretations of model behavior and to consider, instead, the broader sociotechnical context of who is using that model and for what purpose [32, 47]. In that vein, Miller provides best practices on how AI should explain itself to humans, based on 250 articles from the social science literature, recommending explanations be contrastive, be carefully selected for contextual relevance amongst a motley of valid causes, and incorporate human conversational dynamics. [56].

2.2.3 Factors Influencing Human-AI Team Success. Zhang et al. [86] find that human-AI teams find the best results when each member has complementary expertise, noting that humans can evaluate task expertise based on confidence signals embedded in AI communication. In addition, recent work shows the harms of using insufficiently complex machine learning models for a given task [67].

However, updating an AI model to improve performance is not a panacea. Such updates can affect the mental model of the user, which can degrade the user's experience and success [15], corroborated by user responses during GPT 5's release [33]. Zamfirescu-Pereira et al. [84] highlight the challenges of prompt engineering by non-technical end users, likening it to programming with natural language specifications. These challenges echo studies of learning barriers found in end-user programming systems [45]. Finally, Bansal et al. [16] investigate the communication challenges humans and AI agents face while collaborating.

Extending the broader literature on human-AI teaming, our work explores breakdowns novice users face when aiming to use commercially available "AI agents" for tasks they are designed to do — breakdowns that encompass user mental models, trust, and communication styles.

2.3 Evaluations of LLMs and AI Agents

AI has a long history of benchmarks and datasets for evaluating models, which have been extended to agent evaluation systems [24, 40, 64, 82, 88]. Recent work attempts to revisit how we assess AI agents by incorporating personal user data [23], creating new frameworks [40], evaluation environments for computer-based knowledge work [30, 61, 81], agent cards [13, 74], domain-specific benchmarks [39], and cross-domain cognition assessments [63]. Nevertheless, as Salaudeen et al. [66] observe, many AI evaluations suffer from notable validity gaps. Crucially, regardless of their strengths and weaknesses, many evaluations narrowly focus on technical competence and ignore human-centered dimensions such as usability, interpretability, trust, and user satisfaction. This view is espoused by Wallach et al. who argue that measurement tasks in generative AI socio-technical systems lack scientific rigor for them to be valid and propose a framework rooted in social science measurement practices [76].

Broaching beyond technical evaluations of AI agent capabilities, we contribute an evaluation of commercial AI agents with real people, doing tasks representative of how those agents are marketed.

3 Systematic Review of Marketed Use Cases of AI Agents

To understand the use cases and aspirational ideals that the technology industry has envisioned for "AI agents", we started by taxonomizing the litany of products and services being marketed for immediate or near-term use under the "AI agent" banner. 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.

3.1 Methodology

To construct a taxonomy of use-cases for AI agents, we started by surveying the landscape of AI Agent product websites and articles that aggregated AI Agents for distinct business functions (e.g. marketing and customer support) and customer deliverables (e.g. slide decks, short-form videos, and research). During this process, we found that most agents were task or domain-specific as opposed to truly general purpose, so we isolated unique use cases by deliverable and the verbs used to explain the internal process the agent automated. Next, we performed open, inductive coding and extrapolated specific features and interactions into more abstract behaviors. Throughout the process, we brainstormed umbrella words to cluster individual use cases, until we settled on three atomic, orthogonal capability types, of which every agent showcased at least one.

3.1.1 Search Criteria. To source a broad list of of AI agent products marketed for immediate or near-term consumer use, we employed a multi-pronged strategy.

First, we explored several aggregated lists of AI Agent products such as Google Cloud's 601 Real-World AI Use Cases [65], the AI Agent Directory [13], and Product Hunt [35]. Second, we supplemented these pre-curated lists with our own Google searches under the keywords "AI agents" and "LLM agents". Using this process, we ended up with a list $N = 102$ AI agent products and services to taxonomize. Note that our goal was not to find *all* possible AI agent products and services — as doing so would be intractable — but to find a broad set of such agents such that we could create a taxonomy that captures their general envisioned use cases and aspirational ideals.

3.1.2 Exclusion Criteria. We excluded academic articles and software products intended for users with specialized domain knowledge. For example, our exploratory search of agent products led us to two popular programming agents: Lovable [8] and Cursor [3], which are positioned to different customer segments. Lovable targets non-developers looking for tools in the "no-code/low-code" paradigm by dedicating most of the visual interface to a preview of the website they are creating, similar to *What You See Is What You Get (WYSIWYG)* [38, 68] end-user visual programming tools such as Microsoft FrontPage [25]. In contrast, Cursor is built as a fork of a source code editor, Visual Studio Code, used primarily by software engineers, and dedicates most of its visual space to the viewing of source code with a limited visual preview in the AI sidebar. Since our goal was to identify agents that are targeted towards non-technical lay-users (i.e., the canonical "Johnny" persona used in prior work [77, 84]), we excluded Cursor but included Lovable in our final list.

3.1.3 Limitations. "AI agents" are a rapidly evolving product category. New AI agent products are released for public use everyday, and there is much variance in how AI agents are defined and made distinct from chatbots, products with LLM functionality built-in, and virtual assistants. To that end, we make no claim to have reviewed all AI agent products and services nor that our list is emblematic of all future such products. Instead, we have sourced a broad and diverse list of products that are advertised under the banner of "AI agents". Taxonomizing this list affords us an opportunity to understand the industry-envisioned use-cases and aspirational ideals for AI agents in the near-term and, subsequently, evaluate to what extent real end-users can use these products for those marketed use-cases.

3.2 Analysis

To taxonomize the list of AI agents we had sourced, two authors engaged in an open-coding process.

First, following the process used in prior work to categorize use-cases for "pre-trained models" [62], for each of the AI agent products in our broader list, we read descriptions of customer journeys to identify low-level input-output behaviors and deliverables an end-user could expect. From this analysis, the first author independently created an initial set of codes that described each AI agent use case. A second author independently followed the same steps to create their own codes. Both authors met, discussed conflicts, and agreed on a final codebook. The second author then went and applied this final codebook to the full set of AI agents we sourced. Our codebook is listed in Appendix A.

3.3 Taxonomy of AI Agents

We identified three broad categories of immediate / near-term marketed use-cases for AI agents: **Orchestration**, **Creation**, and **Insight**.

Orchestration agents "act" on behalf of users to manipulate other software interfaces. These agents promise to simplify the procedural and motor tasks required to operate user interfaces. **Insight** agents provide allow people to offload high-level cognitive skills and aim help users with analysis, synthesis, decision-making, and recommendations. **Creation** agents generate structured documents with well-defined formats: e.g., emails, websites, interactive applications, and marketing materials. The key difference between **Creation** and **Insight** is that agents focused on creation focus on formatting and presentation of content, as opposed to generating and synthesizing technical content on behalf of users. We note that these categories are not mutually exclusive, and some AI agents could fall under more than one of these categories. Indeed, complex knowledge work often requires all three: breaking an underlying problem into a solution recipe (insight), navigating computing workflows to execute on it (orchestration), and presenting it in a format tailored to a specific end-user (creation).

3.3.1 ***Orchestration.*** Orchestration agents feature GUI automation to perform direct manipulation actions on the user's behalf. They operate in a loop where they read the state of a computer GUI using vision language models, generate commands with a given user objective via the underlying large language model (LLM), and execute them via a controller that simulates human input. This allows agents to produce actions that interact with real-world state. End-users can use such agents to reduce manual labor and increase reliability in such tasks as data entry, meeting note transcription, and unemployment claim appeals, which can be keyboard / click intensive and prone to human error. Notable examples include Agentforce (Salesforce) [10], LiveX AI Agent (LiveX AI) [7], and Trip Matching AI Agent (Expedia) [4].

3.3.2 ***Creation.*** Creation agents focus broadly on helping users compose structured documents for visual and information communication. Modern digital knowledge work involves creating documents with unique structures and formats, often with an emphasis on visual presentation, including legal briefs, slide decks, and websites. Creation agents use an ensemble of tools, ranging from word processors, spreadsheets, and data visualization platforms to produce their deliverables, which provide support for text formatting, layout management, data representation through tables and charts, and graphic production and editing. Notable examples include Lovable [8], used to create apps and websites without any programming experience and Gamma [5], used to create slide decks from prompts.

3.3.3 ***Insight.*** Agents in this category partner with users to transform unstructured information queries into digestible insights. Insight agents perform online Web searches, query web APIs, access custom knowledge bases in addition to knowledge encoded

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 Website Builder; Den (Workspace)
	Presentations	2	Gamma Presentation Builder; Beautiful.AI
	Images	2	L'Oréal AI Agent (text-to-image)
Insight	Information Retrieval	98	Perplexity AI; Cohesity Gaia
	Recommendations	44	Netflix Recommendation Agent; Spotify AI DJ
	Data Analysis	31	Hex (Analytics Agent); ThoughtSpot Spotter
	Synthesis	17	Fullstory Behavioral Data Agent
	Evaluation	6	Wipro Contract Assistant
	Personalization	2	Mercedes-Benz CLA Conversational Agent

Table 1: Taxonomy of AI agents highlighting umbrella use cases, with capabilities and example agents.

in their pre-training data, and combine the result of an indefinite series of stateful prompt-response loops to respond to a user's request for information. These agents simulate unstructured research, knowledge discovery, and decision-making processes that might otherwise done manually by an end-user. Notable examples include Perplexity's Deep Research [12], Deloitte's Care Finder Agent [51], and Spotify's AI DJ [57].

In terms of example tasks, a recent study of Generative AI Chatbot use amongst students [72] found that "Brainstorming ideas", "Searching for facts and information", "Help with coursework", and "Learning a new skill" were among the top use cases. A common thread between all of these tasks is the cognitive thinking required for humans in these tasks: remembering, understanding, applying, analyzing, synthesizing, and evaluating [19].

Our complete taxonomy is listed in Table 1.

4 User Study

We next sought to address RQ2: i.e., "What challenges do end-users face when attempting to use commercial AI agents for their advertised uses?" To that end, we conducted a study to assess to what extent end-users would be able to use these agents for those marketed use-cases in practice.

4.1 Methodology

We conducted a user study with $N = 31$ participants, comprising a think-aloud usability session of 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].

[1]https://en.wikipedia.org/wiki/OpenAI_Operator
[2]https://en.wikipedia.org/wiki/Manus_(AI_agent)

4.2 The Participants

Participants were recruited through social media outreach on X and LinkedIn, as well as through Prolific [3], a research study participant recruitment platform. Twenty three participants identified as men and 8 identified as women. Twenty participants used Generative AI tools daily or almost daily. Our participants belonged to all of our age groups (ranging from 18-24 to 65+), with 25-34 being the most represented group. Twenty five participants had at least a Bachelor's degree. Participants worked in diverse fields, with Student being the most common occupation. Table 2 provides an enumeration of our participant demographics.

4.3 The Agents

The two agents we chose for the task were Manus [53] and Operator [60]. Note that Operator is no longer accessible by itself and has since been integrated into OpenAI's ChatGPT Agent [59].

We chose these agents because they were, at the time of our study, accessible to the general public and had primarily text-based prompt interfaces similar to tools we expected participants to have some exposure to, such as ChatGPT and Claude. Moreover, they supported all three use-cases in our taxonomy (including the ability to directly control user interfaces for orchestration) and they had state-of-the-art feature sets for AI agents at the time we conducted the study (e.g., the ability to independently use a computer).

4.3.1 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 3) consists of 3 full-height panes. On the left is a list of past task threads that take up about 20% of the screen width. In the middle, occupying about 40% of the screen width is a chat thread that contains user input and agent messages. The remaining 40% of the screen width is occupied by a live preview of the agent's computer, with a dial below that

[3]https://prolific.com

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 2: Demographic Data of User Study Participants

allows you to see the complete replay of the computer's actions, and checklist of task progress.

When a user sends an initial prompt, Manus generally provides a detailed summary of the steps it plans to take. In addition to acknowledging user requests, users are shown a detailed log of every action it performs on its computer along with a text summary. It features interactive chips which link to individual Web search results and states of the computer screen (e.g. when it scrolls on the browser). It always terminates a task with a list of any files it generated, which are clickable buttons that cause the tab on the right to show a preview of the corresponding file (replacing the computer screen preview). The user is always able to send prompts even if it is working on a previous prompt, and the new instructions are seamlessly integrated into the plan. One can take control of the computer to complete actions by clicking the "Take Over" button on the bottom right.

4.3.2 Operator. Operator has capabilities comparable to those of Manus, but with a more minimal interface. Its initial screen, as shown in Figure 4, 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

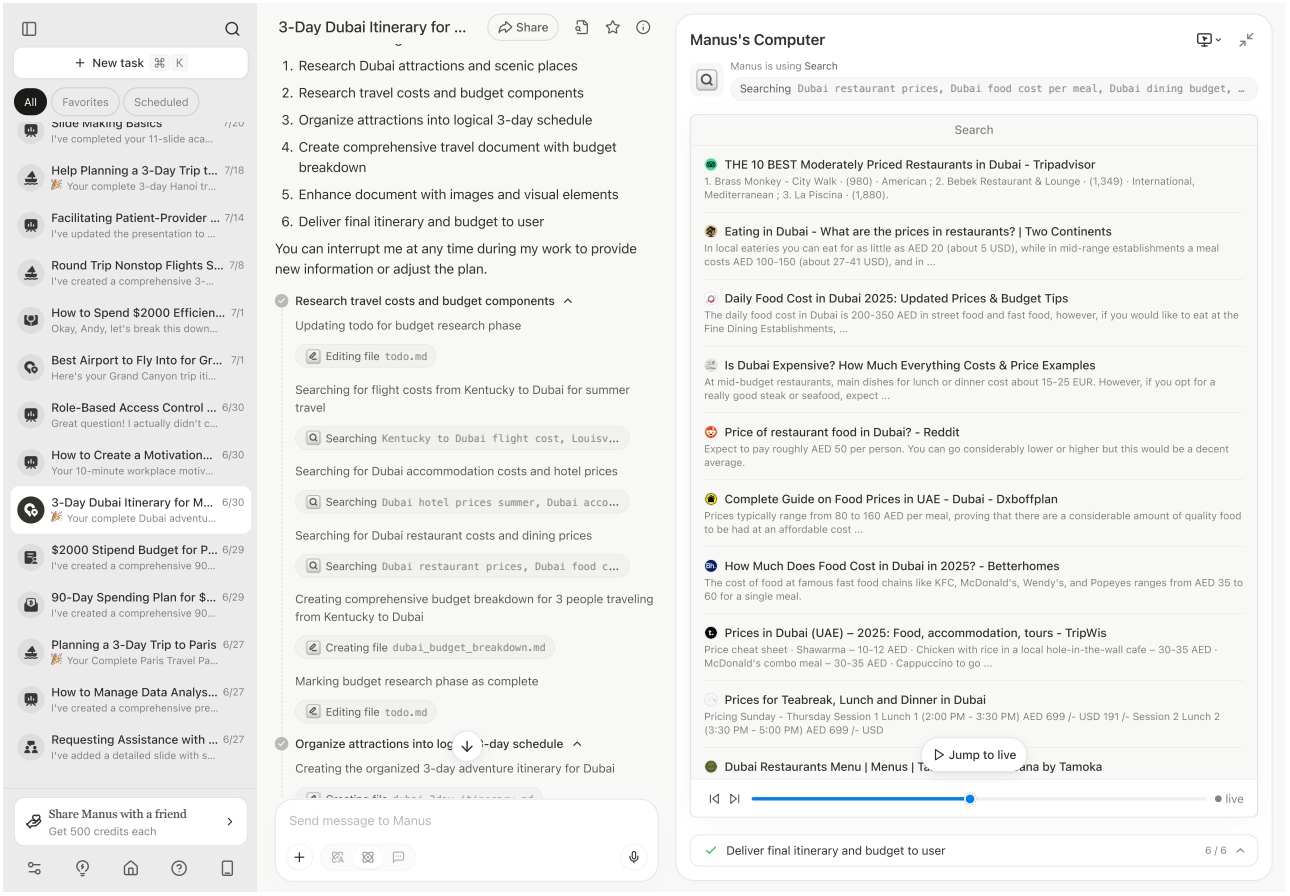

Figure 3: Manus, one of our AI Agents, working on our Holiday Planning task.

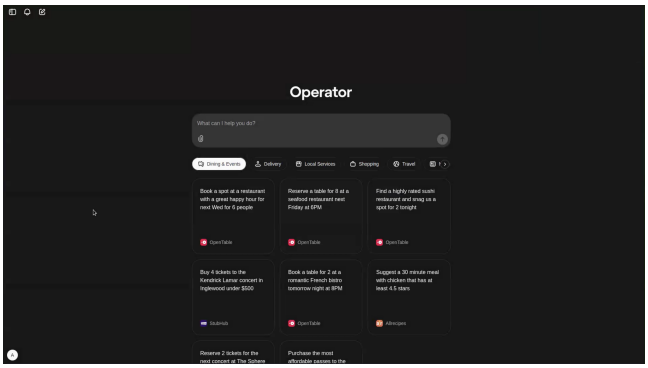

Figure 4: The initial screen on Operator, featuring a text box to enter prompts, and a grid of examples tasks in various categories.

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.

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. Specifying tasks is, of course, a very open-ended activity. To pick useful tasks for our study, we applied the following 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.

Speed. Tasks must be completed relatively quickly. We opted to observe participants working on two tasks across two 20-minute sessions in an hour. This gave users sufficient time to complete each task and 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 did not reflect the care they might place on deliverables whose quality was personally significant.

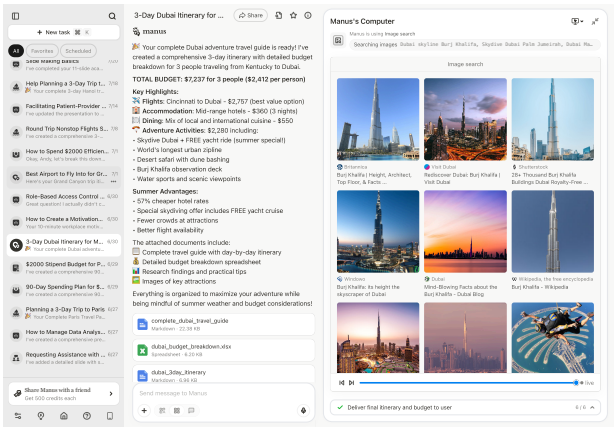

Figure 5: Manus, one of our AI Agents, operating a computer.

4.4.1 ***ORCHESTRATION*** *Task: Holiday Planning.* In this task, participants needed to use 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 3.3.1) in that it requires the agent to complete individual steps by navigating real UIs on an active computer, as can be seen in Figure 5.

First, the user needed to identify the ideal destination for the trip, with or without the help of the AI agent. Second, they needed to choose flights to and from this location, from a list of options that the AI agent found by entering their preferences into and browsing the websites of airlines and flight aggregators. Third, the agent needed to elicit the user's preferences for the trip's activities such as sightseeing and dining, perform a number of searches to retrieve them, and present a personalized list in real time. Finally, the agent needed to combine all of these into a coherent schedule that was suitable for an end-user. The task required planning to consider a number of complex constraints: e.g., seasonal variation in sightseeing attractions, budget constraints, personal preferences for flights, activities, and hotels.

Alignment with Task Criteria. Traveling on holiday is a common activity for people in the United States, where all participants were located during our study. Generally, people project their interests and personal preferences in the locations they choose to travel to and the activities they engage in, which combined with the non-trivial costs of a holiday, raise the emotional stakes of the task. All participants in the pilot study were able to complete this task in 20 minutes.

Agents that claim to perform this type of task must support Web search, Computer Use, or the Model Context Protocol (MCP). This is needed to navigate the appropriate websites and provide the user's travel information to surface relevant flights, housing accommodations, and activity recommendations. Users typically want economical options, which means browsing many websites. Given the number of input data points involved, users are typically unable to provide all the relevant details. The agent must be willing to proactively confirm details or prompt the user if they will affect the itinerary, which requires some memory and autonomy.

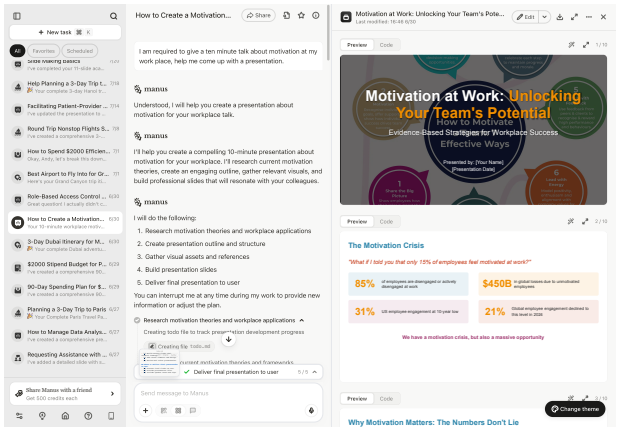

Figure 6: Manus working on the Creation task, Slide Making.

4.4.2 ***CREATION*** *Task: Slide Making.* In this task, portrayed in Figure 6, participants had to prepare a slide deck for a 10-minute presentation on a topic of their choice. It could involve visual design, computer use, or programming, depending on the mechanism used to produce slides.

Alignment with Task Criteria. We selected this task because it had reasonably high *stakes* — we instructed users that the slides produced should be something they would be willing to present themselves, requiring them to think critically about the outputs and if they would be comfortable presenting them. Moreover, the task was *collaborative* in that we expected users to express both stylistic and content preferences, since people can have a strong variation in presentation style. Moreover, preparing and delivering presentations is common in modern digital knowledge work.

Agents that do this task must first be able to research information about the topic of the user's talk to source a suitable quantity of information, and barring that ask the user to provide source content. They must be able to formulate a slide outline that divides the

information being presented into approximately equal chunks that fill the duration of the talk. They must either be able to write code in a visual presentation language such as HTML, Markdown, or ReStructured Text that can be converted to a slide deck, or be able to operate a computer to interact with third-party design software on behalf of the user. Frequently, they must be able to fetch graphics and media. Finally, they must incorporate design principles to lay out the content on each slide while maintaining a consistent theme throughout the presentation. Notable examples of creation-focused agents include Gamma [5], Beautiful AI [1], and Genspark [6].

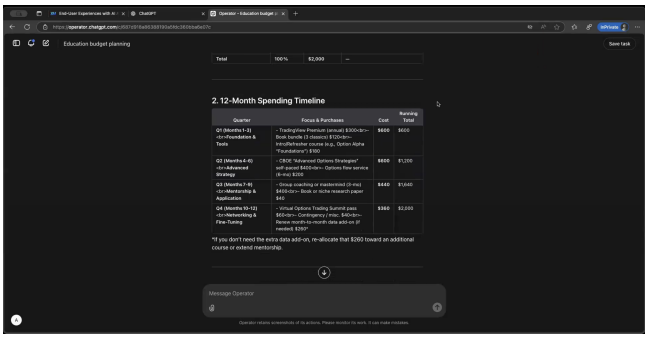

Figure 7: OpenAI Operator, sharing a sample budget with a participant. Participants appreciated its succinct tables.

4.4.3 **INSIGHT** *Task: Professional / Personal Growth Stipend Budgeting.* In this task, the participants used a research assistant agent that would help them make the best use of USD 2,000 to advance their personal or professional development. This task was representative of Insight (Section 3.3.3) tasks in that it required the agent to understand the user's goals and preferences, apply them in performing relevant web searches, analyze each product or activity to see if they were a match and fit the user's budget, and synthesize many instances across numerous websites. Figure 7 illustrates this on Operator.

Alignment with Task Criteria. Although everyone spends money differently, budgeting (and accounting, its institutional analogue) is a task common to people from all walks of life. This task has emotional stakes because it is deeply personal: what each person considers a good use of disposable money is highly dependent on their values, interests, and goals, making a generic budget or lackluster attempt unlikely.

This task requires an AI agent be able to perform research to retrieve products, services, and experiences that would benefit the user's personal and professional goals. It must be able to synthesize information from various websites or their internal memory. Finally, it must be able to determine which combinations of items fall under the user's budget, by incorporating the cost of each item. Since AI Agents only simulate the experience of performing arithmetic calculations, hallucinations can lead to arithmetic errors that can affect the user's trust and reliability. Examples of insight agents are Perplexity [9], a research-focused AI agent that synthesizes information across many Web searches; Cleo [2], which answers financial questions based on spending data; and Beam AI [17], which creates budgets for businesses by consolidating financial data.

4.5 Procedure

The first author conducted all of the interviews, which were 60-minute-long Zoom sessions that were both screen and audio recorded.

For studies conducted on participants who signed up via 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 to be formally entered into the study. Before each interview, the first author randomly chose two tasks, agent tools, and their order for the user, attempting to ensure that all tasks and agent pairs were approximately equally distributed and to eliminate order effects. Table 3 describes the statistics of task-agent combinations across all the interviews.

First, participants were introduced to the study and asked to open a link to the online survey. Before each task, a password manager was used to send temporary links to access each agent. They were read the task description and were given 20 minutes to work on each task. The interviewer clarified task-related questions and communicated intermittently while the participant worked on a task to capture live reactions to events and updates from the agent. In general, the interviewer did not answer questions about the user interface, except for 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 over and above 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. Appendix B contains our interview script.

4.6 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 filled out 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 2. We selected participants to capture a wide breadth 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. The study was approved by an institutional review board (IRB).

4.7 Data & Analysis

Through the study, we collected the following data for analysis:

- Audio and screen recordings of end-users attempting each task
- Online survey responses to a questionnaire with demographic-related questions and responses to the System Usability Scale [20] for each task
- Audio recordings and text transcriptions of the think-aloud session and semi-structured interview

To analyze these data, we again used open inductive coding. Our analytic focus was to answer **RQ2**: i.e., understanding what

Task	Manus	Operator	Total
Holiday Planning	10	10	20
Slide Making	13	10	23
Personal / Professional Development Stipend Budgeting	10	9	19
Total	**33**	**29**	**62**

Table 3: A decomposition of the 62 tasks users completed across 31 sessions.

barriers and usability challenges end-users faced while attempting the tasks we had assigned them to complete with the use of an AI agent. The first and second authors independently coded three interviews, to come up with an initial set of codes. They met weekly and discussed the codes for multiple interviews, merging and deduplicating codes when realizing they were either the same phenomenon or variants that were different experiences altogether. Across these discussions, they authors came up with themes that explained general experiences that we interpreted participants were going through as they navigated agent software. Our codebook can be viewed in Appendix C.

4.8 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.

4.8.1 Agent successes. Out of the three tasks, participants were most successful at Holiday Planning and least successful at Budgeting. Manus excelled at the slide making task, on which Operator performed the worst. Operator was most successful on the Holiday Planning task. Figure 8 presents the mean System Usability Scale (SUS) [20] scores each task and agent received.

Overall, users were impressed and charitable in their perception of both agents' capabilities and usability, even when they could not complete the stipulated task. As P23 stated: *"This is worth it in every sense of the word."* Note that none of our participants had used AI agents prior to our study, so part of this reaction could be attributable to novelty effects. Most users were generally satisfied with their results, found that the agent provided useful information, and praised the agents for their details: "it's very comprehensive and thorough" (P26).

4.8.2 Impressions towards Manus vs. Operator. Each agent had unique strengths and weaknesses that made it better suited for specific tasks.

Manus. All Manus users were impressed by its ability to turn short prompts into detailed and "visually appealing" slide decks. P26 commented that they would not themselves have invested as much effort in slide design: "Compiling this personally would have taken days or weeks."

Yet, in almost all cases, participants were not ready to present the final slide deck, expressing a desire to edit due to concerns with the density of text content and dissatisfaction with the image quality. P31, for example, stated:

> *unless it was like a webinar where I'm giving it over the computer. And everyone else is just watching on their computer screen. That might be acceptable. But I'm thinking, like an actual in person presentation, where these things are gonna be displayed like on a wall, or something behind me.*

P30 shared: *"I don't like the way they just threw the words on top of the images."*

Moreover, a recurring complaint about Manus was the volume of its workflow logs and the speed at which it generated them, which far exceeded the bandwidth of users following along.

For the budgeting and holiday planning tasks, a common request was for links and interactive elements for each recommendation. Without these links, the final deliverable felt incomplete and necessitated additional work from the user to act upon.

Operator. Participants appreciated Operator's minimal user interface and quick responses, which were often presented in formats that were easy to digest. As P19 stated: *"it's very organized in the way it presents data...kind of like a spreadsheet"*. However, participants were less impressed with the slow pace of Computer Use, the frequent errors it encountered, and instances of it hanging without communication. Many users noted that they could perform web searches and actions much more rapidly by controlling their own browser. However, users also observed that speed was not as critical if they could run in the background while they focused on other tasks.

4.8.3 Barrier: Agent capabilities and behaviors don't align with user mental models. AI agents have many unknown capabilities, making it difficult for users to develop effective mental models of what is possible and how to access agent functionality.

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. As P22 stated: *"Well, I'm kind of afraid to ask this a question...Just because it takes so long"*.

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

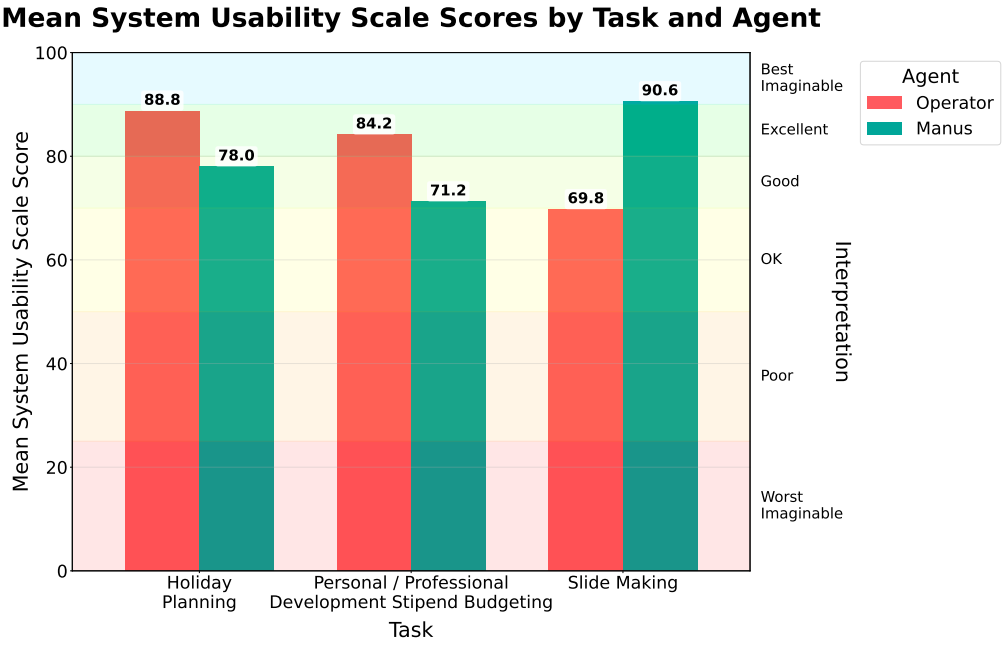

Figure 8: Mean System Usability Scale (SUS) [20] Scores by Task and Agent. The average experience with our chosen agents and tasks was interpreted generally as Good (70-80) and Excellent (80-90), with Slide Making on Operator (69.2, Okay) and on Manus (90.6, Best Imaginable) being notable exceptions.

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. *"You've got to sit there and make sure that your initial prompt is perfect . . . You ask exactly what you're looking for."* (P26) Similarly, P30 echoed: *"If I were more specific, I feel like I would have gotten this on the first try."*

In other cases, users expressed being pleasantly surprised at how much the agent was able to do with relatively little specification on their part: *"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) In these cases, participants were able to gradually update their model of the agent being less like a deterministic tool that required structured hand-holding and more *"as if it were an administrative assistant"* (P9). This confusion — while not stifling — still indicates that the *affordances* of agents were a mystery to users; users' conceptions of what an agent could do 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 expressed confusion about what "Computer use" was and how it worked. Attempting to make slides, Operator repeatedly asked users to use the *Take Over* functionality to sign in to Google Slides, which came across as unexpected and undesirable behavior. One user expected the slide outline and designs to be presented before being asked to log in to a tool. Furthermore, they expected slides to be prepared within Operator, unaware that slide making was not part of the native functionality. In addition, they misunderstood the scope of the *Take Over* functionality, believing it was a request for them to make the slides, as opposed to a request to sign in and return control to the agent. P19 stated of the "Take Over" functionality: "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, and instead wanted the agent to use web APIs for research tasks instead of having the agent control a browser. Computer Use was perceived as being drastically slower than participants' own ability to perform the same searches manually.

Some participants also developed folk models of how interim agent outputs might contribute to an agent's overall speed. For example, noting the large amount of interim progress the agent reported on, P31 wondered if disabling workflow logs would hasten task completion:

> *if I could maybe turn off the you know. Show me what you're thinking part… if that is, you know, causing a slowdown, then I would be interested in having the option of… disabling that to, you know, speed up.*

In short, AI agents are very capable — so much so that the simple, unassuming chat-based interface leaves too much to users' imagination. This results in confusion about agents can do, what they need from the user, how they work, and what the user's role is in the process of completing the task.

4.8.4 Barrier: Agents presume trust without demonstrating competence or security. Several participants expressed a distrust in working with AI agents to perform either specific parts of a task or the task as 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. For example, when reviewing car rental data, P26 stated: *"Car rental Fiat 500 at $10 a day… like, how do they make any money?"*

4.8.5 Barrier: Agents fail to accommodate a diversity of collaboration styles. Many of our participants expressed uncertainty about how to "collaborate" with agents to accomplish tasks. For example, users were uncertain about what they could delegate, how they could exert more 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. P26 shared: "*I would plan manually first, then compare… I still don't trust it completely*". Similarly, P9 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. *"I would call myself like an experienced traveler, so that I don't necessarily need like the currency or 'smiling in public' that kind of thing. So maybe one of the questions for personalizing it could be you know are is this, are you new to traveling? Is it your 1st trip? Will you be with family? Just kind of gauging where you are with travel."*

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 the one shown in Figure 9. However, many users struggled with recovering from these errors.

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

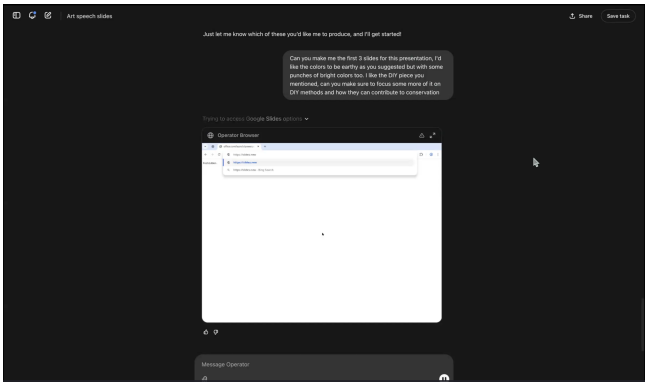

Figure 9: OpenAI Operator, working on the slide making task. Frequently, it was blocked while connecting to online presentation software, potentially due to anti-bot protection.

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 situations required the interviewer's intervention, after many attempts by users to adjust their prompt in anticipation of comprehension difficulties.

4.8.6 Barrier: Agents create an overwhelming amount of communication overhead. Another critical barrier we observed related to the communication overhead agents created: both in users needing to effectively articulate exactly what they wanted out of agents, and in their needing to make sense of the overwhelming amount of output agents produced as they executed tasks at the users' behests.

Users have varying expectations around communication of agent progress. We observed a full spectrum of expectations and preferences when it came to how agents should communicate task progress to users.

Several users (e.g., P16, P24, P31) expressed little interest in the agent communicating intermediate progress: they entirely ignored pages of intermediate logs, stating that their key performance indicators were based solely on the final deliverable. For example, P31 shared: *"I just want the thing done, but not necessarily know how it's done."*

Other users (e.g., P21, P22, P23) appreciated when agents reported on their progress. As P21 noted: *" I always like it when I see the thought process of the AI to know why it chose a certain route."* P23 expressed even more enthusiasm: *"I love to see actually the mind working...I just think that's amazing."* However, while many of these users were *interested* in reading intermediate progress outputs from the agent, they found doing so difficult. These users expressed that key progress milestones were buried within paragraphs of identically formatted text that flew by too quickly, making it hard for them to validate search queries and other GUI actions that agents took.

> *The only thing is, I don't know what some of this stuff means with Manus like where it's saying, "scrolling down" and "clicking element", and I don't know if that's the thing not operating quite right, or if it's just jargon that I just don't, really, I'm not familiar with so. It just looks weird.* –P16

A third subset of users actively disliked the barrage of outputs the agents generated as they completed their tasks, expressing feelings of irritation, overwhelm, and exhaustion. P18, for example, stated: *"It's just non-stop spewing."* Similarly, 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"*. Some participants noted that it was not the sheer volume of output but also that the output was presented in a way that made sensemaking difficult: *"It has the UI that makes it seem as though it's complex. But the information that it's sharing with you isn't so complex."* (P20)

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 and often conditionally intertwined criteria that would be difficult to express into logically sound instructions (e.g., frequent flyer miles, number of stops, time of day). The need to attempt to do so created intense metacognitive demands for users, prompting P11 to speculate about a future brain-computer interface that would eliminate this burden: *"a neural interface... like when typing the prompt, there's often details, I forget whether it's like some preferences or whatnot. So if it can just like directly absorb that to my mind, from like my brain, without me having to type it, then that would just make like a much smoother interface."*

4.8.7 Barrier: Agents lack the metacognitive abilities that enable productive collaboration. Finally, we found that agents lacked the metacognitive abilities to recognize their own limitations and to enable effective oversight by articulating their progress and challenges in terms that 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 the 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 had they taken more initiative in troubleshooting, they would have been able to compensate for the agent's failure to recognize

its limitations, highlighting metacognitive gaps that users must sometimes cover.

Users expressed a desire for agents to express greater metacognitive 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. Andy Grove, long time CEO of Intel, once famously wrote: "Delegation without oversight is abdication." We observed that agents often used tools with which users were unfamiliar, inhibiting users' ability to. We observed this use of unfamiliar tools often in the context of the Slide Making task: agents unexpectedly use programming tools to create slides. P29, for example, was surprised by Manus' decision to use HTML to make slides: *"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, therefore, was pessimistic about success: *"I see a lot of HTML coding and JavaScript. Yeah. So I'm not too confident it's going to come up with what I'm looking for."*

5 Discussion

5.1 How does the tech industry conceive of "AI agents", and can end-users effectively use them?

What exactly are "AI agents", what can they do, and how well can users actually use them for the purposes for which they are advertised? Through a broad survey of the products and services that are being advertised under this banner, we found that the tech industry appears to conceive of AI agents as belonging to one or more of three categories: **Orchestration** agents that operate GUIs on behalf of users, **Insight** agents that summarize and synthesize information for users, and **Creation** agents that convert user intentions into structured documents. Through our user study, we found that the agents we studied are highly capable and appear to already provide value to everyday users — our participants were generally successful at accomplishing their assigned tasks with agents and were impressed by what those agents could do. Both agents demonstrated the ability to formulate plans, make decisions, and deliver on user goals in concert with human users, even if there were rough edges in alignment, control, coordination, and creative work with subjective acceptance criteria. As a result, although our agents were unable to effectively interact with humans as true "partners" with complementary expertise, and instead acted more like talented "lone wolf" task mercenaries, our participants consistently rated the agents as having "Good" to "Excellent" usability on the System Usability Scale.

While this most recent incarnation of "AI agents" bring us closer to realizing the longstanding vision of creating computational thought partners that augment human intellect, our study also identified several critical usability barriers that must be addressed before AI agents can reach their full potential. First, **agent capabilities and behaviors are misaligned with user mental models**. For example, our users were uncertain about how to communicate their intent to agents and lamented being unable to predict what agents would produce based on their initial prompts. Second, **agents presume trust without first establishing credibility**. For example, our users did not feel comfortable delegating some tasks to agents because they did not observe the agent asking the right questions to elicit a user's subjective preferences. Third, **agents fail to accommodate a diversity of collaboration styles**. For example, our users expressed a broad range of preferences for how "proactive" they wanted the agent to be in task execution, but agents did not account for this diversity. Fourth, **agents create an overwhelming amount of communication overhead**. For example, our users had trouble articulating, from scratch, the often complex, conditional, and intertwined preferences they had for tasks. Fifth, **agents lack the metacognitive abilities that enables productive collaboration**. For example, our users noted several instances where agents could not recognize their own limitations and got stuck in repetitive try-fail cycles that could have been easily resolved by users directly.

5.2 Design Implications for Designing End-User AI Agents

The interface agent paradigm challenged many user assumptions internalized through years of experience with primarily direct manipulation interfaces. Our synthesis of end-user usability challenges points us to several implications for designing the next generation of agents, which we outline in this section, and condense in Figure 10.

5.2.1 "Know your user": Agents should get to know the end-user. Each end-user has unique skills, preferences, collaboration styles, epistemological positions (i.e., preferred sources of information) and communication preferences, which are valuable data for agents to be aware of and rely on. Some of these characteristics are stable across tasks, whereas others vary by task. We recommend that they be elicited using recognition-based UIs and ideally in an on-demand fashion to avoid overwhelming the user with a deluge of questions. We hypothesize that collecting these preferences will help build the user trust the agent and feel supported by a deeply personalized experience. In particular, they will address usability barriers around agents' inflexible collaboration styles, communication overhead, and erroneous presumptions of trust. As Horvitz alludes [36], agents must strike a delicate balance between the risk of annoying users by asking too many questions and that of losing the user's trust by making invalid assumptions.

5.2.2 "Know yourself": Agents should have improved "metacognition". The agents we tested had trouble reflecting on the actions they took and the outputs those actions produced. For instance, both Operator and Manus chose to operate a computer for tasks when performing an API-based web search would have been faster; our users also noted that agents would fail to recognize when they encountered errors and fall into unproductive loops. Agents should

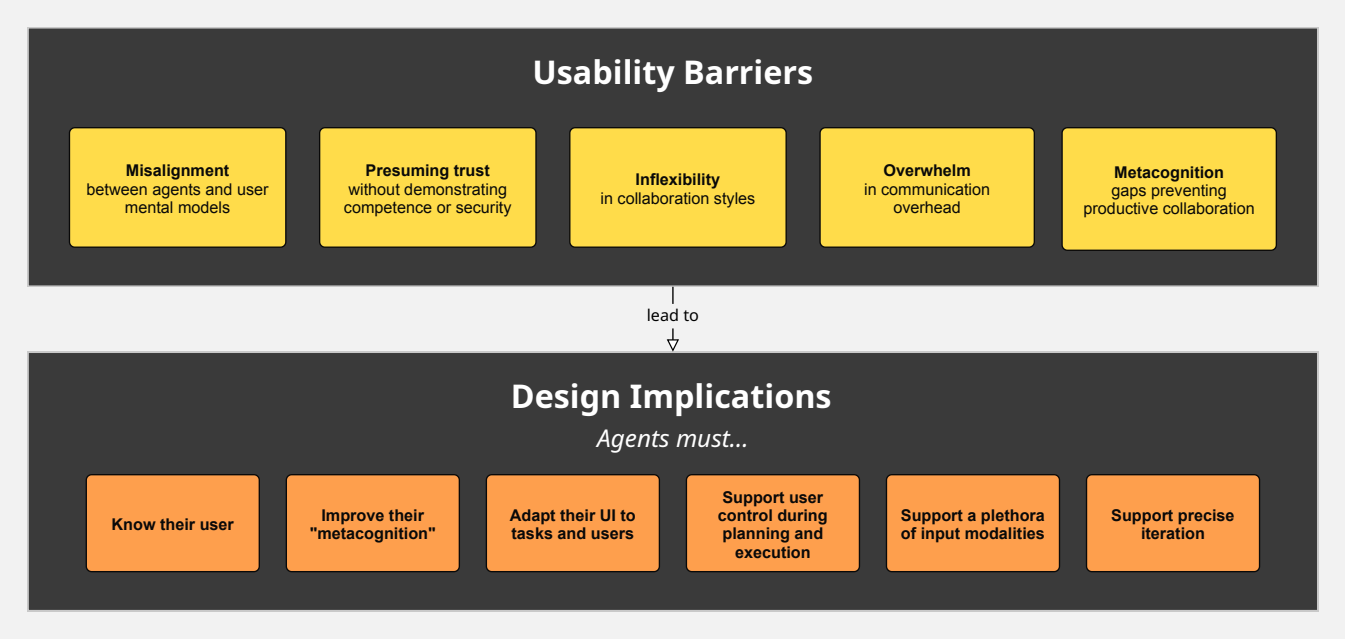

Figure 10: Mapping five empirically discovered usability barriers to six implications for next-generation AI agent design.

be designed with stronger "metacognitive" abilities that help them recognize their own limitations, when they have encountered errors that would be difficult to solve themselves, and when a shift in strategy may be necessary. For example, agents might recognize when the output of a previous action does not align with theirs or the user's expectations, and attempt to diagnose if they should stay the course, change strategies, or ask users for assistance.

5.2.3 "Be adaptable": Agents must adapt their interface based on their knowledge of the task and the end-user. Across all tasks, both agents sent several distinct categories of messages to users: questions to elicit user input, workflow and UI action logs to communicate progress, and links to deliverables for the user to review. Users reported being overwhelmed by the amount of information produced by agents (Section 4.8.6. With knowledge of the *type* of task they are attempting to perform and knowledge of the user's preferences, agents should employ stronger visual hierarchy and principles of progressive disclosure to reduce communication overhead and keep users appropriately abreast of progress without overwhelming them.

For example, **Orchestration** tasks benefit from an increase in screen space dedicated to the computer stream, as the agent performs direct manipulation of third-party software. Insight tasks benefit from rich, interactive elements and citations to information sources, so they can trust the information and organize it more clearly. Finally, creation tasks may have both an "advanced" mode in which users can peruse source code and other implementation details, and a "simple" mode where these details are abstracted away and only the final output is shown.

5.2.4 "Measure twice, cut once": Agents should support user control during both planning and execution. A common challenge faced by users was their inability to control the behavior, collaboration style, task requirements, and many other aspects of an agent's execution plan once they submitted an initial prompt. Coupled with users' unclear mental models of agents, their capabilities, and plan of action, this interaction pattern disconnected the user at a critical moment in their journey from specifying their requirements to obtaining a desired deliverable. Agents presently allow users to "Take Over" only during the execution phase of work, but they should also allow users to "Take Over" during planning.

In the planning phase, the agent might engage with the end-user to develop a shared plan of action, at which point the user would sign off, and *then* the agent would execute. The aforementioned plan would include an enumeration of the steps involved, the tools the agent will use, the expected user actions, and a description of the interim deliverable for upcoming turn of iteration, after which the user potentially provides new directions. Co-creating plans with users could have the added benefit of exposing users to agent capabilities, limitations, and expected outcomes, as well as clarify their role in the process, all of which are needed for smooth collaboration. We expect this to significantly reduce challenges related to user-agent misalignment and support diverse collaboration styles, since the user is explicitly given the opportunity to coauthor agent plans. Critically, it provides an opportunity for the agent to determine if the task cannot be completed, which may occur if it requires capabilities beyond those of both parties [73].

5.2.5 "Show, don't tell": Agents should support a plethora of input modalities, beyond textual prompts. Purely free text-based controls are not commonplace in most software used for knowledge work. They are fragile because they resemble recall-based specification writing without well-known best practices and feedback. In addition, they are error-prone and time-consuming, as corroborated by

several of our study participants. To address these challenges, we propose three recommendations.

Template-based task specification. First, we recommend providing effective prompt templates for well-known tasks. Templates can help scaffold users who are not familiar with prompt engineering and those working on a new task by providing users with tried-and-true recipes that work well. They reduce the likelihood that critical input data is left unspecified, by making omission an explicit decision. This template-based approach aligns with a popular software design philosophy that originated from the Ruby on Rails community: "convention over configuration" — software should aim to reduce the number of decisions users take by making the most common and sensible ones the default.

Recognition over Recall. Second, we recommend that agents support input modalities that prioritize recognition (e.g., selecting from a known list of options) over recall. In addition, to support users working on tasks with multiple serial dependencies, they should surface user input on important dimensions that non-trivially influence a given task's success earlier, especially if these requirements are challenging to incorporate later in the process.

Programming-by-Demonstration. Third, agents could consider supporting programming-by-demonstration (PBD) paradigms [27] for situations where tasks have many intricate requirements and steps that are challenging to explicitly specify. For example, when entering data in legacy software, PBD can help provide agents with the idiosyncratic knowledge needed to efficiently complete them.

5.2.6 "Next time's the charm!": Agents should support precise iteration. Finally, agents should better support iterating on final deliverables. Many users in our study expressed not trusting agents to make small, iterative improvements on the outputs they initially produced. Yet, knowledge work often has subjective and evolving acceptance criteria that makes iteration the rule, not an exception. As a result, agents must simplify the iteration process for the user to maintain trust, empower users [36], and minimize divergence from expectations. Iteration can be better supported by providing the user with context-sensitive controls that they can operate while the agent is executing and after it has produced a deliverable. Providing these controls is especially important to consider when the output has non-textual components (music, slides, and videos come to mind). Examples of such controls can include interface elements that adjust the agent's speed of execution and even pause it, configure the quantity of research based on initial results, skip steps that may now be redundant, and allow the user to provide real-time feedback on the output as it is being constructed.

5.3 Limitations

As with any empirical study of an emerging technology, our study has a number of limitations related to the tasks we developed, the composition of our pool of participants, 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 found in modern digital knowledge work. Next, our group of participants comes exclusively from the US sociocultural context, is young, able, and comprises frequent users of generative AI tools. Finally, we chose the desktop versions of two state-of-the-art 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, finding that they broadly fall into three categories: orchestration, creation, and insight. We next performed a user study where participants attempted tasks in two of three categories on two commercial AI agent products: OpenAI Operator and Manus. While users achieved reasonable success at completing tasks, our study revealed 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. Based on these results, we propose several recommendations to address these barriers. AI agents should: learn user preferences to build trust, understand and account for their own limitations, adapt their interface to the task at-hand, support user control in both the planning and execution phases, allow for non-textual communication of intent, and afford precise iteration over initial outputs.

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] 2025. Beautiful.ai. https://www.beautiful.ai/. Accessed: 2025-09-12.
[2] 2025. Cleo AI. https://web.meetcleo.com/. Accessed: 2025-09-11.
[3] Anysphere 2025. *Cursor - The AI Code Editor.* Anysphere. https://cursor.com/ Accessed: 2025-09-01.
[4] 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.
[5] 2025. Gamma Presentation Builder. https://gamma.app/. Accessed: 2025-09-12.
[6] 2025. Genspark. https://www.genspark.ai/. Accessed: 2025-09-12.
[7] 2025. LiveX AI Agent. https://livex.ai/. Accessed: 2025-09-12.
[8] 2025. *Lovable.* https://lovable.dev/ Accessed: 2025-09-01.
[9] 2025. Perplexity AI. https://www.perplexity.ai/. Accessed: 2025-09-11.
[10] 2025. Salesforce Agentforce. https://www.salesforce.com/agentforce/. Accessed: 2025-09-12.
[11] Eytan Adar. 2018. Bounced Checks at the UI/AI Intersection. In *Human-Computer Interaction Consortium Workshop (HCIC'18).*
[12] Perplexity AI. 2024. *Introducing Perplexity Deep Research.* https://www.perplexity.ai/hub/blog/introducing-perplexity-deep-research Accessed: 2025-09-02.
[13] AI Agents Directory. 2025. AI Agent Marketplace & Directory | Find Top AI Agents & AI Agent Solutions. https://aiagentsdirectory.com/ Accessed: 2025-09-01.
[14] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19).* Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3290605.3300233
[15] 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

[16] 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

[17] Beam.ai. 2025. Budget Generation | AI Agents & Agentic Workflows. https://beam.ai/workflows/budget-generation Accessed: 2025-09-12.

[18] Fabio Bellifemine, Agostino Poggi, and Giovanni Rimassa. 2000. Developing multi-agent systems with JADE. In *International workshop on agent theories, architectures, and languages*. Springer, 89–103.

[19] B. S. Bloom, M. B. Engelhart, E. J. Furst, W. H. Hill, and D. R. Krathwohl. 1956. *Taxonomy of educational objectives. The classification of educational goals. Handbook 1: Cognitive domain.* Longmans Green, New York.

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

[21] 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

[22] Vannevar Bush et al. 1945. As We May Think. *The Atlantic monthly* 176, 1 (1945), 101–108.

[23] 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

[24] 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*.

[25] Microsoft Corporation. 1996. *Microsoft FrontPage*. https://en.wikipedia.org/wiki/Microsoft_FrontPage Web authoring software.

[26] 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

[27] 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.

[28] F. S. de Boer, K. V. Hindriks, W. van der Hoek, and J. J. Ch. Meyer. 2002. Agent Programming with Declarative Goals. arXiv:cs/0207008 [cs.AI] https://arxiv.org/abs/cs/0207008

[29] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. arXiv:1702.08608 [stat.ML] https://arxiv.org/abs/1702.08608

[30] 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

[31] Alpana Dubey, Kumar Abhinav, Sakshi Jain, Veenu Arora, and Asha Puttaveerana. 2020. HACO: A Framework for Developing Human-AI Teaming. In *13th Innovations in Software Engineering Conference (ISEC 2020)*. ACM, Jabalpur, India, 1–9. doi:10.1145/3385032.3385044

[32] Upol Ehsan, Philipp Wintersberger, Q Vera Liao, Elizabeth Anne Watkins, Carina Manger, Hal Daumé III, Andreas Riener, and Mark O Riedl. 2022. Human-Centered Explainable AI (HCXAI): beyond opening the black-box of AI. In *CHI conference on human factors in computing systems extended abstracts*. 1–7.

[33] 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.

[34] John D. Gould, John Conti, and Todd Hovanyecz. 1983. Composing letters with a simulated listening typewriter. *Commun. ACM* 26, 4 (April 1983), 295–308. doi:10.1145/2163.358100

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

[36] 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

[37] 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

[38] Edwin L. Hutchins, James D. Hollan, and Donald A. Norman. 1985. Direct Manipulation Interfaces. In *User Centered System Design: New Perspectives on Human-Computer Interaction*, Donald A. Norman and Stephen W. Draper (Eds.). Lawrence Erlbaum Associates, Hillsdale, NJ, 87–124.

[39] 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

[40] 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

[41] Garry Kasparov. 2010. The Chess Master and the Computer. *The New York Review of Books* (2010). https://www.nybooks.com/articles/2010/02/11/the-chess-master-and-the-computer/

[42] 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

[43] 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

[44] 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)*.

[45] 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

[46] 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

[47] Q Vera Liao and Kush R Varshney. 2021. Human-centered explainable ai (xai): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790* (2021).

[48] J. C. R. Licklider. 1960. Man-Computer Symbiosis. *IRE Transactions on Human Factors in Electronics* HFE-1, 1 (1960), 4–11. doi:10.1109/THFE2.1960.4503259

[49] 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

[50] 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

[51] 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..

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

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

[54] 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.

[55] Tim R. Merritt, Kian Boon Tan, Christopher Ong, Aswin Thomas, Teong Leong Chuah, and Kevin McGee. 2011. Are artificial team-mates scapegoats in computer games. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work* (Hangzhou, China) *(CSCW '11)*. Association for Computing Machinery, New York, NY, USA, 685–688. doi:10.1145/1958824.1958945

[56] 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

[57] Spotify Newsroom. 2023. *Spotify Debuts a New AI DJ, Right in Your Pocket*. https://newsroom.spotify.com/2023-02-22/spotify-debuts-a-new-ai-dj-right-in-your-pocket/ Accessed: 2025-09-12.

[58] 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.

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

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

[61] 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

[62] 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

[63] 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

[64] 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).

[65] 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..

[66] 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

[67] Devansh Saxena, Ji-Youn Jung, Jodi Forlizzi, Kenneth Holstein, and John Zimmerman. 2025. AI Mismatches: Identifying Potential Algorithmic Harms Before AI Development. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–23. doi:10.1145/3706598.3714098

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

[69] Ben Shneiderman. 2022. *Human-centered AI.* Oxford University Press.

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

[71] Yoav Shoham. 1993. Agent-oriented programming. *Artificial Intelligence* 60, 1 (1993), 51–92. doi:10.1016/0004-3702(93)90034-9

[72] Pradyumna Shome and Miuyin Marie Yong Wong. 2024. "Learning Too Much About Me": A User Study on the Security and Privacy of Generative AI Chatbots. In *Proceedings of the Twentieth Symposium on Usable Privacy and Security (SOUPS 2024) — Poster Session*. USENIX Association, Philadelphia, PA, USA. https://www.usenix.org/conference/soups2024/presentation/shome-poster Poster.

[73] Neville A Stanton. 2006. Hierarchical task analysis: Developments, applications, and extensions. *Applied ergonomics* 37, 1 (2006), 55–79.

[74] 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

[75] Blase Ur, Melwyn Pak Yong Ho, Stephen Brawner, Jiyun Lee, Sarah Mennicken, Noah Picard, Diane Schulze, and Michael L Littman. 2016. Trigger-action programming in the wild: An analysis of 200,000 ifttt recipes. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 3227–3231.

[76] 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).

[77] 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.

[78] Wikipedia contributors. 2025. Amazon Alexa. https://en.wikipedia.org/wiki/Amazon_Alexa. Accessed: 2025-09-11.

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

[80] Wikipedia contributors. 2025. Siri. https://en.wikipedia.org/wiki/Siri. Accessed: 2025-09-11.

[81] 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.

[82] 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

[83] 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/

[84] 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

[85] 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

[86] 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

[87] 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

[88] 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 Codebook

Table 4: 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 Script

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** (https://manus.im), a general-purpose AI agent.
- **Operator** (https://operator.chatgpt.com), 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.1>

Interviewer asks the participant questions including and beyond those from Section B.2.

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.1>.

Interviewer asks the participant questions including and beyond those from Section B.2

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.1 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.2 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?

C User Study Codebook

Table 5: 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

Code Name	Definition
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
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.

Code Name	Definition
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 IsDifficult-ToMakeSenseOf	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.
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.

Code Name	**Definition**
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.