

Effects of Before-Use Explanations on Non-Expert Users' Understanding and Task-Planning Strategies in the Use of Agent Skills

MASTER THESIS

Ziyi Qiu

Student Number: 6248314

Design for Interaction

Chair | **Evangelos Niforatos**

Mentor | **Tianhao He**



ABSTRACT

Agent skills provide reusable instruction bundles that guide LLM-based agents through complex workflows. Yet their `SKILL.md` instructions are written primarily for agent execution rather than non-expert users. Users may know that a skill is available without knowing how to begin, what to provide, or when to adjust its behavior. This thesis asks whether a before-use explanation derived from `SKILL.md` and reorganized around users' actions can help bridge this gap.

The research followed a two-phase mixed-methods design. In Phase 1, 15 university students used an agent skill and described what they needed before use based on their overall experience. Their difficulties and preferences guided the design of a concise, user-facing explanation derived from `SKILL.md`. In Phase 2, 24 Master's students took part in a between-subject experiment comparing two pre-task materials: an adapted `SKILL.md` for the control group and the before-use explanation for the experiment group. Both groups used the same agent skill to complete the same design task. The analysis combined reading behavior, measures of understanding and task planning, interaction logs, and interviews.

The experiment group read the material more completely, reported higher understanding, and demonstrated stronger understanding of available workflow patterns and the pattern used during the task. However, the groups did not differ significantly in task-planning strategies. Qualitative findings suggested that the explanation helped participants recognize input requirements, workflow options, defaults, and checking opportunities, while planning also appeared to reflect prior AI-use routines and task uncertainty. These findings suggest that before-use explanations may help not simply by shortening documentation, but by selecting and reorganizing information that users can act on.

Drawing together the explanation design and empirical findings, this thesis proposes a preliminary translation framework for user-facing agent-skill explanations. The framework highlights three forms of translation: making agent-facing language understandable, connecting procedural knowledge to users' tasks, and making decision points and opportunities for intervention visible. From a human-AI interaction design perspective, the findings suggest that improving agent-skill use should begin with an information framework that helps users understand the skill and determine how to act, before focusing on interface-level interaction design or visual presentation.

CONTENTS

Abstract	ii
1 Introduction	1
2 Related Work	3
2.1 Explanations' Influence on User Understanding and Task Planning within LLM-based AI Systems	3
2.2 The Emergence of Agent Skills and Challenges for Non-expert Users	4
2.3 Before-use Explanations as Prior Knowledge for Agent Skill Use	5
3 Study Design	7
4 Methodology	9
4.1 Phase 1: Exploratory Study	9
4.1.1 Phase 1: Participant Recruitment	9
4.1.2 Phase 1: Task Design and Procedure	9
4.1.3 Phase 1: Findings and Implication for Phase 2	10
4.2 Phase 2: Between-subject Experiment	10
4.2.1 Phase 2: Participant Recruitment	11
4.2.2 Phase 2: Before-use Explanation Design	11
4.2.3 Phase 2: Prototype Design	12
4.2.4 Phase 2: Task Design	13
4.3 Phase 2: Measurement	13
4.3.1 Measures for RQ1: Understanding of the Agent Skill	13
4.3.2 Measures for RQ2: Task-planning Strategies When Using the Agent Skill	14
4.4 Phase 2: Experiment Procedure	15
5 Results	16
5.1 Baseline Comparability and Material Engagement	16
5.2 RQ1: Effects on Understanding	18
5.3 RQ2: Effects on Task-Planning Strategies	21
5.4 Exploratory Correlations Between Reading Engagement, Understanding, and Planning	22
6 Discussion	24
6.1 RQ1: Better Understanding Through Concise, Action-Oriented Explanations	24
6.2 RQ2: Task-Planning Strategies: Understanding, Exploration, and Situated Judgment	26

6.3	Design Implications: Translating SKILL .md into User-Facing Explanations	
	27	
6.4	Limitations	29
6.5	Future Work.	30
7	Conclusion	32
	Acknowledgements	34
A	Appendix	35
A.1	Phase 1 participant demographic sheet	35
A.2	Phase 1 introduction process	35
A.3	Phase 1 interview questions.	37
A.4	Phase 1 qualitative analysis and full findings	37
A.5	Phase 2 participant demographic sheet	39
A.6	Phase 2 explanation design	40
A.7	Phase 2 agent skill source, license, and adaptation	42
A.8	Phase 2 design brief.	43
A.9	Phase 2 experimental materials by condition	43
A.10	Phase 2 survey and interview instruments	45
A.11	Additional quantitative results	47

1

INTRODUCTION

Artificial intelligence systems are increasingly becoming productive tools for non-expert users who have limited familiarity and practical experience with AI systems and limited coding-related technical knowledge. Rather than requiring code-related skills, large language models (LLMs) allow these users to state their goals through conversations. As a result, non-expert users can now access AI-supported workflows more easily than before (Anthropic, 2025; Dang et al., 2023; OpenAI, 2025; Strzelecki, 2024).

This broader access has also changed what non-expert users need to understand. One emerging form of LLM-based support is the agent skill: a reusable instruction bundle that guides an agent through specific workflows for certain tasks. Such skills can package procedural knowledge into files such as `SKILL.md` and related resources, allowing agents to follow more structured task procedures than a single prompt would provide (Anthropic, 2026; Li et al., 2026; OpenAI, 2026). For non-expert users, the key challenge is to understand an agent skill's operation logic and use it appropriately.

For these users, agent skills create a translation problem. The information that makes a skill executable for an agent is mainly stored in technical files, scripts, or configuration resources. These materials may be meaningful to agents or to expert users with coding-related knowledge. However, they may not clearly explain to non-expert users what the skill can and cannot do, what input it expects, how the workflow will proceed, or when human intervention may be needed. Therefore, each unfamiliar skill may require non-expert users to rebuild their sense of how to begin, what to provide, and how to steer the interaction again and again.

This translation problem also matters for human oversight during the use of agent skills. If non-expert users have only an insufficient understanding of how a skill works, they may find it harder to plan their interaction, monitor whether the agent is following a suitable path, and decide when to intervene. In higher-risk contexts, limited preparation may also make it harder for users to judge what information is safe to provide or what kinds of tool use require extra oversight (Abadeh et al., 2026; X. Hou et al., 2025; Rado-sevich & Halloran, 2025). Human oversight serves as a broader motivation for this study but is not directly examined; the study first focuses on understanding and task planning

as more fundamental outcomes.

Prior work suggests that explanations can support mental model formation, understanding of system behaviors, and informed decision-making (Lim et al., 2009; Rutjes et al., 2019; Sanchez et al., 2026). Explanations may also shape how users prepare their following actions, like deciding what to ask and choosing whether to follow or ignore the suggestions from AI systems (Cabrera et al., 2023; Das et al., 2023; S. S. Y. Kim et al., 2023). However, existing work has mainly examined explanations of AI outputs, recommendations, or interaction behavior. Less is known about whether information provided before use can help non-expert users understand agent skills and plan how to work with them better. This thesis therefore examines two questions:

RQ1: How do before-use explanations derived from `SKILL.md` influence non-expert users' understanding of agent skills?

RQ2: How do before-use explanations derived from `SKILL.md` influence non-expert users' task-planning strategies when using agent skills?

2

RELATED WORK

2.1. Explanations' Influence on User Understanding and Task Planning within LLM-based AI Systems

LLM-based AI systems increasingly support productivity, creativity, and task completion across different work contexts (Lee & Chung, 2024; Noy & Zhang, 2023; Weber et al., 2024; Zhou & Lee, 2024). LLMs have further lowered the barrier to using such systems. Rather than relying on programming expertise or complex software operations, non-expert users can express their goals and collaborate with the system through natural language (Dang et al., 2023; Kazemitabaar et al., 2023; Liu et al., 2023).

Higher-education students and employees are also increasingly using LLM-based AI systems for learning, writing, research, and other knowledge-intensive tasks (Baig & Yadegaridehkordi, 2024; Sallam, 2023; Strzelecki, 2024). This development forms part of a broader expansion of AI use beyond technical specialists, including non-technical professionals such as users in marketing and human resources (Anthropic, 2025; OpenAI, 2025).

AI literacy offers one starting point for understanding how these non-expert users approach LLM-based systems. It has been defined as a set of competencies for communicating and collaborating with AI, critically evaluating AI, and using AI effectively (Long & Magerko, 2020). Prior work describes AI literacy through four dimensions: knowing and understanding AI, using and applying AI, evaluating and creating AI, and AI ethics (Ng et al., 2021). Within this framework, foundational knowledge for understanding AI serves as prerequisites for other forms of effective use. Prior research also shows that AI literacy can influence the quality of prompt and generated outputs, and higher AI literacy is often associated with more refined and contextually relevant outputs (J. Kim et al., 2025). Although AI literacy is often treated as a relatively stable user attribute, emerging evidence suggests that it can also be shaped during interaction, with explanations playing an important role.

Explanations are reasons or evidence that a system provides for its outputs or pro-

cesses, and they should be meaningful, accurate, and bounded by what the system knows (Phillips et al., 2021). The relationship between explanations and understanding is often discussed through mental model formation. Explanations can help users build mental models of what a system can do, how it behaves, and how its outputs should be interpreted (Rutjes et al., 2019; Sanchez et al., 2026). These mental models can then support more accurate expectations about system behavior (Leichtmann et al., 2023). Lim et al. showed that explanations describing why a system behaved in a particular way improved users' understanding and trust in a context-aware intelligent system (Lim et al., 2009). In LLM-based interaction, explanations may also help users connect their intentions with system responses through clarifying prompt-output relationships, guiding both prompt formulation and result interpretation (Subramonyam et al., 2024; Zamfirescu-Pereira et al., 2023). The value of explanations also depends on timing. Chen et al. found that pre-explanations were especially useful when AI systems showed biased behavior, whereas post-explanations were more favorable when systems were bias-free (C. Chen et al., 2024).

Explanations can also shape task-planning strategies, including how users decide to follow, ignore, or intervene when working with complex systems. In medical consultation contexts, pre-visit question prompt sheets increase patients' question-asking behavior and active participation, suggesting that pre-task information can expand users' perceived opportunities for action (Albada et al., 2012; Dimoska et al., 2008). In search contexts, explanations can help users clarify their own search intent and make more informed navigation choices (Juneja et al., 2024). Similar planning effects appear in human-AI interaction: explanations can shape users' information processing and help them decide how to improve their own decisions or help an AI system understand the task better (Bauer et al., 2023; S. S. Y. Kim et al., 2023; Kulesza et al., 2012). Other work shows that explanations can also affect appropriate reliance on AI-generated next-step suggestions and help users identify system failures in later decision stages (Cabrera et al., 2023; Das et al., 2023; Schoeffer et al., 2024). Overall, prior literature indicates that explanations can shape users' mental models, expectations, and decisions about how to interact with AI systems. Most evidence, however, concerns AI outputs, recommendations, or system behavior during interaction.

2.2. The Emergence of Agent Skills and Challenges for Non-expert Users

Explanations influence user understanding and task-planning strategies, and agent skills provide a context in which explanations might be particularly important. Agent instruction bundles, also known as agent skills, are reusable and versioned bundles that usually include a required `SKILL.md` file together with optional scripts, references, assets, and other supporting resources (OpenAI, 2026). Among these components, `SKILL.md` is particularly important because it often contains the core instructions that shape the skill's purpose, workflow logic, usage guidance, and operational constraints. In practice, users can invoke skills for complex tasks through CLI-based agent tools such as Codex CLI (Codex CLI, 2026), Claude Code (Claude Code, 2026), and Gemini CLI (Gemini CLI, 2026). Some simplified forms of agent skill have also been deployed in web-based agent

platforms (Anthropic, 2026), making them accessible to a broader user group.

Unlike traditional prompt-oriented interaction with LLM-based agents, agent skills provide a more structured way to support complex workflows. By combining core instructions in `SKILL.md` with other optional resources, agent skills allow agents to rely on stored procedural knowledge rather than only a one-off prompt (Anthropic, 2026; OpenAI, 2026). Users may select or call a skill through natural language, but skills may also be invoked automatically when the user's input matches the skill's metadata or intended use case (Anthropic, 2026). As portable file-based bundles, skills can also be shared and reused across tasks or AI systems (OpenAI, 2026). Existing skill repositories on GitHub already contain skills for diverse tasks such as planning, writing, coding, brainstorming, and documentation, each involving different workflows and usage requirements ("anthropics/skills", 2026; "openai/skills", 2026). Therefore, in the context of agent skills, users need to understand more than just AI systems: they need to understand a skill's workflow, input requirements, constraints, and possible human interventions they can implement. However, practical knowledge about a new skill mainly comes from `SKILL.md` and other optional resources, which may be technically meaningful but difficult for non-expert users to understand and use (Chatlatanagulchai et al., 2025; Phillips et al., 2021). This suggests a gap between agent-facing documentation and user-facing explanation.

Besides onboarding, agent skills also create an oversight problem. Because a skill can shape the workflow the agent follows, the files it accesses, the scripts it runs, and the external tools it calls, users who only see the final output may miss the points where human judgment is needed. SkillSafetyBench frames these risks as skill-facing attack surfaces, while work on semantic supply-chain attacks shows that `SKILL.md` files and skill metadata can become channels for harmful instructions (Jin et al., 2026; Saha et al., 2026). This concern becomes sharper when skills interact with local files and external sources, because poorly understood or malicious instructions can lead to unintended tool use, sensitive information disclosure, unsafe file operations, or indirect prompt injection (Abadeh et al., 2026; X. Hou et al., 2025; Radosevich & Halloran, 2025). Research on tool-using agents similarly shows that external tool outputs or untrusted content can redirect agents toward unintended actions (Debenedetti et al., 2024; Zhan et al., 2024). Defense and detection work, such as structured queries or malicious-skill triage, addresses part of this risk at the system level (S. Chen et al., 2024; Y. Hou et al., 2026; Ning et al., 2026). In this setting, agent skills turn AI interaction from single prompts into reusable, file-based workflows with distinct inputs, constraints, and points for human intervention. For non-expert users, the technical documentation and safety-relevant aspects of these workflows can be difficult to interpret before interaction.

2.3. Before-use Explanations as Prior Knowledge for Agent Skill Use

Prior work on explanation timing and instructional support suggests that information provided before task execution may support users' understanding and task planning. Research on explicit instruction timing shows that instructions provided before, during, or after a task can support different learning needs rather than serving a single univer-

sal function (Michaud & Ammar, 2023). Explanations provided before task engagement are mainly associated with preparation, such as expectation setting, mental model formation, and planning. Chen et al. found that before-use explanations can be useful when users may need to anticipate biased or problematic AI behaviors (C. Chen et al., 2024). This timing logic is consistent with Mayer's pre-training principle, which argues that learners benefit from understanding key components before engaging with a complex task ("Pre-training Principle", 2009). Explanations provided during task execution often support coordination, helping users monitor progress and adjust actions in real time (Gammelgård-Larsen et al., 2024; S. S. Y. Kim et al., 2023). Explanations provided after task completion are more associated with reflection, such as evaluating outputs, calibrating trust, and learning for future interactions (Bauer et al., 2023; Lim et al., 2009).

Research on instruction and pre-training further clarifies why before-use explanations may matter. Pre-training interventions can produce consistent learning benefits, but their effects depend on the format of the intervention and its match to the learning outcome (Mesmer-Magnus & Viswesvaran, 2010). Earlier information can help users organize new knowledge, reduce cognitive load, and prepare more effective strategies (O. Chen et al., 2023; Hattie & Donoghue, 2016). Design-related research shares similar insights on human-AI interaction: designers need to treat AI as a design material by understanding its capabilities, limits, and possible effects on design decisions (Yildirim et al., 2022), while unclear capabilities and unpredictable system behavior make interaction difficult to plan in advance (Yang et al., 2020). Practitioner-facing AI guidelines can also support early problem framing and help teams avoid later design failures (Yildirim et al., 2023). Across these studies, such prior knowledge can support users' expectations, mental models, and later task-planning strategies. Its effects depend on both when the information is provided and how it is presented. In current agent-skill use, however, the information needed to build such prior knowledge is mainly contained in SKILL.md and related technical resources. As previous work suggests that these materials can be difficult for non-expert users to interpret and apply (Chatlatanagulchai et al., 2025; Phillips et al., 2021), the available information may not function as effective prior knowledge for the non-expert users who need it.

3

STUDY DESIGN

Prior work has shown that explanations can support users' mental model formation, understanding, and task-related planning (S. S. Y. Kim et al., 2023; Rutjes et al., 2019). However, less attention has been paid to repeated onboarding contexts in which non-expert users must repeatedly adapt to agent skills with distinct workflows and interaction logics, while the only available basis for anticipating how the agent will operate is often the `SKILL.md` file itself, which many non-expert users find difficult to interpret. In these situations, the main question is not only whether users can access a skill, but whether the information contained in `SKILL.md` can be translated into a form that helps them make sense of the skill before use.

Understanding is therefore the first outcome examined in this study. In the context of agent skills, understanding means more than knowing that a skill is available. Users need to know what the skill can do, what input it expects, how its workflow may proceed, and where they may need to guide or check the agent. Because this information is currently mainly communicated through technical skill files, it remains unclear whether non-expert users can form such an understanding before interaction. If a before-use explanation can make these aspects clearer, it may help users approach the skill as a more legible workflow rather than as opaque automation. Therefore, this study first examines how such explanations influence users' understanding of agent skills.

RQ1: How do before-use explanations derived from `SKILL.md` influence non-expert users' understanding of agent skills?

The corresponding hypothesis is:

H1: Before-use explanations derived from `SKILL.md` are expected to improve non-expert users' understanding of agent skills.

The second outcome is task planning. Even when users understand a skill better, they still need to decide how to begin, what information to provide, how to structure

their prompts, and when to monitor or adjust the agent's response. In this sense, task planning concerns whether users can connect the skill's operation logic with their own task and use that understanding to prepare their interaction. Although prior work suggests that explanations and mental models can support task-related planning, it remains unclear whether before-use explanations can help non-expert users plan their interaction with agent skills whose workflows and input requirements vary across tasks. Therefore, this study also examines whether such explanations influence how users prepare and structure their interaction with agent skills.

3

RQ2: How do before-use explanations derived from `SKILL.md` influence non-expert users' task-planning strategies when using agent skills?

The corresponding hypothesis is:

H2: Before-use explanations derived from `SKILL.md` are expected to improve non-expert users' task-planning strategies when using agent skills.

This study used a two-phase mixed-methods design to answer the research questions. The overall study protocol, including both phases, was reviewed and approved by the TU Delft Human Research Ethics Committee (HREC) before participant recruitment and data collection. Phase 1 was an exploratory study that identified users' challenges with agent-skill use and informed the explanation design. Phase 2 was the main between-subject experiment and directly examined the effects of before-use explanations on understanding and task-planning strategies. Human oversight was excluded from the research questions and outcome measures. Although understanding and task planning may support later monitoring and intervention, this study did not assess oversight behavior, appropriate reliance, or safety outcomes.

4

METHODOLOGY

4.1. Phase 1: Exploratory Study

The goal of the Phase 1 study was to identify users' current understanding, task-planning difficulties, before-use explanation needs, and preferred explanation formats. These findings provided the empirical basis for the explanation design used in Phase 2.

4.1.1. Phase 1: Participant Recruitment

A total of 15 higher-education students were recruited, including 13 Master's students and two PhD students (see Appendix A.1, p. 35; Figure A.1, p. 35). All had a low to intermediate level of experience in using LLM-based AI systems. Eight out of fifteen reported prior awareness, current use, or an intention to learn or use agent skills. This group was relevant because the participants regularly encountered AI-supported educational tasks but had limited agent-skill expertise. The sample size followed prior qualitative HCI work on saturation and sample justification (Caine, 2016; Vasileiou et al., 2018).

4.1.2. Phase 1: Task Design and Procedure

The task and agent skill were chosen to be familiar enough for non-expert users, reducing unnecessary cognitive demand. Participants were asked to use the Slides skill with Codex to create presentation slides for a 10-minute Master's course based on a provided research paper.

The overall process was conducted and recorded remotely through Teams. Participants controlled the researcher's laptop to use the skill on Codex and complete the tasks on their own. The research paper (Johansson-Sköldberg et al., 2013) was provided by the researcher in PDF format, and its content was not related to the research focus or research questions, which avoided carryover effects.

Participants first received an introduction to the session and the Codex interface. In the warm-up session, they were shown how to call the Slides skill and how to access SKILL.md. During the task, they were asked to think aloud so the study could capture their thoughts and motivations. After the task, participants completed a semi-structured

interview. The interview outline was based on prior work about mental models (Brachman et al., 2025) and explanation preferences (Cai et al., 2019; Das et al., 2023; Lim et al., 2009; Mucha et al., 2021; van der Waa et al., 2021). The introduction process is provided in Appendix A.2, p. 35 and illustrated in Figure A.2, p. 36, and Figure A.3, p. 36; the interview questions are provided in Appendix A.3, p. 37.

4.1.3. Phase 1: Findings and Implication for Phase 2

Transcripts from recordings and text from interaction logs were coded qualitatively. The coding focused on current understanding and task-planning challenges, task-planning strategies, explanation content needs and format preferences. The following content shows the main findings of the Phase 1 study. Detailed explanations with quotes from interviews are provided in Appendix A.4, p. 37.

Understanding and task-planning challenges

Finding 1: users distinguished agent skills from standard conversational AI interactions.

Finding 2: users experienced a high cognitive burden when starting and felt technically alienated.

Finding 3: users perceived a gap between technical files and user needs.

Task-planning strategies

Finding 4: users often adopted trial-and-error strategies.

Finding 5: users often broke conversational interaction into steps, beginning with a general direction and adding details later.

Explanation content needs

Finding 6: users wanted to know about capability edges and functional boundaries.

Finding 7: users wanted an operational checklist.

Finding 8: users wanted direct and actionable assistance with user input.

Explanation format preferences

Finding 9: users preferred example-based explanations.

Finding 10: users preferred plain-language explanations for non-experts.

Finding 11: users preferred control over information structure, volume, and density.

The findings above were combined with research about instruction design to formulate the Phase 2 before-use explanation based on SKILL.md (see Section 4.2.2).

4.2. Phase 2: Between-subject Experiment

In this phase, participants were asked to interact with an LLM-based prototype and use the provided agent skill to solve a design task. After that, they were asked to complete a survey and an interview. The overall process was conducted and recorded remotely through Zoom, with participants receiving links to the prototype and the survey.

For the between-subject experiment, two conditions were compared:

1. **Control group:** SKILL .md only, provided before participants interacted with the prototype.
2. **Experiment group:** before-use explanation only, provided before participants interacted with the prototype.

Both groups completed the same task using the same prototype. The only intended difference was the information provided before interaction.

4.2.1. Phase 2: Participant Recruitment

A total of 24 Master's students from design-related programs were recruited (see Appendix A.5, p. 39; Figure A.4, p. 40). All had used LLM-based AI systems, with 13 reporting at least basic familiarity with agent skills and the remaining 11 reporting low familiarity. Therefore, their backgrounds matched the design task while retaining the non-expert focus of the study. The sample size was informed by prior work on sample sizes in CHI research (Ortloff et al., 2025). Their attitudes toward AI were collected through a survey based on previous research (Sindermann et al., 2021) (see Appendix A.10, p. 45).

4.2.2. Phase 2: Before-use Explanation Design

The design of the before-use explanation was informed by procedural instruction design. Some researchers (Van Der Meij & Gellevis, 2004) argue that procedures should include goals, prerequisite states, unwanted states, and actions and reactions. Other work emphasizes simple conceptual structure before details (Eirisdottir & Catrambone, 2011), task-oriented instruction (Mayer, 2008), and minimalist instruction (Carroll, 1990; Farkas, 1999).

Inspirations from Procedural Instruction Design

1. For content selection, the explanation included skill capability, prerequisites for operating the skill (i.e. user input), unwanted situation warnings, action and reaction guidance.
2. For structure, it moved from general understanding to action: goal, use context, and working logic first; user input and guidance next; then warnings and intervention guidance.
3. For presentation style, it used concise, plain-language, example-based, task-oriented wording rather than a direct translation of SKILL .md. For example, the SKILL .md's description for *Business Rule Variations Pattern*, "different rules = different stories", should be translated as "for rules, regulations, or condition-centered user stories" in the explanation.

Together with the Phase 1 findings (see Section 4.1.3), the final explanation was organized as follows (the full explanation contained 513 words in total and details are provided in Appendix A.6, p. 40):

Explanation Outline

1. **What This Skill Does** clarified the skill's purpose and capability boundaries.
2. **How It Works** reduced the workflow to checking the idea, choosing a pattern, and generating user stories.
3. **What You Should Provide to Start** converted input requirements into a checklist and prompt example.
4. **Other Tips On How To Improve The Output** described expected outputs, unsuitable situations, and possible corrective actions.

4.2.3. Phase 2: Prototype Design

To simulate the process of using a skill, the prototype represented a situation in which users call a skill, the system reads the markdown file, and the agent follows the guidance in that file (OpenAI, 2026). The prototype was implemented as a lightweight web application using Streamlit. To support remote participation, the local prototype was exposed through an ngrok tunnel, allowing participants to access the same interface through a browser link during the experiment sessions. The AI backend used a fixed, frozen model version, claude-sonnet-4-5-20250929, throughout the experiment, and the same model configuration was applied across conditions so that differences between groups would mainly come from the provided SKILL.md or before-use explanation rather than from changes in the model itself.

Because the way explanations are represented in an interface can influence users' decision-making (Mucha et al., 2021), the prototype interface was simplified to include only chat-like components and to reduce interface-driven effects (see Figure 4.1, p. 12).

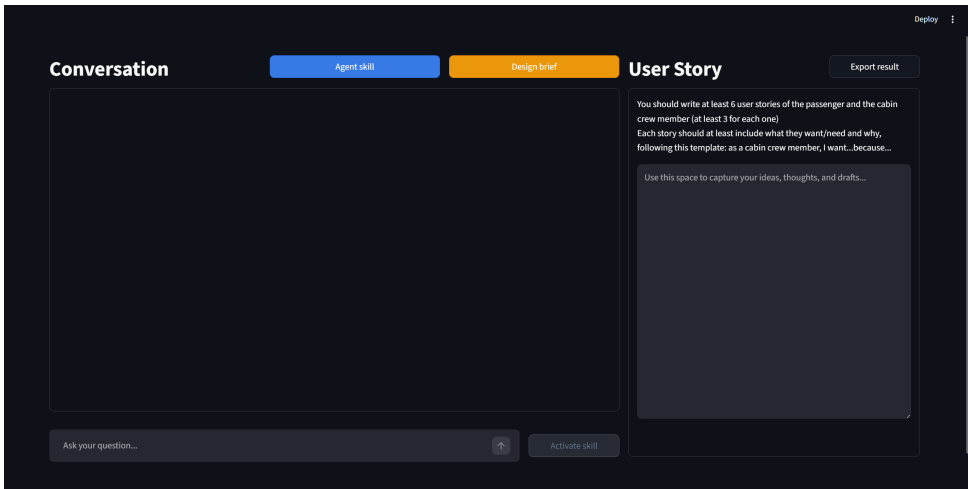


Figure 4.1: Phase 2 prototype interface. The screenshot shows the experimental workspace, including the Conversation panel, Agent skill and Design brief controls, and User Story panel.

4.2.4. Phase 2: Task Design

The task context was grounded in a real-world airline industry challenge, where in-flight cabin waste is generated as a mixed stream and must be managed under regulatory constraints (Consolidated TEXT, 2009; Sustainable Cabin, 2019). This context required rule-based handling rather than simple material-based sorting.

The skill used in Phase 2 was adapted from the open-source agent skill *Epic-breakdown-advisor*, licensed under Attribution-NonCommercial-ShareAlike 4.0 International, which permits adaptation for academic research purposes. The agent skill was selected because it provides different patterns for generating implementable product user stories, which suited the task context. Pattern 3, Business Rule Variations, was treated as the primary reference strategy. Pattern 4, Data Variations, was treated as a secondary or partially appropriate strategy. Pattern 1, Workflow Steps, was set as the default execution logic (the adapted SKILL.md contained 1,920 words in total; see Appendix A.7, p. 42 for more details).

The final design task was summarized as “use the Epic-breakdown-advisor skill to break down an airline in-flight waste separation design idea into valuable user stories that prioritize the regulation of separate ICW & non-ICW” (see Appendix A.8, p. 43 for full-text design brief). The task brief provided the necessary domain context, so participants were not expected to have prior expertise in aviation waste regulation.

4.3. Phase 2: Measurement

To address both research questions, Phase 2 combined self-reported data analysis and observed behavioral data analysis. Data from planning statements, interaction logs (reading duration, reading completeness of the provided SKILL.md or explanation, and participant prompts), post-task surveys, and post-task interviews were used for both qualitative and quantitative analysis. During data collection, reading duration was calculated in seconds, and reading completeness was scored as follows: 0 = not read seriously / mostly skipped, 1 = partially read / quick scan, 2 = complete reading / main content covered.

Baseline comparability between the two groups was also assessed using participants' pre-experiment responses on AI attitudes, AI familiarity, and agent-skill familiarity (see Appendix A.10, p. 45). AI attitude was measured using five items from an existing survey (Sindermann et al., 2021): items on trust and perceived benefit were summed as Accept AI, while items on fear, human destruction, and job loss were summed as Fear AI. AI familiarity and agent-skill familiarity were each measured on a 1–5 self-reported scale.

4.3.1. Measures for RQ1: Understanding of the Agent Skill

Self-reported Understanding: Post-task Survey & Post-task Interview

Self-reported understanding was measured through the post-task survey. This captured how participants reported their understanding of the agent skill. The scale developed by Hoffman et al. (2023) was used to measure users' self-reported understanding of the agent skill based on the provided material (SKILL.md or explanation). The original term “explanation” in the scale was intentionally rephrased as “information I received”, since explanation was not provided for the control group (see Appendix A.10).

An interview was created to measure self-reported understanding qualitatively (see Appendix A.10, p. 45), providing additional support for the self-reported results.

Actual Understanding: Interaction Log & Post-task Interview

Prior work suggests that users' mental models of AI systems include their understanding of system actions and behavior (Brachman et al., 2025), and these mental models shape how users use, rely on, trust, or reject AI systems (Sanchez et al., 2026).

Actual understanding was based on two components of the post-task interview: whether participants noticed that different patterns were available, and whether they correctly identified the pattern used in their interaction and explained why it was used, with interaction logs further checking which pattern had actually been followed. One point was awarded for each correct component. The resulting score ranges from 0, neither component correct, to 2, both components correct. The score was only used to compare the groups on these two components to test if their understanding levels were different. The results do not necessarily represent a high or comprehensive level of agent-skill understanding (even if some of them were scored 2).

4.3.2. Measures for RQ2: Task-planning Strategies When Using the Agent Skill

A coding rubric adapted from Gollwitzer's implementation intention framework (Gollwitzer & Sheeran, 2006) was developed to quantitatively evaluate participants' task-planning strategies using the following principles:

Did the participant define a clear goal for using the agent skill, beyond simply restating the task or asking AI to generate user stories? 0/1/2

Did the participant mention task-related or skill-related situational cues? 0/1/2

Did the participant describe a concrete action or action chain for using, guiding, checking, or refining the agent skill response? 0/1/2

Did the participant link a specific situation, problem, or skill behavior to a specific planned action or intervention? 0/1/2

Self-reported Task-planning: Planning Statement

Before interacting with the prototype, participants were asked to verbally explain how they planned to use the agent skill for the task. This planning statement was coded using the coding rubric above, and the resulting score was used as the measure of participants' self-reported task-planning strategy.

Actual Task-planning: Interaction Logs

Participants' prompts and interaction logs were also collected to analyze their actual planning. This measure was not used to judge whether participants followed their initial plans exactly, but to examine how planning strategies were enacted during interaction. To keep the coding consistent with the planning statement, the same four-dimension rubric was applied.

4.4. Phase 2: Experiment Procedure

The full session lasted approximately 60 minutes, with the approximate duration of each step reported in parentheses.

1. **Consent form** (approximately 2 minutes): participants were first asked to read and sign the consent form.
2. **Experiment introduction** (approximately 3 minutes): participants were informed about the overall process and what they would do in each session.
3. **Design brief** (approximately 5 minutes): participants were informed about the design context, current idea, and their task.
4. **Warm-up session** (approximately 5 minutes): participants were introduced to the prototype interface and condition materials (see Appendix A.9, p. 43; Figure A.5, p. 44; Figure A.6, p. 44; Figure A.7, p. 45) and asked to read the SKILL.md (control group only) or explanation (experiment group only). Their reading duration and completeness of the materials were also collected.
5. **Task-completing session** (approximately 30 minutes): before interacting with the prototype, participants were asked to explain how they would use it for the task (referred to later as the planning statement), including anticipated prompts and their planned process.
6. **Post-task survey session** (approximately 5 minutes): after the task was completed, participants were asked to finish the survey on their understanding of the skill-use process (see Appendix A.10, p. 45).
7. **Post-task interview session** (approximately 10 minutes): participants engaged in a semi-structured interview using the questions listed in Appendix A.10, p. 45.

All recordings, transcripts, survey responses, and interaction logs were anonymized before analysis, and participants were referred to only by participant IDs in the thesis.

5

RESULTS

Overall, the clearest effect of the before-use explanation was observed in participants' understanding of the agent skill. The experiment group showed higher self-reported understanding (median = 26.00 vs. 23.00, exact $p = .024$) and actual understanding (mean rank = 16.25 vs. 8.75, exact $p = .008$). By contrast, group differences in self-reported and actual task planning were not statistically significant (exact $p = .242$ and $.478$, respectively), so the results supported H1 but not H2 at the overall score level. Participants in the experiment group did not spend significantly longer reading the material (median = 238.50 vs. 290.50 seconds, exact $p = .478$), but they showed higher reading completeness (mean rank = 15.50 vs. 9.50, exact $p = .039$).

Before conducting the group comparisons, all variables used in the independent-group analyses were first examined using Shapiro–Wilk normality checks and boxplot-based inspection of distributional shape. Because several variables showed non-normality and many measures were ordinal or discrete, Mann–Whitney U tests were used for group comparisons. Exact two-tailed p values for these comparisons were reported based on SPSS output. For the same reason, exploratory correlations were assessed using Spearman's rank-order correlations rather than Pearson correlations. Each group included $n = 12$ participants. The effect size r was calculated as $|Z|/\sqrt{N}$. Medians and IQRs were reported descriptively for all variables. Boxplots were used to assess whether the group distributions had broadly similar spread and shape. Mann–Whitney U results were interpreted as differences in median location only when the boxplots indicated broadly comparable distributions; when they indicated differences in spread, skewness, or concentration at the floor or ceiling, results were interpreted as rank-distribution differences and reported using mean ranks. Item-level Mann–Whitney tests were treated as exploratory follow-up analyses. Additional quantitative outputs are provided in Appendix A.11, p. 47 (see Figure A.8, p. 47; Figure A.9, p. 48).

5.1. Baseline Comparability and Material Engagement

At baseline, the groups did not differ significantly in AI attitudes or prior AI-related experience (all exact $p \geq .410$, $r = .03$ to $.18$). Participants in the experiment group

showed higher reading completeness (mean rank = 15.50 vs. 9.50, $U = 36.00$, exact $p = .039$, $r = .47$), although they did not spend significantly longer reading the material (mean rank = 11.46 vs. 13.54, $U = 59.50$, exact $p = .478$, $r = .15$). The baseline comparability table (see Table 5.1, p. 17) shows no statistically significant pre-existing differences between the control group and the experiment group in Fear AI, Accept AI, AI familiarity, or agent skill familiarity (all exact $p \geq .410$, $r = .03$ to .18).

Baseline comparability between groups

Measure	Control Median (IQR)	Control Mean Rank	Experiment Median (IQR)	Experiment Mean Rank	U	p	r
FearAI	13.50 (12.00-15.50)	13.71	13.00 (9.50-15.50)	11.29	57.50	.410	.17
AcceptAI	12.50 (10.00-14.25)	11.38	13.00 (12.00-14.25)	13.63	58.50	.443	.16
AI familiarity	3.00 (3.00-3.25)	13.63	3.00 (2.00-3.25)	11.38	58.50	.443	.18
Agent skill familiarity	2.00 (1.00-2.00)	12.33	1.50 (1.00-3.00)	12.67	70.00	.932	.03

Values are median (IQR) and mean rank. Group comparisons use two-tailed exact Mann-Whitney U tests; r is the effect size.

Table 5.1: Baseline comparability between groups*. The table reports medians (IQRs), mean ranks, U , exact p , and r values for Fear AI, Accept AI, AI familiarity, and agent skill familiarity in the control and experiment groups. None of the four baseline comparisons was significant, indicating that the two groups were broadly comparable (all exact $p \geq .410$; $r = .03$ to .18).

Reading duration was descriptively longer in the control group (median = 290.50 seconds, IQR = 167.00–445.50) than in the experiment group (median = 238.50 seconds, IQR = 191.50–270.25), although the control-group distribution was more dispersed and had a longer upper tail. The rank difference was not statistically significant (mean rank = 13.54 vs. 11.46, $U = 59.50$, exact $p = .478$, $r = .15$; see Table 5.2, p. 17).

Reading duration by group

Measure	Control Median (IQR)	Control Mean Rank	Experiment Median (IQR)	Experiment Mean Rank	U	p	r
Reading duration (s)	290.50 (167.00-445.50)	13.54	238.50 (191.50-270.25)	11.46	59.50	.478	.15

Values are median (IQR) and mean rank. Group comparisons use two-tailed exact Mann-Whitney U tests; r is the effect size.

Table 5.2: Reading duration by group*. The table reports medians (IQRs), mean ranks, U , exact p , and r values for reading duration in seconds in the control and experiment groups. The control group had a higher median (290.50 vs. 238.50 seconds) and mean rank (13.54 vs. 11.46), but the difference was not significant ($U = 59.50$, exact $p = .478$, $r = .15$).

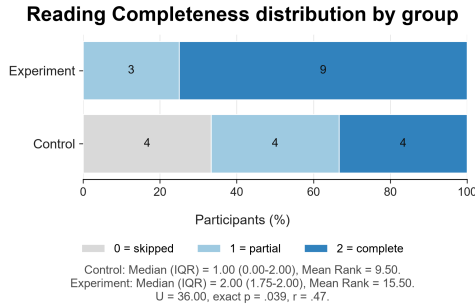


Figure 5.1: Reading completeness distribution by group*. The stacked bars show counts for the control and experiment groups at each reading-completeness level: 0 = skipped, 1 = partial, and 2 = complete. The experiment group had 9 complete and 3 partial readers, whereas the control group had 4 complete, 4 partial, and 4 skipped readers; mean ranks were 15.50 and 9.50, respectively ($U = 36.00$, exact $p = .039$, $r = .47$).

5

Reading completeness was significantly higher in the experiment group than in the control group (mean rank = 15.50 vs. 9.50, $U = 36.00$, exact $p = .039$, $r = .47$; see Figure 5.1, p. 18). Nine participants in the experiment group completely read the 513-word explanation, compared with four participants in the control group who completely read the 1,920-word adapted SKILL.md. Although the explanation was much shorter than the adapted SKILL.md, three experiment-group participants still read it only partially. Among complete readers, reading duration was descriptively longer in the control group than in the experiment group (mean = 526.50 vs. 271.33 seconds; $n = 4$ and $n = 9$, respectively).

Qualitative records showed a similar difference in how participants used the material. In the experiment group, P1 first scanned the headings and then focused on *How It Works* and *What You Should Provide to Start*, while P13 and P20 used the explanation to identify what information to provide, how to structure the prompt, and how the default workflow or pattern options might guide use. In the control group, P10 reported being “a little bit overwhelmed by all of the information” and found the splitting patterns difficult to use for the task, while P22 reported being unsure what instruction to use after partly reading the SKILL.md material. These accounts were consistent with the reading-completeness result.

5.2. RQ1: Effects on Understanding

The experiment group showed higher self-reported understanding (median = 26.00 vs. 23.00, $U = 33.50$, exact $p = .024$, $r = .46$) and actual understanding (mean rank = 16.25 vs. 8.75, $U = 27.00$, exact $p = .008$, $r = .58$) than the control group. As summarized in Table 5.3, p. 19, self-reported understanding was higher in the experiment group (median = 26.00, IQR = 24.00–30.25) than in the control group (median = 23.00, IQR = 20.75–25.25), and the difference was statistically significant (mean rank = 15.71 vs. 9.29, $U = 33.50$, exact $p = .024$, $r = .46$).

Exploratory item-level analyses showed which aspects of understanding differed between groups. Significant group differences appeared on Post-items 1, 2, 5, and 6, which

Self-reported understanding and Actual understanding by group

Measure	Control Median (IQR)	Control Mean Rank	Experiment Median (IQR)	Experiment Mean Rank	U	p	r
Self-reported understanding	23.00 (20.75-25.25)	9.29	26.00 (24.00-30.25)	15.71	33.50	.024	.46
Actual understanding	1.00 (0.00-1.00)	8.75	1.50 (1.00-2.00)	16.25	27.00	.008	.58

Values are median (IQR) and mean rank. Group comparisons use two-tailed exact Mann-Whitney U tests; r is the effect size.

Table 5.3: Self-reported understanding and actual understanding by group*. The table reports medians (IQRs), mean ranks, U , exact p , and r values for the two understanding measures in the control and experiment groups. The experiment group had higher self-reported understanding (median = 26.00 vs. 23.00, $U = 33.50$, exact $p = .024$, $r = .46$) and actual understanding (mean rank = 16.25 vs. 8.75, $U = 27.00$, exact $p = .008$, $r = .58$).

concern knowing how the skill works, satisfaction with the explanation, knowing how to use the skill, and usefulness for the participant's goals (exact $p = .017$, $.039$, $< .001$, and $.008$, respectively). Item 5 showed the strongest item-level effect ($r = .71$). Full item-level medians (IQRs), mean ranks, U , exact p , and r values are reported in Table 5.4, p. 20. Post-items 3, 4, and 7, which concern information detail, information completeness, and perceived agent-skill accuracy, did not show significant group differences (exact $p = .242$, $.630$, and $.266$, respectively). Overall, the experiment group scored higher on items about making sense of and using the skill, rather than on items about whether the information felt detailed or complete.

Actual understanding was also higher in the experiment group than in the control group (mean rank = 16.25 vs. 8.75, $U = 27.00$, exact $p = .008$, $r = .58$; see Figure 5.2, p. 20). A score of 2 indicates correct understanding of both the availability of different patterns and the pattern used, including the reason for its use. The categorical distributions visibly differed between groups. In the control group, five participants scored 0, six scored 1, and one scored 2; in the experiment group, no participant scored 0, six scored 1, and six scored 2. This difference across the two coded components, together with the self-reported understanding result, supported H1.

Interview data illustrated this difference. Several experiment-group participants described the skill in terms of inputs, workflow logic, pattern choices, or default behavior. P13 and P20 connected input structure, default behavior, and pattern choice to how they would use the skill, while P7 and P16 highlighted examples and start-up guidance as useful for understanding. Control-group participants more often completed the task while remaining uncertain about how the skill worked: P10 did not know what the splitting patterns meant, P12 reported reading only the purpose and early steps, and P22 was unsure which instruction to use. This contrast was not absolute, since P24 in the control group noticed that different patterns were available.

Item-level self-reported understanding by group

Measure	Control Median (IQR)	Control Mean Rank	Experiment Median (IQR)	Experiment Mean Rank	U	p	r
Post-item 1 (how the skill works)	2.50 (2.00-3.00)	9.08	4.00 (3.00-4.00)	15.92	31.00	.017	.51
Post-item 2 (satisfaction with information)	2.50 (2.00-3.25)	9.50	4.00 (3.00-4.00)	15.50	36.00	.039	.45
Post-item 3 (information detail)	4.50 (3.75-5.00)	14.25	4.00 (3.00-4.25)	10.75	51.00	.242	.26
Post-item 4 (information completeness)	4.00 (4.00-5.00)	13.21	4.00 (3.75-4.25)	11.79	63.50	.630	.11
Post-item 5 (how to use the skill)	3.00 (2.00-3.00)	7.67	4.50 (4.00-5.00)	17.33	14.00	< .001	.71
Post-item 6 (usefulness for goals)	3.50 (3.00-4.00)	8.75	4.00 (4.00-5.00)	16.25	27.00	.008	.57
Post-item 7 (perceived agent-skill accuracy)	3.00 (2.00-3.25)	10.83	3.00 (3.00-4.00)	14.17	52.00	.266	.25

Values are median (IQR) and mean rank. Group comparisons use two-tailed exact Mann-Whitney U tests; r is the effect size.

Table 5.4: Item-level self-reported understanding by group*. The table reports medians (IQRs), mean ranks, *U*, exact *p*, and *r* values for all seven post-task items in the control and experiment groups. Items 1, 2, 5, and 6 differed between groups (exact *p* = .017, .039, < .001, and .008, respectively); item 5 had the largest effect (*r* = .71). Items 3, 4, and 7 did not differ (exact *p* = .242, .630, and .266, respectively).

Actual understanding distribution by group

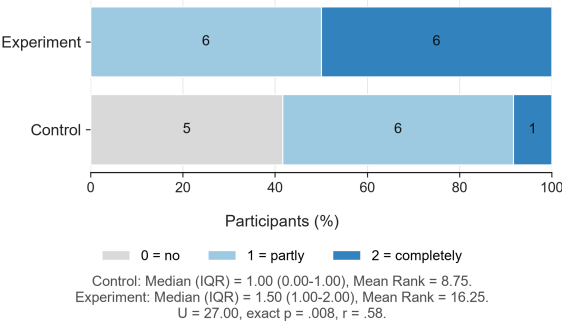


Figure 5.2: Actual understanding distribution by group*. The stacked bars show counts for the control and experiment groups at each coded level: 0 = neither component correct, 1 = one component correct, and 2 = both components correct. The experiment group had 0, 6, and 6 participants at levels 0, 1, and 2, compared with 5, 6, and 1 in the control group; mean ranks were 16.25 and 8.75, respectively (*U* = 27.00, exact *p* = .008, *r* = .58).

5.3. RQ2: Effects on Task-Planning Strategies

The experiment group had higher self-reported task-planning scores (median = 7.00 vs. 4.00, $U = 51.00$, exact $p = .242$, $r = .25$) and actual task-planning scores (median = 7.50 vs. 5.50, $U = 59.50$, exact $p = .478$, $r = .16$) than the control group, but neither difference was statistically significant. The task-planning distributions were broad and partly overlapping.

For self-reported task planning, the experiment group had a higher median score (median = 7.00, IQR = 4.00–7.25) than the control group (median = 4.00, IQR = 3.75–6.25), but the difference was not statistically significant (mean rank = 14.25 vs. 10.75, $U = 51.00$, exact $p = .242$, $r = .25$; see Table 5.5, p. 21). For actual task planning, the experiment group again had a higher median score (median = 7.50, IQR = 4.00–8.00) than the control group (median = 5.50, IQR = 4.00–8.00), but this difference was also not statistically significant (mean rank = 13.54 vs. 11.46, $U = 59.50$, exact $p = .478$, $r = .16$). Therefore, although both results were directionally consistent with H2, the quantitative evidence did not support H2 at the overall score level.

Self-reported task-planning and Actual task-planning by group

Measure	Control Median (IQR)	Control Mean Rank	Experiment Median (IQR)	Experiment Mean Rank	U	p	r
Self-reported task planning	4.00 (3.75-6.25)	10.75	7.00 (4.00-7.25)	14.25	51.00	.242	.25
Actual task planning	5.50 (4.00-8.00)	11.46	7.50 (4.00-8.00)	13.54	59.50	.478	.16

Values are median (IQR) and mean rank. Group comparisons use two-tailed exact Mann-Whitney U tests; r is the effect size.

Table 5.5: Self-reported task-planning and actual task-planning by group*. The table reports medians (IQRs), mean ranks, U , exact p , and r values for the two task-planning measures in the control and experiment groups. The experiment group had higher medians for self-reported task planning (7.00 vs. 4.00) and actual task planning (7.50 vs. 5.50), but neither difference was significant (self-reported: $U = 51.00$, exact $p = .242$, $r = .25$; actual: $U = 59.50$, exact $p = .478$, $r = .16$).

Qualitative planning accounts were also mixed across groups. Among experiment-group participants with higher actual task-planning scores, P8 first asked the system to analyze the task, then used the skill to check the materials, select a suitable workflow, and revise the resulting stories. P13 began with one story and used feedback before trying another workflow option, while P15 compared several workflow options before selecting outputs that best fit the task. P18 reported that the static explanation left only a general impression, but that evaluation criteria encountered during use prompted further thinking about how to check and filter outputs. Higher actual task-planning scores also appeared in the control group. P3 separated task interpretation, audience analysis, review, and story generation; P6 moved from a summary and preliminary plan to small-scale generation, verification, refinement, and final selection; and P21 selected a workflow based on role-specific rules and repeatedly emphasized the core ICW constraints. Lower actual task-planning scores occurred in both groups. P14 used two direct

generation prompts without subsequent review or conditional adjustment, whereas P16 used a simpler, more output-driven sequence with fewer explicit situation–action contingencies. Initial uncertainty did not necessarily indicate lower actual task planning: P22 and P23 expressed uncertainty before use, but their later interaction logs showed more structured planning behavior. This heterogeneity was consistent with the non-significant differences in overall task-planning scores.

5.4. Exploratory Correlations Between Reading Engagement, Understanding, and Planning

Across all participants, exploratory Spearman correlations were used to examine correlations among reading engagement, understanding, and planning outcomes (see Figure 5.3, p. 22). Reading duration was positively correlated with reading completeness, $r_s = .513$, $p = .010$, but it was not significantly correlated with self-reported understanding, actual understanding, or actual task-planning. Reading completeness was positively correlated with actual understanding, $r_s = .543$, $p = .006$, and actual task-planning, $r_s = .439$, $p = .032$. Actual understanding was positively correlated with self-reported task-planning, $r_s = .530$, $p = .008$. Self-reported task-planning was not significantly correlated with actual task-planning, $r_s = .263$, $p = .213$.

5

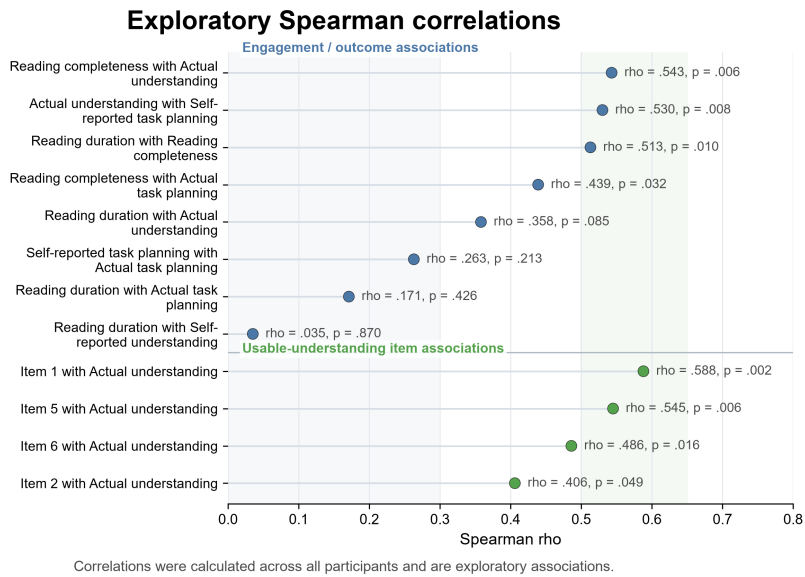


Figure 5.3: Exploratory Spearman correlations*. The plot displays r_s and p values for the correlations among reading engagement, understanding, and task-planning measures, grouped as engagement–outcome and self-reported-understanding-item correlations. The strongest displayed correlation was between Item 1 and actual understanding ($r_s = .588$, $p = .002$); reading completeness was also positively correlated with actual understanding ($r_s = .543$, $p = .006$) and actual task-planning ($r_s = .439$, $p = .032$).

In summary, the results mainly supported H1 but provided limited quantitative sup-

port for H2. The experiment group showed higher self-reported and actual understanding of the agent skill than the control group. Task-planning differences were not statistically significant, although more complete reading was positively correlated with stronger actual task-planning. Qualitative accounts showed the same boundary: higher understanding was more consistently observed in the experiment group, whereas effective planning behaviors appeared in both groups.

6

DISCUSSION

Following the results, the discussion focuses on three connected questions: **why the explanation appeared to support understanding, why this improved understanding did not consistently translate into stronger planning, and what these findings suggest for designing user-facing explanations for agent skills.** Because in this study all participants had design-related backgrounds and worked on a design task, the findings mainly show how before-use explanations shaped understanding and planning from the perspective of designers as one group of non-expert users. This context matters because design work often develops through the co-evolution of problem framing, emerging solutions, and situated judgment, rather than through a fully predefined linear sequence of actions (Dorst, 2019). The following discussion is grounded in this design-facing use case, while also identifying implications for broader efforts to make agent skills more understandable and steerable for non-expert users from other fields.

6.1. RQ1: Better Understanding Through Concise, Action-Oriented Explanations

The experiment group's advantage in understanding appears to come from carefully selecting and reorganizing skill information around users' actions, rather than from shorter text alone. The explanation was substantially shorter than the control material, 513 words compared with 1,920 words in the adapted SKILL.md, but its design logic was selective reorganization rather than simple reduction. Van der Meij and Gellevis describe procedures as support for goals, prerequisite states, actions and reactions, and unwanted states, which maps closely onto the explanation's focus on purpose, input requirements, workflow behavior, outputs, and correction guidance (Van Der Meij & Gellevis, 2004). Eiríksdóttir and Catrambone further distinguish procedural instructions, principles, and examples, showing that instruction design depends on whether the goal is immediate task performance, learning, or transfer (Eiríksdóttir & Catrambone, 2011). In this study, the explanation combined these concerns by giving users enough procedural structure to begin the task while also making the skill's working logic easier to interpret.

Its intended advantage was therefore a stronger alignment between the information presented and the actions users might need to take.

The material engagement results are important here because action-oriented information could support understanding only when users engaged sufficiently with the material. In the experiment group, participants used the explanation to locate headings, input guidance, workflow logic, and pattern options, whereas some control-group participants read less completely and remained unsure how to use the adapted SKILL.md. This pattern suggests that the explanation gave users more recognizable ways to enter and navigate the skill information, moving from reading to action: where to start, what to provide, which workflow options to notice, and which parts of the agent's behavior might be adjustable. This interpretation aligns with research showing that brief steps, simple formatting, and predictable structures can support the processing and navigation of procedural information (Farkas, 1999). Carroll's minimalist approach similarly treats concision as a way to help users begin meaningful tasks more readily, rather than as the indiscriminate removal of information (Carroll, 1990).

The understanding results also went beyond the quantitative score differences. Participants from the experiment group more often described the skill through input requirements, workflow logic, pattern choices, and default behavior, while several control-group participants still found the patterns or instructions unclear. These results suggest that the explanation mainly supported a practical understanding of how to use and steer the skill. From the designers' perspective, this means understanding not only what a skill can produce, but also how its workflow frames a design task and whether that framing fits the design goal. This interpretation is consistent with research on AI as a design material. Yildirim et al. show that designers working with AI need to understand AI capabilities, data dependencies, and system logic as part of design work, often through shared artifacts and examples (Yildirim et al., 2022). Yang et al. similarly argue that human-AI interaction is difficult to design partly because designers face uncertainty about what AI can do and how complex its outputs may be (Yang et al., 2020).

The findings above also connect back to the study's initial motivation. For non-expert users, the main challenge is not simply accessing more skill-related information, but translating agent-facing materials into usable prior knowledge. The needs identified in Phase 1 were therefore not only personal preferences. They specified the kinds of information missing from SKILL.md as a user-facing explanation. This aligns with prior work showing that explanations can help users provide better inputs and collaborate with AI, rather than only understand technical details (S. S. Y. Kim et al., 2023).

In relation to RQ1, these findings indicate that the before-use explanation mainly supported a practical form of understanding through reorganizing the SKILL.md in a concise and action-oriented way: it made the agent skill more legible as a workflow that non-expert users could enter, question, and adjust, rather than as an opaque set of internal instructions. For designers in particular, this matters because understanding the skill also means judging how its workflow may frame the design task and where its outputs may need to be checked against the design situation.

6.2. RQ2: Task-Planning Strategies: Understanding, Exploration, and Situated Judgment

The task planning results show the boundary of the before-use explanation's effect discussed above. The experiment group showed stronger understanding of the agent skill, but not significantly stronger task-planning strategies than the control group. This suggests that understanding may provide a basis for task planning by making action-relevant information more visible, but it does not by itself determine how users formulate plans.

The qualitative task planning results support this interpretation. For example, P13 began with one user story and used feedback from the output before trying another workflow option, while P8 first asked the AI system to analyze the task, then used the skill to select a suitable workflow for the task before generating user stories. These cases suggest that planning developed through probing, comparison, situated feedback, and domain judgment, rather than through the explanation alone. Given that all the participants had design-related backgrounds, these behaviors can be interpreted through reflection-in-action, in which designers make sense of a situation through action and use the consequences of that action to inform the next move (Valkenburg & Dorst, 1998). They also reflect the co-evolution of problem and solution: an emerging output may change how the task is understood, while a revised understanding of the task may redirect the next prompt or workflow choice (Dorst & Cross, 2001). Research on designers learning to co-create with AI similarly highlights boundary testing, self-explanation, and reflection as ways of developing effective working practices with unfamiliar AI tools (Gmeiner et al., 2023). In this study, early outputs from AI functioned not only as results to accept or reject, but also as design materials through which participants judged the task, the workflow, and their next action.

The behavioral results also suggest that task planning might be shaped by factors beyond the before-use material, including participants' prior AI-use habits, the uncertainty of the design task, and feedback from generated outputs during interaction. This interpretation is consistent with accounts of generative-AI use as metacognitively demanding: users must decide what to ask, how much context to provide, whether to inspect an output, and when to rely on or revise it (Tankelevitch et al., 2024). Prompt-based interaction can also create uncertainty about what task to formulate, how to express an intention, and what output to expect (Subramonyam et al., 2024). Non-experts may respond through opportunistic exploration rather than systematic prompt design (Zamfirescu-Pereira et al., 2023), and may prefer having choices available over maintaining continuous, effortful control (Dang et al., 2023). A before-use explanation may therefore clarify inputs, workflow options, and possible intervention points, but it cannot remove the situated decisions users need to make while working with an agent skill.

In relation to RQ2, the findings suggest that before-use explanations did not reliably improve task-planning strategies. Instead, the explanation appeared to provide optional orientation resources that non-expert users could draw on to test the system, interpret outputs, and judge the task as it develops. From a design perspective, the before-use explanations should not be designed to make users follow a fixed planning script, but provide lightweight scaffolds that help them notice options, compare outputs, and decide when to redirect the workflow. This principle of scaffolding without

scripting may also be relevant to other non-expert users who need support in working with agent skills without losing workflow flexibility.

6.3. Design Implications: Translating SKILL.md into User-Facing Explanations

The findings suggest that the design challenge is not simply to shorten SKILL.md, but to translate agent-facing procedural knowledge into information that non-expert users can locate, understand, and use. This audience mismatch was visible in Phase 1, where participants described the material as instructions for the AI rather than for themselves. The problem was not necessarily an absence of relevant procedural knowledge, but that the knowledge was not presented in a form connected to users' questions and actions. In Phase 2, participants also needed to relate the skill's workflow to the way they framed, evaluated, and adjusted a task. Although these observations come from a design context, the underlying audience mismatch may also affect non-experts using agent skills in other domains. This framing is consistent with Liao et al.'s account of explainable-AI design as a translation problem between technical capabilities and user needs. In their study with UX and design practitioners, explanation needs were represented as prototypical questions users might ask about an AI system, rather than as technical properties alone (Liao et al., 2020). Similarly, the present framework treats SKILL.md not only as agent-facing documentation, but as material that must be reorganized around the questions non-expert users need answered before use.

This study offers a preliminary, evidence-informed translation framework for agent skill documentation with three related dimensions. *Semantic translation* makes the agent's logic understandable by converting technical or agent-oriented language into concepts that non-expert users can interpret. *Task translation* connects general capabilities and workflow logic to the information users need to provide, attend to, and evaluate, including whether a workflow fits the task. *Control translation* makes defaults, automated decisions, checkpoints, and opportunities to accept or adjust the default path visible as user-relevant decision points.

The framework synthesizes the Phase 1 findings, the explanation developed in this study, and the Phase 2 findings. However, these findings do not validate the dimensions separately or verify if the same framework could work for other non-expert users apart from designers. Therefore, this framework should be treated as a starting point that future research and design can test, refine, and adapt across agent-skill contexts with different users. Although this framework is preliminary rather than a validated design model, it relates to broader HCI and XAI work on transparency, collaboration, and control. Liao and Vaughan argue that AI transparency should be shaped by stakeholders' goals and contexts, which supports the need to explain agent skills in ways that are meaningful to specific users before interaction begins (Liao & Vaughan, 2024). Kim et al. show that users often value explanations that help them provide better inputs and collaborate with AI, which connects to the task translation dimension (S. S. Y. Kim et al., 2023). The control translation dimension also relates to contestable-AI work, where explanations, interactive control, and opportunities for intervention are treated as design concerns (Alfrink et al., 2023).

Building on these three translation dimensions, the following design implications show how semantic, task, and control translation can be applied to the design of user-facing before-use explanations.

1. Orient users before guiding action.

Before-use explanations for an agent skill should aim to orient users before asking them to actually use the skill. An initial orientation can state the skill's purpose, capability boundaries, and general workflow briefly. After that, action-related guidance can then explain what users should provide, what outputs they can expect, and what they may need to check or correct in detail. This sequence is supported by Van der Meij and Gellevisj's four-component view of procedures, which treats goals, prerequisites, actions and reactions, and unwanted states as distinct forms of support for users (Van Der Meij & Gellevisj, 2004).

2. Make decision points and defaults visible.

Before-use explanations for an agent skill should make workflow features that matter for users' choices visible. Depending on the skill, these may include the skill's default behavior, automatic workflow decisions, available alternatives, and points in the workflow where users can check, confirm, redirect, or stop the process. For designers, these alternatives may foreground different aspects of a design problem and lead them to use the same skill in substantially different ways. In other domains, the relevant alternatives and decision points may differ. Such visibility does not guarantee better planning or control, but it can give non-expert users a clearer basis for deciding when to follow, evaluate, or adjust a skill.

This implication is consistent with Brachman et al.'s study of an agentic chatbot, where users could inspect the system's actions, rationales, tools, inputs, and sources, and used these action descriptions to evaluate the system's answers (Brachman et al., 2025). It is also supported by Das et al.'s work on subgoal-based explanations, which showed that explaining recommended actions through task-level subgoals helped users understand and avoid suboptimal suggestions in a planning task (Das et al., 2023).

3. Offer optional planning and review scaffolds.

Although the before-use explanation did not significantly improve task planning in this study, this does not mean that planning support for agent-skill use is unnecessary. Instead, the findings suggest that such support should be optional rather than prescriptive. Before-use explanations could include input checklists, prompt templates, constraint reminders, and suggested actions for unsuitable outputs. These scaffolds can help users plan and review their interaction without forcing them to follow a fixed sequence. This design implication is consistent with Dang et al.'s finding that writers often preferred having multiple AI suggestions to choose from, while direct prompting could offer more control but also required effort and could interrupt the writing flow (Dang et al., 2023). It also connects to Tankelevitch et al.'s account of generative AI as metacognitively demanding, because users must formulate goals, decompose tasks, evaluate outputs, decide when to rely on them, and adjust strategies (Tankelevitch et al., 2024). Optional support can therefore preserve workflow flexibility while reducing some of the effort involved in prompting, monitoring, and evaluation.

4. Use plain language in translation.

Before-use explanations for an agent skill should make procedural knowledge ac-

cessible without removing the meaning non-expert users need for action. Plain language should not mean removing important workflow logic. Instead, technical or agent-oriented concepts should be translated into user-facing and action-related terms: what the concept means for the user's input, the agent's behavior, the expected output, or the user's decision to check or redirect the process. For example, in this study, the explanation did not only name the patterns, but also explained what kinds of tasks each pattern might suit in an easy-to-understand way. In this sense, plain language is an important part of translation: it makes the same procedural knowledge readable from the user's side rather than only from the agent's side.

These implications show how the preliminary translation framework may guide explanation design without prescribing one fixed workflow, interface, or information format. Its role is to open a design space around understandable agent logic, workflow–task fit, and visible opportunities for adjustment, rather than to define a universal solution.

6.4. Limitations

The understanding measures covered only selected aspects of agent-skill understanding. Actual understanding was limited to whether participants noticed the available workflow patterns and understood the pattern used. However, the study combined self-reported understanding, coded actual understanding, interaction records, and interview data to avoid relying on a single measure.

The task-planning rubric focused on explicit goals, situational cues, planned actions, and situation–action links. It was adapted from Gollwitzer's implementation-intention framework (Gollwitzer & Sheeran, 2006), but it was not pilot-tested or independently validated. It may therefore have missed exploratory, intuitive, or output-driven strategies. However, the planning scores were interpreted together with participants' interaction logs and qualitative accounts.

All coding was conducted by one researcher, so no inter-rater reliability could be established. However, participant records were anonymized and coded in a shuffled order before being reassigned to the two conditions. This reduced the risk that knowledge of group membership influenced the coding, although it did not replace independent coding.

The study included 24 participants, with 12 in each condition. The sample size was informed by prior work on sample sizes in CHI research (Ortloff et al., 2025), and the groups were checked for baseline comparability. However, the sample may still have been too small to detect weaker effects, especially for task-planning outcomes. The non-significant task-planning results should therefore not be interpreted as evidence of no effect.

The study measured only the immediate effects of the explanation in one task session. This made it possible to examine participants' initial responses before repeated use changed their understanding or strategies. However, the study cannot show whether the effects remain, weaken, or grow over time.

The study used one adapted agent skill, one design task, and participants with design-related backgrounds. Keeping the skill and task the same allowed the two conditions to be compared under the same requirements. However, the findings cannot be general-

ized directly to other agent skills, tasks, or non-expert user groups.

The explanation was a composite intervention: its length, wording, content, examples, and structure were changed together. This allowed the explanation to be evaluated as one coherent and practically usable design, but the effects of its individual elements cannot be separated. The study therefore cannot identify the optimal explanation length, format, or translation strategy.

Finally, the study tested only a static before-use explanation and did not compare it with interactive walkthroughs, progressive disclosure, or in-use guidance. Human oversight was also outside the scope of the study. This kept the study focused on understanding and task planning, but the findings do not show whether better understanding leads to better monitoring, intervention, reliance, or safer agent use.

6.5. Future Work

Future research on explanation design. This study identifies what information may need to be translated for non-expert users, but it does not determine how that information should be represented in an interface. This distinction matters because explanation format is itself an interface-design variable: plain text, diagrams, layered cards, examples, and interactive elements may lead users to notice, interpret, and use the same information differently (Liao & Vaughan, 2024; Mucha et al., 2021). Future work could therefore compare plain-language text explanations, workflow diagrams, example-driven guidance, and optional expandable technical detail (Brachman et al., 2025; Nassif & Robillard, 2025). Future studies could also test these design implications separately, for example by comparing different orientation sequences, levels of decision-point visibility, optional planning scaffolds, and plain-language translation strategies. Such work could also study runtime evaluation cues, side-by-side workflow comparison, and reflection checkpoints that help users reconsider the task and the agent's approach while working (Gmeiner et al., 2023). Such studies should evaluate self-reported and actual understanding, cognitive burden, reading engagement, task performance, task-framing decisions, and workflow adjustment. Comparisons across designers and other non-expert groups would also show which implications transfer beyond the present design context. This would help test how the preliminary translation framework can be implemented in concrete interfaces while contributing to broader HCI research on intelligible AI (Abdul et al., 2018; Liao et al., 2020).

Future research on oversight-related behavior. The present study did not test whether before-use explanations change users' monitoring or intervention behavior. From the designers' perspective, oversight may begin before output checking: users may need to notice when an agent's workflow overlooks a key constraint, question how it has organized the task, and redirect the process. The findings suggest that visible defaults, decision points, and checking guidance may help make such judgments possible, but they do not show improved oversight, appropriate reliance, design quality, or safety. This distinction is important because human-centered transparency, meaningful human control, and contestability depend on the stakeholder, context, responsibility, and opportunities available to question and alter a system's course (Alfrink et al., 2023; Cavalcante Siebert et al., 2023; Liao & Vaughan, 2024). Prior work also cautions that explanations may increase awareness or reliance without helping users judge whether an output is correct

(Rader et al., 2018; Schoeffler et al., 2024). Future studies should therefore measure oversight directly: whether designers and other non-expert users monitor the process, detect deviations or missing constraints, question default task organization, check outputs, rely appropriately on the agent, and intervene at suitable points. Such work could compare before-use explanations, runtime transparency, and their combination to test whether the translation framework supports not only understanding, but also observable human oversight.

7

CONCLUSION

Agent skills package procedural knowledge for LLM-based agents, but this knowledge is usually organized for agent execution rather than for non-expert users. This thesis examined whether a before-use explanation derived from `SKILL.md` could help bridge this gap by making skill information more accessible before interaction. Through a two-phase mixed-methods study, it investigated the effects of such explanations on users' understanding and task-planning strategies, and explored how agent-facing skill information can be translated into user-facing explanations.

The findings show that before-use explanations most clearly supported understanding. Participants who received the explanation read the material more completely and showed higher self-reported and demonstrated understanding than those who received the adapted `SKILL.md`. These results do not indicate complete skill understanding, but they suggest that concise, plain-language, and action-oriented procedural information can help users recognize features that matter for use, such as input requirements, workflow options, defaults, and checking opportunities. The evidence for improved task-planning strategies was more limited, suggesting that understanding may support planning but does not determine how users organize their work with AI.

The main contribution of this thesis is both empirical and design-oriented. It provides evidence that before-use explanations can make selected aspects of an agent skill more understandable, and it proposes a preliminary translation framework for user-facing agent-skill explanations. The framework consists of three dimensions: *semantic translation*, which makes agent-oriented language understandable; *task translation*, which connects skill capabilities and workflows to users' tasks; and *control translation*, which makes defaults, alternatives, checkpoints, and opportunities for adjustment visible. More broadly, the study suggests that agent-skill design may need to begin at the level of information: interface interaction and visual presentation can support users only when the underlying information framework helps them understand what the skill does, how it works, and where their judgment may be needed.

These conclusions remain preliminary. The study examined one adapted agent skill, one design task, a small sample of design-related Master's students, and immediate mea-

sures of understanding and planning. Future research should test and refine the framework across different skills, tasks, user groups, explanation formats, and forms of runtime transparency. Such work would help clarify how before-use explanations can support not only understanding, but also more inspectable and steerable agent-skill workflows.

ACKNOWLEDGEMENTS

I would like to thank my chair, Evangelos Niforatos, and my mentor, Tianhao He, for all their guidance and support over the past four months. This graduation project has meant more to me than simply completing a degree. It helped me grow as a designer and deepen my understanding of design thinking, while also changing how I look at AI. I learned to think more critically about when and how AI should be used, and to explore AI-related design questions with both passion and academic care.

Above all, I want to thank my parents for supporting my decision to change fields and pursue my Master's degree abroad. None of what I have achieved would have been possible without your openness toward my choices, your trust in me, and your unwavering support. Thank you for always giving me the freedom and confidence to follow the path I chose.

A

APPENDIX

A.1. Phase 1 participant demographic sheet

PHASE 1 PARTICIPANT DEMOGRAPHIC SHEET				
Participant ID	Age	Professional Background	Educational Background	AI Literacy Level
P1	25	Strategic product design	Master student	Intermediate
P2	24	Architecture design	Master student	Low
P3	25	Creative media	Master student	Low
P4	25	Design for interaction	Master student	Intermediate
P5	28	Industrial design	PhD student	Intermediate
P6	24	Intelligent system design	Master student	Intermediate
P7	25	Architecture design	Master student	Intermediate
P8	23	Creative media	Master student	Low
P9	25	Intelligent robot design	Master student	Intermediate
P10	24	Digital art design	Master student	Intermediate
P11	25	Acoustic Engineer	Master student	Intermediate
P12	24	Communication and media	Master student	Low
P13	26	Smart product service system	PhD student	Intermediate
P14	25	Architecture design	Master student	Low
P15	25	Architecture design	Master student	Low

Figure A.1: Phase 1 participant demographic sheet. The table lists participant ID, age, professional background, educational background, and AI literacy level for the 15 exploratory-study participants.

A.2. Phase 1 introduction process

Participants were introduced to the basic concepts of Codex and agent skills, with researchers directly showing the Codex interface, `SKILL.md`, and local configurations of the skill they would use later (see Figure A.2, p. 36; Figure A.3, p. 36).

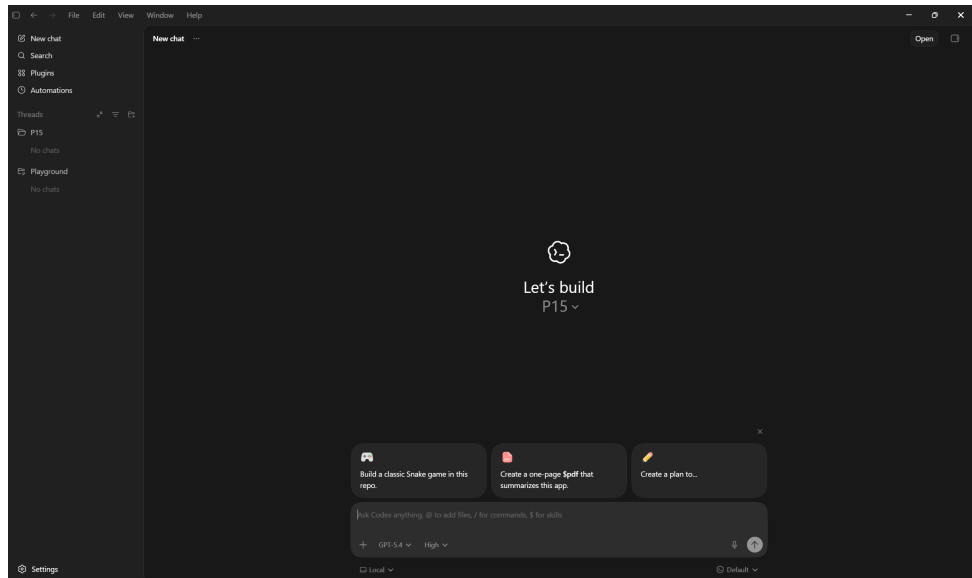


Figure A.2: Phase 1 Codex interface introduction. The screenshot shows the Codex interface presented to participants during the session introduction, before they used the Slides skill in the formal task.

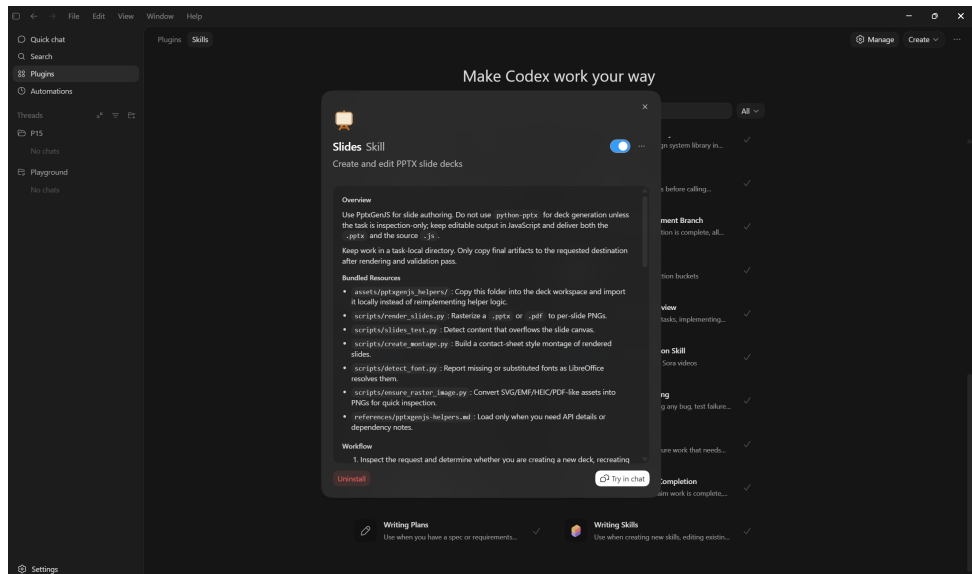


Figure A.3: Phase 1 Slides skill material introduction. The screenshot shows the Slides skill information presented to participants during the introduction before they formally used the skill in the task.

A.3. Phase 1 interview questions

1. Did you find the process challenging? If so, could you elaborate on why?
2. Can you explain how you planned to use the skill just now and why?
3. Reflect on the overall experience, what explanations do you think you need or want to have before going to actual usage of certain skills? (**discuss data and source, Actions & Model, Input & Context, System Response and other explanations if participants think they are important**)
4. If certain before-use explanations are available in this system, how do you expect it to be provided? Do you expect them to explain which parts of the input caused the result (**feature-based**), explain through showing examples (**example-based**) or other forms?

A.4. Phase 1 qualitative analysis and full findings

1 Current understanding and challenges

Distinguishing agent skills from standard conversational AI interactions. Participants viewed them as specialized task-execution tools rather than general-purpose chat. This cognitive shift meant users entered the interaction with a "solution-oriented mindset", perceiving the process as a structured project workflow rather than an open-ended brainstorming session. Participant 13 emphasized the goal-oriented nature of skills: "We're not playing like the conversational AI... Once I think everyone looking for skills and installing skills, they have a quite a specific results-oriented mindset already set into them". Agent skills were described as "allowing your AI to execute a more professional, more repeatable type of workflow" by Participant 15.

High cognitive burden when starting and technical alienation. Fourteen out of 15 participants clearly described confusion or a feeling of getting lost when trying to find instructions but only seeing technical files. Many reported a high learning cost and psychological pressure when seeing coding-related formats and terms. "I kind of couldn't read that thing. My brain just couldn't process it." P11 said. Participants were not willing to spend much time understanding the skill through technical files. "I might only read the overview, then I probably wouldn't want to read more. I wouldn't spend that much time understanding the underlying logic of a tool." (P12) Some participants first expected the file to be a user guide but later found out it was not. P3 said, "It felt more like code-oriented documentation than a beginner guide."

Gap between technical files and user needs. Eight participants found it hard to connect the technical files with their own task planning, even though the files contained information they later wanted. P4 said, "I can sense the structure, but not the practical meaning. I know it has reference value, but I still wouldn't know exactly how to use it." P9 explained the same gap: "I kind of understand the content about how it would complete the task but... It explains things, but it does not feel like a user manual." P8 also felt the file was written more for agents than for users: "I understood the general gist of the file,

but most of it felt like instructions for the AI rather than for me.”

2 Before-use explanation content preferences

Capability edges and functional boundaries. Since participants found it difficult to read the technical files, they wanted clearer explanations of what the agent skill could actually do. Among the capabilities, participants cared about the result most. P3 said, “I’d like to know clearly what result it can produce.” P5 also mentioned a similar need: “Before spending tokens, tell me what result I should expect.” Since the agent might not perfectly complete the task using the skill, P9 wanted to know the ideal result first: “At least then I know what it thinks it can do.” P11 proposed a complementary requirement: “And also, I’d like clearer boundaries of what it cannot do if it’s possible.”

Operational checklist. Besides capabilities, participants wanted to know what input the skill needed to operate normally. P1 said, “It should tell me what inputs or customizable settings users need to provide.” P2 also agreed that knowing the prerequisites to start the process is important: “It’s better to tell me at the beginning what tools or files it needs.” After learning that the skill could also be run locally, participants raised more concerns about security. P3 considered this a requirement for before-use explanations: “Permissions requests or installation needs should be clearer.” Similarly, P7 expressed practical concerns: “I want to know whether it affects the result, my computer, or the original files.”

Direct and actionable assistance with user input. Many participants still found it hard to start when no extra explanation of user prompts was provided before tasks. P2 cared about precise wording: “It would help if it also told me what wording works for each function.” P6 also needed actionable suggestions for how to start: “I like actionable next steps that help me continue. Earlier suggestions were useful because they gave me directions I didn’t think of.” P11 asked for direct help with writing better prompts: “it would be best if it could guide me on how to write prompts or provide input in a way that makes the agent work more efficiently or yields better results.”

3 Before-use explanation format preferences

Preference for example-based explanations. Eleven out of 15 participants said they wanted examples or demonstrations of what the agent skill could produce before the task. When discussing feature-based descriptions of results, P5 clearly preferred concrete examples: “Deliverables matter more than abstract planning descriptions. Tell me concrete outputs, not vague assumptions.” P8 gave a more detailed reason: “Concrete previews are useful. Adjectives are not.” P7 also asked for examples with the precise prompts that led to such results: “The most direct thing is examples with the matching prompt. That is the most intuitive way.”

Plain-language explanations for non-experts. Compared with the technical files, participants clearly preferred explanations in understandable language. P1 said such plain language would improve her understanding of the skill: “If it explains each step in plain language, and what that step is for, I’d understand more.” A similar requirement appeared in P3’s interview: “I hope it can summarize that long list of content it mentioned in the files in a way that’s easier for users to understand.” P6 said he would refuse

to read explanations if they were provided in a coding style. P15 reflected on the motivation behind this preference: “Don’t force me to learn a command system. I prefer natural language over templates. Templates are only fine as optional choices to inspire me.”

Information structure, volume, and density control. Participants did not want long blocks of text when seeking before-use explanations. P1 said, “General short explanations are better than very detailed ones.” Consistent with the preference for example-based explanations, P4 expressed a similar idea: “Examples would be more useful than dense text. People like me won’t read huge tables.” P6 described a selective reading strategy: “I scan the key information instead of reading every detail, only looking for useful ones.” P14 was more direct: “I’d rather understand it through trying it than reading a long explanation.” P12 described his expectation as concise guidance rather than endless reading and also suggested charts, diagrams, or visuals instead of a large block of text to demonstrate the workflow.

4 Task-planning strategies

Trial-and-error strategies. Most participants used this strategy because they were uncertain about the “expected results”, so their first conversational turns often used simple prompts. P2 said, “I’d start with a simple prompt and then see what it generates.” P6 described a similar plan: “I usually first test the most basic version and see how good it is. If quality matters, I will add detailed prompts about audience, timing, pages, style, and logic.” P9 considered perfect prompting at the beginning too costly: “I usually write prompts casually. If it misunderstands me, I rewrite them. I use trial and error rather than trying to perfect prompts first for a lower cost.”

Breaking conversational interaction into steps: general direction first, details later. P8 said, “I split the workflow into separate procedures so I can control it more.” Because of limited understanding of the skill’s capabilities, P9 decided to work step by step: “I first asked for an outline because I was not sure whether it could directly give me the final output.” P13 described a stronger planning-first approach: “only after I have a clear outline, clear structure, and page-to-page content will I call the skill for final output.”

A.5. Phase 2 participant demographic sheet

PHASE 2 PARTICIPANT DEMOGRAPHIC SHEET

Participant ID	Age	Professional Background	Education Background	AI familiarity	Agent skill familiarity
P1	25-34	Design	Master	Basic	Basic
P2	25-34	Design for interaction	Master	Intermediate	Low
P3	25-34	Game design	Master	Proficient	Low
P4	25-34	Design for interaction	Master	Intermediate	Low
P5	25-34	Media & Design	Master	Basic	Low
P6	25-34	Architecture & Computation Design	Master	Proficient	Intermediate
P7	25-34	Industrial Design	Master	Proficient	Intermediate
P8	18-24	Architecture Design	Master	Intermediate	Low
P9	18-24	Urban Planning Design	Master	Intermediate	Basic
P10	18-24	Design (Strategic & Interaction)	Master	Intermediate	Low
P11	25-34	Acoustics/Sound Design	Master	Proficient	Intermediate
P12	25-34	Integrated Product Design	Master	Intermediate	Basic
P13	25-34	Strategic Product Design	Master	Intermediate	Basic
P14	25-34	Landscape Design	Master	Intermediate	Low
P15	25-34	Media & Design	Master	Basic	Low
P16	25-34	Integrated Product Design	Master	Basic	Low
P17	18-24	Media & Design	Master	Intermediate	Basic
P18	25-34	Interaction Design	Master	Proficient	Intermediate
P19	25-34	Architecture Design & Technology	Master	Intermediate	Basic
P20	25-34	Design for Interaction	Master	Intermediate	Intermediate
P21	25-34	Integrated design	Master	Intermediate	Basic
P22	25-34	Production design	Master	Basic	Low
P23	25-34	Interaction design	Master	Proficient	Intermediate
P24	25-34	Interaction design	Master	Intermediate	Low

Figure A.4: Phase 2 participant demographic sheet. The table lists participant ID, age, professional background, educational background, AI familiarity, and agent skill familiarity for the 24 participants.

A.6. Phase 2 explanation design

1 What This Skill Does

This skill helps you break down a product or service idea into smaller, implementable user stories. It is useful when an idea is too broad, complex, or difficult to directly turn into detailed design.

2 How It Works

Step 1: Check the Idea

The skill first checks whether your idea is clear and sufficient to generate user stories.

Step 2: Choose a Splitting Pattern

A splitting pattern is the logic used to break down the idea into user stories, and different patterns show different priorities.

The skill provides the following six patterns to break down the idea. If you do not ask for a specific pattern, the skill will use Pattern 1 Workflow Steps by default. You may check Section 3 and 4 for details on how to use different patterns.

Pattern 1 Workflow Steps — for creating process or sequence-focused user stories.
i.e. users need to first...then...

Pattern 2 Operations — for action-focused user stories.
i.e. users need to do... in a certain way...

Pattern 3 Business Rule Variations — for rules, regulations, or condition-centered user stories.

i.e. under the regulation of...

Pattern 4 Data Variations — for data, item, or material type-centered user stories.

i.e. for different users, the items should be divided by...

Pattern 5 Data Entry Methods — for different input or interface-centered user stories.

i.e. the UI of... for users should be...

Pattern 6 Defer Performance — for "making it work first, then improving performance later" type of user stories.

i.e. it should be functional first, with performance improved later.

Step 3: Generate User Stories

Then the skill will generate user stories based on the pattern chosen in Step 2.

3 What You Should Provide to Start

To get a useful result, provide:

- the design background or context
- the product or service idea
- target users or stakeholders
- the main goal
- Optional: important criteria or constraints
- Optional: Please use XXX pattern to break down the idea.

If you do not ask for a specific pattern to break down the idea, the skill will use Pattern 1: Workflow Steps, which may not always fit your task. You can also specify which pattern you want the skill to use.

Example:

Please use this skill to break down my service idea into smaller user stories.

The context is: design background.

The target users are: stakeholders.

The main goal is: goal.

Optional: Important constraints are: constraints.

Optional: Please use Operations pattern to break down the idea.

4 Other Tips On How To Improve The Output

The output usually includes:

- Smaller user stories of target stakeholders
- Acceptance criteria
- A short rationale for the breakdown.
- Workload estimation of turning each story into implementation

Guide the skill more explicitly if you think:

- the output misses important rules or constraints
- the output only follows a general workflow
- the stories do not match the real complexity of your task
- you want the result to focus on a specific logic, such as rules, operations, or data types

Example:

Please add constraints into consideration and revise the output.

Please check whether another splitting pattern would fit this task better.

A.7. Phase 2 agent skill source, license, and adaptation**Epic-breakdown-advisor source**

The agent skill used in this study was adapted from epic-breakdown-advisor, an open-source skill in Dean Peters' Product-Manager-Skills GitHub repository. The repository provides a collection of product management skills and frameworks designed for use with AI agent environments such as Claude Code, Codex, and other agent-based tools.

License information

The repository is released under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0). This license allows non-commercial use, sharing, and adaptation of the material, provided that appropriate attribution is given and adapted materials are shared under compatible terms. The skill was therefore adapted for academic research purposes only.

Original skill overview

The original epic-breakdown-advisor skill is an interactive product management skill that guides users in breaking down large epics into smaller user stories. It is based on Richard Lawrence's Humanizing Work methodology and uses a systematic story-splitting process. The skill first validates whether an epic is suitable for splitting, then applies nine splitting patterns in sequence, such as Workflow Steps, Operations, Business Rule Variations, and Data Variations. Its goal is to help users create vertical slices that preserve user value rather than splitting work only by technical layers.

Skill adaptation

The original epic-breakdown-advisor skill was designed to help product managers break down large epics into user stories using the Humanizing Work story-splitting methodology. It originally guided the agent through nine splitting patterns in sequence.

For this study, the skill was adapted to better fit a design-oriented experimental task. First, the wording was changed from "epics" to "product/service requirements" so that participants from product, service, and system design backgrounds could understand the task more naturally. Second, the available splitting patterns were reduced from nine to six to lower unnecessary complexity within the limited task time. Third, Pattern 1, Workflow Steps, was set as the default execution logic unless users explicitly requested another pattern.

This default setting was introduced intentionally for measurement. The Phase 2 task required participants to recognize rule-based differences between ICW and non-ICW handling. Therefore, Pattern 3, Business Rule Variations, was treated as the most appropriate reference strategy. By setting Workflow Steps as the default, the study could observe whether participants understood the skill well enough to guide or correct the agent toward a more suitable pattern.

The original and adapted SKILL .md files are provided in the accompanying attachments.

A.8. Phase 2 design brief

1. Regulatory Context and Problem

In airline operations, cabin waste separation is not only a service design issue. It is strongly shaped by regulations. Under EU laws, International Catering Waste (ICW), such as food waste from international flights, must be treated as high-risk waste and disposed of through controlled processes.

Importantly, originally recyclable items can also become ICW if they are contaminated by food waste or other ICW materials. For example, a plastic cup may require different handling depending on whether it is clean or contaminated.

2. Current Idea

Based on regulatory requirements, the airline adopts a simplified color-based classification system:

- Grey = ICW : food waste and contaminated materials requiring regulated disposal
- Green = Non-ICW : materials not subject to ICW handling rules

Passengers and cabin crew members should collaborate during the waste separation process to enhance efficiency.

3. Your Task

User stories describe what stakeholders want or need and why during a service, for example: as a cabin crew member, I want...because...

For the current service idea of color-based classification system, you as a designer are asked to use the agent skill to help you generate ideas for possible user stories from the passengers and the cabin crew members (at least 3 stories for each one) during the cabin service, with a focus on ICW-related regulations above.

A.9. Phase 2 experimental materials by condition

The two groups interacted with the same prototype. The only difference was whether clicking the blue button labeled “Agent skill” displayed the adapted SKILL .md (control group) or the before-use explanation (experiment group).

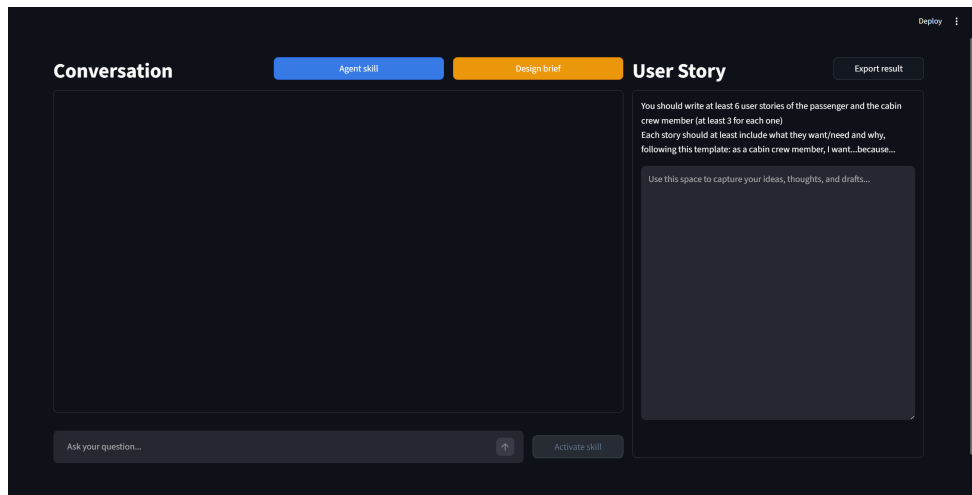


Figure A.5: Phase 2 prototype interface. The screenshot shows the experimental workspace, including the Conversation panel, Agent skill and Design brief controls, and User Story panel.

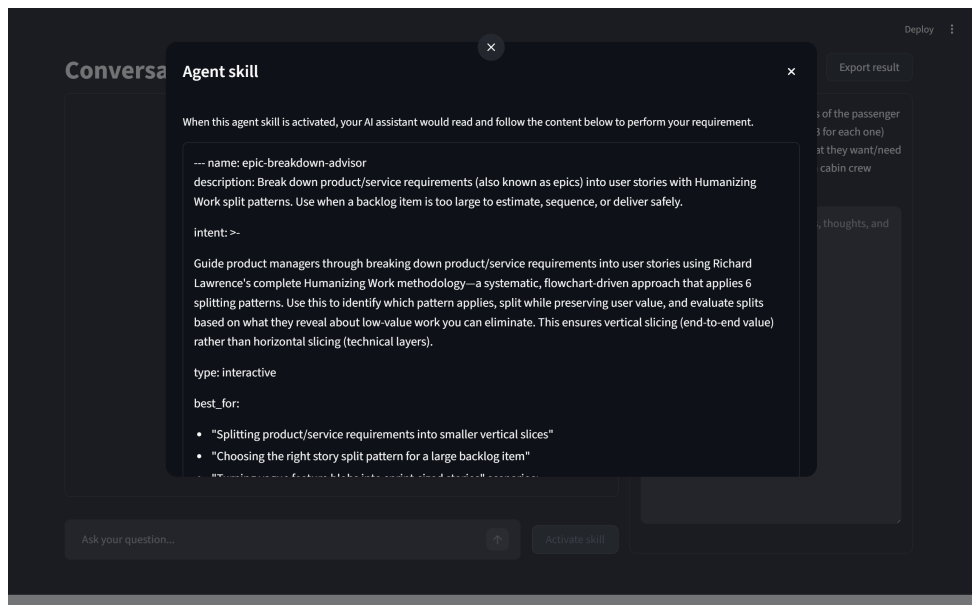


Figure A.6: Control group SKILL .md material. The screenshot shows the adapted SKILL .md presented to participants in the control group during the warm-up session.

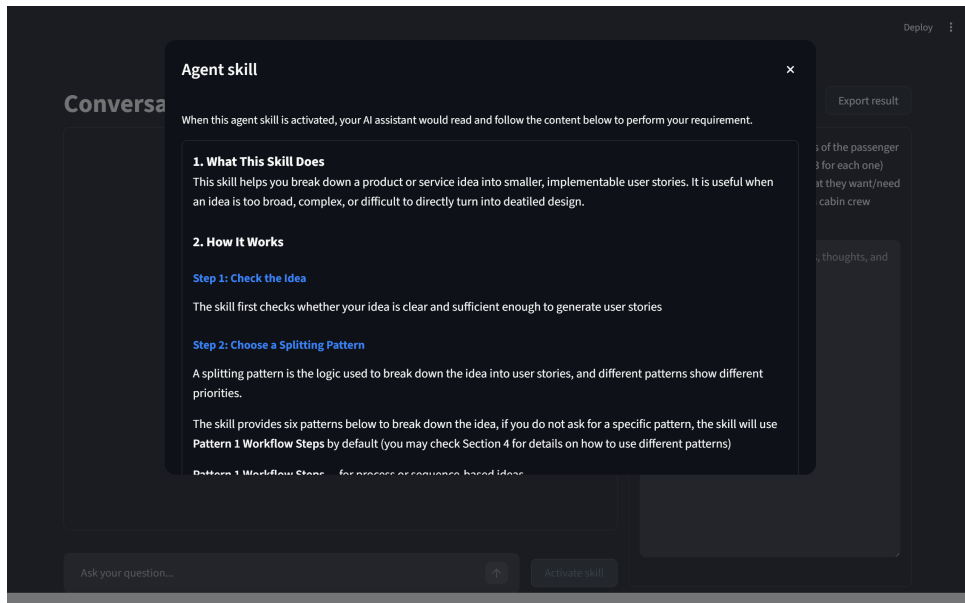


Figure A.7: Experiment group before-use explanation. The screenshot shows the before-use explanation presented to participants in the experiment group during the warm-up session.

A.10. Phase 2 survey and interview instruments

1 Pre-task survey

Demographic information

1. What is your gender?

- Woman
- Man
- Non-binary
- Prefer not to say

2. What is your age?

- Under 18
- 18–24
- 25–34
- 35–44
- 45–54
- Above 54

3. What is your current level of education?
 - High school or equivalent
 - Bachelor
 - Master
 - PhD or higher
4. What is your professional or academic background?
 - Open-ended response

Attitude toward artificial intelligence

Response scale: 0 = strongly disagree, 10 = strongly agree

1. I fear artificial intelligence.
2. I trust artificial intelligence.
3. Artificial intelligence will destroy humankind.
4. Artificial intelligence will benefit humankind.
5. Artificial intelligence will cause many job losses.

Scores for Item 2 and Item 4 were summed to create a new variable called “Accept AI,” and scores for Item 1, Item 3, and Item 5 were summed to create a new variable called “Fear AI.”

Self-reported knowledge about artificial intelligence and agent skills

Response scale: 1=Low, 2=Basic, 3=Intermediate, 4=Proficient, 5=Advanced.

1. I am familiar with artificial intelligence and know how to use it well.
2. I am familiar with agent skills and know how to use them well.

2 Post-task survey

Response scale: 1 = strongly disagree, 5 = strongly agree

1. From the information I received, I know how the agent skill works.
2. The information I received about how the agent skill works is satisfying.
3. The information I received about how the agent skill works has sufficient detail.
4. The information I received about how the agent skill works seems complete.
5. The information I received about how the agent skill works tells me how to use it.
6. The information I received about how the agent skill works is useful to my goals.

7. The information I received about the agent skill shows me how accurate the agent skill is.

3 Post-task interview questions

1. How do you perceive your understanding of the skill now? What helped with the understanding, if any?
2. Do you still remember how you planned to use the agent skill? Could you elaborate on the reasons behind your planning?
3. Do you think the information we provided about the agent skill had an effect on your planning? Please elaborate on how.
4. Did you notice the skill actually offered different patterns to generate user stories? Do you know which one your AI assistant followed when creating the final implementation? Do you know why it followed that one?

A.11. Additional quantitative results

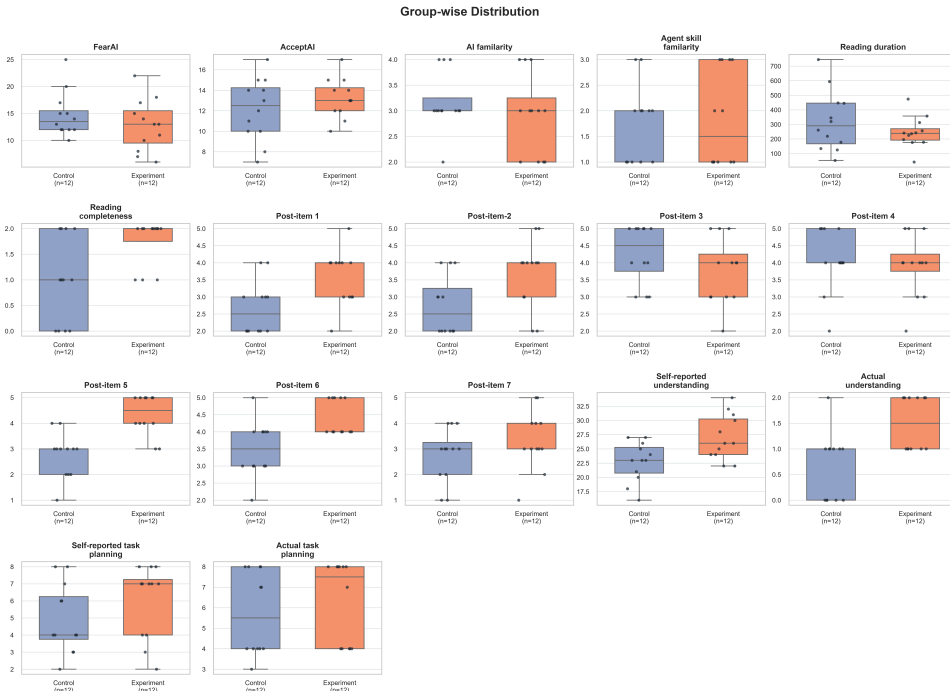


Figure A.8: Group-wise distribution. The multi-panel boxplots show the 17 variables from Fear AI through actual task-planning for the control and experiment groups, with jittered participant-level points.

Shapiro-Wilk normality checks

Variable	Control N	Control W	Control p	Control result	Experiment N	Experiment W	Experiment p	Experiment result
FearAI	12	0.847	.034	Non-normal	12	0.975	.953	Not rejected
AcceptAI	12	0.968	.884	Not rejected	12	0.979	.978	Not rejected
AI familiarity	12	0.753	.003	Non-normal	12	0.824	.018	Non-normal
Agent skill familiarity	12	0.807	.011	Non-normal	12	0.737	.002	Non-normal
Reading duration	12	0.947	.590	Not rejected	12	0.937	.459	Not rejected
Reading completeness	12	0.811	.012	Non-normal	12	0.552	< .001	Non-normal
Post-item 1	12	0.777	.005	Non-normal	12	0.875	.077	Not rejected
Post-item-2	12	0.764	.004	Non-normal	12	0.863	.053	Not rejected
Post-item 3	12	0.764	.004	Non-normal	12	0.894	.134	Not rejected
Post-item 4	12	0.818	.015	Non-normal	12	0.865	.056	Not rejected
Post-item 5	12	0.884	.099	Not rejected	12	0.777	.005	Non-normal
Post-item 6	12	0.886	.106	Not rejected	12	0.640	< .001	Non-normal
Post-item 7	12	0.872	.069	Not rejected	12	0.920	.286	Not rejected
Self-reported understanding	12	0.940	.499	Not rejected	12	0.937	.458	Not rejected
Actual understanding	12	0.784	.006	Non-normal	12	0.650	< .001	Non-normal
Self-reported task planning	12	0.909	.204	Not rejected	12	0.810	.012	Non-normal
Actual task planning	12	0.781	.006	Non-normal	12	0.682	< .001	Non-normal

Shapiro-Wilk tests were conducted separately by group (n=12 each). p < .05 indicates evidence against normality.

Figure A.9: Shapiro-Wilk normality checks. The table reports *N*, *W*, *p*, and the test result for each of the 17 variables, calculated separately for the control and experiment groups (*n* = 12 per group).

BIBLIOGRAPHY

- Abadeh, P., Lochner, M., Ansari, T., & Zarrinkalam, F. (2026, April). MCP-Scanner: Detecting Security Risks in Model Context Protocol Systems. In *Proceedings of the 2026 ACM/IEEE 7th International Workshop on Engineering and Cybersecurity of Critical Systems* (pp. 41–48). <https://doi.org/10.1145/3786160.3788471>
- Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M. (2018). Trends and Trajectories for Explainable, Accountable and Intelligible Systems. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–18. <https://doi.org/10.1145/3173574.3174156>
- Albada, A., van Dulmen, S., Ausems, M. G. E. M., & Bensing, J. M. (2012). A pre-visit website with question prompt sheet for counselees facilitates communication in the first consultation for breast cancer genetic counseling: Findings from a randomized controlled trial. *Genetics in Medicine*, 14(5), 535–542. <https://doi.org/10.1038/gim.2011.42>
- Alfrink, K., Keller, I., Kortuem, G., & Doorn, N. (2023). Contestable AI by Design: Towards a Framework. *Minds and Machines*, 33(4), 613–639. <https://doi.org/10.1007/s11023-022-09611-z>
- Anthropic. (2025). Anthropic Economic Index report: Uneven geographic and enterprise AI adoption. *Anthropic*. <https://www.anthropic.com/research/anthropic-economic-index-september-2025-report>
- Anthropic. (2026). Agent Skills. <https://platform.claude.com/docs/en/agents-and-tools/agent-skills/overview>
- Anthropics/skills. (2026, June). <https://github.com/anthropics/skills>
- Baig, M. I., & Yadegaridehkordi, E. (2024). ChatGPT in the higher education: A systematic literature review and research challenges. *International Journal of Educational Research*, 127, 102411. <https://doi.org/10.1016/j.ijer.2024.102411>
- Bauer, K., von Zahn, M., & Hinz, O. (2023). Expl(AI)ned: The Impact of Explainable Artificial Intelligence on Users' Information Processing. *Information Systems Research*, 34(4), 1582–1602. <https://doi.org/10.1287/isre.2023.1199>
- Brachman, M., Kunde, S., Miller, S., Fucs, A., Dempsey, S., Jabbour, J., & Geyer, W. (2025). Building Appropriate Mental Models: What Users Know and Want to Know about an Agentic AI Chatbot. *Proceedings of the 30th International Conference on Intelligent User Interfaces*, 247–264. <https://doi.org/10.1145/3708359.3712071>
- Cabrera, Á. A., Perer, A., & Hong, J. I. (2023). Improving Human-AI Collaboration With Descriptions of AI Behavior. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1), 1–21. <https://doi.org/10.1145/3579612>
- Cai, C. J., Jongejan, J., & Holbrook, J. (2019, March). The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (pp. 258–262). <https://doi.org/10.1145/3301275.3302289>

- Caine, K. (2016, May). Local Standards for Sample Size at CHI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 981–992). <https://doi.org/10.1145/2858036.2858498>
- Carroll, J. M. (1990, June). *The Nurnberg Funnel: Designing Minimalist Instruction for Practical Computer Skill* (M. Lesk, Ed.). MIT Press.
- Cavalcante Siebert, L., Lupetti, M. L., Aizenberg, E., Beckers, N., Zgonnikov, A., Veluwenkamp, H., Abbink, D., Giaccardi, E., Houben, G.-J., Jonker, C. M., van den Hoven, J., Forster, D., & Lagendijk, R. L. (2023). Meaningful human control: Actionable properties for AI system development. *AI and Ethics*, 3(1), 241–255. <https://doi.org/10.1007/s43681-022-00167-3>
- Chatlatanagulchai, W., Li, H., Kashiwa, Y., Reid, B., Thonglek, K., Leelaprute, P., Rungsawang, A., Manaskasemsak, B., Adams, B., Hassan, A. E., & Iida, H. (2025, November). Agent READMEs: An Empirical Study of Context Files for Agentic Coding [arXiv:2511.12884 [cs]]. <https://doi.org/10.48550/arXiv.2511.12884>
- Chen, C., Liao, M., & Sundar, S. S. (2024, September). When to Explain? Exploring the Effects of Explanation Timing on User Perceptions and Trust in AI systems. In *Proceedings of the Second International Symposium on Trustworthy Autonomous Systems* (pp. 1–17). <https://doi.org/10.1145/3686038.3686066>
- Chen, O., Retnowati, E., Chan, B. K. Y., & Kalyuga, S. (2023). The effect of worked examples on learning solution steps and knowledge transfer. *Educational Psychology*, 43(8), 914–928. <https://doi.org/10.1080/01443410.2023.2273762>
- Chen, S., Piet, J., Sitawarin, C., & Wagner, D. (2024, September). StruQ: Defending Against Prompt Injection with Structured Queries [arXiv:2402.06363 [cs.CR]]. <https://doi.org/10.48550/arXiv.2402.06363>
- Claude Code. (2026). Claude Code by Anthropic | AI Coding Agent, Terminal, IDE. <https://claude.com/product/claude-code>
- Codex CLI. (2026). CLI – Codex | OpenAI Developers. <https://developers.openai.com/codex/cli>
- Consolidated TEXT. (2009). Consolidated TEXT: 32009R1069 — EN — 15.07.2019. *EUR-Lex*. <https://eur-lex.europa.eu/eli/reg/2009/1069/2019-07-15/eng>
- Dang, H., Goller, S., Lehmann, F., & Buschek, D. (2023, April). Choice Over Control: How Users Write with Large Language Models using Diegetic and Non-Diegetic Prompting. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1–17). <https://doi.org/10.1145/3544548.3580969>
- Das, D., Kim, B., & Chernova, S. (2023, March). Subgoal-Based Explanations for Unreliable Intelligent Decision Support Systems. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (pp. 240–250). <https://doi.org/10.1145/3581641.3584055>
- Debenedetti, E., Zhang, J., Balunović, M., Beurer-Kellner, L., Fischer, M., & Tramèr, F. (2024, November). AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents [arXiv:2406.13352 [cs.CR]]. <https://doi.org/10.48550/arXiv.2406.13352>
- Dimoska, A., Tattersall, M. H. N., Butow, P. N., Shepherd, H., & Kinnarsley, P. (2008). Can a "prompt list" empower cancer patients to ask relevant questions? *Cancer*, 113(2), 225–237. <https://doi.org/10.1002/cncr.23543>

- Dorst, K. (2019). Co-evolution and emergence in design. *Design Studies*, 65, 60–77. <https://doi.org/10.1016/j.destud.2019.10.005>
- Dorst, K., & Cross, N. (2001). Creativity in the Design Process: Co-evolution of Problem–Solution. *Design Studies*, 22(5), 425–437. [https://doi.org/10.1016/S0142-694X\(01\)00009-6](https://doi.org/10.1016/S0142-694X(01)00009-6)
- Eiríksdóttir, E., & Catrambone, R. (2011). Procedural Instructions, Principles, and Examples - Elsa Eiríksdóttir, Richard Catrambone, 2011. <https://journals-sagepub-com.tudelft.idm.oclc.org/doi/full/10.1177/0018720811419154>
- Farkas, D. (1999). The Logical and Rhetorical Construction of Procedural Discourse. *Technical Communication*, 46.
- Gammelgård-Larsen, A., van Berkel, N., Skov, M. B., & Kjeldskov, J. (2024, May). Designing for Human-AI Interaction: Comparing Intermittent, Continuous, and Proactive Interactions for a Music Application. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (pp. 1–8). <https://doi.org/10.1145/3613905.3650886>
- Gemini CLI. (2026). Gemini CLI | Gemini for Google Cloud. <https://docs.cloud.google.com/gemini/docs/codeassist/gemini-cli>
- Gmeiner, F., Yang, H., Yao, L., Holstein, K., & Martelaro, N. (2023). Exploring Challenges and Opportunities to Support Designers in Learning to Co-create with AI-based Manufacturing Design Tools. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–20. <https://doi.org/10.1145/3544548.3580999>
- Gollwitzer, P. M., & Sheeran, P. (2006, January). Implementation Intentions and Goal Achievement: A Meta-analysis of Effects and Processes. In *Advances in Experimental Social Psychology* (pp. 69–119, Vol. 38). Academic Press. [https://doi.org/10.1016/S0065-2601\(06\)38002-1](https://doi.org/10.1016/S0065-2601(06)38002-1)
- Hattie, J. A. C., & Donoghue, G. M. (2016). Learning strategies: A synthesis and conceptual model. *NPJ Science of Learning*, 1, 16013. <https://doi.org/10.1038/npjscilearn.2016.13>
- Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2023). Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science*, 5. <https://doi.org/10.3389/fcomp.2023.1096257>
- Hou, X., Zhao, Y., Wang, S., & Wang, H. (2025). Model Context Protocol (MCP): Landscape, Security Threats, and Future Research Directions. *ACM Transactions on Software Engineering and Methodology*, 0(ja). <https://doi.org/10.1145/3796519>
- Hou, Y., Yang, Z., Pang, Z., & Ma, X. (2026, May). SkillSieve: A Hierarchical Triage Framework for Detecting Malicious AI Agent Skills [arXiv:2604.06550 [cs.CR]]. <https://doi.org/10.48550/arXiv.2604.06550>
- Jin, C., Wang, A., Wei, Z., Wang, K., Zeng, B., Zhang, Q., Yang, C., Qu, J., Hu, X., & Xu, X. (2026, May). SkillSafetyBench: Evaluating Agent Safety under Skill-Facing Attack Surfaces [arXiv:2605.12015 [cs.CR]]. <https://doi.org/10.48550/arXiv.2605.12015>
- Johansson-Sköldberg, U., Woodilla, J., & Çetinkaya, M. (2013). Design thinking: Past, present and possible futures. *Creativity and Innovation Management*, 22(2), 121–146. <https://doi.org/10.1111/caim.12023>

- Juneja, P., Zhang, W., Smith-Renner, A. M., Lamba, H., Tetreault, J., & Jaimes, A. (2024, May). Dissecting users' needs for search result explanations. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (pp. 1–17). <https://doi.org/10.1145/3613904.3642059>
- Kazemitabaar, M., Chow, J., Ma, C. K. T., Ericson, B. J., Weintrop, D., & Grossman, T. (2023, April). Studying the effect of AI Code Generators on Supporting Novice Learners in Introductory Programming. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1–23). <https://doi.org/10.1145/3544548.3580919>
- Kim, J., Yu, S., Lee, S.-S., & Detrick, R. (2025). Students' prompt patterns and its effects in AI-assisted academic writing: Focusing on students' level of AI literacy. *Journal of Research on Technology in Education*, 0(0), 1–18. <https://doi.org/10.1080/15391523.2025.2456043>
- Kim, S. S. Y., Watkins, E. A., Russakovsky, O., Fong, R., & Monroy-Hernández, A. (2023, April). "Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1–17). <https://doi.org/10.1145/3544548.3581001>
- Kulesza, T., Stumpf, S., Burnett, M., & Kwan, I. (2012, May). Tell me more? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1–10). <https://doi.org/10.1145/2207676.2207678>
- Lee, B. C., & Chung, J. (2024). An empirical investigation of the impact of ChatGPT on creativity. *Nature Human Behaviour*, 8(10), 1906–1914. <https://doi.org/10.1038/s41562-024-01953-1>
- Leichtmann, B., Humer, C., Hinterreiter, A., Streit, M., & Mara, M. (2023). Effects of Explainable Artificial Intelligence on trust and human behavior in a high-risk decision task. *Computers in Human Behavior*, 139, 107539. <https://doi.org/10.1016/j.chb.2022.107539>
- Li, X., Chen, W., Liu, Y., Zheng, S., Chen, X., He, Y., Li, Y., You, B., Shen, H., Sun, J., Wang, S., Li, B., Zeng, Q., Wang, D., Zhao, X., Wang, Y., Chaim, R. B., Di, Z., Gao, Y., ... Lee, H.-c. (2026, March). SkillsBench: Benchmarking How Well Agent Skills Work Across Diverse Tasks [arXiv:2602.12670 [cs]]. <https://doi.org/10.48550/arXiv.2602.12670>
- Liao, Q. V., Gruen, D., & Miller, S. (2020). Questioning the AI: Informing Design Practices for Explainable AI User Experiences. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–15. <https://doi.org/10.1145/3313831.3376590>
- Liao, Q. V., & Vaughan, J. W. (2024). AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. *Harvard Data Science Review*, (Special Issue 5). <https://doi.org/10.1162/99608f92.8036d03b>
- Lim, B. Y., Dey, A. K., & Avrahami, D. (2009, April). Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2119–2128). <https://doi.org/10.1145/1518701.1519023>

- Liu, S.-C., Wang, S., Chang, T., Lin, W., Hsiung, C.-W., Hsieh, Y.-C., Cheng, Y.-P., Luo, S.-H., & Zhang, J. (2023, December). JarviX: A LLM No code Platform for Tabular Data Analysis and Optimization. In M. Wang & I. Zitouni (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track* (pp. 622–630). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.emnlp-industry.59>
- Long, D., & Magerko, B. (2020, April). What is AI Literacy? Competencies and Design Considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). <https://doi.org/10.1145/3313831.3376727>
- Mayer, R. E. (2008). Applying the science of learning: Evidence-based principles for the design of multimedia instruction. *The American Psychologist*, 63(8), 760–769. <https://doi.org/10.1037/0003-066X.63.8.760>
- Mesmer-Magnus, J., & Viswesvaran, C. (2010). The role of pre-training interventions in learning: A meta-analysis and integrative review. *Human Resource Management Review*, 20(4), 261–282. <https://doi.org/10.1016/j.hrmr.2010.05.001>
- Michaud, G., & Ammar, A. (2023). Explicit Instruction within a Task: Before, During, or After? *Studies in Second Language Acquisition*, 45(2), 442–460. <https://doi.org/10.1017/S0272263122000316>
- Mucha, H., Robert, S., Breitschwerdt, R., & Fellmann, M. (2021, May). Interfaces for Explanations in Human-AI Interaction: Proposing a Design Evaluation Approach. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–6). <https://doi.org/10.1145/3411763.3451759>
- Nassif, M., & Robillard, M. P. (2025). Non-Linear Software Documentation with Interactive Code Examples. *ACM Transactions on Software Engineering and Methodology*, 34(2), 1–32. <https://doi.org/10.1145/3702976>
- Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, 2, 100041. <https://doi.org/10.1016/j.caeai.2021.100041>
- Ning, Y., Zhang, Z., Lal, Y. K., Gou, B., Li, J., Ruan, W., Ye, C., Gupta, R., Yang, D., Su, Y., & Sun, H. (2026, June). SkillHarm: Lifecycle-Aware Skill-Based Attacks via Automated Construction [arXiv:2606.02540 [cs.CL]]. <https://doi.org/10.48550/arXiv.2606.02540>
- Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654), 187–192. <https://doi.org/10.1126/science.adh2586>
- OpenAI. (2025, June). The state of enterprise AI. *OpenAI*. <https://openai.com/index/the-state-of-enterprise-ai-2025-report/>
- OpenAI. (2026). Agent Skills Overview. <https://agentskills.io/home>
- Openai/skills. (2026, June). <https://github.com/openai/skills>
- Ortloff, A.-M., Martius, F., Meier, M., Raimbault, T., Geierhaas, L., & Smith, M. (2025, April). Small, Medium, Large? A Meta-Study of Effect Sizes at CHI to Aid Interpretation of Effect Sizes and Power Calculation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (pp. 1–28). <https://doi.org/10.1145/3706598.3713671>

- Phillips, P. J., Hahn, C., Fontana, P., Yates, A., Greene, K. K., Broniatowski, D. A., & Przybicki, M. A. (2021, September). *Four Principles of Explainable Artificial Intelligence* (NIST Interagency/Internal Report (NISTIR) No. 8312). National Institute of Standards and Technology. Gaithersburg, MD. <https://doi.org/10.6028/NIST.IR.8312>
- Pre-training Principle. (2009). In R. E. Mayer (Ed.), *Multimedia Learning* (2nd ed., pp. 189–199). Cambridge University Press. <https://doi.org/10.1017/CBO9780511811678.014>
- Rader, E., Cotter, K., & Cho, J. (2018). Explanations as Mechanisms for Supporting Algorithmic Transparency. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13. <https://doi.org/10.1145/3173574.3173677>
- Radosevich, B., & Halloran, J. (2025, April). MCP Safety Audit: LLMs with the Model Context Protocol Allow Major Security Exploits [arXiv:2504.03767 [cs.CR]]. <https://doi.org/10.48550/arXiv.2504.03767>
- Rutjes, H., Willemsen, M., & IJsselsteijn, W. (2019). Considerations on explainable AI and users' mental models: CHI 2019 Workshop. *Where is the Human? Bridging the Gap Between AI and HCI*.
- Saha, S., Faghieh, K., & Feizi, S. (2026, May). Under the Hood of SKILL.md: Semantic Supply-chain Attacks on AI Agent Skill Registry [arXiv:2605.11418 [cs.AI]]. <https://doi.org/10.48550/arXiv.2605.11418>
- Sallam, M. (2023). ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare*, 11(6), 887. <https://doi.org/10.3390/healthcare11060887>
- Sanchez, T., Vereschak, O., & Deroy, O. (2026, March). Mental Models in Human-AI Interaction: Systematic Review of Empirical Methodologies and Guidelines. In *Proceedings of the 31st International Conference on Intelligent User Interfaces* (pp. 663–682). <https://doi.org/10.1145/3742413.3789223>
- Schoeffer, J., De-Arteaga, M., & Köhl, N. (2024, May). Explanations, Fairness, and Appropriate Reliance in Human-AI Decision-Making. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (pp. 1–18). <https://doi.org/10.1145/3613904.3642621>
- Sindermann, C., Sha, P., Zhou, M., Wernicke, J., Schmitt, H. S., Li, M., Sariyska, R., Stavrou, M., Becker, B., & Montag, C. (2021). Assessing the Attitude Towards Artificial Intelligence: Introduction of a Short Measure in German, Chinese, and English Language. *KI - Künstliche Intelligenz*, 35(1), 109–118. <https://doi.org/10.1007/s13218-020-00689-0>
- Strzelecki, A. (2024). Students' Acceptance of ChatGPT in Higher Education: An Extended Unified Theory of Acceptance and Use of Technology. *Innovative Higher Education*, 49(2), 223–245. <https://doi.org/10.1007/s10755-023-09686-1>
- Subramonyam, H., Pea, R., Pondoc, C., Agrawala, M., & Seifert, C. (2024). Bridging the Gulf of Envisioning: Cognitive Challenges in Prompt Based Interactions with LLMs. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–19. <https://doi.org/10.1145/3613904.3642754>
- Sustainable Cabin. (2019). Sustainable Cabin. *International Air Transport Association*. <https://www.iata.org/en/programs/sustainability/cabin-waste/>

- Tankelevitch, L., Kewenig, V., Simkute, A., Scott, A. E., Sarkar, A., Sellen, A., & Rintel, S. (2024). The Metacognitive Demands and Opportunities of Generative AI. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–24. <https://doi.org/10.1145/3613904.3642902>
- Valkenburg, R., & Dorst, K. (1998). The Reflective Practice of Design Teams. *Design Studies*, 19(3), 249–271. [https://doi.org/10.1016/S0142-694X\(98\)00011-8](https://doi.org/10.1016/S0142-694X(98)00011-8)
- Van Der Meij, H., & Gellevij, M. (2004). The four components of a procedure. *IEEE Transactions on Professional Communication*, 47(1), 5–14. <https://doi.org/10.1109/TPC.2004.824292>
- van der Waa, J., Nieuwburg, E., Cremers, A., & Neerincx, M. (2021). Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, 291, 103404. <https://doi.org/10.1016/j.artint.2020.103404>
- Vasileiou, K., Barnett, J., Thorpe, S., & Young, T. (2018). Characterising and justifying sample size sufficiency in interview-based studies: Systematic analysis of qualitative health research over a 15-year period. *BMC Medical Research Methodology*, 18, 148. <https://doi.org/10.1186/s12874-018-0594-7>
- Weber, T., Brandmaier, M., Schmidt, A., & Mayer, S. (2024). Significant Productivity Gains through Programming with Large Language Models. *Proceedings of the ACM on Human-Computer Interaction*, 8(EICS), 1–29. <https://doi.org/10.1145/3661145>
- Yang, Q., Steinfeld, A., Rosé, C., & Zimmerman, J. (2020, April). Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). <https://doi.org/10.1145/3313831.3376301>
- Yildirim, N., Kass, A., Tung, T., Upton, C., Costello, D., Giusti, R., Lacin, S., Lovic, S., O'Neill, J. M., Meehan, R. O., Ó Loideáin, E., Pini, A., Corcoran, M., Hayes, J., Cahalane, D. J., Shivhare, G., Castoro, L., Caruso, G., Oh, C., . . . Zimmerman, J. (2022, April). How Experienced Designers of Enterprise Applications Engage AI as a Design Material. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). <https://doi.org/10.1145/3491102.3517491>
- Yildirim, N., Pushkarna, M., Goyal, N., Wattenberg, M., & Viégas, F. (2023, April). Investigating How Practitioners Use Human-AI Guidelines: A Case Study on the People + AI Guidebook. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). <https://doi.org/10.1145/3544548.3580900>
- Zamfirescu-Pereira, J., Wong, R. Y., Hartmann, B., & Yang, Q. (2023). Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–21. <https://doi.org/10.1145/3544548.3581388>
- Zhan, Q., Liang, Z., Ying, Z., & Kang, D. (2024, August). InjecAgent: Benchmarking Indirect Prompt Injections in Tool-Integrated Large Language Model Agents [arXiv:2403.02691 [cs.CL]]. <https://doi.org/10.48550/arXiv.2403.02691>
- Zhou, E., & Lee, D. (2024). Generative artificial intelligence, human creativity, and art. *PNAS Nexus*, 3(3), pgae052. <https://doi.org/10.1093/pnasnexus/pgae052>

