# Can AI Design a Good UX Study?

Presented by Theresa Wilkinson

## Can AI Design a Good UX Study?





Agenda

1. Executive Summary
2. Research Question Findings
3. Participant Findings
4. Sample Size Findings
5. Usability Tasks Findings

# Executive Summary

- Compared AI-generated UX research plans from **ChatGPT, Claude, Gemini, and Copilot**.
- All four AI tools produced usable research plans but varied in quality, completeness, and consistency.
- The largest differences were observed in participant selection, sample size, research questions, and usability tasks.
- AI is a valuable tool for accelerating research planning but requires human review before implementation.
- AI reduced drafting time but increased review and refinement time.

## Implications for UX Researchers

- AI can accelerate early UX research planning.
- AI consistently identifies core UX research components.
- AI-generated recommendations vary in quality, completeness, and practicality.
- Human expertise remains essential to validate and refine AI-generated research plans.

## Study Overview

- Evaluated the ability of **ChatGPT, Claude, Gemini, and Copilot** to design a qualitative UX research study.
- Compared AI-generated research plans for a veterinary appointment scheduling application.
- Assessed AI-generated deliverables across research questions, participant criteria, sample size, usability tasks, rationale, assumptions, and recommendations.

# Research Questions

1. Can AI generate a complete and effective UX research plan?
2. How do AI-generated research plans differ in quality and completeness?
3. Which UX research components are consistently identified across AI models?
4. Where do AI-generated recommendations require human expertise and refinement?
5. Which AI model produces the most practical and usable deliverables with the least amount of editing required?
6. How well do AI-generated research plans align with established UX research best practices?

# FINDINGS

## Overall Findings

- All four AI models produced usable starting points for UX research planning.
- AI models consistently identified core UX research components but differed in completeness, quality, and practicality.
- No single AI generated a complete research plan.
- Human expertise remained necessary to validate and refine AI-generated recommendations.

## Recommendations

- Use AI to accelerate early UX research planning.
- Validate AI-generated research plans before implementation.
- Apply human expertise to refine AI-generated recommendations.

# Research Question Findings

- ChatGPT generated the most comprehensive research questions.
- Only Gemini included accessibility as a research objective.
- No AI model identified readability or comprehension.
- Several research questions were overly verbose and required editing.

| AI | Impressions |
|---|---|
| ChatGPT | **4** – Generated the most comprehensive set of research questions, covering core tasks, accessibility, digital literacy, terminology, navigation, error recovery, reminders, and trust, although several questions were verbose and overlapping. |
| Claude | **3.5** – Generated focused research questions covering the primary workflows, accessibility, reminders, and multi-pet management, but included one compound question that mixed a research question with interpretation. |
| Gemini | **3** – Generated fewer, broader research questions that emphasized task success, digital literacy, accessibility, and information architecture, resulting in less comprehensive coverage than the other AI models. |
| Copilot | **3** – Generated focused research questions covering the application's primary workflows and accessibility requirements, but several questions included explanatory rationale and provided less comprehensive coverage than ChatGPT. |

# Participant Findings

- Claude generated the most comprehensive participant criteria.
- Participant recruitment recommendations varied across AI models.
- Only one AI model recommended a broad adult age range (18–75+).
- No AI model fully addressed participant diversity, accessibility, and platform experience.
- Human review was needed to ensure representative participant recruitment.

| AI | Impressions |
|---|---|
| ChatGPT | **4** – Strong coverage of digital literacy, accessibility, pet ownership, and assistive technology, but no age range or platform mix. |
| Claude | **5** – The most complete. Includes age range, digital literacy, platforms, prescription refill experience, and accessibility cohort. |
| Gemini | **3.5** – Good segmentation, but missing age diversity and platform mix. |
| Copilot | **3.5** – Good participant characteristics, but the focus on iterative rounds belongs more under sample size than participant criteria, and it lacks age diversity and platforms. |

# Sample Size Findings

- Sample size recommendations ranged from 8 to 48 participants.
- Testing recommendations ranged from one to six rounds.
- ChatGPT recommended the most practical sample size; Copilot recommended the largest.
- Several AI models recommended iterative testing without considering budget or timeline constraints.
- Human judgment was needed to align sample size recommendations with project constraints.

| AI | Impressions |
|---|---|
| ChatGPT | ***5*** – Recommended a practical sample size (8–12 participants) with a single round of testing. |
| Claude | ***3.5*** – Recommended a larger sample (13–15 participants) and two rounds of testing. |
| Gemini | ***3.5*** – Recommended a moderate sample (12–15 participants) with one to two rounds of testing. |
| Copilot | ***3.5*** – Recommended the largest sample (20–48 participants) and four to six rounds of iterative testing. |

# Usability Tasks Findings

- ChatGPT generated the most comprehensive set of usability tasks.
- Task coverage varied considerably across AI models.
- Copilot emphasized realistic, multi-step user scenarios.
- Accessibility-related tasks were included inconsistently.
- Human review was needed to ensure comprehensive task coverage.

| AI | Impressions |
|---|---|
| ChatGPT | **5** – Most comprehensive task set covering core workflows, error recovery, reminder settings, and accessibility. |
| Claude | **3.5** – Focused on the primary workflows with realistic task scenarios but fewer tasks overall. |
| Gemini | **3.5** – Included realistic, scenario-based tasks with a strong accessibility task, but fewer overall tasks. |
| Copilot | **4** – Focused on core workflows with detailed scenarios, but omitted accessibility testing and several secondary tasks. |