Can AI Design a Good UX Study?

Evaluating AI's Ability to Develop High-Quality UX Research Plans

Research Plan

By Theresa Wilkinson

Study Objectives

This study evaluated whether ChatGPT, Claude, Gemini, and Copilot could develop a usable qualitative UX research plan. The goal was to compare the completeness, accuracy, quality, and practicality of their recommendations and identify where human expertise was still needed.

Study Design

Each AI will receive the same project scenario and identical prompt to create a consistent basis for comparison. The responses will be reviewed independently and compared across research questions, participant criteria, sample-size recommendations, usability tasks, rationale, and assumptions.

Study Prompt

To ensure a fair comparison, each AI model will receive the identical project scenario and prompt. Using the same prompt for all AI models should minimize variability and allow differences in the responses to be attributed to the AI models rather than differences in the instructions provided.

You are a Senior UX Researcher. You have been asked to design a usability study for a new veterinary appointment scheduling application. The application allows pet owners to schedule appointments, select a veterinarian, manage multiple pets, request prescription refills, view appointment history, and receive appointment reminders. The application serves adults with varying levels of digital literacy and must comply with WCAG 2.2 accessibility guidelines.

Developing the following deliverables:

1. Research questions.
2. Participant criteria, including the recommended sample size and rationale.
3. Realistic usability task scenarios that align with the research questions.
4. For each deliverable, explain the rationale behind your recommendations and identify any assumptions you made.

Evaluation Framework

The AI-generated deliverables were evaluated for completeness, accuracy, quality, practicality, and the amount of editing required. Research questions, participant criteria, sample-size recommendations, and usability tasks were reviewed separately. Each area was rated using the same five-point scale, with half-point scores used when a response fell between two rating levels.

Study Questions

1. Can AI generate good research questions?
2. Can AI identify the appropriate participants?
3. Can AI recommend the correct sample size?
4. Can AI justify the sample size?
5. Can AI create good usability tasks?

Evaluation Framework

Each AI-generated deliverable was evaluated using a consistent set of questions designed to assess its

completeness, accuracy, quality, practicality, and overall usefulness. The following criteria were applied to each section of the research plan to ensure a fair and consistent comparison across all AI models.

Research Questions

- **Completeness:** Did it provide enough research questions?
- **Accuracy:** Are the questions appropriate for a usability study?
- **Quality:** Are they useful and well written?
- **Practicality:** Would I actually use these?
- **Editing required:** How much would I have to change?

Participant Criteria

- Did it identify the right users?
- Was the sample size reasonable?
- Were accessibility needs considered?
- Would I recruit from this plan?

Usability Tasks

- Do the tasks reflect realistic user goals?
- Do they answer the research questions?
- Are they unbiased?
- Are they complete?

Scale

Each evaluation criterion was rated using a 5-point scale. Half-point ratings were assigned when a deliverable fell between two rating levels.

- **5 – Excellent:** Fully meets expectations with little or no editing required.
- **4 – Good:** Minor issues that require minimal editing.
- **3 – Acceptable:** Meets basic expectations but requires moderate editing.
- **2 – Needs Improvement:** Significant weaknesses requiring substantial revision.
- **1 – Poor:** Does not adequately meet the evaluation criterion.

Limitations

- This study evaluated only four AI models available at the time of the research.
- Results reflect the responses generated from a single prompt and project scenario.
- AI models evolve over time, so results may differ as models are updated.
- The evaluation was conducted by a single experienced UX researcher using predefined evaluation criteria.
- The study focused on qualitative UX research planning and did not evaluate AI performance during study moderation, data analysis, or report writing.