Can AI Accurately Analyze UX Research Data?

Evaluating AI's Ability to Analyze Quantitative UX Research Data

Test Plan

By Theresa Wilkinson

Study Objectives

This study will evaluate whether ChatGPT, Claude, Gemini, and Copilot can analyze UX evaluation results. The goal is to compare the quality of their analyses, identify differences in how each AI interprets the same dataset, and determine where human expertise remains essential for producing evidence-based conclusions and recommendations.

Study Design

Each AI will receive the same evaluation results and identical prompt to ensure a consistent basis for comparison. The prompt will instruct each AI to analyze the evaluation results and produce an executive summary, key conclusions, implications for UX researchers, limitations, and recommendations. The responses will be reviewed independently and compared across each deliverable.

Study Prompt

The responses will be reviewed independently and compared across executive summaries, key conclusions, implications for UX researchers, limitations, recommendations, and the overall depth and quality of the analysis.

Prompt: Review and analyze the attached evaluation results. Write an executive summary, identify the key conclusions, discuss the implications for UX researchers, identify limitations, and provide recommendations based only on the data provided. Do not make assumptions beyond the data.

Evaluation Framework

The AI-generated analyses will be evaluated for completeness, accuracy, quality, practicality, and the amount of editing required. The evaluation will also assess whether each AI meaningfully analyzes the evaluation results by identifying patterns, supporting conclusions with evidence, and synthesizing findings across the dataset. Executive summaries, key conclusions, implications for UX researchers, limitations, and recommendations will be reviewed separately. Each area will be rated using the same five-point scale, with half-point scores assigned when a response falls between two rating levels.

Research Questions

1. Can AI meaningfully analyze UX evaluation results rather than simply summarize them?
2. Can AI identify meaningful patterns and relationships within UX evaluation results?
3. Can AI support its conclusions with evidence from the evaluation data?
4. Can AI synthesize findings into meaningful conclusions and implications for UX researchers?
5. Can AI generate new insights beyond those already identified in the evaluation results?

Evaluation Criteria

- Completeness: Did it summarize the study accurately?
- Accuracy: Was it supported by the evaluation results?
- Quality: Was it clear and well organized?

- Practicality: Would I use it as written?
- Editing required: How much would I have to change?
- Analysis: Did the AI meaningfully analyze the evaluation results?

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 will evaluate only four AI models available at the time of the research.
- Results will reflect responses generated from a single prompt and evaluation dataset.
- AI models evolve over time, so results may differ as models are updated.

- The evaluation will be conducted by a single experienced UX researcher using predefined evaluation criteria.
- The study will evaluate AI's ability to analyze UX evaluation results and will not assess AI performance in other research activities, such as study planning, participant recruitment, moderation, or data collection.