Can AI Accurately Interpret UX Research Data?

Evaluating AI's Ability to Interpret UX Research Results

Test Plan

By Theresa Wilkinson

Study Objectives

The objective of this study was to evaluate the ability of four generative AI models—ChatGPT, Claude, Gemini, and Microsoft Copilot—to analyze UX evaluation results and generate evidence-based research summaries.

Specifically, the study examined each model's ability to:

- Analyze evaluation results rather than summarize individual findings.
- Identify patterns, similarities, and differences across AI models.
- Develop evidence-based conclusions supported by the data.
- Discuss implications for UX researchers.
- Identify study limitations.
- Provide actionable recommendations.
- Produce clear, accurate, and practical deliverables requiring minimal editing.

Study Design

Four generative AI models were evaluated using the same UX evaluation dataset and standardized prompts. Each model independently analyzed the evaluation results and generated a written report.

Responses were compared using a predefined evaluation framework to assess the completeness, accuracy, quality, practicality, and editing required for each deliverable.

Study Prompt

Each AI model received the same evaluation results and was asked to analyze the findings using standardized prompts.

Example prompts included:

- Review and analyze the attached evaluation results.
- Write an executive summary.
- Identify key conclusions.
- Discuss the implications for UX researchers.
- Identify study limitations.
- Provide evidence-based recommendations.
- Based all conclusions solely on the evidence provided. Do not make assumptions beyond the data.

Evaluation Framework

Each response was evaluated using a structured framework designed to assess the quality and usefulness of AI-generated research analysis.

Responses were evaluated for:

- Completeness
- Accuracy
- Quality
- Practicality
- Editing Required

Both the individual deliverables and the overall usefulness of each response were compared across all four AI models.

Study Questions

The study sought to answer the following questions:

- Can generative AI accurately analyze UX evaluation results?
- Can AI identify meaningful patterns across multiple evaluation datasets?
- Can AI distinguish between evidence-supported findings and unsupported assumptions?
- Can AI produce practical research deliverables suitable for UX professionals?
- Which AI model produces the highest-quality analytical output?

Analysis Materials

Each AI model analyzed the same materials:

- AI evaluation results spreadsheet
- Individual model evaluations
- Comparative evaluation data
- Standardized evaluation criteria
- Standardized prompts

Analysis Tasks

Each AI model was asked to:

1. Review the evaluation results.
2. Analyze patterns across all four AI models.
3. Write an executive summary.
4. Identify key conclusions.
5. Discuss implications for UX researchers.
6. Identify study limitations.
7. Provide evidence-based recommendations.

Limitations

This study evaluated AI performance using a single UX evaluation dataset and a standardized set of prompts.

Results reflect the capabilities of the evaluated AI models at the time of testing and may vary as models are updated. The study focused on the quality of AI-generated analysis rather than response speed, cost, or user preference. Conclusions are limited to the evidence generated during this evaluation and should not be generalized beyond the scope of the study.