# Can AI Accurately Analyze UX Research Data?

## Comparing ChatGPT, Claude, Gemini, and Copilot





Presented by Theresa Wilkinson

# Can AI Accurately Analyze UX Research Data?

## Agenda

1. Executive Overview
2. Implications for UX Researchers
3. Study Overview
4. Findings
5. Recommendations

# Executive Overview

- All four AI models identified some meaningful findings, but the **analyses varied substantially in accuracy, completeness, and organization**.
- Quantitative errors, missed findings, and unsupported or inaccurate interpretations were identified.
- Report organization made analyses difficult to review and verify.
- **None of the AI models recommended inferential statistical analysis** to evaluate differences among the three age groups.
- Human verification was required before the analyses could be trusted.
- In some cases, **verifying the AI output required as much or more effort than performing the analysis directly.**
- This need for human verification is consistent with current professional AI workflows that use AI to accelerate analysis while retaining human review before the output is used [1].

[1] Source: The Rundown, *10 Ways The Rundown Uses AI*

# Implications for UX Researchers

- AI can assist with quantitative UX research analysis, but the output should not be accepted without verification.
- Researchers need sufficient quantitative knowledge to recognize calculation errors and unsupported interpretations.
- AI-generated findings should be checked against the source data before they are reported to stakeholders.
- AI may provide greater value as a research assistant working collaboratively with an experienced researcher than as an independent analyst.
- This human-in-the-loop approach is also used in professional AI workflows, where AI-generated work is reviewed and edited before it is used [1] .

[1] Source: The Rundown, *10 Ways The Rundown Uses AI*

# Study Overview

- Evaluated the ability of ChatGPT, Claude, Gemini, and Copilot to analyze the same participant-level first-click usability dataset.
- Dataset included 30 participants: 10 ages 45–54, 10 ages 55–64, and 10 ages 65+.
- Each model received the same dataset and analysis prompt.
- AI outputs were reviewed against the source data to identify accurate findings, errors, omissions, and differences in analysis.
- The planned 5-point evaluation framework was not used in the final analysis; the AI outputs were instead evaluated directly against the source data for accuracy, completeness, and quality of analysis.

# FINDINGS

## Overall Findings

- All four AI models identified at least some important usability findings.
- The models differed substantially in what they reported, emphasized, and calculated.
- Errors were not consistent across models; different models produced different types of problems.
- No AI analysis could be accepted as accurate without checking it against the source data.
- The evaluation shifted from simply comparing AI reports to evaluating whether the analyses could be trusted without extensive human verification.

## Overall Takeaway

AI generated useful observations, but reliability varied enough that human review remained essential.

## Report Organization Findings

- All four reports were difficult to review and verify.
- Findings were not consistently presented in a clear, traceable structure.
- Some reports mixed tasks, age groups, and findings instead of following the requested organization.
- Poor organization increased the effort required to compare reported findings with the source data.

## Overall Takeaway

Report organization was not simply a presentation issue; it directly affected the ability to verify the analysis.

## Quantitative Accuracy Findings

- Multiple quantitative errors were identified.
- Some models accurately calculated many results but still included incorrect values.
- Some important results were omitted.
- Copilot primarily summarized the results and provided little numerical evidence to support its conclusions.
- Because errors differed across models, reported values had to be checked individually against the dataset.

## Overall Takeaway

AI-generated quantitative results could not be assumed to be correct.

## Time-on-Task Findings

- Time-on-task produced some of the clearest quantitative problems.
- Incorrect or substantially inflated time values appeared in the AI analyses.
- Some incorrect time values affected the interpretation of whether participants struggled with a task.
- Means, medians, and individual participant times had to be checked against the dataset.
- These were substantive calculation errors rather than minor rounding differences.

## Overall Takeaway

Incorrect time calculations could lead directly to incorrect conclusions about usability problems.

## Statistical Analysis Findings

- None of the four AI models recommended conducting an inferential statistical analysis to evaluate age-group differences.
- Some models instead relied on general cautions about the small sample size.
- A one-way ANOVA was an appropriate analysis to consider for comparing time-on-task across the three independent age groups.
- The ANOVA was **not provided to the AI models**, so the finding is that they did not independently recommend an appropriate inferential test.

## Overall Takeaway

The models identified descriptive patterns but did not independently move from descriptive analysis to appropriate inferential testing.

## Human Verification Findings

- Calculations had to be checked against the original dataset.
- Reported findings had to be reviewed for accuracy.
- Missed findings had to be identified manually.
- Interpretations had to be checked to determine whether they were supported by the data.
- The inconsistent organization of the reports increased verification time.
- At times, performing the analysis directly would have been faster than verifying the AI output.

## Overall Takeaway

The researcher remained the quality control throughout the analysis.

# Recommendations

- Use AI as an analysis assistant rather than as an independent source of quantitative findings.
- Verify calculations against the source data before reporting results.
- Require AI to present quantitative findings in a consistent, traceable structure.
- Distinguish descriptive patterns from statistically supported conclusions.
- Ask AI to recommend appropriate statistical tests when analyzing quantitative UX datasets.
- Consider using multiple AI models to cross-check analyses rather than relying on a single model, while still performing a final human review. The Rundown describes a similar parallel-review workflow in which several AI models review the same material before the results are consolidated and manually checked [1].
- Apply experienced researcher judgment before using AI-generated findings or recommendations.

[1] Source: The Rundown, *10 Ways The Rundown Uses AI*