

Can AI Accurately Analyze UX Survey Data?

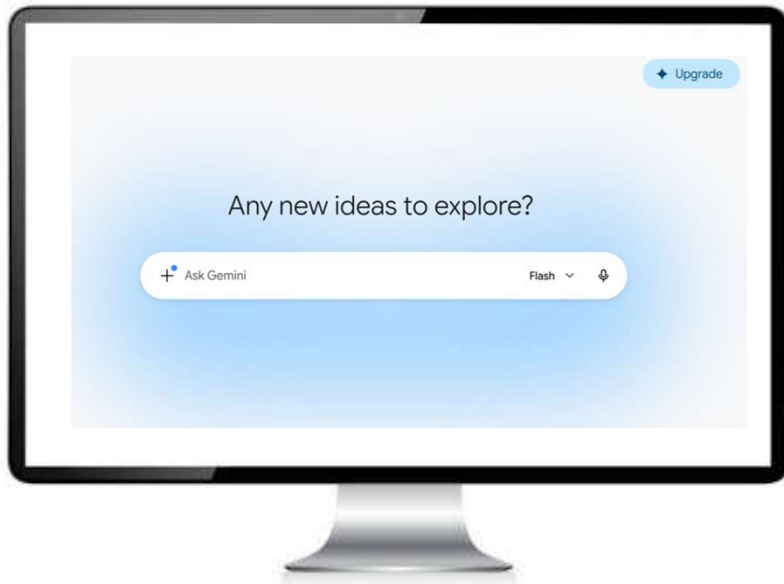
Survey Findings

*What four AI tools got right, what they got wrong, and why
human verification was needed*



Presented by Theresa Wilkinson

Can AI Accurately Analyze UX Survey Data? Survey Findings



Agenda

1. Executive Summary
2. Summary of Key Results
3. Comparison Across the Four AIs
4. Overall Takeaway
5. Human Verification Required
6. Findings by Question
7. Results by AI

Executive Summary

Purpose & Background

This study evaluated how accurately ChatGPT, Claude, Gemini, and Copilot analyzed the same UX research survey data.

Each AI's output was compared with the researcher's verified results to identify accurate findings, errors, and the human verification needed to use its analysis.

Methods / Participants

AI tools evaluated: ChatGPT, Claude, Gemini, and Copilot.

Each received the same survey data and prompt. Their counts and analyses of written responses were compared with the researcher's verified analysis.

The source survey contained responses from **34 UX research professionals**.

Summary of Key Results

- All four AIs correctly counted Q1–Q3 and Q7–Q8.
- All four made counting errors on Q4 and Q5. On Q5, all four correctly counted the same **3 of 9 activities (33%)**.
- Copilot's Q6 counts were incorrect; the other three AIs' counts were correct.
- None produced usable groupings and counts for Q9's written responses. Q10 was also marked as a failure but did not receive a separate detailed review.

Overall Takeaway

The AIs accurately counted several survey questions, but their results became **less reliable when they had to track multiple distinct response items or group written answers**. Errors appeared in both the counts and the categories used to describe written responses. **Each AI output required human verification.**

Comparison Across the Four AIs

- All four correctly counted Q1–Q3 and Q7–Q8.
- All four made errors on Q4 and Q5.
- Their Q9 problems differed, but none produced usable categories and counts.
- Correct results on some questions did not predict accuracy across the rest of the survey.

Overall Takeaway

The AIs accurately counted several survey questions, but their results became less reliable when they had to track multiple distinct response items or group written answers.

Errors appeared in both the counts and the categories used to describe written responses. Each AI output required human verification.

Human Verification Required

The researcher checked the AIs' counts against the survey data and rebuilt the written-response categories. The additional statistical analyses produced by the AIs were **not evaluated**.

Key Finding

Correct counts on some questions did not make an AI's complete analysis reliable.

Findings by Question

Results by Question

- **Q4:** Every AI miscounted three AI tool entries.
- **Q5:** All four AIs made counting errors. Copilot counted 6 of 9 activities correctly; Claude counted 5 of 9; ChatGPT and Gemini each counted 4 of 9.
- **Q6:** Copilot's counts were incorrect; the other three AIs' counts were correct.
- **Q9:** None produced usable categories and counts for the written responses.
- **Q10:** Marked as a failure because the AIs could not correctly group the written responses. A separate detailed review was not completed.

Overall Takeaway

All four AIs made errors in the questions that required them to count separate response items or group written ideas.

Questions 1–3: Counting Respondent Information

All four AIs reported the correct counts for:

- Q1: Respondents' primary roles
- Q2: Years of UX/research experience
- Q3: Frequency of AI use for UX research work

Key Finding

All four AIs counted every response category in these three questions correctly.

Questions 4: Counting AI Tool Selections

Respondents could select more than one tool. Each AI miscounted **three tool entries**. All four got the Gemini and Perplexity counts wrong.

AI	Tool entries counted incorrectly
ChatGPT	Claude, Gemini, Perplexity
Claude	Claude, Gemini, Perplexity
Gemini	Claude, Gemini, Perplexity
Copilot	Copilot, Gemini, Perplexity

Key Finding

All four AIs made counting errors on Q4.

Question 5: Counting Trust Ratings

All four AIs made errors counting the nine research activities. Copilot counted 6 of 9 (66.7%) correctly; Claude counted 5 of 9 (55.6%); ChatGPT and Gemini each counted 4 of 9 (44.4%). Each activity's results required verification against the survey data.

AI	Activities counted correctly
ChatGPT	4 of 9
Claude	5 of 9
Gemini	4 of 9
Copilot	6 of 9

Key Finding

Every AI made errors counting Q5's separate research activities, although the number of errors varied by AI.

Question 6: Counting Human Verification Responses

ChatGPT, Claude, and Gemini reported the correct Q6 counts. Copilot's Q6 counts were incorrect.

Key Finding

Copilot was the only AI that failed to count Q6 correctly.

Questions 7 and 8 — Counting

All four AIs reported the correct counts for both questions.

Key Finding

Every AI counted Q7 and Q8 accurately.

Question 9: Written Responses

The researcher identified **94 mentions** across distinct categories. None of the four AIs produced categories and counts suitable for that analysis.

- ChatGPT's counts were difficult to follow and included incorrect respondent totals.
- Claude's Q9 categories and counts were difficult to use; it counted respondents rather than mentions.
- Gemini did not group the comments.
- Copilot grouped some themes but did not count them.

Key Finding

None of the AIs reliably separated and counted the distinct ideas in Q9's written responses.

Questions 10 — Written Responses

The researcher identified **57 mentions**, including **44 mentions of AI reliability concerns**. Q10 was marked as a failure because the AI outputs did not provide usable groupings.

A separate detailed review of each AI's Q10 output was not completed.

Key Finding

The AI outputs did not provide usable groupings for Q10.

Results by AI

ChatGPT Results

- Correct counts: Q1–Q3, Q7–Q8, and 4 of 9 Q5 activities.
- Q4 errors: Claude, Gemini, and Perplexity counts.
- Q9: Counts were difficult to follow and included incorrect respondent totals. The output did not provide usable counts of distinct ideas.

Key Finding

ChatGPT produced many correct counts, but its Q4, Q5, and Q9 results required verification.

Claude Results

- Correct counts: Q1–Q3, Q6–Q8, and 5 of 9 Q5 activities.
- Q4 errors: Claude, Gemini, and Perplexity counts.
- Q9: Produced counts but counted respondents rather than mentions.

Key Finding

Claude's Q9 counts did not measure the distinct ideas the researcher needed to count.

Gemini Results

- Correct counts: Q1–Q3, Q6–Q8, and 4 of 9 Q5 activities.
- Q4 errors: Claude, Gemini, and Perplexity counts.
- Q9: Did not group the comments.

Key Finding

Gemini counted several questions correctly but did not complete the Q9 grouping task.

Copilot Results

- Correct counts: Q1–Q3, Q7–Q8, and 6 of 9 Q5 activities.
- Q4 errors: Copilot, Gemini, and Perplexity counts.
- Q6: All counts were incorrect.
- Q9: Grouped some themes but did not count them.

Key Finding

Copilot counted the most Q5 activities correctly.

Survey Questions

Appendix

Survey Questions

1. What is your primary role?
2. How many years of professional UX/research experience do you have?
3. How often do you currently use generative AI for UX research-related work?
4. Which AI tools have you used for UX research work? (Select all that apply.)
5. How much would you trust AI to help with each of the following? [Developing research questions]
need each of these
6. How much human review do you think AI-generated UX research work requires?
7. How comfortable would you be presenting AI-generated research findings to stakeholders if you had reviewed the work yourself?
8. Have you ever found an error or an unsupported conclusion in AI-generated UX research work?
9. What UX research activity do you trust AI with the MOST, and why?
10. What UX research activity do you trust AI with the LEAST, and why?