

Can AI Accurately Analyze UX Survey Data?

Survey Findings

*What two 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 of Dovetail and Perplexity
4. Overall Takeaway
5. Human Verification Required
6. Findings by Question
7. Results by AI

Executive Summary

Purpose & Background

This study evaluated how accurately Dovetail and Perplexity 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: Dovetail and Perplexity.

Both tools 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

- Both tools correctly counted Q1–Q3 and Q6–Q8.
- Both made errors on Q4 and on the same four Q5 activities. Each correctly counted 5 of 9 (55%) Q5 activities.
- Perplexity grouped and counted Q9 responses, but its counts were off by one or two. Dovetail did not provide usable counts of the distinct ideas.
- Neither tool provided usable Q10 groupings and counts.

Overall Takeaway

Both tools correctly counted several survey questions, but **neither produced a complete analysis that could be used without human verification.**

Comparison of Dovetail and Perplexity

Result	Dovetail	Perplexity
Q1–Q3, Q6–Q8	Correct	Correct
Q4 tool entries miscounted	3 of 5 (60%)	4 of 5 (80%)
Q5 activities counted correctly	5 of 9 (55.5%)	5 of 9 (55.5%)
Q9 written responses	No usable counts of distinct ideas	Counts close, but off by one or two
Q10 written responses	No usable groupings and counts	Confusing groupings; no usable result

Overall Takeaway

Perplexity came closer on Q9. Both tools made counting errors elsewhere, including the same four Q5 activities.

Human Verification Required

The researcher checked both tools' counts against the verified survey analysis and assessed whether their written-response categories captured the distinct ideas participants mentioned. The additional statistical analyses in the AI outputs were not evaluated.

Key Finding

Correct answers on several questions did not establish that the full analysis was accurate.

Findings by Question

Results by Question

- **Q1–Q3:** Both tools counted all response categories correctly.
- **Q4:** Dovetail miscounted three tool entries; Perplexity miscounted all five entries.
- **Q5:** Both counted 5 of 9 activities correctly and missed the same four.
- **Q6–Q8:** Both reported the correct counts.
- **Q9:** Perplexity came close to the verified counts. Dovetail did not provide usable counts of distinct ideas.
- **Q10:** Neither produced usable groupings and counts.

Overall Takeaway

Both tools needed human verification for the questions involving multiple selections, separate trust activities, and written responses.

Questions 1-3: Counting Respondent Information

Dovetail and Perplexity both reported the correct counts for:

- Q1: Primary role
- Q2: Years of UX/research experience
- Q3: Frequency of AI use for UX research work

Key Finding

Both tools counted every response category in these three questions correctly.

Question 4: Counting AI Tool Selections

Respondents could select more than one tool. Both AIs made counting errors.

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

Key Finding

Both tools correctly counted research questions, participant screeners, qualitative analysis, patterns across findings, and conclusions. **Both miscounted usability test tasks, quantitative analysis, recommendations, and stakeholder summaries or reports.**

Question 5: Counting Trust Ratings

Both tools correctly counted research questions, participant screeners, qualitative analysis, patterns across findings, and conclusions. Both miscounted usability test tasks, quantitative analysis, recommendations, and stakeholder summaries or reports.

AI	Activities counted correctly
Dovetail	5 of 9 (55.5%)
Perplexity	5 of 9 (55.5%)

Key Finding

The tools made errors on the same four research activities.

Questions 7 & 8 — Counting Responses

Both tools reported the correct counts for comfort presenting reviewed AI findings (Q7) and experience finding AI errors or unsupported conclusions (Q8).

Key Finding

Both tools counted Q7 and Q8 correctly.

Question 9: Written Responses

The researcher identified **94 mentions** across distinct categories.

- Perplexity grouped and counted the responses, but its counts were off by one or two.
- Dovetail provided extensive explanation but did not provide usable counts of the distinct ideas.

Key Finding

Neither tool reproduced the researcher's Q9 analysis exactly. Perplexity came closest.

Questions 10 — Written Responses

The researcher identified **57 mentions**, including **44 mentions of AI reliability concerns**.

- Dovetail did not provide usable counts of distinct ideas.
- Perplexity's groupings were confusing and did not match the researcher's categories.

Key Finding

Neither tool provided usable Q10 groupings and counts.

Results by AI

Dovetail Results

- **Correct counts:** Q1–Q3, Q6–Q8, and 5 of 9 Q5 activities.
- **Q4 errors:** Claude, Gemini, and Perplexity counts.
- **Q5 errors:** Usability test tasks, quantitative analysis, recommendations, and stakeholder summaries or reports.
- **Q9–Q10:** Extensive explanation, but no usable counts of distinct ideas

Key Finding

Dovetail's detailed report still required the researcher to check its counts and rebuild the written-response analysis.

Perplexity Results

- **Correct counts:** Q1–Q3, Q6–Q8, and 5 of 9 Q5 activities.
- **Q4 errors:** ChatGPT, Claude, Gemini, and Microsoft Copilot.
- **Q5 errors:** Usability test tasks, quantitative analysis, recommendations, and stakeholder summaries or reports. (The same four activities Dovetail miscounted.)
- **Q9:** Counts came close but were off by one or two.
- **Q10:** Groupings were confusing and did not match the researcher's categories.

Key Finding

Perplexity came closest on Q9, but its analysis still needed human verification.

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?