Can AI Analyze UX Evaluation Results?

Presented by Theresa Wilkinson

Can AI Analyze UX Evaluation Results?

Agenda

1. Executive Overview
2. Areas of Agreement and Disagreement
3. Key Conclusions
4. Implications for UX Researchers
5. Limitations
6. Recommendations

Executive Overview

- All 4 AI models produced usable initial analyses covering core veterinary-app workflows.
- All 4 included digital-literacy variation, assistive-technology users, and accessibility considerations.
- The models agreed more on **what should be tested** than on **how the study should be conducted**.
- Sample-size recommendations ranged from **8–12 participants to approximately 20–48 participants**.
- Task coverage ranged from **4 to 8 tasks**, with accessibility and error-recovery testing handled inconsistently.
- No model produced a complete plan without gaps.
- AI can accelerate research planning, but human review is required to ensure completeness, practicality, accessibility coverage, and alignment between research questions and tasks.

Areas of Agreement and Disagreement

Agreement

- All 4 models recommended moderated usability testing.
- All 4 covered appointment scheduling, provider selection, multi-pet management, prescription refills, and appointment history.
- All 4 included participants with different levels of digital literacy.
- All 4 recommended including assistive-technology users.
- All 4 connected participant recruitment to study objectives.
- All 4 used realistic task scenarios focused on observable behavior.

Disagreement

- Sample-size recommendations ranged from **8–12 to approximately 20–48 participants**.
- **3 of 4 models** recommended one or two rounds with approximately **8–15 participants**; Copilot recommended **4–6 iterative rounds**.
- Research-question coverage ranged from **4 to 16 questions**.
- Task coverage ranged from **4 to 8 tasks**.
- Only **2 of 4 models** included tasks specifically designed to expose accessibility failures.
- Models differed substantially in methodological detail, metrics, assumptions, and supporting rationale.

Key Conclusions

- Consensus existed around the primary workflows and participant characteristics, but not around study structure.
- Sample-size recommendations represented the largest disagreement and would significantly affect budget and timeline.
- Accessibility was universally acknowledged but inconsistently operationalized.
- Only ChatGPT and Gemini included dedicated tasks designed to expose accessibility or error-recovery problems.
- Gemini provided the strongest explicit connection between research questions, tasks, and specific WCAG 2.2 criteria.
- Claude provided the most developed metrics, assumptions, risks, and participant criteria.
- ChatGPT provided the broadest task coverage and clearest task-level measures.
- Copilot provided the strongest iterative-testing rationale but recommended the largest and most expensive study.
- No single model was strongest across every evaluation category.

Implications for UX Researchers

- AI-generated research plans should be treated as starting drafts rather than final study plans.
- Comparing multiple outputs can reveal gaps that may not be apparent in a single model's response.
- Researchers must verify that every research question is addressed by at least one task.
- Recruiting assistive-technology users is not sufficient unless tasks are designed to expose accessibility barriers.
- Sample size should follow the selected research approach, available budget, timeline, and number of planned iterations.
- Named standards and methodologies should be independently verified before use.
- Human judgment remains necessary to balance methodological rigor with practical project constraints.

Limitations

- The evaluation compared proposed research plans, not completed usability studies.
- No participant performance, satisfaction, error, or outcome data were available.
- The analysis cannot determine which plan would produce better research findings in practice.
- The evaluation covered one veterinary-app scenario and may not generalize to other products or industries.
- The original prompt conditions, model versions, and generation settings could not be independently confirmed.

Recommendations

- Combine the strongest elements from all 4 models rather than adopting a single plan unchanged.
- Use approximately **12–15 participants across two rounds** as a practical baseline when resources permit.
- Include clear participant segments for digital literacy, pet ownership, platform experience, and assistive-technology use.
- Include at least one dedicated accessibility task and one error-recovery task.
- Map every research question directly to one or more usability tasks.
- Measure task success, time on task, errors, hesitations, confidence, and post-task ease.
- Consider SUS or UMUX-Lite as a standardized post-session measure.
- Integrate relevant WCAG 2.2 criteria into realistic tasks rather than relying only on a separate accessibility audit.
- Independently review all AI-generated recommendations before fielding the study.