JUX Journal of User Experience

Vol. 21, Issue 4, November 2026, pp. 1–4

The Illusion of Rigor: AI-Driven Product Teams Need Context Architecture to Protect Product Quality

Walky Goode
Amazon
525 Market Street
San Francisco, CA 94105
walkygo@amazon.com

Product development once progressed in distinct stages, with teams constantly pushing to shorten release cycles from months down to two-week sprints. The cycle was built on cross-functional collaboration, in which product teams worked with research teams to explore and test an idea, with design creating mockups and engineering building the product. Today, AI and code-generation tools compress product development much further by collapsing not just timelines but disciplines.

Nowadays, builder teams can use AI as a fast, super-intelligent co-worker. Vibe coding lets builders instruct AI agents to spin up fully interactive environments directly in front of testers. They can deploy agents to gather customer data at scale, transcribe feedback, and quickly crunch patterns. A team can access a synthesized report instantly. But when both the research brief and report summary are AI-driven, the reports only reflect what the raw data contains. Human guidance remains essential to interpret customer feedback through the lens of product context, organizational nuance, and relevant theoretical grounding.

Copyright © 2026-2027, User Experience Professionals Association and the authors. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. URL: http://www.upassoc.org.

In 2023, in a personal injury lawsuit, the plaintiff's attorney used ChatGPT to conduct research and draft a court submission (*Mata v. Avianca, Inc.*, 2023). In researching and drafting briefs, lawyers consulted legal encyclopedias and databases such as Westlaw and LexisNexis Academic, using tech tools to assist them. In this particular case, ChatGPT invented six non-existent court decisions complete with fabricated judicial quotes and internal citations. The opposing counsel couldn't locate several cases cited in the brief and alerted the court. When the plaintiff's attorney failed to produce the full text of the source material, the judge sanctioned them.

Verification is part of the legal process. The opposing counsel must read every word and check every citation. With an audit system in place, failures are loud and quickly identified. In product development, there is no opposing counsel. When an AI generates a satisfactory research summary for a product team, the idea passes straight to the roadmap. Without an audit system, failures are silent and won't surface until weeks later.

In a recent study, a product team sought input on a problem. Their team had launched a new personalized onboarding workflow that was receiving negative feedback. The new system was designed to help previous customers start using the new product quickly by skipping specific learning modules. The team's assumption was that returning users remembered how the product worked. The team used AI to pilot test the concept. The study compared a control group against a pilot group. The control group studied the learning modules regardless of past user history. The pilot group skipped a subset of modules if they had used the product within the previous 6 months.

Because the product had changed during that time, the pilot group's mental model was incorrect. The AI-generated experimental design was flawed from the beginning; the design treated both groups as comparable. Worse, because the pilot group was not exposed to the learning modules, they didn't know a new capability even existed. The pilot group didn't provide any feedback on it. Yet the AI interpreted the absence of negative feedback as a lack of friction. Its summary report was mainly positive. The tool synthesized feedback in the transcript, but the tool lacked the context to recognize that the participants did not comment on a feature because they didn't know it existed.

The team shipped the new onboarding workflow, which helped users start using the product quickly but created a new problem. Returning users were systematically routed around the new capability, making it difficult for them to discover or adopt it. Furthermore, those users who managed to discover the new capability were confused about how to use it because they had skipped the learning modules.

The product team had followed a standard product development framework that included discovery and validation. Their team trusted the AI's output because it appeared rigorous, but they didn't have the domain expertise to audit the conditions behind the data. Once my team identified the issues, they were able to fix the problem.

The previous example illustrates one of AI's limitations: An automated summary is only as reliable as the transcripts it processes, and transcripts without context are incomplete. Qualitative research doesn't start at the analysis and synthesis phase. Qualitative research starts at setup with the experimental design and thrives during data collection, during which a human interprets data that is not captured in text.

An experienced researcher reads between the lines. Researchers watch gestures and pick up tone. We know when to probe further to understand why hesitation or contradiction arises. We try to interpret what causes a pause: cognitive load, confusion, or a different mental model. My team jokes that we're the masters of awkward silence, letting silences extend for 6 to 8 seconds to allow a participant to fill it with their perspective. In addition to empathy, a human researcher also brings the team's internal context into the session. My team ran a proof-of-concept study to test a new creative workflow. There were two workflow designs being evaluated: light, in-context prompts versus a fully guided walkthrough. Our product team was split on the trade-off between user experience and engineering effort. The researcher understood the internal debate and knew to probe user hesitation. Human intuition helped the product team decide what to build.

AI moderation tools are valuable tools that help scale our work. A tool can force-multiply the reach of a single researcher by reaching thousands of users across the globe and multiple

languages. However, an AI moderator does not experience the context of internal team debates about a specific product strategy or understand historical business decisions. An AI tool can dynamically follow up on a participant's response, whereas a human moderator guides the participant toward unstated product questions.

Without a human in the loop, automated interviews collect data faster. However, the issues around using AI are compounded when teams stack automated tools throughout the research phase, using AI to draft briefs, collect data, synthesize transcripts, and provide recommendations. When AI lacks the full context, an incompletely scoped research brief leads to a flawed discussion guide, surface-level moderation, and misleading transcripts. Yet because AI excels at writing, it will produce a research summary that appears rigorous but is flawed.

To prevent context failures from undermining the value of AI, our team embeds context architecture directly into the product lifecycle. We create context architecture before a research project even starts. We apply embedded domain knowledge and strategic constraints at the front end of our process. We start with standardized research brief templates covering product history, business table stakes, design principles, and technical constraints. We use scoping documents to guide our immersion and align with stakeholders to understand internal team debates and key decision risks. We review internal research repositories to ensure we don't overlook past findings. When relevant, we review external research to understand our users' full ecosystem. We also partner closely with our quantitative peers to integrate behavioral data. We introduce rigor upstream to set a solid contextual foundation for our AI models.

On the back end, our team acts as an active auditor. We use AI tools to transcribe and assist in data analysis. We evaluate the AI's outputs against reality in the form of high-level notes from a live session or by spot-checking an output against a recorded session. We don't accept AI summaries at face value. Instead, we use established prompts to question the summaries or AI research agents configured to raise the standard for acceptance. We use AI agents to challenge automated conclusions or propose alternative explanations. This downstream audit helps confirm that our insights have been pressure-tested with human domain expertise and intentional stress-testing.

An end-to-end approach establishes the value of the research discipline in the AI era. Context architecture is not about slowing the work or doing manual data processing; rather, context architecture is about shaping the context AI relies on. A researcher who steers the upstream inputs and evaluates the downstream outputs shifts from executing tasks to orchestrating strategy. As a result, the shift ensures that rapid product development is grounded in methodological rigor and human understanding.

AI will only continue to accelerate product development, and the value of research will no longer be measured by how fast researchers run studies or generate reports. Our value will be measured by the quality of the contextual architecture we build. Teams that embed domain knowledge upstream and rigorously audit outputs downstream ensure they're not only moving fast, but they're moving in the right direction.

AI will always be able to generate credible answers, but only human researchers can ask the right questions, interpret nuanced human signals, and apply sound judgment. Our role is to set guardrails to ensure that—as product development accelerates—the product strategy remains human-centered.

References

Mata v. Avianca, Inc., 678 F.Supp.3d 443 (S.D.N.Y. 2023). https://www.law.berkeley.edu/wp-content/uploads/archive/2025/12/Mata-v-Avianca-Inc.pdf

About the Author

Walky Goode is Senior Manager of Research Strategy within PXT Central Science at Amazon. She leads a team conducting qualitative research to uncover critical opportunities across complex enterprise platforms and emerging technology workflows. With over a decade of UX research leadership experience at Google and Amazon, she specializes in building research systems that inform organizational decision-making. She holds an MSc in Industrial Engineering and a PhD in Cognitive Psychology.