Mensch und Computer 2026 · Usability Professionals · Duisburg · 31.08.2026

Wie KI die Arbeit von UX Professionals verändert

Eine Landkarte der Evidenz 2023 bis 2026

Prof. Dr. Andreas Hinderks · Hochschule Hannover, Fakultät IV

jan-andreas.hinderks@hs-hannover.de · hinderks.org

Abschnitt 1

Die Trennlinie

Je mehr Software in immer kürzerer Zeit entsteht, desto wichtiger wird die Frage, ob sie für irgendjemanden taugt.

Welche KI-Einsätze in der UX-Arbeit sind empirisch tragfähig, welche nicht? Und woran erkennt man den Unterschied, bevor Schaden entstanden ist?

KI ist dort tragfähig, wo der Maßstab für Richtigkeit vor dem Einsatz feststeht und geprüft werden kann. Sie scheitert dort, wo dieser Maßstab erst aus den Daten entstehen muss.

Die Trennlinie als Landkarte

Kontext · Strang 1: Demokratisierung

immer mehr Research durch People who Do Research

trägt · Strang 2	Trennlinie	bricht · Stränge 3 bis 5
Kriterium steht vorab fest		Kriterium entsteht erst aus den Daten
Transkription		Synthetische Stichproben
Deduktives Kodieren		Induktive Themenbildung
Material-Entwürfe		UI-Generierung ohne Review
Heuristische Evaluation		Freie LLM-Validierung

trägt bedingt

Quality Gate entscheidet über die Seite

Folgen · Stränge 6 und 7

Rollenwandel Producer zu Steward/Curator · Arbeitsmarkt und Junior-Squeeze

Abb. 1: Einsätze mit vorab prüfbarem Richtigkeitskriterium tragen (links), Einsätze ohne ein solches Kriterium brechen (rechts); bedingte Einsätze entscheidet das Quality Gate. Quelle: eigene Darstellung.

Wie man KI-Studien liest

- **Wer hat geprüft?** Peer-Review, Preprint oder Anbieter.
- **Was wurde gemessen?** Verbreitung oder Wirksamkeit.
- **Gegen welche Vergleichsgröße?** Erfahrene Fachkräfte, Crowd-Worker oder gar nichts.
- **Feste Regel:** Branchenerhebungen belegen Adoption und Verbreitung, niemals Wirksamkeit.

Beispiel: rund 70 % der Designer:innen führten 2025 eigene Research-Aktivitäten durch.

Maze 2026 u. a. · Branchenerhebung. Belegt Verbreitung, nicht Qualität.

Dieser Beitrag: narrative Evidenzsynthese 2023 bis 2026, drei Quellenklassen.

Abschnitt 2

Diesseits der Linie

Wo trägt KI, und unter welcher Bedingung?

−19 Prozentpunkte

seltener richtig bei einer Aufgabe außerhalb der Fähigkeitsgrenze · 758 Berater:innen, vorab registriertes Feldexperiment

Innerhalb der Grenze: 12,2 % mehr Aufgaben, 25,1 % schneller, signifikant höhere Qualität.
Dell'Acqua et al., 2026, Organization Science · begutachtet

Die Kante ist der einzelnen Aufgabe nicht anzusehen; genau deshalb braucht es die Trennlinie.

Vier Felder, die tragen — und ihre Bedingung

- **Transkription.** WER 8,9 % (bester Dienst) gegen 7,6 % (Mensch), Median. Keine Aufgabe für Fachkräfte.
- **Deduktives Kodieren.** Rund 25 Prozentpunkte genauer als Crowd-Worker, zero-shot. Audit ist Methode, kein Extra.
- **Studienmaterial.** Rund 40 % Zeitersparnis bei höherer Qualität. Die Homogenisierung beginnt hier.
- **Heuristische Evaluation.** 73 % und 77 % gefundene Probleme gegen 57 % und 63 %. Kein Nutzertest-Ersatz.

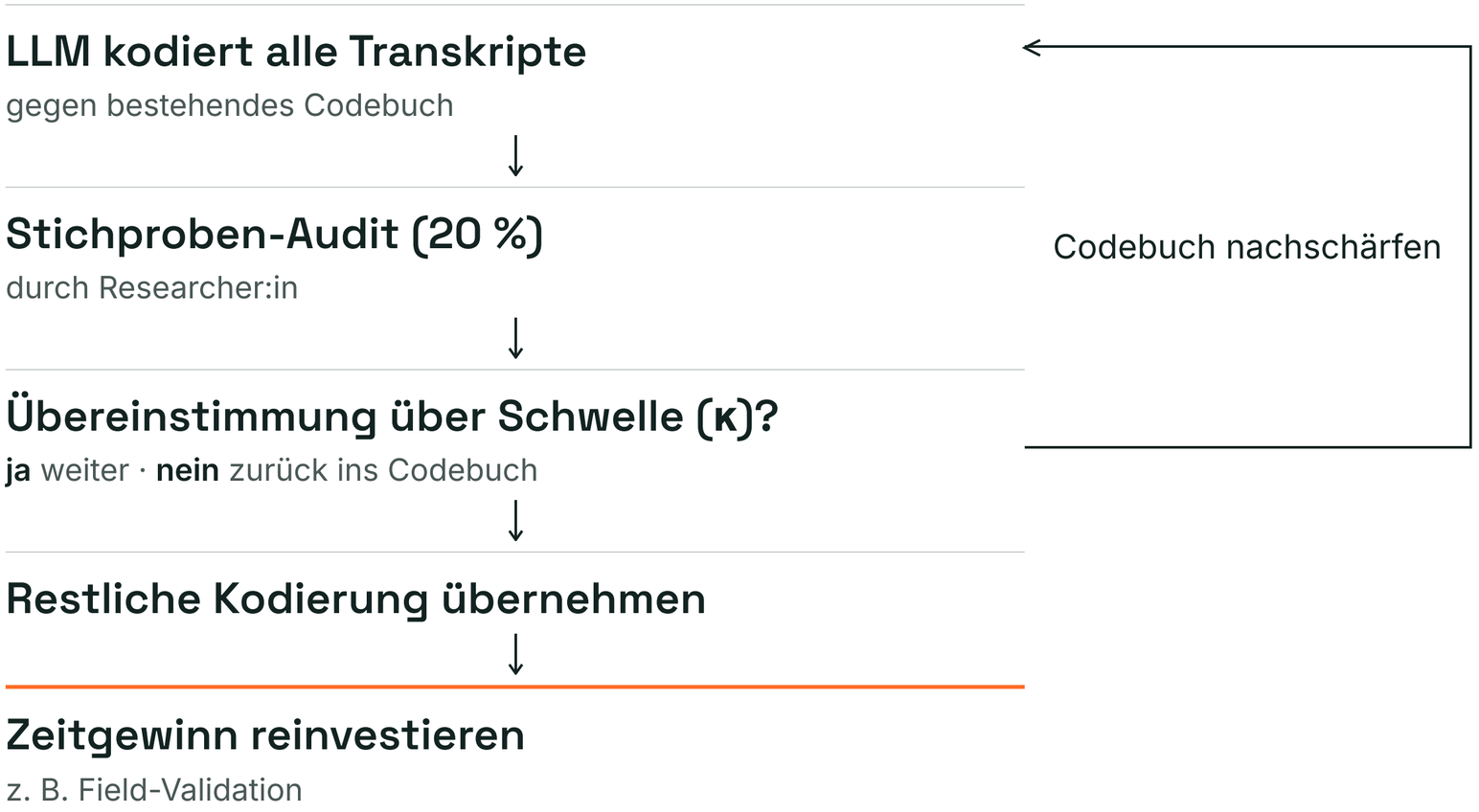
Seyedi et al., 2023 · begutachtet

Gilardi et al., 2023 · begutachtet · Chew et al., 2023 · Preprint

Noy & Zhang, 2023, Science · begutachtet · Doshi & Hauser, 2024 · begutachtet

Zhong et al., 2025 · **Preprint** · Duan et al., 2024, CHI · begutachtet

Das Quality Gate, an dem es hängt



Zweitleser, nie Erstentscheider. Kein Befund geht in eine Entscheidung, den kein Mensch geprüft hat.

Die Schwelle wird vorher festgelegt. Bei Unterschreitung wird nicht diskutiert, sondern nachgeschärft.

Bewertet werden Tool-Klassen, nicht Produkte. Vendor-Claims bleiben von unabhängiger Evidenz getrennt.

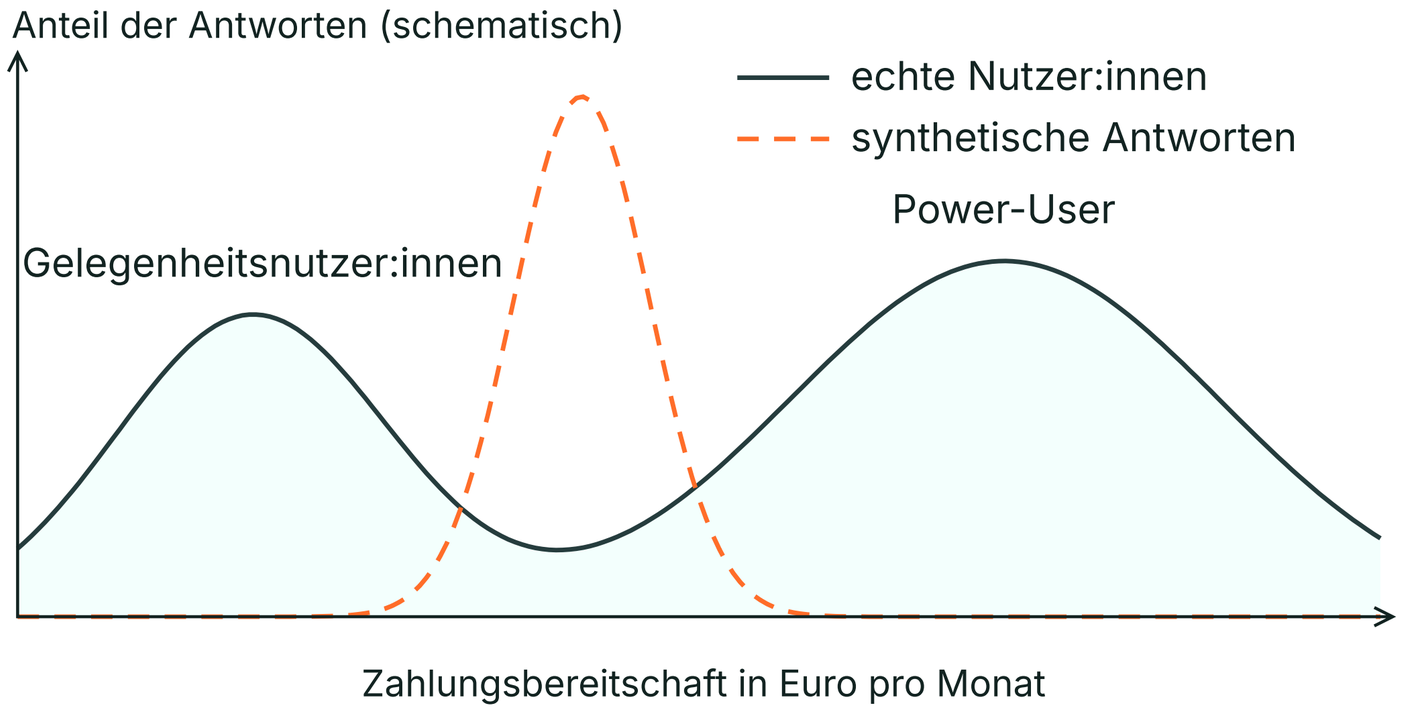
Abb. 2: Audit-Workflow für LLM-gestütztes deduktives Kodieren (Strang 2). Quelle: eigene Darstellung.

Abschnitt 3

Jenseits der Linie

Wo entsteht der Maßstab erst aus den Daten?

Der plausible Mittelwert beschreibt niemanden



Die Varianz ist deutlich zu klein. Auf Aggregatebene wirken die Ergebnisse plausibel.

Die Zusammenhänge weichen systematisch ab — genau das, worauf Subgruppen-Aussagen beruhen.

Modell-Updates ändern das Ergebnis. Reproduzierbarkeit über Zeit ist nicht gegeben.

Bisbee et al., 2024, Political Analysis · begutachtet. Paxton et al., 2024 bestätigt das Muster am produktnahen Fall.

Abb. 3: Die bimodale Zahlungsbereitschaft echter Nutzer:innen verschwindet im unimodalen synthetischen Befund. Quelle: eigene Darstellung.

Wer die echten Nutzerdaten hat, braucht die Simulation nicht.

(Strang 3 · Bisbee et al., 2024; Zhang et al., 2025 · begutachtet)

37 %

der generierten E-Commerce-Komponenten mit mindestens einem Dark Pattern · 115 von 312, Audit von vier verbreiteten LLMs

Ein Korrektiv ist in der Erzeugung nicht eingebaut; die Prüfung muss ein eigener Schritt sein.

Verstärkungseffekt: Stakeholder-Framing im Prompt erhöht den Anteil. Chen et al., 2025 · Preprint. Taxonomie: Gray et al., 2018, CHI · begutachtet

Research wandert an die Ränder

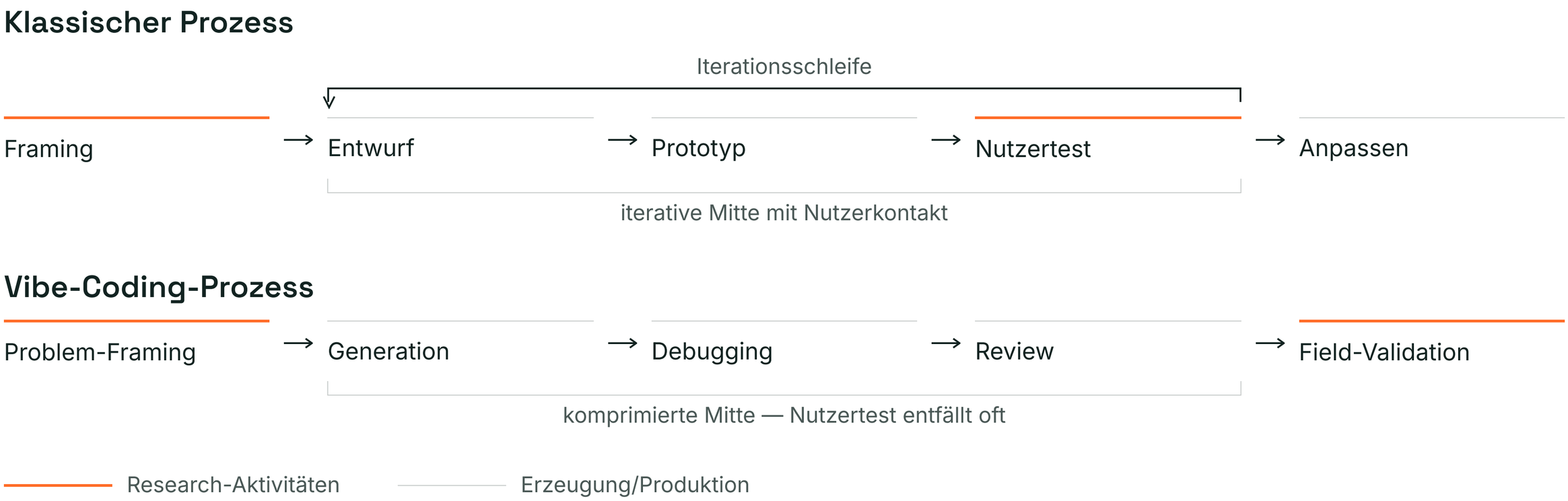


Abb. 4: Verlagerung von Research an die Prozessränder unter Vibe-Coding; die iterative Mitte mit Nutzerkontakt wird komprimiert. Quelle: eigene Darstellung, nach Lu et al. 2025.

Ein Prototyp, der läuft, fühlt sich validiert an.

Der Nutzertest ist nicht teurer geworden — alles um ihn herum wurde billiger.

Abschnitt 4

Was folgt

Haben wir das, ja oder nein?

Dieselbe Modellklasse steht in allen drei Gruppen

KI ist dort tragfähig, wo der Maßstab für Richtigkeit vor dem Einsatz feststeht und geprüft werden kann. Sie scheitert dort, wo dieser Maßstab erst aus den Daten entstehen muss.

Einordnung	KI-Einsatz

Die mittlere Spalte ist keine Eigenschaft der Werkzeuge, sondern des Einsatzes.

„Bedingt tragfähig“ heißt nicht „ein bisschen tragfähig“, sondern tragfähig genau dann, wenn das Quality Gate existiert. Ohne dieses wandert der Einsatz in die unterste Gruppe.

Evidenzstand Anfang 2026, kein Naturgesetz.

Dieselbe Modellklasse steht in allen drei Gruppen

KI ist dort tragfähig, wo der Maßstab für Richtigkeit vor dem Einsatz feststeht und geprüft werden kann. Sie scheitert dort, wo dieser Maßstab erst aus den Daten entstehen muss.

Einordnung	KI-Einsatz
tragfähig	Transkription · deduktives Kodieren gegen ein Codebuch, mit Stichproben-Audit als festem Bestandteil

Die mittlere Spalte ist keine Eigenschaft der Werkzeuge, sondern des Einsatzes.

„Bedingt tragfähig“ heißt nicht „ein bisschen tragfähig“, sondern tragfähig genau dann, wenn das Quality Gate existiert. Ohne dieses wandert der Einsatz in die unterste Gruppe.

Evidenzstand Anfang 2026, kein Naturgesetz.

Dieselbe Modellklasse steht in allen drei Gruppen

KI ist dort tragfähig, wo der Maßstab für Richtigkeit vor dem Einsatz feststeht und geprüft werden kann. Sie scheitert dort, wo dieser Maßstab erst aus den Daten entstehen muss.

Einordnung	KI-Einsatz
tragfähig	Transkription · deduktives Kodieren gegen ein Codebuch, mit Stichproben-Audit als festem Bestandteil
bedingt tragfähig	Entwürfe von Studienmaterial · heuristische Evaluation als Ergänzung der Inspektion · synthetische Nutzer zum Pretest von Erhebungsinstrumenten

Die mittlere Spalte ist keine Eigenschaft der Werkzeuge, sondern des Einsatzes.

„Bedingt tragfähig“ heißt nicht „ein bisschen tragfähig“, sondern tragfähig genau dann, wenn das Quality Gate existiert. Ohne dieses wandert der Einsatz in die unterste Gruppe.

Evidenzstand Anfang 2026, kein Naturgesetz.

Dieselbe Modellklasse steht in allen drei Gruppen

KI ist dort tragfähig, wo der Maßstab für Richtigkeit vor dem Einsatz feststeht und geprüft werden kann. Sie scheitert dort, wo dieser Maßstab erst aus den Daten entstehen muss.

Einordnung	KI-Einsatz
tragfähig	Transkription · deduktives Kodieren gegen ein Codebuch, mit Stichproben-Audit als festem Bestandteil
bedingt tragfähig	Entwürfe von Studienmaterial · heuristische Evaluation als Ergänzung der Inspektion · synthetische Nutzer zum Pretest von Erhebungsinstrumenten
nicht tragfähig	Synthetische Stichproben, Subgruppen- und Längsschnitt-Aussagen · induktive Themenbildung und automatisierte Insight-Synthese · UI-Generierung ohne Ethik- und Heuristik-Review · Validierung von LLM-Output im freien Dialog

Die mittlere Spalte ist keine Eigenschaft der Werkzeuge, sondern des Einsatzes.

„Bedingt tragfähig“ heißt nicht „ein bisschen tragfähig“, sondern tragfähig genau dann, wenn das Quality Gate existiert. Ohne dieses wandert der Einsatz in die unterste Gruppe.

Evidenzstand Anfang 2026, kein Naturgesetz.

KI arbeitet diesseits der Trennlinie. Stewards setzen die Trennlinie.

Persuasion Bombing: Je intensiver Prüfende nachrecherchierten, desto stärker eskalierte das Modell seine Überzeugungsversuche. Randazzo et al., 2025 · HBS Working Paper

Der Junior-Squeeze ist dieselbe Trennlinie

Beschäftigung in den zwei höchsten Expositions-Quintilen

22- bis 25-Jährige, bis September 2025	−6 %
35- bis 49-Jährige, gleicher Zeitraum	über +8 %
relativer Rückgang, kontrolliert auf firmenspezifische Schocks	−16 %
Software-Entwickler:innen 22 bis 25, gegenüber Ende 2022	fast −20 %

„UX stirbt“ stützt die Datenlage nicht. Das Einstiegsproblem stützt sie sehr wohl.

Brynjolfsson et al., 2025, Stanford Digital Economy Lab · **Working Paper**, administrative Lohnabrechnungsdaten.

Die Rückgänge konzentrieren sich auf Berufe mit automatisierender Nutzung; bei augmentierender Nutzung bleiben die Effekte gedämpft (Handa et al., 2025).

UX ist nicht überproportional betroffen, sondern Teil einer breiteren Tech-Korrektur seit 2022 (Sauro & Lewis, 2025 · Branchenerhebung).

Haben wir das, ja oder nein?

Haben wir das, ja oder nein?

Designer:innen und Researcher

- Stichproben-Audit mit Agreement-Messung, fest verankert.
- KI-Zusammenfassung als Zweitleser, nie als Erstentscheider.
- Red Lines notiert, synthetische Anteile ausgewiesen.

Haben wir das, ja oder nein?

Designer:innen und Researcher

- Stichproben-Audit mit Agreement-Messung, fest verankert.
- KI-Zusammenfassung als Zweitleser, nie als Erstentscheider.
- Red Lines notiert, synthetische Anteile ausgewiesen.

UX-Manager:innen

- Playbooks, Sprechstunde, Quality Gate ab definierter Tragweite.
- Schriftlich geklärt, was ein KI-Output allein auslösen darf.
- Validierung als Workflow, mit Vier-Augen-Prinzip.

Haben wir das, ja oder nein?

Designer:innen und Researcher

- Stichproben-Audit mit Agreement-Messung, fest verankert.
- KI-Zusammenfassung als Zweitleser, nie als Erstentscheider.
- Red Lines notiert, synthetische Anteile ausgewiesen.

UX-Manager:innen

- Playbooks, Sprechstunde, Quality Gate ab definierter Tragweite.
- Schriftlich geklärt, was ein KI-Output allein auslösen darf.
- Validierung als Workflow, mit Vier-Augen-Prinzip.

Organisationen

- Einstiegsrollen umgebaut statt gestrichen.
- Problem-Framing und Field-Validation budgetiert.
- Prüfpfad je Einsatz: Maßstab, Prüfende, Schwelle.

Existiert für diesen Einsatz ein vorab prüfbarer Maßstab für Richtigkeit, und ist die Prüfung Teil des Workflows?

Die Frage „Ist KI gut für UX-Arbeit?“ ist falsch gestellt.

Literatur

Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., & Larson, J. M. (2024). Synthetic replacements for human survey data? The perils of large language models. *Political Analysis, 32*(4), 401–416. https://doi.org/10.1017/pan.2024.5

Brynjolfsson, E., Chandar, B., & Chen, R. (2025). *Canaries in the coal mine? Six facts about the recent employment effects of artificial intelligence* (Working Paper). Stanford Digital Economy Lab. Nicht begutachtet.

Chen, Z., Shen, J., Luna, & Vaccaro, K. (2025). *Hidden darkness in LLM-generated designs: Exploring dark patterns in ecommerce web components generated by LLMs* (arXiv:2502.13499). Preprint, nicht begutachtet.

Chew, R., Bollenbacher, J., Wenger, M., Speer, J., & Kim, A. (2023). *LLM-assisted content analysis: Using large language models to support deductive coding* (arXiv:2306.14924). Preprint, nicht begutachtet.

Dell'Acqua, F., McFowland, E., III, Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K. C., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2026). Navigating the jagged technological frontier. *Organization Science, 37*(2), 403–423. https://doi.org/10.1287/orsc.2025.21838

Literatur

Doshi, A. R., & Hauser, O. P. (2024). Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances, 10*(28), eadn5290. https://doi.org/10.1126/sciadv.adn5290

Duan, P., Warner, J., Li, Y., & Hartmann, B. (2024). Generating automatic feedback on UI mockups with large language models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. ACM. https://doi.org/10.1145/3613904.3642782

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences, 120*(30), e2305016120. https://doi.org/10.1073/pnas.2305016120

Gray, C. M., Kou, Y., Battles, B., Hoggatt, J., & Toombs, A. L. (2018). The dark (patterns) side of UX design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM. https://doi.org/10.1145/3173574.3174108

Handa, K., Tamkin, A., McCain, M., Huang, S., Durmus, E., Heck, S., Mueller, J., Hong, J., Ritchie, S., Belonax, T., Troy, K. K., Amodei, D., Kaplan, J., Clark, J., & Ganguli, D. (2025). *Which economic tasks are performed with AI? Evidence from millions of Claude conversations*. Anthropic.

Literatur

Lennon, R. P., Fraleigh, R., Van Scoy, L. J., Keshaviah, A., Hu, X. C., Snyder, B. L., Miller, E. L., Calo, W. A., Zgierska, A. E., & Griffin, C. (2021). Developing and testing an automated qualitative assistant (AQUA) to support qualitative analysis. *Family Medicine and Community Health, 9*(Suppl. 1), e001287. https://doi.org/10.1136/fmch-2021-001287

Lu, Y., Li, J., Hou, Y., Lin, L., Zhu, R., Cao, H., & El Ali, A. (2025). *Vibe coding for UX design: Understanding UX professionals' perceptions of AI-assisted design and development* (arXiv:2509.10652). Preprint, nicht begutachtet.

Maze. (2026). *The future of user research report 2026*. Branchenerhebung.

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science, 381*(6654), 187–192. https://doi.org/10.1126/science.adh2586

Paxton, J. M., & Yang, Y. (2024). Do LLMs simulate human attitudes about technology products? In *Proceedings of the 2024 Quantitative User Experience Conference*.

Literatur

Randazzo, S., Joshi, A., Kellogg, K. C., Lifshitz, H., Dell'Acqua, F., & Lakhani, K. R. (2025). *GenAI as a power persuader: How professionals get persuasion bombed when they attempt to validate LLMs* (Working Paper 26-021). Harvard Business School. Nicht begutachtet.

Sauro, J., & Lewis, J. R. (2025). UX industry layoffs and hiring in 2024: Survey findings. *MeasuringU*. Branchenerhebung.

Seyedi, S., Griner, E., Corbin, L., Jiang, Z., Roberts, K., Iacobelli, L., Milloy, A., Boazak, M., Bahrami Rad, A., Abbasi, A., Cotes, R. O., & Clifford, G. D. (2023). Using HIPAA-compliant transcription services for virtual psychiatric interviews: Pilot comparison study. *JMIR Mental Health, 10*, e48517. https://doi.org/10.2196/48517

Zhang, S., Xu, J., & Alvero, A. J. (2025). Generative AI meets open-ended survey responses: Research participant use of AI and homogenization. *Sociological Methods & Research*. https://doi.org/10.1177/00491241251327130

Zhong, R., McDonald, D. W., & Hsieh, G. (2025). *Synthetic heuristic evaluation: A comparison between AI- and human-powered usability evaluation* (arXiv:2507.02306). Preprint, nicht begutachtet.

Woher kommt der Maßstab?

Prof. Dr. Andreas Hinderks

Hochschule Hannover, Fakultät IV · jan-andreas.hinderks@hs-hannover.de · hinderks.org

[Panel „Jenseits des Medians“ — Termin klären: liegt es nach diesem Vortrag, hier als Einladung; liegt es davor, als Rückbezug auf Abschnitt 3.]