Evaluating Generative AI in the Lab: Methodological Challenges and Guidelines

Hyerim Park*
BMW Group
Munich, Germany
University of Stuttgart
Stuttgart, Germany
hyerim.park@bmw.de

Khanh Huynh*
BMW Group
Munich, Germany
LMU Munich
Munich, Germany
khanh.huynh@bmw.de

Malin Eiband
BMW Group
Munich, Germany
malin.eiband@bmw.de

Jeremy Dillmann
BMW Group
Munich, Germany
jeremy.dillmann@bmw.de

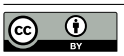
Sven Mayer
TU Dortmund University
Dortmund, Germany
Research Center Trustworthy Data Science and Security
Dortmund, Germany
info@sven-mayer.com

Michael Sedlmair
University of Stuttgart
Stuttgart, Germany
michael.sedlmair@visus.uni-stuttgart.de

Abstract

Generative AI (GenAI) systems are inherently non-deterministic, producing varied outputs even for identical inputs. While this variability is central to their appeal, it challenges established HCI evaluation practices that typically assume consistent and predictable system behavior. Designing controlled lab studies under such conditions therefore remains a key methodological challenge. We present a reflective multi-case analysis of four lab-based user studies with GenAI-integrated prototypes, spanning conversational in-car assistant systems and image generation tools for design workflows. Through cross-case reflection and thematic analysis across all study phases, we identify five methodological challenges and propose eighteen practice-oriented recommendations, organized into five guidelines. These challenges represent methodological constructs that are either amplified, redefined, or newly introduced by GenAI's stochastic nature: (C1) reliance on familiar interaction patterns, (C2) fidelity–control trade-offs, (C3) feedback and trust, (C4) gaps in usability evaluation, and (C5) interpretive ambiguity between interface and system issues. Our guidelines address these challenges through strategies such as reframing onboarding to help participants manage unpredictability, extending evaluation with constructs such as trust and intent alignment, and logging system events, including hallucinations and latency, to support transparent analysis. This work contributes (1) a methodological reflection on how GenAI's stochastic nature unsettles lab-based HCI evaluation and (2) eighteen recommendations that help researchers design more transparent, robust, and comparable studies of GenAI systems in controlled settings.

*The first two authors contributed equally to this research.



IUI '26, Paphos, Cyprus

ACM ISBN 979-8-4007-1984-4/26/03
https://doi.org/10.1145/3742413.3789065

CCS Concepts

• **Human-centered computing** → **User studies**; **HCI design and evaluation methods**; **Empirical studies in HCI**.

Keywords

Generative AI, large language models (LLMs), user studies, methodology, human–AI interaction

ACM Reference Format:
Hyerim Park, Khanh Huynh, Malin Eiband, Jeremy Dillmann, Sven Mayer, and Michael Sedlmair. 2026. Evaluating Generative AI in the Lab: Methodological Challenges and Guidelines. In *31st International Conference on Intelligent User Interfaces (IUI '26), March 23–26, 2026, Paphos, Cyprus.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3742413.3789065

1 Introduction

Generative AI (GenAI) technologies are increasingly integrated into interactive systems across domains—from productivity tools to creative applications—by generating diverse forms of content such as text, images, and voice [1, 32, 55]. As these systems become more common in everyday contexts, evaluating their usability and user experience has become an important topic in HCI. However, GenAI also introduces challenges for established evaluation practices, particularly in controlled lab studies, which are among the core methods in HCI research [28, 49, 68].

Unlike rule-based systems, such as traditional chatbots or in-car voice assistants, which produce fixed responses based on predefined commands, decision trees, or state machines [56, 59, 64], GenAI models generate open-ended and context-dependent outputs that can differ with each interaction [20, 36, 45, 46, 57, 75]. This non-determinism, while enabling new forms of generative interaction, also disrupts key methodological assumptions that underlie lab-based evaluation, such as control, consistency, and comparability [31, 44]. Without adapted approaches, researchers risk drawing misleading conclusions about usability, trust, or user behavior when stochastic system behavior is mistaken for interface design flaws. We do not claim that these challenges are unique to GenAI. Similar

challenges can arise when evaluating other adaptive or intelligent systems [65, 67]. However, in generative systems, where output variability tends to be higher, these challenges are amplified and reframed, and additional challenges can emerge. Designing studies that remain transparent, robust, and comparable under such stochastic conditions is thus an important step toward more reliable GenAI evaluation.

This unpredictability affects all stages of user studies, from task definition and prototype development to data collection and interpretation, yet how to plan and conduct such studies remains underexplored. Numerous GenAI studies focus on system performance or user-facing outcomes [30, 40, 55, 70, 74], while the methodological decisions and trade-offs behind evaluation design are rarely discussed. Some acknowledge issues such as hallucinations, shifting output styles, or novelty effects that influence trust and user experience [23, 73], yet these aspects are often mentioned only as study limitations rather than systematically examined through internal materials and reflections. Making such methodological reasoning explicit can help researchers anticipate challenges and design more rigorous evaluations of GenAI systems in controlled settings.

To address these issues, we conducted a **reflective multi-case study** of four controlled lab-based user studies involving GenAI-integrated prototypes. The cases span two domains—LLM-based conversational in-car assistants and GenAI image tools for professional design workflows—and vary in fidelity, modality, and participant groups. We adopted a multi-case approach because methodological challenges manifest differently across systems and study designs, and a single case would not capture this diversity or reveal recurring patterns [18]. Our goal was not to report user-facing outcomes such as task performance or satisfaction measures, but to analyze the methodological decisions, adaptations, and tensions that emerged throughout the research process. Drawing on study materials, researcher notes, and team discussions, we reflected on how GenAI's stochastic nature shaped study planning, prototyping, data collection, and analysis. Our work addresses two research questions:

RQ1 What recurring methodological challenges arise when evaluating GenAI systems in controlled lab settings?

RQ2 How can these challenges be addressed in the design and execution of such studies?

Through cross-case reflection, affinity diagramming, and inductive thematic analysis [6], we identified **five methodological challenges (C1–C5)** that complicate established HCI evaluation practices. These challenges represent methodological constructs that are either **amplified, redefined, or newly introduced** by the generative and non-deterministic nature of GenAI systems: (C1) amplified reliance on familiar interaction patterns, (C2) amplified trade-offs between fidelity and experimental control, (C3) redefined feedback loops and user trust, (C4) new methodological gaps in usability evaluation, and (C5) amplified interpretive ambiguity between interface and system behavior. Building on these findings, we propose **five methodological guidelines (G1–G5)**, each linked to one of the challenges, and eighteen practice-oriented recommendations that offer actionable strategies for designing, conducting, and analyzing GenAI user studies. The guidelines include preparing participants for unpredictable system behavior, aligning prototype fidelity to study goals, improving feedback interpretability and user trust, adapting evaluation strategies to capture GenAI-specific experiences, and building flexibility into study design and analysis.

This paper contributes (1) a methodological reflection based on four GenAI-integrated lab studies that reveal how stochastic model behavior challenges established evaluation practices, and (2) eighteen concrete recommendations, structured under five guidelines, to support the planning and execution of GenAI user studies in controlled research settings.

2 Related Work

User studies are a foundational method in HCI for evaluating interactive systems [34, 36]. By typically combining quantitative measures (e.g., surveys, task logs) with qualitative feedback data, user studies help assess usability, user experience, and task performance across diverse interface types [21, 45, 53]. Traditional evaluations often rely on controlled lab experiments, measuring completion time, error rates, or subjective satisfaction, and are well-suited to systems with well-defined tasks and stable behavior [21, 34, 45].

2.1 Methodological Shifts in HCI Evaluation

As interactive systems become more adaptive, open-ended, and embedded in dynamic contexts, conventional evaluation methods such as usability testing and short-term lab studies often prove insufficient. Poppe et al. [43] emphasize that systems involving novel sensing technologies and shifting user/system initiative require longitudinal observations and context-sensitive evaluation. Similarly, Greenberg et al. [21] argue that standardized usability testing may hinder innovation or overlook the critical experiential dimensions, especially in systems supporting exploration, creativity, or collaboration. Brdnik et al. [9], in a review of IUI papers from 2012 to 2022, found that many evaluations still rely on conventional experiments and questionnaires, with limited attention to metrics suited for system adaptivity or the dynamics of human-AI co-adaption. Similar limitations have been observed in newer interaction paradigms such as voice interfaces, AR/VR, and IoT [12].

In response, the HCI community has adopted a range of complementary methods, including in-the-wild studies [51], longitudinal deployments [27], Wizard-of-Oz studies [29], simulation and modeling approaches [39], and more interpretive, mixed-method designs that combine usage data with reflective user feedback [14]. These approaches are often applied to systems that are complex, adaptive, or open-ended, or that are embedded in real-world settings where traditional lab evaluations may be insufficient.

The information visualization community provides a documented and explicit example of methodological evolution in response to similar challenges. Visualization tools are often used for exploratory tasks like analyzing large datasets or generating insights where no single correct answer or predefined success criterion exists [33, 42]. In such contexts, traditional metrics like task completion time and error rates may misrepresent how users interact or derive value from the system over time. To address these challenges, the BELIV (Beyond Time and Errors) workshop series [3][1] was established to promote evaluation approaches designed for the unique

[1] https://beliv-workshop.github.io/

challenges faced by visualization systems, emphasizing user engagement, exploration, and sense making beyond traditional performance metrics. Building on this foundation, researchers have proposed structured evaluation frameworks. Lam et al. [33] introduced a taxonomy of seven evaluation scenarios to guide method selection based on system type and research goal. Munzner's nested model [38] defines multiple design levels, such as data abstraction, visual encoding, interaction techniques, and domain tasks, clarifying what is being evaluated and how. Sedlmair et al. [54] reflect on the practical challenges of real-world design studies, emphasizing the "messiness" and need for context-specific adaptation when evaluating visualization systems outside controlled lab settings. Together, these methodological considerations highlight challenges that arise when evaluating complex, adaptive, or user-driven systems in HCI [17, 54]. GenAI shares similar characteristics: its outputs are inherently variable and context-dependent, making them difficult to evaluate using fixed metrics or predefined task goals.

Such methodological shifts have also emerged during periods of disruption or modality-specific innovation. For example, Schmidt et al. [53] explored remote and out-of-the-lab evaluation strategies in response to pandemic-related constraints, proposing alternatives to in-person testing such as browser-based prototypes and the reuse of existing datasets. Similarly, new interaction modalities have prompted adaptations in evaluation practice. Voice user interfaces (VUIs), for instance, prompted new methods focused on conversational timing, speech clarity, and context-aware interaction [30]. Tools such as HEUROBOX were developed to identify voice-specific usability issues [19], and standard instruments like the System Usability Scale (SUS) were adapted to assess qualities such as naturalness, politeness, and conversational flow [24]. These efforts reflect a broader push toward inclusive and context-sensitive evaluation strategies for emerging interaction paradigms [61].

Despite this progress, limited attention has been paid to the methodological implications of evaluating non-deterministic GenAI systems in controlled lab settings. Our work addresses this gap by reflecting on researcher decisions, trade-offs, and adaptations in the context of user studies involving GenAI.

2.2 Evaluating GenAI Systems in HCI

GenAI systems, including LLMs and image generators, introduce distinct methodological challenges for HCI evaluation. Unlike deterministic systems, GenAI tools are inherently variable, producing different outputs for the same input prompt, even under similar conditions [10, 55, 58]. This variability complicates evaluation in multiple ways. Traditional metrics, such as task completion time, accuracy, and error rates, rely on consistent system behavior and clear success criteria; yet, many GenAI use cases—including creative generation, idea exploration, and open-ended problem solving—lack a single correct outcome. Moreover, responses may vary in quality, structure, length, and completeness, making comparisons across participants difficult. GenAI systems are also known to produce hallucinations, that is, outputs that appear plausible but are factually incorrect. This can undermine output-based measures of reliability or satisfaction, particularly when users initially trust the response.

Prior work in HCI and explainable AI (XAI) has explored user study methodologies for evaluating AI systems. For instance, Rong et al. [52] review how XAI research primarily focuses on interpreting decision boundaries and building user trust in deterministic systems, often centered on decision-support tools or classifiers with predictable behaviors. In contrast, GenAI systems produce open-ended, multi-modal, and variable outputs, introducing unpredictability and interpretive ambiguity that are not central in XAI. Interactions with GenAI systems are typically iterative and involve a co-construction of meaning between user and system, with user expectations and satisfaction shaped by subjective, context-dependent factors. Together, these differences present challenges that call for rethinking user study design in HCI.

Existing user studies investigating GenAI systems also highlight several distinct challenges. While GenAI systems excel at generating general content, they often struggle with domain-specific understanding and fine-grained detail. As a result, outputs tend to be more generalized, lacking the depth and expertise required in specific fields [23, 35, 63, 70]. Prior work further shows that participants' mental models of and skepticism towards GenAI frequently shape their responses in user studies. Participants often draw on their prior experiences with GenAI—both positive and negative—when forming opinions during studies. A recurring concern is GenAI's tendency to produce seemingly plausible but factually incorrect content [23, 70]. In addition, reliance on AI can introduce cognitive biases, such as confirmation bias or the uncritical acceptance of agreeable responses [23, 72, 73]. Interviews have also reported novelty effects associated with so-called "advanced" AI tools, introducing another source of bias. Despite recognizing these issues, many studies acknowledge them only as limitations and fall short of proposing comprehensive methodological responses. To address these challenges, researchers have employed alternative approaches in GenAI studies. One common strategy is to use pre-generated outputs to control variability and ensure consistency. Other studies adopt Wizard of Oz methodologies, in which human facilitators simulate GenAI capabilities in real-time. However, this approach introduces its own limitations, such as delays in response and gaps in domain expertise [47]. Overall, most prior work emphasizes real-world usage scenarios and treats unpredictability and non-determinism primarily as limitations, rather than examining their implications for lab-based evaluation of generative systems.

Beyond GenAI-specific work, our contribution builds on a broader body of HCI research examining how evaluation methods need to be adapted for systems with uncertain, opaque, or partially simulated behavior. Wizard-of-Oz studies have long highlighted trade-offs between experimental control, realism, and interpretability, as well as challenges of attribution when system behavior is mediated by human or hybrid components [15, 48, 60]. More recent work has extended these concerns to machine learning (ML)–enabled systems, demonstrating that realistically simulating ML errors in Wizard-of-Oz studies is itself methodologically challenging and has consequential effects on user experience evaluation [25]. Additionally, critiques of usability evaluation methods argue against viewing them as "fixed recipes," urging the adaptation of methodological resources to align with specific system properties and research goals [69]. Building on these perspectives, we present guidelines as modular methodological resources for designing and interpreting lab-based studies of GenAI systems.

3 Methodology: A Multi-Case Study with Reflection and Thematic Analysis

We adopted a **multi-case study approach** supported by reflective and inductive thematic analysis to examine methodological challenges in evaluating GenAI systems through lab-based user studies. This approach enabled us to capture methodological challenges across varied GenAI systems and study designs while reflecting on our research decisions. A multi-case perspective supported the identification of both recurring patterns and context-specific nuances across studies [22]. By combining Eisenhardt's **systematic multi-case framework** [18] with **affinity diagramming** [4] for initial structuring and **inductive thematic analysis** [7, 8], we moved from concrete study observations to broader methodological insights. The resulting five challenges (C1–C5) represent **recurring methodological tensions** in lab-based GenAI evaluation and form the foundation for the guidelines presented in Section 5. In summary, this analytic process combined affinity diagramming for initial structuring with thematic analysis to synthesize cross-case methodological challenges.

3.1 Case Selection and Context

We analyzed four user studies conducted between *January 2024* and *September 2025*, covering both conversational and visual GenAI systems. We selected these cases to support cross-case comparison, following established guidance for multi-case research [18], rather than to maximize the number of challenges identified. Accordingly, our inclusion criteria were guided by methodological considerations and practical access constraints. We included only lab-based user studies that we conducted ourselves, involved direct user interaction with a GenAI system, and empirically evaluated that interaction with participants. Because all authors were involved in the studies, we had detailed insight into design rationale, trade-offs, and study-planning decisions, supporting retrospective methodological analysis. The cases varied in system type, participant group, and prototype fidelity—from in-car voice assistants to design-oriented image generation tools—providing sufficient diversity to examine recurring methodological challenges across contexts. This scale aligns with established guidance for multi-case research, which suggests that a small number of heterogeneous cases can provide depth and analytical comparability [18, 22]. As such, the resulting challenges should be understood as illustrative rather than exhaustive, highlighting recurring methodological tensions rather than providing a complete landscape of GenAI evaluation issues. We approached reflection as a method for methodological inquiry rather than subjective introspection, triangulating researcher memos, study logs, and artifacts to ensure transparency. Each case featured different combinations of GenAI models, interaction modalities, user goals, and study designs. This diversity offered a comparative basis for identifying recurring methodological tensions across distinct study contexts. An overview of the cases is provided in Table 1, with detailed descriptions in Section A.2.

3.2 Data Collection

We collected a range of internal research materials, including study plans, interview guides, prototype specifications, and observational notes from user sessions. Researcher reflections and memos written during and after each study, together with meeting summaries and study logs, were used to capture the reasoning behind methodological decisions. These materials enabled us to reconstruct both the practical procedures and the rationale behind specific design choices. Each researcher independently identified methodological issues from the studies they primarily planned and conducted. All observations and notes were then consolidated into a shared Figma workspace, where we collaboratively externalized, organized, and discussed emerging methodological patterns. To support collaborative organization of the collected materials, we employed **affinity mapping techniques** [4] to iteratively group, split, and reorganize methodological observations through team discussion.

3.3 Analysis

Our analysis followed an inductive thematic analysis applied across the four lab-based studies. Rather than starting from predefined categories, we iteratively developed themes grounded in the collected data. The analysis builds on materials generated through an initial affinity diagramming step, which supported the organization and externalization of methodological observations before formal thematic analysis. This process unfolded in three stages: (1) organizing methodological observations within each case, (2) collaboratively clustering and comparing patterns across cases, and (3) synthesizing broader methodological challenges emerging from the analysis [18].

Step 1: Organizing Case Observations. In the first stage, we organized all methodological observations and reflections according to study phases (e.g., research planning, prototyping, participant interaction, data collection, and analysis), supported by affinity diagramming. This phase-based coding allowed us to identify when and where methodological challenges occurred within the study process and supported early visualization of emerging patterns across projects. Section A.1 shows this early phase-based clustering.

Step 2: Collaborative Coding and Clustering. Next, building on this initial organization, we conducted **collaborative coding and clustering** to identify recurring methodological patterns. Using the materials generated through affinity diagramming, the authors iteratively compared and merged related analytic codes while discussing conceptual relationships across studies. This process moved from detailed, case-specific observations (codes) to conceptual clusters (subthemes) that captured methodological issues recurring across multiple contexts.

Step 3: Thematic Synthesis. Finally, through **thematic synthesis**, we abstracted these conceptual clusters into five **higher-level methodological challenges (C1–C5)**. This synthesis combined descriptive coding with reflective interpretation, focusing on methodological tensions we encountered repeatedly, such as participant familiarity, prototype fidelity, interpretability of feedback, metric validity, and confounding system factors. Figure 1 shows this progression from initial analytic codes (specific observations) to focused subthemes (conceptually related methodological issues) and five higher-order themes (challenges).

Table 1: Summary of the four lab-based studies highlighting study context, user groups, primary interaction modalities, adopted GenAI models, prototype fidelity, evaluation methods, and study timeframe.

Study Aspect	Case A	Case B	Case C	Case D
Study Overview	A multimodal conversational in-car assistant powered by an LLM, exploring interactions across different driving-related use cases.	A multimodal conversational LLM-based in-car assistant with integrated GUI interaction, focusing on users' references to visual elements.	A paper-based prototype of an early-stage GenAI image generation tool, targeting professional designers' input strategies.	A fully functional GenAI image generation tool deployed in professional design practice to support interactive workflows.
User Groups	Drivers	Drivers and passengers	Professional designers	Professional designers and design students
Interaction Modalities	Voice and GUI interaction (touchscreen)	Voice and GUI interaction (touchscreen)	Text input, scribbling, and handwritten annotations	Text input and visual inputs (scribbling and annotations), or combinations of both
AI Models Used	GPT-4o	GPT-4o	DALL·E 2	DALL·E 3, GPT-image-1, Flux.1 Kontext Pro, GPT-4o
Prototype Type and Fidelity	Fully functional	Fully functional	Paper-based prototype	Fully functional
Evaluation Methods	Semi-structured interviews and usability surveys (Likert scale)	Semi-structured interviews and usability surveys (Likert scale)	Interviews and think-aloud methods	Comparative study, interviews, and custom surveys (Likert scale)
Study Timeframe	January 2024–April 2024	March 2024–August 2024	August 2024–January 2025	March 2025–September 2025

4 Results: Challenges Identified Across Case Studies

Our inductive thematic analysis of the four lab-based GenAI studies revealed **five methodological challenges (C1–C5)** that complicate conventional HCI evaluation practices. These challenges reflect methodological constructs that are either **amplified, redefined, or newly introduced** by the generative and non-deterministic nature of GenAI systems. They emerged through iterative comparison of authors' reflections, study artifacts, and participant observations, capturing tensions spanning study planning, prototyping, user interaction, evaluation, and interpretation. Together, they illustrate how GenAI systems reshape established assumptions about what can be controlled, measured, and meaningfully interpreted in lab research. Figure 2 provides an overview of the core challenges and the corresponding methodological recommendations. In the following, we use *output variability* to describe the degree to which a system may produce different outputs for the same input under similar conditions—it is stochastic, but not arbitrary or completely unpredictable.

4.1 C1. GenAI amplifies user reliance on familiar interaction patterns

Users in HCI studies often default to familiar interaction strategies when introduced to novel systems. In GenAI contexts, this tendency becomes more pronounced, not merely due to habit or bias, but because the system's stochastic feedback prevents stable learning. Without predictable input–output mappings—as found in conventional interfaces that follow rule-based or deterministic logic—users struggle to infer how the system interprets their actions and instead rely on previously learned strategies, such as favoring text over visual input in GenAI image tools or conventional phrasing in voice commands. Across our GenAI cases, this reliance was reinforced by inconsistent responsiveness: participants preferred input modes that seemed more legible to the system, even when novel modalities were available. For example, participants in **Cases A and B** often relied on their expectations of how in-car voice assistants typically behave, referring to GUI elements as if the systems followed deterministic interaction rules. Similarly, in **Cases C and D**, some designers preferred text prompts over sketches or annotations, citing prior training and confidence in text-based interaction. This reliance limited participants' exploration of new affordances and, consequently, the study's ability to evaluate novel interaction designs.

Takeaway: GenAI's unpredictability **amplifies** users' reliance on familiar input modes, reducing their willingness to explore new affordances and constraining what lab studies can reveal about interaction designs.

4.2 C2. Trade-offs between fidelity and control are amplified by GenAI's output variability

Interactive system studies often aim to balance **experimental control** with **ecological realism**—a long-standing methodological tension in HCI. In GenAI evaluations, however, this balance becomes especially difficult because stochastic outputs introduce additional, uncontrollable variability on top of existing system behavior. Low-fidelity setups (e.g., scripted responses or Wizard-of-Oz methods) can increase consistency but may limit the generative qualities that characterize GenAI, while high-fidelity prototypes capture more authentic behavior but can introduce noise and unpredictability that complicate study outcomes. In **Case D**, low-fidelity prototypes increased consistency but limited opportunities for exploration. Conversely, in **Cases A, B, and D**, fully functional prototypes enabled genuine generative interactions yet led to latency and unexpected responses. Participants expressed confusion when outputs deviated from expectations, despite prior briefings about system limitations.

Takeaway: This challenge extends a familiar HCI problem—balancing fidelity and control—but GenAI's higher degree of output

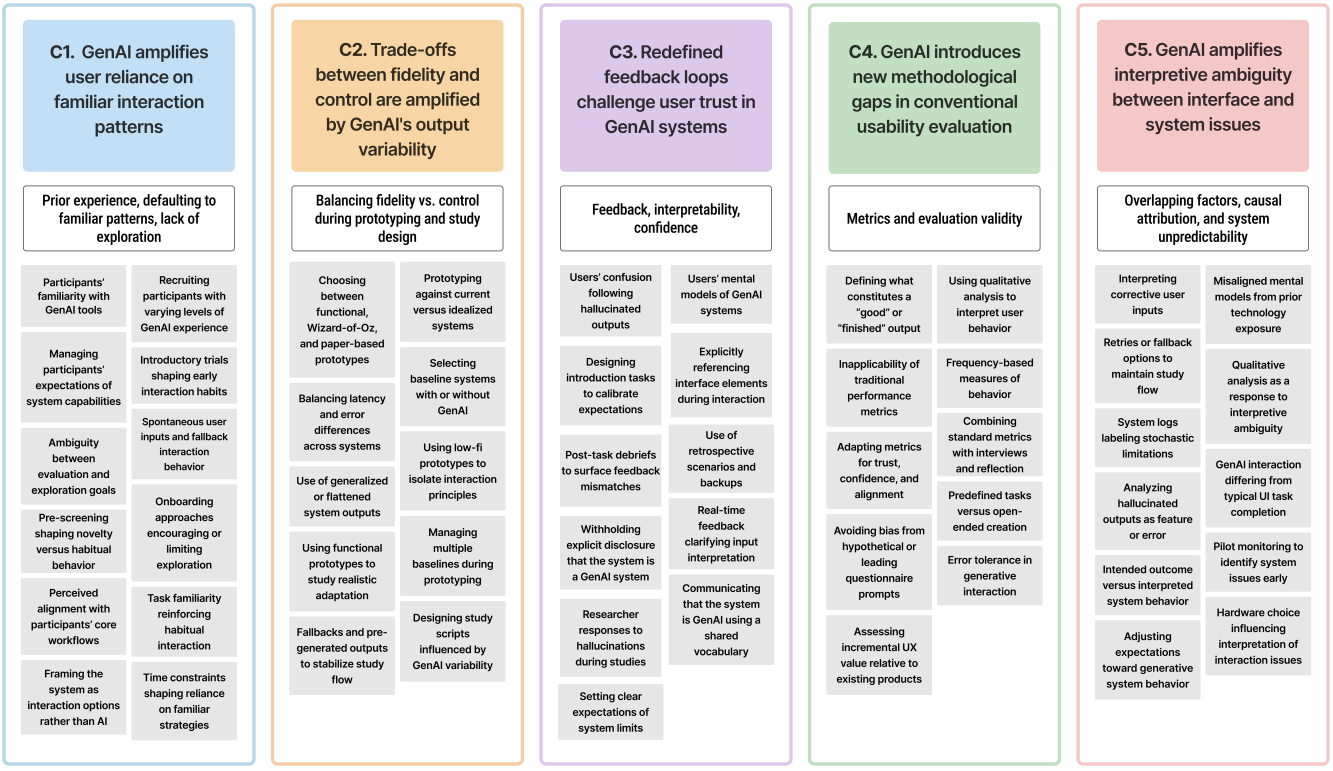


Figure 1: Visualization of our inductive thematic coding process from initial analytic codes (light gray boxes) to subthemes (white boxes) and five higher-order themes (C1–C5). Each column represents one methodological challenge that emerged through iterative comparison and clustering across studies. Detailed documentation of affinity notes, analytic codes, subthemes, themes, and methodological reflections is provided in the supplementary material.

variability **amplifies** this impact, making design choices methodological in nature, as they directly affect study reliability and determine which types of interactions can be meaningfully observed.

4.3 C3. Redefined feedback loops challenge user trust in GenAI systems

In conventional interfaces, feedback mechanisms are typically designed to be interpretable, enabling users to infer the relationship between their input and the resulting output. GenAI systems complicate this loop, as stochastic generation and opaque model behavior make it unclear how inputs are processed or acknowledged. Across our studies, participants frequently expressed doubt about whether the system recognized or understood their input, especially when using less familiar modalities (e.g., voice in the car study, scribbles in the design study). In **Case B**, conversational hallucinations led participants to question whether miscommunication stemmed from their phrasing or from the model's unpredictable generation of irrelevant responses. In **Case D**, scribble-based prompts sometimes yielded mismatched or irrelevant images, leading participants to wonder whether the system had misunderstood their visual input or failed to align it with the intended concept. In contrast, text prompts provided more consistent and traceable responses, which participants perceived as more reliable.

Takeaway: In GenAI systems, feedback is not simply less reliable but **redefined**. Users may hesitate to explore unfamiliar modalities not because of interface flaws, but because feedback can be unstable or ambiguous, which can disrupt trust and engagement during the study itself. This instability can limit the extent to which evaluators can reliably infer from observed interaction behavior.

4.4 C4. GenAI introduces new methodological gaps in conventional usability evaluation

Standard usability measures such as the System Usability Scale (SUS) or the User Experience Questionnaire (UEQ) were developed under the assumption of systems that produce relatively consistent and reproducible responses. These instruments assume clear input–output relationships and stable performance, allowing numerical scores to reflect interface usability. GenAI systems violate these assumptions: stochastic and context-dependent outputs mean that user ratings often capture variability in system behavior rather than interface design quality. In **Case B**, SUS ratings were frequently influenced by factors such as hallucinations, latency, or unexpected output variation rather than interface design, as indicated by follow-up interviews. Similarly, in **Case D**, low UEQ scores for scribble-based tasks were sometimes attributed to confusing or incoherent outputs rather than to interaction design flaws.

Such findings suggest that established usability metrics may not reliably separate interface-related issues from variability arising from generative system behavior. To support a valid interpretation, qualitative observations and mixed methods (e.g., interviews, think-aloud protocols) were essential for contextualizing numerical results.

Takeaway: GenAI creates a **new methodological gap** in usability evaluation. Traditional metrics assume consistency and determinism, yet user frustration often arises from model unpredictability rather than interface design, highlighting the need for new or hybrid evaluation approaches that account for stochastic behavior.

4.5 C5. GenAI amplifies interpretive ambiguity between interface and system issues

In conventional systems, usability breakdowns can often be traced to specific causes, such as interface design issues or users' interaction mistakes. In GenAI studies, however, overlapping factors—interface affordances, user strategies, and model behavior—make it challenging to determine the origin of observed problems. Unlike the interaction-level uncertainty described in Section 4.3, this challenge instead arises during **post-hoc interpretation**, when researchers attempt to attribute causes to observed outcomes. Across our cases, participant confusion or task failures often stemmed from intertwined system- and interface-level factors, blurring evaluative judgment. This ambiguity is further amplified when variability in system output makes it harder to attribute observed behavior to interface design decisions rather than underlying model behavior. In **Case D**, irrelevant image outputs from scribbles could signal either poor interface affordances or stochastic model behavior. Similarly, in **Case B**, hallucinated or delayed responses disrupted task flow, making it unclear whether confusion arose from design limitations or the model's unpredictable processing. Without precise system logging or detailed observational notes marking critical events (which tend to occur more frequently in GenAI systems than in conventional ones), tracing the source of such breakdowns and interpreting their methodological implications reliably becomes more difficult.

Takeaway: GenAI's unpredictability **amplifies** existing interpretive challenges in usability evaluation. Overlapping effects from user behavior, interface design, and model responses blur causal boundaries, making it harder to determine whether observed issues reflect design flaws or generative variability.

5 Guidelines for Designing and Evaluating GenAI Systems in Controlled Studies

Building on the five methodological challenges (C1–C5) identified in controlled lab studies, we developed **five methodological guidelines (G1–G5)**, each accompanied by a set of **practice-oriented recommendations**. While the challenges highlight how GenAI complicates established evaluation practices, the guidelines provide actionable strategies for designing, conducting, and analyzing GenAI lab studies more effectively. Each guideline responds to a core methodological issue, including preparing participants for stochastic systems (G1), balancing control and fidelity (G2), improving feedback interpretability (G3), adapting usability metrics (G4), and strengthening post-hoc interpretation (G5). They aim to help HCI and UX researchers anticipate, document, and mitigate the unique methodological tensions introduced by GenAI's non-deterministic behavior. Table 2 summarizes the five guidelines and their corresponding recommendations, while Figure 2 illustrates how these guidelines relate to the five challenges across study phases.

5.1 G1. Prepare Participants for Non-Deterministic System Behavior

Participants often approach GenAI systems with expectations shaped by prior experiences with deterministic interfaces (e.g., voice assistants or design tools). When outputs vary across identical inputs, these expectations can lead to confusion, frustration, or misattribution of errors to the interface. Preparing participants for stochastic behavior helps them interpret variability as an inherent property of the system rather than as a failure. The four recommendations under G1 address participant preparation as a gradual process, starting with shaping expectations during system introduction (R1.1), accounting for prior experience (R1.2), guiding initial interactions through structured onboarding (R1.3), and concluding with free exploration in a low-pressure setting (R1.4).

R1.1 Frame the system around interaction possibilities. How a system is introduced influences how participants approach it. Framing it as "AI-powered" or referencing commercial tools (e.g., ChatGPT, Midjourney) can narrow participants' expectations and lead them to rely on familiar input patterns even when other modalities are available. To encourage broader exploration, present the system in terms of its **input possibilities, supported tasks, and limitations** rather than emphasizing its AI identity. In **Case C and D**, where the system was introduced as an AI tool, users often defaulted to text prompts despite having access to scribbles and annotations, reflecting prior experience with text-based GenAI tools that participants cited when reasoning about their interaction strategies. In contrast, in **Cases A and B**, the systems were presented through their task functions, and information about the underlying LLMs was disclosed only after the study, which appeared to promote more varied and open-ended interaction.

R1.2 Screen for prior interaction experience. Participants' previous exposure to specific input modalities can shape both their confidence and their evaluation of usability. Without accounting for this, it becomes difficult to distinguish genuine usability issues from those arising from **unfamiliarity** (e.g., hesitation or confusion) or **over-familiarity** (e.g., reluctance to explore alternative inputs). We therefore recommend **pre-screening participants** for prior experience with modalities such as in-car voice assistants, stylus input, or prompt-based systems to help support more accurate interpretation of observed behaviors. In **Case B**, the absence of such screening meant that participants defaulted to typical home-assistant commands when they were uncertain how to proceed. In contrast, **Case A** recruited experienced in-car voice assistant users, which enabled clearer differentiation between skill-based challenges and system-related issues. Familiarity also shaped exploration. In **Cases C and D**, participants with strong prior experience using text-based GenAI tools often defaulted to text prompts, while those without prior stylus experience struggled initially. Together, these familiarity effects complicated attribution, making it difficult

Table 2: Overview of the five methodological guidelines and their associated recommendations, showing how they address core challenges in evaluating GenAI systems.

G#	Guideline	Key Recommendations
G1	Prepare participants for non-deterministic system behavior *(Participant readiness)*	**R1.1** Frame the system around interaction possibilities **R1.2** Screen for prior interaction experience **R1.3** Design contextual onboarding to promote exploration **R1.4** Offer a low-pressure trial phase before formal tasks
G2	Align prototype fidelity to study goals *(Fidelity & control)*	**R2.1** Choose prototype fidelity according to study objectives **R2.2** Manage unpredictability through selective backend adjustments **R2.3** Prepare fallback strategies proactively to sustain study flow **R2.4** Document system behavior and contextual variables **R2.5** Design tasks that reflect GenAI's exploratory nature
G3	Improve feedback interpretability and user trust *(Feedback & trust)*	**R3.1** Make the system feedback loop interpretable across input and output **R3.2** Provide real-time input feedback for immediate transparency **R3.3** Use post-task debriefs to identify mismatches between user intent and system behavior
G4	Adapt evaluation strategies to capture GenAI-specific experiences *(Evaluation strategies)*	**R4.1** Expand evaluation metrics to capture GenAI-specific constructs **R4.2** Pair standardized metrics with qualitative reflections
G5	Build flexibility into GenAI study design and analysis *(Researcher adaptation)*	**R5.1** Anticipate system issues through pilot testing and live monitoring **R5.2** Respond flexibly to system failures to preserve study continuity **R5.3** Label system limitations in logs to ensure transparent analysis

to determine whether performance issues stemmed from the input method itself, habitual interaction patterns, or device unfamiliarity.

R1.3 Design contextual onboarding to encourage exploration. Even when prior experience is taken into account, participants still require orientation to the specific interaction context of the study. Structured onboarding tasks that mirror the main study activities help participants understand how to interact with the system and adapt existing habits to new modalities. When onboarding aligns with task goals—such as editing images or navigating maps—participants are more willing to experiment with unfamiliar input methods. In **Case D**, a warm-up task that incorporated text, scribbles, and annotations closely mirrored the main study and helped participants use a broader range of inputs later. By contrast, in **Case B**, a scripted onboarding that focused on a generic voice example did not sufficiently prepare participants to use the GUI-referencing feature, leading them to rely on familiar command-style interactions. Aligning onboarding more closely with actual study tasks can therefore help participants bridge the gap between prior experience and the intended interaction design, promoting richer exploration and engagement.

R1.4 Offer a low-pressure trial phase before formal tasks. After completing structured onboarding, participants often still need space to explore new input methods without performance pressure. A short **exploratory trial phase** before formal tasks allows participants to internalize what they learned during onboarding and to build confidence in how the system interprets their actions. This phase can include unstructured interaction as well as light demonstrations, such as showing how a scribble is recognized or presenting example prompts. In **Case B**, participants who lacked early opportunities for open-ended exploration worried about providing "incorrect" input, whereas in **Case D**, those who were able to experiment freely beforehand became more confident using scribbles and annotations during the main tasks. Providing a low-pressure exploration phase helps participants translate structured learning into flexible engagement, reducing anxiety and fostering more authentic interaction with unfamiliar inputs.

5.2 G2. Align Prototype Fidelity to Study Goals

Balancing prototype fidelity with experimental control is a recurring challenge in evaluating GenAI systems (see C2, Section 4.2). High-fidelity prototypes enable authentic interactions but introduce unpredictable outputs that compromise consistency, whereas low-fidelity setups increase control but sacrifice the generative qualities that define GenAI. This trade-off, therefore, becomes a methodological choice rather than a purely technical one. The five recommendations for G2 outline how to align fidelity with study goals, address system variability, and maintain both experimental rigor and ecological validity.

R2.1 Choose prototype fidelity according to study objectives. Prototype fidelity should align with the main research goal, whether to explore input preferences, observe adaptation to system behavior, or evaluate output quality. In early-stage investigations that focus on **input exploration** or interaction patterns, simplified or non-functional prototypes help isolate user strategies. This avoids

confounding effects such as latency or unpredictable output. In **Case C**, a paper-based prototype was used to examine input choices (text, scribbles, annotations) without system interference, enabling clearer observation of interaction patterns. In contrast, in **Case D**, a fully functional image generation prototype allowed researchers to observe real-time adaptation but also revealed issues with latency and output errors. Choosing fidelity with intention helps ensure methodological coherence between research goals, study conditions, and interpretive validity.

R2.2 Manage unpredictability through selective backend adjustments. Functional prototypes are essential when studying how users adapt to GenAI systems in realistic conditions, such as refining prompts or responding to variable outputs. Although they reduce experimental control, they reveal interaction patterns that only emerge during live system use. To balance realism and reliability, researchers can manage unpredictability through selective adjustments to the backend, such as refining prompts, tuning model parameters, or adding contextual guidance. In **Case D**, a fully functional prototype supported real-time image generation from text, scribbles, and annotations. To reduce excessive output variability while preserving generative behavior, backend prompts were constrained during closed-ended tasks to better align with task goals, whereas fewer prompt constraints were applied during open-ended phases to allow broader exploration. This setup enabled observation of how participants refined their inputs in response to varied outputs under different degrees of system control. In **Case A**, backend adjustments to the voice assistant prompt (e.g., including navigation examples) reduced hallucinations without diminishing the system's perceived authenticity. Balancing fidelity and control through targeted technical adjustments helps researchers capture GenAI-specific interaction dynamics while maintaining interpretive validity across participants.

R2.3 Prepare fallback strategies proactively to sustain study flow. Even with backend control, GenAI systems can fail to produce coherent or timely outputs. Such disruptions can interrupt task flow and frustrate participants, thereby reducing the reliability of collected data. We therefore recommend designing **fallback strategies**, such as pre-generated outputs, scripted alternatives, or structured opportunities to re-prompt, to maintain task continuity when system breakdowns occur. In **Case D**, participants were allowed to re-prompt the system when image generation failed, which helped them remain engaged and complete the task despite interruptions. During analysis, the research team noted that preparing additional fallback materials, such as pre-generated images, could further support continuity in similar studies. By contrast, in **Case B**, the absence of fallback options forced participants to restart interactions manually, leading to frustration and loss of focus. Overall, proactive fallback planning supports both data quality and participant engagement, while documenting such events enables richer post-study analysis of system breakdowns and recovery strategies.

R2.4 Document system behavior and contextual variables. Because GenAI systems can produce variable and unpredictable responses, detailed documentation of system behavior is essential for credible interpretation and replication. Logging prompts, outputs, latency, and contextual variables allows researchers to understand how system performance shapes participants' experiences and to attribute observed behaviors more accurately. In **Case D**, extensive event logging and synchronized voice recordings enabled the research team to trace each interaction and examine how users adapted to system responses in real time. This comprehensive documentation provided valuable insights during analysis and helped differentiate between user behavior, interface design, and stochastic model variation. Overall, thorough and transparent logging ensures that GenAI evaluation remains interpretable and reproducible, enabling researchers to distinguish between design-related issues and variability inherent to generative systems.

R2.5 Design tasks that reflect GenAI's exploratory nature. Study tasks should represent how users naturally engage with GenAI systems through iteration, experimentation, and adaptation. Restricting interaction to single attempts limits realism and prevents observation of how participants refine their input in response to system outputs. In **Case D**, participants frequently modified prompts or sketches in response to system feedback but were constrained by a study-imposed three-iteration limit, which shaped their interaction strategies. By contrast, **Case B** deliberately incorporated an open-ended task that allowed participants to freely ask questions to the in-car assistant, revealing more spontaneous exploration patterns. Similarly, in **Case A**, participants interacted with the system through loosely defined goals, such as setting destinations, asking about car functions, or initiating casual conversation. To support hesitant participants, the research team prepared a small set of fallback ideas to help maintain engagement. Designing tasks that support structured yet flexible iteration better captures the exploratory nature of GenAI use and provides more ecologically valid insights into user adaptation. Choosing appropriate prototype fidelity, managing unpredictability transparently, and enabling iterative exploration together help preserve both experimental rigor and the authentic dynamics of generative interaction.

5.3 G3. Improve Feedback Interpretability and User Trust

As identified in Challenge C3 (Section 4.3), GenAI systems **redefine** feedback. Their responses are not always consistent or interpretable, which can make it difficult for participants to form stable mental models or trust the system's behavior. Unlike conventional interfaces, where the relationship between input and output is transparent, GenAI feedback can be ambiguous, delayed, or seemingly unrelated to the input. The three recommendations under G3 address feedback interpretability as a continuous process: clarifying how the feedback loop operates across input and output (R3.1), providing real-time cues that confirm input recognition across modalities (R3.2), and calibrating participant expectations regarding feedback reliability (R3.3).

R3.1 Make system feedback loop interpretable across input and output. Participants often struggle to understand whether their input was received and why outputs vary, particularly when feedback is delayed or inconsistent. Establishing transparency throughout the feedback loop—by acknowledging inputs and contextualizing output variation—can help participants build trust and maintain

engagement. Providing clear indicators of input recognition (e.g., progress cues or acknowledgment tones), along with brief explanations of why outputs differ, can support a more predictable interaction flow. In **Case B**, latency and missing confirmation cues led some participants to repeat voice commands, unsure whether their input had been processed. In **Case D**, participants initially viewed inconsistent image results as system errors; however, brief clarifications that the model was intentionally exploring multiple interpretations helped participants understand output variation as a design feature rather than as an error. In **Case A**, subtle acknowledgments of recognized speech and contextually adaptive phrasing appeared to encourage participants to view response differences as flexibility rather than faults. Across these cases, these examples suggest that making feedback interpretable across both input and output can reduce uncertainty and support user confidence.

R3.2 Provide real-time input feedback for immediate transparency. Moment-to-moment uncertainty often arises when participants are unsure whether the system has captured their input or is still processing it. This hesitation can disrupt task flow, particularly in modalities such as voice or sketches where input recognition is less visible. We recommend implementing **real-time feedback cues** that visualize how the system interprets user input, such as speech-to-text transcriptions, highlighted drawings, or short textual summaries of recognized content. Such cues can provide immediate reassurance that input has been processed and allow participants to stay focused on evaluating system behavior. In **Cases A and B**, on-screen transcriptions improved transparency but did not fully prevent repeated queries during latency delays, as users sometimes remained unsure when the assistant was "listening." In **Case D**, participants tended to trust text prompts more than scribbles because textual input offered visible acknowledgment, while sketches lacked explicit confirmation. Providing timely, modality-specific feedback can therefore help minimize hesitation and support smoother, more confident engagement.

R3.3 Use post-task debriefs to identify mismatches between user intent and system behavior. Not all misunderstandings between participants and GenAI systems are visible during interaction. Participants may misinterpret responses, question their own input, or blame themselves without expressing this uncertainty in real-time. Structured post-task debriefs can help identify and surface hidden mismatches between user intent and system behavior. We recommend a brief set of follow-up questions after each task, such as "What did you expect the system to do?" or "What do you think the system understood?" These reflections can reveal unspoken confusion and help researchers interpret observed behaviors more accurately. In **Case D**, participants sometimes attributed unexpected images to their "bad writing," later explaining that they were unsure whether the input had been recognized. Similarly, in **Case B**, post-task interviews revealed uncertainty about whether spoken inputs were understood when responses were delayed or off-topic. Post-task debriefs thus provide important context for interpreting user behavior beyond what is directly observable during the study.

5.4 G4. Adapt Evaluation Strategies to Capture GenAI-Specific User Experiences

As discussed in Challenge C4 (Section 4.4), evaluating GenAI systems using traditional usability metrics, such as SUS or UEQ, can lead to incomplete or potentially misleading conclusions. Because GenAI outputs are variable and sometimes hallucinated, user frustration or confusion may stem from model behavior rather than from interface design alone. Conventional usability scales assume consistent, deterministic system responses, which can limit their ability to capture GenAI-specific phenomena such as unpredictability, trust, or intent alignment. The two recommendations under G4 outline how to adapt evaluation strategies to these conditions: by extending what is measured with GenAI-specific constructs (R4.1) and by combining standardized metrics with qualitative reflections for deeper interpretability (R4.2).

R4.1 Expand evaluation metrics to capture GenAI-specific constructs. Traditional usability scales quantify satisfaction and efficiency but often overlook experiential factors central to GenAI interaction, such as trust, confidence, intent alignment, and comfort with uncertainty. We therefore recommend integrating targeted questions such as "Did you trust the system's output?", "How confident were you that your input was understood?", or "Did the result match your intent?" to better reflect these GenAI-specific aspects of user experience. In **Case B**, post-task reflections indicated that participants' satisfaction was driven less by response accuracy and more by how "understood" they felt by the voice assistant. Similarly, in **Case D**, participants sometimes rated the same image output differently depending on whether they believed the system had captured their intent. Including items that capture perceived understanding or intent alignment (e.g., "I felt the system understood what I meant") can provide quantifiable yet context-specific data that complements standard usability metrics.

R4.2 Pair standardized metrics with qualitative reflections. While SUS and UEQ summarize user perceptions quantitatively, they rarely reveal whether ratings primarily reflect interface design or GenAI-specific variability. We recommend pairing standardized scores with short, open-ended reflections after each task, complemented by observation or think-aloud protocols. Follow-up prompts such as "What did you expect to happen?" or "Was the response what you intended?" can help clarify how participants interpreted their experiences. In **Cases A and B**, SUS and UEQ offered a general overview of usability, but interviews suggested that lower scores were often influenced by latency or hallucinations rather than by input design. In **Cases C and D**, think-aloud sessions revealed uncertainty about whether scribbles were interpreted, even in the absence of formal metrics. Combining structured ratings with qualitative insights thus helps ensure that findings reflect both the measurable and interpretive aspects of GenAI interactions.

5.5 G5. Build Flexibility into GenAI Study Design and Analysis

As discussed in Challenge C5 (Section 4.5), distinguishing usability issues from GenAI-specific limitations, such as hallucinations, latency, or backend instability, proved difficult in our studies. These ambiguities complicate analysis and call for flexible, reflexive study

designs. Researchers, therefore, need to be prepared to adapt tasks, logging, or analytic strategies in response to unexpected system behavior or participant confusion. The three recommendations under G5 highlight how flexibility can be integrated throughout a study: proactively monitoring system reliability (R5.1), maintaining task flow during disruptions (R5.2), and labeling the system's limitations to support accurate interpretation (R5.3).

R5.1 Anticipate system issues through pilot testing and adapt during live monitoring. GenAI systems can fail unpredictably, producing long delays, repeated hallucinations, or instability that disrupts the task flow. We recommend **monitoring system behavior continuously** during both pilot testing and live sessions to help detect issues early and enable timely adjustments. This can include logging outputs in real-time and preparing adaptive responses, such as prompt modifications or fallback content. In **Case B**, pilot feedback identified latency as a major issue, and iterative adjustments helped reduce delays that might otherwise have been mistaken for usability flaws. In **Case D**, pre-study testing revealed performance differences between two image-generation models, and the team alternated between the two models to maintain acceptable responsiveness while preserving realistic interaction behavior. Continuous monitoring and real-time intervention can help ensure that technical failures do not unduly distort user evaluation.

R5.2 Respond flexibly to system failures to preserve study continuity. Unlike R2.3 (Section 5.2), which emphasizes proactively preparing fallback strategies before a study to reduce unpredictability, this recommendation focuses on **reactive adaptation during data collection**. When GenAI systems fail or generate unusable outputs, continuing the original task can frustrate participants and compromise data quality. Researchers may therefore need to adjust tasks, repeat inputs, or offer workarounds to sustain engagement and support meaningful data collection despite disruptions. We recommend preparing **alternative task paths**, such as allowing retries, providing pre-generated outputs, or skipping tasks when failures persist, as well as **time-boxing activities** to balance flexibility with session duration. All in-session adaptations should be **documented** so they can be considered during later analysis. In **Case D**, participants could retry image generation up to three times; if failures persisted, they were guided to skip the task. This approach helped maintain flow and later prompted discussion about the potential value of pre-generated examples for continuity. In **Case B**, participants were allowed to retry tasks or restart voice input using the push-to-talk button, providing a simple yet effective recovery mechanism. Such reactive flexibility can prevent technical breakdowns from derailing studies and help ensure that adaptive decisions are captured and reflected during subsequent interpretation.

R5.3 Label system limitations in logs to ensure transparent analysis. To interpret study results accurately, researchers need to separate interface-related challenges from system-side issues. We recommend systematically labeling known limitations, such as "latency > 3s," "output failures," or "hallucination detected," within session logs. Such annotations help clarify when participant hesitation or performance drops are attributable to system behavior rather than interface design. In **Case D**, logging which model generated each image helped differentiate latency-related pauses from genuine interaction difficulties. In **Case B**, separating sessions with and without hallucinations clarified which usability scores reflected user experience versus technical artifacts. These annotations create an "audit trail" for later analysis, supporting more reliable interpretation and facilitating replication by other researchers. Labeling the system's limitations in this way can enhance analytic transparency and strengthen confidence in the reported findings.

6 Discussion

Our findings highlight challenges and recommendations that illustrate how GenAI systems are reshaping controlled lab studies and HCI evaluation. Below, we discuss how these shifts affect study design, user trust, and evaluation metrics, and we reflect on how a reflective multi-case approach helps surface emerging methodological challenges.

6.1 How GenAI Evaluation Extends and Reframes Prior HCI Evaluation

Our results extend long-standing HCI debates about fidelity, control, and trust. As in prior work on adaptive or novel-sensing systems that require context-sensitive and longitudinal observation [27, 43, 50], GenAI variability disrupts the tight input–output mappings assumed by many lab protocols. It also intensifies the fidelity–control trade-offs familiar from Wizard-of-Oz and prototyping research [29]. GenAI does not merely introduce non-determinism that complicates study design; it *can* also generate new challenges while amplifying and reframing existing ones.

Several tensions identified in our analysis—such as users reverting to familiar interaction strategies, difficulties interpreting system behavior, and trade-offs between control and realism—have also been discussed in evaluations of other adaptive or intelligent systems, including speech recognition and recommendation systems [11, 13, 62]. Our contribution is to articulate how the high degree of output variability in GenAI systems amplifies these tensions by coupling probabilistic generation with feedback ambiguity. For example, variability may support exploration in some contexts while undermining trust in others, including through hallucinations. We therefore frame these challenges as amplified or reframed methodological concerns, or issues that emerge specifically in the context of generative systems, rather than as entirely unprecedented problems. Challenges C1–C3 primarily stem from stochastic outputs and the lack of clear input–output mappings. Unpredictable feedback can discourage deviation from familiar strategies (C1) and disrupt fidelity–control trade-offs (C2) and feedback loops (C3), thereby undermining both user confidence and experimental control. Challenges C4–C5 arise from the opacity of system behavior and the entanglement of model and interface effects, complicating evaluation and prompting a reconsideration of how success and usability are defined.

One could argue that non-determinism ultimately reflects training data and tunable randomness (e.g., temperature or sampling strategies). However, we contend that this unpredictability is precisely what makes these systems worth studying: variability enables open-ended dialogue, exploration of alternatives, and forms of collaboration that deterministic systems cannot support. Rather than suppressing variability, researchers should examine whether

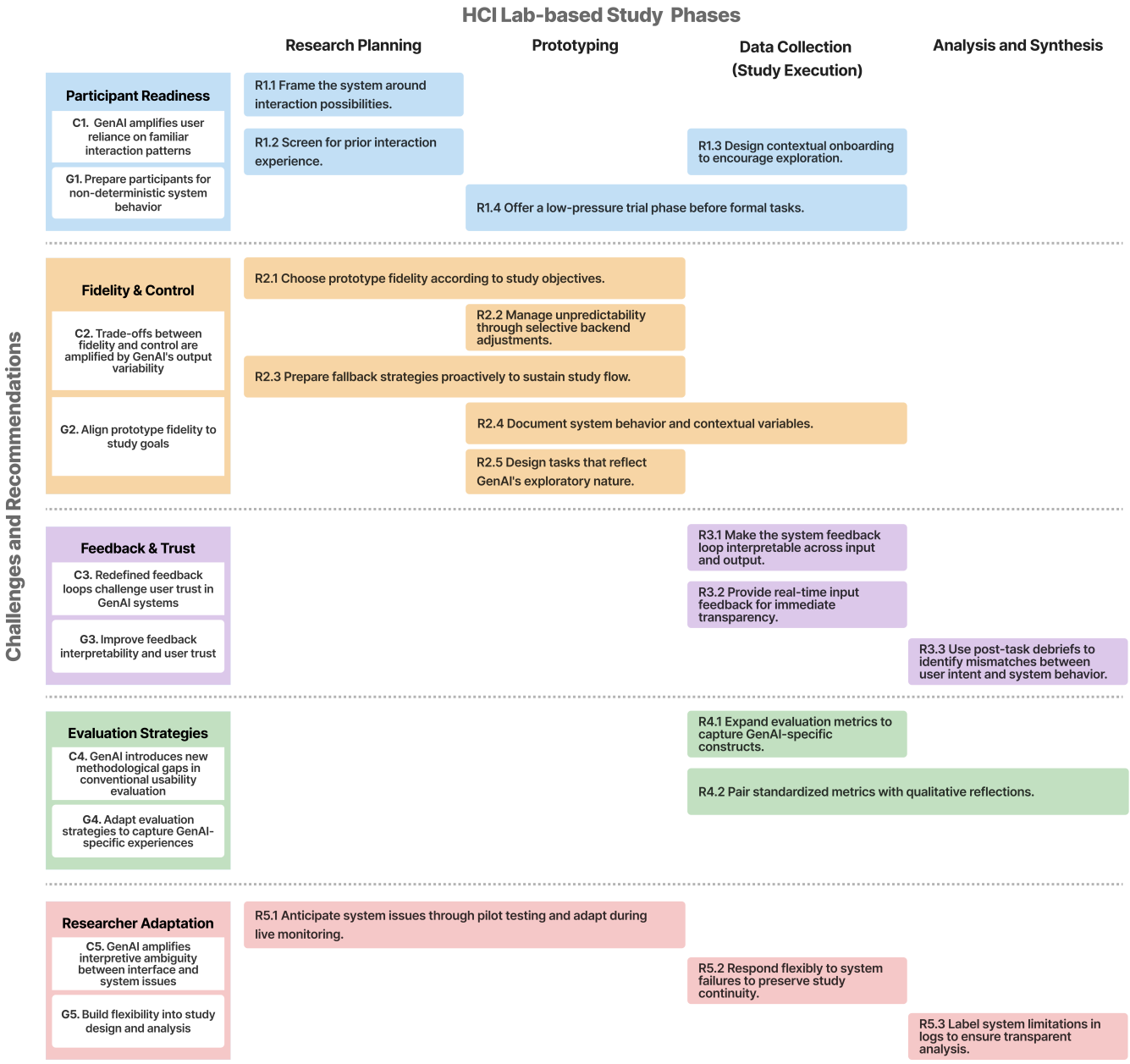


Figure 2: Methodological challenges (C1–C5) and corresponding guidelines (G1–G5) with all eighteen recommendations (R1.1–R5.3) for evaluating GenAI systems in HCI lab studies. The figure illustrates where each recommendation primarily applies across research phases, from planning and prototyping to data collection and analysis.

systems support exploration and recovery, and how users adapt interaction patterns over multiple steps to reach their goals. In considering new approaches to studying GenAI, we argue that methodological transparency is critical. We examined our own research process—including the decisions made, trade-offs encountered, and evolving strategies—and how these choices shaped study procedures and outcomes. Decisions such as constraining inputs, simulating outputs, or tolerating variability were not neutral in their effects; they influenced what could be observed and how findings were interpreted. We therefore recommend reflective case comparisons as a useful way to surface methodological insights for emerging technologies. Looking ahead, as models expand their context windows and adopt more agentic behaviors, additional challenges will likely require longitudinal and in-the-wild studies to understand how trust, adaptation, and success criteria evolve over time.

6.2 Study Design Strategies for User Trust and Confidence

Even when the underlying GenAI model is opaque, a well-designed study can help foster transparency for participants. Across our case studies, three strategies proved particularly effective. First, onboarding that emphasized input possibilities rather than the system's "AI" identity encouraged exploration and reduced hesitation when trying unfamiliar input methods. Second, during interaction, real-time feedback cues—such as transcribed speech or visual confirmations of recognized input—helped participants understand how their actions were interpreted and reassured them that their input had been received. These strategies were especially important when participants encountered unexpected outputs that stemmed not from their input or interface design, but from the system's generative behavior. Making this distinction explicit during the study helped reduce confusion and provided a clearer basis for interpreting participants' responses. Finally, structured post-task reflection questions revealed mismatches between participants' intentions and the system's responses that were not apparent from interaction logs alone. These strategies illustrate how transparency can be purposefully designed into GenAI evaluations, helping participants navigate uncertainty while enabling researchers to interpret behavior more reliably.

6.3 Rethinking What and How We Measure in GenAI Evaluation

Conventional usability scales (e.g., SUS, UEQ) remain useful for comparability, but in our GenAI studies, they often lacked expressiveness for the underlying issues that users encountered. Low usability scores could result from interface design, unpredictable outputs, hallucinations, or mismatches between user intentions and system responses. Relying solely on these standard metrics may therefore fail to capture the nuanced sources of user frustration and can lead to misinterpretation. To address this limitation, we suggest extending standard scales with GenAI-specific constructs such as user trust, confidence, and intent alignment. Adding short probes after each task (e.g., "What did you expect to happen?" or "How do you think the system understood your input?") helped contextualize numerical ratings and revealed mismatches that were otherwise invisible in logs or scale data. Our studies further showed that different methods illuminate different layers of interaction. Usability ratings summarized perceptions of ease, efficiency, and satisfaction; interviews explained participants' reasoning behind those scores; and observations highlighted challenges that emerged during real interactions. While each method contributes a distinct perspective, relying on a single approach can lead to blind spots. This aligns with long-standing calls in HCI and the social sciences for methodological triangulation, in which multiple methods are combined to compensate for individual limitations and build a more complete picture of interaction [37, 66]. For GenAI evaluation, mixed-method strategies, such as combining metrics with reflections or triangulating observation, can support more reliable interpretation and help distinguish interface-related issues from GenAI-specific effects [71].

6.4 Limitations and Future Work

Our findings are derived from four short-term, lab-based studies with specific prototypes and user groups. This scope allowed us to identify recurring methodological challenges, but it constrains the empirical generalizability of our findings and the techno-ecological validity of our conclusions. Accordingly, our contribution should be understood as methodological insights derived through cross-case reasoning, rather than as claims that are empirically replicable or statistically generalizable. Nonetheless, highlighting these patterns within constrained settings offers a valuable starting point for understanding how GenAI complicates established lab-based evaluation practices. Additionally, GenAI systems and users' mental models evolve rapidly. Some challenges may diminish as model capabilities improve or as users become more proficient with GenAI tools. We did not study long-term adaptation, in-the-wild use, or a broader range of application domains, all of which represent important directions for future work and motivate extending analyses to a more diverse set of case studies. Such extensions may surface additional challenges or complement those reported here. We therefore do not aim to provide an exhaustive review but instead report a reflective cross-case analysis grounded in an in-depth understanding of the research process. To make our analysis steps and underlying reflections transparent, we provide the coding schema and theme descriptions in the supplementary materials. While these materials document how the analysis was conducted, reproducing the cross-case analysis depends on access to a similarly structured study process. For this reason, we encourage documenting methodological trade-offs and reflections in a comparable manner.

Future studies could explore several additional avenues. One potential direction is to conduct comparative user studies of different evaluation strategies for GenAI systems. Variables such as fallback strategies (retrying versus pre-generated outputs), onboarding framings (AI-highlighted versus task-highlighted instructions), and feedback cues (verbatim transcripts versus interpreted summaries) may influence how participants develop trust, confidence, and input strategies. Such comparisons can help clarify when particular study designs lead to distinct user behaviors and refine best practices in GenAI evaluation. Another direction is to explore GenAI not only as a system under evaluation, but also as a research tool within HCI studies. Recent work has examined the use of LLMs as simulated participants [26] and as aids in qualitative analysis, such as thematic analysis [16]. Other studies have compared how human analysts classify qualitative data with how LLMs generate classifications and reasoning for the same material [2]. These efforts raise additional methodological questions about validity, interpretation, and researcher responsibility, suggesting that reflective evaluation practices will remain important as GenAI becomes increasingly embedded in the research process itself. Finally, we view our contribution as an initial step rather than a definitive account. Additional challenges and recommendations are likely to emerge as GenAI systems become more prevalent in everyday contexts, underscoring the need to revisit and refine evaluation practices over time.

7 Conclusion

In this work, we identified five recurring methodological challenges (C1–C5) in evaluating GenAI systems in controlled lab settings

through a reflective analysis of four case studies. These challenges reveal how GenAI's stochastic, non-deterministic behavior complicates established assumptions in HCI evaluation. In doing so, they expose tensions between control and realism, between deterministic expectations and probabilistic system behavior, and between user interpretation and model variability. Building on these insights, we proposed five methodological guidelines (G1–G5) and eighteen practice-oriented recommendations to support researchers in designing, conducting, and analyzing GenAI user studies more effectively. These guidelines include preparing participants for unpredictable behavior, aligning prototype fidelity with study goals, improving feedback interpretability, adapting evaluation metrics to account for stochasticity, and incorporating flexibility and transparency into study design and analysis. Overall, our work foregrounds the research process itself—its design choices, trade-offs, and interpretive challenges—when studying generative systems. Rather than offering a final framework, we aim to provide a foundation for continued methodological reflection as GenAI technologies become more integrated into everyday life. Continually revisiting and expanding these methodological discussions will be essential for building reliable and transparent HCI research on GenAI-integrated systems.

Acknowledgments

The authors used generative AI tools to improve readability (e.g., spelling, rephrasing, and formatting). All substantive content, analysis, and conclusions were developed by the authors, who retain full responsibility for the work.

References

[1] Buthayna AlMulla, Maram Assi, and Safwat Hassan. 2025. Understanding the Challenges and Promises of Developing Generative AI Apps: An Empirical Study. *arXiv preprint arXiv:2506.16453* (2025). doi:10.48550/arXiv.2506.16453

[2] Muneera Bano, Didar Zowghi, and Jon Whittle. 2023. Exploring qualitative research using LLMs. *arXiv preprint arXiv:2306.13298* (2023). doi:10.48550/arXiv.2306.13298

[3] Enrico Bertini, Catherine Plaisant, and Giuseppe Santucci. 2007. BELIV'06: beyond time and errors; novel evaluation methods for information visualization. *Interactions* 14, 3 (2007), 59–60. doi:10.1145/1358628.1358955

[4] Hugh Beyer and Karen Holtzblatt. 1999. Contextual design. *Interactions* 6, 1 (Jan. 1999), 32–42. doi:10.1145/291224.291229

[5] Ann Blandford, Dominic Furniss, and Stephann Makri. 2016. *Qualitative HCI research: Going behind the scenes.* Morgan & Claypool Publishers.

[6] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101. doi:10.1191/1478088706qp063oa

[7] Virginia Braun and Victoria Clarke. 2012. Thematic analysis. In *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological.*, Harris Cooper, Paul M. Camic, Debra L. Long, A. T. Panter, David Rindskopf, and Kenneth J. Sher (Eds.). American Psychological Association, Washington, 57–71. doi:10.1037/13620-004

[8] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative research in sport, exercise and health* 11, 4 (2019), 589–597. doi:10.1080/2159676X.2019.1628806

[9] Saša Brdnik, Tjaša Heričko, and Boštjan Šumak. 2022. Intelligent user interfaces and their evaluation: a systematic mapping study. *Sensors* 22, 15 (2022), 5830. doi:10.3390/s22155830

[10] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc., 1877–1901. https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[11] Daniel Buschek, Malin Eiband, and Heinrich Hussmann. 2022. How to Support Users in Understanding Intelligent Systems? An Analysis and Conceptual Framework of User Questions Considering User Mindsets, Involvement, and Knowledge Outcomes. *ACM Trans. Interact. Intell. Syst.* 12, 4, Article 29 (Nov. 2022), 27 pages. doi:10.1145/3519264

[12] Marc Gonzalez Capdevila, Toni Granollers Saltiveri, Juan Enrique Garrido, Octávio Henrique Müller, and Leonardo Coelho Ruas. 2021. Do current user testing practices meet the needs of the new interactive paradigms?. In *Proceedings of the XXI International Conference on Human Computer Interaction* (Málaga, Spain) *(Interacción '21)*. Association for Computing Machinery, New York, NY, USA, Article 22, 9 pages. doi:10.1145/3471391.3471416

[13] Chen Chen, Ella T Lifset, Yichen Han, Arkajyoti Roy, Michael Hogarth, Alison A Moore, Emilia Farcas, and Nadir Weibel. 2023. Screen or No Screen? Lessons Learnt from a Real-World Deployment Study of Using Voice Assistants With and Without Touchscreen for Older Adults. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility* (New York, NY, USA) *(ASSETS '23)*. Association for Computing Machinery, New York, NY, USA, Article 52, 21 pages. doi:10.1145/3597638.3608378

[14] John W Creswell and Vicki L Plano Clark. 2017. *Designing and conducting mixed methods research.* Sage publications.

[15] N. Dahlbäck, A. Jönsson, and L. Ahrenberg. 1993. Wizard of Oz studies — why and how. *Know.-Based Syst.* 6, 4 (Dec. 1993), 258–266. doi:10.1016/0950-7051(93)90017-N

[16] Shih-Chieh Dai, Aiping Xiong, and Lun-Wei Ku. 2023. LLM-in-the-loop: Leveraging large language model for thematic analysis. *arXiv preprint arXiv:2310.15100* (2023). doi:10.48550/arXiv.2310.15100

[17] Peter Dalsgaard and Kim Halskov. 2012. Reflective design documentation. In *Proceedings of the Designing Interactive Systems Conference* (Newcastle Upon Tyne, United Kingdom) *(DIS '12)*. Association for Computing Machinery, New York, NY, USA, 428–437. doi:10.1145/2317956.2318020

[18] Kathleen M Eisenhardt. 1989. Building theories from case study research. *Academy of management review* 14, 4 (1989), 532–550. doi:10.2307/258557

[19] Lokesh Fulfagar, Anupriya Gupta, Arpit Mathur, and Abhishek Shrivastava. 2021. Development and Evaluation of Usability Heuristics for Voice User Interfaces. In *Design for Tomorrow – Volume 1*. Smart Innovation, Systems and Technologies, Vol. 221. Springer, 375–385. doi:10.1007/978-981-16-0041-8_32

[20] Frederic Gmeiner, Kenneth Holstein, and Nikolas Martelaro. 2025. Prototyping Multimodal GenAI Real-Time Agents with Counterfactual Replays and Hybrid Wizard-of-Oz. *arXiv preprint arXiv:2510.06872* (2025). doi:10.48550/arXiv.2510.06872

[21] Saul Greenberg and Bill Buxton. 2008. Usability evaluation considered harmful (some of the time). In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. Association for Computing Machinery, New York, NY, USA, 111–120. doi:10.1145/1357054.1357074

[22] Johanna Gustafsson. 2017. Single Case Studies vs. Multiple Case Studies: A Comparative Study. Literature review, Academy of Business, Engineering and Science, Halmstad University, Sweden.

[23] Jessica He, Stephanie Houde, Gabriel E. Gonzalez, Darío Andrés Silva Moran, Steven I. Ross, Michael Muller, and Justin D. Weisz. 2024. AI and the Future of Collaborative Work: Group Ideation with an LLM in a Virtual Canvas. In *Proceedings of the 3rd Annual Meeting of the Symposium on Human-Computer Interaction for Work* (Newcastle upon Tyne, United Kingdom) *(CHIWORK '24)*. Association for Computing Machinery, New York, NY, USA, Article 9, 14 pages. doi:10.1145/3663384.3663398

[24] Adriana Lorena Iniguez-Carrillo, Laura Sanely Gaytan-Lugo, Miguel Angel Garcia-Ruiz, and Rocio Maciel-Arellano. 2021. Usability questionnaires to evaluate voice user interfaces. *IEEE Latin America Transactions* 19, 9 (2021), 1468–1477. doi:10.1109/TLA.2021.9468439

[25] Anniek Jansen and Sara Colombo. 2022. Wizard of Errors: Introducing and Evaluating Machine Learning Errors in Wizard of Oz Studies. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 426, 7 pages. doi:10.1145/3491101.3519684

[26] Shivani Kapania, William Agnew, Motahhare Eslami, Hoda Heidari, and Sarah E Fox. 2025. Simulacrum of Stories: Examining Large Language Models as Qualitative Research Participants. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 489, 17 pages. doi:10.1145/3706598.3713220

[27] Maria Kjærup, Mikael B Skov, Peter Axel Nielsen, Jesper Kjeldskov, Jens Gerken, and Harald Reiterer. 2021. Longitudinal studies in HCI research: a review of CHI publications from 1982–2019. In *Some Edited Volume Title or Proceedings (if applicable)*. Springer. doi:10.1007/978-3-030-67322-2_2

[28] Jesper Kjeldskov and Mikael B. Skov. 2014. Was it worth the hassle? ten years of mobile HCI research discussions on lab and field evaluations. In *Proceedings of the 16th International Conference on Human-Computer Interaction with Mobile Devices & Services* (Toronto, ON, Canada) *(MobileHCI '14)*. Association for Computing

Machinery, New York, NY, USA, 43–52. doi:10.1145/2628363.2628398
[29] Scott R. Klemmer, Anoop K. Sinha, Jack Chen, James A. Landay, Nadeem Aboobaker, and Annie Wang. 2000. Suede: a Wizard of Oz prototyping tool for speech user interfaces. In *Proceedings of the 13th Annual ACM Symposium on User Interface Software and Technology* (San Diego, California, USA) *(UIST '00)*. Association for Computing Machinery, New York, NY, USA, 1–10. doi:10.1145/354401.354406
[30] Kristina Kölln, Jana Deutschländer, Andreas M Klein, Maria Rauschenberger, and Dominique Winter. 2022. Identifying User Experience Aspects for Voice User Interfaces with Intensive Users. In *WEBIST*. 385–393. doi:10.5220/0011383300003318
[31] Thomas Kosch and Sebastian Feger. 2024. Risk or Chance? Large Language Models and Reproducibility in HCI Research. *Interactions* 31, 6 (Oct. 2024), 44–49. doi:10.1145/3695765
[32] Sebastian Krakowski. 2025. Human-AI agency in the age of generative AI. *Information and Organization* 35, 1 (2025), 100560. doi:10.1016/j.infoandorg.2025.100560
[33] Heidi Lam, Enrico Bertini, Petra Isenberg, Catherine Plaisant, and Sheelagh Carpendale. 2012. Empirical studies in information visualization: Seven scenarios. *IEEE transactions on visualization and computer graphics* 18, 9 (2012), 1520–1536. doi:10.1109/TVCG.2011.279
[34] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. 2010. *Research Methods in Human-Computer Interaction*. Wiley Publishing.
[35] Chaeyeon Lee, Chungnyeong Lee, Sangyong Kim, Yongsoon Choi, and Jusub Kim. 2025. StorageChat Timeline: A Generative AI-Based Art Appreciation System for Enhancing Immersion and Exploratory Experience. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 921, 2 pages. doi:10.1145/3706599.3721349
[36] I. Scott MacKenzie. 2015. User studies and usability evaluations: from research to products. In *Proceedings of the 41st Graphics Interface Conference (GI '15)*. Canadian Information Processing Society, CAN, 1–8.
[37] Joseph E McGrath. 1995. Methodology matters: Doing research in the behavioral and social sciences. In *Readings in human–computer interaction*. Elsevier, 152–169. doi:10.1016/B978-0-08-051574-8.50019-4
[38] Tamara Munzner. 2009. A nested model for visualization design and validation. *IEEE transactions on visualization and computer graphics* 15, 6 (2009), 921–928. doi:10.1109/TVCG.2009.111
[39] Roderick Murray-Smith, Antti Oulasvirta, Andrew Howes, Jörg Müller, Aleksi Ikkala, Miroslav Bachinski, Arthur Fleig, Florian Fischer, and Markus Klar. 2022. What simulation can do for HCI research. *Interactions* 29, 6 (Nov. 2022), 48–53. doi:10.1145/3564038
[40] Hyerim Park, Joscha Eirich, Andre Luckow, and Michael Sedlmair. 2024. "We Are Visual Thinkers, Not Verbal Thinkers!": A Thematic Analysis of How Professional Designers Use Generative AI Image Generation Tools. In *Proceedings of the 13th Nordic Conference on Human-Computer Interaction* (Uppsala, Sweden) *(NordiCHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 35, 14 pages. doi:10.1145/3679318.3685370
[41] Tibor Petzoldt, Hanna Bellem, and Josef F Krems. 2014. The critical tracking task: a potentially useful method to assess driver distraction? *Human factors* 56, 4 (2014), 789–808. doi:10.1177/0018720813501864
[42] Catherine Plaisant. 2004. The challenge of information visualization evaluation. In *Proceedings of the Working Conference on Advanced Visual Interfaces* (Gallipoli, Italy) *(AVI '04)*. Association for Computing Machinery, New York, NY, USA, 109–116. doi:10.1145/989863.989880
[43] Ronald Poppe, Rutger Rienks, and Betsy van Dijk. 2007. Evaluating the future of HCI: challenges for the evaluation of emerging applications. In *Artificial Intelligence for Human Computing: ICMI 2006 and IJCAI 2007 International Workshops, Banff, Canada, November 3, 2006, Hyderabad, India, January 6, 2007, Revised Selected and Invited Papers*. Springer, 234–250. doi:10.1007/978-3-540-72348-6_12
[44] Ali Ebrahimi Pourasad and Walid Maalej. 2025. Does GenAI Make Usability Testing Obsolete?. In *2025 IEEE/ACM 47th International Conference on Software Engineering (ICSE)*. IEEE, 437–449. doi:10.1109/ICSE55347.2025.00138
[45] Jenny Preece, Yvonne Rogers, Helen Sharp, David Benyon, Simon Holland, and Tom Carey. 1994. *Human-Computer Interaction*. Addison-Wesley Longman Ltd., GBR.
[46] Anna Ravera and Cristina Gena. 2025. On the usability of generative AI: Human generative AI. *arXiv preprint arXiv:2502.17714* (2025). doi:10.48550/arXiv.2502.17714
[47] Jude Rayan, Dhruv Kanetkar, Yifan Gong, Yuewen Yang, Srishti Palani, Haijun Xia, and Steven P. Dow. 2024. Exploring the Potential for Generative AI-based Conversational Cues for Real-Time Collaborative Ideation. In *Proceedings of the 16th Conference on Creativity & Cognition* (Chicago, IL, USA) *(C&C '24)*. Association for Computing Machinery, New York, NY, USA, 117–131. doi:10.1145/3635636.3656184
[48] Laurel D. Riek. 2012. Wizard of Oz studies in HRI: a systematic review and new reporting guidelines. *J. Hum.-Robot Interact.* 1, 1 (July 2012), 119–136. doi:10.5898/JHRI.1.1.Riek
[49] Alexander Rind. 2011. Some Whys and Hows of Experiments in Human–Computer Interaction. *Foundations and Trends® in Human-Computer Interaction* 5 (Jan. 2011), 299–373. doi:10.1561/1100000043
[50] Yvonne Rogers and Paul Marshall. 2017. Case Studies: Designing and Evaluating Technologies for Use in the Wild. In *Research in the Wild*, Yvonne Rogers and Paul Marshall (Eds.). Springer International Publishing, Cham, 33–67. doi:10.1007/978-3-031-02220-3_4
[51] Yvonne Rogers, Paul Marshall, and John M. Carroll. 2017. *Research in the Wild*. Morgan & Claypool Publishers.
[52] Yao Rong, Tobias Leemann, Thai-Trang Nguyen, Lisa Fiedler, Peizhu Qian, Vaibhav Unhelkar, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci. 2024. Towards Human-Centered Explainable AI: A Survey of User Studies for Model Explanations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 4 (April 2024), 2104–2122. doi:10.1109/TPAMI.2023.3331846
[53] Albrecht Schmidt, Florian Alt, and Ville Mäkelä. 2021. Evaluation in Human-Computer Interaction – Beyond Lab Studies. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 142, 4 pages. doi:10.1145/3411763.3445022
[54] Michael Sedlmair, Miriah Meyer, and Tamara Munzner. 2012. Design study methodology: Reflections from the trenches and the stacks. *IEEE transactions on visualization and computer graphics* 18, 12 (2012), 2431–2440. doi:10.1109/TVCG.2012.213
[55] Jingyu Shi, Rahul Jain, Hyungjun Doh, Ryo Suzuki, and Karthik Ramani. 2023. An HCI-centric survey and taxonomy of human-generative-AI interactions. *arXiv preprint arXiv:2310.07127* (2023). doi:10.48550/arXiv.2310.07127
[56] Hassan Silkhi, Brahim Bakkas, and Khalid Housni. 2024. Comparative Analysis of Rule-Based Chatbot Development Tools for Education Orientation: A RAD Approach. In *Proceedings of the 7th International Conference on Networking, Intelligent Systems and Security* (Meknes, AA, Morocco) *(NISS '24)*. Association for Computing Machinery, New York, NY, USA, Article 51, 7 pages. doi:10.1145/3659677.3659825
[57] Auste Simkute, Lev Tankelevitch, Viktor Kewenig, Ava Elizabeth Scott, Abigail Sellen, and Sean Rintel. 2025. Ironies of generative AI: understanding and mitigating productivity loss in Human-AI interaction. *International Journal of Human–Computer Interaction* 41, 5 (2025), 2898–2919. doi:10.1080/10447318.2024.2405782
[58] Hannu Simonen, Atte Kiviniemi, and Jonas Oppenlaender. 2025. An Initial Exploration of Default Images in Text-to-Image Generation. *arXiv preprint arXiv:2505.09166* (2025). doi:10.48550/arXiv.2505.09166
[59] Sonali Uttam Singh and Akbar Siami Namin. 2025. A survey on chatbots and large language models: Testing and evaluation techniques. *Natural Language Processing Journal* (2025), 100128. doi:10.1016/j.nlp.2025.100128
[60] Aaron Steinfeld, Odest Chadwicke Jenkins, and Brian Scassellati. 2009. The oz of wizard: simulating the human for interaction research. In *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction* (La Jolla, California, USA) *(HRI '09)*. Association for Computing Machinery, New York, NY, USA, 101–108. doi:10.1145/1514095.1514115
[61] Brodrick Stigall and Kelly Caine. 2020. A systematic review of human factors literature about voice user interfaces and older adults. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 64. SAGE Publications Sage CA: Los Angeles, CA, 13–17. doi:10.1177/1071181320641004
[62] Elias Storms, Oscar Alvarado, and Luciana Monteiro-Krebs. 2022. 'Transparency is Meant for Control' and Vice Versa: Learning from Co-designing and Evaluating Algorithmic News Recommenders. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 405 (Nov. 2022), 24 pages. doi:10.1145/3555130
[63] Jiajia Su and Zhongjun He. 2024. Enhancing User Experience Evaluation of Graphic Art Style Games through Collaboration with Generative AI. In *Proceedings of the 2024 5th International Conference on Computer Science and Management Technology (ICCSMT '24)*. Association for Computing Machinery, New York, NY, USA, 31–38. doi:10.1145/3708036.3708042
[64] Tee Hean Tan. 2025. Rule-Based vs. AI-Driven: Comparing PolyAQG Framework and Generative AI Models. In *Proceedings of the 2024 8th International Conference on Natural Language Processing and Information Retrieval (NLPIR '24)*. Association for Computing Machinery, New York, NY, USA, 298–303. doi:10.1145/3711542.3711583
[65] Kashyap Todi, Gilles Bailly, Luis Leiva, and Antti Oulasvirta. 2021. Adapting User Interfaces with Model-based Reinforcement Learning. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 573, 13 pages. doi:10.1145/3411764.3445497
[66] Koen van Turnhout, Arthur Bennis, Sabine Craenmehr, Robert Holwerda, Marjolein Jacobs, Ralph Niels, Lambert Zaad, Stijn Hoppenbrouwers, Dick Lenior, and René Bakker. 2014. Design patterns for mixed-method research in HCI. In *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational* (Helsinki, Finland) *(NordiCHI '14)*. Association for Computing Machinery, New York, NY, USA, 361–370. doi:10.1145/2639189.2639220
[67] Sarah Theres Völkel, Christina Schneegass, Malin Eiband, and Daniel Buschek. 2020. What is "intelligent" in intelligent user interfaces? a meta-analysis of 25

years of IUI. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 477–487. doi:10.1145/3377325.3377500

[68] C. G. Wolf, J. M. Carroll, T. K. Landauer, B. E. John, and J. Whiteside. 1989. The role of laboratory experiments in HCI: help, hindrance, or ho-hum?. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '89)*. Association for Computing Machinery, New York, NY, USA, 265–268. doi:10.1145/67449.67500

[69] Alan Woolrych, Kasper Hornbæk, Erik Frøkjær, and Gilbert Cockton. 2011. Ingredients and meals rather than recipes: A proposal for research that does not treat usability evaluation methods as indivisible wholes. *International Journal of Human-Computer Interaction* 27, 10 (2011), 940–970. doi:10.1080/10447318.2011.555314

[70] Lixiang Yan, Vanessa Echeverria, Gloria Milena Fernandez-Nieto, Yueqiao Jin, Zachari Swiecki, Linxuan Zhao, Dragan Gašević, and Roberto Martinez-Maldonado. 2024. Human-AI Collaboration in Thematic Analysis using ChatGPT: A User Study and Design Recommendations. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 191, 7 pages. doi:10.1145/3613905.3650732

[71] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3290605.3300509

[72] Bhada Yun, Dana Feng, Ace S. Chen, Afshin Nikzad, and Niloufar Salehi. 2025. Generative AI in Knowledge Work: Design Implications for Data Navigation and Decision-Making. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 634, 19 pages. doi:10.1145/3706598.3713337

[73] Sojeong Yun and Youn-kyung Lim. 2025. User Experience with LLM-powered Conversational Recommendation Systems: A Case of Music Recommendation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 898, 15 pages. doi:10.1145/3706598.3713347

[74] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 437, 21 pages. doi:10.1145/3544548.3581388

[75] Qingxiao Zheng, Minrui Chen, Pranav Sharma, Yiliu Tang, Mehul Oswal, Yiren Liu, and Yun Huang. 2025. EvAlignUX: Advancing UX Evaluation through LLM-Supported Metrics Exploration. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 1051, 25 pages. doi:10.1145/3706598.3714045

A Appendix

A.1 Initial Stage of Analysis

The initial collaborative workspace was used in the first round of analysis. In this phase, we brainstormed and collected methodological observations from four user studies, organizing them according to study phases (e.g., research planning, prototyping, data collection, and analysis). This initial mapping served as an analytic scaffold, helping us identify where and when methodological issues emerged and supporting early sensemaking around recurring patterns (see Figure 3).

A.2 Overview of Case Studies

This appendix provides an overview of the four lab-based user studies analyzed in our multi-case reflection. Each case represents a distinct evaluation context for GenAI systems, ranging from early concept exploration to studies involving fully functional prototypes. The studies vary in domain, prototype fidelity, and participant group, collectively illustrating recurring methodological challenges encountered when evaluating GenAI systems in controlled lab settings.

A.2.1 Case Study A: An LLM-Based Conversational Car Assistant.

Study Objective. This study investigated how users interact with an LLM-based conversational car assistant during various driving-related tasks. We examined conversation flow (single-turn versus multi-turn), language style (command-based versus natural language), task completion, and recovery from system errors. Additionally, we assessed distraction levels during interaction and overall usability. We were particularly interested in how users responded when the system could not handle certain requests that were not yet implemented in the prototype.

Study Procedure. We recruited 30 participants for our study, which was conducted in a standing vehicle with participants seated in the driver's seat. Participants were given structured tasks related to driving and car controls, each followed by a post-task interview and evaluation. The tasks included navigating to a destination with a stop along the route, controlling the windows and lights via speech, asking about car functionalities typically covered in the car manual, and engaging in free conversation on a topic of their choice. These tasks were introduced in a way that guided participants while still allowing them as much freedom as possible, including the option to go beyond the system's limits. Additionally, participants completed the Critical Tracking Task [41] on a screen positioned in front of the car to simulate driving-equivalent cognitive load for half of the task duration.

Data Collection and Analysis. We collected qualitative data through post-task interviews, as well as interviews conducted at the beginning and end of each session. Quantitative measures included the deviation in the Critical Tracking Task, feedback gathered using Likert-scale items from the UEQ, and system errors recorded through observations made by the study team. Moreover, we logged user utterances, system replies, and the number of conversational turns.

A.2.2 Case Study B: An LLM-Based Conversational Navigation Assistant Referencing the Display.

Study Objective. In this case, we investigated user interaction patterns in a multimodal LLM-based VUI within the automotive navigation context. We used a system that integrates GPT-4's multimodal capabilities with screenshots of the car's central display. Participants could ask questions about visual elements on the map or other GUI components shown on the central display. Using this system in a stationary vehicle, we conducted a user study with 21 participants. Through structured tasks and post-task questionnaires, we analyzed how users verbally described visual elements, including map features and interface icons, and assessed system usability using the SUS. We created a taxonomy that categorizes the linguistic structures participants used when referencing visual elements through the VUI.

Study Procedure. All participants interacted with the multimodal LLM-based system via speech to complete three tasks while seated in the driver's seat of a stationary vehicle. The three tasks varied in goals and interaction complexity. The first involved searching for charging stations within the navigation system. The second required participants to reference a lake shown on the map. The final task was an open exploration activity, where participants

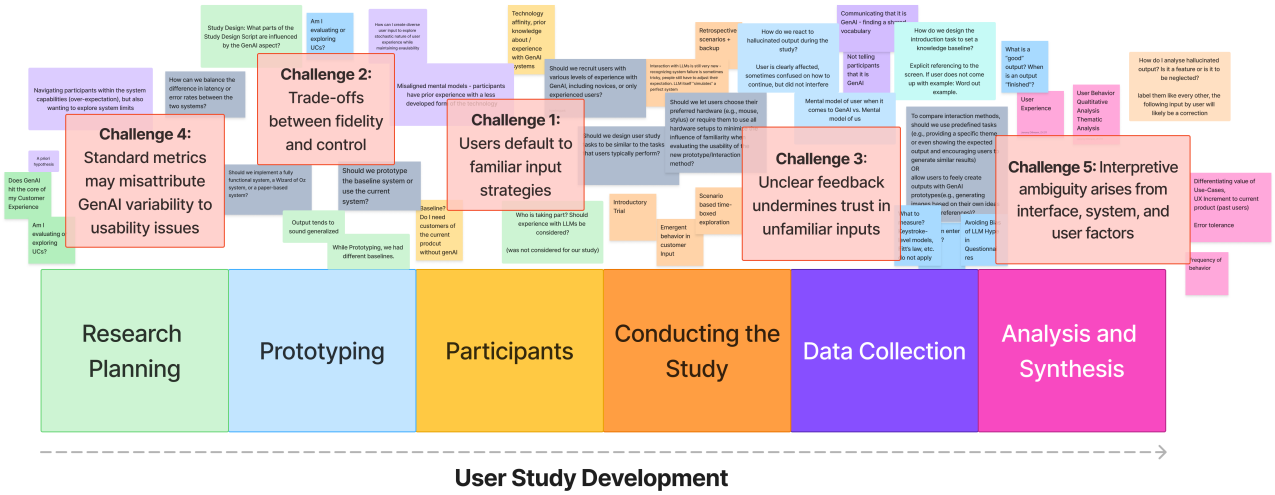


Figure 3: Overview of the initial collaborative workspace used during an early stage of analysis. Methodological observations were externalized and provisionally grouped to support shared sensemaking and explore emerging patterns. These materials were iteratively added, merged, and reorganized to explore potential subthemes and overarching patterns, informed by principles of thematic analysis [6]. Challenge labels (C1–C5) were added to make the connection between early affinity notes and the resulting challenges explicit.

could ask questions about the elements displayed on the interface. This procedure allowed us to examine how users structured spatial references across different interaction types.

Data Collection and Analysis. We collected transcripts of user–LLM interactions along with the contextual images provided to the model. Post-task interviews yielded both quantitative usability assessments via the SUS questionnaire and qualitative insights through semi-structured discussions. To analyze user behavior, we identified all utterances containing spatial references and applied thematic analysis [5], resulting in a taxonomy of reference types and interaction patterns.

Key Differences from Case Study A. This case builds on the conversational in-car assistant studied in Case Study A by focusing on how users reference visual elements in the GUI. This focus allowed us to examine users' referencing strategies and how successful references affect usability. By narrowing the interaction context, this study highlights challenges that arise specifically when conversational input is grounded in shared visual representations.

A.2.3 Case Study C: A Paper-Based Exploration of Visual Input Methods for GenAI Image Tools.

Study Objectives. This study explored alternative interaction methods for GenAI image tools, which are typically centered on text prompts. To support more visual forms of input, we introduced scribble- and annotation-based interaction techniques that allow users to draw, handwrite, or annotate visual elements directly onto images, aligning with common design practices.

Prototype Design and Development. We designed a hybrid input interface combining text prompts with scribbles and annotations. The front end was implemented using JavaScript and React and connected to several GenAI backends (Stable Diffusion, DALL·E 2, and GPT-4o). Early testing revealed latency and recognition errors in freehand input, which hindered the isolation of interaction behavior from system performance. We therefore adopted a paper-based prototype to simulate these interactions, enabling controlled evaluation of input strategies without interference from model variability.

Study Procedure. A qualitative study was conducted with seven professional designers. Each participant compared three input methods (text prompts, scribbles, and annotations) across six design tasks. Five tasks focused on predefined refinement categories (adding objects, increasing complexity, making global changes, adjusting layout, and modifying texture), and one was open-ended. Participants used pen or keyboard input freely. A think-aloud protocol and post-task interviews were used to capture reasoning and reflections.

Data Collection and Analysis. Qualitative data were collected from three sources: think-aloud protocols, post-task interviews, and user-generated artifacts (e.g., annotated sketches and text prompts). Think-aloud sessions captured participants' real-time prompting strategies, while interviews provided retrospective reflections on usability, preferences, and the perceived value of each input method. No structured questionnaires or quantitative metrics (e.g., Likert scales or task completion times) were used. Instead, the analysis was grounded in inductive thematic analysis [6], focusing on patterns in behavior, input preferences, and recurring challenges across participants.

A.2.4 Case Study D: A Functional Prototype User Study for a GenAI Image Generation Tool.

Study Objectives. Building on the findings from the earlier paper-based study (Case Study C), this follow-up study evaluated user interactions with a fully functional prototype of the GenAI image tool. While the previous study focused on input preferences in a static, controlled setup, this study observed how designers interacted with the system in real-time. The goal was to understand how real-time generative feedback affected user strategies, tool preferences, and iterative design behavior, as well as whether previously reported preferences for visual input methods persisted under dynamic feedback conditions.

Study Procedure. This study expanded on the earlier paper-based work by introducing an interactive system that generates images in real time in response to user input. While Case Study C employed a low-fidelity prototype to investigate preferences in a fully controlled setting, Case Study D introduced a fully functional system with real-time, generative output, which enabled observation of how designers adapted their strategies dynamically to the system's stochastic behavior. Latency, error handling, and responsiveness—factors abstracted away in the previous study—became key aspects of this evaluation.

Data Collection and Analysis. Participants included professional designers and design students who performed image refinement and creation tasks using text, scribbles, and annotations as input modalities. We collected usage logs, screen recordings, and post-task interview data. Quantitative measures included input completion times, the NASA-TLX, the UEQ, and a custom survey assessing perceived intent alignment. Quantitative results were used descriptively to contextualize the findings, while qualitative data were analyzed to capture behavioral adaptation and iterative strategies during live system interaction.

Key Differences from Case Study C. Compared to Case Study C, this study examined real-time interaction with a fully functional GenAI system rather than simulated input. This enabled observation of how live feedback shaped iteration, trust, and engagement, offering insights into the usability and methodological implications of evaluating functional GenAI image tools in controlled settings.