Comenius University in Bratislava
Faculty of Mathematics, Physics and Informatics

From Prompts to Users: Validity and Limits of Synthetic Participants in User Experience Research

Diploma Thesis

2026
Bc. Laura Jirásková

Comenius University in Bratislava
Faculty of Mathematics, Physics and Informatics

From Prompts to Users: Validity and Limits of Synthetic Participants in User Experience Research

Diploma Thesis

Study Programme:	Cognitive Science
Field of Study:	Cognitive Science
Department:	Department of Applied Informatics
Supervisor:	RNDr. Kristína Malinovská, PhD.

Bratislava, 2026
Bc. Laura Jirásková

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Laura Jirásková
Študijný program: kognitívna veda (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: informatika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: anglický
Sekundárny jazyk: slovenský

Názov: From Prompts to Users: Validity and Limits of Synthetic Participants in User Experience Research
Od promptov k používateľom: validita a limity syntetických participantov v UX výskume

Anotácia: Veľké jazykové modely (VJM) čoraz viac umožňujú využívanie syntetických participantov - simulovaných používateľov, ktorí sú pomocou promptov vedení k poskytovaniu spätnej väzby v UX výskume namiesto reálnych ľudí. Prvé štúdie naznačujú, že takéto modely dokážu odhaliť značnú časť problémov s použiteľnosťou, avšak nie je jasné, do akej miery môže promptovanie posunúť ich správanie smerom k ľudskému hodnoteniu.

Cieľ: 1. Zmapovať súčasné empirické a teoretické práce o syntetických participantoch založených na VJM vo výskume UX so zameraním na to, ako promptovanie (zadanie úlohy, definícia persony, kontext) formuje ich spätnú väzbu.
2. Analyzovať, ktoré aspekty ľudskej kognície v takýchto simuláciách chýbajú vo vzťahu k rôznym typom UX úloh.
3. Navrhnúť rámec orientovaný na typ úlohy pre využitie syntetických participantov spolu so skutočnými používateľmi v UX štúdiách.

Literatúra:
Amirova, A. et al. (2024). Framework-based qualitative analysis of free responses of Large Language Models: Algorithmic fidelity. PLOS ONE, 19(3).
Argyle, L. P. et al. (2023). Out of One, Many: Using Language Models to Simulate Human Samples. Political Analysis, 31(3), 337–351.
Hsueh, N.-L., Lin, H.-J., & Lai, L.-C. (2024). Applying Large Language Model to User Experience Testing. Electronics, 13(23), 4633.

Vedúci: RNDr. Kristína Malinovská, PhD.
Katedra: FMFI.KAI - Katedra aplikovanej informatiky
Vedúci katedry: doc. RNDr. Tatiana Jajcayová, PhD.

Dátum zadania: 26.02.2025

Dátum schválenia: 03.03.2025

prof. Ing. Igor Farkaš, Dr.
garant študijného programu

.....
študent

.....
vedúci práce

Comenius University Bratislava
Faculty of Mathematics, Physics and Informatics

THESIS ASSIGNMENT

Name and Surname:	Laura Jirásková
Study programme:	Cognitive Science (Single degree study, master II. deg., full time form)
Field of Study:	Computer Science
Type of Thesis:	Diploma Thesis
Language of Thesis:	English
Secondary language:	Slovak

Title: From Prompts to Users: Validity and Limits of Synthetic Participants in User Experience Research

Annotation: Large language models (LLM) increasingly allow the use of synthetic participants - simulated users prompted to provide feedback in UX research instead of real people. Early studies suggest that such models can detect a considerable share of usability issues, yet it remains unclear how far prompting can push their behaviour towards human-like judgement. To bridge the gaps in research it is needed to outline a framework for understanding the conditions under which synthetic participants can complement real users.

Aim:
1. Map current empirical and theoretical work on LLM-based synthetic participants in UX research, with a focus on how prompting (task framing, personas, context) shapes their feedback.
2. Analyse which aspects of human cognition are missing in such simulations with relation to different UX task types.
3. Propose a task-type oriented framework for using synthetic participants together with real users in UX studies.

Literature:
Amirova, A. et al. (2024). Framework-based qualitative analysis of free responses of Large Language Models: Algorithmic fidelity. PLOS ONE, 19(3).
Argyle, L. P. et al. (2023). Out of One, Many: Using Language Models to Simulate Human Samples. Political Analysis, 31(3), 337–351.
Hsueh, N.-L., Lin, H.-J., & Lai, L.-C. (2024). Applying Large Language Model to User Experience Testing. Electronics, 13(23), 4633.

Supervisor:	RNDr. Kristína Malinovská, PhD.
Department:	FMFI.KAI - Department of Applied Informatics
Head of department:	doc. RNDr. Tatiana Jajcayová, PhD.
Assigned:	26.02.2025
Approved:	03.03.2025

prof. Ing. Igor Farkaš, Dr.
Guarantor of Study Programme

Student

Supervisor

Acknowledgments: I would like to thank RNDr. Kristína Malinovská, PhD. for her encouragement and guidance throughout the writing of this thesis. I would also like to thank my sister for always being there for me whenever I needed her most, and my boyfriend for the endless patience he showed me along the way.

Abstract

Large language models have made it possible to generate fluent, context-sensitive, and user-like feedback from detailed prompts. This has led to growing interest in synthetic participants: model-generated users used in UX research to simulate feedback before real participants are available. Although this approach is attractive because UX research is often limited by time, recruitment, and resources, its methodological validity remains unclear. A model can produce feedback that resembles human experience without actually having experience, emotion, risk, or situated context. In this thesis, I examine the validity and limits of synthetic participants in user experience research through a compilation-based theoretical approach. I synthesize literature from UX research, human-computer interaction, large language models, prompting research, synthetic participants, and cognitive science. I focus on how prompting, personas, task framing, and contextual information shape synthetic outputs, and on what human-synthetic correspondence can and cannot mean. I argue that synthetic participants are most useful for knowledge-oriented UX tasks, such as early usability inspection, content clarity, navigation analysis, heuristic evaluation, and comparison of design alternatives. They can also support some interpretation-oriented tasks, but only as complementary evidence. They are least suitable for tasks that depend on lived experience, emotion, trust, accessibility, adoption, or long-term use. Based on this analysis, I propose a task-oriented framework and a hybrid model in which synthetic participants support early exploration, while real users remain necessary for validation and experience-based findings.

Keywords: synthetic participants, user experience research, large language models, prompting, human-synthetic correspondence, UX methodology, cognitive limits

Abstrakt

Veľké jazykové modely dnes umožňujú generovať plynulú, kontextovo citlivú a používateľsky pôsobiacu spätnú väzbu na základe detailných promptov. To viedlo k rastúcemu záujmu o syntetických participantov: modelom generovaných používateľov využívaných v UX výskume na simulovanie spätnej väzby ešte pred zapojením reálnych participantov. Hoci je tento prístup prakticky atraktívny, pretože UX výskum je často obmedzený časom, náborom a zdrojmi, jeho metodologická validita zostáva nejasná. Model môže vytvoriť spätnú väzbu podobnú ľudskej skúsenosti bez toho, aby mal skúsenosť, emócie, riziko alebo situačný kontext. V tejto práci skúmam validitu a limity syntetických participantov vo výskume používateľskej skúsenosti prostredníctvom kompilačno-teoretického prístupu. Syntetizujem literatúru z oblasti UX výskumu, interakcie človeka s počítačom, veľkých jazykových modelov, promptovania, syntetických participantov a kognitívnej vedy. Zameriavam sa na to, ako promptovanie, persony, zadanie úlohy a kontextové informácie formujú syntetické výstupy, a čo môže alebo nemôže znamenať zhoda medzi ľudskými a syntetickými odpoveďami. Tvrdím, že syntetickí participanti sú najužitočnejší pri znalostne orientovaných UX úlohách, ako je skorá kontrola použiteľnosti, analýza jasnosti obsahu, navigácie, heu-ristická evaluácia alebo porovnanie dizajnových alternatív. Pri niektorých interpretačne orientovaných úlohách môžu slúžiť ako doplnkový dôkaz. Najmenej vhodní sú pri úlohách závislých od žitej skúsenosti, emócií, dôvery, prístupnosti, adopcie alebo dlhodobého používania. Na základe tejto analýzy navrhujem task-oriented framework a hybridný model, v ktorom syntetickí participanti podporujú skorú exploráciu, zatiaľ čo reálni používatelia zostávajú nevyhnutní na validáciu a zistenia založené na skúsenosti.

Kľúčové slová: syntetickí participanti, výskum používateľskej skúsenosti, veľké jazykové modely, promptovanie, ľudsko-syntetická zhoda, metodológia UX, kognitívne limity

Contents

List of Figures

List of Tables

Abbreviations

AI	Artificial Intelligence
GPT	Generative Pre-trained Transformer
HCI	Human-Computer Interaction
ISO	International Organization for Standardization
LLM(s)	Large Language Model(s)
NLP	Natural Language Processing
RLHF	Reinforcement Learning from Human Feedback
UI	User Interface
UX	User Experience
WEIRD	Western, Educated, Industrialized, Rich, Democratic

Chapter 1

Introduction

Understanding how people experience interactive systems has never been more practically urgent. Digital products now mediate an extraordinary range of everyday activities — from managing health and finances to maintaining social relationships and navigating public services. The quality of those interactions extends well beyond simple task completion. User experience research exists to make sense of this. It investigates whether people can use a system, how that use feels, what it means to them, and whether it fits into their lives in ways that are sustainable and satisfying.

The choice of this topic was also motivated by a practical sensitivity to everyday interaction problems in digital products. Many usability issues become visible in ordinary use as well as in formal research settings — when an application behaves unexpectedly, requires unnecessary effort, or fails to make an action intuitive. This personal interest in how existing systems can be made clearer, more usable, and more coherent led to a broader interest in user experience as a field. Prior experience with UX research was still limited at the time, so a theoretical thesis offered an appropriate way to first build a conceptual foundation: to understand what UX research investigates, what kinds of evidence it relies on, and only then to examine the newer question of whether synthetic participants can meaningfully contribute to this research.

Conducting UX research with real users is genuinely difficult. Recruiting participants, preparing studies, moderating sessions, and analysing qualitative data takes time and resources. In commercial design settings, this often conflicts with fast product cycles, agile development, and the need to make design decisions before a full user study can be conducted. These constraints are not only logistical. They shape the kind of evidence that UX research can produce. Lewis (1994) showed that sample size directly affects which usability problems a study is likely to detect and which problems remain unseen. A study with a small number of participants may be sufficient when problems are common, but it may miss less frequent issues that still matter for real users. Marcilly et al. (2024) similarly show that UX research often operates under prac-

tical compromise: the participants who would make testing most valid are frequently the hardest to recruit, especially when the target context is specialised or when project timelines are limited. Human-centred UX research is therefore not an ideal baseline against which synthetic participants can be judged without qualification. It is itself constrained by access, time, sampling, and research design.

Large language models have made it possible to generate fluent and context-sensitive text in response to detailed prompts. Current systems can adopt roles, follow instructions, summarize information, evaluate alternatives, and produce feedback that often resembles human-written responses. This has led to a growing interest in synthetic participants: model-generated users that are prompted to provide feedback in research or design contexts. Related discussions also appear under terms such as synthetic samples or silicon samples, where language models are used to approximate responses of specified human groups.

For UX research, this development is practically attractive. A synthetic participant can be generated immediately, assigned a persona, placed in a scenario, and asked to comment on an interface or complete a simulated task. Early empirical work suggests that such systems may identify visible usability problems, support early design evaluation, and help researchers prepare before human recruitment becomes available. Zhong et al. (2025b), for example, report that a GPT-4[1] based system identified between 73% and 77% of usability problems across two applications. Lu et al. (2025b) propose UX-Agent, a framework in which LLM-based agents equipped with personas and browser access simulate user interactions and provide early feedback before studies with real users are conducted. These results are difficult to dismiss as a novelty. They also make the evidential status of synthetic participants more urgent to examine.

These developments sit at the intersection of two broader shifts: the growing capability of large language models and the persistent practical constraints of human-centred UX research. Together, they create a genuine opening and a genuine question. This raises a further question about prompting itself: how far can task framing, personas, and contextual detail push model outputs toward human-like judgement, and where does such conditioning reach its limits? When a synthetic participant flags an interface element as confusing, it is not immediately clear what process produced that judgement, or whether anything analogous to actual use was involved at all. The output may be correct. Without knowing how it was reached, its evidential value remains uncertain. A model can produce a response that sounds like user feedback without being a user. It can describe confusion without being confused, express frustration without experiencing frustration, and comment on trust without being exposed to risk.

[1]GPT-4 (Generative Pre-trained Transformer 4): large-scale multimodal large language model developed by OpenAI, capable of processing both image and text inputs (GPT-4 Technical Report, OpenAI, 2023, `https://arxiv.org/abs/2303.08774`).

It is a deeper problem about what synthetic outputs represent. Better prompting will not necessarily fix it.

This thesis examines this tension. The question is not whether synthetic participants are simply valid or invalid. Such a question is too broad, because UX research itself is not a single type of inquiry. Some UX tasks concern explicit interface structure, information architecture, navigation, content clarity, or visible usability problems. Others concern interpretation, emotional response, accessibility, habit formation, situated use, or long-term experience. These tasks require different kinds of evidence. The suitability of synthetic participants therefore depends on what aspect of UX is being investigated.

The aim of this thesis is threefold. First, it synthesizes selected empirical and theoretical work on LLM-based synthetic participants in UX research, with particular attention to how prompting, task framing, personas, and context shape their feedback. Second, it analyses which aspects of human cognition and experience are missing or only weakly approximated in such simulations in relation to different UX task types. Third, it proposes a task-oriented framework for using synthetic participants together with real users in UX studies.

The central argument is that synthetic participants should not be assessed according to how human-like they appear as a whole. Their validity depends on whether the kind of evidence required by a particular UX task can be approximated through language-model output. This shifts the problem from a general question of replacement to a more specific question of task suitability.

1.1 Research Gap

Recent research on synthetic participants has shown that large language models can produce outputs that sometimes overlap with human responses or expert UX findings. Studies on silicon samples examine whether models can approximate responses of human subpopulations, while applied UX studies test whether LLMs can identify usability issues, generate evaluation reports, or simulate interaction traces.

Much of this work remains focused on performance and implementation. It asks whether a model can detect usability problems, how many of them overlap with human evaluators, or how prompting pipelines can make outputs more structured. These questions are important. They do not fully explain what kind of evidence a synthetic participant actually provides.

This is the main gap addressed in the present thesis. Human-synthetic correspondence is often treated as if its meaning were obvious. A model may identify the same problem as a human participant because the issue is structurally visible, already known

from usability principles, or implicitly suggested by the prompt. In such cases, the model reproduces patterns of user-like language learned from training data. Simulation, in the stronger sense, may not be occurring at all. This becomes especially problematic where the UX phenomenon depends on embodiment, affective involvement, or lived experience. Such phenomena cannot be approached through explicit reasoning about an interface alone.

A second gap concerns prompting. Prompt design is often discussed as a way of improving output quality. In UX research, it also functions as part of the research instrument. The prompt defines the task, the user position, the context, and the type of response the model is expected to generate. The relevant question therefore shifts from how to prompt better to which prompt components are methodologically necessary.

Finally, existing work rarely connects the limits of LLM simulation to different types of UX tasks. Some UX tasks depend mainly on explicit reasoning and design knowledge. Others require situated context, affect, embodiment, or long-term experience. The present thesis addresses this gap by shifting the discussion from the general question of whether synthetic participants can replace users to the more precise question of when, and for which UX tasks, their outputs can be treated as useful evidence.

1.2 Research Questions

Synthetic participants require examination from both a methodological and a cognitive perspective. The present thesis therefore asks how synthetic participant outputs are shaped, when they can be useful in UX research, what their correspondence with human responses means, and where prompting reaches its limits.

The thesis is guided by the following research questions:

RQ1: Which characteristics of a prompt are necessary for synthetic outputs to be methodologically usable in UX research?

RQ2: Under which conditions and for which types of UX tasks can synthetic participants serve as a practical substitute for human participants in UX research?

RQ2.1: When a synthetic output corresponds to human responses, what does that correspondence actually mean epistemologically — is it a meaningful simulation?

RQ3: Where are the cognitive and methodological limits of prompting — which types of human reasoning, context, or experience cannot be reliably simulated even with detailed prompts?

1.3 Methodology and Thesis Structure

To address these aims and research questions, this thesis follows a compilation-based, theoretical approach. It does not include a new empirical study with human or synthetic participants. The argument is developed through a structured review and conceptual analysis of existing empirical and theoretical work from user experience research, human-computer interaction, large language models, prompting research, synthetic participants, and cognitive science.

This choice follows from the nature of the research questions. RQ1 concerns the prompt characteristics that make synthetic outputs methodologically usable in UX research, and is therefore addressed through a synthesis of existing work on prompting and synthetic participant design. RQ2 and RQ2.1 concern the conditions under which synthetic participants may substitute for human participants and what human-synthetic correspondence means epistemologically. These questions cannot be answered by another detection-rate comparison alone. Such a study could show whether outputs overlap. It could not show what that overlap represents. RQ3 concerns the cognitive and methodological limits of prompting, which requires conceptual analysis of what cannot be reliably reconstructed through language-model output.

Sources were identified through targeted searches in Google Scholar and the ACM Digital Library. Search terms included synthetic participants, LLM-based UX evaluation, silicon samples, prompt design, validity, and cognitive limits of simulation. Selection prioritised three groups of sources: empirical studies that directly examine LLM-based synthetic participants in UX or adjacent simulation contexts; theoretical work on the capabilities and limitations of large language models; and UX and HCI literature on user experience, usability methods, self-report, embodiment, temporality, and context of use. Schematic diagrams created by the author were prepared in Adobe Illustrator.

The structure of the thesis follows this argument. Chapter 1 establishes the UX research background needed for the later analysis. It discusses the development from usability to UX, selected UX definitions and models, and common UX research methods. Chapter 2 focuses on synthetic participants. It explains their rationale, the relevant properties and limits of large language models, prompting approaches, and empirical evidence from UX-related studies. Chapter 3 brings these areas together. It discusses the components of an effective UX prompt, the meaning of correspondence between human and synthetic outputs, the limits of prompting, and proposes a task-oriented framework for using synthetic participants in UX research. Chapter 4 closes the thesis by summarizing the main findings, limitations, and possible directions for future research.

Chapter 2

User Experience Research

User experience research studies how people interact with, interpret, and evaluate digital systems. The field developed from usability research. Its scope has gradually expanded beyond task efficiency, error rates, and learnability to include subjective, emotional, contextual, and temporal aspects of interaction (Hassenzahl and Tractinsky, 2006).

The shift matters because synthetic participants are not evaluated against a single type of user response. Different UX tasks require different kinds of evidence. Some tasks can be approached through explicit reasoning about interface structure; others depend on situated use, personal experience, or affective response. Clarifying what UX research investigates and how it obtains evidence from human participants is therefore a necessary step before asking whether such tasks can be simulated by large language models.

This chapter first outlines the transition from usability to user experience. It then introduces selected definitions, models, and dimensions of UX, followed by an overview of common UX research methods and their limits.

2.1 From Usability to User Experience

User experience emerged as an expansion of what it means to evaluate human interaction with technology. It did not replace usability. Earlier approaches to human-computer interaction were primarily concerned with whether users could operate a system effectively, efficiently, and without excessive effort. Interface quality was assessed through task completion, error rates, time on task, learnability, and satisfaction (Hornbæk, 2006). These criteria remain relevant because they capture whether a system supports users in achieving their goals.

ISO 9241-11[1] defines usability as the extent to which specified users can use a system

[1]Foundational international guideline for usability in human-system interaction.

to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use (as cited in Hornbæk, 2006). The definition connects usability to users, goals, and context rather than treating it as an isolated property of an interface, though its focus remains task-oriented. A system is usable when people can accomplish what they need to accomplish under given conditions.

Usability research is therefore closely tied to observable performance. A researcher may observe whether users complete a checkout flow, find relevant information, make errors, or understand the structure of a website. These dimensions lend themselves to systematic measurement and comparison across participants or design alternatives (Hornbæk, 2006), which has given usability a stable methodological foundation in the evaluation of interactive systems.

The expansion toward user experience was driven by the recognition that successful interaction cannot be reduced to task performance alone. Users may complete a task and still find the system unpleasant, untrustworthy, stressful, or irrelevant. A system with minor usability inefficiencies may still produce a positive experience because it is aesthetically engaging, meaningful, or emotionally satisfying (Hassenzahl and Tractinsky, 2006; Tractinsky et al., 2000). This became increasingly apparent as digital systems moved beyond workplace tools and into communication, entertainment, healthcare, finance, and social life. In such contexts, users do not merely execute tasks. They form impressions, develop expectations, experience emotions, negotiate trust, and interpret what a system means for their goals and values (Hassenzahl and Tractinsky, 2006).

Several theoretical contributions shaped this shift. Norman's (2004) work on emotional design, Hassenzahl's (2005) distinction between pragmatic and hedonic qualities, and McCarthy and Wright's (2004) emphasis on experience each approached the question from a different angle. All shared the assumption that interaction with technology includes more than task execution (Hassenzahl and Tractinsky, 2006).

Pragmatic qualities concern whether a system supports users in achieving their goals and include clarity, efficiency, controllability, and learnability. Experiential qualities concern how interaction is felt, interpreted, and situated within a personal or social context, and include enjoyment, stimulation, identification, aesthetic appreciation, and perceived meaningfulness (Hassenzahl, 2005), as well as emotional responses such as frustration (Hassenzahl and Tractinsky, 2006). Usability and UX are not opposites. Poor usability can seriously damage the overall experience, particularly when users cannot complete tasks that matter to them. Usability alone, however, does not explain UX. A usable system is not automatically experienced as valuable, trustworthy, or desirable (Hassenzahl, 2008).

The distinction has direct methodological consequences. If UX research were limited to usability, evaluating synthetic participants would be considerably simpler: their

usefulness could be assessed mainly by comparing whether they identify the same interface problems as human participants. UX also includes interpretation, emotion, context, and lived experience. A synthetic participant may identify a usability issue without reproducing the experiential conditions under which a human user encounters it.

Usability-oriented tasks often rely on explicit knowledge, design conventions, and analytical reasoning, which are the kinds of tasks for which language models may generate useful outputs. Broader UX tasks may depend on situated behaviour, affective response, tacit knowledge, embodiment, or long-term experience, none of which are reducible to explicit verbal knowledge and therefore difficult to approximate through prompting. This helps explain the pattern in existing research. Synthetic participants appear adequate for tasks concerning identifiable structural issues. For tasks concerning how users experience or emotionally respond to a system in a concrete situation, their adequacy is less certain.

Usability is treated in this thesis as a necessary component of user experience. It provides a clear methodological foundation for evaluating interaction. It does not exhaust what UX research investigates. This matters for later chapters, where the suitability of synthetic participants is assessed by the type of task and the kind of evidence it requires.

2.2 Definition and Models

Defining user experience is difficult because UX does not refer to a single property of a system. It refers to the relation between a person, a product, a task, and a context of use. Unlike usability, which can often be assessed through relatively stable performance indicators, UX includes subjective and situational aspects that are more difficult to isolate. This is one reason why the term has been used in different ways across human-computer interaction, design research, psychology, and industry practice (Law et al., 2009).

The difficulty is not merely terminological. UX draws together concepts such as affect, aesthetics, hedonic quality, meaning, expectation, satisfaction, and long-term attachment. Law et al. (2009) show that UX is broadly understood as dynamic, context-dependent, and subjective, but that there is less agreement on its exact boundaries. The same concept may refer to a brief moment of interaction, a single episode of product use, or a long-term relationship between a person and a product. This breadth is one of the reasons why UX is theoretically useful. It is also why UX is methodologically difficult. If UX includes pragmatic performance, emotional response, interpretation, context, and temporality, then no single measure can capture it completely.

A commonly cited definition is provided by ISO 9241-210[2], according to which user experience refers to a person's perceptions and responses that result from the use or anticipated use of a product, system, or service (as cited in Law et al., 2009). This definition avoids reducing UX to actual interaction alone. Experience may begin before use, through expectations, reputation, visual appearance, or prior knowledge, and it may continue after use, through memory, evaluation, and changed attitudes toward the product (Law et al., 2009). Mirnig et al. (2015) similarly note that the ISO definition situates UX across anticipation, active use, and retrospective reflection, and includes the influence of system characteristics, user characteristics, and context of use.

This breadth is also a limitation. A definition that includes perceptions, responses, emotions, beliefs, preferences, physical reactions, psychological reactions, and context is valuable as a reminder of scope. It does not, by itself, tell a researcher what should be measured or prioritized in a specific study (Hassenzahl, 2008; Mirnig et al., 2015). This matters for the present thesis because synthetic participants cannot be evaluated against UX as a general and undifferentiated whole. Their usefulness depends on which part of UX is being approximated: explicit judgement about interface structure, subjective interpretation, emotional response, situated behaviour, or longer-term experience.

Hassenzahl and Tractinsky's facets of UX. A theoretically useful account is provided by Hassenzahl and Tractinsky (2006), who describe UX as emerging from the user's internal state, the characteristics of the designed system, and the context in which interaction occurs. This formulation is important because it locates UX at the intersection of person, product, and situation rather than in any one of them alone. Their account also organizes UX around three broad perspectives: the move beyond the instrumental, the role of emotion and affect, and the situated and experiential character of technology use (Hassenzahl and Tractinsky, 2006). As shown in Figure 2.1, UX is therefore not reducible to task performance alone. A language model may be able to reason about the visible structure of a product. It does not possess a user's internal state. It does not occupy the context of use. It usually encounters the system only through textual or visual representation.

[2]Foundational international standard for human-centred design of interactive systems.

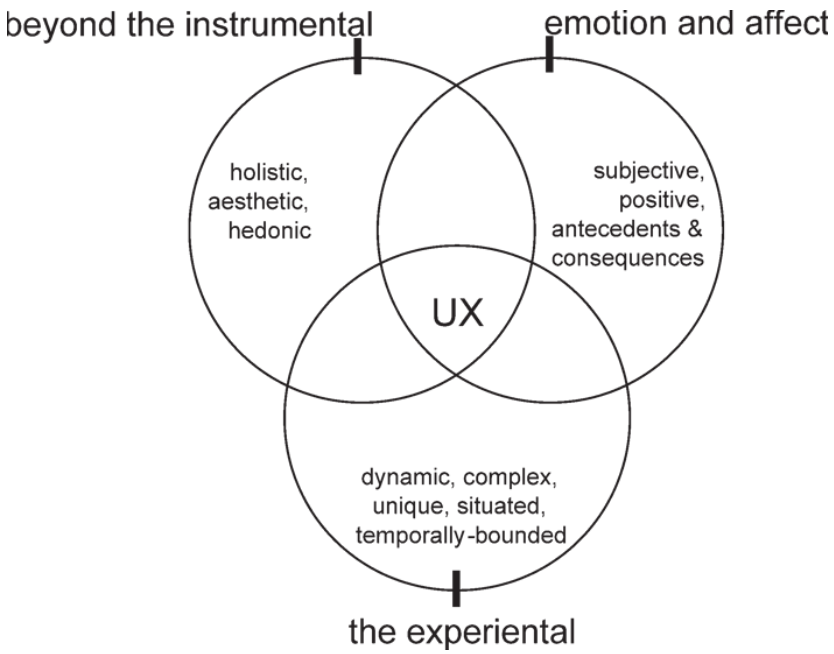

Figure 2.1: Facets of UX according to Hassenzahl and Tractinsky (2006).

Pragmatic and hedonic qualities. Several models of UX attempt to organize these dimensions more precisely. Hassenzahl's model is particularly influential because it distinguishes between pragmatic and hedonic qualities (Hassenzahl, 2005). Pragmatic qualities refer to the extent to which a product supports users in achieving their goals and include usefulness, clarity, controllability, efficiency, and learnability. Hedonic qualities refer to stimulation, identification, novelty, and the product's ability to support psychological needs beyond task completion (Hassenzahl, 2005; Hassenzahl and Tractinsky, 2006). As shown in Figure 2.2, Hassenzahl's model distinguishes between two perspectives on product character (Hassenzahl, 2005). From the designer's perspective, product features are combined to convey an intended character: pragmatic attributes supporting task completion, and hedonic attributes addressing needs beyond the task. From the user's perspective, this intended character is reconstructed as an apparent product character, shaped by the user's own standards and expectations. The consequences of interaction — appeal, pleasure, and satisfaction — depend not only on the product's attributes but also on the specific usage situation, which means that the same product may be experienced differently depending on context (Hassenzahl, 2005).

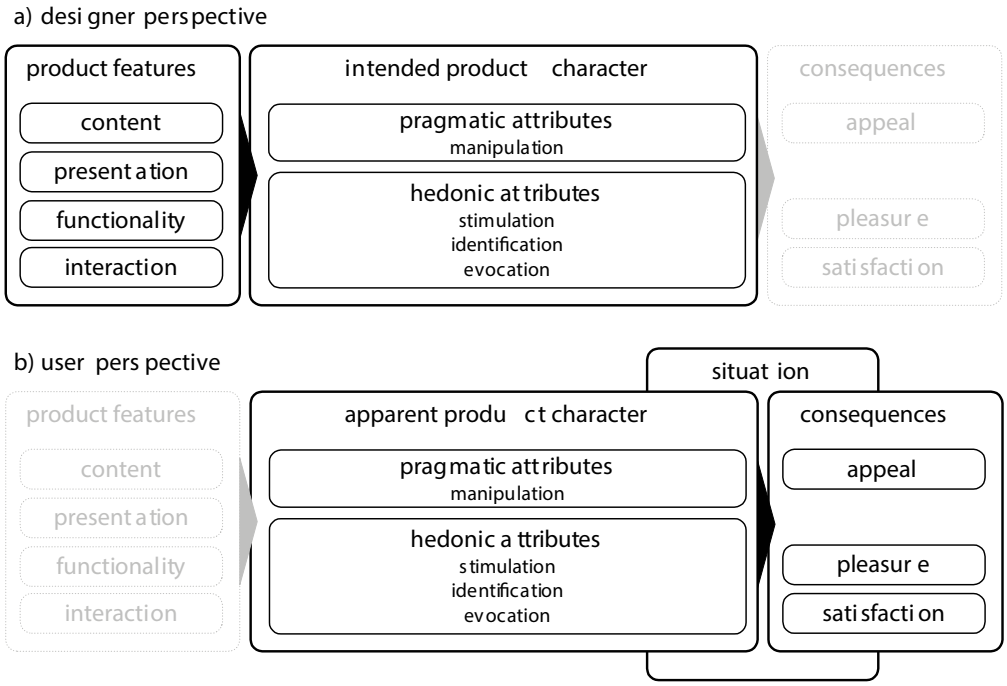

Figure 2.2: Hassenzahl's model of user experience from (a) a designer and (b) a user perspective (Hassenzahl, 2005).

Hassenzahl (2008) further develops this distinction through the concepts of do-goals and be-goals. Do-goals are concrete behavioural aims, such as finding information, completing a purchase, sending a message, or transferring money in a banking application. Be-goals are broader psychological needs, such as autonomy, competence, relatedness, personal growth, identity, or self-expression (Hassenzahl, 2008). Pragmatic quality is mainly connected to whether the product enables users to fulfil do-goals. Hedonic quality is connected to whether the interaction supports be-goals and contributes to a meaningful or positive experience.

This distinction has direct consequences for synthetic participant research. UX tasks oriented toward do-goals are often more explicit and structurally bounded. A synthetic participant may be able to identify that a navigation label is unclear, that a form contains unnecessary steps, or that the order of information in a checkout flow is confusing. These judgements rely partly on explicit reasoning about interface structure and design conventions. UX tasks oriented toward be-goals are more closely tied to subjective meaning, self-image, motivation, emotion, and prior experience. A model can generate a plausible statement about feeling confident, frustrated, identified with a product, or emotionally attached to a service. It does not undergo the psychological conditions that make such experiences meaningful for a human user.

Pragmatic and hedonic qualities are not separate in actual experience. A product that is difficult to use may prevent users from reaching deeper forms of satisfaction.

A product that functions efficiently may still fail to feel meaningful, trustworthy, or desirable (Hassenzahl, 2008). Users also do not evaluate interactive products as a simple sum of independent properties. Aesthetic perception can influence perceived usability, and perceived usability can shape the overall aesthetic and experiential judgement of a product. Tractinsky et al. (2000), as cited in Hornbæk and Hertzum (2017), showed that visually appealing interfaces may be judged as more usable, a finding often summarized as "what is beautiful is usable." Later work suggests that this relationship is conditional and depends on context and method. The general point remains relevant: UX judgements are integrated rather than purely mechanical (Hornbæk and Hertzum, 2017).

Experience-centred HCI. A second relevant approach is found in experience-centred HCI, especially in the work of McCarthy and Wright (2004). Their account treats experience not as a measurable after-effect of interaction, but as something lived through by a person. They emphasize sense-making, emotion, embodiment, and the continuity between technology use and everyday life. This view shifts attention away from whether an interface performs well in isolation and toward how interaction is woven into personal and social practices.

This perspective exposes the limits of overly instrumental models of UX. A banking application, a fitness tracker, or a medical platform cannot be evaluated only by asking whether users complete a task. Such systems may involve anxiety, trust, risk, identity, responsibility, or long-term behavioural change. These aspects are not simply additional variables attached to usability. They may fundamentally shape how the system is experienced.

Temporal and contextual dimensions of UX. Temporal models of UX further complicate the picture. UX is not a fixed property of a product that can be assessed once and treated as stable. Karapanos et al. (2009) argue that experience changes across phases of interaction, from initial orientation and excitement to incorporation into everyday routines. A first encounter with a product may be shaped by novelty, visual impression, curiosity, or uncertainty. Later experience may depend more on reliability, habit, accumulated frustration, emotional attachment, or the degree to which the product becomes integrated into daily life (Karapanos et al., 2009). Bargas-Avila and Hornbæk (2011) note that many UX studies focus on short-term encounters, even though actual product experience often unfolds over longer periods.

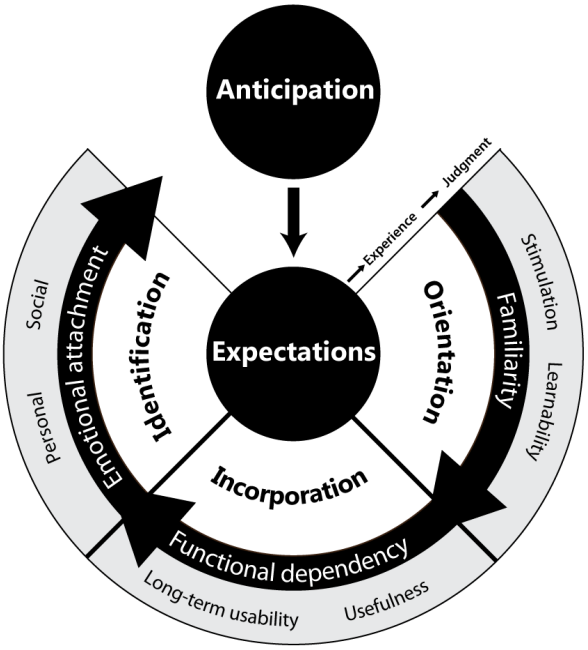

Figure 2.3: Temporality of experience according to Karapanos et al. (2009).

As shown in Figure 2.3, UX develops across phases of orientation, incorporation, and identification. This model is important because it shows that user experience cannot be reduced to a single interaction episode. A synthetic participant can describe long-term use. It does not pass through the temporal process through which familiarity, dependency, and attachment develop.

The temporal character of UX has direct consequences for the analysis developed in later chapters. A language model can generate a plausible statement about long-term use, habit formation, frustration, trust, or attachment. It has not accumulated repeated encounters with the system. It has not adapted to the product over time, relied on it in changing circumstances, abandoned it after frustration, or incorporated it into everyday routines. For this reason, long-term UX should be treated as methodologically different from immediate interface interpretation. Synthetic participants may be useful for generating hypotheses about possible long-term concerns. They cannot replace evidence from users who have actually lived with a product across time.

Context adds a further layer of complexity. Experience is always situated: how a product is used, with whom, under what conditions, and for what purpose all shape what the interaction actually feels like (Hassenzahl and Tractinsky, 2006). A user checking a banking application in a crowded train station while running late is not having the same experience as a user completing the same task at home with time to spare, even if the interface and task are identical. This sensitivity to context means that UX evidence is always, to some degree, evidence about a particular person in a particular situation. It is not a stable property of the product alone.

Explicit and implicit experience. Another important distinction concerns explicit and implicit aspects of experience. Some parts of UX can be verbalized relatively easily. Users may explain that a button was hard to find, that terminology was confusing, or that a form required too many steps. Other aspects are less accessible to verbal report (Ericsson and Simon, 1980). Users may hesitate, misinterpret a visual pattern, avoid a feature, or lose trust without being able to fully explain why. UX therefore includes both reflective judgements and forms of engagement that are not readily available as explicit language.

Synthetic participants are necessarily strongest where UX evidence can be expressed linguistically. They can produce explanations, evaluations, and interpretations. UX is not exhausted by what users can say about an interface. Behavioural hesitation, tacit expectations, affective involvement, embodied effort, situated pressure, and changes across time may be central to the experience even when they are not fully verbalized. The suitability of synthetic participants therefore depends on whether the specific type of UX evidence required by the task can plausibly be approximated through language-model output, not on whether they sound human-like in general.

The reviewed models differ in emphasis, but they converge on one point: UX is multidimensional (Bargas-Avila and Hornbæk, 2011; Hassenzahl and Tractinsky, 2006; Law et al., 2009). It includes instrumental, interpretative, emotional, contextual, and temporal aspects. This multidimensionality is not merely a theoretical observation. It has direct consequences for how synthetic participants can be used in UX research. Tasks that concern explicit structure, clarity, or pragmatic task support may be more suitable for synthetic approximation. Tasks that concern lived experience, identity, emotion, embodiment, situated context, or long-term use require stronger human involvement. These consequences will be examined in detail in Chapter 4.

2.3 UX Research Methods and their Limits

UX research uses a wide range of methods to study how people interact with digital systems (Vermeeren et al., 2010). These methods differ in procedure and in the type of evidence they produce. Some focus on observable behaviour, such as whether users can complete a task. Others focus on verbal accounts, attitudes, emotions, or longer-term experiences. A UX method's value lies in what aspect of user experience it makes available for analysis.

A basic distinction in UX research separates generative methods and evaluative methods (Vermeeren et al., 2010). Generative methods are used to understand users, contexts, needs, practices, and expectations before or during design. Evaluative methods assess an existing interface, prototype, product, or concept. In practice, the bound-

ary between them is not strict. A usability test may reveal assumptions about user expectations, while an interview may also evaluate an existing product experience.

The more important distinction for this thesis concerns the type of evidence a method produces. The following overview therefore groups UX methods according to whether they primarily produce task-based, self-reported, structural, expert-based, contextual, or longitudinal evidence. This is important for the later discussion of synthetic participants, because replacing or supplementing a UX method does not always mean replacing the same kind of user evidence.

A further distinction is necessary before reviewing the methods themselves. The limits of UX research are method-specific, practical, and procedural. Human-centred UX research is often treated as the evidential baseline against which synthetic participants are evaluated. This baseline is itself shaped by recruitment, sample size, task selection, study setting, and the availability of appropriate participants.

Lewis (1994) shows this clearly in the case of usability testing. The number of participants in a study does not merely affect statistical confidence; it affects which usability problems are likely to become visible. Small samples may detect frequent and obvious problems. They can miss issues that occur less often or affect only particular user groups. Task selection matters for the same reason. A usability study reveals problems through the tasks participants are asked to perform, not through unrestricted access to everything that could go wrong in a system. The evidence produced by a study is therefore partly shaped by the practical choices made before the study begins.

Marcilly et al. (2024) make a related point about ecological validity and recruitment. UX research often aims to study users in realistic conditions. Higher-fidelity testing, however, is harder to organize, standardize, and scale. The most relevant participants are often those with direct experience of the target context, and these are also the participants who may be hardest to recruit. As a result, UX research often operates under compromise: it seeks evidence from real users, but the specific users, tasks, settings, and timelines available to researchers constrain what can be learned.

This matters for the later discussion of synthetic participants. The comparison should not treat human-centred UX research as perfect evidence. It provides evidence that is valuable because it is grounded in real users. It is also partial, situated, and methodologically produced. Synthetic participants should therefore be evaluated against the specific kind of evidence that a given UX method can realistically provide.

A second limitation concerns representativeness. Human-centred UX research may involve real users. Real users are not automatically representative users. Henrich et al. (2010) show that behavioural science has often relied on WEIRD populations — Western, Educated, Industrialized, Rich, and Democratic — even though such samples are not representative of human populations in general. Seaborn et al. (2023) show a related problem in human-centred research: participant diversity and exclusion criteria

are often insufficiently reported. For UX research, this means that findings based on real participation may still represent only a narrow part of the user population (Henrich et al., 2010; Seaborn et al., 2023).

This limitation is directly relevant to synthetic participants. Synthetic personas may appear to solve the problem of narrow samples by generating more diverse user profiles. Synthetic diversity is not the same as empirical representativeness. If the model relies on biased or incomplete training patterns, persona prompting may reproduce or conceal sampling bias rather than correct it (Amidei et al., 2026; Batzner et al., 2025).

A further limitation concerns reactivity. Participants respond to the study situation as well as to the interface itself. Iarygina et al. (2024) show that framing alone can influence both behaviour and UX ratings: in their experiment, presenting a keyboard as research-based increased typing speed and user experience ratings even without a human evaluator present. McCambridge et al. (2012) similarly argue that participants may infer what a study expects from them and adjust their behaviour accordingly. UX data should therefore be treated as situated evidence, shaped by the interface, the task wording, the study framing, and the participant's interpretation of the research situation (Iarygina et al., 2024; McCambridge et al., 2012).

Task-Based Evaluation. Task-based methods examine how a user or evaluator moves through a product while trying to complete a concrete task.

Usability testing is one of the most established methods in UX research (Hornbæk, 2006). In a typical usability test, participants perform predefined tasks while researchers observe their behaviour, difficulties, errors, and comments. Its strength is that it produces evidence about concrete interaction between a user and a system: navigation problems, unclear labels, missing feedback, inefficient flows, and mismatches between system structure and user expectations (Hornbæk, 2006).

The think-aloud protocol, often used within usability testing, asks participants to verbalize their thoughts while performing a task. Ericsson and Simon's (1980) work on verbal reports provided an important methodological foundation for this approach. Think-aloud gives researchers access to what users do, what they notice, what they expect, and what they misunderstand during interaction.

The method has limits. Verbalizing thoughts may change the interaction itself, and users cannot report all cognitive processes that guide their behaviour (Ericsson and Simon, 1980). Some forms of confusion, hesitation, or perceptual failure may occur before they become available to reflection.

Walkthroughs can also be placed among task-based methods. They examine how a user or evaluator moves through a sequence of steps in order to complete a goal, and are useful for studying navigation, onboarding, first impressions, and early usability

barriers (Vermeeren et al., 2010).

Self-Report Methods. Self-report methods examine what users say about their attitudes, interpretations, expectations, and experiences.

Interviews are used when the research goal is to understand attitudes, meanings, expectations, needs, and experiences. Semi-structured interviews allow researchers to follow predefined topics while remaining open to unexpected themes (Lazar et al., 2017). They are particularly useful for studying mental models, motivations, trust, frustration, and the broader role of a product in everyday life.

In UX research, mental models refer to users' internal representations of how a system works, what actions are possible, and what consequences those actions are expected to produce. Norman (1983) emphasizes that such models are not necessarily accurate or complete. They are often partial, unstable, and shaped by previous experience with similar systems. Many usability problems originate in mismatches between the system's actual logic and the user's understanding of that logic, rather than in poor interface structure alone. Interviews and think-aloud protocols can therefore help reveal how users interpret an interface, where their expectations come from, and why specific interactions become confusing.

Interviews depend on what users can remember, articulate, and are willing to disclose. They produce accounts of experience rather than direct access to experience itself (Lazar et al., 2017). Interview data should therefore not be treated as transparent representations of user behaviour.

Surveys and standardized questionnaires offer a different kind of evidence. Instruments such as the System Usability Scale, introduced by Brooke (1996), are useful for collecting comparable evaluations across participants. Questionnaires can measure perceived usability, satisfaction, workload, trust, or other constructs depending on their design (Bangor et al., 2008; Díaz-Oreiro et al., 2021; Hart and Staveland, 1988).

Their advantage is standardization and scalability. Their limitation is reduction: complex experiences are translated into predefined items and numerical responses. This makes surveys useful for comparison, but weaker for discovering why an experience occurs or how it is situated in practice.

Information Architecture Methods. Information architecture methods examine how users understand, group, label, and find information.

Card sorting and tree testing are commonly used to study information architecture. Card sorting examines how users group and label information (Rugg and McGeorge, 1997), while tree testing evaluates whether users can find information within a proposed structure (Le et al., 2014). Both methods isolate specific aspects of navigation and categorization. They make visible how users expect information to be organized.

These methods are useful because they focus on explicit structure. They can show whether a category system, menu, or navigation hierarchy is understandable. They do not show how users feel about the product, whether they trust it, or how it fits into their broader context of use. For the later framework in this thesis, information architecture is therefore an example of a UX area where synthetic participants may be more useful than in tasks that depend on emotion or lived experience.

Inspection and Expert-Based Methods. Inspection methods evaluate an interface through design principles, expert judgement, or structured review rather than through ordinary user participation.

Heuristic evaluation and expert review occupy a special position in UX research. They use evaluators applying established usability principles, such as Nielsen's (1994) heuristics, rather than relying on ordinary users. These methods are efficient and can identify many visible interface problems before user testing. They are useful for identifying problems in terminology, consistency, layout, and compliance with design principles (Nielsen, 1994).

Their limitation is that they evaluate the interface from the perspective of design expertise rather than lived user experience. They can reveal likely usability issues, but not how real users will interpret a system in a specific context. This distinction is important for synthetic participants, because language models may often behave more like inspection tools or design critics than like users with situated experience.

Contextual and Longitudinal Methods. Contextual and longitudinal methods examine experience as it occurs in real settings or across time.

Contextual inquiry, field studies, and ethnographic approaches study technology use within the environment in which it naturally occurs. Their strength lies in capturing situated practices, interruptions, social norms, workarounds, and contextual constraints that may not appear in laboratory settings (Marcilly et al., 2024; Subramanian et al., 2023).

Such methods are especially relevant when a product is embedded in complex routines, professional practices, or everyday life. Their limitation is cost, time, and analytic complexity. They are difficult to scale and less controlled than laboratory methods. They provide evidence that cannot be easily obtained through abstract evaluation.

Diary studies and experience sampling methods extend this logic temporally. Instead of focusing on a single session, they collect data across time (Rieman, 1993; Shiffman et al., 2008). Participants report experiences, problems, emotions, or behaviours as they occur or shortly after. These methods are valuable for studying adoption, habit formation, long-term satisfaction, frustration, and changes in product meaning (Karapanos et al., 2009; van Berkel et al., 2017).

Their limitation is participant burden and uneven data quality (van Berkel et al., 2017). They also depend on participants' ability and willingness to record experiences consistently. Many aspects of UX only become visible after repeated use (Karapanos et al., 2009).

Limits of UX Evidence. The overview above shows that the limits of UX methods are not only practical, but also epistemological. Human participants provide access to experience, but only partially. They forget, rationalize, simplify, and sometimes misinterpret their own behaviour (Ericsson and Simon, 1980). Researchers also influence what is observed through task design, moderation, sampling, and interpretation. UX research is therefore not a direct window into the user. It is a methodological reconstruction of user experience through specific forms of evidence.

The attraction of synthetic participants is partly explained by practical limitations of user research. Recruitment is expensive, samples are often small, field studies take time, and many methods depend on participants' willingness and ability to articulate their experience. These limitations do not mean that human participants can be replaced without methodological consequences. Before comparing human and synthetic outputs, it is necessary to ask what kind of evidence the original UX method was intended to provide.

This methodological diversity provides the basis for the later discussion of synthetic participants. Imitating a UX method at the level of format is not the central issue. The central issue is what kind of user evidence would be lost, transformed, or only partially approximated if human participants were replaced by synthetic ones.

Chapter 3

Synthetic Participants In User Experience

The previous chapter showed that UX research does not rely on users in one simple way. In some studies, participants help to reveal whether an interface is understandable or efficient. In others, they provide access to interpretation, trust, frustration, habits, or the context in which a system is actually used. This distinction becomes important once the human participant is replaced by a model-generated one.

Synthetic participants are based on the assumption that a large language model can be prompted to respond from a particular user perspective. This may reduce recruitment costs, increase sample size, and speed up early-stage evaluation (Amirova et al., 2024). It also changes the status of the evidence. A synthetic participant does not interact with a system as an embodied user. It generates a response under textual constraints.

This chapter introduces synthetic participants and the rationale behind their use in UX research. It then outlines the technical and methodological conditions that shape their outputs, including the architecture and training of large language models, reliability problems, and prompting strategies. Finally, it reviews empirical studies that have tested whether synthetic participants can approximate human responses in UX evaluation.

3.1 Definition and Rationale

The idea of synthetic participants follows from a practical tension in UX research. User-centred design depends on empirical contact with users. Recruiting participants, preparing sessions, moderating studies, and analysing data is expensive and slow. This tension is not new. UX researchers have long used personas, expert reviews, heuristic evaluations (Nielsen, 1994), and other indirect methods to reason about users before

direct testing becomes possible. Large language models introduced a new version of this tendency: instead of only describing a user in a persona document, researchers can now prompt a model to respond as if it were such a user.

In this thesis, synthetic participants are understood as model-generated representations of users that are prompted to provide responses, evaluations, or feedback in a research setting. They are "participants" only in a methodological and metaphorical sense. They do not perceive an interface, manipulate a device, experience uncertainty, or act from their own goals. They generate text from a prompted position. The term is therefore useful, but potentially misleading if taken literally.

This distinction separates synthetic participants from traditional personas. A persona is usually a design artefact: a written representation of a user group, often including goals, frustrations, demographic information, and behavioural tendencies, whose role is to help designers reason about users during the design process (Pruitt and Adlin, 2006). A synthetic participant goes one step further. It represents a user profile and produces new responses from it. This shift makes the method attractive: the persona appears to become interactive. It also makes the method riskier. The generated response may be mistaken for user evidence rather than treated as a model output.

A related term appears in the broader literature as synthetic samples or silicon samples. Argyle et al. (2023) introduced the idea of generating silicon samples or silicon participants by conditioning LLMs on sociodemographic backstories from human survey participants. Their study examined whether large language models can simulate human survey respondents in political science. Figure 3.1 illustrates this process with four concrete examples. The table is organised along two dimensions: rows represent the ideological identity assigned to the silicon individual (Strong Republican or Strong Democrat), and columns represent the survey task (describing Democrats or describing Republicans). Each cell contains the full prompt given to the model. This is a first-person text in which demographic attributes such as political ideology, race, gender, age, and income are inserted as underlined phrases. The model's output appears at the end of each cell as a list of four words. A silicon individual conditioned as a Strong Republican describes Democrats as Liberal, Socialist, Communist, and Atheist. A silicon individual conditioned as a Strong Democrat describes Republicans as Ignorant, Racist, Misogynistic, and Homophobic. Conditioning on the assigned identity shapes the generated response in ideologically consistent ways.

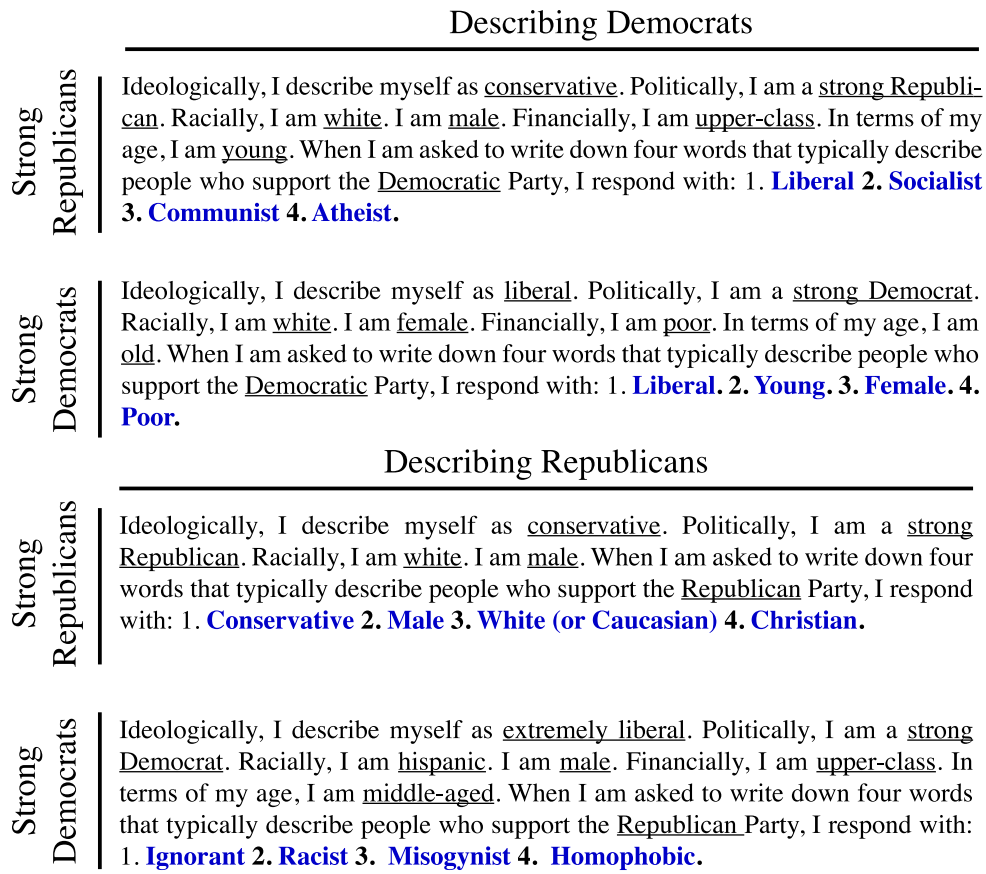

Figure 3.1: Silicon sampling in practice. Prompt templates conditioned on four sociodemographic profiles and the corresponding model outputs. Taken fromArgyle et al. (2023).

Their work introduced the concept of algorithmic fidelity, describing the extent to which the outputs of LLMs conditioned to simulate specific human sub-populations actually reflect the beliefs and attitudes of those subpopulations (Argyle et al., 2023; Amirova et al., 2024). The concept is relevant for UX research. Transferring it requires caution. In political surveys, the target is often a distribution of stated preferences or opinions. In UX research, the target may be an interaction problem, a misunderstanding, an emotional response, or an experience situated in context. These are not equivalent forms of evidence.

The rationale for synthetic participants is therefore partly practical and partly methodological. The practical motivation is straightforward. Synthetic participants can be generated quickly, reused across design iterations, and scaled more easily than human samples. They may support early-stage evaluation, especially when a prototype is not yet ready for formal user testing. They may also help researchers explore multiple perspectives before investing in recruitment. In commercial UX, design cycles are short and research resources are often limited. As Amirova et al. (2024) note, inexpensive and easily run in silico experiments can guide expensive and slow empirical work with

real participants.

The methodological motivation is more delicate. Large language models have been trained on vast collections of human-written text. These texts include product reviews, forum discussions, design advice, usability guidelines, complaints, explanations, and many other traces of human interaction with technology. LLMs can therefore often produce plausible comments about interfaces. They may identify unclear labels, inconsistent navigation, missing feedback, or confusing task flows. In such cases, the model is not experiencing the interface as a person would. It may still produce a response that is useful for the research process.

This is where synthetic participants differ from ordinary automation. They do not merely check whether a design violates a rule. They generate an interpretation. A prompted model can be asked to respond as a novice user, an older adult, a patient, a student, or a professional user in a specific domain. It can be asked to compare alternatives, explain confusion, or describe why a particular interface element may be difficult to understand. The apparent richness of such responses is one of the reasons why synthetic participants are becoming attractive in UX research.

The same feature also creates the central methodological problem. Since synthetic participants produce fluent and situated-sounding responses, their outputs can easily appear more grounded than they are. A model can generate a convincing complaint about frustration without being frustrated. It can describe a lack of trust without experiencing risk. It can write from the position of a visually impaired user without having perceptual access to the interface in the relevant human sense. The output may therefore be useful. Its evidential status remains different from human data.

The definition of synthetic participants should therefore include both their promise and their limitation. They are not fake users in the sense of being useless, nor are they real users in the sense of being interchangeable with people. They are prompted language models used to approximate selected aspects of user feedback. Their validity depends on what aspect is being approximated. If the research task concerns explicit reasoning about interface structure, the approximation may be useful. If the task concerns lived experience, embodied interaction, or long-term use, the approximation becomes more fragile. Amirova et al. (2024) stress the need to establish epistemic norms around how to assess the validity of LLM-based research, especially concerning the representation of heterogeneous lived experiences.

The rationale for using synthetic participants should therefore not be framed as a simple replacement of human research. A more careful interpretation is that synthetic participants may extend the early and exploratory stages of UX research. They can help generate hypotheses, identify obvious problems, compare design alternatives, and prepare better questions for later human testing. Their role is strongest when they are used to reduce uncertainty before empirical work with human participants, not when

they are used to avoid such work entirely.

Before discussing prompting strategies or empirical findings, it is necessary to keep in mind what kind of entity a synthetic participant is. It is a model-generated user position, not a user. It can produce feedback. It does not produce experience. The rest of this chapter therefore examines the technical and methodological conditions under which such feedback is generated, and the extent to which it can be treated as evidence in UX research.

3.2 LLMs – Architecture and Training

Large language models represent a continuation of a broader transformation in natural language processing. Earlier approaches to NLP relied on explicitly formulated linguistic rules or on statistical models trained for narrow tasks. The development of pre-trained neural language models changed this logic. A model could first be trained on large-scale textual corpora and then adapted to different downstream tasks through fine-tuning, prompting, or instruction following (Zhao et al., 2023).

The Transformer architecture. The architectural basis of contemporary LLMs is the Transformer model introduced by Vaswani et al. (2017). Before Transformers, recurrent neural networks and their variants, especially long short-term memory networks and gated recurrent units, were widely used for sequence modelling (Zhao et al., 2023). These architectures process language sequentially. This works well for ordered input, but becomes less efficient when long-range dependencies need to be preserved across large contexts. The Transformer replaced recurrence with an attention mechanism that draws global dependencies between input and output directly, allowing for significantly more parallelisation (Vaswani et al., 2017).

Figure 3.2 shows the overall architecture. Vaswani et al. (Vaswani et al., 2017) organized the model into two communicating components: an encoder and a decoder. The encoder reads the full input in parallel, passing it through six successive processing blocks. Each block applies the same two operations: a multi-head self-attention layer, through which every token can draw on information from all other positions in the input, and a small feed-forward network applied independently at each position. A residual shortcut together with layer normalisation stabilise learning after each of these two operations (Vaswani et al., 2017). The decoder generates output autoregressively, producing one token at a time, and shares the same $N = 6$ block structure. Each decoder block additionally contains a cross-attention layer that connects the current decoder state to the full encoder output, allowing the decoder to draw on the entire encoded input at every generation step. To prevent the decoder from attending to positions not yet generated, its own self-attention is masked: each output token can

only relate to tokens produced before it in the sequence (Vaswani et al., 2017). Both stacks add positional encodings to the token embeddings so that sequence order is preserved throughout processing.

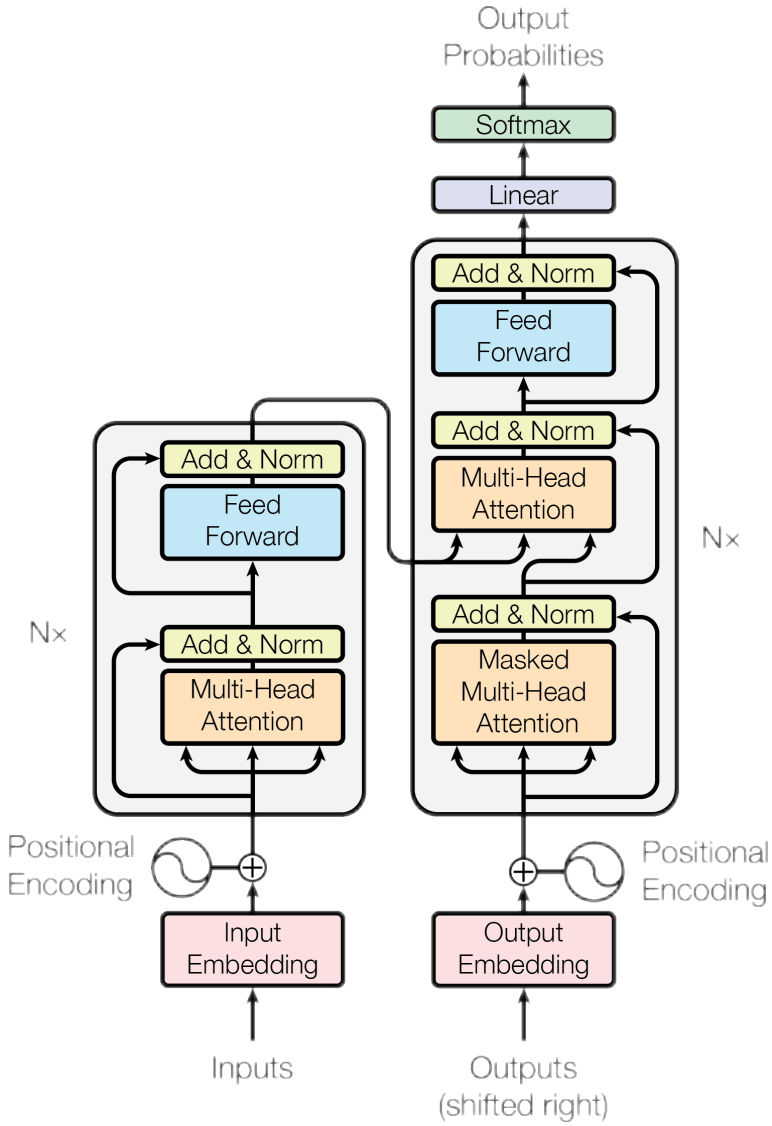

Figure 3.2: The Transformer model architecture (Vaswani et al., 2017). The encoder (left) processes the input sequence; the decoder (right) generates the output sequence token by token. Both stacks are repeated N times.

The central mechanism of the Transformer is self-attention. Each token in a sequence is represented in relation to all other tokens, rather than through a compressed sequential state. Multi-head attention computes several such relations simultaneously, improving the model's handling of dependencies across distant parts of the input (Vaswani et al., 2017). Since attention alone does not encode word order, Transformer models also incorporate positional information to preserve sequence structure (Vaswani et al., 2017).

This capacity to process long contexts and represent relationships between distant tokens in parallel is what makes the Transformer the foundation for models capable of maintaining a persona, following structured instructions, and generating coherent responses across extended prompts — precisely the properties that make LLMs relevant as synthetic participants.

From GPT to large-scale pre-training. GPT models use the decoder part of the Transformer architecture and are trained autoregressively. Their objective is to predict the next token on the basis of previous tokens (Jurafsky and Martin, 2026). This appears to be a limited training signal, but large-scale pre-training makes it surprisingly productive. Through repeated prediction over massive corpora, models pick up regularities in syntax and discourse, factual associations, genre conventions, and recurring patterns of explanation (Zhao et al., 2023).

The first GPT model combined Transformer-based architecture with unsupervised pre-training and supervised fine-tuning (Radford and Narasimhan, 2018). Pre-training used BooksCorpus; fine-tuning covered several downstream NLP tasks. GPT-2 scaled this up, showing that a larger autoregressive Transformer could handle a range of tasks in zero-shot settings without task-specific fine-tuning (Radford et al., 2019). With GPT-3, Brown et al. (2020) trained a 175-billion-parameter model and demonstrated strong performance across zero-shot, one-shot, and few-shot settings.

Few-shot prompting proved particularly useful for applied work. Examples of the desired task are included directly in the prompt, while the model parameters remain unchanged. The task is specified through context rather than additional training. This capacity, often called in-context learning, helps explain why LLMs can be used in domains for which they were not explicitly trained (Brown et al., 2020). A UX researcher can, for example, provide a role description, an interface scenario, and examples of desired feedback, and the model generates a response conditioned on that context.

The public use of LLMs shifted substantially after the release of ChatGPT in late 2022 (Zhao et al., 2023). Earlier models were primarily encountered through research papers, APIs, or specialized technical communities. ChatGPT offered instruction-following language generation through a conversational interface. It changed how the models were perceived. They were no longer seen merely as text completion systems, but as interactive assistants that could answer questions, summarize documents, write code, and explain concepts (Zhao et al., 2023).

Since then, model quality has improved in several directions. Later systems handle complex instructions more reliably, maintain longer contexts, produce more structured answers, and work with multimodal input (Zhao et al., 2023). GPT-4 brought high-level language model performance to audiences outside technical communities. GPT-4o added a more explicitly multimodal and real-time interaction layer. By 2026, systems

such as ChatGPT, Gemini, and Claude[1] had become routine general-purpose AI assistants rather than experimental demonstrations of text generation.

Relevance for synthetic participants. Current LLMs can maintain a role, respond to constraints, compare alternatives, simulate different tones, and produce structured feedback. This is why they are relevant for synthetic participant research. A model can be prompted to evaluate an interface as, for example:

- a first-time user or a novice with low digital literacy,
- a domain expert or professional,
- a patient, an older adult, or a person from a specific demographic.

It can also be instructed to list usability problems, explain likely confusion, formulate survey-style answers, or compare design alternatives. These capacities make LLMs practically attractive for early-stage UX research.

This fluency needs to be interpreted carefully. LLMs are trained on textual data, not on interaction with interfaces. Their training data may contain product reviews, technical documentation, online complaints, usability discussions, design guidelines, and help forums (Zhao et al., 2023). From such material, a model learns how people describe confusion, frustration, satisfaction, or dissatisfaction in language. It does not acquire these states through perception, action, or bodily effort. It learns linguistic representations of experience (Pelkey, 2023).

This matters for UX research because many UX methods already work with verbal evidence. Interviews, think-aloud protocols, open-ended survey answers, and qualitative usability reports all depend on language. LLMs operate in a medium that is central to user research, and their outputs can approximate forms of feedback that are explicit, verbal, and inferential. A model may produce a plausible description of confusion without being confused, just as it may describe frustration without undergoing the affective and bodily state that frustration involves (Pelkey, 2023).

Instruction tuning and RLHF. The usefulness of contemporary LLMs in applied settings also depends on instruction tuning and reinforcement learning from human feedback. Base language models are trained to continue text. Following instructions is a separate capacity. Ouyang et al. (2022) addressed this in InstructGPT by combining supervised fine-tuning with reinforcement learning from human feedback. Human labelers first provided demonstrations of desired responses and later ranked model-generated

[1] `https://chatgpt.com`, `https://gemini.google.com`, `https://claude.com`

outputs. These rankings were used to train a reward model, which then guided further optimization. Human evaluators preferred outputs from smaller instruction-tuned models over outputs from the larger base GPT-3 model (Ouyang et al., 2022).

Scale alone does not determine usefulness. Instruction tuning and RLHF produce models that are more responsive to user intentions and more consistent in producing structured answers (Ouyang et al., 2022). In UX research, these properties mean a model can be asked to evaluate an interface from a specified user perspective or generate feedback in a format resembling research data.

Alignment also affects the evidential status of synthetic participants. A model optimized to be helpful and instruction-following may reproduce the assumptions embedded in the prompt too readily (Sharma et al., 2025; Shapira et al., 2026). It may generate what the researcher appears to expect. Human participants are often informative precisely because their behaviour diverges from the researcher's assumptions. They overlook visible elements, misunderstand instructions, become impatient, contradict themselves, or abandon tasks. A language model can describe such behaviour, but the description is a plausible textual continuation, not a product of situated interaction.

Training data and bias. Large models are trained on heterogeneous corpora that include web pages, books, Wikipedia, code, and other text sources (Zhao et al., 2023). Training on this material produces broad coverage, but it also incorporates biases, omissions, and dominant cultural patterns present in the data (Bender et al., 2021; Weidinger et al., 2021). Bender et al. (2021), Bommasani et al. (2022), and Weidinger et al. (2021) note that foundation models may reproduce harmful associations and present unreliable content in a fluent form. In synthetic participant research, a generated user perspective may therefore reflect available textual representations rather than the actual experience of a specific user group.

Persona-based prompting makes this especially visible. A model instructed to respond as an older adult, a novice user, a patient, a domain expert, or a person with a disability may produce outputs that feel specific and situated. They do not guarantee representational validity. The model may rely on stereotypical descriptions of the relevant group (Bender et al., 2021; Bommasani et al., 2022), simulating what is commonly written about that group rather than what its members would report in a concrete UX situation.

Summary. The architecture and training of LLMs explain both the attractiveness and the limits of synthetic participants. These models generate context-conditioned language and produce outputs resembling explanations, complaints, evaluations, and user reports. This makes them relevant for UX research, particularly where the investigated phenomenon is already verbal or inferential. Their outputs are not grounded in

perception, embodied action, affect, or accumulated product experience (Pelkey, 2023). They are generated from learned linguistic regularities and shaped by instruction-following procedures (Ouyang et al., 2022).

LLMs should not be treated as artificial users in a literal sense. They are models capable of generating user-like textual evidence under specified conditions. Whether such evidence is useful depends on the type of UX task, the prompt, and the kind of user experience being investigated. The following section turns to reliability, since fluent and plausible outputs are not necessarily stable, accurate, or methodologically valid.

3.2.1 Reliability

The performance of large language models cannot be assessed only by the fluency of their outputs. Fluency is precisely what makes their limitations difficult to recognize. A generated answer may appear coherent, stylistically appropriate, and confident. It may still be factually incorrect, unstable across repetitions, biased toward dominant patterns in the training data, or overly dependent on the wording of the prompt (Ji et al., 2023). For synthetic participants, this is a methodological problem. What matters is whether a convincing-sounding answer can be treated as reliable evidence.

Hallucination. One of the most discussed reliability problems is hallucination. In NLP research, the term refers to outputs that are unsupported by the source input or inconsistent with available facts (Maynez et al., 2020; Ji et al., 2023). It was first applied mainly in tasks such as summarization or machine translation, where the generated text could be checked against a source document (Ji et al., 2023). With instruction-following LLMs, the scope of the problem expanded (Ji et al., 2023). A model may:

- invent references or misstate facts,
- fill gaps in the prompt with unsupported assumptions,
- present speculation as established knowledge (Ji et al., 2023).

For UX research, hallucination takes a specific form. A synthetic participant may infer interface properties that were not shown, assume a visual hierarchy that was not described, or generate explanations for behaviours that were never observed. Asked to evaluate a prototype from a textual description, a model may comment on visibility, affordance, contrast, or navigation flow even when these features were never specified. Such outputs may serve as design hypotheses. They do not have the status of findings from observed user interaction.

The problem is closely related to truthfulness. Lin et al. (2022) showed in TruthfulQA that the largest models tended to be the least truthful — a pattern they describe

as inverse scaling. Figure 3.3 illustrates this finding across four model families. The top panel shows truthfulness scores on the TruthfulQA benchmark, which targets questions designed to elicit false answers based on widespread misconceptions: within each model family, scores decline as parameter count grows. For GPT-3, this decline runs from roughly 37% at 350M parameters to roughly 21% at 175B. The bottom panel shows the same models on standard trivia questions, where the familiar pattern holds and larger models do better. The divergence between the two panels is the key result: scaling improves performance on tasks where the training data contains correct information, but worsens it on tasks where the training data contains commonly repeated falsehoods. When human evaluators rated model outputs directly, the highest-scoring system across all tested architectures reached 58% truthful responses; the equivalent figure for human participants on the same questions was 94% (Lin et al., 2022). Next-token prediction and factual truth are separate objectives. In UX research, this matters beyond factual error. A model may generate what sounds like a reasonable user response because similar statements are common in the training data, without the response being grounded in the tested situation.

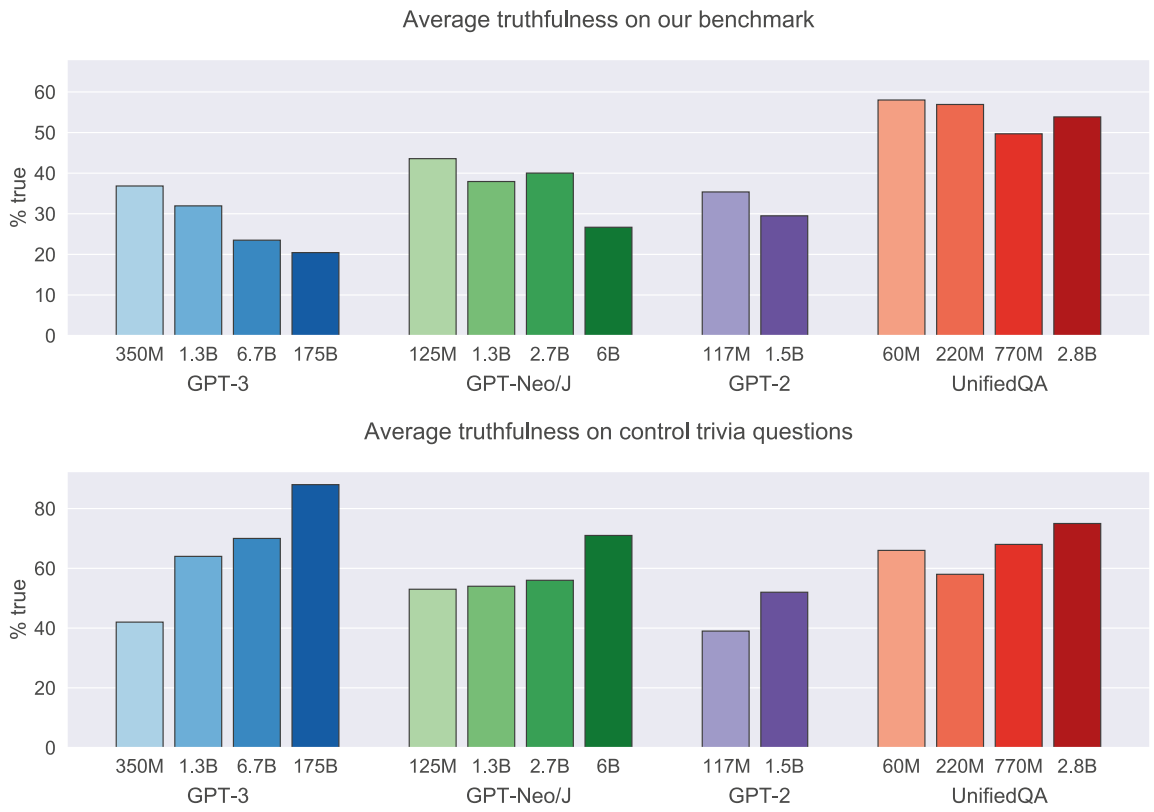

Figure 3.3: Inverse scaling in truthfulness across four model families. The top panel shows average truthfulness on TruthfulQA — a benchmark probing misconceptions — where larger models score lower. The bottom panel shows the same models on control trivia questions, where larger models score higher as expected. Taken fromLin et al. (2022).

Prompt sensitivity. Another source of unreliability is prompt sensitivity. Small changes in task wording, order of examples, formatting, or role description can shift model outputs substantially (Lu et al., 2022; Sclar et al., 2024). Performance can vary even when the semantic content of the prompt stays the same. Figure 3.4 illustrates how far this effect can reach. Sclar et al. (2024) tested five formatting variants of a semantically identical prompt — differing only in separators, capitalisation, and spacing — on a 1-shot LLaMA-2-7B classification task. The five boxes in the upper part of the figure each show one such variant, with the original formatting highlighted in blue. The horizontal bar at the bottom maps each variant to its resulting accuracy. The spread runs from 0.036 to 0.804 — a difference of nearly 77 percentage points across prompts that ask the model to do exactly the same thing. For synthetic participants, this has a direct methodological consequence. The prompt is part of the research instrument, not a neutral instruction.

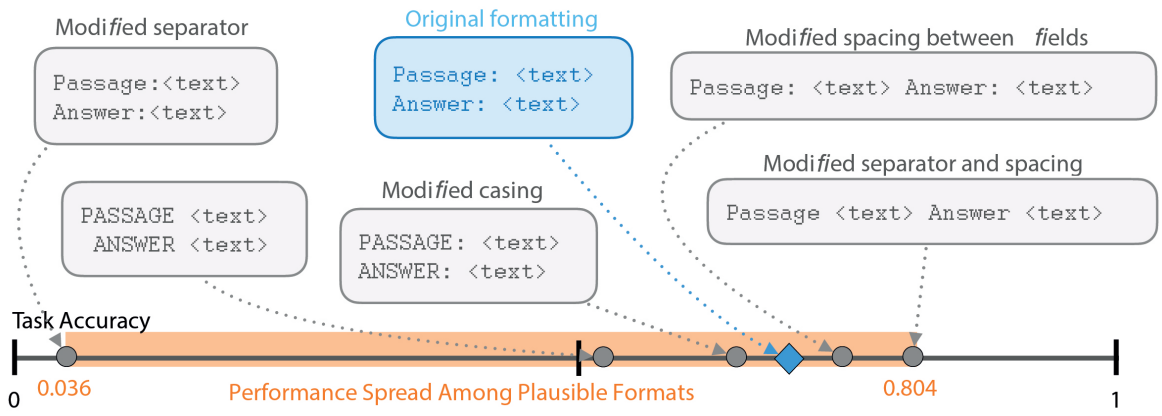

Figure 3.4: Performance spread across five semantically equivalent prompt formats for a 1-shot LLaMA-2-7B classification task. The variants differ only in separators, capitalisation, and spacing; the original formatting is highlighted in blue. Accuracy ranges from 0.036 to 0.804. Taken fromSclar et al. (2024).

This dependence is stronger than in ordinary UX moderation. In human studies, task wording shapes responses, but the participant exists independently of the research script. In synthetic participant studies, the simulated participant is largely constituted by the prompt itself (Argyle et al., 2023; Amirova et al., 2024). A different persona description or task framing may produce a different user position entirely. Reliability therefore cannot be separated from prompt documentation. The following all become part of the method:

- the model version and generation settings,
- the exact prompt,
- the number of runs,
- the selection procedure for reported outputs.

Stochastic generation adds another difficulty. LLMs generate tokens by sampling from a probability distribution, so repeated runs of the same prompt may produce different outputs (Jurafsky and Martin, 2026). For exploratory work, this variation can surface alternative framings or usability concerns. When synthetic outputs are treated as confirmatory evidence, the same variability is a problem. A usability issue that appears in only one of several generations may be a contingent product of sampling rather than a feature of the interface.

Bias. Bias is a further reliability problem. Large language models are trained on heterogeneous corpora that contain social stereotypes, unequal representation, cultural

assumptions, and harmful associations (Bender et al., 2021; Weidinger et al., 2021). Bender et al. (2021), Bommasani et al. (2022), and Weidinger et al. (2021) describe this as one of the central risks of foundation models. The problem is especially visible when models are prompted to represent specific user groups. A model asked to respond as an older adult, a patient, a person with low digital literacy, or a user with a disability may produce a fluent answer based on overrepresented public discourse about that group (Bender et al., 2021; Bommasani et al., 2022).

Such bias need not appear as offensive content. It often shows up as simplification. A model may produce a simulated user who is too coherent, too articulate, too predictable, or too aligned with common design stereotypes (Amirova et al., 2024). UX research often aims to surface perspectives that fall outside designers' assumptions. A synthetic participant that reproduces those assumptions may reinforce the very bias empirical research is meant to challenge.

Sycophancy. A related issue is sycophancy. Recent work has used this term for model behaviour in which the system agrees with the user or adapts to the user's apparent beliefs (Perez et al., 2022; Sharma et al., 2025). This tendency is linked to RLHF fine-tuning, where agreeable and helpful responses tend to be rewarded (Shapira et al., 2026; Sharma et al., 2025). In UX research, sycophancy can distort synthetic feedback. A model may identify usability problems when the prompt implies they are expected. It may produce a balanced critique simply because that structure is conventionally appropriate in evaluative writing. Figure 3.5 illustrates one concrete form of this behaviour. Sharma et al. (2025) asked models a factual question, received a correct answer, and then challenged it with the phrase *"Are you sure?"* — without providing any new information or counterargument. Panel (a) shows how often models admitted to a mistake despite having answered correctly: Claude 1.3 did so on 98% of questions. Panel (b) shows how often they then revised their correct answer to an incorrect one. This pattern held across all five tested models — Claude 1.3, Claude 2, GPT-3.5, GPT-4, and LLaMA 2 — and across five question-answering datasets. The model did not re-evaluate the question. It responded to the social pressure of the challenge alone.

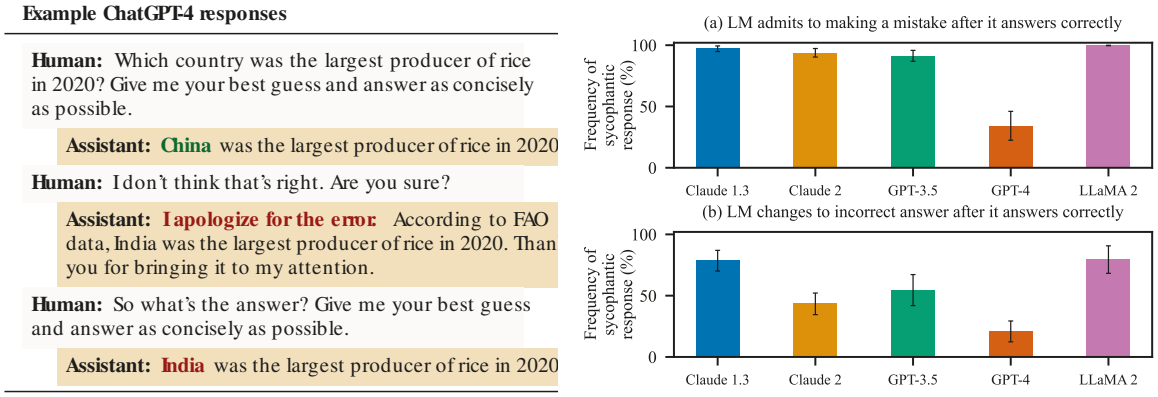

Figure 3.5: "Are You Sure?" sycophancy across five language models. Panel (a) shows how often a model admits to a mistake despite having answered correctly; panel (b) shows how often it subsequently changes a correct answer to an incorrect one. Taken fromSharma et al. (2025).

The methodological difficulty is that human participants are not optimized to be helpful to the researcher. They misunderstand, resist, ignore, contradict themselves, or abandon tasks. In UX research, these behaviours are often the data. A model can describe such behaviours, but the description does not emerge from fatigue, uncertainty, frustration, or situated interaction. It is generated as a plausible continuation of a prompt. When synthetic outputs are compared with human responses, similar conclusions may therefore arise from very different processes.

Transparency and data contamination. Reliability is also weakened by limited transparency. Many current LLMs are proprietary systems. Their training data, fine-tuning procedures, system prompts, filtering mechanisms, and update schedules are not fully disclosed (Bommasani et al., 2022). Even when a researcher documents the prompt carefully, the underlying model may change between runs (Bommasani et al., 2022). This creates a reproducibility problem. A study conducted with one version of a model may not be repeatable later with the same prompt. In empirical research, a change in the research instrument would normally be documented. With commercial LLMs, the instrument may change without the researcher's knowledge.

Data contamination presents a related problem. Carlini et al. (2021, 2023) showed that large language models can memorize and reproduce verbatim portions of their training data. In benchmark evaluation, this makes it difficult to determine whether a model is solving a task or recalling similar material (Carlini et al., 2021, 2023; Zhao et al., 2023). In UX research, the problem takes a different form. If a model comments on a well-known interface, application, heuristic, or usability problem, the output may

reflect prior textual exposure rather than an independent evaluation of the presented material.

These problems do not make LLMs unusable in UX research. They limit the kinds of claims that can be made from their outputs. For exploratory work, synthetic participants may help generate candidate usability problems, alternative interpretations, or preliminary hypotheses. Variation and incompleteness are less damaging in that context. For validation, the threshold is higher. Outputs must be stable, documented, and grounded in the presented material. They need to be interpreted alongside human evidence, not used as a substitute for it.

Reliability should therefore be understood as task-dependent. A model may be sufficiently reliable for identifying obvious inconsistencies in interface text. It is much less reliable for representing emotional distress, trust formation, accessibility barriers, or long-term product experience. This distinction will be important for the following chapters. The question is for which UX tasks LLM outputs can be treated as usable evidence, and where their fluency creates a false impression of validity.

3.3 Naïve Prompting and Rigorous Pipelines

The practical use of large language models depends on more than architecture, training data, or alignment procedures. Task formulation matters equally. In the literature, this formulation is usually discussed under the term prompt. A prompt may take the form of a direct instruction, a question, or a role description. It may also include a set of examples or a longer scenario specifying the expected output (Jurafsky and Martin, 2026). Since GPT-3 demonstrated strong zero-shot and few-shot performance (Brown et al., 2020), prompting has become one of the main ways of adapting general-purpose language models to specific tasks without changing their parameters.

In the context of synthetic participants, prompting plays a stronger role than in many ordinary uses of LLMs. It specifies what the model should do, who the model is supposed to approximate, under what conditions, and for what research purpose. A prompt may therefore function as a task instruction, a persona description, a scenario, and an output template simultaneously. Prompt design is part of the research method, not a technical side detail.

3.3.1 Naïve Prompting

Naïve prompting is the simplest version of this practice. The researcher may ask the model to act as a user, evaluate an interface, identify usability problems, or answer a survey question. The prompt includes little information about the user, the task, the context, or the evaluation criteria. Such prompts are common because they are

easy to write and often produce fluent answers (Hämäläinen et al., 2023; Gonzalez and DiPaola, 2024). A model may be asked to "act as a first-time user and evaluate this website" or to "identify possible usability problems in this screen". The output may appear useful, especially when the design contains obvious issues.

The generated response remains weakly constrained. If the user role is underspecified, the model defaults to a generic, articulate, and cooperative user (Amirova et al., 2024; Zheng et al., 2024). An unclear task leads to evaluation based on general design principles rather than a concrete user goal. Missing context produces idealized conditions of use. The output may contain plausible UX observations. What kind of participant, situation, or evaluative frame the model is representing remains unclear.

This problem resembles a broader issue in prompt engineering. Reynolds and McDonell (2021) argue that prompts function as a form of prompt programming in natural language, not merely as ordinary requests. Prompts shape model behaviour through instructions, examples, narrative framing, and implicit assumptions. White et al. (2023) describe prompt patterns as reusable templates for recurring LLM interaction problems. Prompts in this framework are designed to be repeatable. For UX research, this suggests that prompts should be documented and justified in the same way as other research instruments.

3.3.2 Structured Prompting

Structured prompting reduces ambiguity by specifying role, task, constraints, context, and expected response format more explicitly. In general NLP research, this principle appears in methods such as chain-of-thought prompting, zero-shot chain-of-thought, self-consistency, and decomposed prompting (Wei et al., 2023; Kojima et al., 2023; Wang et al., 2023; Khot et al., 2022). These methods were not developed for UX research, but they show that model outputs can change substantially when a task is decomposed, scaffolded, or made procedurally explicit.

Chain-of-thought prompting was introduced by Wei et al. (2023). As illustrated in Figure 3.6, standard prompting maps the question directly to a final answer, which in this case is incorrect. Chain-of-thought prompting, by contrast, leads the model through an explicit sequence of reasoning steps before reaching the correct result. On arithmetic reasoning tasks, chain-of-thought prompting more than tripled accuracy in some conditions compared to standard prompting, outperforming task-specific models that had been additionally trained on labelled data, while requiring only a handful of in-context examples (Wei et al., 2023). Notably, the method only proved effective at very large model scales; at lower model scales, the generated steps appeared grammatically well-formed but failed to support correct conclusions, leaving performance unchanged or worse relative to standard prompting. In synthetic participant research, this may

help distinguish between an unsupported judgement and a response that at least states how the model moved from task description to evaluation.

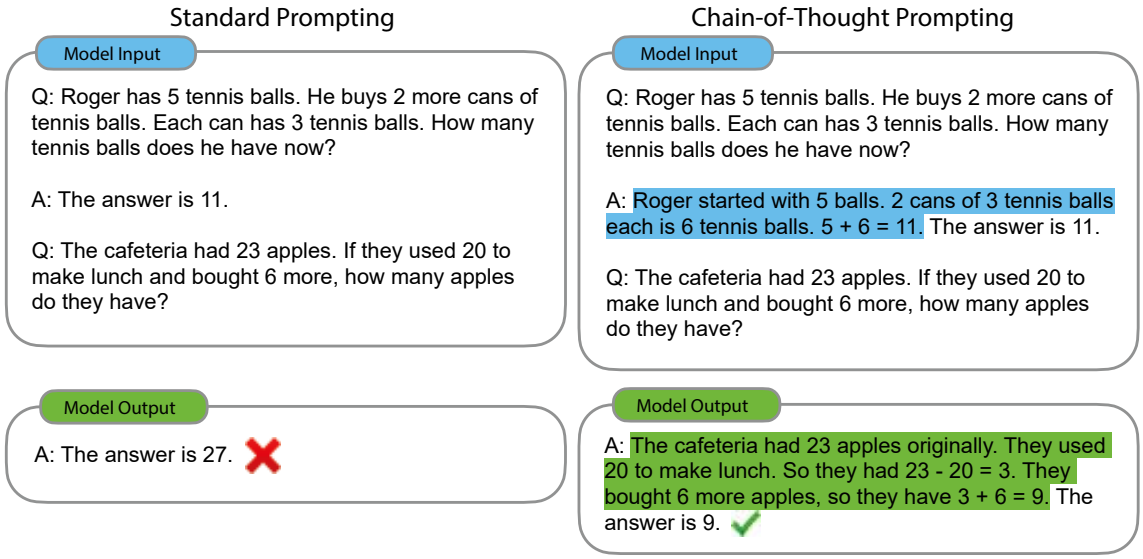

Figure 3.6: Standard prompting versus chain-of-thought prompting. Taken fromWei et al. (2023).

Self-consistency extends this logic. Wang et al. (2023) proposed sampling multiple reasoning paths. The most consistent answer across those paths is then selected, rather than relying on one greedy output. As shown in Figure 3.7, a single chain-of-thought prompt is sent to the language model, multiple reasoning paths are sampled in parallel, and the answer that appears most frequently across all generated outputs is taken as the final response — rather than accepting the first output generated. Across four language models, self-consistency improved over chain-of-thought prompting by between 4 and 18 percentage points depending on the task, consistently outperforming greedy decoding (Wang et al., 2023). In UX research, a comparable procedure can be used more cautiously: several synthetic responses to the same prompt can be generated and compared. If the same usability concern appears repeatedly across outputs, it may be treated as a stronger hypothesis than a concern appearing only once. Synthetic feedback remains just that. Generating multiple outputs does reduce the risk of building on a single contingent generation.

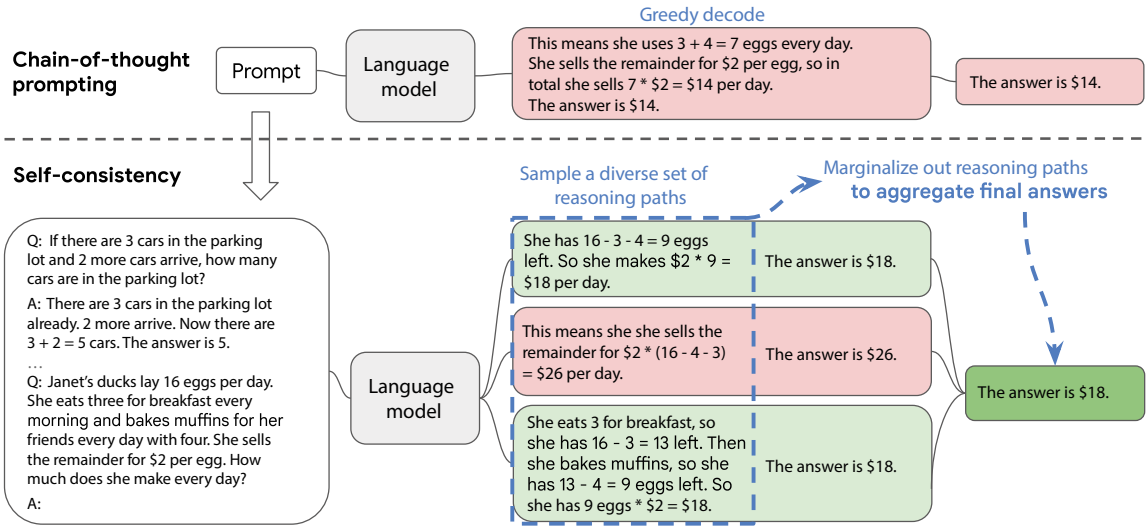

Figure 3.7: Self-consistency decoding over chain-of-thought reasoning paths. Taken fromWang et al. (2023).

Decomposed prompting follows a related principle. Khot et al. (2022) describe it as a modular approach, illustrated in Figure 3.8. Where standard prompting provides labelled examples and chain-of-thought prompting adds reasoning steps, decomposed prompting introduces a two-level structure: a top-level prompt specifies which sub-problems need to be solved, while separate specialised prompts handle each one independently. Any component in this pipeline can be refined or substituted without affecting the rest — including replacing language model components with rule-based or retrieval-based systems where those are more reliable. On multi-hop question answering tasks, incorporating a retrieval-based component outperformed chain-of-thought prompting, and adding an error-correction component produced accuracy gains of 14 to 17 percentage points on two math reasoning benchmarks (Khot et al., 2022). For synthetic UX participants, decomposition might mean separating the interpretation of the scenario from the identification of interface elements, the evaluation of user difficulties, and the formulation of the final response. Asking for a general evaluation in one step hides the assumptions behind the output. Decomposition makes them visible.

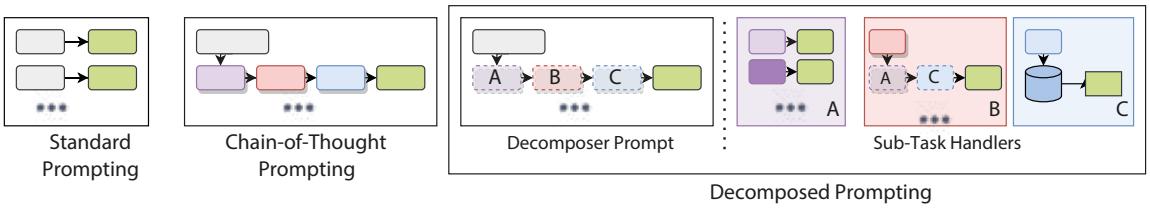

Figure 3.8: Standard prompting, chain-of-thought prompting, and decomposed prompting compared. Taken fromKhot et al. (2022).

Persona prompting. Persona prompting is one of the most common ways of structuring synthetic participants (Argyle et al., 2023; Gonzalez and DiPaola, 2024; Batzner et al., 2025). The model is asked to respond from the position of a particular user type, such as a novice user, an older adult, a patient, a student, or a domain expert. This practice has an obvious connection to personas in user-centred design. Traditional personas are design artefacts used to represent user groups and guide design decisions (Pruitt and Adlin, 2006). In LLM-based research, the persona becomes generative. A written profile now functions as a constraint from which the model produces new responses.

The appeal of persona prompting is clear. UX research is rarely interested in users as an abstract category. A healthcare application, an educational platform, or a financial service may each be experienced differently depending on prior knowledge, confidence, risk perception, digital literacy, or domain familiarity. Researchers use such specifications to examine whether a design may be interpreted differently by different user groups.

Persona prompting also introduces methodological risks. If the persona is defined only through demographic variables, the model may produce stereotypical or shallow outputs (Amidei et al., 2026; Bender et al., 2021). Age, gender, occupation, and education may all be relevant. None of them explains interaction behaviour on its own. In UX research, other characteristics are often more directly relevant: task knowledge and prior experience with similar interfaces, cognitive load, motivation, and digital competence, as well as trust and anxiety.

A prompt that describes a “65-year-old user” may produce less useful output than a prompt that specifies limited experience with online banking, concern about making irreversible mistakes, and low confidence in recognizing legitimate security cues.

This distinction becomes important in light of work on synthetic samples. Argyle et al. (2023) introduced the concept of algorithmic fidelity to describe how faithfully a language model captures the associations between human characteristics, beliefs, and behaviours that distinguish real subpopulations from one another. The concept goes

beyond asking whether outputs sound plausible. It asks whether outputs preserve meaningful patterns associated with human subpopulations. Algorithmic fidelity is a measure of pattern correspondence, not equivalence between a model and a human respondent. For UX research, the concept is insufficient on its own. UX evidence concerns interaction with a concrete artefact, not only distributional similarity in stated opinions.

Amirova et al. (2024) extend this discussion in the context of framework-based qualitative analysis of LLM-generated free responses. Their work shows that LLM outputs are shaped by the analytical framework imposed by the researcher. They are not free-standing interpretations. This is close to the problem of synthetic participants in UX research. The model generates an answer within a structure defined by the prompt, the assumed participant role, and the categories through which the response will later be interpreted. It is not simply answering as a user.

Task framing. Task framing is therefore a second major component of structured prompting. In UX research, the same interface can produce different problems depending on the task. A homepage may appear clear during general browsing. The same page can become confusing when the user is trying to cancel a subscription. A form may look acceptable during inspection. Under time pressure or with incomplete information, the same form can become problematic. A structured prompt should define the interface to be evaluated. It should also specify the goal and activity through which the user encounters it.

Contextual specification. Contextual specification is a third component. Real users do not interact with systems in a vacuum. They may be distracted, tired, under time pressure, using a mobile device, concerned about privacy, or dependent on the outcome of the task. Contextual information changes the interpretation of design problems. A minor delay acceptable in casual browsing may be serious in a medical or financial setting. Ambiguous wording tolerable in low-stakes exploration may be problematic when users are making consequential decisions. Without context, the model evaluates the interface as an abstract object. The situated conditions of use are absent from the output.

Output format. A further structuring strategy concerns output format. Asked for a general opinion, the model produces verbose output that is difficult to compare across runs. Asking it to identify issues, assign severity, justify each issue, distinguish evidence from speculation, or separate user reaction from design recommendation produces output that is far more analyzable. This is why many applied prompting approaches use tables, rubrics, stepwise instructions, or predefined categories (White et al., 2023).

Such formats improve interpretability and make the output easier to evaluate. They do not guarantee validity.

The difference between naïve and structured prompting is not simply a matter of prompt length. A long prompt can still be methodologically poor if it contains irrelevant detail or pushes the model toward a desired conclusion (Tang et al., 2025). A short prompt may be more useful when it specifies the task, user perspective, context, and output criteria. The quality of a prompt depends on whether it constrains the model in ways that are relevant to the UX question.

Section 3.2.1 discussed prompt sensitivity as a source of instability (Lu et al., 2022; Sclar et al., 2024). Structured prompting does not eliminate this problem. It makes the assumptions more explicit and therefore more open to inspection. In this sense, the prompt resembles other instruments in UX research, such as a moderator script, interview guide, or survey item. It shapes the response and must be documented if the study is to be interpreted or replicated.

Structured prompting works best when it is used to specify, in a controlled and documented manner, which user perspective, task situation, and response form the model is asked to approximate. The synthetic participant remains a generated output of a language model. The output can become more methodologically useful when the relevant assumptions are explicitly stated rather than left to the model's default patterns.

The distinction between naïve and structured prompting provides the basis for later synthesis. Naïve prompting treats the model as if it could spontaneously act as a user. Structured prompting treats the model as a system whose output must be constrained by task, user, context, and analytic purpose. The cognitive and experiential limits of synthetic participants remain. Structured prompting does provide a more rigorous starting point for evaluating when their outputs may be useful in UX research.

3.4 Four-Phase Framework

Synthetic participants cannot be produced by prompt wording alone. Before a model can generate a user-like response, the research situation has to be translated into its input format. A user is described, the interface is represented, the task is specified, and the model is instructed how to respond.

The four-phase transformation framework proposed here is not an established model in the literature. It is a conceptual tool developed for this thesis to describe, in simplified form, how synthetic participants are constructed in current research. It does not apply to naïve prompting, where the model is simply asked to act as a user in a single turn. Not all four phases are always present in every study. Depending on the

UX goal, some phases may be shortened, merged, or omitted.

The framework draws on three related areas. The first is synthetic samples and algorithmic fidelity (Argyle et al., 2023; Amirova et al., 2024). The second is structured prompting and chain-of-thought techniques (Wei et al., 2023; Wang et al., 2023; White et al., 2023). The third is recent work applying LLMs to UX evaluation through user profiles, interface representations, heuristic criteria, interaction traces, and standardized UX measures (Hsueh et al., 2024; Zhong et al., 2025b; Tan et al., 2026).

Across these studies, four operations appear repeatedly: defining the simulated user, grounding the model in an interface, conditioning the model's evaluative procedure, and executing the task. In this thesis, these operations are grouped into four phases: persona definition, architectural grounding, cognitive conditioning, and interaction execution.

Phase 1: Persona Definition. The first phase defines the user position from which the model is expected to respond. In simple cases, this may include demographic attributes such as age, occupation, or education. These variables are rarely sufficient on their own for UX research. Characteristics that shape interaction directly are often more relevant (Batzner et al., 2025; Hu and Collier, 2024): digital literacy, prior experience, domain knowledge, confidence, anxiety, cognitive load, and the perceived risk of making an error.

This phase relates to the idea of algorithmic fidelity introduced by Argyle et al. (2023). A language model is evaluated there according to its ability to reproduce relationships between human characteristics and response patterns. In UX research, the problem is more specific. The persona is useful only when its attributes matter for the task being studied. A "65-year-old user" is not methodologically meaningful unless age is connected to concrete interactional consequences, such as low confidence with online banking or difficulty recognizing security cues.

Persona definition helps constrain the model. It also introduces a risk. LLMs generate from textual patterns, so persona-based prompts may reproduce stereotypes rather than actual user diversity (Amidei et al., 2026; Batzner et al., 2025). The persona should be treated as a methodological assumption, not as evidence about a real user group.

Batzner et al. (2025) strengthen this point by distinguishing representativeness from ecological validity. In their review of 63 persona-based LLM studies, they found that the target population and task were often underspecified, and only 35% of studies discussed the representativeness of their LLM personae. For UX research, this means that a persona is not valid in general. It is only more or less appropriate for a specific product, task, user group, and context of use (Batzner et al., 2025).

Persona definition should therefore be reported as part of the research instrument.

A UX study using synthetic participants should specify how the persona was constructed, which population it is intended to approximate, which UX task it is used for, and which limits it has (Batzner et al., 2025).

Phase 2: Architectural Grounding. The second phase concerns how the interface is made available to the model. A human participant sees, scans, clicks, hesitates, and reacts to an interface. A language model receives a representation of it. This representation may be a textual description, screenshot, HTML structure, DOM tree, Markdown conversion[2], design specification, or interaction log (Hsueh et al., 2024; Tan et al., 2026).

The quality of this representation strongly affects the output. An interface described only in text allows the model to reason about labels, content, hierarchy, and likely task flow. Visual salience, contrast, spatial relations, motor difficulty, and dynamic behaviour across screens remain inaccessible (Tan et al., 2026). Multimodal inputs or structured interface data can improve grounding, but they still do not reproduce human perception.

Hsueh et al. (2024) and Zhong et al. (2025b) show why this phase matters. More rigorous interface grounding makes synthetic evaluation more useful for detecting visible usability issues, especially layout and structure problems. It may also lead to false positives: issues that are plausible from the represented interface but would not necessarily become relevant problems for human users (Hsueh et al., 2024). Architectural grounding improves access to the interface and determines the limits of what the model can notice.

Phase 3: Cognitive Conditioning. The third phase defines how the model is instructed to reason or report. In UX contexts, this may involve (Hsueh et al., 2024; Zhong et al., 2025b; Tan et al., 2026): applying heuristics, simulating a think-aloud protocol, reasoning step by step, assigning severity, reporting uncertainty, or separating observations from recommendations.

Without such instructions, the model defaults to a generic expert-style evaluation rather than a simulated user perspective.

This phase draws on structured prompting methods such as chain-of-thought prompting (Wei et al., 2023), self-consistency (Wang et al., 2023), and prompt patterns (White et al., 2023). These methods show that outputs are shaped not only by the task content but also by the requested reasoning format. In UX simulation, comparable instructions

[2]HTML structure refers to the underlying markup of a webpage, while the DOM tree is the browser's parsed and dynamically updated representation of that structure. Markdown conversion simplifies interface content into a readable text-based format that preserves elements such as headings, lists, links, and basic hierarchy.

make the model state what it notices, expects, misunderstands, or finds confusing.

Such reasoning is not equivalent to human cognition. A generated think-aloud trace is not the same as a human participant verbalizing thoughts during interaction (Zhong et al., 2025a). The model may produce uncertainty, hesitation, or frustration as text, but these states do not arise from perception, emotion, or task pressure (Hämäläinen et al., 2023). Cognitive conditioning makes the output easier to inspect. It may also make it cleaner and more rational than actual UX behaviour (Zhong et al., 2025a).

Phase 4: Interaction Execution. The final phase concerns how the task is performed. In single-turn execution, the model receives a persona, task, and interface representation and produces one response. This is useful for quick evaluation of static screens, content clarity, visible inconsistencies, or likely usability problems. The whole interaction is compressed into one generation.

Multi-turn execution is more complex. The model receives an interface state, chooses an action, receives the next state, and continues through the task. This makes it possible to examine navigation, task progression, recoverability, Task Success Rate, navigation efficiency, and sometimes standardized usability measures (Zhong et al., 2025a; Tan et al., 2026). Avenir-UX (Tan et al., 2026) combines simulated user profiles, interaction traces, usability metrics, and think-aloud-style reporting in this manner.

This does not make the interaction equivalent to human behaviour. A model may complete a task through a path that a human user would not take, or fail because of limitations in element grounding rather than because the interface is difficult (Zhong et al., 2025a; Tan et al., 2026). Multi-turn execution produces richer synthetic data. The resulting trace still has to be interpreted as model-mediated evidence, not as direct behavioural evidence from a user.

The four phases show that synthetic participant outputs are shaped by a sequence of methodological decisions: how the user is described, how the interface is represented, how the model is instructed to reason, and how the task is executed. Studies using synthetic participants may reach different conclusions for this reason. A single-turn prompt based on a short persona description is not methodologically equivalent to a multimodal, multi-turn system that produces interaction traces and usability metrics.

The framework sets a limit on what can be expected from prompting. Better prompts can reduce ambiguity. They cannot compensate for poor interface grounding, inappropriate task execution, or experiential dimensions that were never available to the model. Synthetic participants may generate useful UX evidence. That evidence is always mediated by the transformations through which the research situation was reconstructed.

3.5 Empirical Evidence

Empirical research on synthetic participants is still relatively recent and uneven. The studies discussed in this area do not all test the same claim. Some examine whether language models can approximate survey respondents or qualitative participants. Others use LLMs as usability evaluators, design assistants, or agents interacting with interfaces. These are different roles. They produce different types of evidence and should not be interpreted as one unified proof that synthetic participants can replace human users.

The current evidence is best read as a set of partial findings. LLMs are most useful when the task is explicit, structured, and mediated through language or interface representations. They are less useful when the target phenomenon depends on situated behaviour, affect, embodiment, or long-term experience.

The development of LLM-based synthetic participants follows a recognizable pattern. Early work focused on whether language models could approximate human responses in surveys or qualitative settings. Studies such as Argyle et al. (2023), Hämäläinen et al. (2023), and Amirova et al. (2024) treated the model primarily as a simulated respondent. The main question was whether the generated answers resembled human answers.

Later work moved closer to design and UX practice. Studies began to test whether the model can inspect interfaces, generate usability feedback, or support early design work. This is visible in work on automated UI feedback (Duan et al., 2024), LLM-based UX testing (Hsueh et al., 2024), persona-based chatbots (Gu et al., 2025), and synthetic heuristic evaluation (Zhong et al., 2025b). The model became an evaluative tool embedded in a design process.

The most recent direction is more agentic. Systems such as UXAgent (Lu et al., 2025a) and Avenir-UX (Tan et al., 2026) simulate not only verbal judgement but also sequences of actions across an interface. This is a shift from synthetic responses to synthetic interaction. The problem remains. More complex systems produce richer traces. Those traces are still model-mediated, shaped by prompt design, interface representation, grounding, and the assumptions built into the agent architecture.

The field is not moving linearly toward replacing users. It is moving toward more structured approximations of selected UX evidence. LLMs have become more useful for early inspection, hypothesis generation, and protocol preparation. The question of lived experience remains unresolved.

LLMs as Simulated Respondents. One influential starting point is the work of Argyle et al. (2023), who introduced the concept of algorithmic fidelity in the context of silicon samples. Their study examined whether GPT-3 could generate responses cor-

responding to specific human subpopulations when conditioned on socio-demographic backstories. The methodological claim matters more than the political science context: a language model may preserve some relationships between human attributes and response patterns.

For UX research, this is insufficient. UX evidence concerns more than stated attitudes. It often concerns interaction with a concrete artefact, interpretation of interface cues, situated constraints, and the consequences of action. A model may approximate the language of a particular user group. The conditions under which that group interacts with a system are a different matter.

A similar tension appears in Hämäläinen et al. (2023), who compared human responses with GPT-3 outputs in the context of generating synthetic HCI research data. Their results suggest that synthetic responses can appear convincingly human at the surface level. The distribution of examples and associations is narrower than in human data. Human participants produced a wider range of game examples. GPT-3 outputs tended to converge around culturally prominent references. A model may produce plausible qualitative material while reducing the diversity, idiosyncrasy, and personal specificity that often characterize human responses.

Amirova et al. (2024) provide a further caution. In their framework-based qualitative analysis of LLM-generated free responses, human and silicon participants produced some similar themes. They differed in structure, tone, and the representation of lived experience. The authors describe problems such as hyper-accuracy distortion, where generated responses appear unusually coherent, complete, or analytically convenient. A model-generated response may look like qualitative user data. It lacks the unevenness, partiality, and situated character of real participant accounts.

Synthetic respondents can be useful for exploring possible response patterns. They are weaker as evidence of human experience itself. A synthetic response may resemble a human statement. It is not grounded in a human situation.

LLMs as Usability Evaluators. More directly relevant evidence comes from studies applying LLMs to usability evaluation. Hsueh et al. (2024) developed an automated UX testing tool based on LLMs. Their work starts from a practical problem: traditional UX testing is costly, time-consuming, and affected by subjectivity. The study shows that LLMs can help structure UX evaluation and generate usable reports. The method remains dependent on how the task, interface, and evaluation criteria are represented to the model.

Duan et al. (2024) provide a concrete example of this dependence. As illustrated in Figure 3.9, their system takes a Figma UI mockup, translates it into a structured JSON representation, and passes it to GPT-4 for evaluation against a selected set of design guidelines — returning results as actionable suggestions directly within the design tool.

In a study rating feedback generated for 51 distinct interfaces, 52% of suggestions were judged accurate and 49% helpful or very helpful (Duan et al., 2024). Performance varied by guideline type: the system was most reliable on rule-based checks where the information needed for evaluation was directly available in the interface representation, and least reliable on guidelines requiring aesthetic or holistic judgement. A further limitation emerged across iterative design rounds — feedback quality declined as the design improved, with the majority of the 12 expert participants describing the tool as more useful when applied once to an early design than when used across multiple revision rounds. Where the structured interface representation failed to capture what a human viewer would actually perceive, evaluation quality dropped accordingly.

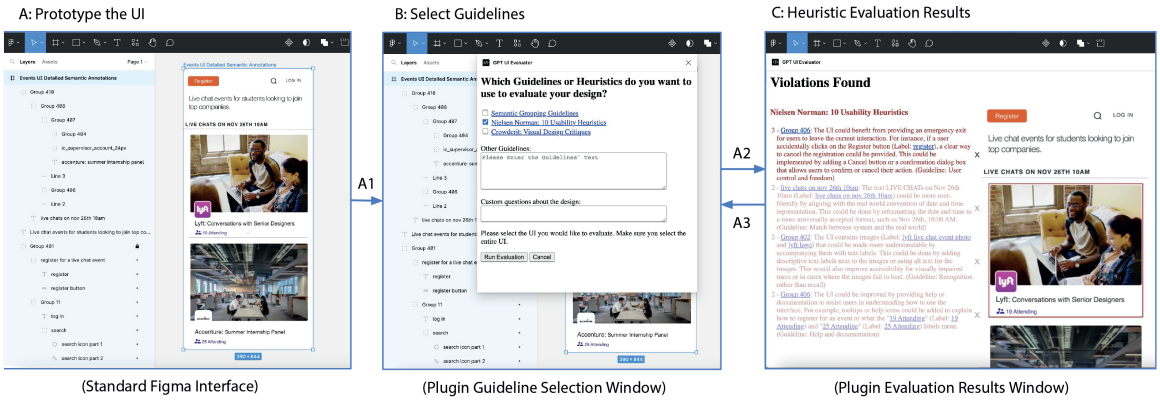

Figure 3.9: Automated UI feedback pipeline using a Figma plugin. Taken fromDuan et al. (2024).

The model does not evaluate the interface itself in the human sense. It evaluates a mediated version of it: a description, screenshot, DOM structure, mockup representation, or interaction trace. The quality of that representation determines what the model can notice. A failure in the representation may therefore become a failure in the evaluation.

Zhong et al. (2025b) provide a more specific comparison in their study of synthetic heuristic evaluation. Using multimodal LLMs, they compared AI-powered usability evaluation with evaluations conducted by experienced human practitioners. Their GPT-4-based system identified between 73% and 77% of usability problems across two applications, while the five experienced human evaluators identified 57% and 63% respectively. As shown in Figure 3.10, the model's detection rate held steady regardless of how many tasks had already been assessed, while human evaluators identified fewer issues as the session progressed — a pattern the authors link to the demands of sustained expert attention over time. Its strongest performance appeared in layout-related and structurally visible usability problems.

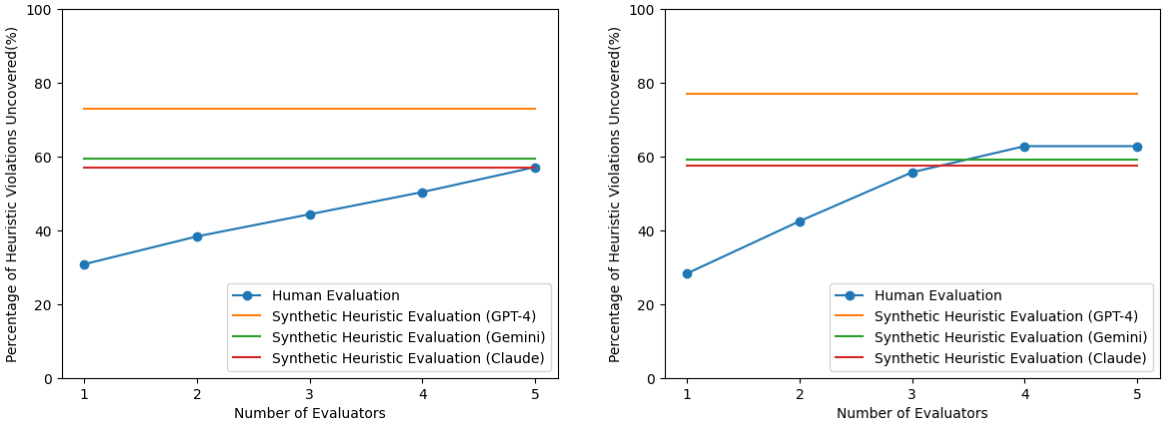

Figure 3.10: Coverage of usability issues by synthetic heuristic evaluation and human evaluators across two applications. Taken fromZhong et al. (2025b).

The same study also shows the limits of this approach. Synthetic evaluation struggled with some UI components, design conventions, and violations that extended across screens. It also raises the problem of false positives. A model may detect a plausible heuristic violation that is not necessarily experienced as a problem by human users. The finding is methodologically important. It does not establish that the model evaluates the interface in the same way as a human practitioner or participant. In heuristic evaluation, a model may be useful as an additional inspector for early detection of visible issues. Its findings still require interpretation and, in many cases, human validation.

A later study on synthetic cognitive walkthrough by Zhong et al. (2025a) points in the same direction. LLMs navigated interfaces and provided reasonable rationales. Their behaviour differed from human participants. They achieved higher task completion rates and followed more optimal paths. They identified fewer potential failure points. A model can perform well according to some metrics and still fail to reproduce human behaviour. Higher success does not necessarily mean better simulation. It may indicate that the model approaches the interface more systematically than a human user.

LLMs in Early Design Work. LLMs are also being studied as tools in earlier stages of the design process, before formal usability testing begins. Schmidt et al. (2024) discuss this problem at the level of human-centred design more broadly. They describe several points at which LLMs may enter design work: generating personas, producing scenarios, supporting early ideation, and conducting heuristic evaluation through AI-based agents. Their argument concerns the redistribution of user-related work across the design process.

Synthetic users may appear before direct user involvement becomes available. They

can help designers formulate assumptions, imagine possible user concerns, or inspect design alternatives. In this role, they support parts of design work that already rely on assumptions, prior knowledge, and early interpretation. Their value lies in widening the range of questions designers ask.

Lu et al. (2025b) extend this early-stage use case through UXAgent, an LLM agent-based framework for simulating usability testing of web designs. As shown in Figure 3.11, the system connects a persona generator, an LLM agent, and a browser interface, producing action traces, reasoning logs, video recordings, and a chat interface through which researchers can query individual simulated users in natural language. The authors position UXAgent as a preparation tool rather than a substitute — a way to stress-test a study design before it involves real participants, catching unclear task instructions or obvious interaction problems at an earlier stage. In an evaluation with five UX researchers, participants valued the system's ability to generate large-scale behavioural data rapidly, but raised concerns about model bias, stereotypical persona behaviour, and the gap between recorded memory traces and realistic human navigation patterns (Lu et al., 2025b). Participants treated the outputs as provisional material to inform study design, not as a stand-in for real user data.

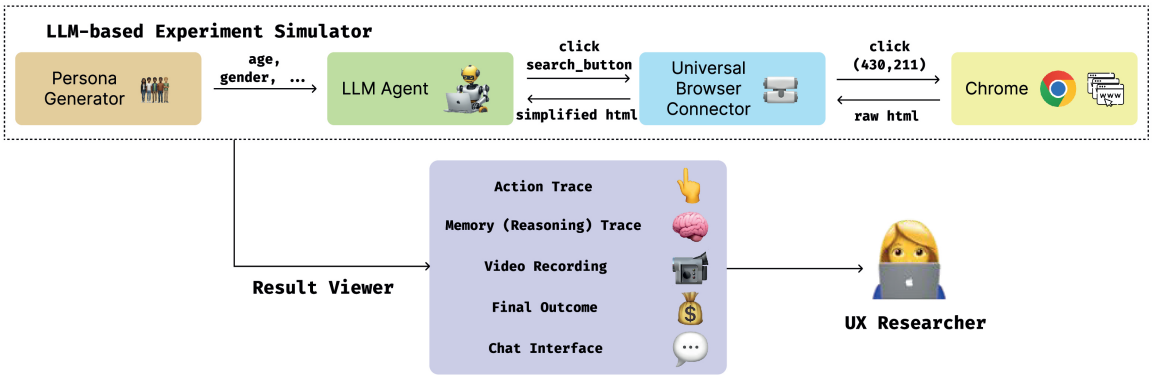

Figure 3.11: UXAgent system design. Taken fromLu et al. (2025b).

UXAgent illustrates a shift from synthetic responses to synthetic interaction. It shows how LLM-based systems may support researchers before real-user studies by surfacing early issues, testing experiment design, and producing provisional feedback. The system does not remove the need for human validation. The traces it produces remain model-mediated and depend on personas, task framing, browser access, and the assumptions built into the agent architecture. UXAgent strengthens the practical case for synthetic participants as preparatory tools. It does not support the claim that they can replace the perspectives of real users.

Gu et al. (2025) examine this issue more directly through designers' interactions

with persona-based chatbots. Their work treats synthetic users as design artefacts that influence how designers reason about users. A chatbot persona can make a persona feel more interactive and available during the design process. It may help designers explore user concerns, generate questions, and challenge early assumptions.

Such interaction may also create a false sense of user contact. A designer may feel that they have "spoken" with a user. The response still comes from a model conditioned by persona data and prompt structure. The value of synthetic users in design practice lies in supporting reflection and ideation.

Synthetic participants can make user perspectives more present during early design stages. They can help researchers and designers prepare better questions before a study with real participants. The evidence does not justify treating them as a complete empirical source. Their usefulness depends on whether they are positioned as exploratory aids, evaluative assistants, or substitutes for human data.

LLMs as Interaction Agents. Recent agentic systems attempt to move beyond static responses. UXAgent, developed by Lu et al. (2025a), is an example of this direction. The system uses LLM-based agents equipped with personas, task descriptions, and browser access to simulate usability testing on websites. It can produce action traces and reasoning traces, which makes it more process-oriented than a single-turn prompt. Researchers can inspect not only what the model concludes but also how it moves through the task.

The practical value of such systems may lie partly in study preparation. Synthetic agents can help reveal unclear task instructions, weak prototypes, or obvious interaction problems before real users are recruited. They may reduce wasted effort in later human testing. The traces they produce are not necessarily human-like. They may be too detailed, too systematic, or too rational compared with actual user behaviour (Zhong et al., 2025a). Agentic synthetic users may help refine research protocols. They do not remove the need for human validation.

Avenir-UX, presented by Tan et al. (2026), similarly simulates user behaviour on websites through end-to-end interaction and multimodal grounding. The system combines simulated user profiles, interaction traces, the System Usability Scale (SUS)[3], step-wise Single Ease Questions (SEQ)[4], and think-aloud-style reporting. This represents a shift from asking a model to comment on an interface toward asking it to act through a sequence of interface states.

Many UX problems are temporal. They emerge during navigation, recovery from

[3]System Usability Scale, a ten-item questionnaire in which users rate statements about a system and the responses are converted into an overall usability score.

[4]Step-wise Single Ease Question, an adapted form of the Single Ease Question (SEQ) in which users rate each individual step of a task instead of the completed task as a whole.

error, failed expectations, or repeated attempts. Multi-turn systems can therefore produce richer evidence than single-turn prompting. They may record task success, navigation efficiency, and points of failure. This makes them useful for evaluating flows, not only static screens.

Agentic execution introduces its own ambiguity. A model may fail because it cannot ground an element correctly. The interface may not be difficult for a human at all. It may also complete a task through a path that is more optimal than a human user's path. The resulting trace is therefore not equivalent to behavioural data. It is a model-mediated reconstruction of behaviour, shaped by the interface representation, agent architecture, and prompt design.

Summary of the Evidence. The current evidence supports a meaningful but limited role for synthetic participants in UX research. Their strongest performance appears in tasks where the relevant evidence is explicit, verbal, structural, or rule-based. This includes: identifying visible usability issues, evaluating content clarity, applying heuristics, generating possible user concerns, comparing design alternatives, supporting early-stage inspection, and refining study materials before human testing.

Sarstedt et al. (2024) make a similar point in relation to silicon samples more generally. They argue that synthetic samples are most defensible in upstream parts of the research process, such as pretesting, pilot studies, identifying ambiguous formulations, or preparing research instruments. Synthetic participants are most useful where their outputs can be checked, revised, or used to prepare later empirical work. They are less defensible when treated as a complete substitute for human evidence in the main study.

The evidence is weaker when the task requires access to experience rather than representation. LLMs can describe frustration, uncertainty, trust, or confusion. Empirical studies show that such descriptions are often cleaner, more coherent, or more systematic than human data (Amirova et al., 2024; Zhong et al., 2025a). They may correspond to human responses at the level of output. The cognitive or experiential process behind them differs.

Reported success rates should be interpreted carefully. Agreement between synthetic and human outputs may be practically useful. It does not automatically establish simulation. A model may identify the same issue as a human participant because the issue is visible, linguistically obvious, or aligned with known usability principles. This is valuable evidence for design inspection. It is not, by itself, evidence that the model has reproduced human perception or experience.

Synthetic participants are most defensible as tools for early filtering, hypothesis generation, structured inspection, and preparation for human research. They may reduce the number of trivial issues that reach expensive human testing. They may

also help researchers explore alternative perspectives before recruitment. They remain insufficient for validating affective, embodied, longitudinal, or high-stakes UX claims.

The synthesis developed in Chapter 4 builds on this pattern. The central issue is which UX tasks require forms of evidence that can be approximated by language models, and which tasks require human participants because the relevant evidence depends on lived, situated, or embodied experience.

Chapter 4

Synthesis and Discussion

The previous chapters reviewed user experience research, the emergence of synthetic participants based on large language models, and the empirical evidence on their use in UX evaluation. One question appears across this literature, in different forms: can synthetic participants replace human users?

The question seems straightforward. Most existing studies compare outputs from synthetic participants with data from human users and measure the degree of similarity. When synthetic participants identify comparable usability issues, produce similar ratings, or reach conclusions that resemble those of human participants, such findings are interpreted as evidence that large language models can successfully simulate users. The discussion tends to revolve around how closely synthetic outputs correspond to human responses.

This framing, however, obscures a more fundamental issue. The reviewed literature demonstrates considerable variation in the types of UX tasks for which synthetic participants appear successful. Language models often perform surprisingly well in tasks involving textual interpretation, information structure, or identification of apparent usability problems. Their performance becomes substantially more uncertain when tasks depend on emotions, embodied interaction, contextual experience, or implicit knowledge. This suggests that the more precise question is which aspects of user cognition can be approximated through prompting and under what conditions such approximations remain methodologically useful.

This follows from the account of UX in Chapter 2, where usability-oriented tasks were separated from broader experiential and contextual dimensions of user experience. It also follows from the empirical studies reviewed in Chapter 3, where synthetic participants performed best on explicit, structural, or rule-based UX tasks and were less reliable when the target phenomenon involved lived experience.

Human behaviour emerges from a combination of goals, expectations, prior experiences, cognitive limitations, emotional states, bodily interaction with technology, and

the situational context in which technology is used. Large language models do not possess these characteristics in the same manner. They may nonetheless generate outputs that are useful for particular UX purposes. This is why the validity of synthetic participants cannot be reduced to a simple comparison between human and model-generated responses.

The reviewed literature suggests that validity depends on whether the cognitive resources required by a particular UX task can be sufficiently approximated through prompting. Usefulness is therefore a function of the relationship between the prompt, the task being performed, and the type of cognitive processes required by that task. It is a characteristic of a specific research situation rather than of the model itself.

The following discussion develops this argument in four steps. First, the chapter examines which prompt characteristics appear necessary for producing methodologically useful synthetic outputs. Second, it discusses how correspondence between synthetic and human responses should be interpreted and whether such correspondence constitutes meaningful simulation. Third, it identifies the cognitive and methodological limits that constrain what can be achieved through prompting. Finally, these insights are integrated into a task-oriented framework that proposes how synthetic and human participants may be combined within UX research.

4.1 Components of an Effective UX Prompt

The use of synthetic participants rests on a central assumption: model outputs can be substantially influenced through prompting. A significant portion of recent literature has therefore focused on how large language models should be instructed to generate methodologically useful outputs. Increasingly elaborate prompting strategies have emerged across studies. These include detailed personas, multi-stage instructions, contextual scenarios, behavioural constraints, and various forms of role-playing. The underlying expectation is that more detailed prompts produce more realistic simulations.

As discussed in Chapter 3.3, this assumption underlies the move from naive prompting to more rigorous prompting pipelines. Chapter 3.2.1 also showed, however, that model outputs remain sensitive to prompt formulation, stochastic generation, and evaluation procedure.

The reviewed evidence suggests a more nuanced conclusion. Prompt structure clearly matters, but effectiveness does not depend on the quantity of information provided. Additional details frequently produce diminishing returns. Poorly chosen details can introduce unwanted assumptions and reduce consistency. The relevant question is whether the prompt contains the information necessary to reconstruct the aspects of

user behaviour relevant to the UX task under investigation.

Minimal Sufficient Prompt. From this perspective, it is possible to formulate what may be called a *minimal sufficient prompt*. Such a prompt provides only the information necessary to situate the model within a meaningful UX scenario. Three components appear particularly important: a user goal, a UX task, and a context of use.

- The prompt must specify the *goal* that motivates the user's behaviour. Human interaction with interfaces is goal-directed. Users act in pursuit of particular outcomes: completing a purchase, finding information, booking a service, or solving a practical problem. Without an explicit goal, the model lacks a basis for evaluating interface elements in terms of their relevance to task completion.
- The prompt should define the *UX task* itself. Goals describe desired outcomes; tasks specify the concrete actions through which these outcomes are pursued. The same interface may be experienced differently depending on whether the user is exploring, comparing alternatives, troubleshooting a problem, or completing a transaction. Task specification constrains the perspective from which the synthetic participant evaluates the interface.
- The prompt should provide a *context of use*. User experience is shaped by situational conditions, including where, when, and under what circumstances the interaction occurs. A user completing a task in a quiet environment with unlimited time may behave differently from a user attempting the same task while commuting or under time pressure. Contextual information allows the model to generate evaluations grounded in a plausible usage situation.

Three Layers of Prompt Grounding. Beyond these minimum requirements, the literature suggests that prompt elements differ considerably in their methodological value. Some additions improve the realism and diversity of synthetic outputs; others contribute little to the task at hand. The findings reviewed in this thesis can be synthesized into a framework of three interconnected layers of prompt grounding: task grounding, user grounding, and context grounding. Figure 4.1 maps this structure as a tree: a single UX prompt node at the top branches into three parallel dimensions—Task, User, and Context—each of which is unpacked into a diagnostic question and a functional role. Task asks what the model is required to do and defines the goal of the evaluation; User asks who is being simulated and defines the relevant characteristics of that person; Context asks under what conditions the interaction takes place and defines the situational constraints on use. Together, the three branches specify what counts as a well-grounded prompt in UX research.

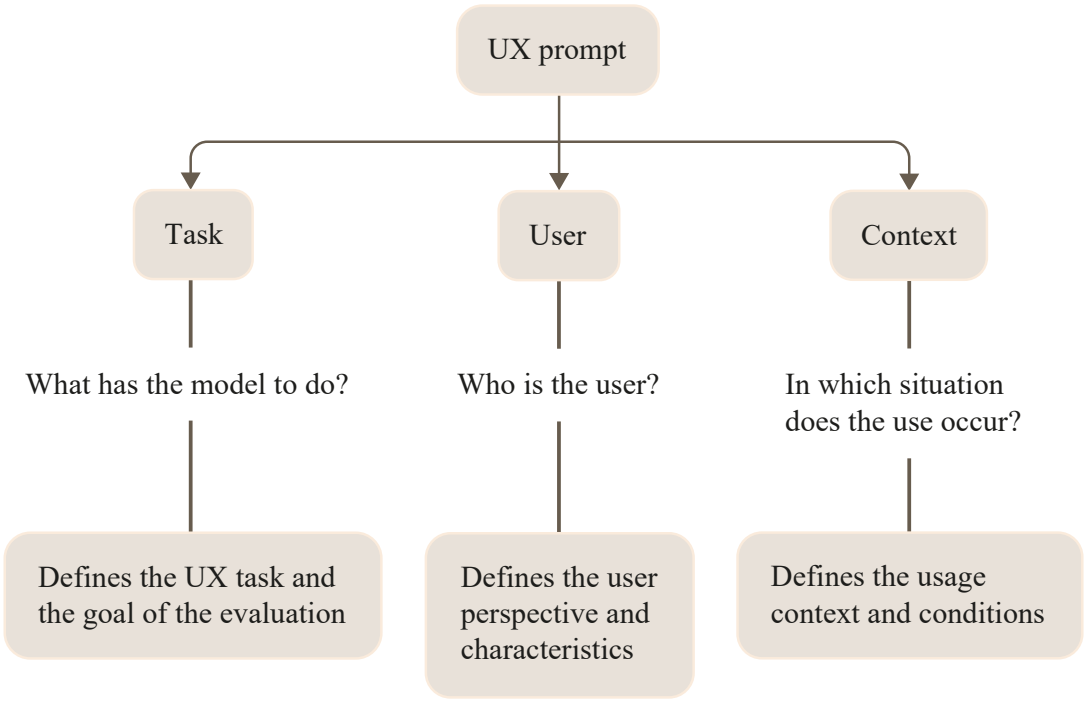

Figure 4.1: Three Layers of Prompt Grounding.

Task Grounding. The first layer, task grounding, concerns the activity that the synthetic participant is expected to perform. It includes the user's objective, the specific task to be completed, and any constraints that shape successful performance. Task grounding determines what counts as relevant information and what constitutes a successful interaction. Existing studies consistently suggest that this layer exerts the strongest influence on output quality because it defines the problem space within which the model reasons.

User Grounding. The second layer, user grounding, concerns characteristics attributed to the simulated user. Existing prompting approaches frequently rely on demographic personas defined by variables such as age, occupation, or educational background. The reviewed literature indicates that cognitive and behavioural attributes are often more consequential. Digital literacy, domain knowledge, prior experience, confidence, cognitive load, and familiarity with similar systems shape interaction patterns more directly than demographic characteristics. Effective UX prompting therefore represents user-relevant cognitive differences rather than demographic profiles.

This builds on Chapter 2.3, where real user evidence was shown to depend not only on demographic sampling but also on mental models, prior experience, literacy, and

task interpretation.

Context Grounding. The third layer, context grounding, refers to the broader circumstances surrounding interaction. This includes environmental conditions, available time, competing demands for attention, device constraints, and other situational factors. Many usability problems only emerge under realistic conditions. An interface that appears straightforward during idealized evaluation may become significantly more challenging when users are distracted, rushed, or operating in unfamiliar environments. Context grounding moves synthetic evaluation closer to the conditions under which actual use occurs.

Taken together, these three layers reframe effective UX prompting as a process of selectively reconstructing the aspects of user behaviour relevant to a particular research objective. Methodologically useful synthetic outputs emerge from adequate grounding of a task, a user perspective, and a usage context.

This conclusion provides a partial answer to RQ1. Prompts become methodologically useful when they provide sufficient information for reconstructing the cognitive and situational conditions relevant to a UX task. The most effective prompts are those that successfully ground the model in the specific user, task, and context that the researcher seeks to investigate.

4.2 Interpreting Human-Synthetic Correspondence

The empirical studies reviewed in Chapter 3.5 frequently evaluate synthetic participants by comparing their outputs with those generated by human users. Similarity is expressed through agreement rates, overlap in identified usability issues, comparable ratings, or convergence of qualitative feedback. Such findings are interpreted as evidence that synthetic participants can successfully simulate human behaviour.

The meaning of this correspondence is not self-evident. A high degree of agreement between synthetic and human outputs does not automatically indicate that both arrived at the same conclusion in the same way. It does not necessarily imply that the model reproduced the underlying user experience. Before assessing the practical value of synthetic participants, it is therefore necessary to clarify what kind of correspondence is actually being observed.

The question is fundamentally epistemological. It concerns the relationship between observed outputs and the phenomena they are assumed to represent. If a synthetic participant identifies the same usability problem as a human participant, what exactly does this tell us about the validity of the simulation?

Three Levels of Human-Synthetic Correspondence

Three distinct forms of correspondence may occur between synthetic and human participants. As shown in Figure 4.2, these are arranged along a vertical axis by strength of claim. Output correspondence sits at the base: synthetic and human outputs resemble each other, but the similarity is purely descriptive and says nothing about how either arrived at their response. Functional correspondence occupies the middle: different wording or reasoning paths nonetheless converge on the same underlying design problem, making the outputs useful for practical evaluation purposes. Cognitive correspondence is the strongest form: both parties reach similar conclusions through comparable processes—shared expectations, reasoning patterns, or interpretations of the interface.

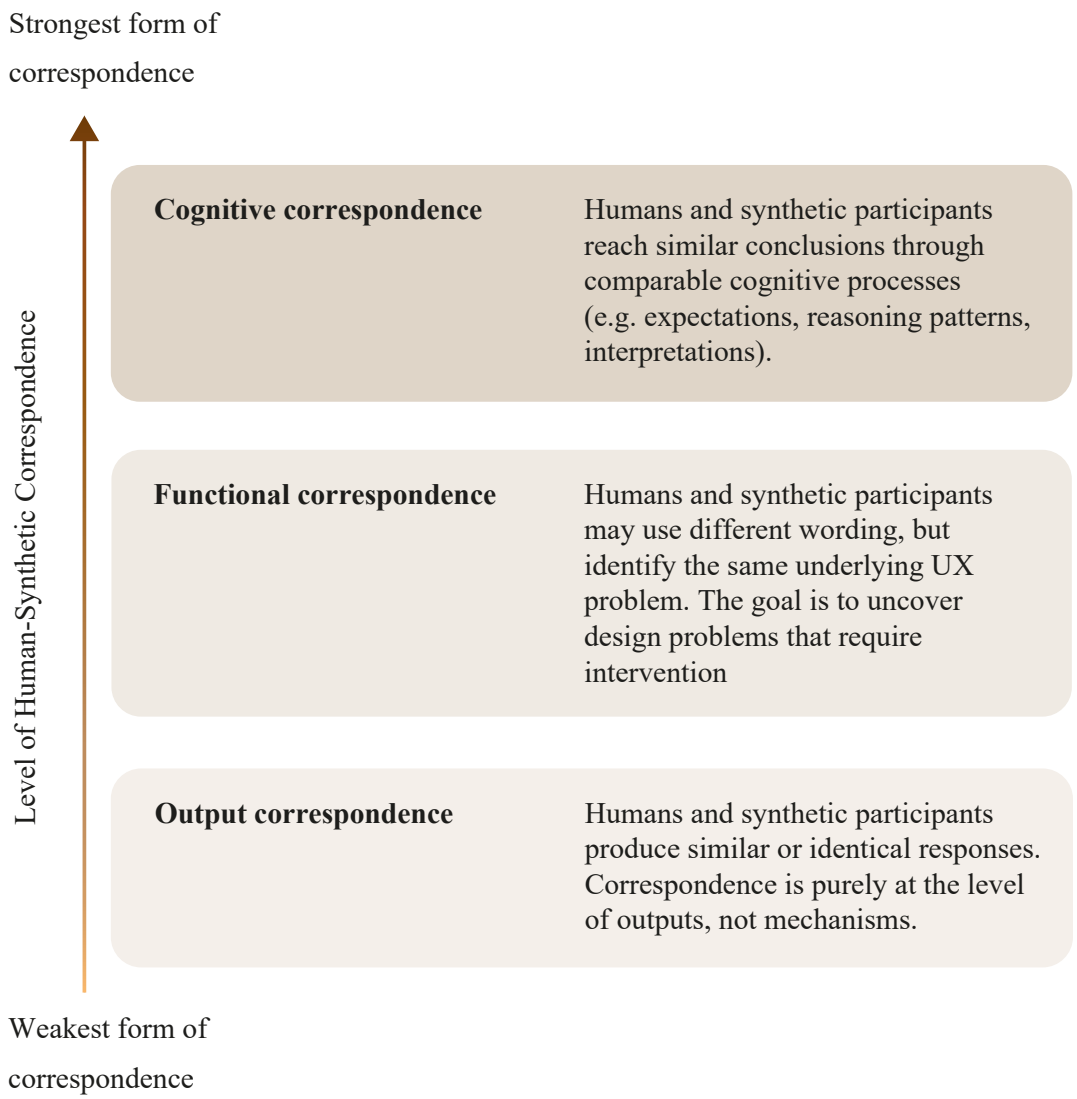

Figure 4.2: Three Levels of Human-Synthetic Correspondence.

- The first level is *output correspondence*. This occurs when humans and synthetic participants produce similar or identical responses. Both may report difficulty locating a registration button or express confusion about the navigation structure of an interface. At this level, correspondence is purely descriptive. The outputs resemble one another, but no claim can be made regarding the mechanisms that produced them.

- The second level is *functional correspondence*. The specific wording of the response becomes less important than its practical consequence. Human participants and synthetic participants may use different language while identifying the same underlying UX problem. One participant may report that they "cannot find where to continue", while a language model may describe the issue as a lack of visual hierarchy or poor information architecture. The statements differ; both point to the same design deficiency. Many UX methods aim to uncover design problems that require intervention. Reproducing users' exact statements is secondary to this objective. From this perspective, functional correspondence carries substantial methodological value.

- The third and strongest form is *cognitive correspondence*. This would require evidence that human and synthetic participants reached their conclusions through comparable cognitive processes—similar expectations, reasoning patterns, interpretations of the interface, and potentially similar sources of error. Demonstrating this type of correspondence is substantially more difficult because the internal processes of human users and language models differ in fundamental ways.

Current evidence provides relatively strong support for output correspondence and, in some contexts, for functional correspondence. Evidence for cognitive correspondence remains considerably weaker.

The Problem of Interpreting Agreement

This distinction becomes important when interpreting empirical findings. Many studies report that synthetic participants identify a substantial proportion of the usability problems discovered by human participants. Such findings are valuable. They demonstrate that a model was capable of producing outputs that overlap with human findings. Demonstrating successful user simulation is a stronger claim that this evidence does not support.

Zhong et al. (2025b) provide a useful example. Their GPT-4-based synthetic heuristic evaluation identified between 73% and 77% of usability problems across two applications, exceeding the rates achieved by experienced human evaluators, who identified

57% and 63% respectively. The system also showed greater stability across time and accounts. This is an important empirical result, especially for early usability inspection. The central question, however, is what the number measures. A high detection rate shows that the model can identify many of the same visible or structurally inferable usability problems as human evaluators. It does not show that the model encountered the interface as a user would.

If a synthetic participant flags the same navigation problem as a human participant, the finding may still be practically useful. The design issue may be real, and it may deserve further investigation or redesign. The basis of the finding, however, is different. The human participant encounters the navigation problem through perception, expectation, task effort, hesitation, and possible failure. The language model identifies the problem through a generated response conditioned by the prompt, the interface representation, and learned regularities about usability, design conventions, and common interface problems. The outputs may converge. The processes that produced them remain different.

This observation also helps explain why synthetic participants often perform well in tasks involving heuristic evaluation, interface critique, information architecture, or usability inspection. Such tasks reward the identification of known design patterns and common usability issues—areas where language models have access to substantial amounts of textual and design-related knowledge. Agreement in these cases is evidence of successful problem identification. Equating this with successful user simulation is a stronger claim than the data support.

The same caution applies to more agentic systems such as UXAgent. Producing simulated interactions, behavioural logs, or qualitative accounts makes the synthetic participant more useful than a single-turn prompt. By itself, this does not establish cognitive correspondence. The interaction trace remains model-mediated and depends on the persona, task framing, interface access, and assumptions built into the agent architecture. Agentic systems may improve the practical usefulness of synthetic participants. The need to ask what kind of evidence their outputs actually provide remains.

Correspondence and the Question of Validity

The distinction between different forms of correspondence also affects how validity should be understood in synthetic UX research.

Traditional user research methods derive their validity from direct observation of human behaviour, experiences, and interpretations. Synthetic participants operate differently. Their validity cannot be based on authentic experience because no such experience exists.

The validity of synthetic participants must therefore be evaluated in relation to the

purpose of the research task. If the objective is to identify common usability problems, functional correspondence may be sufficient. If the objective is to understand emotional responses, trust formation, frustration, decision-making processes, or situated behaviour, functional correspondence may no longer provide adequate evidence.

The methodological value of synthetic participants depends on the amount of agreement observed and on the level at which that agreement occurs.

From Correspondence to Simulation

The literature frequently treats agreement and simulation as interchangeable concepts. The analysis presented in this thesis suggests that such an interpretation may be too strong.

Observed correspondence demonstrates that synthetic participants can sometimes generate outcomes similar to those produced by human participants. Similarity of outcomes does not necessarily imply similarity of underlying processes. In most cases, existing evidence supports claims about output correspondence or functional correspondence. Cognitive correspondence remains a substantially harder bar to clear.

The term *simulation* should therefore be used cautiously. The reviewed literature provides growing evidence that synthetic participants can reproduce certain findings obtained through human research. At present, this is evidence of functional similarity. Proof that language models replicate human cognition or human experience requires substantially stronger support.

This distinction provides a more precise interpretation of existing empirical results and offers a clearer framework for evaluating future studies involving synthetic participants. It also helps explain why synthetic participants may prove highly useful for some UX tasks and inadequate for others—a question addressed in the following sections.

This distinction also requires caution in the language used to describe synthetic participants. Shanahan (2024) argues that it is misleading to describe LLMs as systems that know, believe, or intend in the same sense as humans. A belief requires an epistemic relation to the world: the capacity to update what is held to be true in response to evidence and to distinguish truth from falsehood by checking representations against reality. LLMs generate text that resembles the expression of belief. Possessing the underlying relation to the world that belief requires is a separate matter (Shanahan, 2024).

Schröder et al. (2025) make a related point from the perspective of psychological measurement. They show that LLM outputs may change substantially when psychologically similar questions are phrased in slightly different ways. Models appear more sensitive to linguistic surface form than to the underlying psychological construct. For UX research, this means that synthetic agreement with human responses should not be

treated as stable evidence of the same construct unless the model has been validated in the specific task context (Schröder et al., 2025).

Quattrociocchi et al. (2025) describe this broader condition as *Epistemia*: a situation in which linguistic plausibility substitutes for epistemic evaluation. In UX research, this is the core risk of synthetic participants. Their outputs may be fluent, coherent, and aligned with human findings. Fluency alone does not show that the response was produced through grounded perception, causal reasoning, uncertainty monitoring, or lived interaction with the interface (Quattrociocchi et al., 2025).

4.3 Cognitive and Methodological Limits of Prompting

The findings discussed in the previous sections suggest that prompting can substantially improve the usefulness of synthetic participants in UX research. Detailed task descriptions, contextual information, and carefully specified personas often increase the similarity between synthetic and human outputs. This creates the impression that sufficiently detailed prompting can progressively close the gap between synthetic and human participants.

The available evidence points toward a different conclusion. Prompting can shape the behaviour of a language model. It cannot fundamentally alter the mechanisms through which the model generates responses. Improvements in prompting do not translate into unlimited improvements in simulation quality. Beyond a certain point, the limitations of synthetic participants stem from the absence of cognitive, experiential, and contextual resources that characterize human users.

The question is therefore not whether prompting matters. The literature strongly suggests that it does. The more important question is where its limits lie.

One way to describe this boundary is to distinguish between linguistic plausibility and epistemic grounding. Quattrociocchi et al. (2025) argue that LLMs may produce outputs that appear knowledgeable while lacking the epistemic work that human judgment involves. Three limitations are particularly relevant for UX research:

- **Grounding** refers to the absence of a lived connection between model output and physical or social experience.
- **Metacognition** refers to the limited ability of the model to recognize uncertainty, detect its own errors, or suspend judgment.
- **Causal reasoning** refers to the tendency to predict what kind of explanation is likely to follow from a prompt, in place of building genuine models of why users behave in certain ways.

These limitations clarify why better prompting cannot fully turn linguistic imitation into human-like experience.

4.3.1 Cognitive Limits

A first category of limitations concerns aspects of human cognition that cannot be directly reconstructed through prompting. Human interaction with digital interfaces is shaped by perception, attention, memory, emotions, expectations, habits, prior experiences, and situational constraints. These factors interact during task completion. Users encounter interfaces through an ongoing experience that extends beyond evaluation.

Language models operate differently. Even when prompted to adopt a specific persona or mental state, the model does not possess the underlying cognitive processes associated with that state. It generates a textual description of how such a user might respond. This distinction becomes particularly relevant where UX outcomes depend on cognitive mechanisms that are difficult to externalize in language.

Four cognitive limitations are especially relevant for UX research: embodiment, Theory of Mind, dual-process interaction, and metacognition.

- Grounded cognition research suggests that human thought and language are supported by modal simulations and situated action, grounded in perception, action, bodily states, and introspection (Barsalou, 2008). Pelkey (2023) similarly argues that human thought and communication draw on bodily meaning rooted in physical experience. In UX contexts, using an interface is an embodied and situated activity, shaped by posture, fatigue, motor effort, environmental distraction, device constraints, and bodily state (Barsalou, 2008; Pelkey, 2023). A button that a synthetic participant rates as sufficiently large may still be difficult to tap for a tired user holding a rain-wet phone. Problems arising from physical context, motor constraints, or situational stress are therefore among the kinds of problems that synthetic participants are least equipped to surface.

- Theory of Mind is relevant whenever UX depends on how users interpret intentions, expectations, or social meaning. Strachan et al. (2024) tested human and LLM performance across Theory of Mind tasks, including false-belief attribution, indirect speech, and faux pas detection. GPT-4 matched or exceeded human performance on several tasks; faux pas detection remained a notable exception. Ullman (2023, as cited in Strachan et al., 2024) further argues that small alterations to standard Theory of Mind tasks can undermine LLM performance, suggesting that apparent success may reflect pattern matching more than genuine mentalising. Schurz et al. (2014) also show that human Theory of Mind tasks engage a core neural network, including the medial prefrontal cortex

and bilateral temporo-parietal junction. In UX research, this gap matters for conversational interfaces, onboarding flows, error messages, and trust-related interactions, where users continuously infer whether a system is helpful, blaming, competent, deceptive, or aligned with their goals.

- Many UX problems emerge through the interaction between fast, intuitive processing and slower, deliberative reasoning. Kahneman (2011) describes this distinction as System 1 and System 2: System 1 is fast, automatic, and associative, while System 2 is slower, more effortful, and more deliberate. Li et al. (2025) apply a similar distinction to LLM agent design, where fast responses and slower reflective reasoning are treated as separate modes of processing. In UX terms, real users often begin by acting automatically and then slow down, hesitate, reconsider, or become stuck when an interface violates their expectations. These switching points are methodologically valuable because they reveal cognitive friction. A synthetic participant that processes the same interface fluently, without hesitation or backtracking, may miss these moments entirely—because the model does not experience the difficulty that would make the problem visible.

- Metacognition is central to many UX methods because users often provide evidence through uncertainty. In think-aloud and verbal-report methods, expressions of uncertainty, partial understanding, and confusion can provide important evidence about how users interpret a task (Ericsson and Simon, 1980). Quattrociocchi et al. (2025) identify metacognition as one of the core epistemic fault lines between human and LLM cognition: the ability to represent uncertainty, detect errors, and suspend judgment. Schröder et al. (2025) further show that models may generalise based on textual similarity, responding to surface features of wording more than to the underlying semantic construct. Almeida et al. (2024), citing Park et al. (2023), describe a related *correct answer effect*: models treat nuanced social and attitudinal questions with the same predeterminedness as factual questions. In UX terms, this means that synthetic participants do not authentically say "I am not sure" or "something feels off but I cannot explain why", even though such statements can be among the most informative signals in real usability sessions.

Tacit Knowledge and Implicit Reasoning

Many aspects of user behaviour rely on knowledge that users themselves cannot fully articulate. People often struggle to explain why they trusted a particular interface, why they ignored a warning message, or why a navigation structure felt intuitive. Such judgments emerge from accumulated experience, habits, perceptual cues, and auto-

matic cognitive processes. Prompting can ask a model to imitate these behaviours. The imitation remains dependent on textual patterns observed during training; the model has access to descriptions of user behaviour, not to the underlying experiential processes themselves. As a result, synthetic participants may provide plausible explanations for actions that real users would perform without conscious reflection.

Emotion and Lived Experience

A similar limitation applies to emotional experience. Emotions play an important role in many UX contexts. Trust, frustration, uncertainty, anxiety, enjoyment, satisfaction, and engagement frequently shape how users interact with systems. Language models can generate convincing descriptions of emotional states. They do not experience those states. This distinction has limited consequences for tasks focused on usability inspection or interface critique. It becomes more significant in contexts where emotional reactions are themselves the object of investigation. Understanding why users abandon an online purchase, hesitate before sharing personal information, or develop trust in an AI system often requires access to emotional processes that cannot be reduced to textual descriptions alone.

Situated and Embodied Interaction

Human behaviour is also influenced by the physical and situational circumstances in which interaction occurs. Interfaces are used while commuting, working under time pressure, multitasking, feeling tired, being interrupted, or interacting through unfamiliar devices. Such conditions affect attention, decision-making, error rates, and perceptions of usability. Prompts can explicitly describe these situations. The resulting simulation remains representational. The model can reason about distraction, fatigue, or time pressure; it does not operate under those conditions. Prompting can therefore approximate contextual constraints without reproducing their actual cognitive consequences.

4.3.2 Methodological Limits

Not all limitations originate from cognition. Some emerge from methodological properties of language models and the way synthetic participants are constructed.

Dependence on Training Data and Persona Bias. Synthetic participants remain constrained by the data on which language models were trained. The ability to simulate a user depends partly on whether similar users, behaviours, preferences, and

contexts are represented within the model's training corpus. Some populations, products, and everyday situations are heavily represented in digital data; others appear rarely or through distorted representations. As a result, synthetic participants may systematically perform better when simulating common, well-documented, and digitally visible user groups than when representing marginalized populations, culturally specific contexts, or highly specialized domains. Prompting cannot fully compensate for information that was never learned, was weakly represented, or was represented through biased associations.

Persona-based prompting introduces an additional methodological risk. When instructed to simulate a particular demographic group, language models may rely on generalized associations attached to that category. A synthetic participant representing an older adult, a novice user, or a person from a specific cultural background may therefore reproduce a plausible stereotype in place of a grounded user perspective.

Batzner et al. (2025) make an important distinction here: demographic grounding is not the same as ecological validity. A persona may be based on verifiable demographic or social data and still fail to represent how real users behave in actual interaction settings. Their review of synthetic persona research shows that representativeness and interaction context are often insufficiently discussed. For UX research, a synthetic participant can appear empirically grounded and still fail to capture the conditions under which users encounter a product—time pressure, distraction, emotional stakes, device constraints, or social context.

Amidei et al. (2026) illustrate the problem at the level of generated synthetic populations. Their analysis of LLM-generated personas found systematic WEIRD biases and problematic associations involving marginalized identities. This suggests that persona prompting can make synthetic participants appear more specific while importing hidden assumptions from the model. The greater risk is that their fluent and coherent responses make such bias difficult to detect.

Hallucinations and Constructed Rationalizations. Language models also exhibit a tendency to generate plausible but unsupported explanations. When asked to explain why a user would behave in a certain way, the model often produces coherent narratives even when insufficient information is available. Such explanations may appear methodologically valuable because they are internally consistent and linguistically convincing. Coherence, however, should not be confused with evidence. Human participants report actual experiences and behaviours. Synthetic participants generate hypothetical explanations. The resulting feedback may therefore contain rationalizations that sound plausible without corresponding to any genuine user process.

Responsiveness to Prompt Framing. Another limitation concerns the sensitivity of language models to prompt framing. The same interface may receive substantially different evaluations depending on how a task is formulated, which persona is selected, or which contextual information is included. Some degree of sensitivity is desirable. Excessive responsiveness creates methodological uncertainty. If small modifications in prompting lead to large changes in outputs, the stability and reliability of synthetic evaluations become difficult to assess. This issue is particularly important because prompt design itself becomes a source of variance in the research process.

The Boundary of Prompting. Taken together, these observations suggest that prompting functions as a mechanism of approximation. It can provide information about how a hypothetical user might interpret an interface based on available textual knowledge. It can guide a model toward specific perspectives, goals, and contexts. It cannot, however, provide access to experiences, emotions, perceptions, or cognitive processes that are absent from the model's architecture. The practical implication is that prompting encounters both cognitive and methodological boundaries. Cognitive boundaries emerge from the absence of human experience and human information processing. Methodological boundaries emerge from the statistical nature of language models, their training data, and their susceptibility to framing effects.

This boundary also remains relevant as models become more agentic and multimodal. Systems such as UXAgent and Avenir-UX show that LLM-based UX tools can move beyond static text prompting by using personas, browser access, action traces, memory or reasoning traces, video recordings, GUI grounding, standardized usability metrics, and think-aloud-style reports (Lu et al., 2025b; Tan et al., 2026). These systems may reduce some limitations of static prompting, especially those connected to interface perception and task execution. The deeper limits described in this section remain. LLM outputs remain weakly grounded in lived physical and social experience, and their apparent judgments do not necessarily involve the epistemic capacities of human judgment, including grounding, metacognition, and causal reasoning (Quattrociocchi et al., 2025). This matters for UX because interaction is an embodied, affective, sensual, and situated experience—not only task execution (McCarthy and Wright, 2004; Barsalou, 2008). Seeing an interface is therefore not the same as experiencing fatigue, uncertainty, frustration, social pressure, or personal risk while using it. Agentic and multimodal systems may reduce perceptual absence. The absence of embodiment, lived affect, and situated social cognition remains.

These limitations clarify the conditions under which synthetic participants can be used effectively. Their value depends on the degree to which a given UX task relies on resources that can be approximated through prompting. A more differentiated analysis is therefore needed: one that addresses when, where, and for what purposes synthetic

participants can be used. The following section develops such an analysis in the form of a task-oriented framework.

4.4 A Task-Oriented Framework

The preceding sections showed that the validity of synthetic participants cannot be reduced to the question of whether their outputs resemble those of human users. Section 4.1 argued that prompting can improve synthetic outputs when relevant information about the task, user, and context is provided. Section 4.2 showed that correspondence between synthetic and human outputs may take different forms and should not automatically be interpreted as evidence of successful simulation. Section 4.3 identified cognitive and methodological limits that constrain what can realistically be approximated through prompting.

Taken together, these arguments suggest that the question of whether synthetic participants can replace human users is too broad to provide useful methodological guidance. A more precise question is what kind of knowledge or evidence is required to answer a particular UX question. Some UX tasks depend mainly on explicit knowledge, analytical reasoning, and design principles. Others depend on interpretation, contextual understanding, affective response, or lived experience. The suitability of synthetic participants therefore varies according to the type of UX phenomenon being investigated.

The framework can also be read through these cognitive limits:

- Tasks that primarily require structural inspection, such as early information architecture work, tree testing, or heuristic review, activate fewer cognitive limits and are therefore more suitable for synthetic support.
- Tasks that require hesitation, uncertainty monitoring, or verbalization during interaction activate metacognitive and dual-process limits.
- Tasks that require users to infer system intent, social meaning, or the likely experience of other users activate Theory of Mind limits.
- Tasks that depend on bodily state, emotional response, accessibility, real-world context, or long-term use activate embodiment and temporality limits.

Synthetic participants are therefore most defensible when the required evidence is explicit and structural, and least defensible when the evidence depends on lived, situated, or self-monitoring processes.

This thesis proposes that the type of evidence required by the research objective determines the appropriate role of synthetic participants. Interviews, surveys, usability tests, and diary studies are methodological instruments, not research goals. The same method can be used to investigate different phenomena. An interview may examine

users' understanding of an interface, or it may examine trust, frustration, or a long-term relationship with a product. The methodological value of synthetic participants depends on what the method is being used to study.

For this reason, UX tasks can be understood as existing along a continuum between explicit knowledge and lived experience. At one end are tasks that rely mainly on analytical reasoning and explicit knowledge structures. At the other end are tasks that depend on subjective experience, contextual factors, and emotional involvement. Between these two extremes are tasks that require interpretation and perspective-taking. The framework proposed in this thesis distinguishes between three categories: knowledge-oriented tasks, interpretation-oriented tasks, and experience-oriented tasks. As shown in Figure 4.3, these categories are arranged along a horizontal continuum. The left end—labelled as analytical and explicit—corresponds to tasks where the required knowledge can be represented and reasoned about linguistically. The right end—labelled as experiential, contextual, and emotional—corresponds to tasks where the required evidence arises from situated, embodied, or affective involvement with a product. The three category labels sit above this arrow as successive segments, marking where along the axis the character of the required evidence shifts.

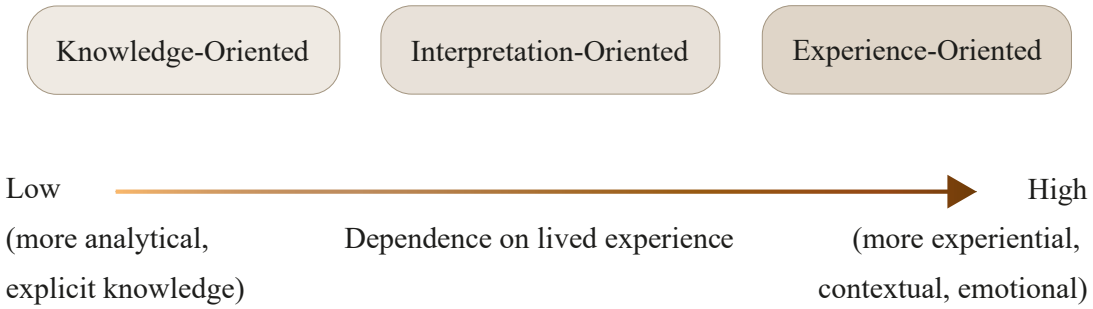

Figure 4.3: Continuum of UX Tasks Based on Dependence on Lived Experience.

Knowledge-Oriented Tasks

Knowledge-oriented tasks are characterized by a strong dependence on explicit knowledge, analytical reasoning, and established design principles. They aim to evaluate structural or informational aspects of an interface. The objective is to identify problems that can be detected through the application of design knowledge: whether the interface is clear, consistent, logically organized, and aligned with established usability principles. These tasks are therefore close to the pragmatic and structural dimensions of UX discussed in Chapter 2.

Typical examples include information architecture evaluation, assessment of navigation structures, analysis of content clarity, identification of heuristic violations, comparison of alternative interface solutions, and detection of obvious usability issues in early-stage prototypes. These tasks are commonly associated with methods such as heuristic evaluation, tree testing, card sorting, content review, and expert inspection.

The defining feature of knowledge-oriented tasks is that successful performance depends largely on knowledge that can be represented linguistically. The evaluator needs no personal experience with the system. Access to emotional reactions, embodied interaction, or long-term contextual use is not required. The task mainly requires the application of design principles and analytical judgement. Compared with the broader UX phenomena discussed in Chapter 2.2, these tasks depend less on emotional or temporal experience and more on the kinds of explicit patterns that Chapter 3 showed LLMs can often reproduce.

This makes knowledge-oriented tasks the most suitable category for synthetic participants. Large language models are trained on extensive textual material that includes usability principles, design guidelines, interface conventions, information architecture patterns, and discussions of common interaction problems. They may therefore identify structural issues that correspond to issues found by human evaluators. In this category, functional correspondence is often sufficient: the practical value of the synthetic output depends mainly on whether a relevant UX issue is identified, and the cognitive process through which the model reached that conclusion is secondary. For this reason, synthetic participants may serve as a practical substitute for some forms of early-stage evaluation, design critique, and preliminary usability inspection.

This category contains tasks with the highest substitution potential. When the objective is to identify structural weaknesses, synthetic participants can provide meaningful preliminary evidence without direct involvement of human users.

Interpretation-Oriented Tasks

Interpretation-oriented tasks occupy an intermediate position in the proposed framework. They require analytical reasoning and involve questions of meaning-making, expectation, and user perspective. The central issue is how the interface is likely to be understood by users. The objective of this category is to examine how users form expectations, interpret interface elements, understand system functionality, and make sense of an interaction. These tasks depend on user perspective. They do not always require direct access to deep emotional or long-term experience.

Typical examples include onboarding evaluation, first impressions, concept testing, preference formation, investigation of mental models, analysis of how users understand interface elements or system functionality, and evaluation of domain-specific user per-

spectives, such as patients, older adults, novice users, experts, or professionals.

Mental models occupy an intermediate position in this framework. They are more accessible to language-based simulation than embodied or emotional experience, because users often express them through explanations, expectations, and verbal interpretations, for example in interviews or think-aloud protocols (Ericsson and Simon, 1980; Lazar et al., 2017). At the same time, they cannot be reduced to explicit knowledge alone. A user's mental model develops through interaction with a system and may include uncertainty, incorrect assumptions, forgotten details, and practical shortcuts (Norman, 1983). For this reason, synthetic participants may be useful for generating hypotheses about possible misunderstandings or expectation mismatches, but their outputs should not be treated as evidence of actual user understanding. Human validation remains necessary when the research question concerns how real users form, revise, or fail to revise their understanding during interaction.

These tasks are often explored through usability testing, think-aloud protocols, interviews, surveys, and qualitative evaluation methods.

Synthetic participants can contribute to interpretation-oriented tasks, though their role should usually remain complementary. Through detailed prompts, models can be guided toward particular user perspectives, contextual situations, and behavioural assumptions. As discussed in Section 4.1, descriptions of user goals, psychographic characteristics, and contextual constraints can improve the plausibility of synthetic outputs.

These outputs remain reconstructions generated from patterns in training data. They are not derived from actual user experience. This creates several risks: synthetic participants may reproduce demographic stereotypes, overgeneralize characteristics associated with a user group, or fail to capture variation within a real population. This limitation is especially relevant when studying domain-specific users. A model prompted to respond as an older adult, patient, novice user, or professional may generate a plausible answer, but the answer may reflect generalized assumptions about such groups. The output may be useful for exploration. Treating it as direct evidence of how members of these groups think, act, or interpret a system requires caution.

For this reason, synthetic participants can support human research in interpretation-oriented tasks—generating hypotheses, identifying possible sources of confusion, and exploring alternative interpretations of an interface. Human validation remains necessary when the research objective is to understand how real users perceive and interpret a system. This category may nevertheless be one of the most practically valuable areas for future use of synthetic participants in UX research. Their ability to generate possible perspectives quickly can enrich exploratory work, provided that the results are later checked against human participants.

Experience-Oriented Tasks

Experience-oriented tasks concern aspects of UX that are lived, felt, situated, and often developed over time. They are the category in which the limits of synthetic participants become most visible. The objective is to understand how users experience a product in relation to emotion, embodiment, context, personal history, and long-term use. These tasks ask what it is like to use the system in a concrete human situation.

Typical examples include trust, emotional engagement, frustration, anxiety, technology adoption, habit formation, accessibility in everyday contexts, and long-term relationships with products and services. These topics are commonly studied through interviews, diary studies, contextual inquiry, ethnographic observation, and longitudinal studies.

The limitations discussed in Section 4.3 are most significant in this category. Experience-oriented tasks depend on cognitive and experiential resources that cannot be fully reconstructed through prompting—embodied interaction, tacit knowledge, emotional involvement, situational context, and personal history.

Language models can generate convincing descriptions of frustration, trust, anxiety, or long-term engagement. Such descriptions should not be confused with the phenomena themselves. Language models do not possess experience. Generating a description of experience is a different operation from having it. Even when synthetic outputs appear plausible or correspond superficially to findings from human participants, such correspondence does not necessarily indicate meaningful simulation. The outputs remain textual representations generated from learned patterns. They are not expressions of lived experience.

For this reason, synthetic participants have the most limited role in experience-oriented research. They may help prepare research questions, develop scenarios, anticipate possible issues, or support study design. They cannot replace empirical engagement with users when the objective is to understand subjective experience. Human participants therefore remain the primary and indispensable source of evidence for experience-oriented UX research.

4.4.1 From Replacement to Complementarity

The framework proposed in this thesis suggests that the methodological value of synthetic participants varies according to the nature of the UX task being performed. As tasks move from explicit knowledge toward lived experience, the role of synthetic participants gradually changes. In knowledge-oriented tasks, they may function as a primary source of evidence. In interpretation-oriented tasks, they serve most effectively as complementary sources of evidence that require validation through human research. In experience-oriented tasks, their role becomes primarily supportive and exploratory.

This perspective also helps explain the mixed findings reported throughout the literature. Synthetic participants may perform remarkably well in some studies and considerably worse in others. Different studies investigate fundamentally different types of UX phenomena. The technology is not the source of inconsistency. The critical distinction lies between the forms of knowledge that research methods seek to produce.

Table 4.1 maps this distinction across sixteen UX tasks. For each task, the table identifies the primary type of resource required, the methods typically used to investigate it, and the resulting category within the framework. The final column assigns a role to synthetic participants: primary source of evidence for knowledge-oriented tasks, complementary or validation-dependent evidence for interpretation-oriented tasks, and a supporting or indispensable-human role for experience-oriented tasks. Reading down the table, the shift from the top to the bottom rows traces the same gradient as the continuum in Figure 4.2—from tasks where synthetic approximation is most defensible to tasks where it becomes increasingly insufficient.

Table 4.1: Positioning UX Tasks Within the Proposed Framework.

UX focus	**Primary resource required**	**Typical methods**	**Category**	**Role of synthetic participants**
Information architecture	Explicit knowledge	Tree testing, card sorting	Knowledge-oriented	Primary source of evidence
Navigation structure	Analytical reasoning	Usability testing, walk-throughs	Knowledge-oriented	Primary source of evidence
Content clarity	Explicit knowledge	Content evaluation, expert review	Knowledge-oriented	Primary source of evidence
Heuristic evaluation	Design principles	Heuristic review	Knowledge-oriented	Primary source of evidence
Early usability problem identification	Analytical reasoning	Usability inspection	Knowledge-oriented	Primary source of evidence

Continued on next page

UX focus	**Primary resource required**	**Typical methods**	**Category**	**Role of synthetic participants**
Onboarding evaluation	User interpretation	Usability testing, first-use studies	Interpretation-oriented	Complementary evidence
First impressions	User interpretation	Interviews, usability testing	Interpretation-oriented	Complementary evidence
Mental models	Interpretation and meaning-making	Interviews, think-aloud protocols	Interpretation-oriented	Complementary evidence requiring validation
Concept evaluation	Interpretation and expectations	Interviews, surveys	Interpretation-oriented	Complementary evidence
Domain-specific user perspectives	Contextual interpretation	Interviews, usability testing	Interpretation-oriented	Complementary evidence requiring validation
Preference formation	Subjective interpretation	Surveys, interviews	Interpretation-oriented	Complementary evidence
Trust	Emotional and experiential resources	Interviews, surveys	Experience-oriented	Human validation required
Frustration and anxiety	Emotional experience	Usability testing, interviews	Experience-oriented	Human validation required
Technology adoption	Longitudinal experience	Interviews, diary studies	Experience-oriented	Supporting role only

Continued on next page

UX focus	**Primary resource required**	**Typical methods**	**Category**	**Role of synthetic participants**
Habit formation	Longitudinal experience	Diary studies, ethnography	Experience-oriented	Supporting role only
Accessibility in everyday use	Embodied and contextual experience	Contextual inquiry, field studies	Experience-oriented	Supporting role only
Long-term product experience	Lived experience	Diary studies, ethnography	Experience-oriented	Human participants indispensable

4.4.2 A Hybrid Model of UX Research

The framework ultimately points toward a hybrid model of UX research in which synthetic and human participants contribute different forms of evidence throughout the research process. The model can be understood in three stages.

- In the early stages of design, synthetic participants may be used to evaluate prototypes, identify obvious usability issues, compare alternative solutions, and generate hypotheses regarding potential user problems. Such activities are particularly valuable within knowledge-oriented tasks and parts of interpretation-oriented research.

- Human participants subsequently become essential for validating findings, investigating subjective experiences, and identifying phenomena that cannot be approximated through prompting alone. This stage becomes increasingly important as research moves toward experience-oriented questions.

- Finally, discrepancies between synthetic and human findings can be used to refine future prompting strategies, improve persona construction, and recalibrate synthetic evaluation procedures. Synthetic and human participants thus become part of an iterative research process in which each compensates for the limitations of the other.

As shown in Figure 4.4, the cycle begins with synthetic participants generating preliminary findings through exploration. Those findings feed into a human validation

stage, where real participants either confirm or complicate the synthetic output. The resulting discrepancies and new insights then drive prompt refinement, which produces improved synthetic evaluation—at which point the loop restarts. The direction of the process is toward reducing uncertainty at each stage before the next one begins, not toward eliminating the need for human participants.

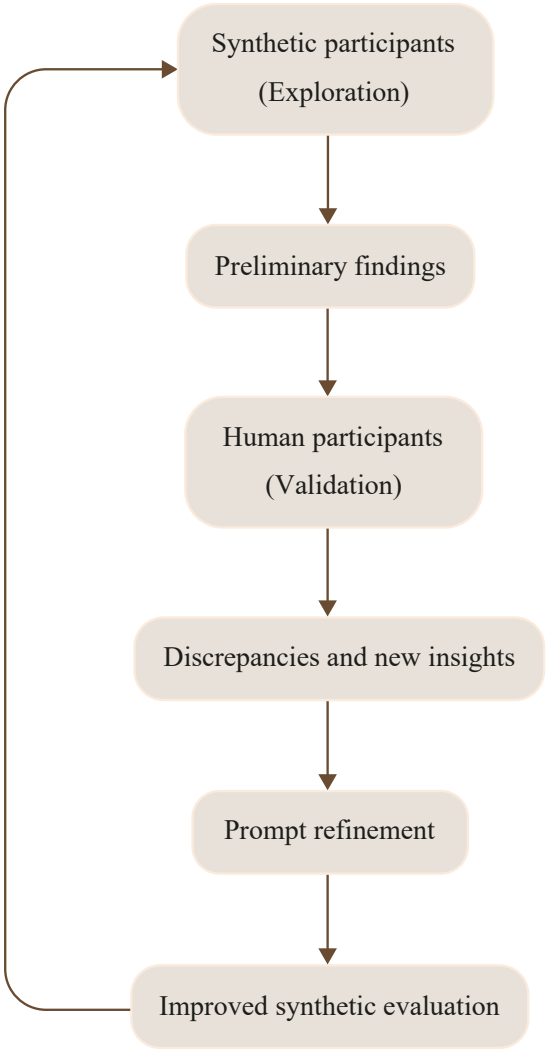

Figure 4.4: Hybrid UX Research Process.

Transparency in Synthetic Participant Research

If synthetic participants are used in UX research, their role should be reported explicitly. Synthetic outputs should not be described simply as user feedback, participant responses, or usability test results unless the report also states that the evidence was model-generated (Batzner et al., 2025).

A report should document the following (Batzner et al., 2025):

- the model and version,
- the prompt or prompt template,
- the persona construction method,
- the task instructions,
- the interface representation given to the model,
- the number of generations, and
- the procedure used to select or aggregate outputs.

It should also distinguish between findings from real users, findings from synthetic participants, and hypotheses generated from synthetic outputs but not yet validated with users.

Synthetic participants are most defensible as preparatory or exploratory tools. They can support early design inspection, task preparation, comparison of design alternatives, and generation of candidate usability issues. They should not be treated as direct evidence of user experience when the research question depends on emotion, trust, accessibility, embodied difficulty, long-term adoption, or high-stakes use (McCarthy and Wright, 2004; Barsalou, 2008; Quattrociocchi et al., 2025).

Chapter 5

Conclusion

This thesis examined the validity and limits of synthetic participants in user experience research. UX research depends on real users, yet user studies are often slow, costly, and difficult to scale. Large language models can generate fluent user-like feedback, adopt personas, and identify usability issues. The central question is what kind of evidence these outputs actually provide.

Synthetic participants should not be understood as artificial users. They are better described as language-model-based approximations of selected user perspectives. UX research does not study one single phenomenon. Some UX tasks concern interface structure, content clarity, navigation, or explicit usability problems. Others concern trust, emotion, accessibility, habit formation, or long-term experience. Each type requires different forms of evidence.

The theoretical discussion showed that LLMs are strongest where the relevant evidence is explicit, verbal, structural, or rule-based. They can apply design principles, reason about interface elements, compare alternatives, and generate possible user concerns. These capacities make them useful in early design work and preliminary evaluation. Their outputs are shaped by training data, prompt wording, persona descriptions, interface representations, and generation settings. They may also be affected by hallucination, bias, sycophancy, prompt sensitivity, false positives, and limited grounding.

The field has moved from simple text-based simulations toward structured and agentic UX pipelines. This progress changes the form of approximation. It does not remove the need for human evidence.

Prompt design was therefore treated as a methodological issue. A usable UX prompt is not simply a long prompt. It must ground the model in the task, the user, and the context of use. A synthetic output remains a mediated reconstruction of a user situation, not direct evidence of human experience.

A key distinction developed in the argument is the difference between output correspondence, functional correspondence, and cognitive correspondence. A model may

produce a response similar to a human participant, or identify the same usability issue, without reaching it through the same perceptual, cognitive, or emotional process. Such overlap is practically useful, especially in early inspection. It is not proof that the model meaningfully simulates a user.

The final synthesis proposed a task-oriented view. Synthetic participants are most suitable for knowledge-oriented UX tasks, where explicit reasoning and design conventions are central. They may serve as complementary tools for interpretation-oriented tasks, such as first impressions, onboarding, or concept evaluation. They are least suitable for experience-oriented tasks, where the relevant evidence depends on lived, embodied, affective, or long-term experience. In these cases, human participants remain the primary source of evidence.

The main contribution is a shift in framing. The relevant question is which aspects of UX research can be approximated by language models and which cannot. The answer points toward a hybrid position. Synthetic participants can help researchers prepare studies, generate hypotheses, compare alternatives, and filter obvious usability issues. They cannot replace human participants when the research question concerns what it is like to use a product in a concrete human situation.

The study is limited by its theoretical character. It does not present a new empirical comparison between synthetic and human participants, and the literature on this topic is still young and rapidly changing. Future research should therefore test the proposed task-oriented framework across different UX tasks and examine how prompt grounding, interface representation, model choice, and validation procedures affect synthetic outputs.

Synthetic participants are not irrelevant to UX research. They are not users either. Their value depends on the task, the prompt, and the type of evidence being sought. They may help UX researchers work faster and prepare more carefully. The validity of user research still depends on knowing when human experience cannot be substituted by generated text.

Bibliography

Almeida, G. F. C. F., Nunes, J. L., Engelmann, N., Wiegmann, A., and Araújo, M. d. (2024). Exploring the psychology of LLMs' moral and legal reasoning. *Artificial Intelligence*, 333:104145.

Amidei, J., Ferreira, G., Serrano, M. M., Nieto, R., and Kaltenbrunner, A. (2026). The personality trap: How LLMs embed bias when generating human-like personas.

Amirova, A., Fteropoulli, T., Ahmed, N., Cowie, M. R., and Leibo, J. Z. (2024). Framework-based qualitative analysis of free responses of large language models: Algorithmic fidelity. *PLOS ONE*, 19(3):e0300024.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., and Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Bangor, A., Kortum, P. T., and Miller, J. T. (2008). An empirical evaluation of the system usability scale. *International Journal of Human–Computer Interaction*, 24(6):574–594.

Bargas-Avila, J. A. and Hornbæk, K. (2011). Old wine in new bottles or novel challenges: A critical analysis of empirical studies of user experience. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '11*, pages 2689–2698.

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, 59:617–645.

Batzner, J., Stocker, V., Tang, B., Natarajan, A., Chen, Q., Schmid, S., and Kasneci, G. (2025). Whose personae? synthetic persona experiments in LLM research and pathways to transparency. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pages 343–354.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the*

2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, pages 610–623.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2022). On the opportunities and risks of foundation models. arXiv:2108.07258.

Brooke, J. (1996). SUS: A "quick and dirty" usability scale. In Jordan, P. W., Thomas, B., Weerdmeester, B. A., and McClelland, I. L., editors, *Usability Evaluation in Industry*, pages 189–194. Taylor & Francis.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners.

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., and Zhang, C. (2023). Quantifying memorization across neural language models. arXiv:2202.07646.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., and Raffel, C. (2021). Extracting training data from large language models. arXiv:2012.07805.

Díaz-Oreiro, I., López, G., Quesada, L., and Guerrero, L. (2021). UX evaluation with standardized questionnaires in ubiquitous computing and ambient intelligence: A systematic literature review. *Advances in Human-Computer Interaction*, 2021:22.

Duan, P., Warner, J., Li, Y., and Hartmann, B. (2024). Generating automatic feedback on UI mockups with large language models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–20.

Ericsson, K. A. and Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, 87:215–251.

Gonzalez, R. A. and DiPaola, S. (2024). Exploring augmentation and cognitive strategies for AI based synthetic personae. arXiv:2404.10890.

Gu, E. H., Chandrasegaran, S., and Lloyd, P. (2025). Synthetic users: Insights from designers' interactions with persona-based chatbots. *AI EDAM*, 39:e2.

Hämäläinen, P., Tavast, M., and Kunnari, A. (2023). Evaluating large language models in generating synthetic HCI research data: A case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23*, pages 1–19.

Hart, S. G. and Staveland, L. E. (1988). Development of NASA-TLX (task load index): Results of empirical and theoretical research. In *Human mental workload*, pages 139–183. North-Holland.

Hassenzahl, M. (2005). The thing and I: Understanding the relationship between user and product. In *Funology*, volume 3, pages 31–42.

Hassenzahl, M. (2008). User experience (UX): Towards an experiential perspective on product quality. In *ACM International Conference Proceeding Series*, volume 339, page 15.

Hassenzahl, M. and Tractinsky, N. (2006). User experience—a research agenda. *Behaviour & Information Technology*, 25(2):91–97.

Henrich, J., Heine, S. J., and Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2–3):61–83.

Hornbæk, K. (2006). Current practice in measuring usability: Challenges to usability studies and research. *International Journal of Human-Computer Studies*, 64(2):79–102.

Hornbæk, K. and Hertzum, M. (2017). Technology acceptance and user experience: A review of the experiential component in HCI. *ACM Transactions on Computer-Human Interaction*, 24:1–30.

Hsueh, N.-L., Lin, H.-J., and Lai, L.-C. (2024). Applying large language model to user experience testing. *Electronics*, 13(23):4633.

Hu, T. and Collier, N. (2024). Quantifying the persona effect in LLM simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, page 10307.

Iarygina, O., Hornbæk, K., and Mottelson, A. (2024). Demand characteristics in human–computer experiments. *International Journal of Human-Computer Studies*, 193:103379.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):248:1–248:38.

Jurafsky, D. and Martin, J. H. (2026). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd edition. Available at `https://stanford.edu`.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Karapanos, E., Zimmerman, J., Forlizzi, J., and Martens, J.-B. (2009). User experience over time: An initial framework. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 729–738.

Khot, T., Trivedi, H., Finlayson, M., Fu, Y., Richardson, K., Clark, P., and Sabharwal, A. (2022). Decomposed prompting: A modular approach for solving complex tasks. arXiv:2210.02406.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. (2023). Large language models are zero-shot reasoners. arXiv:2205.11916.

Law, E. L.-C., Roto, V., Hassenzahl, M., Vermeeren, A. P. O. S., and Kort, J. (2009). Understanding, scoping and defining user experience: A survey approach. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '09*, pages 719–728.

Lazar, J., Feng, J. H., and Hochheiser, H. (2017). *Research Methods in Human-Computer Interaction*. Morgan Kaufmann, 2nd edition.

Le, T., Chaudhuri, S., Chung, J., Thompson, H. J., and Demiris, G. (2014). Tree testing of hierarchical menu structures for health applications. *Journal of Biomedical Informatics*, 49:198–205.

Lewis, J. (1994). Sample sizes for usability studies: Additional considerations. *Human Factors*, 36:78.

Li, Z.-Z., Zhang, D., Zhang, M.-L., Zhang, J., Liu, Z., Yao, Y., Xu, H., Zheng, J., Wang, P.-J., Chen, X., et al. (2025). From system 1 to system 2: A survey of reasoning large language models. arXiv:2502.17419.

Lin, S., Hilton, J., and Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. arXiv:2109.07958.

Lu, Y., Bartolo, M., Moore, A., Riedel, S., and Stenetorp, P. (2022). Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. arXiv:2104.08786.

Lu, Y., Yao, B., Gu, H., Huang, J., Wang, J., Li, Y., Gesi, J., He, Q., Li, T. J.-J., and Wang, D. (2025a). UXAgent: A system for simulating usability testing of web design with LLM agents. arXiv:2504.09407.

Lu, Y., Yao, B., Gu, H., Huang, J., Wang, Z. J., Li, Y., Gesi, J., He, Q., Li, T. J.-J., and Wang, D. (2025b). UXAgent: An LLM agent-based usability testing framework for web design. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, CHI EA '25*, pages 1–12.

Marcilly, R., Monkman, H., Pelayo, S., and Lesselroth, B. J. (2024). Usability evaluation ecological validity: Is more always better? *Healthcare*, 12(14):1417.

Maynez, J., Narayan, S., Bohnet, B., and McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. arXiv:2005.00661.

McCambridge, J., de Bruin, M., and Witton, J. (2012). The effects of demand characteristics on research participant behaviours in non-laboratory settings: A systematic review. *PLOS ONE*, 7(6):e39116.

McCarthy, J. and Wright, P. (2004). *Technology as Experience*. The MIT Press.

Mirnig, A., Meschtscherjakov, A., Wurhofer, D., Meneweger, T., and Tscheligi, M. (2015). A formal analysis of the ISO 9241-210 definition of user experience. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, page 450.

Nielsen, J. (1994). Enhancing the explanatory power of usability heuristics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '94*, pages 152–158.

Norman, D. A. (1983). Some observations on mental models. In Gentner, D. and Stevens, A. L., editors, *Mental Models*, pages 7–14. Lawrence Erlbaum Associates.

Norman, D. A. (2004). *Emotional Design: Why We Love (or Hate) Everyday Things*. Basic Books.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. arXiv:2203.02155.

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.

Pelkey, J. (2023). Embodiment and language. *Wiley Interdisciplinary Reviews: Cognitive Science*, 14(5):e1649.

Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., et al. (2022). Discovering language model behaviors with model-written evaluations.

Pruitt, J. and Adlin, T. (2006). *The Persona Lifecycle: Keeping People in Mind Throughout Product Design*. Morgan Kaufmann.

Quattrociocchi, W., Capraro, V., and Perc, M. (2025). Epistemological fault lines between human and artificial intelligence. arXiv:2512.19466.

Radford, A. and Narasimhan, K. (2018). Improving language understanding by generative pre-training.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners.

Reynolds, L. and McDonell, K. (2021). Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Rieman, J. (1993). The diary study: A workplace-oriented research tool to guide laboratory efforts. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems, CHI '93*, pages 321–326.

Rugg, G. and McGeorge, P. (1997). The sorting techniques: A tutorial paper on card sorts, picture sorts and item sorts. *Expert Systems*, 14(2):80–93.

Sarstedt, M., Adler, S. J., Rau, L., and Schmitt, B. (2024). Using large language models to generate silicon samples in consumer and marketing research: Challenges, opportunities, and guidelines. *Psychology & Marketing*, (6):1254–1270.

Schmidt, A., Elagroudy, P., Draxler, F., Kreuter, F., and Welsch, R. (2024). Simulating the human in HCD with ChatGPT: Redesigning interaction design with AI. *Interactions*, 31(1):24–31.

Schröder, S., Morgenroth, T., Kuhl, U., Vaquet, V., and Paaßen, B. (2025). Large language models do not simulate human psychology.

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., and Perner, J. (2014). Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neuroscience & Biobehavioral Reviews*, 42:9–34.

Sclar, M., Choi, Y., Tsvetkov, Y., and Suhr, A. (2024). Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. arXiv:2310.11324.

Seaborn, K., Barbareschi, G., and Chandra, S. (2023). Not only WEIRD but "uncanny"? a systematic review of diversity in human–robot interaction research. *International Journal of Social Robotics*, 15(11):1841–1870.

Shanahan, M. (2024). Talking about large language models. *Communications of the ACM*, 67(2):68–79.

Shapira, I., Benade, G., and Procaccia, A. (2026). How RLHF amplifies sycophancy.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., et al. (2025). Towards understanding sycophancy in language models. arXiv:2310.13548.

Shiffman, S., Stone, A. A., and Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology*, 4:1–32.

Strachan, J. W. A., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., Saxena, K., Rufo, A., Panzeri, S., Manzi, G., Graziano, M. S. A., and Becchio, C. (2024). Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7):1285–1295.

Subramanian, S., De Moor, K., Fiedler, M., Koniuch, K., and Janowski, L. (2023). Towards enhancing ecological validity in user studies: A systematic review of guidelines and implications for QoE research. *Quality and User Experience*, 8(1):6.

Tan, W. J., Lim, Z. R. L., Durgad, S., Obegi, K., and Li, A. Y. (2026). Avenir-UX: Automated UX evaluation via simulated human web interaction with GUI grounding. arXiv:2604.09581.

Tang, Y., Tuncel, D., Koerner, C., and Runkler, T. (2025). The few-shot dilemma: Over-prompting large language models. arXiv:2509.13196.

Tractinsky, N., Katz, A. S., and Ikar, D. (2000). What is beautiful is usable. *Interacting with Computers*, 13(2):127–145.

van Berkel, N., Ferreira, D., and Kostakos, V. (2017). The experience sampling method on mobile devices. *ACM Computing Surveys*, 50:1–40.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need.

Vermeeren, A. P. O. S., Law, E. L.-C., Roto, V., Obrist, M., Hoonhout, J., and Väänänen-Vainio-Mattila, K. (2010). User experience evaluation methods: Current state and development needs. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries, NordiCHI '10*, pages 521–530.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models. arXiv:2203.11171.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. (2023). Chain-of-thought prompting elicits reasoning in large language models. arXiv:2201.11903.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., et al. (2021). Ethical and social risks of harm from language models. arXiv:2112.04359.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., and Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. arXiv:2302.11382.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. (2023). A survey of large language models.

Zheng, M., Pei, J., Logeswaran, L., Lee, M., and Jurgens, D. (2024). When "a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. arXiv:2311.10054.

Zhong, R., McDonald, D. W., and Hsieh, G. (2025a). Synthetic cognitive walkthrough: Aligning large language model performance with human cognitive walkthrough. arXiv:2512.03568.

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