CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

Contrasting LLM-Assisted and Traditional UX Research in a B2B SaaS Environment

LLM-Assisted UX Research using Naturally Occurring Organizational Data and Synthetic Users

Master's thesis in Computer Science and Engineering

MAJA LARSSON
NADINE LINDÉN MALMBERG

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2026

Master's thesis 2026

Contrasting LLM-Assisted and Traditional UX Research in a B2B SaaS Environment

LLM-Assisted UX Research using Naturally Occurring Organizational Data and Synthetic Users

MAJA LARSSON
NADINE LINDÉN MALMBERG

UNIVERSITY OF GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Computer Science and Engineering
Chalmers University of Technology
University of Gothenburg
Gothenburg, Sweden 2026

Contrasting LLM-Assisted and Traditional UX Research in a B2B SaaS Environment
LLM-Assisted UX Research using Naturally Occurring Organizational Data and Synthetic Users
MAJA LARSSON & NADINE LINDÉN MALMBERG

Supervisor: Jasmina Maric, Interaction Design and Software Engineering,
Advisor: Carl-Johan Kihlbom, Teamtailor
Examiner: Aris Alissandrakis, Interaction Design and Software Engineering

Master's Thesis 2026
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in LaTeX
Gothenburg, Sweden 2026

Contrasting LLM-Assisted and Traditional UX Research in a B2B SaaS Environment
LLM-Assisted UX Research using Naturally Occurring Organizational Data and Synthetic Users
MAJA LARSSON & NADINE LINDÉN MALMBERG
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg

Abstract

User Experience research in Business-to-Business Software-as-a-Service (B2B SaaS) organizations relies on methods that are resource-intensive to sustain alongside continuous product development. Meanwhile, these organizations continuously generate large volumes of naturally occurring data such as customer support conversations, internal feedback channels, and design artifacts, that are rarely treated as systematic research input. While large language models (LLMs) are increasingly used to analyze data elicited for research, their ability to surface UX insight from naturally occurring organizational data, and how that insight compares to traditional methods, have received little systematic attention. This thesis investigates that question through a two-track study at Teamtailor, a B2B SaaS applicant tracking system (ATS) used by over 12,000 organizations. Track 1 conducted manual user research: a survey of 437 respondents and seven usability tests on a proposed redesign of the job creation flow. Track 2 developed a multi-agent LLM system that analyzed the same design context using different data. For the current product it drew on naturally occurring Slack feedback and Intercom support conversations; for the redesign it used synthetic-persona evaluation of the prototype in Figma. The two tracks were then systematically contrasted. The approaches surfaced different but complementary insights. LLM-assisted analysis was more effective at detecting silent system behaviors and recurring workflow constraints distributed across large data volumes; manual research was more effective at surfacing domain knowledge, missing functionality, and insights requiring contextual reasoning grounded in direct user engagement. The two tracks converged most strongly on the most critical usability issues, all of which informed design changes. Output quality depended heavily on conditions set before analysis: personas grounded in prior research, prompts framed with neutral domain context, and data structured consistently enough to carry meaningful distinctions. These findings support a hybrid model in which LLM-assisted analysis serves as a continuous breadth-first layer and human research provides the depth-first layer of contextual and interpretive understanding.

Keywords: LLM, UX research, naturally occurring data, B2B SaaS, synthetic personas, Human-AI collaboration, Research through Design.

Acknowledgements

We would like to thank Teamtailor for welcoming us and providing the access, resources, and trust that made this research possible. Working within a real product context gave the thesis a grounding that would have been difficult to achieve otherwise.

A particular thank you to Carl-Johan Kihlbom for his guidance throughout the project. His product knowledge and willingness to engage with our questions at every stage were invaluable and his enthusiasm for the work made the collaboration genuinely enjoyable.

We would also like to thank our academic supervisor Jasmina Maric for her consistent support and thoughtful feedback. Her input helped us sharpen both the argument and the presentation of the work, and her availability throughout the process made a real difference during the more demanding stretches of the thesis.

Finally, thank you to the Teamtailor users and customers who gave their time to participate in the interviews, survey, and usability testing. This work would not exist without them.

Maja Larsson and Nadine Lindén Malmberg, Gothenburg, 2026-06-10

Contents

List of Figures

List of Tables

1

Introduction

Over the past two decades, software has become the primary infrastructure through which organizations handle their work. In the Business to Business (B2B) context in particular, this shift has software at the center of everyday activity with employees routinely depending on digital products to complete core tasks across multiple workflows. As this dependency has grown, the expectations users bring to these tools have risen. It is no longer only about that software should function but that it should be intuitive and efficient.

This expectation has highlighted the importance of Interaction Design as a discipline which shapes the relationship between people and the digital systems they use. Where early software design prioritized technical capability over usability, interaction design focuses on the user's goals and behaviors as primary concerns in the design process. Within this discipline, User Experience (UX) research serves as the empirical foundation with a set of methods and practices through which designers develop evidence-based understanding of who their users are and what they need. Another effect of this digital shift is that B2B Software as a Service (SaaS) organizations generate large volumes of data in the course of their ordinary operations. They have customer feedback in internal messaging channels, support conversations and design artifacts produced during product development. This data is produced continuously and without the overhead of active data collection, yet it is rarely treated as a systematic source of UX insight.

1.1 Research Gap and Aim

UX research in B2B SaaS contexts relies on a well-established set of methods such as interviews, surveys, usability testing and observational studies. These methods produce rich, contextually grounded insight but are resource-intensive, requiring significant time and analytical effort that many product teams struggle to sustain alongside continuous development cycles [1]. As B2B SaaS platforms scale globally, this constraint becomes more complicated because the user base diversifies and the product complexity grows. This causes the demand for continuous research to increase while the capacity to conduct it often does not [2].

Recent developments in artificial intelligence (AI), specifically in large language model (LLM) capabilities, have opened new possibilities for addressing this constraint. A growing body of research has begun to examine whether LLMs can

perform qualitative analysis at a level comparable to human researchers, finding promising results alongside characteristic limitations around interpretive depth and contextual sensitivity [3], [4].

However, this body of work has focused almost exclusively on elicited data which is material produced specifically for research purposes, such as interview transcripts and survey responses. B2B SaaS organizations simultaneously generate a large stream of data through their day-to-day operations. This data exists independently of any research process, yet it contains information about user needs and recurring friction. Whether LLMs can bridge the distance between this naturally occurring organizational data and meaningful UX insight has received little systematic attention. This is the gap the thesis addresses.

This thesis investigates whether large language models can be used to inductively analyze naturally occurring data and whether this process can surface meaningful UX insights. The findings are contrasted against independently conducted manual user research on the same design context examining where the two approaches converge, diverge and contradict one another. The thesis contributes both an empirical investigation and a methodological protocol that practitioners and researchers in the field can build upon.

In the course of the study, the redesign context presented a methodological constraint. No naturally occurring data existed for the prototype, which had not yet been released. Rather than treating this as a limitation, the study used synthetic persona evaluation as an alternative input source for that track. This introduced a third emerging research question (RQ3) that is presented together with the two original research questions in Section 1.3.

1.2 Stakeholders

This study was conducted at Teamtailor, a B2B SaaS company offering applicant tracking software (ATS) to over 12,000 organizations worldwide [5]. Teamtailor supports companies ranging from small startups to large enterprises, each with distinct recruitment needs and workflows. This scale creates a continuous challenge of how to develop and validate a user-centred design (UCD) that works across a diverse user base without the capacity to conduct dedicated user research for every product change.

In their day-to-day operations, Teamtailor generates large volumes of operational data. Customer-facing teams collect informal product feedback in internal Slack channels, support staff handle inbound user problems through Intercom and designers produce wireframes and annotated mockups in Figma as a natural byproduct of the design process. Despite the volume and relevance of this material, it is rarely treated as a systematic source of UX insight. The product team instead mainly relies on their own judgment.

For Teamtailor, this thesis serves the purpose of performing UX research on an ongoing redesign of their core job posting flow. This is a workflow that affects the

majority of their users and that had not been evaluated with structured user research before launch. The thesis was conducted in close collaboration with Teamtailor's design and product teams, with access to internal data sources, design artifacts and product expertise throughout. From an academic standpoint, this context provides a representative and focused setting in which to examine the broader research questions this thesis set out to answer.

1.3 Research questions

Building on the gap established in Section 1.1, this leads to the following problem statement and research questions:

Problem Statement: Whether LLMs can turn the large volumes of naturally occurring operational data that B2B SaaS organizations generate through daily operations into usable UX insight, and what conditions shape that capacity, remains underexplored. As a result, these organizations have little basis for judging whether the operational data they already produce can serve as a systematic research input.

RQ1: How do the insights surfaced by LLM-assisted analysis of naturally occurring data relate to those from traditional UX research?

RQ2: What conditions shape the usefulness of LLM-assisted analysis on naturally occurring data?

In the course of the study, the redesign context presented a methodological constraint. No naturally occurring data existed for the prototype, which had not yet been released. Rather than treating this as a limitation, the study used synthetic persona evaluation as an alternative input source for that track. This introduced a third emerging research question:

RQ3: What conditions shape the reliability and usefulness of prototype evaluation with synthetic users?

1.4 Scope and Limitations

This thesis focuses exclusively on unstructured textual data across two naturally occurring textual sources (Slack feedback channel and Intercom support conversations) for the current design, and a Figma prototype evaluated via synthetic personas for the redesign. Quantitative behavioral data, such as product analytics or clickstream data falls outside the scope of the study. The analysis is bounded to a specific, active design context rather than the product as a whole. This scoping is deliberate as it allows for a focused evaluation of whether LLM-assisted analysis surfaces insights relevant to a concrete design problem. The thesis does not argue that LLM-assisted analysis of naturally occurring data can or should replace traditional UX research

methods. Rather, it examines what such analysis is capable of producing and what conditions shape its usefulness.

This study is also subject to some limitations that should be noted from start. The analysis is conducted within a single organizational context and makes no claim of statistical generalisability. The multi-agent LLM system (hereafter "LLM system") was constrained to predefined output schemas, which reduces the risk of hallucinated findings but may suppress patterns that do not align neatly with those categories. The redesign evaluation was conducted on a Figma prototype rather than a live product. The manual research drew on one survey and 7 usability tests. While these are sample sizes that are conventional for qualitative research they are small enough that the absence of a finding carries less weight than its presence.

1.5 Thesis Outline

Chapter 1 introduces the research context, establishes the gap this study addresses, and presents the research questions. Chapter 2 situates the study in its practical and industry context. Chapter 3 establishes the theoretical foundations. Chapter 4 describes the two parallel research tracks and their methods. Chapter 5 documents the process of conducting the two research tracks and their analysis. Chapter 6 presents the findings from both tracks across both design contexts. Chapter 7 discusses what the contrast between the different approaches reveals about LLM-assisted UX research. Chapter 8 concludes the thesis and outlines directions for future research.

2

Background

This chapter situates the study within its practical and industry context. It introduces Teamtailor and the B2B SaaS environment in which the research takes place, outlines the shifting expectations of B2B users and the resulting pressure on User Experience practice.

2.1 Teamtailor

This thesis is conducted in collaboration with Teamtailor, a global leader in recruitment software. Teamtailor provides an Applicant Tracking System (ATS), which is a software designed to manage and optimize recruitment data. Currently, Teamtailor supports over 12,000 companies worldwide, which creates the challenge of developing a user-centric design for a wide variety of users.

2.1.1 Applicant Tracking Systems

Initially conceived as databases for storing candidate information and tracking hiring pipeline stages, Applicant Tracking Systems (ATS) have evolved into strategic tools that shape how organizations define, attract, assess and hire talent [6]. Today's systems extend far beyond administrative record-keeping. They coordinate communication between recruiters and candidates, integrate with career sites and sourcing channels, automate screening logic and increasingly incorporate AI-assisted matching and ranking functionality [7].

What makes ATS software distinctive from a UX perspective is its structural multi-stakeholder character. A single ATS instance is used by recruiters (who manage day-to-day hiring activities), hiring managers (who assess candidates for their teams), administrators (who configure the system) and in some flows by candidates themselves (who interact with public-facing career pages and application forms). These stakeholder groups have overlapping but not identical needs, which creates the tension between serving the *buyer*, the *primary user*, and the *end beneficiary*, roles that are rarely occupied by the same person. In addition to this there is also the tension between the needs of small-to-medium businesses (SMBs) and enterprises.

Among the workflows that an ATS supports, job posting is a particularly friction-rich flow. Creating a job advertisement involves a combination of repetitive administrative steps (title, location, department), strategic content decisions (tone of

voice and selling points) and configuration work (approval flows and distribution channels). This combination of tasks (some fast and habitual, others slow and high-stakes) is the part of B2B workflows this thesis is concerned with and is the specific redesign context in which the study is conducted.

2.1.2 Redesign Context

Prior to the start of this Master's thesis, Teamtailor had begun planning a significant redesign of their job setup. A plan and preliminary mockups were already in place when the thesis began. As this redesign establishes the design context against which both the AI-assisted and manual research pipelines are evaluated, an understanding of the changes introduced is essential for interpreting the findings presented in Chapter 6. The existing job setup is illustrated as a user flow in Figure 2.1.

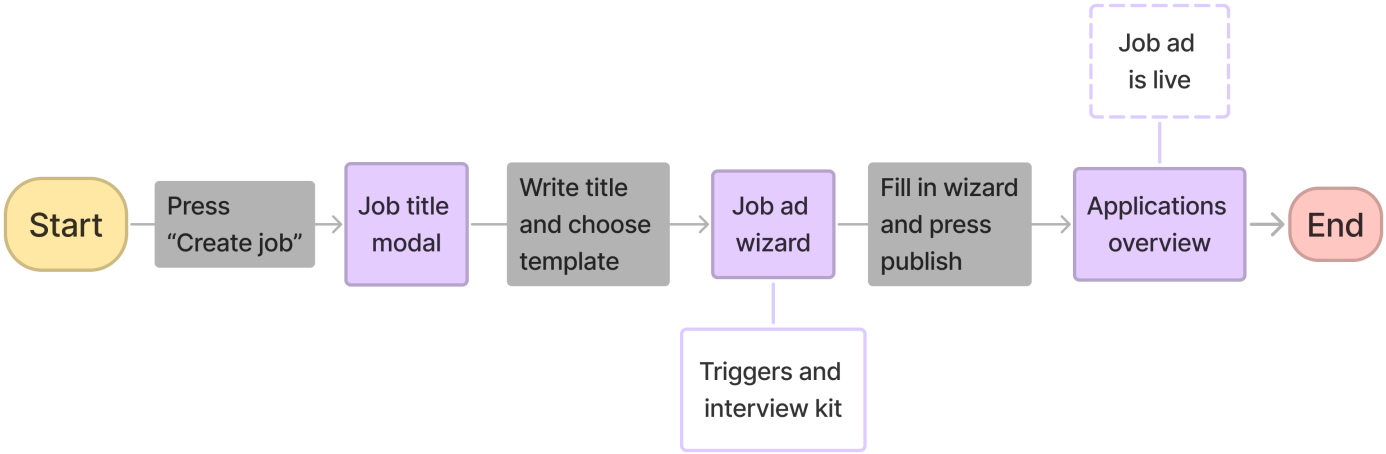


Figure 2.1: User flow of creating a job in the current design.

In the current flow, a job is created through a multi-step wizard. Upon completing the wizard, the user creates the job which posts an advertisement on the company's career page and simultaneously creates an applications overview page. From this page, recruiters can manage incoming applications and source candidates from the existing candidate database.

The redesign introduces a conceptual separation between a job and a job advertisement(ad). In the new model, a user creating a job can choose to do so without immediately publishing a job ad, enabling them to begin sourcing from their existing candidate pool before investing time in the details of the advertisement. Alternatively, they can create a job ad directly, following a flow similar to the current one. The redesigned flow is illustrated in Figure 2.2.

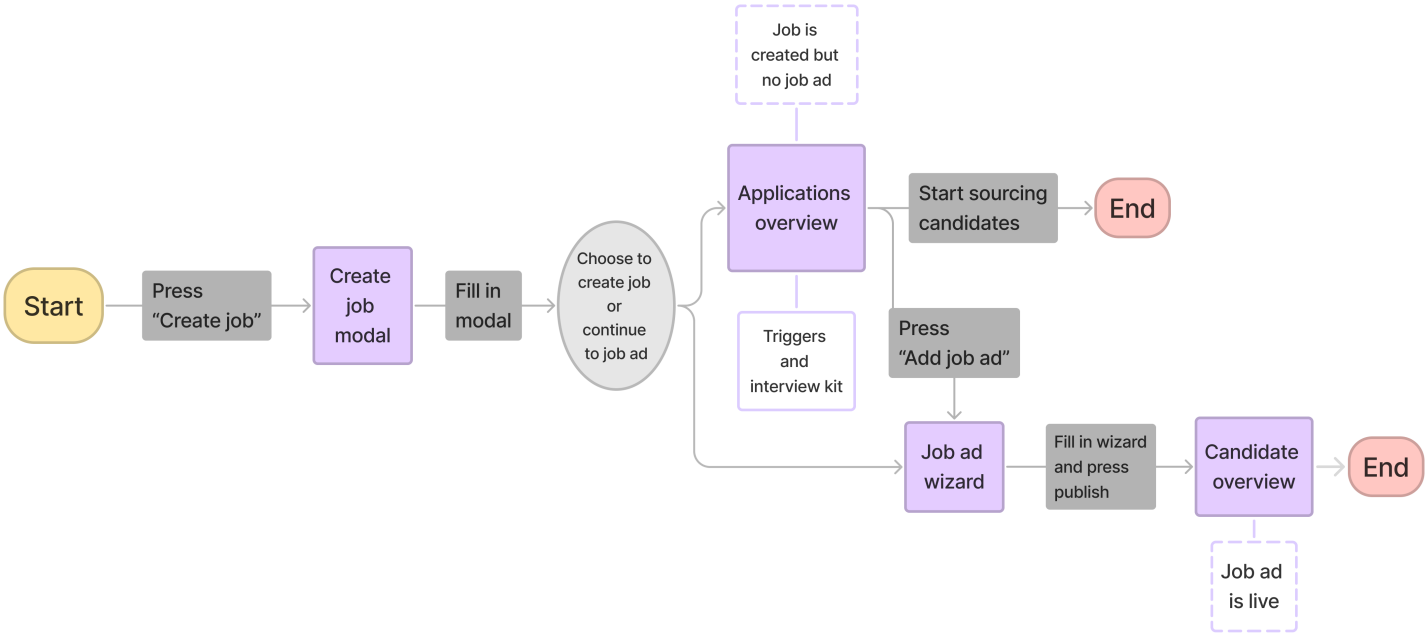
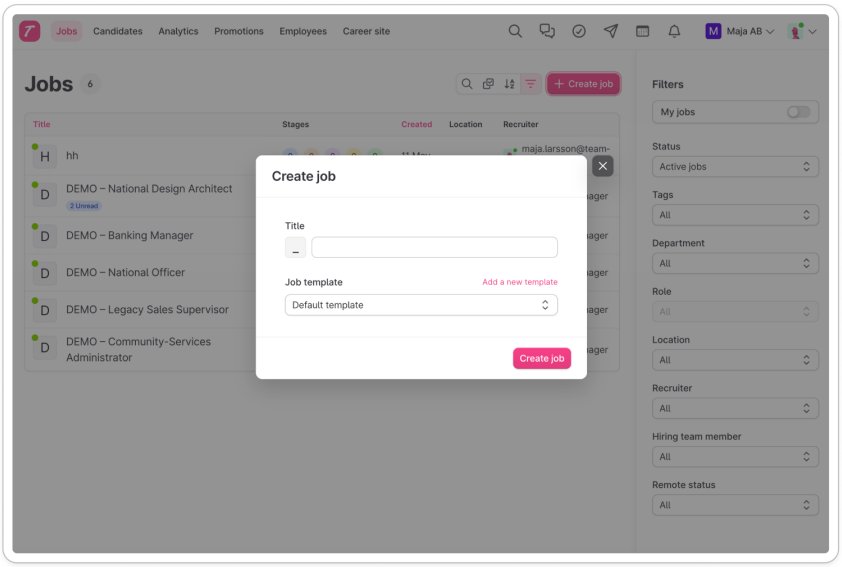


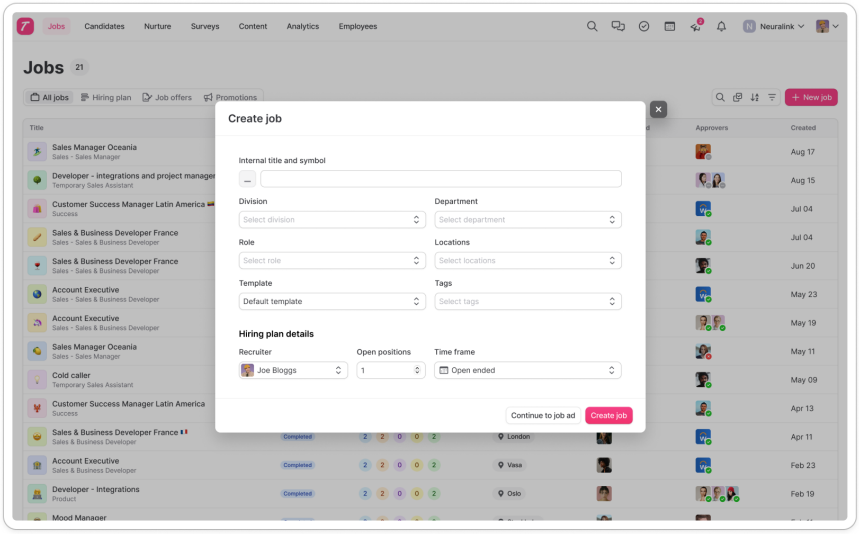
Figure 2.2: User flow of creating a job in the new design.

The most fundamental distinction between the two flows is that, in the current design, a job cannot exist independently of a job ad. In the redesign, the two are decoupled. Another key difference concerns the placement of two configuration features: the interview kit, which consists of interview questions associated with a given job, and stage configuration, which defines the recruitment stages candidates move through. Stage configuration can also include automated actions, known as triggers, which are activated when a candidate enters a specific stage. For example, when a candidate is moved from screening to an interview stage, a trigger can automatically send an email informing them that they have progressed to the next step. In the current flow, both the interview kit and stage configuration are set up within the job ad wizard, whereas in the redesigned flow they have been relocated to the applications overview page.

The primary focus of this research is the change in flow and user perceptions of the new job wizard. Figures 2.3-2.5 below present the current and redesigned versions of key pages.

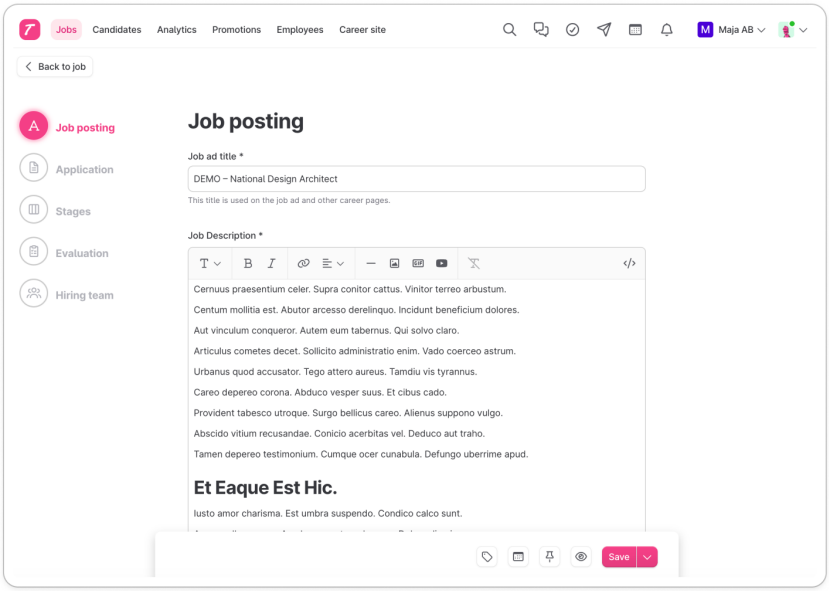


(a) The current start dialog box when creating a job.

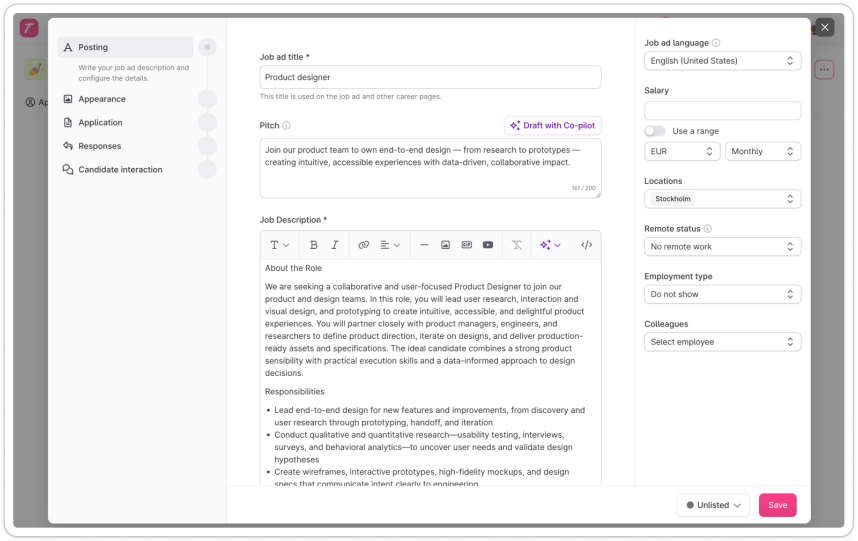


(b) The proposed new start dialog box design.

Figure 2.3: Comparison of the start dialog box layout for job creation.

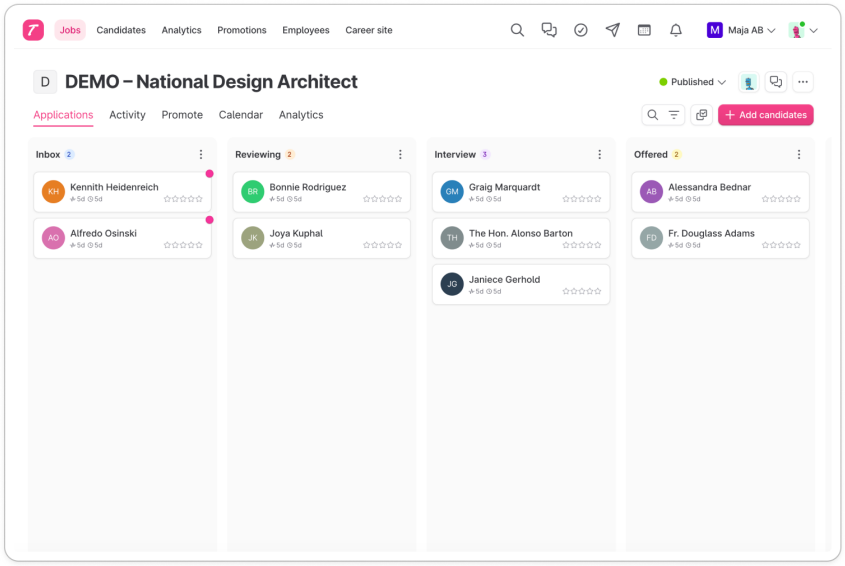


(a) The current wizard when creating a job ad.

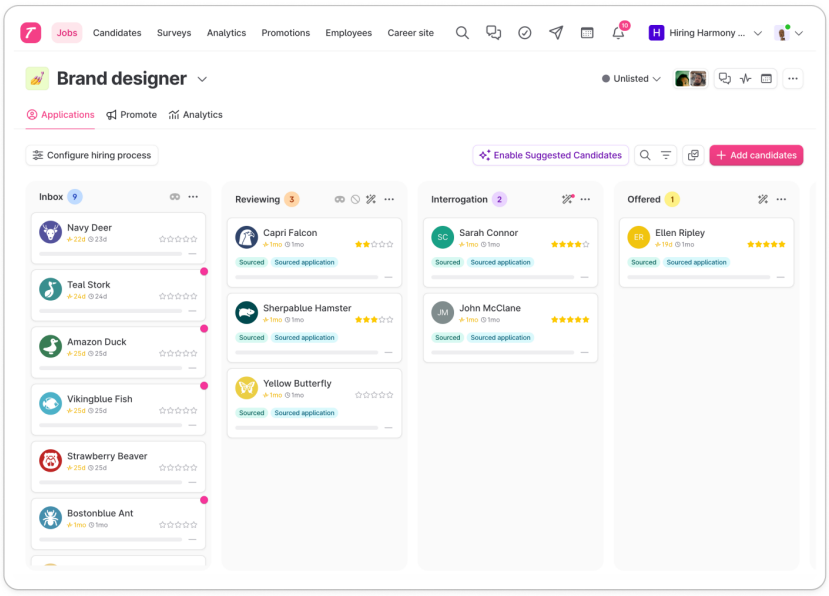


(b) The proposed new wizard design.

Figure 2.4: Comparison of the wizard layout for job ad creation.



(a) The current applications view.



(b) The proposed new applications view design.

Figure 2.5: Comparison of the applications view after creating a job.

Figure 2.3 compares the initial job creation dialog box in the current and redesigned flows. The current design, shown on the left, is minimal because most configuration is deferred to the subsequent job ad wizard. The redesigned version, shown on the right, requires the user to provide more information at this earlier stage. This reflects the redesign's separation of job creation from job ad creation.

Figure 2.4 shows the wizards used for creating a job ad in each respective design. The current wizard is presented as a full page and is reached directly from the start dialog box. In the new design, the wizard takes the form of a dialog box and can be accessed either from the start dialog box or via a dropdown menu after the job has been created.

Figure 2.5 presents the application views side by side. The two versions are largely similar, with mainly one key difference which lies in the button "configure hiring process".

While additional pages are affected by the redesign, the figures above have been selected to provide a foundational understanding of the changes most relevant to topics discussed in the following chapters.

2.1.3 Teamtailor's Internal Data Sources

Naturally occurring data refers to data that is produced as a byproduct of normal organizational activity, independently of any research process. Unlike interview transcripts or survey responses, which are created because a researcher decided to collect them, naturally occurring data would exist regardless of whether anyone planned to analyze it. It captures real behavior and real friction in authentic contexts, but it is not structured with research in mind.

Two of the internal data sources utilized as inputs for this study (Slack and Intercom) are naturally occurring in this specific sense. The company uses Slack for internal communication, including dedicated feedback channels where customer-facing teams collect and share product feedback informally. Intercom handles customer support through chat-based conversations between customers and support staff. Both sources provide an unstructured, historical record of real user friction within the existing product.

The third data source, the Figma redesign wireframes, is different in kind: it is not naturally occurring data. Because the redesign has not yet been deployed to production, no historical, user-generated operational data exists for it. Instead, the Figma wireframes serve as a design artifact. The subsequent evaluation data is generated through synthetic personas rather than live users. Because no naturally occurring data about the redesign can exist, evaluating it required a different kind of input altogether: the prototype itself, explored through synthetic personas. It is included here as a system input, but it represents a synthetic evaluation environment rather than a trace of natural organizational activity. The distinct methodological treatment of these naturally occurring and synthetic data is described in Chapter 4.

2.2 Evolution of B2B SaaS User Expectations

The past decade has witnessed significant growth of SaaS solutions within B2B environments. Modern B2B users increasingly expect consumer-grade experiences, challenging the assumption that enterprise software users will tolerate complexity in exchange for functionality [8].

This shift is evident in platforms like Teamtailor, where user expectations include:

E1. Minimal onboarding friction: Users expect to complete core tasks (e.g., creating a job posting) within minutes, without consulting documentation or training materials.

E2. Visual clarity: Modern B2B interfaces employ card-based layouts, drag-and-drop interactions and contextual guidance.

E3. Immediate feedback: Real-time validation, progress indicators and confirmation messages reduce uncertainty.

Meeting these expectations requires continuous user research and validation. However, as B2B SaaS platforms scale globally, traditional research methods face significant practical constraints.

2.2.1 Structural Characteristics of B2B SaaS UX

B2B SaaS UX differs from consumer UX in ways that matter for how research is designed and how findings are acted upon. Several differences are particularly relevant to this study.

B2B software typically supports *role-based, multi-user workflows* rather than isolated individual tasks. Actions performed by one user often have downstream consequences for others: a recruiter's job posting configuration affects how a hiring manager reviews candidates, which in turn affects what an administrator reports on. Research methods that focus on a single user in isolation risk missing the interdependencies that define actual use.

B2B procurement also often prioritizes feature coverage over user experience, as the buyer is rarely the end-user. Tools accumulate complex features to satisfy buyers rather than addressing user needs.

Lastly, the scale and diversity of user organizations vary widely. Teamtailor, for instance, supports companies ranging from fast-growing startups with a single in-house recruiter to enterprises with dedicated recruitment operations teams. A feature that works well for one organization may be difficult or unusable for another without configuration. This variation places an importance on continuous, diverse user input. Exactly the kind of input that resource-constrained UX teams find difficult to sustain [2] and that naturally occurring data could potentially provide at a relevant scale.

3

Theory

This chapter establishes the theoretical foundations of the study. It opens with Human-Computer Interaction as the broadest disciplinary frame, with particular attention to Human-AI Collaboration in order to understand the relationship between researcher and model. From there it moves on to Interaction Design and the conceptual frameworks for treating naturally occurring data as legitimate research material. The chapter closes with an examination of Large Language Models and their application in UX research contexts.

3.1 Human-Computer Interaction

HCI is an interdisciplinary field about the design and implementation of interactive computing systems for human use [9]. HCI emerged as a field in the early 1980s when the arrival of personal computing made the quality of the human-machine interface a practical concern for a broad population of non-expert users [10]. Early HCI research focused primarily on whether users could complete tasks efficiently and without error. Over the following decades the field evolved and shifted its focus from isolated task performance toward a richer understanding of how people experience and derive meaning from interactive systems in everyday and organizational contexts [9]. This broadening is directly relevant to the present study as the questions this thesis asks about how AI systems can support UX research workflows are fundamentally HCI questions. The thesis explores how a specific class of human-computer interaction, between a researcher and an LLM, can be designed and understood.

3.1.1 Human-AI Collaboration

A central theoretical commitment of this thesis is that the LLM functions as a collaborator within the analytical process rather than as a tool that performs analysis on the researcher's behalf. This distinction is a methodological choice that draws on an established body of work in human–computer interaction and AI design.

The foundations of this view can be traced to Licklider's vision of "man–computer symbiosis" [11], which proposed that the productive relationship between human and machine lies in complementary capability rather than substitution. Horvitz later formalized related ideas through the principle of *mixed-initiative interaction*, arguing that effective human–AI systems are those in which control of the interaction

is shared and context-sensitive, with each party contributing where they are most competent [12]. Shneiderman's framing of *human-centred AI* extends this further, arguing that the goal of AI design should be trustworthy systems that amplify rather than replace human capability [13].

Two distinctions from this literature are particularly useful. The first is between *augmentation* and *automation*. Automation frames AI as a substitute: a task done by a machine that would otherwise be done by a human. Augmentation frames AI as an amplifier: a process in which machine capability is combined with human judgment to produce outcomes that neither could achieve alone. This thesis operates firmly in the augmentation area. The LLM handles material at a scale and pace that a single researcher cannot, while the researcher brings the contextual and interpretive judgment that the LLM cannot.

The second is between *task-level* and *process-level* collaboration. At the task level, a human and an AI each perform discrete steps and hand off to one another. At the process level they interact iteratively. The researcher shapes what the AI does, the AI's output shapes what the researcher investigates next and the analytical trajectory emerges from this back-and-forth.

Literature also documents how human–AI collaboration can fail when users defer excessively to AI outputs (a phenomenon known as *automation bias*) when system limitations are poorly surfaced, or when the interaction pattern implicitly treats the AI as authoritative rather than as a contributor [14]. These failure modes are particularly relevant in qualitative research, where the interpretive authority of the researcher is central to the credibility of the findings. Maintaining a genuinely collaborative stance is therefore a necessary condition for the study's validity.

3.2 Interaction Design and UX Research

Interaction design is a discipline concerned with shaping the behavior of products and systems to support human goals [15]. While Interaction Design encompasses both physical and digital contexts, it has developed significant overlap with HCI in the domain of interactive software systems [9]. Within interaction design, UX research constitutes the empirical foundation of the discipline. UX research is concerned with understanding the behaviors and needs of users through observation and other feedback methodologies [15]. Its core purpose is to inform design decisions by grounding them in empirical evidence rather than assumption. Within the UCD lifecycle, research activities span the full product development process from early discovery through to post-launch evaluation.

3.2.1 Naturally Occurring Data in Qualitative Research

A meaningful distinction in qualitative research methodology is that between *elicited data* and *naturally occurring data*. Elicited data is produced through direct researcher–participant interaction: the interview question prompts the answer, the survey question frames the response, etc. Naturally occurring data, by contrast, exists indepen-

dently of the research process. It is data that would have been created regardless of whether a researcher planned to analyze it [16].

Examples of naturally occurring data include forum posts, customer support tickets, meeting transcripts, and messages exchanged in organizational communication tools. Silverman argues that naturally occurring data has important advantages as it avoids the reactivity that can distort elicited data (participants behaving differently because they know they are being studied), and it captures behavior and language in authentic contexts [16].

In a B2B SaaS company, a considerable volume of naturally occurring data accumulates as a by-product of normal operations: users report issues and requests in support channels, internal teams discuss product decisions in messaging platforms, and designs and their justifications are iterated in design tools. Despite the abundance of this material and its apparent relevance to UX concerns, it is rarely treated as a systematic research input. Organizational data does not accumulate in a single, structured repository. It is distributed across communication tools, ticketing systems, knowledge bases and design files, each with its own conventions and access patterns. Unlike interview transcripts which are bounded and purposive, naturally occurring data is continuous and produced at a volume that exceeds the capacity of manual qualitative analysis. In resource-constrained UX teams, where maintaining continuous research is already difficult alongside fast product cycles [2], allocating analyst time to such material has historically been hard to justify.

Elicited data also carries with it the researcher's framing: the question defines the topic and the range of relevant response. Naturally occurring data does not. A support ticket, a Slack thread or a call log must be interpreted without a pre-established frame. Doing so requires domain knowledge to distinguish, for example, a lone complaint from a recurring usability pattern or an internal design debate from a confirmed design rationale.

Recent developments in large language models change this as LLMs now are able to process heterogeneous, unstructured text at a scale that is impractical for a single analyst, reducing the cost of engaging with continuous data streams [17]. In addition, LLMs can be primed with contextual information through system-level instructions, allowing domain knowledge to be encoded into the analytical process rather than held implicitly by the researcher. This does not eliminate the interpretation barrier but it lowers it as the model can be oriented toward the analytical lens relevant to the research question before any data is introduced.

A further consideration concerns the scope of the category itself. The methodological tradition from which the term *naturally occurring data* is drawn has been developed primarily in relation to public or semi-public corpora such as broadcast media, public forum posts, recorded conversations and institutional records [16]. Applying the category to internal organizational data, as this thesis does, is a conceptual extension rather than a direct inheritance. Internal data shares the defining property of having been produced independently of the research process, but it is typically more fragmented, more informal and subject to different access and ethical considerations than naturally occurring data in the classical sense. This thesis therefore treats in-

ternal organizational data as naturally occurring in the methodological sense, while acknowledging that the category is here being extended into a context in which its practical advantages and constraints are not yet well understood.

3.3 Large Language Models

Large Language Models (LLMs) are a class of artificial intelligence systems trained on vast corpora of text with the objective of learning to predict and generate coherent, contextually appropriate language. Contemporary LLMs are built on the *transformer* architecture, introduced by Vaswani et al., which uses a mechanism called self-attention to model relationships between words across long sequences [18]. Through exposure to hundreds of billions of tokens of text, these models develop broad linguistic competence and a form of implicit world knowledge that enables them to perform a wide range of language tasks, such as summarisation, classification, question answering and qualitative reasoning, without task-specific retraining. The most prominent LLMs in contemporary use include OpenAI's GPT series, Google's Gemini, and Anthropic's Claude [19]. Claude, the model used in this study, is developed by Anthropic with an explicit focus on safety and interpretability. It is designed to follow nuanced instructions and produce outputs that reflect stated constraints and contextual priorities [20].

3.3.1 Prompt Engineering

While LLMs are capable of responding to open-ended natural language input, the quality and relevance of their outputs are shaped by how queries are formulated. This practice, known as *prompt engineering*, involves crafting instructions that guide the model toward a desired output format, reasoning style, or level of specificity [21].

In the context of research, prompt engineering takes on methodological significance. A poorly constructed prompt may yield generic or superficial outputs; a carefully designed one can orient the model to apply a specific analytical lens, adhere to a particular framework, or flag uncertainty rather than fabricate conclusions. Researchers using LLMs for qualitative analysis must therefore treat prompt design as an explicit methodological decision similar to the design of an interview guide or survey structure.

The sensitivity of LLM outputs to prompt formulation also introduces a reproducibility concern. Unlike a fixed algorithm, the same model may produce different outputs in response to subtly different phrasings of the same underlying question. This does not exclude LLMs from being used as research tools, but it requires researchers to document their prompts and account for potential variability in the research design.

3.3.2 Contextual Priming via System Instructions

Beyond individual prompts, LLMs can be configured at a higher level through persistent system instructions that establish the context and constraints for an entire session. In this study, this is applied through a `CLAUDE.md` file. This is a structured

document provided to Claude at the outset of each analysis session [22]. It specifies the research context, relevant background on Teamtailor and its users, the analytical framework to apply and the expected output format.

This approach effectively encodes domain knowledge and methodological constraints into the model's context, reducing the risk of generic or irrelevant outputs. It also enables a degree of consistency across sessions, as the same foundational instructions are applied each time.

3.3.3 Model Context Protocol

Model Context Protocol (MCP) is an open protocol introduced by Anthropic that standardizes how AI language models connect to and interact with external tools and data sources [23]. Prior to MCP, integrating an AI model with external services required custom, one-off implementations for each connection. MCP addresses this by providing a universal interface through which a model can communicate with any number of external services in a consistent way.

An MCP server exposes a defined set of tools that the AI model can invoke during a conversation or task. When the model determines that an external resource is needed, it issues a structured call to the relevant server, which executes the requested action and returns the result. This allows the model to interact with real-world systems such as databases, APIs, or productivity tools, without those integrations being hardcoded into the model itself.

In the context of this thesis, MCP was used in order to retrieve and analyze data from sources directly. This architecture was central to the LLM system, as it allowed the model to access the naturally occurring data described in the previous section without requiring manual data extraction or preprocessing.

3.3.4 LLMs for Qualitative Analysis

A recent body of work has begun to examine how LLMs perform when applied to qualitative analysis tasks that have traditionally required human interpretive work. This literature is still young, but several patterns have emerged.

De Paoli demonstrates that GPT-based models can produce thematic analyzes of semi-structured interview data that are recognizable as such, identifying plausible codes and organizing them into themes [4]. However, De Paoli also notes characteristic limitations: the model tends toward descriptive rather than interpretive themes, surfaces themes at inconsistent levels of abstraction and lacks the contextual sensitivity that a human analyst develops through immersion in the data. The output is recognizable as thematic analysis without necessarily matching the depth of a human-led version.

Xiao et al. apply GPT-3 to an expert-drafted codebook for deductive coding, reporting fair-to-substantial agreement with expert coders (Cohen's $\kappa = 0.38$–0.61 across two coding dimensions) [24]. Agreement varied with prompt design and they observe the largest when the codebook was accompanied by worked examples rather

than code descriptions alone. This supports the view that LLM output quality is dependent on how the model is instructed rather than a fixed property of the model itself.

Where the existing literature is thinnest is in the analysis of naturally occurring data. Most published LLM-assisted qualitative studies to date work from purpose-collected material: interview transcripts, survey responses, focus group recordings. How LLMs handle fragmented, multi-source organizational data, where context is implicit and relevant signal is mixed with operational noise, remains largely unexplored. This thesis contributes to that gap by applying LLM-assisted analysis to exactly such data.

3.3.5 Synthetic Users in UX

A particularly visible use case for LLMs in UX research is the *synthetic user*: an artificial entity generated by an LLM that aims to emulate human behavior and produce interview-style responses, attitudinal feedback or simulated reactions in place of recruiting real participants. Commercial services market synthetic users as alternatives to traditional research participants, promising reduced cost and increased efficiency, with one provider going as far as to advertise "user research without the user". A 2025 industry survey cited by Salminen et al. found that 69% of market researchers had already used synthetic data in their work, and 70% expected synthetic responses to constitute more than half of all consumer data collection within 3 years [25]. The appeal is rooted in real constraints. They also report that more than half of UX teams face time pressure, nearly half face budget pressure, and around a third struggle to recruit participants.

Critical engagement with synthetic users in the academic literature has been overwhelmingly cautious and several concerns are directly relevant to the present study. The first is validation. Synthetic user services typically operate as proprietary black boxes. Vendors claim periodic comparison against live testing without disclosing the methodologies used, which makes independent verification effectively impossible [25]. Compounding this, Chapman has argued that LLMs are unlikely to do a decent job of simulating specific groups of people in specific contexts, which is precisely the purpose of user research in the first place [25]. Another concern is what Salminen et al. term *circularity*: because LLMs are pattern-synthesis engines, the user representations they produce cannot exceed the patterns present in their training data and they tend to default to stereotypical and generic characterisations rather than the specific, situated experiences user research is designed to surface [25]. Another concern is *convincing mimicry*: synthetic outputs are linguistically fluent and superficially plausible, which raises the risk that designers trust them precisely because they read well, even when the content lacks validity [25].

These concerns are reinforced by empirical work on LLM behavior. Tao et al., in an evaluation of five Generative Pre-training Transformer (GPT) models against the Integrated Values Surveys covering 107 countries, find that the cultural values expressed by the models cluster closely with English-speaking and Protestant European countries and are most distant from African-Islamic ones [26]. Cultural

prompting reduces but does not eliminate this bias, and for 19–29% of the countries studied it fails to improve alignment or actively makes it worse [26]. The implication for synthetic users is direct: a model whose default responses approximate Finland or the Netherlands cannot be assumed to represent users from contexts outside that distribution, however confidently the prompt instructs it to do so. This aligns with broader observations that AI-generated user representations may reflect a WEIRD (Western, Educated, Industrialised, Rich, Democratic) worldview rather than the cultural diversity of real user populations, and that they struggle to capture the unpredictability and qualitative richness of behavior that interviews and observational studies make visible [27].

Not all of the literature is dismissive. Xiang et al. develop SimUser, an LLM-based tool for generating heuristic usability feedback during prototyping and report that simulated users can match human feedback with similarity ranging from 35.7% to 100% depending on the user group and usability category, while explicitly framing the tool as a complement to rather than replacement for human usability tests [28]. A systematic review by Pehar of 52 studies on synthetic users in human-centred design reaches a comparable conclusion: synthetic users can accelerate early design iterations and broaden creative exploration, but subjective qualities such as emotional resonance, cultural nuance, and genuinely novel user behavior remain difficult to synthesize, leading most authors to advocate hybrid approaches in which AI augments rather than replaces real user input [29]. The overall picture across the literature is therefore that synthetic users' useful range is narrower than commercial framing suggests. Their failure modes disproportionately affect exactly the populations and insights that user research is most needed to surface.

3.3.6 Industry Tooling

Alongside the academic discussion of AI in UX research, a parallel and much faster-moving development has unfolded in commercial tooling. Research repositories such as Dovetail and Marvin now ship with built-in AI features for transcription, tagging, clustering, and theme generation. Unmoderated testing platforms such as Maze and UserTesting embed AI-assisted analysis of session recordings and open-text responses. Beyond purpose-built tools, general-purpose LLMs (ChatGPT, Claude, Gemini) are increasingly used ad hoc by researchers for tasks such as affinity mapping, survey synthesis and early-stage insight generation. By 2025, a substantial majority of UX practitioners report using AI in at least part of their research workflow, with adoption concentrated in the analysis and synthesis phases [30].

A characteristic of this tooling landscape is that commercial AI features overwhelmingly target *elicited* data such as interview transcripts, survey responses, usability session recordings that has been produced specifically for research purposes. Naturally occurring organizational data such as support tickets, internal messaging threads and design documents is occasionally acknowledged in tool marketing as a potential source of insight, but dedicated tooling for analysing it systematically as part of UX research is considerably less mature. This asymmetry reinforces the gap the present study seeks to address.

Furthermore, the claims made about AI-in-UX tooling tend to emphasize speed and scale rather than analytical depth or validity. Marketing language foregrounds time saved, transcripts analyzed and themes surfaced, while questions about whether the resulting themes are sound, whether they reflect the underlying data accurately, and how a researcher should calibrate trust in them are typically left to the user. This is consistent with the limitations Lu et al. identify, where simplistic automation can undermine rather than support the analytical workflow [31].

3.3.7 Potential and Limits of AI in UX

Literature suggests that AI possesses the capacity to automate significant portions of the design process by using large datasets to build robust models that mimic human design workflows. As observed by Wells, the primary value of AI lies in its ability to handle repetitive tasks, thereby allowing practitioners to focus on the more creative aspects of the design process [32]. AI can also enhance the effectiveness of digital artifacts by utilizing data-driven insights to inform strategic design decisions.

Despite this potential, Stige et al. write that the complete removal of the human designer is unlikely [33] . Specifically, the initial stages of the User-Centred Design process, such as understanding the context of use and specifying requirements, remain resource intensive and require direct interaction with end-users. While the resource-demanding nature of these stages makes them ideal candidates for automation, they remain significantly under-researched in current literature.

While the benefits of AI in UX research are discussed theoretically, empirical studies reveal that simplistic automation often fails to meet the needs of practitioners. Lu et al. identify several critical limitations [31]:

L1. Workflow Integrity: When qualitative analysis is automated without sufficient refinement, the outcome can be counterproductive: established research workflows are disrupted, teams face coordination trouble, and both the speed and quality of the analytical process can decline rather than improve.

L2. Obstruction of Empathy: AI systems designed to provide "synthetic user" information are often rejected by practitioners because they can obstruct the core empathy-building goal of UX research.

L3. Reinforcement of Stereotypes: Because AI draws on statistical patterns, the user profiles it generates tend to reflect average behavior rather than real diversity. This risks reinforcing existing assumptions about users instead of surfacing the nuanced differences that design decisions depend on.

Beyond these procedural issues, there is a deeper cognitive challenge: AI systems often struggle with complex problem-solving that is highly contextual, social, and relational [32]. Consequently, the literature calls for a shift toward a "human-centred AI perspective" that supports the specific mindsets and goals of designers rather than attempting to automate the entire UX process.

The areas covered in this chapter establish the theoretical foundation from which this study's design is derived. HCI and Human-AI Collaboration provide the framing of

AI as a collaborator that amplifies rather than replaces human analytical capacity. Interaction design and UX research supply the methods and frameworks through which both tracks produce and interpret findings. The literature on large language models and AI in UX research establishes both what LLMs can do in qualitative analysis contexts and where their characteristic limitations lie. Chapter 4 describes how these theoretical positions are translated into concrete methodological choices across the two research tracks.

4

Methods

This thesis employs two parallel research tracks with distinct methods and purposes that together answer the research questions. The tracks are not symmetric: Track 1 produces empirical findings, through surveys and usability testing with real users, while Track 2 yields a methodological contribution in addition to empirical findings. Track 2 applies a multi-agent LLM system to Teamtailor's own existing data and artifacts in two modes: analyzing naturally occurring operational data (Slack and Intercom) to study the current product, and running synthetic-persona evaluations of the redesign prototype in Figma. Because the LLM system itself is an outcome of the study, Track 2 spans both building and applying it. The methodological knowledge gained through its development is documented in Chapter 5, while the empirical findings are produced by a single run of the finalized system, reported in Chapter 6.

Track 1 consisted of manual user research where traditional qualitative user research was conducted on both the current design and the redesigned prototype. This track included the following approaches:

- One product survey with 437 respondents (current design)
- 7 usability tests (new design)

Qualitative methods like usability testing capture the why behind user behavior and domain expertise that users bring to their work. These methods are essential for understanding the situated context of how job creation actually happens at different companies and across different user roles. The survey complements this by capturing perceptions at scale. The empirical findings from Track 1 are also described in Chapter 6.

Track 2 was the AI-assisted research where we developed a multi-agent LLM system to systematically process naturally occurring organizational data and simulate user interactions with the prototype. This track included:

- Automated content analysis of Slack feedback channels and Intercom support conversations (current design)
- Synthetic persona walkthroughs of the Figma prototype (new design)

LLMs can process unstructured text at a scale impractical for manual analysis. This makes it possible to surface recurring patterns that appear across many user interac-

tions but might never emerge in a small number of interviews. LLM agents can also systematically navigate a prototype following predefined task scenarios, producing consistent evaluation data. Together, these provide a pattern-detection layer that complements human research. The two tracks are intentionally different because this thesis is fundamentally about how LLM-assisted and manual research complement each other, not how one replaces the other.

4.1 Research Methods and Their Trade-offs

UX research methods are commonly divided into two broad categories: *generative* and *evaluative*. Generative research, which includes methods such as semi-structured interviews and contextual inquiry, aims to uncover new understanding about users' lives, goals and contexts. It is exploratory in nature and produces rich, qualitative insight. Evaluative research, by contrast, assesses how well an existing or proposed design meets user needs through methods such as usability testing, heuristic evaluation, and surveys [15].

There is also the distinction between qualitative and quantitative data. Qualitative methods such as interviews provide deep contextual understanding but are resource-intensive: recruiting participants, conducting sessions and analyzing transcripts can require significant time investment, often limiting sample sizes. Quantitative methods such as surveys scale more easily and allow for statistical analysis, but tend to surface *what* users do rather than *why*. In resource-constrained B2B SaaS environments, these trade-offs make it difficult for UX teams to maintain continuous research cadences alongside fast product development cycles [2].

4.1.1 Research Through Design

Research Through Design (RtD) is a research approach developed within the interaction design and HCI communities in which the act of designing constitutes the research process itself [34]. Introduced by Zimmerman et al., RtD positions design artifacts as the primary vehicle through which knowledge is produced and communicated. Unlike research *about* design, which studies existing artifacts and practices from the outside, or research *for* design, which generates knowledge intended to inform future design work, RtD produces knowledge *through* the iterative process of making [34].

RtD has been applied across a range of contexts in HCI and interaction design research, from the design of interactive systems and physical computing artifacts to methodological tools and research instruments [34]. Its application to AI-assisted research pipelines as design artifacts is less established. This study extends the RtD framing to treat the LLM system itself as the design artifact. The pipeline (its agent architecture, prompt design, contextual priming strategy and MCP integration) provides answers to RQ2 and RQ3, together with the conditions under which it succeeds and fails. The empirical findings that emerge from running this pipeline in turn, help answer RQ1. The contribution of the thesis therefore lies not only

in what the pipeline produces but in the pipeline itself, which is intended to be transferable to comparable B2B SaaS contexts.

4.1.2 Thematic Analysis

A central method for making sense of qualitative data in UX research is thematic analysis. Braun and Clarke define thematic analysis as a method for identifying and analyzing patterns within qualitative data [35]. Their widely adopted six-phase framework consists of: familiarising oneself with the data, generating initial codes, searching for themes, reviewing themes, defining and naming themes, and producing the final report. Thematic analysis is particularly well suited to UX research because it is flexible enough to work across a wide variety of data types (from interview transcripts and open survey responses to written feedback and support correspondence) without being tied to a specific theoretical model. This flexibility is directly relevant to the present study as it provides the methodological lens through which both manually collected data and AI-generated outputs can be interpreted. Thematic analysis involves a degree of interpretive judgment at each stage. The researcher's choices about what counts as a meaningful pattern, how to name it and how to bound it relative to other themes are not mechanical decisions but reflective ones. This characteristic becomes significant when an AI system performs analogous operations, raising questions about the nature and validity of the patterns it identifies.

4.1.3 Personas

Personas are a method in interaction design for representing the diversity of a user population in a form that is concrete enough to inform design decisions. Introduced by Cooper as a response to the tendency of product teams to design for a generic or idealized user, a persona is a profile constructed from empirical research that captures the goals, behaviors and contextual constraints of a particular user segment rather than any single individual [36]. Pruitt and Adlin elaborate on this foundation by treating personas as living research artifacts that document analytical work done at one point in the design process and make it available to future stages of that process [37]. This means the value of a persona lies not only in what it communicates to a design team but in how it carries domain knowledge forward across the different phases of a project.

In conventional UX practice, this knowledge transfer operates within the design team. A researcher conducts interviews, synthesizes the findings into personas and shares them with designers who use them to evaluate design decisions against a specific user's perspective. This thesis extends that function in a direction that the existing literature does not address. Here, personas serve not only as communication artifacts within a research team but as the operational instructions that define the perspective of an LLM agent. Each persona description is provided to the model as a system-level prompt and the model reasons about and evaluates the prototype from within that defined perspective. The quality of the agent's output is therefore dependent on the quality of the persona. A description that is behaviorally specific and grounded in observed user patterns will produce more detailed findings, while

a generic or abstractly defined persona will tend to produce the kind of averaged, stereotypical observations that Salminen et al. identify as a characteristic limitation of AI-generated user representations [25].

4.1.4 Contextual Interviews

To understand how users interact with the existing job posting flow in practice, five contextual interviews were conducted with Teamtailor customers. Contextual inquiry was chosen as the primary method for this phase because it enables observation of users performing real tasks in their authentic work environments while preserving the ability to ask clarifying questions as behavior unfolds [38]. Unlike retrospective interviews, in which participants reconstruct past experiences from memory, contextual inquiry captures actual workflow behavior alongside the reasoning and workarounds that accompany it [39]. This makes it particularly well-suited to investigating a multi-step task such as job creation, where habitual patterns and minor friction points are unlikely to be surfaced through direct questioning alone.

A defining feature of this type of qualitative work is that it relies on relatively few participants, since each session generates a large volume of detailed data and is analytically demanding. Therefore it is important to be mindful when choosing participants and make sure they reflect the wider user base.

4.1.5 Surveys

Surveys offer the inverse trade-off to contextual interviews. They cannot capture observed behavior or the reasoning behind it, but they can reach a far larger and more diverse population than is practical to interview, and they support distributional claims that interview data cannot. In a B2B SaaS context, this makes surveys particularly useful for understanding whether patterns observed in a small number of interviews generalize across user segments.

The intercept survey design used here captures responses in close temporal proximity to the experience being evaluated [40], which reduces the recall bias associated with retrospective surveys. The survey consisted of 7 questions combining multiple-choice, Likert-scale and open-ended response formats. The combination of closed and open-ended items was chosen to enable both distributional analysis across the full response set and thematic analysis of the qualitative responses. The full set of survey questions is provided in Appendix A.2.

4.1.6 Usability Testing

Following the current-design research, usability testing was conducted on a high-fidelity Figma prototype of the redesigned job posting flow. Usability testing is an evaluative method in which representative users attempt to complete defined tasks using a product or prototype while their behavior and reactions are observed [41]. It is distinguished from generative methods such as interviews by its focus on observable task performance rather than reported attitudes. This makes it an appropriate

method for identifying specific friction points and breakdowns in a proposed design before it reaches production [42].

In this study, usability testing was conducted on a Figma prototype that was high-fidelity in appearance but limited in interactivity, since not all click targets in the design were wired up. The findings it produces reflect the prototype as encountered, which means observed friction can stem from constraints of the prototype itself as well as from the design under evaluation. In addition to this, the small sample sizes typical of usability testing mean that a friction point not raised by a participant may simply not have been encountered in the tasks performed.

5

Process and Analysis

This chapter documents how the two research tracks introduced in Chapter 4 were carried out. Contextual interviews came first and informed both tracks: they grounded the survey and usability test design in Track 1 and produced the personas used as agent instructions in Track 2. The remainder of the chapter follows the two tracks in turn, with Track 2 documented as an iterative construction process in line with the Research Through Design framing (Section 4.1.1). Figure 5.1 provides an overview of the research tracks.

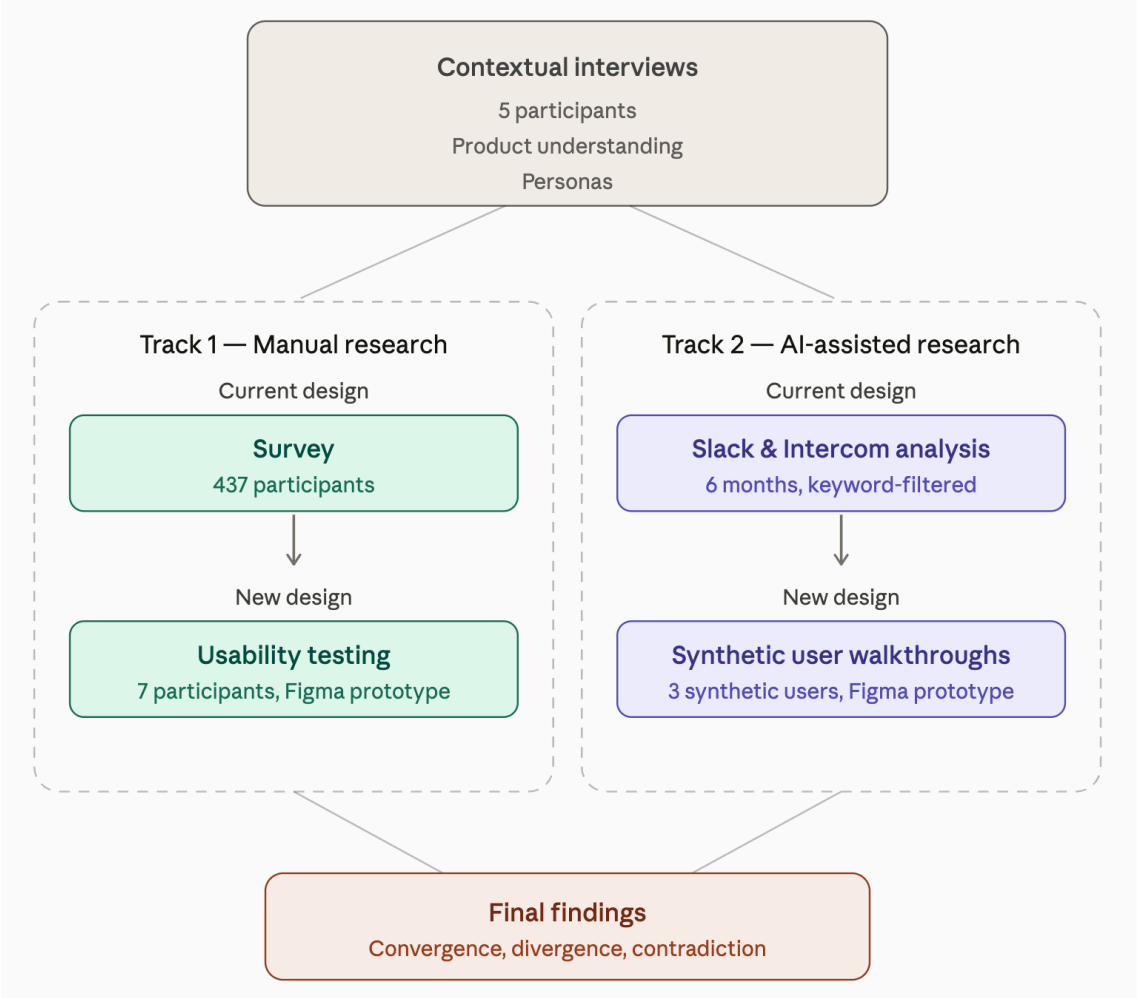


Figure 5.1: Overview of the research conducted in this Master thesis.

Some collected material from the manual research falls outside the scope of this thesis or is subject to confidentiality with Teamtailor and is not reported in full. Similarly, findings produced by the multi-agent LLM system before the final architecture are not reported in full, as they served to inform subsequent iterations. The findings reported in the results are presented in full in Appendix B.1.

5.1 Contextual Interview Sessions

All contextual interviews were conducted remotely via Zoom. Participants were asked to share their screen and to use the live Teamtailor product throughout the session. Remote video-mediated interviews have been established as a methodologically credible alternative to co-located sessions, provided that the participant can share their screen and perform the task in their natural work environment. In practice, doing it remotely often increases ecological validity by placing the user at their own machine within their normal work context [43]. Participants received a brief introduction to the session format and were informed that their participation was voluntary and that data would be used for academic and product research purposes only.

Each session followed a semi-structured format, structured in 3 stages. The session opened with introductory questions about the participant's role, their frequency of use, their typical recruitment workflows, and their level of familiarity with Teamtailor. This stage was designed to establish the organizational and behavioral context needed to interpret what followed. Participants were then asked to share their screen and to create a job posting from start to finish as they would in their normal work, while thinking aloud throughout. The think-aloud protocol, in which participants verbalize their thoughts continuously as they perform a task, is a well-established method in HCI and usability research for making the otherwise invisible reasoning process clear to the observer [42]. Participants were encouraged to narrate freely rather than explain or justify their actions, and the interviewer intervened only to ask brief clarifying questions when behavior was ambiguous or when a notable reaction was not talked about. The session concluded with reflective questions covering recurring pain points and functionality considered essential to the job creation process. The full interview guide is provided in Appendix A.1. Sessions were documented through live note-taking by the researcher. Where participants gave verbal consent, sessions were also transcribed to support a more detailed analysis.

The analysis was conducted iteratively after each session rather than deferred until all interviews were complete. Each session produced a structured summary delivered to Teamtailor comprising 3 components: a friction point inventory documenting specific breakdowns observed or reported during the task, a feature interaction map tracing which product features the participant engaged with and in what sequence and a set of key insights synthesizing the patterns most relevant to the product team. This per-session structure follows the principle of progressive analysis in qualitative fieldwork, in which early analytical work informs subsequent data collection rather than waiting until the full dataset is assembled [44]. In the context of this thesis, the outputs of this stage were used as input to Teamtailor's product team and as

the basis for scoping the survey instrument and the usability test tasks described in the following sections.

One example of how the contextual interviews shaped the subsequent research is the inclusion of the survey question "How would you describe your role in job postings at your company?" The interviews made it clear that participants had very different experiences depending on how much they relied on templates and how frequently they used the product. At one large enterprise retailer, store managers only entered the product to post a job from a template, without modifying anything. At a smaller company, the participant valued the ability to customise the flow and make it more personal. Understanding how participants used the product was therefore necessary to interpret their survey responses meaningfully. The interviews also showed that Likert scales alone would not be sufficient, since participants with similar overall ratings often had very different reasoning behind those ratings.

In the usability testing, automatic triggers (actions that fire automatically when a candidate is moved to a new stage) were perceived very differently by different participants. This, combined with the fact that the process of adding them was being redesigned, led to the inclusion of a dedicated task focused entirely on triggers.

5.2 Persona Construction

The contextual interviews established that Teamtailor's user base is not uniform and that the differences between user segments are significant enough to produce different experiences of the same interface. A recruiter at an enterprise company using Teamtailor daily approaches the job creation flow with different expectations and failure modes than an HR generalist at a small company who opens the product a few times a year.

This observation was reinforced through conversations with the Teamtailor product team and by attending internal design discussions during the thesis project. It became clear that how to serve different user segments was an active concern for the product team. The tension between the needs of small-to-medium businesses (SMBs) and enterprise customers shaped product discussions and was reflected in the redesign decisions being made during the thesis period. This context made apparent that any analytical approach reasoning from a generic or averaged user perspective would produce findings of limited value to the team.

The 3 personas were constructed from the patterns that emerged across the contextual interviews and survey data. Each persona was grounded in the specific behavioral segments the research identified and each persona was used as direct contextual input to the synthetic users in Track 2.

P1: Small-to-Medium-Business HR Manager

The SMB HR Manager works at a thirty-person e-commerce startup and carries sole responsibility for all HR activity, including onboarding, payroll and hiring. Job postings are infrequent, occurring 4 to 5 times per year, which means the product

is encountered as a relatively unfamiliar tool on each visit rather than as a practiced daily environment. Technical confidence is limited and the persona's primary expectation of the interface is that it should be quick and self-explanatory. When a step is unclear, the characteristic response is to begin clicking through available options in the hope that something produces the desired outcome. This exploratory fallback behavior is a direct signal of a low tolerance for ambiguity and an interface that relies on prior knowledge or familiarity to navigate.

P2: Enterprise Recruiter

The Enterprise Recruiter is a full-time recruiter at a 600-person SaaS company who uses Teamtailor as a core work tool every day. At a volume of twenty to thirty job postings per month, this persona has developed strong expectations about how recruitment workflows should be structured and where specific functionality should be located. Speed and efficiency are the dominant concerns. The cost of any unnecessary step, obscured option or workflow interruption is severe because it shows up across a high volume of repeated interactions. Unlike the SMB HR Manager, frustration here does not produce exploratory clicking but impatience with tooling that introduces friction the persona considers avoidable. This persona knows what is needed and expects the product to provide it with minimal friction.

P3: Mid-Market Hiring Manager

The Mid-Market Hiring Manager is an engineering manager at a hundred-and-fifty person financial tech company who opens Teamtailor approximately once per quarter when the team needs to grow. Recruitment is not a core competency. The persona is comfortable with software in general but does not have familiarity with recruiting terminology or the conceptual model that defines an ATS workflow. Navigation is driven by what appears logical rather than by product knowledge, which makes the persona's experience highly sensitive to whether the interface's structure matches intuitive expectations. When it does not, confusion follows rather than frustration and the persona is likely to stall at decision points that a more experienced user would resolve immediately.

5.2.1 Rationale for Segment Selection

The 3 personas span two dimensions that structure Teamtailor's customer base: company size and depth of product familiarity. The SMB HR Manager represents the small, infrequent-use end of the spectrum; the Enterprise Recruiter represents the high-volume, expert-user end; and the Mid-Market Hiring Manager occupies a middle position characterized by moderate company size and low recruiting expertise. Crucially, the 3 personas also represent meaningfully different failure modes when encountering a new or unfamiliar design. The SMB HR Manager resorts to exploratory clicking; the Enterprise Recruiter loses patience with unnecessary steps; the Mid-Market Hiring Manager stalls when interface logic does not match mental models. Holding all 3 perspectives simultaneously during prototype evaluation therefore produces a richer picture of where a design is brittle than any single user

profile could, and ensures that findings are not implicitly optimized for one segment at the expense of others.

5.3 Track 1: Manual User Research

Manual user research was conducted in two stages with distinct roles. The first stage was exploratory where contextual interviews were used to build a foundational understanding of the product, of how users approach the job posting flow in practice and of which aspects of the experience were most important for them. This understanding was not treated as a body of findings in itself, but as the basis for designing the subsequent research instruments. The second stage was evaluative: a survey and usability testing were used to generate the specific findings reported in this thesis, with the survey addressing the current design at scale and usability testing examining the redesigned flow on a high-fidelity prototype. The combination was selected so that the evaluative methods could be grounded in a prior understanding of how the product is actually used, rather than relying on assumptions about which behaviors and pain points were worth measuring [15].

5.3.1 In Product Survey

After the contextual interviews, a survey was deployed within the product and triggered automatically after users completed a job posting.

Responses were collected from 437 participants over the course of a week through Teamtailor's customer service tool, Intercom. The qualitative answers were transferred onto digital post-its in a shared workspace, and simultaneously cleaned to remove duplicate and redundant entries, for example identical responses submitted to different questions by the same user. The cleaned dataset was then analyzed thematically. The Likert-scale responses were collected for Teamtailor's internal use and are not analyzed in this study. Only the 244 qualitative answers form the basis of the analysis. In Figure 5.2 is the full cleaned qualitative dataset before grouping sorted by question. The questions from the survey can be found in Appendix A.2.

Figure 5.3 is after grouping by the feature they are referring to. Individual post-its are intentionally illegible to preserve participant anonymity. The figures are included to illustrate the scale and structure of the analysis rather than its content.

The analysis was conducted collaboratively by both researchers. Each response was first read independently, after which the researchers worked together to group the post-its by the product feature or interaction they referred to, such as templates, AI features, integrations or the job creation flow as a whole. Grouping by feature rather than by sentiment was a deliberate choice, as it allowed positive and negative responses about the same feature to be examined side by side, surfacing cases where the same functionality produced very different experiences for different users.

Once the feature-level themes were established, the findings were filtered against the redesign context of this thesis (see Section 2.1.2). Because the survey covered the job posting experience broadly, not all themes were equally relevant. The thematic

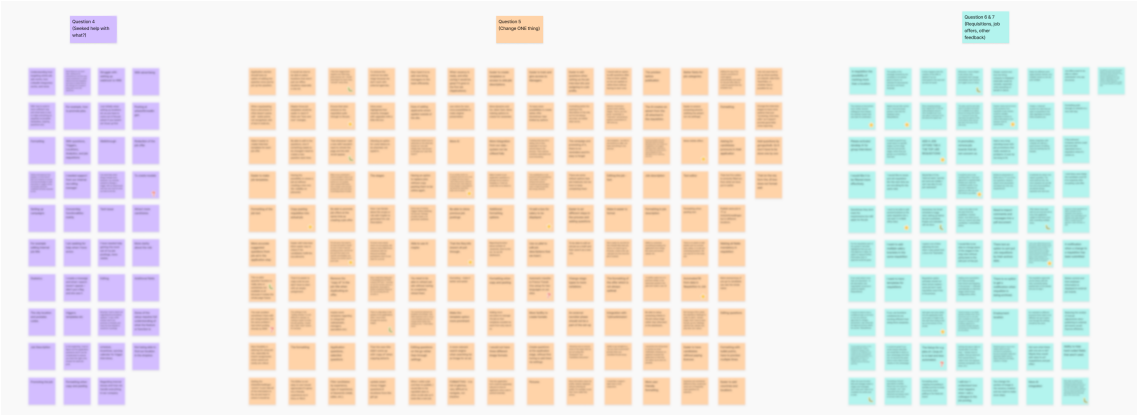


Figure 5.2: Qualitative responses transferred onto digital post-its prior to thematic analysis. Purple = Question 4, Orange = Question 5, Turquoise = Question 6 and 7.

Figure 5.3: Qualitative responses after thematic analysis by product feature.

grouping served as an analytical tool for making sense of the data as a whole by surfacing patterns, contradictions and the range of user experiences within each feature area. From this thematic understanding, individual findings were selected for inclusion in the Results chapter based on their relevance to the redesign: whether they identified a friction point, an unmet need, or a user behavior that the redesign could address. No frequency threshold was applied so a response raised by a single participant could constitute a finding if it pointed to something meaningful for the design.

Working collaboratively throughout the grouping and theming process served two purposes. First, it allowed disagreements about where a response belonged or what it indicated to be resolved through discussion rather than by individual judgment. Second, it ensured that both researchers developed a shared understanding of the dataset, which was important for both a deeper understanding of the product and the design work that followed in the thesis.

5.3.2 Figma Prototype Usability Testing

Following the current-design research, usability testing was conducted on a high-fidelity Figma prototype of the redesigned job posting flow. Sessions were conducted remotely through video call, following the same format as the contextual interviews. Participants were asked to share their screen and to interact with the Figma prototype directly. The use of a high-fidelity prototype rather than a live product is standard practice when evaluating a redesign that has not yet been built. 7 participants took part in the usability testing sessions and their profiles are summarized in Table 5.1.

Participant	**Role**	**Company size**	**Usage profile**
Pt1	Recruiter / Admin	Enterprise	Multi-country, heavy template use
Pt2	Implementation Lead	Enterprise (1000+ stores)	Template-governed, centralized setup
Pt3	Recruiter	SMB (70 employees)	Low volume, manual personalized approach
Pt4	Recruiter	SMB (12–13 employees)	Heavy template use
Pt5	Recruiter / Advanced user	Enterprise	Process-heavy, integration-aware
Pt6	Recruiter	SMB	Frequent Teamtailor user, copy-job driven workflow
Pt7	Talent Acquisition Manager	SMB (350 employees)	Prefer candidate sourcing over application

Table 5.1: Usability study participants by role, organization size, and usage profile.

Participants were asked to complete a set of predefined tasks designed to reflect the core scenarios within the redesigned flow: creating a job posting, configuring automation triggers, and locating the interview kit. Participants were encouraged to think aloud throughout each task, verbalizing their intentions and reactions as they navigated the prototype. The think-aloud method is particularly valuable in prototype evaluation because it makes visible the gap between the designer's intended interaction model and the user's actual mental model, which observation alone cannot reveal [45].

The researchers observed each session, taking structured notes on task completion status, points of hesitation or confusion and explicit comments. Where participants gave verbal consent, sessions were recorded for subsequent review. Following each

usability session, observations and participant quotes were compiled into a shared findings document and organized thematically by the area of the interface they concerned such as navigation, template handling or the job creation flow. Both researchers reviewed the compiled material and identified findings relevant to the ongoing redesign. As with the survey analysis, no frequency threshold was applied so a finding raised by a single participant was considered relevant if it pointed to a friction point, unmet need or design decision requiring reconsideration. Table 5.2 provides a selection of findings illustrating how observations were interpreted.

Finding	**Representative quote**
No visible publish button	"I would just want the publish job button. I don't see the purpose of unlisted or archive at this point." — P1
Inverted CTA hierarchy on Create Job modal	"I have no idea which one of these I would choose. So I assume this [Create Job button] would publish the ad, but this one [Continue to job ad button] would be to further change things within the job ad." — P2
No next/continue button; sidebar navigation not discoverable	"There's no clear way to move on to the next section. I'd always look for a big pink button." — P6
Template selection should come first in the flow	"For our store managers, we don't want them to have to fill out all of that. Template should be first, then it populates the fields." — P2
Preview should be final check including application form	"I guess it makes sense that it has to do with appearance, but personally I would preview it as the very last thing." — P2

Table 5.2: Examples of usability testing findings from the redesign evaluation with representative quotes.

Because the usability sessions were conducted exclusively on the redesigned prototype, all findings were directly relevant to the redesign context and are included in full in Appendix B.1..

5.4 Track 2: LLM-Assisted Research

The second research track was designed to investigate whether an LLM system could surface meaningful UX insights from naturally occurring organizational data and to produce findings that could be evaluated against the manual research described in Section 5.3. The system was developed iteratively through the phases described in the following sections, moving from data source evaluation through prompt refinement to a final architecture. The resulting system is presented in Section 6.1.

The structure of the final LLM system was developed through the Research Through Design approach introduced in Section 4.1.1. Under this framing, the system itself is the primary design artifact of the study and knowledge is produced through the process of making and revising it rather than through the application of a predetermined method. Each iteration surfaced new understanding about what the system could and could not do. It informed which data sources yielded actionable feedback and how context framing affected the model's output. These observations are contributions to the research questions the study is designed to answer.

With the manual research establishing the analytical frame, development of the multi-agent LLM system began with a phase focused on understanding what available data sources to use.

5.4.1 The API Verification Layer

Two approaches were tested for incorporating the Teamtailor API into the multi-agent LLM system. The first treated the API as an independent discovery layer, with an agent traversing the product to surface potential issues directly. In practice the outputs reflected the conditions of the data rather than genuine usability problems. The agent ran on a demo instance of the product, where standard fields are frequently left empty out of convenience. The agent flagged, for example, that none of the published jobs had an employment type set. This read as a potential usability problem but was an artifact of how the demo instance had been configured. Running the agent against a live customer instance could have avoided this particular problem, but introduced data access risks that fell outside the scope of what was appropriate for this study.

The second approach used the API as a verification layer. After the Slack and Intercom analysis had surfaced recurring friction points, an agent was used to check whether those problems were reproducible in the product. This also proved limited. The API exposed data about the state of objects within the product but could not capture the interaction layer where friction actually occurred. The verification layer was removed from the system on this basis.

5.4.2 Slack and Intercom as Data Sources

In contrast to the API, the Slack feedback channel and Intercom support conversations proved to be good sources. Both were scoped to key words relevant to the job creation and setup flow to match the constraints of the manual research. The challenge these sources posed was not availability but volume and noise. Filtering for relevant problems required iterative refinement of the scoping criteria applied during retrieval. Multiple Slack channels were tested to assess which contained feedback relevant to the job creation flow. Also different time constraints were applied to both sources to understand the cut off for relevant material.

5.4.3 Direct Prompting to Persona-Based Agents

As we realized there were no naturally occurring data on the new design the design of Track 2 was split into two distinct technical execution pipelines. Pipeline A, which handled the multi-source analysis of naturally occurring data for the current design , and Pipeline B, which executed the persona-driven synthetic user walkthroughs of the redesign prototype.

The initial approach to Pipeline B (new design evaluation) used direct contextual prompting, asking the model to evaluate the Figma prototype and surface usability issues without a defined user perspective beyond a general description of the product context. This made the outputs lean toward general statements about the interface rather than specific friction grounded in user behavior. A representative output from an early run was the observation that it is hard to know what department to choose when creating a job posting. This was not supported by the manual research, which found that frequent Teamtailor users generally understand their organization's department structure well and do not experience department selection as a meaningful source of confusion. The personas constructed in Section 5.2 were therefore provided to the model as system-level instructions rather than generated by the model itself. This distinction is important as the personas were not synthetic users asked to produce simulated responses from scratch, but empirically grounded descriptions used to orient the model's analytical perspective toward a specific segment. The shift produced a measurable change in output specificity. Where direct prompting had surfaced observations applicable to any user of any product, persona-based agents surfaced observations tied to the particular expectations and failure modes of each segment.

5.4.4 Prompt Refinement and Context Sensitivity

With the persona-based approach established, a further phase of iteration focused on how the model's broader context window was structured. This phase revealed that the type of background context provided to the model had a substantial effect on the attitude it adopted. Providing context that included the strategic rationale for the redesign, like explaining the design goals and the reasoning behind specific decisions, produced outputs that reframed documented usability problems as intentional design choices. Removing this justificatory framing in favor of neutral domain context describing the product without explaining or defending specific decisions, restored a more analytical approach. The implications of this finding for how prompts should be understood as research instruments are discussed in Section 7.5.

The iterations described in this chapter resulted in the pipeline architecture presented in Section 6.1. The API verification layer was dropped, Slack and Intercom served as the data sources for the current design analysis, persona-based agents grounded in the manual research drove the redesign evaluation and prompts were structured with neutral domain context rather than justificatory framing. Each of these decisions emerged from a specific observation during iteration rather than from a prior specification, which is consistent with the Research Through Design framing adopted in this study.

5.5 Cross-Track Analysis

This section describes the procedure through which findings from the two tracks were brought together, matched, classified and coded. It is intended to support the results presented in Chapter 6.

5.5.1 Timing and Independence

The evaluation of the findings was conducted after both tracks had completed their analyses. The LLM-based multi-agent system from Track 2 produced its outputs through a single finalized run. The manual research findings from Track 1 were synthesized from the survey and usability testing sessions prior to contrasting.

However, the construction of the LLM system was shaped by the manual research in methodological ways. The personas used to instantiate the redesign evaluation agents were built from Track 1 interview data, the scoping of data sources was informed by the domain knowledge the manual research produced and the prompt design drew on understanding of the product and its user segments that emerged through Track 1. So while the different tracks were not informed of each other specific outputs, there was methodological influence throughout the thesis.

5.5.2 Assembling the Finding Sets

The selection of findings for inclusion in the thesis was guided by relevance to the job creation and setup flow and this boundary was applied consistently across all methods.

The inputs to the matching of the findings were the structured Markdown produced by the finalized LLM system and the set of synthesized findings extracted from the Track 1 process documented in Section 5.3. Prior to matching, both sets were reviewed in full to ensure that only findings relevant to the job creation and setup flow were included, consistent with the scoping constraint applied to both tracks throughout the study. This produced a finding set of 29 items from Track 2 and 34 items from Track 1, which are used in Chapter 6. All findings are listed in Appendix B.1

5.5.3 Matching Procedure

One researcher conducted the primary matching, working through each finding from both tracks systematically. Findings were matched on the basis of *thematic similarity* rather than exact matches. Two findings were considered to address the same issue if they pointed to the same underlying usability problem, even when the framing or level of specificity differed between tracks.

This approach reflects the nature of the two data sources. The LLM system produced findings from Intercom support tickets, Slack messages and Figma node inspection while the manual research produced findings from interview transcripts, survey responses and observed usability test behavior. The same problem therefore

rarely appeared in identical language across the two tracks. Thematic matching was necessary to compare findings that were substantively equivalent but expressed differently.

The second researcher reviewed the complete classification in full. Where disagreements arose regarding whether two findings should be matched or kept separate, these were discussed jointly and resolved by consensus, drawing on shared knowledge of the product and the data from which each finding originated.

5.5.4 Classification Criteria

Once matching was complete, each finding or matched pair was assigned to one of six categories. The decision rule for each category is as follows.

Convergent findings are those for which both tracks independently identified the same underlying usability problem.

LLM-only findings are those present in the multi-agent LLM system output for which no thematically equivalent finding existed in the manual research after full review of both sets.

Human-only findings are those present in the manual research synthesis for which no thematically equivalent finding existed in the multi-agent LLM system output.

LLM misreadings are findings produced by the multi-agent LLM system that were judged, upon review, to be incorrect or based on a false premise. The judgment drew on domain knowledge of the product and, where relevant, on input from the Teamtailor product team regarding intended behavior or system architecture. The reasoning behind each misreading classification is documented individually in Section 6.2.4.

Human misreadings are findings from the manual research that were judged, upon review, to rest on an incorrect assumption about the product's capabilities.

Contradictions are findings where both tracks addressed the same issue but reached opposing conclusions about it.

5.5.5 Development of Problem Nature Categories

To examine what kinds of problems each approach surfaced, findings were additionally coded according to a set of problem nature categories. These categories were not defined in advance. After the classification procedure was complete, both researchers reviewed the full set of 51 unique findings jointly and identified recurring types of problem and observation through thematic analysis. Categories were named to reflect the patterns present in the data rather than using set categories from the start.

This inductive process produced 7 categories: UI and interaction design, navigation and information architecture, silent system behavior, workflow and operational constraints, missing functionality, domain and contextual knowledge, and positive

feedback. Each finding was then assigned to the category that best described the nature of the issue it documented. Where a finding could plausibly belong to more than one category, it was assigned to the category that most closely described the primary problem rather than a consequence.

5.5.6 Limitations of the Analysis

Because the primary matching was conducted by one researcher and reviewed by the other, rather than coded independently by both before discussion, formal inter-rater reliability was not calculated. This is consistent with the interpretive, qualitative nature of the study and the Research Through Design framing adopted throughout but it means that the classification judgments presented in Chapter 6 rest on reasoned researcher consensus rather than a quantified agreement measure. Readers should interpret the category boundaries accordingly. The distinctions between, for example, a convergent finding and an LLM-only finding reflect the researchers' thematic judgment rather than a quantifiable truth.

6

Results

The findings presented in this chapter were produced by a single run of the finalized multi-agent system described in Section 6.1, conducted after the iterative development process documented in Chapter 5 was complete. The outputs of that development process, including discarded approaches and early system versions, do not form part of the findings presented here. This chapter presents all findings generated by that single run alongside the key findings from the manual research (Track 1), limited to those relating directly to the job setup flow.

6.1 Final Structure of multi-agent LLM system

Analysis was implemented as a Python orchestration system delegating tasks to Claude through the Claude Code CLI (session-based authentication). Slack, Intercom and Figma Desktop MCP connections enabled direct data source querying. All agents executed asynchronously (Python asyncio), with parallel execution within the system. Turn-by-turn execution traces (prompts, responses, token usage) were persisted to timestamped JSON files. Outputs were exported as JSON (for programmatic analysis) and Markdown (for human review). Analysis was iteratively refined through prompt optimization based on output coherence. The final system can be found in Figure 6.1.

6.1.1 Current Design Analysis

Automated content analysis was applied to naturally occurring data using LLM agents. Slack messages from customer success (spanning six months), Intercom support conversations (filtered by job-creation-related keywords). Agents were instructed to surface patterns inductively without predefined category lists. Output schemas were structured (critical_issues, positive_signals, findings, top_themes) but the patterns within them emerged from the data. Severity was assessed based on frequency, downstream workflow impact, and cross-source corroboration. Each source was analyzed independently, allowing corroboration to emerge during synthesis rather than being assumed at collection. The Intercom agent processed keywords in parallel batches to reduce latency, then merged results with quote-based deduplication. The Slack agent queried recent conversations directly via MCP. Pipeline A (Current Design) processed these two sources in parallel and fed findings to synthesis A.

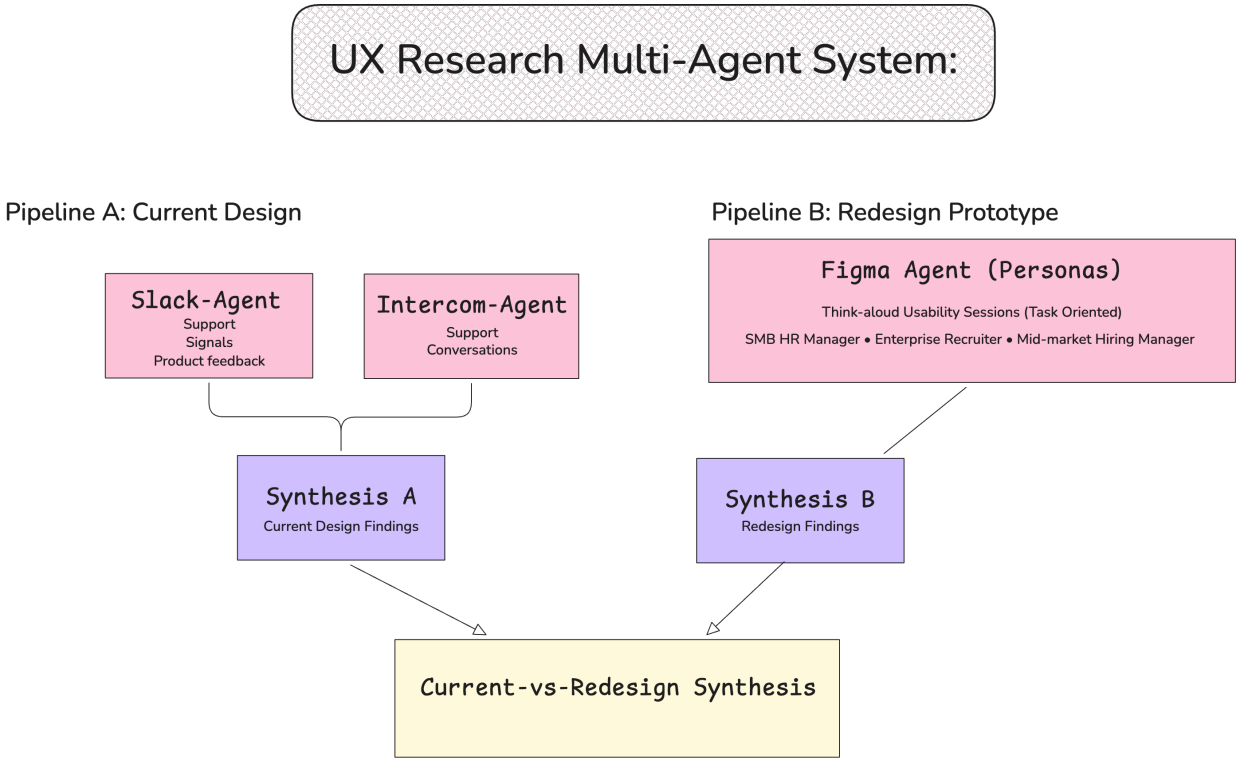


Figure 6.1: Final LLM pipeline overview.

6.1.2 Redesign Evaluation

The 3 personas defined in manual research (SMB HR Manager, Enterprise Recruiter, Mid-market Hiring Manager) were instantiated as LLM agents and navigated the Figma redesign prototype through 4 predefined task scenarios: (1) Create a job ad, (2) Add a send message trigger, (3) Add a reject message trigger, (4) Find the interview kit. Giving the same starting point as the real users during manual user testing.

Evaluation occurred in two stages. Stage 1: Haiku-model agents conducted think-aloud narration using Figma Desktop MCP to inspect design nodes, producing structured JSON output for each task including task completion status, difficulty rating, screens visited, think-aloud narration (tagged with observation types: expectation_met, confusion, friction, positive_surprise, prototype_dead_end), friction_points, positive_moments, key_quotes, and prototype_gaps. Stage 2: Sonnet-model synthesis agent analyzed task results to produce persona-level summary including overall_impression, comparison to current design, critical_issues, positive_signals, and source_summary.

Pipeline B (Redesign) executed these persona sessions in parallel and fed findings to synthesis.

6.1.3 Synthesis and Contrast

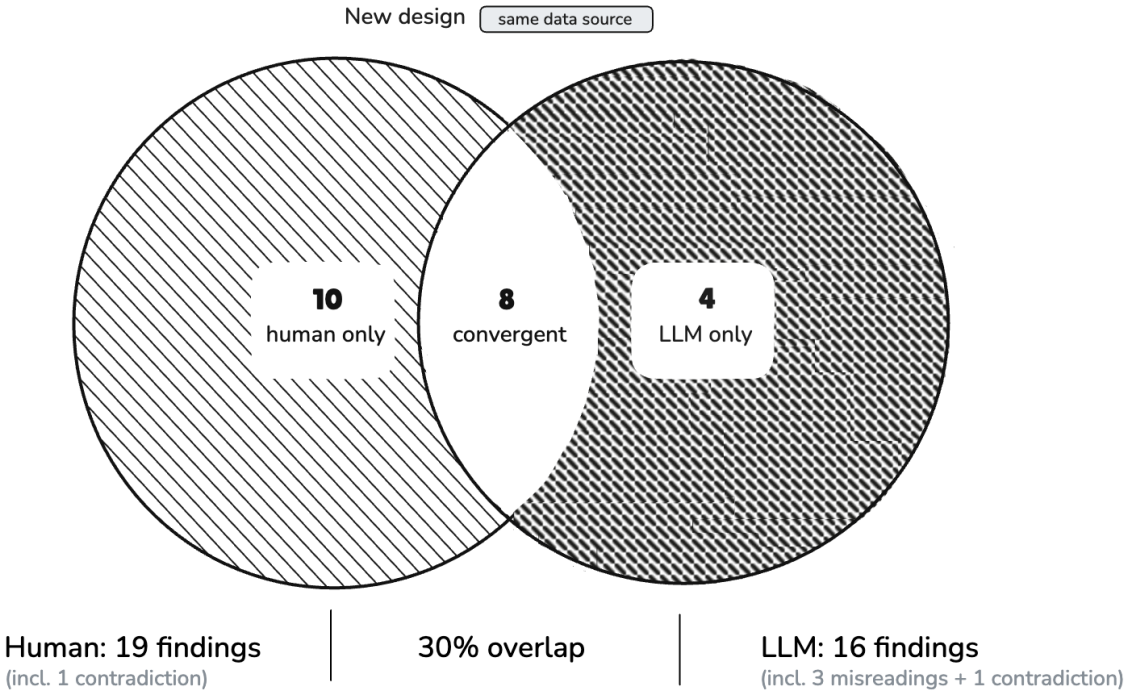
Comparative synthesis took current design findings and redesign findings as input and produced a verdict matrix classifying each current design problem against redesign outcomes: fixed (redesign solves the problem), partially fixed (architectural change addresses root cause but execution remains incomplete), worse (redesign introduces new regression), or unknown (not tested in prototype). This comparison functioned as a structured framework for identifying coverage gaps and divergences between the two design generations.

6.2 Findings overview

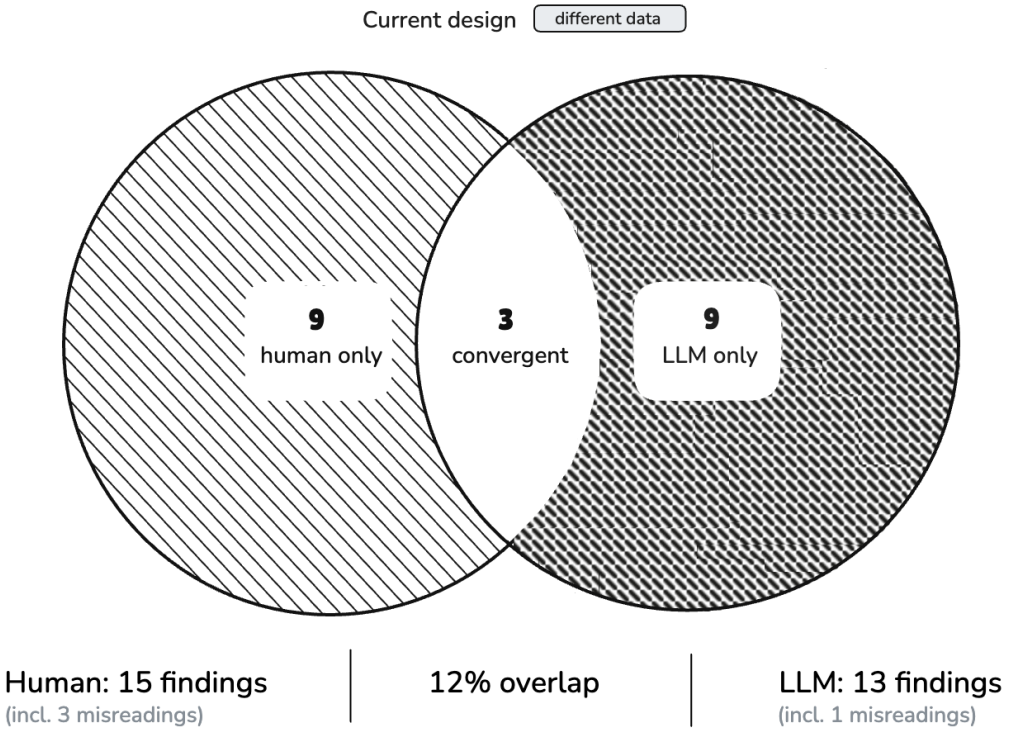
The findings presented in this section concern two design contexts: the current Teamtailor design and the proposed new design introduced in Section 2.1.2. Both contexts were evaluated independently by the LLM system and the manual research before findings were contrasted. In this section, findings from the manual research (Track 1) will be referred to human findings.

The findings were produced by two independent sources. The manual research (Track 1) which included one survey with 437 respondents and 7 usability tests, described in Section 5.3 produced 34 findings. The LLM system (Track 2) produced 29 findings using the finalized architecture described in Section 6.1. In Figure 6.2 the contrast between the two tracks are visualized for the different design contexts.

On the current design, where the LLM system worked from naturally occurring textual data and the manual research from interviews and a survey, the two tracks surfaced similar amounts of unique findings but only overlapped on 3 as seen in Figure 6.2. On the redesign, where the LLM system took the form of synthetic personas walking through the prototype, the pattern is different: the manual research surfaced more unique findings and the LLM system produced a larger share of incorrect or misclassified observations. However, the overlap is more than double from the current design findings.



(a) Results regarding the new design



(b) Results regarding the current design

Figure 6.2: Visualization of the findings the two tracks surfaced, divided by design context.

6.2.1 Convergent findings

The findings in Table 6.1 were identified independently by both the LLM system (Track 2) and the manual research (Track 1), suggesting they represent issues legible to both analytical approaches.

Table 6.1: Convergent findings: identified independently by both LLM-assisted and manual research.

Finding	**Design**
LinkedIn location maps wrong silently	Current
Hiring team access can't be set during creation	Current
Manual entry pain point: users want requisition data to flow into job post	Current
Inverted CTA hierarchy on Create Job modal	New
No visible Publish button	New
No Next/Continue button, sidebar navigation not discoverable	New
Send message trigger: 0% completion	New
Location field on two screens with no signal that data carried over	New
Desktop + mobile preview on Appearance step	New
Autosave "Last saved: Just now" indicator	New
Data carry-over between steps works seamlessly	New

On the current design, 3 findings appeared independently in both analyzes. An example of these concerns the location accuracy for the LinkedIn integration seen at the very top of Table 6.1. When a job is published in Teamtailor, the system can automatically post it to LinkedIn. Users expect the location displayed on LinkedIn to match precisely what they entered during job creation. However, the system sometimes silently maps the wrong location onto the LinkedIn post, with no in-product indication that this has occurred. For recruiters relying on LinkedIn to attract local candidates, this means jobs are being advertised in the wrong location without their knowledge. The LLM system surfaced this from a recurring pattern in Intercom tickets in which customers reported that posted jobs appeared on LinkedIn with incorrect locations. The manual research arrived at the same finding through an interview in which an enterprise customer described checking LinkedIn manually after each post as a workaround.

On the new design, 8 findings appeared in both the LLM system persona walkthroughs and the manual prototype testing, and 3 of these were noted by a majority of human participants. The most consequential is the inverted call-to-action hierarchy on the Create Job modal (see Figure 2.3b), where *Create job* is styled as the primary action while every LLM persona wanted to use *Continue to job ad*. 6 out of 7 human participants in the usability testing also expressed doubt or uncertainty surrounding these two buttons.

The absence of a visible Publish button similarly registered in both analyzes: every LLM persona completed the full creation flow without locating a way to publish and 4 of 7 human participants were also confused on how to publish the ad. These two findings were one of the most critical ones during the manual research and also led to actual iterations of the new design at Teamtailor.

The higher rate of convergence between the LLM system and human findings for the new design contrasted to the current is likely attributable to the difference in data type. For the current design, the LLM system analyzed Slack feedback and Intercom support conversations which is a large and varied corpus with differing formats and topics. For the new design, the LLM system instead analyzed the Figma prototype, which is fundamentally similar in nature to what human participants were exposed to during interviews. This alignment in data source likely explains why the LLM system and human findings converged more frequently.

6.2.2 LLM only findings

The findings in Table 6.2 were surfaced exclusively by LLM-assisted research and did not appear in the manual research.

Table 6.2: Findings surfaced only by LLM-assisted analysis (Track 2).

Finding	**Design**
Location silently defaults to HQ if left empty	Current
No autosave	Current
Requisition approval creates problems for long drawn processes	Current
Protected template restrictions hidden until after creation	Current
Generic save errors don't identify which field failed	Current
Internal job name can't be set during creation, requires a return visit	Current

Continued on next page

Continued from previous page

Finding	**Design**
Recruiter can't be reassigned after delegated creation	Current
No post-creation guidance, users stall at the handoff moment	Current
Dark mode renders job description editor header text black, invisible	Current
"Symbol" field on internal title: no explanation, two personas ignored it	New
"Role" dropdown ambiguous vs "Internal title"	New
Write with Copilot feature	New
Salary range toggle	New

The LLM system surfaced 9 findings on the current design that did not appear in the manual research. The LLM findings on the current design are, with few exceptions, concrete and smaller friction points rather than broader UX problems. Issues of this kind rarely surface in interviews, likely because participants tend to focus on the overall flow rather than individual system behaviors. Yet, they are consistently present in Intercom conversations and internal Slack feedback, suggesting that they accumulate into meaningful pain points over time. Individually they may appear minor, but taken together they can reveal blind spots that interview-based research systematically misses.

The LLM-only findings in the new design were fewer in number and more variable in significance. 2 of 3 were positive observations that do not add analytical weight beyond confirming that these features are visible and legible to a first-time user. The third, the unexplained *Symbol* field on the internal title screen, represents the LLM noticing a UI element that human participants passed over without comment. Taken together, the LLM-only findings on the new design offer less depth than their counterparts on the current design. This outcome is to be expected, as the current design benefited from more extensive data.

6.2.3 Human only findings

The findings in Table 6.3 emerged exclusively from the manual research and were not surfaced by the LLM system.

Table 6.3: Findings surfaced only by manual research (Track 1).

Finding	**Design**
AI features not used much in practice	Current
Heavy reliance on template for job creation	Current
Importance of custom fields (organizational context)	Current
Calendar / end-date handling major pain point, jobs stay live too long	Current
Cannot make custom fields mandatory	Current
Text editor friction	Current
No ability to edit a question once added; no bulk-add	Current
Stage management feels buggy and too high-level	Current
Strong demand for better AI-assisted features across setup	Current
General thoughts on job / job-ad split in the new design	New
Domain context: laws for different countries affect setup decisions	New
Job description field missing in requisition	New
Template selection should come first in the flow	New
Preview should be final check including application form, not just appearance step	New
Confusion around timeframe field	New
Accessibility: trigger color coding insufficient for color-blind users	New
Weekend exclusion from triggers	New
Appreciated the research and testing approach before launch	New
Consolidated view of all triggers was appreciated	New

On the current design, the manual research surfaced 9 findings that the LLM system did not. Many of them represent a behavioral or contextual observation about product usage, often relying on knowledge of regulatory or organizational structures

supplied by participants during interviews. Naturally occurring data, such as support conversations, tends to capture system exceptions and specific malfunctions rather than baseline usage patterns. For example, users report broken templates or malfunctioning custom fields, but rarely articulate their underlying reliance on these elements. Identifying these foundational workflows therefore required direct engagement with users.

On the new design, 10 findings appeared in the manual research that the LLM personas did not surface. The pattern across the new design human-only findings is broader than the pattern on the current design. Where the current design human-only findings were almost uniformly strategic and behavioral, the redesign findings span strategic reasoning, domain knowledge, accessibility, and concrete defects. While the LLM persona walkthroughs covered the most prominent issues in the prototype it missed several categories of insight that real users raised through a combination of personal experience, domain expertise, and direct interaction with the design.

6.2.4 LLM misreadings

The findings in Table 6.4 are findings produced by the LLM system that were, upon review, incorrect or misclassified.

Table 6.4: Findings classified as LLM misreadings.

Finding	**Design**
Publish state invisible, career site sync broken (bug, not UX)	Current
Interview kit: 0% completion, no escape hatch (prototype problem)	New
Reject message activation state opaque (prototype problem)	New
No escape from modal to pipeline (prototype problem, raised valid workflow question)	New

4 findings produced by the LLM system were, upon review, incorrect or misclassified. Rather than reflecting a general failure of reasoning, these misreadings were driven by the specific constraints of the two data sources and testing environments: analyzing support tickets for a live product versus analyzing user interactions with a limited-fidelity prototype.

For the current design, the LLM system analyzed customer support conversations from Intercom, which led it to misinterpret technical and contextual limitations as user experience flaws. First, based on user complaints, the LLM system characterized a broken career site synchronization as a UX issue, failing to recognize that the

root cause was a backend system bug. This error highlights a limitation of analyzing naturally occurring support data: The LLM cannot reliably distinguish between the symptoms users describe in Intercom and the underlying system causes that produce them, nor can it infer missing contextual knowledge from the text alone.

The 3 misreadings associated with the new design stemmed entirely from the constraints of the prototype testing environment. The LLM flagged 3 specific issues: the lack of an "escape hatch" in the Interview kit modal, an opaque activation state for the Reject message and the absence of navigation from a modal back to the pipeline or job overview. In each instance, the LLM system's logic was internally consistent but rested on a literal misinterpretation of the prototype's fidelity. Because the Figma prototype lacked click targets for these specific interactions, the LLM system concluded that the underlying functionality was missing from the proposed design. Human participants encountered the exact same constraints, but drew upon their understanding of how prototypes differ from production systems to mentally fill the gaps and discount the missing interactions.

6.2.5 Human misreading

The findings in Table 6.5 are findings from the manual research that were, upon review, based on incorrect assumptions about the product's capabilities.

Table 6.5: Findings classified as human misreadings.

Finding	**Design**
Wish to make fields mandatory in Requisitions	Current
There is no option to add details such as contract type, salary, or pre-screening questions	Current
Would be nice to remove unused steps such as *Evaluation* and *Candidate Interaction*	Current

3 findings from the manual research were, on review, based on incorrect assumptions about the product's capabilities. In each case, a survey respondent reported the absence of functionality that in fact exists within the current design. The findings concern the ability to make requisition documentation fields mandatory, the availability of structured fields for contract type, salary, and pre-screening questions, and the option to remove or hide pipeline steps such as *Evaluation* and *Candidate Interaction* that are not in active use.

All 3 misreadings originate from the survey rather than the contextual interviews or usability testing. Likely this is because the survey was unmoderated, leaving respondents no opportunity to test their assumptions against the actual interface. In the interviews and usability testing sessions, participants raised similar misunderstandings on occasion but these were identified and corrected in real time through researcher clarification and did not make it into the final findings.

6.2.6 Contradictory findings

The findings in Table 6.6 represent direct disagreements between the two tracks on the same issue.

Table 6.6: Findings classified as contradictions.

Finding	Design
Sidebar progress checkmarks: LLM personas praised, human users confused by them	New

Across all the findings contrasted, there was one direct contradiction between the LLM system and the manual research. The new design sidebar progress indicator, which displays a checkmark next to each completed step in the wizard, was praised by all 3 personas in the LLM walkthroughs as providing reliable position awareness in a multi-step flow. 2 of the 7 human participants in the usability testing found the same checkmarks confusing. They did not match either of the participants' mental models of what completion meant, and one participant could not determine whether a checkmark indicated that a step had been visited, that all required fields on it had been filled, or that the step had been validated.

6.2.7 Distribution by Problem Nature

To examine what kinds of problems each approach surfaced, the 51 unique findings were coded according to 7 "problem nature" categories: UI and interaction design, navigation and information architecture, silent system behavior, workflow and operational constraints, missing functionality, domain and contextual knowledge, and positive feedback. The resulting distribution is presented in Figure 6.3 and details on nature, design and source of all findings can be found in Appendix B.1.

UI and interaction design was the largest category (n=12) with almost equal parts convergent, LLM-only and human-only findings, suggesting that surface-level interface problems were legible to both the LLM system and human participants. Silent system behavior and workflow and operational constraints, by contrast, skewed toward LLM-only findings, indicating that the LLM system was more sensitive to issues that manifest through system logic rather than visible interface elements. Missing functionality and domain and contextual knowledge showed the opposite pattern, with human-only findings dominating both categories.

The positive feedback category (n=7) is distributed across convergent, LLM-only, and human-only sources, indicating that both approaches were capable of registering design strengths alongside problems.

Separating the findings by design context reveals a more nuanced picture than the combined distribution. As shown in Figure 6.4, the two design contexts produced different profiles of findings both in total volume and in the nature of issues identified.

In the current design, missing functionality constituted the largest category (n=7), followed by workflow and operational constraints, UI and interaction, and domain

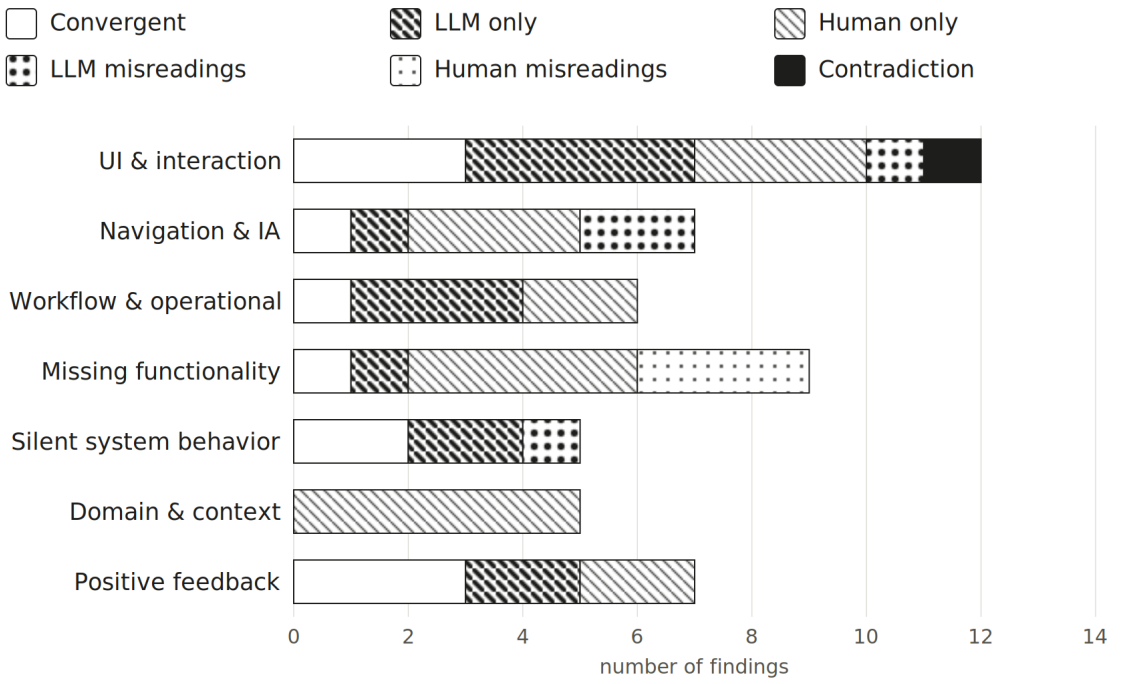
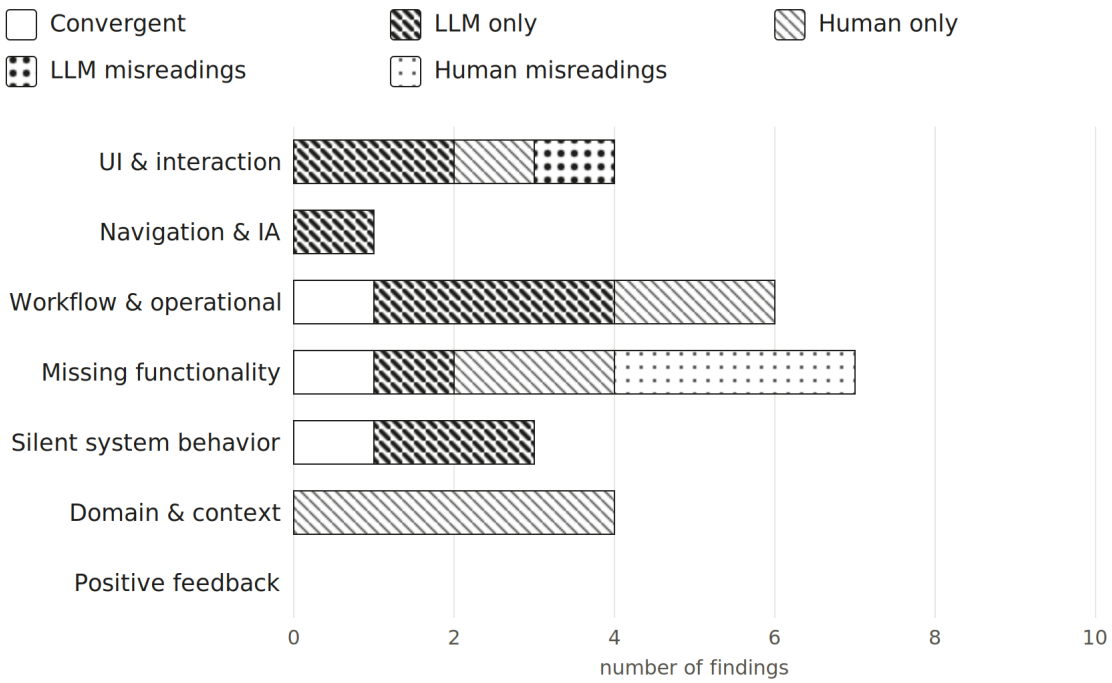


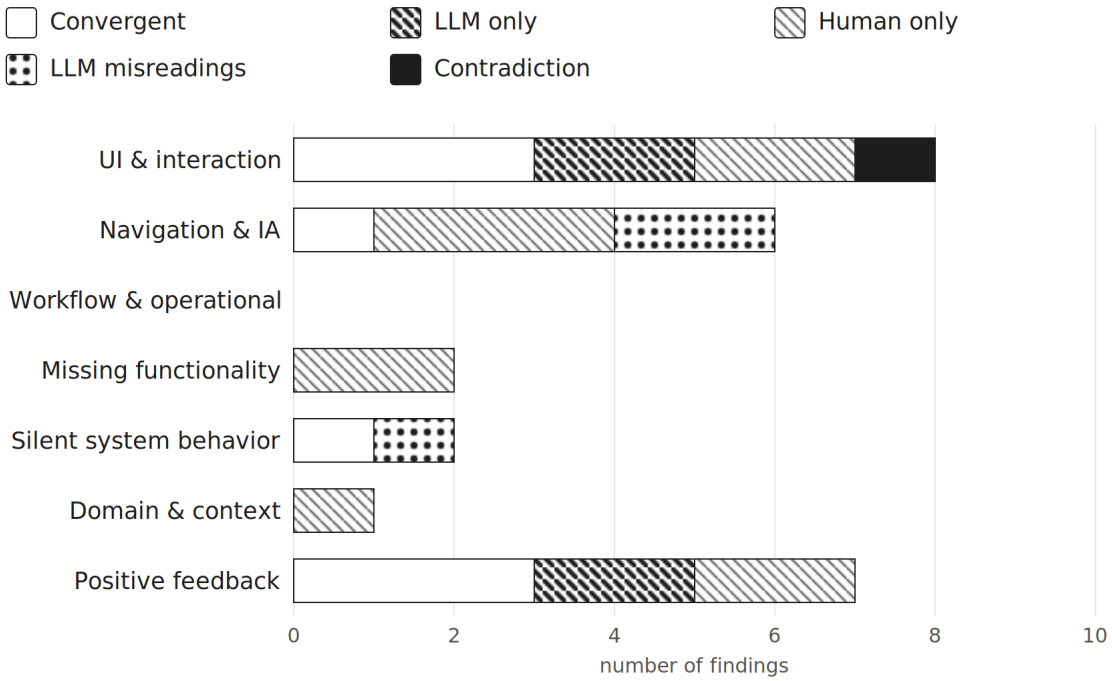
Figure 6.3: Distribution of the 51 unique findings by problem nature, broken down by source.

and contextual knowledge. This pattern suggests that the current design's primary problems are structural; the system has process constraints that obstruct recruiters' work, lacks features that users expect, and requires organizational knowledge to navigate effectively. The LLM system was the dominant source of workflow findings, surfacing 3 of the 6 on its own, while domain and contextual findings were human-only.

The new design produced a different profile. Workflow issues dropped to zero, and positive feedback became one of the two largest categories (n=7). However, UI and interaction problems increased (n=8) and navigation and information architecture emerged as a significant category (n=6) where it had been almost absent in the current design. This shift suggests that the new design introduced new usability friction.



(a) Current design (25 findings)



(b) New design (26 findings)

Figure 6.4: Distribution of findings by problem nature, broken down by source for the different designs.

7

Discussion

The results in Chapter 6 document what the two research tracks surfaced about the current and redesigned job posting flow at Teamtailor. This chapter takes the next step and asks what the comparison between those tracks reveals about LLM-assisted research on naturally occurring data as a UX research approach. Six insights (Sections 7.1-7.6) are drawn from the data and together they form the answer this thesis offers to its research question.

7.1 Convergence on Critical Issues

The two tracks produced 63 findings in total (29 from the LLM system and 34 from the manual research). Because the 11 convergent and 1 contradictory findings were each identified by both tracks, these correspond to 51 unique findings. One of the most interesting observations is that the LLM and the human researchers independently identified the same issues at the top of the severity list. Of the eleven convergent findings reported in Table 6.1, the 3 that human participants noticed most frequently were also surfaced by the LLM personas as critical. All 3 of these findings drove design discussions and changes in the final design. These included: the inverted CTA hierarchy on the Create Job modal (raised by 6/7 human participants and all 3 LLM personas), the absence of a visible Publish button (4/7 humans and all 3 LLM personas), and the undiscoverable sidebar navigation (4/7 humans and all 3 LLM personas). The fact that two methods working independently reached the same conclusions on these critical items is the clearest evidence in this study that the LLM-assisted research produces meaningful output rather than only superficial noise.

The LLM-assisted evaluation of the redesign was conducted before the manual usability testing data was synthesized, and the redesign's critical issues it surfaced were the same ones that human participants later struggled with most visibly. This positions LLM-assisted analysis as a prospective evaluation layer. For a product team operating under the time pressure that B2B SaaS environments often have, the ability to surface a credible short list of high-risk design assumptions in minutes, once the foundational research infrastructure and persona frameworks are established, demonstrates the growing potential of LLM-assisted research to effectively support the design cycle.

However, the convergence rate differs sharply between the two design contexts: 8 of

26 findings were convergent in the redesign, but only 3 of 25 in the current design. The most likely explanation is the difference in input data. For the redesign, both methods were working from the same artifact, the Figma prototype, and producing comparable kinds of evidence. For the current design, the LLM processed approximately 6,000 Slack messages and 400 Intercom support conversations, while the manual research relied on a survey. Variations in input naturally lead to distinct findings. Therefore, a high rate of convergence also relies on the alignment of the data they process. Also, the convergence overall was weaker than on critical issues. Across all 51 findings the methods agreed on just under a quarter of items.

7.2 Scale and Nuance as Complementary Lenses

The distribution by nature of issue (Figures 6.3 and 6.4) shows that the methods surface different categories of problems. AI-only findings cluster in two categories. The first is silent system behavior, where the problem is invisible at the level of the interface but discoverable in the accumulated user reports: location silently defaulting to HQ when left empty, generic save errors that fail to identify the offending field, dark mode rendering the editor header invisible. The second is workflow and operational constraints that emerge only across many cases: protected template restrictions hidden until after creation, recruiter reassignment blocked after delegated creation, no post-creation guidance at the hand-off moment. These problems share the property of being discoverable from a corpus of user-reported friction but rarely volunteered in a single interview.

Human-only findings mainly cluster in categories requiring domain and contextual knowledge or identifying missing functionality. As shown in Table 6.3, these insights include legal differences across countries, specific accessibility concerns and the systemic 'end-date' handling problem. While the consequences of these issues exist within naturally occurring data such as a support ticket noting a job was left live too long, the data itself is 'the wrong shape' to make the root cause legible. It records the symptom but not always the cause or the workaround a user built to bypass a design flaw. Bridging the gap from a data trace to a substantive design problem requires the contextual reasoning unique to human interviews. In other words, LLM-assisted analysis of naturally occurring data extends the researcher's *breadth-first* reach: it makes the field of view larger and turns a dispersed body of organizational text into a navigable map of recurring issues. Human research extends the *depth-first* reach: it follows the user beneath the interface layer into how they actually work. Neither method is a substitute for the other because they work best in different areas. What this study contributes empirically is evidence that the gap between LLM and human analysis is a result of the distinct analytical scopes each method is equipped to handle. LLM is naturally suited for detecting recurring system patterns across large data volumes, while human researchers are naturally suited for uncovering reasoning and organizational constraints.

7.3 AI Errors as Diagnostic Signals

4 findings produced by the LLM-system were classified as misreadings. 3 of these (specifically concerning the interview kit, reject message activation and modal navigation) stemmed from the limited fidelity of the Figma prototype. While the LLM's literal interpretation of these missing functionalities and navigation was technically an error, two of these 'failures' functioned as diagnostic signals that surfaced valid workflow concerns. The clearest example was the modal navigation finding. The LLM flagged the inability to navigate from a job ad modal back to the candidate pipeline as a missing feature. Although human participants did not report this, likely because they mentally 'filled in the gaps' of the prototype, the LLM's literal misreading prompted a design discussion about a genuine workflow gap. The discussion centered around if it was reasonable to ask users to navigate back and forth between the modal and the candidates pipeline when in the previous design everything was done in one flow. In this sense, the LLMs lack of human 'forgiveness' for prototype limitations served to expose structural information architecture questions that had not been articulated in the manual research.

When the LLM's framing of a problem was incorrect, the process of explaining why it was wrong forced the research team to articulate design assumptions that had previously been held implicit. This suggests that value lies in both convergent and divergent findings. While convergence confirms a known issue, divergence forces a 'defense of reasoning' where the underlying logic of the design must be justified against a critique. This means that an AI finding should also be judged on the transparency of the logic that produced a finding. A 'coherent misreading', one based on a logical analysis of the available evidence, is often more valuable than a correct finding reached by chance. This is because a logical error explicitly exposes design assumptions that human researchers can then isolate and test.

7.4 The Importance of Domain Expertise

A consistent pattern across the analysis is that the quality of the LLM's output was proportional to the quality of what was fed into it before any data was processed. The personas that drove the redesign evaluation (SMB HR Manager, Midmarket Hiring Manager, Enterprise Recruiter) were not generated by the model but constructed from the manual research and from conversations with the Teamtailor product team. The segment specific findings the LLM system produced were only possible because the segmentation had already been done by humans.

This has implications for how LLM-based UX tooling should be understood. As mentioned in Chapter 2, Lu et al. and Salminen et al. identify the tendency toward generic, 'average user' observations as a characteristic limitation of LLM-generated personas [25], [31]. The results here suggest that this limitation is not inevitable as it might reflect the absence of grounding rather than a fundamental limit of the technology. When the personas were anchored in real product knowledge and prior research the LLM system produced findings that were differentiated by user

segment. The implication is that LLMs' outputs may improve when they are treated as analytical tools operating on human defined inputs instead of generative tools expected to produce insight from scratch. The same dependency appears more concretely in the human-only findings (Table 6.3). Items such as the country by country legal variation in setup decisions and the accessibility limitation of the trigger color coding were not findings the system could have produced from the available data, because the relevant context was not present in the corpus. The LLM system did not see these issues because it was not configured to look for them while the human participants raised them when prompted by direct questioning. In other words, this indicates that what might look like a limit of LLM analysis may be a limit of what the researcher brought to the analytical setup.

The intuitive expectation is that AI lowers the expertise threshold for doing qualitative analysis. The model takes on more of the interpretive load and the researcher needs to know less. The data here suggest the opposite. An expert researcher, familiar with the product and its user segments, is likely to produce context specific, concrete findings. A less experienced practitioner working with the same tool may tend to receive output that retains the same surface fluency but focused on more generic observations. The technical and analytical threshold for conducting rigorous AI-assisted research may be equally demanding as, yet distinct from, traditional manual research. In addition to UX knowledge the researcher must also know enough about how the AI's outputs weaken under various conditions to interpret them correctly.

7.5 Context as a Research Instrument

As described in Section 5.4.4, including justificatory framing in the model's context shifted its analytical posture from evaluative to defensive.

This shows a useful distinction between *domain context* and *justificatory context*. Domain context describes the product, the users, the workflows and the constraints needed to interpret the material. For example that Teamtailor serves SMB, mid-market and enterprise customers with different approval workflows or that jobs must reach a published state to appear on the career site. Justificatory context explains why a design direction was chosen, what investment has been made in it or why it is considered important. The first kind of context helps the model reason within the correct domain while the second risks, as mentioned, shifting it from evaluation to rationalization. This suggests that prompts should be treated with the same care as interview guides or the design of surveys to avoid introducing systematic bias.

7.6 AI-Ready Data as Research Infrastructure

The LLM's 'analytical reach' depends largely on how well the data is organized. As mentioned, on the current design the LLM system read Slack channels and Intercom tickets and had no problem when the data at times were clear and descriptive. However, when data lacked tags for company size or issue type, the system struggled

to find meaningful patterns. This organizational gap may be another reason why the LLM and human findings matched more often for the redesign than for the current product. While the redesign evaluation used a single, clear prototype, the data for the current design was a messy mix of different channels and formats. This forced the LLM system to spend more effort sorting through 'noise' than uncovering useful UX insights.

However, the Figma prototype also suffered from not being optimized for LLM analysis. As mentioned earlier, 3 of the 4 misreadings produced on the redesign were not really analytical errors. The LLM's reasoning in each case was internally consistent but the prototype lacked the click targets that would have demonstrated the relevant interactions and the system concluded that the underlying functionality was missing from the design. This is the same readiness principle operating on a different artifact. Just as ambiguous Slack channel entries lower the LLM's ability to segment data, missing click targets in a prototype degrade its ability to evaluate a design. In both cases the LLM's strength of applying consistent logic across a corpus or an artifact becomes a liability when the input is incomplete. The model cannot distinguish missing signal from absent feature in the way a human researcher or participant can.

While inconsistency is expected in naturally occurring data, small improvements in data hygiene could improve the quality of LLM output. For B2B SaaS organizations, investments in data organization such as clear naming conventions and consistent metadata has historically served only operational purposes. However, this study demonstrates that such hygiene takes on new value when internal data is treated as 'research infrastructure'. By labeling support tickets consistently, structuring communication channels to distinguish between customer segments and making sure prototypes have sufficient fidelity, organizations create the conditions for LLMs to produce far more specific results. This is a structural precondition since no amount of prompt refinement compensates for a corpus where the relevant distinctions are not present. Practically, this means that the question of whether an organization can benefit from LLM-assisted research on naturally occurring data is partly a question that has to be answered before the research begins, in how the organization has chosen to structure its data over the preceding years.

7.7 Implications for UX in B2B SaaS

The results of this study have practical consequences for both individual UX researchers and the B2B SaaS organizations they work within. For UX researchers, the most immediate shift is in what counts as a researchable corpus. Slack channels and Intercom transcripts have historically been outside the UX researcher's domain because the cost of analyzing them at scale. This study demonstrates that LLMs make that material more accessible, provided the researcher adapts their responsibilities. The role shifts from collector and analyzer of purpose-built data to designer of the analytical setup. Writing the prompts that will operate on the corpus, deciding which organizational data sources are worth including and knowing when to escalate from breadth-first AI synthesis to depth-first interviewing. These are not

competencies that standard UX training currently covers like prompt engineering and reading AI errors diagnostically. This connects to the fact that AI-assisted research may raise the expertise threshold, discussed in Section 7.4. A researcher who understands the product and the conditions under which the LLM's output quality decreases will produce more usable findings.

For organizations, the most concrete implication concerns data hygiene. Teams that maintain structured support channels and sufficiently detailed prototypes can transition to continuous evaluation against live organizational data. If the structure is not there, that is the first investment that should be made, not the AI tooling itself.

In addition to this, the justificatory context bias identified in Section 7.5 is a risk in how results are interpreted. A team already committed to a roadmap may unconsciously frame their prompts in ways that invite rationalization rather than critique, or may selectively weight AI findings that confirm existing decisions. Establishing explicit protocols around analytical neutrality, particularly in contexts where the team has a stake in the outcome, is a practical safeguard this study suggests but cannot fully prescribe.

Lastly, the insights discussed in this chapter point toward a hybrid research model in which LLM-assisted analysis serves as a continuous breadth-first layer surfacing recurring patterns across organizational data at low marginal cost, while human research provides the depth-first layer of contextual and interpretive understanding. A visualization of this research model is presented in Figure 7.1.

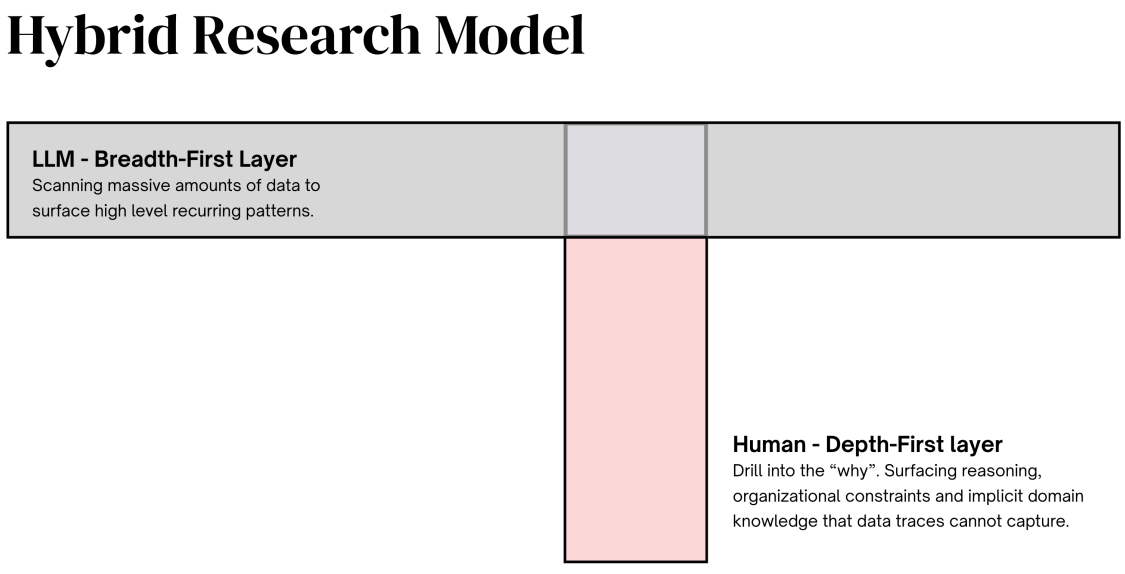


Figure 7.1: Visual representation of the hybrid research model.

7.8 Limitations

The study examined one product redesign within one organization using one research team. The Teamtailor context is broadly representative of B2B SaaS but not of all

software contexts. Consumer products, products without internal customer-facing channels and products in highly regulated domains all impose different conditions. Within the B2B SaaS scope, the findings are most likely to transfer to organizations whose internal data has been structured along the lines described in Section 7.6 and least likely to transfer to organizations whose data is spread out across inconsistent conventions.

A few methodological limitations should be noted. The LLM system was constrained to predefined output schemas (critical issues, positive signals, findings and top themes), which reduces the risk of hallucinated findings but may suppress patterns that do not align with these categories. The redesign evaluation was conducted on a Figma prototype with limited fidelity, and the consequences of this for AI-based prototype evaluation are discussed in Section 6.2.4. Beyond fidelity, the evaluation was also narrow in scope. The personas completed only 4 predefined task scenarios, leaving broader workflows such as approval chains, bulk posting and team collaboration untested. This accounts for the number of current design problems that mapped to unknown verdicts in the comparison matrix and means the redesign evaluation should be understood as a targeted evaluation of specific flows rather than a comprehensive assessment of the new design. The manual research drew on five contextual interviews, one survey, and 7 usability tests. These sample sizes are conventional for qualitative research but small enough that the absence of a finding in the manual research carries less weight than its presence.

The context bias finding described in Section 7.5 is consistent and reproducible within this study but has not been controlled across multiple organizations or model families and is best treated as a hypothesis worth investigating further. A controlled comparison randomizing prompt structure and model choice would strengthen the claim about prompt design and output quality but would require resources beyond the scope of this thesis. A related validity constraint concerns the synthetic persona walkthroughs. The think-aloud sessions produced by the LLM agents reflect the reasoning patterns and linguistic tendencies of the underlying model rather than the spontaneous, embodied behavior of real users encountering an unfamiliar interface. The personas were grounded in empirical research but the agents instantiated from them do not replicate the hesitations or exploratory clicking that characterizes genuine first-time use.

7.9 Ethics

Participants in the contextual interviews, usability tests, and survey were informed verbally of the purpose of the study and consented on that basis. No signed consent forms were collected, and the project did not go through a university ethics review. However, internal approval was provided by Teamtailor and participants were not identified by name in the analysis or in this thesis. Also direct quotes were paraphrased where identification might have been possible from context. Survey responses were also processed through an anonymization script before analysis, removing names, company names and other identifying fields so that the data entered the pipeline in de-identified form.

The more difficult ethical questions concern the naturally occurring operational data analyzed by Pipeline A. The Slack messages and Intercom support conversations used in the study were not produced for research purposes. The employees whose messages appeared in the corpus and the customers whose support tickets were processed, had not provided explicit research consent for that material to be used as academic research input. Their original agreement was to employment conditions or terms of service that likely cover operational use of these records, not necessarily academic analysis. This issue is not unique to the present study. Tools like MCP make it easy to connect an AI model directly to systems such as Slack or a support inbox, removing much of the manual effort that previously limited how operational data was reused. When access becomes this frictionless, data gathered for one purpose can be repurposed for another almost by default, before anyone has explicitly considered whether that use is appropriate. The consent gap described here is therefore likely to become a recurring problem as these pipelines spread. The current ease of connecting a model to a data source moves faster than the consent and governance practices that should accompany it. Several factors narrow, but do not remove, this concern in this study. The Slack corpus was scoped to a single internal feedback channel, #product-feedback, rather than drawn from general work communications and the Intercom material was limited to inbound support tickets filtered by keyword. The study was also conducted with internal approval from Teamtailor, which owns and operates these systems for business purposes. However, internal approval should not be treated as equivalent to informed consent from every employee or customer represented in the data. The justification is one of proportionality. The data was work-related, it stayed within the systems where it was produced and the analysis served a bounded purpose whose benefits return to roughly the same population the data came from. That is a more defensible position than analysing general communication data without restriction, but it does not resolve the underlying consent question.

Several technical choices in the pipeline design created ethical and technical limitations that should be addressed in future implementations. First, Intercom data was not anonymized before being made available through the pipeline. The pipeline instructed the language model to reduce customer identities to a company tier as part of its analysis output, but this anonymization occurred after tool-mediated retrieval rather than before analysis. The use of an MCP connection does not mean that the full Intercom corpus was automatically transmitted to the model provider. Rather, the relevant exposure concerns the specific tool calls made during the agent run. Tool definitions, tool inputs and returned tool results became part of the agent interaction. In practice, any customer text returned by the Intercom MCP tools, potentially including names or identifying descriptions, could therefore be passed back to Claude as tool results before systematic redaction was applied. The instruction was also qualified as “if possible”, making anonymization discretionary rather than guaranteed. A more cautious implementation would treat systematic anonymization as a precondition of the pipeline, either by redacting data before it is exposed to the agent or by ensuring that the MCP tools themselves return only anonymized excerpts. A related issue concerns trace retention. The system’s tracing mechanism saved short content previews of tool results, including fragments of Slack messages

and Intercom conversations, to local JSON files during each run. Future implementations should define an explicit retention and deletion policy for such trace files rather than leaving them as passive artifacts of the infrastructure.

A further limitation concerns the autonomy of the agents. The agents made decisions about which messages to read, which threads to follow, and how deeply to traverse conversations without per-call human review. A human researcher reading Slack uses judgement about relevance and sensitivity at the moment of each access. An automated agent does not exercise judgement in the same way. This does not change the consent situation, but it affects the legibility of data access: what was actually seen during a run is distributed across the agent's reasoning and trace outputs rather than recorded in a human researcher's notes.

Another consideration concerns the UX research profession itself. The synthetic think-aloud sessions in Pipeline B raise no privacy concerns, since they involved constructed personas rather than human participants, but they reinforce the broader question of how much weight AI-generated findings should carry in UX research. The accountability for a finding that originates in a LLM analysis pipeline is less clearly located than for a finding that originates in a named researcher's interview notes. Once such findings begin to inform design decisions, the question of who is answerable for them, particularly when they are wrong, is not resolved by the protocol described here. A related concern is that the method's dependence on expert judgement, established in Section 7.4, limits some of the efficiency gains it appears to offer. If the output is only trustworthy when guided by an experienced researcher, then the method is less a replacement for research expertise than a way of amplifying it. This also creates a training concern: if junior researchers learn the craft through workflows in which the model produces the first draft of an analysis, they may have fewer opportunities to develop the interpretive judgement on which the method itself depends. The cost of producing a finding has dropped, but the cost of representing a user well has not.

Finally, the broader implications of the method should be treated carefully. The same workflow that allows a UX researcher to surface useful patterns from employee Slack and customer support text could also make it easier for organizations to analyze these channels routinely in ways that resemble monitoring rather than research. The argument of this thesis is that LLM-assisted analysis of operational data has value when conducted carefully and with appropriate skepticism. Organizations adopting this approach should be explicit, internally and to their customers, about which corpora are being analyzed and for what purposes.

8

Conclusion

This thesis set out to investigate whether large language models can be used to surface meaningful UX insights from naturally occurring data in a B2B SaaS context and to characterize the opportunities and limitations of doing so.

Addressing RQ1, the main finding is that LLM-assisted analysis and manual research have different but complementary reach. LLM-only findings clustered in silent system behavior and workflow constraints while human-only findings clustered in domain knowledge and missing functionality. The two tracks converged most strongly on the most critical issues, all of which informed design changes. The opportunities of the approach are however limited by several conditions. Domain expertise had to be applied into the analytical setup before any data was processed. Without it, the model defaulted to generic observations and the LLM system's scale amplified noise rather than usability signals. Prompt design itself functioned as a research instrument where justificatory framing led the model to rationalize design decisions while neutral domain framing produced more critical analysis. The pipeline's reach was also limited by the structural readiness of the underlying data. Noteworthy is also that the LLM's misreadings carried analytical weight, surfacing workflow questions that the manual research had not brought up.

In response to RQ2, the study developed and validated a multi-agent pipeline in which LLMs analyze naturally occurring organizational data through several design decisions that proved central to its usefulness. Prompts were framed with neutral domain context rather than justificatory framing. MCP integration allowed agents to access data sources directly without manual extraction. Together these conditions allowed the pipeline to surface analytically meaningful UX insights inductively from Slack feedback and Intercom support conversations.

Finally addressing RQ3, the synthetic user evaluation demonstrated that reliability and usefulness depend heavily on two conditions. First, personas should be grounded in prior empirical research rather than generated from scratch. When anchored in real product knowledge, the agents produced differentiated, segment-specific findings that challenge the generic-output critique common in the literature. Second, the fidelity of the prototype itself functions as a ceiling on what synthetic evaluation can surface. 3 of 4 misreadings in this study were attributable to missing click targets rather than analytical failure. Notably, even coherent misreadings carried diagnostic value, surfacing structural design assumptions that the manual research had not articulated.

As mentioned, the results in this study point towards adoption of the hybrid research model that is described in Section 7.7. The conditions under which this model works are as important as the model itself. It depends on user segmentation grounded in prior research, on prompts treated as analytical instruments and on internal data structured finely enough to carry meaningful distinctions. For B2B SaaS organizations, this reframes data hygiene in tools like Slack, Intercom and Figma as research infrastructure.

8.1 Future Research

The contribution of this thesis opens onto a set of further questions. For example, about the context-bias finding described in Section 7.5. The result was reproducible within the present study but its scope remains to be established. Whether the same shift from critical to defensive analysis occurs in other domains where the operator has a stake in the conclusions such as policy analysis or medical diagnosis is a hypothesis the present study can only suggest. Another question concerns the mechanism of the bias itself. Which features of a prompt actually carry the bias and in what proportions? Is it the quantity of narrative context or the specificity with which desired outcomes are articulated? A controlled comparison varying features independently would clarify what about justificatory context produces the effect. It would also open the possibility of writing prompts that retain the necessary domain grounding while neutralizing the bias. Furthermore, the architecture of the method itself should be thoroughly tested. As mentioned, the model runs in this study all used Claude and the convergence and divergence patterns observed may not reproduce unchanged across other model families and sizes.

This study has shown that LLM-assisted analysis of naturally occurring data can produce meaningful UX insight, but that its usefulness is shaped by the conditions under which it operates. The questions worth pursuing next are therefore not primarily about what LLMs can do but about what researchers and organizations need to put in place for them to do it well.

Bibliography

[1] G. C. Guerino, S. Martinelli, J. Choma, G. C. L. Leal, R. Balancieri, and L. Zaina, "Challenges of UX research and long-term UX: A multiple case study with software startups," in *Proceedings of the 13th Nordic Conference on Human-Computer Interaction (NordCHI 2024)*, 2024. DOI: 10.1145/3679318.3685381. [Online]. Available: https://dl.acm.org/doi/10.1145/3679318.3685381.

[2] A. Tkalich, E. Klotins, T. Sporsem, et al., "User feedback in continuous software engineering: Revealing the state-of-practice," *Empirical Software Engineering*, vol. 30, p. 79, 2025. DOI: 10.1007/s10664-024-10557-2.

[3] Q. Wang, M. Erqsous, K. Barner, and L. Mauriello, "LATA: A pilot study on LLM-assisted thematic analysis," in *Proceedings of the ACM on Human-Computer Interaction (CSCW 2025)*, 2025. [Online]. Available: https://www.eecis.udel.edu/~mlm/docs/2025-Wang-CSCW-LATA-Paper.pdf.

[4] S. De Paoli, "Performing an inductive thematic analysis of semi-structured interviews with a large language model: An exploration and provocation on the limits of the approach," *Social Science Computer Review*, vol. 42, no. 4, pp. 997–1019, 2024, Published online December 2023; formally assigned to Vol. 42 No. 4 in 2024. DOI: 10.1177/08944393231220483.

[5] Teamtailor, *Teamtailor — recruitment software & applicant tracking system*, https://www.teamtailor.com, Accessed: April 2026, 2026.

[6] A. H. AL-Qassem, K. Agha, M. Vij, H. Elrehail, and R. Agarwal, "Leading talent management: Empirical investigation on applicant tracking system (ats) on e-recruitment performance," in *2023 International Conference on Business Analytics for Technology and Security (ICBATS)*, 2023, pp. 1–5. DOI: 10.1109/ICBATS57792.2023.10111172.

[7] S. Laumer, C. Maier, and A. Eckhardt, "The impact of business process management and applicant tracking systems on recruiting process performance: An empirical study," *Journal of Business Economics*, vol. 85, no. 4, pp. 421–453, 2015. DOI: 10.1007/s11573-014-0758-9.

[8] F. B. Council, "How b2b companies can create a more consumer-like business experience," *Forbes*, Oct. 2024, Accessed: May 15, 2026.

[9] A. Dix, J. Finlay, G. D. Abowd, and R. Beale, *Human-Computer Interaction*, 3rd ed. Harlow, England: Pearson Education, 2004.

[10] J. M. Carroll, *HCI Models, Theories, and Frameworks: Toward a Multidisciplinary Science*. San Francisco, CA: Morgan Kaufmann, 2003.

[11] J. C. R. Licklider, "Man-computer symbiosis," *IRE Transactions on Human Factors in Electronics*, vol. HFE-1, pp. 4–11, 1960. DOI: 10.1109/THFE2.1960.4503259.

[12] E. Horvitz, "Principles of mixed-initiative user interfaces," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '99)*, ACM, 1999, pp. 159–166. DOI: 10.1145/302979.303030.

[13] B. Shneiderman, "Human-centered artificial intelligence: Reliable, safe & trustworthy," *International Journal of Human–Computer Interaction*, vol. 36, no. 6, pp. 495–504, 2020. DOI: 10.1080/10447318.2020.1741118.

[14] R. Parasuraman and D. H. Manzey, "Complacency and bias in human use of automation: An attentional integration," *Human Factors*, vol. 52, no. 3, pp. 381–410, 2010. DOI: 10.1177/0018720810376055.

[15] J. Preece, Y. Rogers, and H. Sharp, *Interaction Design: Beyond Human-Computer Interaction*, 4th ed. Chichester, England: Wiley, 2015.

[16] D. Silverman, *Interpreting Qualitative Data*, 5th ed. London: SAGE Publications, 2014.

[17] P. Törnberg et al., "Scaling hermeneutics: A guide to qualitative coding with LLMs for reflexive content analysis," *EPJ Data Science*, vol. 14, 2025. DOI: 10.1140/epjds/s13688-025-00548-8.

[18] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017. [Online]. Available: https://arxiv.org/abs/1706.03762.

[19] T. Tully, J. Redfern, D. Das, and D. Xiao, "2025: The state of generative ai in the enterprise," Menlo Ventures, Tech. Rep., Dec. 2025, Accessed: 2026-04-05. [Online]. Available: https://menlovc.com/wp-content/uploads/2025/12/menlo_ventures_enterprise_ai_report-2025-123125.pdf.

[20] Anthropic, *Claude: An AI assistant by Anthropic*, https://www.anthropic.com, Accessed: 2026-02-01, 2024.

[21] J. White et al., "A prompt pattern catalog to enhance prompt engineering with ChatGPT," *arXiv preprint arXiv:2302.11382*, 2023. [Online]. Available: https://arxiv.org/abs/2302.11382.

[22] Anthropic, *Using CLAUDE.MD files: Customizing Claude Code for your codebase*, https://claude.com/blog/using-claude-md-files, Accessed: 2026-05-10, 2026.

[23] Anthropic, *Introducing the model context protocol*, https://www.anthropic.com/news/model-context-protocol, Accessed: 2026-05-10, Nov. 2024.

[24] Z. Xiao, X. Yuan, Q. V. Liao, R. Abdelghani, and P.-Y. Oudeyer, "Supporting qualitative analysis with large language models: Combining codebook with GPT-3 for deductive coding," in *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23 Companion)*, ACM, 2023, pp. 75–78. DOI: 10.1145/3581754.3584136.

[25] J. Salminen, D. Amin, S.-G. Jung, and B. J. Jansen, "The use of large language models in HCI: A critical analysis of synthetic users," in *Proceedings of the Augmented Humans International Conference 2025 (AHs '25)*, Masdar City, Abu Dhabi, United Arab Emirates: Association for Computing Machinery, 2025, pp. 413–417, ISBN: 979-8-4007-1566-2. DOI: 10.1145/3745900.3746108.

[26] Y. Tao, O. Viberg, R. S. Baker, and R. F. Kizilcec, "Cultural bias and cultural alignment of large language models," *PNAS Nexus*, vol. 3, no. 9, pgae346, 2024, Preprint available as arXiv:2311.14096. DOI: 10.1093/pnasnexus/pgae346.

[27] D. M. Russell, "The challenges of synthetic users in UX research," *Interactions*, Dec. 2025, ACM Interactions Blog. DOI: 10.1145/3779007. [Online]. Available: https://interactions.acm.org/blog/view/the-challenges-of-synthetic-users-in-ux-research.

[28] W. Xiang et al., "SimUser: Generating usability feedback by simulating various users interacting with mobile applications," in *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*, Honolulu, HI, USA: Association for Computing Machinery, 2024, ISBN: 979-8-4007-0330-0. DOI: 10.1145/3613904.3642481.

[29] F. Pehar, "AI as synthetic users in human-centred design: A systematic review," in *Proceedings of the 48th MIPRO ICT and Electronics Convention (MIPRO 2025)*, Opatija, Croatia: IEEE, 2025, ISBN: 979-8-3315-3597-1. DOI: 10.1109/MIPRO65660.2025.11131969.

[30] Maze, *8 UX research trends shaping 2025 (and what to watch in 2026)*, Accessed April 2026, 2025. [Online]. Available: https://maze.co/blog/ux-research-trends/.

[31] Y. Lu, Y. Yang, and C. Zhang, "AI assistance for UX: A literature review through human-centered AI," in *arXiv preprint arXiv:2402.06089*, v2, 2024. [Online]. Available: https://arxiv.org/abs/2402.06089.

[32] S. S. Wells, "The impact of AI on UX: Challenges and opportunities," Master's Thesis, Lindenwood University, Saint Charles, Missouri, May 2024. [Online]. Available: https://digitalcommons.lindenwood.edu/theses/995.

[33] Å. Stige, E. D. Zamani, P. Mikalef, and Y. Zhu, "Artificial intelligence (AI) for user experience (UX) design: A systematic literature review and future research agenda," *Information Technology & People*, vol. 37, no. 6, pp. 2324–2352, 2024. DOI: 10.1108/ITP-07-2022-0519.

[34] J. Zimmerman, J. Forlizzi, and S. Evenson, "Research through design as a method for interaction design research in HCI," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2007)*, 2007, pp. 493–502. DOI: 10.1145/1240624.1240704. [Online]. Available: https://dl.acm.org/doi/10.1145/1240624.1240704.

[35] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative Research in Psychology*, vol. 3, no. 2, pp. 77–101, 2006. DOI: 10.1191/1478088706qp063oa.

[36] A. Cooper, *The Inmates Are Running the Asylum: Why High-Tech Products Drive Us Crazy and How to Restore the Sanity*. Indianapolis, IN: Sams, 1999, ISBN: 978-0672316494.

[37] J. Pruitt and T. Adlin, *The Persona Lifecycle: Keeping People in Mind Throughout Product Design*. San Francisco, CA: Morgan Kaufmann, 2006, ISBN: 978-0125662513.

[38] H. Beyer and K. Holtzblatt, *Contextual Design: Defining Customer-Centered Systems*. San Francisco, CA: Morgan Kaufmann, 1997.

[39] K. Holtzblatt, J. B. Wendell, and S. Wood, *Rapid Contextual Design: A How-to Guide to Key Techniques for User-Centered Design*. San Francisco, CA: Morgan Kaufmann, 2005.

[40] J. Sauro and J. R. Lewis, *Quantifying the User Experience: Practical Statistics for User Research*. Waltham, MA: Morgan Kaufmann, 2012.

[41] J. S. Dumas and J. C. Redish, *A Practical Guide to Usability Testing*. Norwood, NJ: Ablex, 1993.

[42] J. Nielsen, *Usability Engineering*. San Diego, CA: Academic Press, 1993.

[43] B. Lobe, D. Morgan, and K. A. Hoffman, "Qualitative data collection in an era of social distancing," *International Journal of Qualitative Methods*, vol. 19, 2020. DOI: 10.1177/1609406920937875.

[44] M. B. Miles, A. M. Huberman, and J. Saldaña, *Qualitative Data Analysis: A Methods Sourcebook*, 3rd ed. Thousand Oaks, CA: SAGE Publications, 2014.

[45] K. A. Ericsson and H. A. Simon, *Protocol Analysis: Verbal Reports as Data*, revised. Cambridge, MA: MIT Press, 1993.

A

Appendix A (Track 1)

A.1 Interview Questions

The following questions formed the semi-structured interview guide used during the contextual interviews described in Section 5.1. Questions were used as a flexible guide rather than a fixed script; follow-up questions were asked as needed based on participant responses and observed behavior during the screen-sharing task.

Opening: Role and Context

1. Can you describe your role and how job postings fit into your typical workday?
2. How often do you create new job postings in Teamtailor?
3. Do you tend to create postings yourself, or is this shared with others on your team?
4. How long have you been using Teamtailor?

Task: Create a Job Posting (think-aloud)

5. Please share your screen and create a job posting as you normally would. Talk through what you are doing and thinking as you go.

(Researcher observes and asks brief clarifying questions as needed during the task. Example probes: "What were you expecting to happen there?", "What does that mean to you?", "How do you usually handle that?")

Reflection: Satisfaction and Pain Points

6. How did that feel compared to your usual experience?
7. Were there any steps that felt unclear or took longer than expected?
8. Are there things you regularly need to do that the current flow does not make easy?
9. If you could change one thing about this process, what would it be?
10. Are there any features or steps you consider essential that you would not want to lose?

A.2 Survey Questions

1. How would you describe your role in job postings at your company?

- ○ I set up job postings regularly myself
- ○ I occasionally set up job postings
- ○ I mainly use templates set up by someone else
- ○ This is one of my first times setting up a job posting

2. How easy to understand is the current step-by-step process for creating a new job?
(1 = Very Confusing, 5 = Very Easy)

3. While setting up a job posting, how confident do you feel that you are doing things correctly?
(1 = Very uncertain, 5 = Completely confident)

4. Did you need to seek help at any point (e.g., colleague, documentation, support)?
(Yes / No)

→ If yes: What did you need help with? *(Open text - optional)*

5. If you could change ONE thing about the job setup process, what would it be?
(Open text - optional)

6. Do you have any other observations or feedback regarding the current workflow for job creation?
(Open text - optional)

7. If you use the Requisitions or Job offers add ons, do you have any observations or feedback regarding these functionalities?
(Open text - optional)

B

Appendix B (Final Findings)

B.1 Full Findings Coding

The table below presents all 51 findings from both design contexts, coded by source and nature of issue. Human research findings are limited to those relating directly to the job setup flow, as this was the scope of the AI analysis and the primary focus of the comparative evaluation.

Table B.1: All findings coded by source and nature of issue.

Finding	**Design**	**Source**	**Nature**
Current design			
LinkedIn location maps wrong silently	Current	Convergent	Silent system behavior
Hiring team access can't be set during creation	Current	Convergent	Missing functionality
Manual entry pain point, users want requisition data to flow into job post	Current	Convergent	Workflow & operational
Location silently defaults to HQ if left empty	Current	AI only	Silent system behavior
No autosave	Current	AI only	Missing functionality
Requisition approval creates problems for long drawn processes	Current	AI only	Workflow & operational
Protected template restrictions hidden until after creation	Current	AI only	Silent system behavior
Generic save errors don't identify which field failed	Current	AI only	UI & interaction

Continued on next page

Continued from previous page

Finding	**Design**	**Source**	**Nature**
Internal job name can't be set during creation, requires a return visit	Current	AI only	Workflow & operational
Recruiter can't be reassigned after delegated creation	Current	AI only	Workflow & operational
No post-creation guidance, users stall at the handoff moment	Current	AI only	Navigation & IA
Dark mode renders job description editor header text black, invisible	Current	AI only	UI & interaction
Publish state invisible, career site sync broken (bug, not UX)	Current	AI misreadings	UI & interaction
AI features not used much in practice	Current	Human only	Domain & context
Heavy reliance on template for job creation	Current	Human only	Domain & context
Importance of custom fields (organisational context)	Current	Human only	Domain & context
Calendar / end-date handling major pain point, jobs stay live too long	Current	Human only	Workflow & operational
Cannot make custom fields mandatory	Current	Human only	Missing functionality
Text editor friction	Current	Human only	UI & interaction
No ability to edit a question once added; no bulk-add	Current	Human only	Missing functionality
Stage management feels buggy and too high-level	Current	Human only	Workflow & operational
Strong demand for better AI-assisted features across setup	Current	Human only	Domain & context
Documentation to be included in the requisition should be a mandatory field	Current	Human misreadings	Missing functionality

Continued on next page

Continued from previous page

Finding	**Design**	**Source**	**Nature**
No option to add details such as contract type, salary, or pre-screening questions only possible via the main text field	Current	Human mis-readings	Missing functionality
Would be nice to remove unused steps such as *Evaluation* and *Candidate Interaction*	Current	Human mis-readings	Missing functionality
New design			
No visible Publish button (4/7 human users)	New	Convergent	UI & interaction
Inverted CTA hierarchy on Create Job modal (6/7 human users most critical)	New	Convergent	UI & interaction
No Next/Continue button, sidebar navigation not discoverable (4/7 human users)	New	Convergent	Navigation & IA
Send message trigger: 0% completion (2/7 users)	New	Convergent	Silent system behavior
Location field on two screens with no signal that data carried over (1/7)	New	Convergent	UI & interaction
Desktop + mobile preview on Appearance step (1/7)	New	Convergent	Positive feedback
Autosave "Last saved: Just now" indicator (1/7)	New	Convergent	Positive feedback
Data carry-over between steps works seamlessly (1/7)	New	Convergent	Positive feedback
"Symbol" field on internal title: no explanation, two personas ignored it	New	AI only	UI & interaction
"Role" dropdown ambiguous vs "Internal title"	New	AI only	UI & interaction
Write with Copilot feature	New	AI only	Positive feedback

Continued on next page

Continued from previous page

Finding	**Design**	**Source**	**Nature**
Salary range toggle	New	AI only	Positive feedback
Interview kit: 0% completion, no escape hatch (prototype problem)	New	AI misreadings	Navigation & IA
Reject message activation state opaque (prototype problem)	New	AI misreadings	Silent system behavior
No escape from modal to pipeline (prototype problem, raised valid workflow question)	New	AI misreadings	Navigation & IA
General thoughts on job / job-ad split in the new design	New	Human only	Navigation & IA
Domain context: laws for different countries affect setup decisions	New	Human only	Domain & context
Job description field missing in requisition	New	Human only	Missing functionality
Template selection should come first in the flow	New	Human only	Navigation & IA
Preview should be final check including application form, not just appearance step	New	Human only	Navigation & IA
Confusion around timeframe field	New	Human only	UI & interaction
Accessibility: trigger color coding insufficient for colorblind users	New	Human only	UI & interaction
Weekend exclusion from triggers	New	Human only	Missing functionality
Appreciated the research and testing approach before launch	New	Human only	Positive feedback
Consolidated view of all triggers was appreciated	New	Human only	Positive feedback
Sidebar progress checkmarks: AI praised, 2/7 human users confused by them	New	Contradiction	UI & interaction