

TU Dortmund University
Joseph von Fraunhofer Str. 25
D-44227 Dortmund
Department for Informatics

Bachelor Thesis

Design and Implementation of a Reflective Well-Being App

Dennis Pfeiffer

dennis.pfeiffer@tu-dortmund.de

Course of Study: Computer Science

First Examiner: Prof. Dr. Sven Mayer

Second Examiner: Prof. Dr. Mario Botsch

Commenced: November 17, 2025

Completed: March 17, 2026

“Design and Implementation of a Reflective Well-Being App”
2026-06-15

© 2026 Dennis Pfeiffer

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 License (CC BY-SA 4.0):
<http://creativecommons.org/licenses/by-sa/4.0/>



Typesetting: PDF-L^AT_EX 2_ε
Character Count (excluding spaces): 89207

Abstract

Mood-tracking apps can support mental health self-management by helping users monitor their emotional states and identify patterns in their daily lives. However, most existing apps focus on data collection and visualisation while offering little personalised guidance, and many raise serious privacy concerns by sharing sensitive data with third-party services. This thesis presents MoodScape, a reflective well-being app for Android that combines context-aware mood tracking with personalised, LLM-generated recommendations while preserving user privacy through a locally hosted open-source language model. MoodScape collects mood entries alongside contextual signals from weather, music, health, and social interaction APIs and uses a locally deployed Ollama model (llama3.2:1b) on a university-managed server to generate tailored recommendations—without transmitting any mood or health data to commercial cloud providers.

The system was evaluated in a two-week AB/BA crossover field study with ten participants. Each participant experienced both a tailored recommendation condition, in which the LLM drew on their personal mood and context data, and a generic baseline condition, in counterbalanced order. Perceived recommendation quality, system usability (SUS), user experience (UEQ), and self-reported reflection were assessed through standardised questionnaires.

The results show a consistent descriptive pattern favouring the tailored condition: participants rated the LLM-generated recommendations higher than the baseline on all four comparison items, with the largest difference on perceived personalisation ($M = 2.80$ vs. $M = 2.00$), which also reached statistical significance in a supplementary Wilcoxon signed-rank test ($p = .039$, $r = .65$). However, absolute satisfaction levels remained moderate across both conditions, and engagement with the app varied considerably across participants (4–38 mood entries). MoodScape achieved above-average usability ($M_{\text{SUS}} = 74.75$) and positive user experience scores, with participants particularly valuing the data exploration features. Reflection items indicate that the app supported data exploration and, to a degree, self-reflection, though deeper behavioural change did not emerge within the study period.

These findings demonstrate that privacy-preserving, locally hosted LLMs are technically viable for generating mood-related recommendations in a GDPR-compliant architecture and that the personalisation approach itself is perceived positively. The quality gap compared to larger commercial models, however, constrained the practical impact. As open-source language models continue to improve, the approach demonstrated by MoodScape offers a promising path towards AI-powered well-being support that respects user privacy.

Contents

1. Introduction	11
2. Related Work	13
2.1. Mood-tracking and mental-health apps	13
2.2. Privacy of mental-health and mood-tracking apps	14
2.3. Data protection and GDPR for mental-health apps	15
2.4. Usability and UX evaluation in HCI	16
2.5. Reflectivity and technology-supported reflection	16
3. App Design & Implementation	19
3.1. System Architecture	19
3.2. Frontend	24
3.3. Backend	32
4. Study Design	35
4.1. Apparatus	35
4.2. Procedure	35
4.3. Measurements	35
4.4. Participants	36
5. Results	39
5.1. Study Compliance and Data Overview	39
5.2. Perceived Quality of Correlation Insights	40
5.3. Perceived Quality of Recommendations	41
5.4. System Usability	44
5.5. User Experience	44
5.6. Reflection on App Usage	46
5.7. Qualitative Feedback	48
6. Discussion	49
6.1. Perceived Quality of Recommendations and Correlations	49
6.2. Usability and User Experience	50
6.3. MoodScape as a Reflective Well-Being App	51
6.4. Privacy and the Local LLM Architecture	52
6.5. Reasons for Not Following Recommendations	52
6.6. Limitations	53
7. Conclusion	55
7.1. Outlook	56
Bibliography	57
A. Full Survey Questionnaires	63
B. Informed Consent Form	83

C. Additional Visualizations	87
C.1. Age Distribution of Participants	87
C.2. Other Demographic Data	88
D. Background	89
D.1. Wilcoxon Signed-Rank Tests	89
D.2. Ollama format Schema	89
D.3. Cleaned Payload Example	90
D.4. Example Baseline Recommendation	92
D.5. Example Tailored Recommendation	93
D.6. Guiding Recommendations sent with every Baseline Recommendation Payload . . .	94
D.7. Additional <i>MoodScape</i> Screenshots	97

List of Figures

3.1.	Client–server architecture of <i>MoodScape</i>	20
3.2.	Login- and Permission screen	25
3.3.	Overview screen and details views in <i>MoodScape</i>	26
3.4.	Multi-step entry flow. Users first select the entry type and mood, then specify influencing factors and context, and optionally add notes. Screen (c) is scrollable.	27
3.5.	Additional screens shown when Social Interaction is selected as an influence, in order of appearance.	28
3.6.	Screens exclusive to daily recaps.	29
3.7.	Mood analytics view showing monthly trends and distributions of mood labels.	30
3.8.	Scrollable recap analytics view: (a) calendar with mood icons per day, (b) Recap History showing the monthly distribution of recap moods.	31
3.9.	Recommendation screen in <i>MoodScape</i> (tailored condition shown).	32
4.1.	Self-reported factors influencing mood among study participants (n=10)	37
5.1.	Mood entries per participant, split by study week. Colour indicates the recommendation condition (teal = tailored, rust = baseline); hatching distinguishes Week 2 from Week 1. P06 did not create any entries in Week 2.	40
5.2.	Mean ratings for perceived quality of correlation insights by condition. Error bars represent one standard deviation.	41
5.3.	Mean ratings for perceived recommendation quality by condition. Error bars represent one standard deviation.	42
5.4.	Individual participant ratings for correlation insights (left) and recommendations (right) across baseline (B) and tailored (T) conditions. Higher values indicate more positive ratings on a 1–5 scale.	43
5.5.	Individual SUS scores per participant, colour-coded by adjective rating. The dashed lines indicate the empirical average (68) and Grade A threshold (80.3) [40]. Mean and SD rounded up.	44
5.6.	Mean UEQ scale scores with standard deviations. Background shading indicates the UEQ benchmark categories [42].	45
5.7.	UEQ scores per participant across all six scales. Warmer (greener) colours indicate more positive ratings.	46
5.8.	Mean agreement ratings for reflection statements, grouped by category. Individual data points are shown as dots. The dashed line indicates the neutral midpoint (4).	47
C.1.	Age distribution of study participants.	87
C.2.	(a) Gender Distribution, (b) Occupation, (c) Prior Mood Tracking Experience and (d) Regular App Usage	88
D.1.	Additional <i>MoodScape</i> screens.	97

List of Tables

3.1. Fields stored for mood entries and daily recaps in MoodScape	21
3.2. Core backend endpoints for user and entry management	33
5.1. Mood entries per participant and study week. “Cond.” indicates whether the participant received tailored (T) or baseline (B) recommendations in that week. Active days denotes the number of distinct days with at least one mood entry.	39
5.2. Perceived quality of correlation insights by condition ($n = 10$ per condition)	40
5.3. Perceived quality of recommendations by condition ($n = 10$ per condition)	41
5.4. Reasons for not following recommendations by condition (multiple selections allowed) . .	43
5.5. UEQ scale scores ($n = 10$; scale range: -3 to $+3$)	45
5.6. Reflection on app usage ($n = 10$; 7-point Likert scale)	47
D.1. Wilcoxon signed-rank test results for baseline vs. tailored condition ($n = 10$ pairs, two-sided)	89

1. Introduction

Mental health conditions such as depression and anxiety affect hundreds of millions of people worldwide and now account for a substantial share of the global disease burden [16, 51]. Fan et al. [16] report that depressive disorders alone rank among the leading causes of disability globally, while Yang et al. [51] highlight the continued rise in prevalence across young people. At the same time, public mental health services in many countries struggle with long waiting times and limited capacity, leaving a significant gap between the demand for professional support and its availability [17]. Digital mental health tools—in particular mood-tracking apps—have emerged as a promising complement to traditional care, enabling users to monitor their emotional states, identify patterns, and engage in self-management [11, 43]. However, Caldeira et al. [11] found that most existing mood-tracking apps focus primarily on data collection and basic visualisation, offering little support for interpreting the recorded data or translating insights into concrete, personalised actions. Schueller et al. [43] confirmed this through interviews with 22 users who reported a clear desire for app-facilitated recommendations, rather than only charts and statistics. Moreover, many of these apps raise serious privacy concerns: Iwaya et al. [25] found that a majority of popular mental-health apps exhibit unnecessary permissions, insecure cryptography, and leaks of personal data in logs and web requests, while Huckvale et al. [22] reported that 92% of apps for depression and smoking cessation transmitted data to third parties, often in ways not disclosed to users. Independent audits by organisations such as the Mozilla Foundation [33] have similarly concluded that most reviewed mental-health apps fall into their worst privacy categories. This creates a tension between the potential benefits of AI-powered personalisation and the need to protect highly sensitive mood and health data—a tension that is particularly acute under the strict requirements of the European General Data Protection Regulation (GDPR) [24].

Prior work has explored various approaches to enriching mood-tracking apps beyond simple logging. Polhemus et al. [36] analysed user reviews of depression self-management apps and found that tracking and visualisation features help users increase self-awareness and identify trends, but that users consistently request more personalised feedback and guidance. Context-aware systems that integrate mood data with signals such as location, activity, sleep, or other contextual factors have shown promise in helping users discover correlations that might not be obvious from raw logs alone; Burns et al. [10] demonstrated this with *Mobilyze!*, a context-sensing mobile intervention for depression that used machine learning to predict patients' mood and contextual states from phone sensor data. Large language models (LLMs) offer a new avenue for generating personalised, natural-language recommendations based on contextual data. However, deploying LLMs in the mental health domain introduces a fundamental trade-off: cloud-based commercial models such as GPT-4 provide strong performance but require transmitting sensitive data to third-party servers, raising confidentiality and regulatory concerns [6], while locally hosted open-source models can keep data within institutional infrastructure but may offer lower or less predictable output quality, especially on constrained hardware [27]. Research on technology-supported reflection further highlights that effective self-tracking tools should not merely collect data but actively scaffold the user's sense-making process—moving from passive tracking towards reflective well-being support. Bentvelzen et al. [5] developed and validated the Technology-Supported Reflection Inventory (TSRI) specifically to measure whether interactive systems support such reflective engagement. To date, there has been relatively little work on combining context-aware LLM-generated recommendations with privacy-preserving, locally hosted architectures, and empirical evidence on how users perceive such recommendations compared to generic alternatives remains scarce.

This thesis presents *MoodScape*, a reflective well-being app for Android that addresses this gap in the field of Human-Computer Interaction (HCI). *MoodScape* tracks users' moods alongside contextual data from weather, music, health, and social interaction APIs, and uses a locally hosted open-source LLM (Ollama,

1. Introduction

llama3.2:1b) running on a university-managed server to generate personalised recommendations—without sending any mood or health data to commercial cloud providers. The app was evaluated in a two-week AB/BA crossover field study with ten participants, in which each participant experienced both a tailored recommendation condition (drawing on their personal mood and context data) and a generic baseline condition, in counterbalanced order. The study assessed perceived recommendation quality, system usability (SUS), user experience (UEQ), and self-reported reflection on app usage. What distinguishes this work from prior approaches is the explicit combination of three design goals: (1) context-aware, LLM-generated recommendations for mood improvement, (2) a privacy-by-design architecture that keeps all data processing within a controlled institutional environment, and (3) an exploratory empirical assessment of how users perceive tailored recommendations compared to a curated baseline in terms of personalisation and helpfulness.

The results show a consistent descriptive pattern favouring the tailored condition: participants rated the LLM-generated recommendations higher than the baseline on all four comparison items, with the largest difference on perceived personalisation ($M = 2.80$ vs. $M = 2.00$). However, absolute satisfaction ratings remained moderate across both conditions, and the barriers to following the recommendations were nearly equally distributed, suggesting that model quality—rather than the availability of personal data—remains the central bottleneck. MoodScape achieved above-average usability (SUS $M = 74.75$) and positive user experience ratings, with participants particularly valuing the data exploration and correlation features. Engagement with the app varied considerably across participants, ranging from 4 to 38 mood entries over the study period, which underlines the importance of individual differences in interpreting the results. The reflection items indicate that MoodScape successfully supported data exploration and, to a degree, self-reflection, though deeper behavioural change did not manifest within the two-week study period. These findings suggest that privacy-preserving, locally hosted LLMs are technically viable for generating mood-related recommendations and that the personalisation approach itself is perceived positively, but that the quality gap compared to larger commercial models constrains the practical impact. As open-source models continue to improve, the approach demonstrated by MoodScape may become increasingly competitive, offering a path towards AI-powered well-being support that respects users' data privacy.

2. Related Work

More people are living with depression, anxiety and other mental health conditions than ever before, and these disorders now account for a substantial share of the global disease burden [16, 51]. Fan et al. [16] report that depressive disorders rank among the leading causes of disability globally, while Yang et al. [51] highlight the continued rise in prevalence across all age groups. Public mental health services struggle to keep up: as Fiorillo [17] notes, waiting times for psychotherapy are long, and many patients cannot access the help they need when they need it. This situation has sparked growing interest in digital mental health tools, such as mood-tracking apps, that aim to support self-reflection, symptom monitoring and self-management alongside traditional care [11, 43].

2.1. Mood-tracking and mental-health apps

Mood-tracking apps have become increasingly common over the last years as one way to support people with mental health problems and complement traditional care [11, 32]. Caldeira et al. [11] conducted an in-depth feature analysis of 32 commercially available mood-tracking apps and found that most support the collection and reflection stages of self-tracking, while McCloud et al. [32] showed that digital mental health interventions can achieve meaningful therapeutic effects.

These apps allow users to record their emotional state over time, often using simple rating scales, icons or emojis, optionally combined with free-text notes and activity tags. They are used both by people with diagnosed mental health conditions and by members of the general population who want to better understand their emotional patterns and well-being. Typical features include daily logging, reminders, and visualisations such as calendars or line charts that show how mood changes across days or weeks [11, 49].

People usually start using mood-tracking apps in response to stress, negative life events or worsening mental health, with the goal of gaining insight and a sense of control over their emotions. In an interview study with 22 participants, Schueller et al. [43] found that users are particularly interested in recognising patterns—such as how sleep, work, social interactions or physical activity influence their mood—and in using these insights to adjust their daily routines. Qualitative studies also report that some users employ mood logs as a memory aid in therapy, for example by bringing their recorded moods and notes to sessions to make it easier to discuss symptoms and triggers [11, 13, 49].

Research suggests that mobile mood-monitoring can provide several benefits. Repeated mood tracking can increase self-awareness, help users identify personal triggers and early warning signs, and support communication between patients and clinicians [11, 43]. In a mixed-methods study with young people, Dubad et al. [13] found that use of a mood-monitoring app significantly reduced momentary negative mood and retrospectively assessed impulsivity, and that the app aided communication, promoted empowerment, and alleviated memory difficulties in clinical appointments. At the same time, mood-tracking is generally considered a low-threshold intervention that can be integrated into everyday life with relatively little effort [11, 43].

Despite these benefits, existing mood-tracking apps also show important limitations. Many tools put most of the emphasis on data collection and basic visualisation, but offer little help in interpreting the recorded information or translating it into concrete actions [11, 43, 49]. Schueller et al. [43] identified the lack of app-facilitated recommendations as a key gap, reporting that participants wanted their tracked data to be

2. Related Work

translated into “solutions or recommendations.” Similarly, Caldeira et al. [11] found that only 7 of 32 analysed apps provided features to support future actions. Prior work therefore highlights a persistent divide between tracking and reflection on the one hand and actionable, personalised support on the other.

Some approaches explore the combination of mood tracking with contextual information—such as activity, sleep, location, weather or music listening—to automatically discover correlations that might not be obvious to users. Burns et al. [10] pioneered this approach with *Mobilyze!*, a context-aware mobile intervention for depression that used machine learning to predict patients’ mood and contextual states from phone sensor data, achieving promising accuracy rates and preliminary evidence of efficacy. Polhemus et al. [36] found through a qualitative analysis of user reviews that people consistently request the ability to annotate mood entries with contextual information and to discover correlations across data streams. However, such designs are still less common than simple tracking apps, and there remains a need for systems that not only collect and visualise mood data, but also help users understand how different ual factors relate to their emotional state and what they can practically do with this knowledge [36, 43].

Beyond research prototypes, a wide range of commercial mood-tracking apps illustrates these design patterns in practice. Diary-style apps typically focus on very low-effort input (selecting a mood and a few activities) combined with calendar or “year in pixels” visualisations and simple statistics over time [11]. Some apps allow exporting data or sharing summaries with clinicians, which can help structure conversations in therapy [11, 43], but they still rarely provide explicit, personalised recommendations based on the collected mood and information.

2.2. Privacy of mental-health and mood-tracking apps

Mental health and mood-tracking apps process highly sensitive information—including symptoms of depression and anxiety, personal thoughts, and daily emotional states—which makes their privacy properties particularly critical [24, 25]. In an empirical privacy analysis of 27 top-ranked mental health apps from the Google Play Store, Iwaya et al. [25] found a range of data protection problems, such as unnecessary permissions, insecure cryptography, and leaks of personal data. Their study also revealed that many apps enable user profiling because they lack robust mechanisms against linkability, detectability and identifiability of users across different data sources [12, 25].

Several investigations of privacy policies and data flows show that mental-health apps frequently share data with third parties, including analytics and advertising providers, often in ways that are not clearly communicated to users. Huckvale et al. [22] assessed 36 apps for depression and smoking cessation and found that 92% transmitted data to third parties—while only 69% even had a privacy policy. Parker et al. [35] similarly reported that nearly half of the 61 prominent mental-health apps they reviewed lacked a privacy policy entirely. Reviews by independent organisations have repeatedly confirmed these findings: the Mozilla Foundation [33] reported that 20 of 32 reviewed apps earned their “Privacy Not Included” warning label, Consumer Reports [19] found that apps routinely shared unique device identifiers with third parties including Facebook, and a Brookings Institution analysis [34] called for stronger regulatory frameworks. For mood-tracking apps in particular, this creates a risk that intimate details about users’ emotional states, relationships or diagnoses could be combined with other data and used for profiling or targeted advertising [25, 35].

Even the mere fact that someone uses a mood-tracking app can implicitly reveal that they might be experiencing mental-health problems. As Iwaya et al. [25] point out, apps’ development practices do not provide foolproof mechanisms against detectability, which, in combination with opaque data flows and extensive third-party tracking, raises significant concerns about the confidentiality and potential misuse of users’ mental-health information [12, 33].

2.3. Data protection and GDPR for mental-health apps

In the European Union, data about mental health and mood qualify as special-category personal data under the GDPR, which triggers stricter requirements for their processing [24, 38]. Ienca and Malgieri [24] introduced the notion of “mental data”—defined as any data that can be organised and processed to make inferences about the mental states of a person—and analysed existing legal protections, arguing that the GDPR is an adequate tool to mitigate risks related to mental data processing but that interpreters should focus on processing characteristics rather than merely on the category of data at issue. Mood logs and information about symptoms are considered health-related data, so controllers must ensure that processing complies with specific conditions in Article 9 GDPR and is supported by a valid legal basis [24, 38]. In addition, core principles in Article 5 GDPR—such as lawfulness, fairness and transparency, purpose limitation, data minimisation and storage limitation—require that mental-health data are collected only for clearly specified purposes, not retained longer than necessary, and not used in ways that are incompatible with users’ expectations [24, 38, 48].

Several legal and policy analyses highlight that many current mental-health apps struggle to fully comply with these principles in practice [24, 39]. Common issues include vague or incomplete privacy notices that do not clearly explain purposes and recipients, unclear roles and responsibilities between app providers and third parties, and extensive sharing of data with external analytics or advertising services that is difficult to reconcile with purpose limitation and data minimisation. Huckvale et al. [22] found that the correspondence between apps’ stated privacy policies and their actual data-sharing behaviour was often poor, while Parker et al. [35] reported that many apps lacked even basic privacy disclosures. Enforcing core GDPR rights—such as access (Art. 15), rectification (Art. 16) and erasure (Art. 17)—is technically challenging when data are distributed across multiple cloud services or have been used to train machine-learning models, making it hard to remove individual users’ contributions [24, 25, 38].

For mental-health and mood-tracking apps, data protection by design and by default, as required by Article 25 GDPR, is therefore particularly important [24, 38, 48]. Recommended measures include strict data minimisation, encryption in transit and at rest, pseudonymisation of user identifiers, and limiting the transfer of raw health data to third-country cloud providers in line with Articles 44–49 GDPR [38, 45, 48]. Industry practitioners have begun exploring on-device LLM inference as a privacy-preserving alternative for mental health applications [46], though peer-reviewed evaluations of this approach remain scarce.

At the same time, the European Data Protection Board [15] has emphasised that AI-based health applications must be designed in a way that respects data subject rights, including the right to obtain information about how their data are used and the ability to have data deleted when they withdraw consent. For mood-tracking systems that use machine learning to analyse patterns and generate recommendations, this creates a tension between data-hungry models and the need to keep processing transparent, controllable and revocable for individual users [24, 25].

Under Article 9 GDPR, processing of special-category data such as mental-health information is prohibited unless specific conditions are met, for example explicit consent or processing for health-care purposes under appropriate safeguards [24, 38]. Article 5 GDPR further requires that controllers adhere to principles such as purpose limitation, data minimisation and storage limitation, which means that mental-health apps should only collect the data that are strictly necessary for clearly defined purposes and should not retain them indefinitely [25, 39, 48]. As Ienca and Malgieri [24] argue, these principles are often difficult to reconcile with extensive behavioural tracking, broad secondary uses of data and complex ecosystems of third-party services [25, 39].

These considerations suggest that, from a privacy-by-design perspective, architectures which avoid sending sensitive mood data to large external AI providers and instead keep analysis within a tightly controlled institutional environment can better support GDPR principles such as data minimisation, storage limitation, and data protection by design [24, 25, 45, 48]. For MoodScape, this motivates running the recommendation model on a university-managed server rather than in a commercial cloud.

2.4. Usability and UX evaluation in HCI

Usability and user experience are critical factors for the effectiveness of mental-health and mood-tracking apps, because users need to be able to integrate these tools into their daily routines over longer periods [36]. Many mental-health apps show high initial interest but suffer from low long-term engagement, which is often attributed to usability issues, cognitive load, or a lack of perceived usefulness [49]. As a result, studies in this area commonly supplement clinical or behavioural outcomes with standardised usability and UX questionnaires [9, 23].

The System Usability Scale (SUS) is one of the most widely used instruments to assess the perceived usability of interactive systems, including mobile health applications [8]. SUS consists of ten items with 5-point Likert responses and yields a single score between 0 and 100. Several authors have proposed benchmark ranges, adjective ratings, and letter grades to support interpretation, for example treating scores around 68 as “average” usability and scores above roughly 80 as “excellent” [20, 29, 40]. SUS has been applied in numerous mHealth and digital mental-health studies to provide a quick, comparable measure of how easy and pleasant an app is to use [9, 23].

Beyond pure usability, user experience questionnaires aim to capture both pragmatic aspects (goal-directed usefulness and usability) and hedonic aspects (stimulation, novelty, enjoyment) of interaction [42]. Hassenzahl [21] proposed the distinction between pragmatic and hedonic quality as fundamental dimensions of user experience. The User Experience Questionnaire (UEQ), developed by Laugwitz et al. [28], is a commonly used tool in HCI and mHealth research that measures user experience along six dimensions (attractiveness, perspicuity, efficiency, dependability, stimulation, novelty) using semantic differential items. Schrepp et al. [42] provide a comprehensive handbook for applying and interpreting UEQ results. Several studies have used UEQ or its variants to evaluate digital health applications, arguing that positive hedonic qualities such as stimulation and novelty can influence engagement and adherence, especially in self-management s [9, 50].

In evaluations of mental-health and mood-tracking apps, combining instruments like SUS and UEQ with domain-specific outcome measures (e.g. symptom change, self-reflection, adherence) is therefore a common strategy to obtain a more holistic picture of how users experience and adopt these tools [32]. Beyond usability and user experience, a growing body of work in HCI has identified *reflection* as a key design goal for personal informatics systems such as mood-tracking apps.

2.5. Reflectivity and technology-supported reflection

A central goal of MoodScape, as expressed in the thesis title, is to support *reflective* well-being. This section reviews how the concept of reflection has been defined and operationalised in HCI research, outlines a framework for understanding different levels of reflective engagement, and introduces the Technology-Supported Reflection Inventory (TSRI) as a validated instrument for evaluating whether an interactive system effectively supports reflection.

2.5.1. Defining reflection in HCI

Despite its prominence as a design goal, reflection lacks a single agreed-upon definition in Human-Computer Interaction. In a review of how reflection is used across HCI publications, Baumer et al. [3] found that fewer than half of the surveyed papers provided an explicit definition, and the definitions that were given varied considerably in scope. Most researchers, however, draw on two complementary traditions.

The first tradition originates in Schön’s concept of the *reflective practitioner* [41]. Schön distinguishes between *reflection-in-action*, which occurs during an ongoing experience as the individual notices unexpected outcomes and adjusts their approach in real time, and *reflection-on-action*, a retrospective process of

consciously reviewing and analysing a past experience to draw lessons from it. While reflection-in-action is spontaneous and embedded in the flow of activity, reflection-on-action is deliberate and typically happens after the event, for example when a user reviews their mood logs at the end of a week and considers what patterns emerge.

The second tradition emphasises reflection as a learning process that moves through stages of increasing depth. Dewey, Moon, and Mezirow each conceptualise reflection as a hindsight process that begins with noticing a puzzling or unexpected experience and progresses, through conscious examination, towards a change in understanding or behaviour. Drawing on these perspectives, Baumer [2] synthesised three dimensions of reflection relevant to technology design: *breakdown* (encountering doubt or puzzlement that triggers reflection), *inquiry* (consciously reviewing and examining the experience), and *transformation* (changing one's assumptions or behaviour as a result).

2.5.2. Levels of reflection

Building on these theoretical roots, Fleck and Fitzpatrick [18] proposed a framework of five consecutive levels that describe a spectrum of reflective thought, from no reflection to critical reflection:

- R0 — Description.** A description or statement about events without further elaboration or explanation. At this level, a user merely revisits what happened (e.g., looking at a list of logged moods) without interpreting the data. Fleck and Fitzpatrick consider this level pre-reflective.
- R1 — Descriptive reflection.** Description that includes justification or reasons for action, but in a reportive way. The user begins to articulate explanations (e.g., “I was in a bad mood because I slept poorly”), but no alternative explanations are explored and the analysis remains surface-level.
- R2 — Dialogic reflection.** Looking for relationships between pieces of experience or knowledge. The user considers different explanations, examines correlations between data points, and engages in cycles of interpreting and questioning (e.g., “My mood tends to be better on days with more steps and social interaction—could exercise or the social contact be the driver?”).
- R3 — Transformative reflection.** Revisiting an event or knowledge with the intent to reorganise one's perspective or do something differently. The user moves from recognising patterns to actively planning change (e.g., “I will try to walk more on workdays to see if it helps my mood”).
- R4 — Critical reflection.** Reflection that considers social and ethical issues beyond the immediate personal , engaging with the broader picture.

A key insight from this framework is the qualitative gap between lower levels of reflection (R0–R2), which are situated within the displayed data and the user's immediate , and higher levels (R3–R4), which require the user to engage with perspectives that transcend the data at hand [18]. For personal informatics systems such as mood-tracking apps, this implies that simply providing data visualisations may support R0–R2 (revisiting and exploring data), but that meaningful behavioural change (R3) typically requires additional scaffolding, such as guided prompts, goal-setting features, or personalised recommendations that help users translate observed patterns into concrete actions.

2.5.3. Reflection in personal informatics

The concept of reflection is also central to models of personal informatics use. Li et al. [30] proposed a five-stage model in which reflection follows data collection and integration and precedes action. Epstein et al. [14] extended this model to a Lived Informatics framework, emphasising that reflection does not occur automatically when users are given access to their data but must be actively encouraged by the system's design. Research on fitness trackers and health-monitoring apps has similarly found that while users develop

2. Related Work

tracking habits relatively quickly, meaningful reflection on the tracked data—and especially behavioural changes based on that reflection—requires deliberate support and sustained engagement over time [11, 43].

These findings are directly relevant to mood-tracking apps like MoodScape. While logging daily moods and viewing trend visualisations may support lower levels of reflection (revisiting and data exploration), the gap between exploring data and acting on insights remains a known challenge. Features such as -aware correlation summaries and personalised recommendations can be understood as attempts to scaffold the transition from data exploration (R2) towards transformative reflection (R3).

2.5.4. The Technology-Supported Reflection Inventory (TSRI)

A persistent challenge in designing for reflection has been the lack of validated instruments to measure whether a system actually supports reflective engagement. To address this, Bentvelzen et al. [5] developed the Technology-Supported Reflection Inventory (TSRI), a nine-item questionnaire that evaluates how effectively an interactive system supports reflection. The scale was developed in a multi-stage process: an initial pool of items was derived from theoretical reflection literature reviewed, refined through expert review, and then reduced to nine items via exploratory factor analysis ($n = 176$). The instrument was subsequently validated through confirmatory factor analysis ($n = 56$) and test-retest reliability assessment ($n = 20$), yielding acceptable psychometric properties (overall Cronbach's $\alpha = 0.78$; $TLI = 0.97$, $RMSEA = 0.04$).

The nine items are scored on a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree) and load onto three factors:

- **Insight** ($\alpha = 0.74$): Whether the system prompted personal change, such as serving as a “wake-up call,” leading to changes in how users approach things, or providing ideas for overcoming challenges (items 1–3).
- **Exploration** ($\alpha = 0.72$): Whether the system supported engagement with one's own data, including enjoyment of data exploration, ease of getting an overview, and ease of reviewing long-term data (items 4–6).
- **Comparison** ($\alpha = 0.73$): Whether the system facilitated social reflection, such as reflecting on data with others, discussing data with others, or thinking about how one's data relates to that of other people (items 7–9).

The TSRI is designed as an artefact-centric instrument for relative comparison of designs within the same context, rather than as an absolute measure of an individual's reflective disposition [5]. It has since been adopted by other researchers evaluating reflection-support tools; for example, Jörke et al. [26] used it in their work on Pearl, a technology probe for machine-assisted reflection on personal data. Its three-factor structure maps naturally onto different aspects of reflective engagement: Insight captures higher-level reflection (aligning broadly with R3 in Fleck and Fitzpatrick's framework), Exploration captures data-driven sense-making (R1–R2), and Comparison captures the social dimension of reflection that can deepen individual sense-making through dialogue with others.

For the evaluation of MoodScape, the nine TSRI items were included in the end survey to assess whether the system supported reflection across these three dimensions. This allows the findings to be interpreted not only in terms of overall app satisfaction and usability, but also in relation to a validated, theoretically grounded framework for technology-supported reflection—directly connecting the study outcomes to the reflective well-being goal stated in the thesis title.

3. App Design & Implementation

This chapter describes the design and implementation of MoodScape. The Android frontend was adapted from an existing prototype developed in a previous Master’s thesis [4], with changes to integrate the new recommendation logic and the Ollama-based backend. The backend and the tailored recommendation pipeline were implemented from scratch as part of this Bachelor’s thesis.

3.1. System Architecture

3.1.1. Overview

The original MoodScape prototype used a React Native frontend and a Node.js backend with a cloud-based GPT-3.5 API [4]. In this work, the system was reimplemented with an Android-only JavaScript and Kotlin frontend and a new Python/FastAPI backend that runs a local LLM on a university server to avoid sending mood data to commercial cloud providers.

MoodScape follows a client–server architecture (Figure 3.1). The Android app runs on the user’s smartphone and communicates with a university-managed backend server via FastAPI. The server stores pseudonymised mood entries and contextual data in a database, integrates external data sources (weather, music, health), and runs a locally hosted AI model that generates recommendations.

A fully on-device LLM on the smartphone could be ideal from a privacy perspective, but is currently too resource-intensive for typical end-user devices. As a compromise, MoodScape uses a locally hosted Ollama model (llama3.2:1b¹) on a university server to keep data out of commercial cloud infrastructures and retain institutional control over storage and deletion. This design avoids sending raw mood and context data to providers such as OpenAI and supports the GDPR principles of data minimisation, purpose limitation and data protection by design [24, 38, 48].

The backend and the Ollama service run on the same virtual server, provisioned with 8 GB of RAM and 8 virtual CPU cores (QEMU Virtual CPU, version 2.5+). Compared to cloud-based models with access to large GPU clusters, this environment is relatively constrained, which limits the size and complexity of the models that can be used; the 1-billion-parameter model was chosen accordingly. The design therefore explicitly trades some raw model capacity for increased privacy and data control—an aspect that is important when interpreting the quality of the generated recommendations.

¹<https://ollama.com/library/llama3.2:1b> Last accessed 10.03.26

3. App Design & Implementation

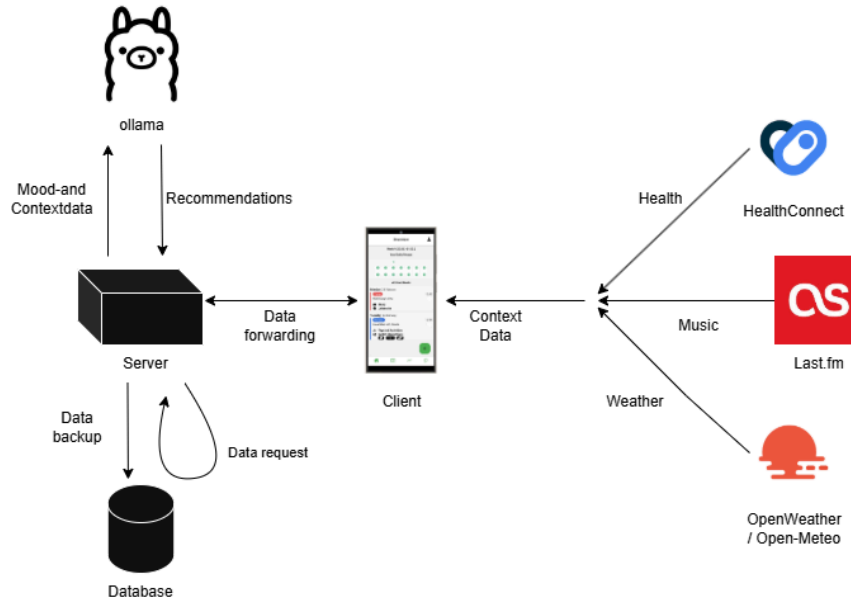


Figure 3.1.: Client–server architecture of *MoodScape*.

MoodScape collects several types of context data on the client side and stores them in the form of mood entries and daily recaps in the backend database. The Android client queries external APIs when users create mood or recap entries: OpenWeather and Open-Meteo for weather conditions, Last.fm for recently played music (including aggregated genres and artists), and the Health Connect API for health and activity data. OpenWeather is used to retrieve current weather for mood entries and same-day recaps, whereas Open-Meteo is used for recaps referring to the previous day because it offers free, keyless access to recent historical weather data, while historical data in OpenWeather’s free tier are more restricted². Last.fm was selected instead of the Spotify Web API because, during the development period, Spotify tightened its developer mode access and test-user limits for the Web API, which made it impractical to support all study participants with a research app, whereas Last.fm’s API remained accessible for small projects and can aggregate listening data from multiple music apps, not only Spotify³. Health Connect was chosen because it provides a unified interface for accessing health and fitness data from multiple apps on Android, allowing MoodScape to obtain daily step counts and sleep duration via a single, permission-managed integration⁴. Access to all three context sources (location for weather, Last.fm, and Health Connect) was optional; if a participant did not grant location access, connect a service, or approve Health Connect permissions, the app skipped the corresponding API calls and the affected metrics were not included in the the model input.

Table 3.1 summarises the context sources and the fields stored in the database.

²<https://openweathermap.org/price>, <https://open-meteo.com> Last accessed 10.03.26

³<https://developer.spotify.com/blog/2026-02-06-update-on-developer-access-and-platform-security> Last accessed 10.03.26

⁴<https://play.google.com/store/apps/details?id=com.google.android.apps.healthdata> Last accessed 10.03.26

Table 3.1.: Fields stored for mood entries and daily recaps in MoodScape

Field	Mood entries	Daily recaps
Date	Timestamp of the momentary mood entry.	Timestamp of the recap day.
mood	Mood label describing the user’s current emotional state at the moment of the entry.	Mood label for the day as a whole.
Notes	Optional free-text note describing the specific situation.	Optional free-text note summarising the day.
InfluenceContext	Self-observed factors that influenced the current situation.	Self-observed factors that influenced the day overall.
socialLoad	–	Numeric indicator of perceived social load for the day (0–100).
Type	Type of social interaction (e.g. <i>In Person</i>) for this situation.	–
Amount	Size of the interaction (e.g. <i>1 to 1</i> , small group).	–
Person	List of interaction partners (e.g. <i>Partner</i> , <i>Friend(s)</i>).	–
StartDate	Start time of the reported situation.	–
EndDate	End time of the reported situation.	–
weather	Current weather at the time and place of the mood entry (location, temperature, humidity, pressure, wind speed, brief description).	Weather summary for the recorded day (location, afternoon temperature and daily min/max, humidity, pressure, wind speed, brief description).
health	Aggregated daily health data from Health Connect: steps taken, walking distance, sleep duration for the previous night, and estimated calories burned.	
music	Aggregated listening data from Last.fm: list of recently played tracks, total listening time, most listened genres, and most listened artists for the selected day.	

3.1.2. AI Recommendation Pipeline

For recommendation generation, the server prepares a compact representation of the user’s recent mood and context data and sends it as a prompt to the local Ollama instance (llama3.2:1b). Under this implementation, no raw data leaves the client–server architecture; the LLM is only accessible from within the backend environment.

To keep prompts within the model’s practical limits and reduce hallucinations, the payload is trimmed and pre-processed before being sent to Ollama. Dates are removed because the model has no access to the current date and tended to become temporally confused when confronted with specific timestamps. Less central context features such as humidity and air pressure are omitted to shorten the prompt, while health data are reduced to daily step count and sleep duration as coarse indicators of physical activity and rest. For music, the backend selects up to three genres and two artists per day to preserve a sense of personalisation without overloading the model.

In initial tests, sending full, unfiltered mood and context histories resulted in frequent invalid outputs or hallucinations, even with more restrictive prompts. By contrast, the trimmed representation led to more stable and interpretable responses. This reflects a general trade-off between prompt richness and output reliability that is particularly relevant under the given hardware constraints.

3. App Design & Implementation

For the A/B study, MoodScape implements two recommendation conditions: a baseline condition with generic mood-support recommendations and a tailored condition that incorporates the user's recent mood and context data. Both conditions use the same locally hosted Ollama model (llama3.2:1b).

Baseline Recommendation Prompt

In the baseline condition, MoodScape uses a static wellbeing prompt that does not include any user-specific mood or context data. The backend passes three requested recommendation categories (for example, *Social*, *Sport*, *Music*) and a fixed list of example suggestions to the model (see Appendix D.6). The model is instructed to generate exactly one concise but actionable suggestion per requested category, using the examples only as inspiration and without copying text verbatim. The example suggestions were reused from the previous Master's thesis [4].

```
BASELINE_PROMPT = """
You are a wellbeing recommendation generator.

You will receive exactly three requested categories.
Generate exactly one explanation per category.

Rules:
- Do not invent new categories.
- suggestion["type"] must correspond to the provided categories.
- Use example_suggestions only as inspiration.
- Do not copy example text verbatim.
- Be concise but actionable.
"""
```

Listing 3.1: Baseline recommendation system prompt (excerpt)

Alongside this system prompt, the backend sends a fixed list of generic example activities (for instance, going for a walk, meeting friends, or doing a short workout) covering categories such as social contact, physical activity, time in nature, music, meditation, and digital detox. The model is free to combine and adapt these ideas, but it does not see any of the user's mood entries or context data in the baseline condition. Like the tailored condition described below, the baseline prompt also instructs the model to return its output in a fixed JSON schema containing an introduction and three numbered suggestions, each with a type and an explanation field (see Appendix D.2). Using the same output format for both conditions ensures that the backend can parse and display the results through a single code path, and that visual presentation on the recommendation screen is unified across conditions.

Tailored Recommendation Prompt

In the tailored condition, the model receives a compact, preprocessed summary of the user's recent mood entries and daily recaps, including the previously discussed contextual signals such as health metrics, weather, music listening history, and social activity. Metrics are only included in the payload if the corresponding data source is available (for example, Health Connect permissions have been granted and Last.fm username was set) and the value is non-zero; otherwise, the respective fields are omitted. The system prompt frames the task as analysing mood and context data to identify meaningful patterns and to provide personalised, empathetic, and actionable suggestions, while enforcing strict constraints on tone, uncertainty, and response structure. It instructs the model to avoid causal claims, to acknowledge when data are limited, and to address the person directly using "you" instead of referring to "the user". In addition, the prompt describes the available data fields in detail so that the model knows which aspects (for example, mood labels, self-observed influences, steps, sleep, genres, artists, or social load) it can draw on when formulating recommendations.

```

SYSTEM_PROMPT_V8_CONSTRAINTS = """
You are a mood analysis assistant. Your task is to analyze mood and context data
to find meaningful patterns and provide personalized, empathetic and actionable suggestions.
When data is limited do NOT make up data, but suggest general wellbeing suggestions instead.

ANALYSIS TASK (match the provided JSON structure):
Examine the provided mood entries and recaps. Use the information that exists in the fields.

Available fields you may see:
- mood: the reported mood label
- context: Influences such as Weather, Physical Activities, Music, Social Interactions, Hobbys
- social: details about a social interaction {type, amount, person}
- socialLoad: degree of social interaction
- metrics: numeric/context values for health{steps, sleep} and weather{temp_c, weather}
- music: listening context {genres, artists}

CAUSALITY AND CAUTION:
- Never claim cause-effect. Use wording like "might", "could", "worth trying".
- If there are fewer than 5 total entries, explicitly say the data is limited and avoid strong
conclusions.
Only draw conclusions that are directly supported by the provided fields.

TONE AND PERSPECTIVE RULES:
- Address the person directly using "you".
- Do NOT refer to "the user".
- Do NOT describe the data from an outside perspective.
- Speak as if you are giving personal feedback directly to the person.
- Use supportive and empathetic language.
- Do NOT EVER try to chat with the user
- Do NOT refer to "today" or "this day", instead refer to "recent entries"

Rules:
1. Pick the 3 MOST RELEVANT categories supported by the user's data
2. Each explanation MUST be 2-4 sentences - do not leave any explanation at just 1 sentence
3. Match categories to actual observed data patterns
4. If data is limited, acknowledge this and provide general wellness suggestions.
5. If fewer than 3 mood entries are provided, explicitly state that
patterns are not yet reliable and provide general wellness suggestions
without claiming correlations

Before writing your response, identify the single most notable
pattern specific to the user. Reference it explicitly by name
in your introduction and build at least one suggestion directly
around it.

Analyze the mood data below and provide your response:
"""

```

Listing 3.2: Tailored recommendation system prompt (excerpt)

Compared to the baseline condition, the tailored prompt imposes several additional constraints. First, it requires the model to select three recommendation categories that are actually supported by the user's recent

3. App Design & Implementation

data—rather than freely choosing generic topics—and to build at least one suggestion around the most notable detected pattern. Second, it enforces strict tone and length rules: each suggestion must be accompanied by an explanation of two to four sentences that uses cautious, non-causal language such as “might” or “could”, and the model must address the user directly rather than describing data from an outside perspective. Although both conditions share the same JSON output schema (as described above), the tailored prompt places a higher structural and cognitive load on the model because it must simultaneously analyse user data, select relevant categories, and produce data-grounded explanations—all within the same small local model (llama3.2:1b) running on a constrained CPU-only server. This asymmetry in prompt complexity is important to keep in mind when interpreting differences in perceived recommendation quality between conditions.

This prompt design draws on general recommendations for structuring prompts for large language models, such as assigning a clear role, specifying behavioural constraints, and requesting an explicit, machine-readable output format [44]. The concrete wording and level of detail were iteratively refined based on pilot runs with the local llama3.2:1b model until the outputs were syntactically valid and interpretable under the given hardware constraints. Recent work also suggests that overly long or weakly structured prompts can degrade output quality and increase hallucinations, especially for smaller models with limited compute [1]. The final prompt was therefore designed to explicitly define the model’s role, enforce cautious language and data-driven reasoning, and require a strict JSON output schema, with the aim of allowing the backend to reliably parse the responses and mitigating typical failure modes of local LLM deployments [1, 44].

3.2. Frontend

The Android frontend of MoodScape was implemented in Kotlin and JavaScript, reusing the general navigation structure of the original React Native prototype but introducing new layouts and flows for this thesis [4]. The app is organised into sections for the overview, recap analytics, mood analytics, and recommendations, accessible via a tab-based navigation bar at the bottom of the screen.

3.2.1. Login and permissions

When opening MoodScape for the first time, participants were presented with the lightweight login screen in Figure 3.2(a), where they only had to enter their assigned participant code. No email address or personally identifying information was collected. After logging in, participants were shown the permissions screen in Figure 3.2(b), which explained the optional context integrations used for tailored recommendations. They could choose to (1) connect a Last.fm account to share aggregated listening history by entering their Last.fm username, (2) grant location access to attach local weather conditions to entries, and (3) grant Health Connect read access for health and fitness data. All options were opt-in; participants could proceed without enabling any of them, in which case MoodScape only stored the self-submitted mood inputs and omitted the corresponding context metrics from the model input.

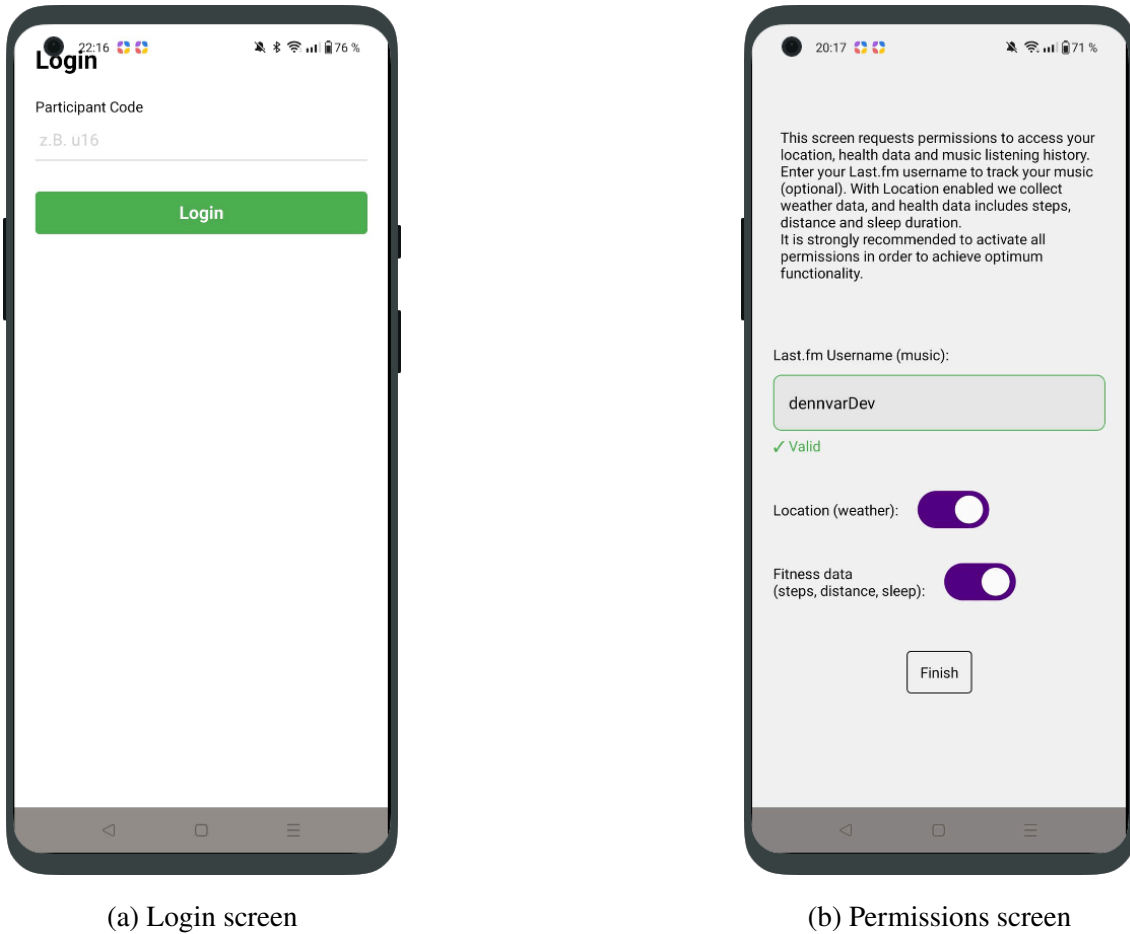


Figure 3.2.: Login- and Permission screen

3.2.2. Overview and Navigation

The overview screen provides a central entry point to the app, showing all mood entries of the current week in a scrollable list and the daily recaps of the current week above, with the current day explicitly highlighted. The daily recap section displays two green dots arranged vertically when nothing has been logged for a given day. When a recap has been logged, the top dot is filled with the mood icon corresponding to the logged mood, and the bottom dot is replaced by one influencing factor in green along with a “+n” indicator showing how many additional factors were selected. If social interactions were selected as an influencing factor, the bottom dot is instead replaced by a visual representation of the social load slider value, making it possible to see at a glance whether the amount of social contact felt too low, appropriate, or overwhelming. In the top-right corner of the overview screen, an account icon opens an account screen where participants can see their participant code, reopen the permissions screen in Figure 3.2(b) to grant context permissions (Last.fm, location, and Health Connect) after the initial onboarding, and, if desired, log out of the app. This keeps privacy-related settings accessible without cluttering the main tracking interface.

Users can log new moods or recaps via a “+” button in the bottom right corner. Tapping a mood entry in the overview opens the mood detail view (Figure 3.3(b)), where users can review the context data specified in Table 3.1, except for Type, Amount, Person, and Start/End Date, which are only shown as icons on the saved mood card. The mood with its corresponding colour, the influencing factors, the notes, and—for weather—the temperature and weather description are displayed on this card.

3. App Design & Implementation

Tapping a day in the recap section opens the recap detail view (Figure D.1(c)). This view shows the recap mood, influencing factors, and notes, and it additionally includes the social load indicator, which is only available for recaps and summarises the perceived intensity of social interactions for that day.

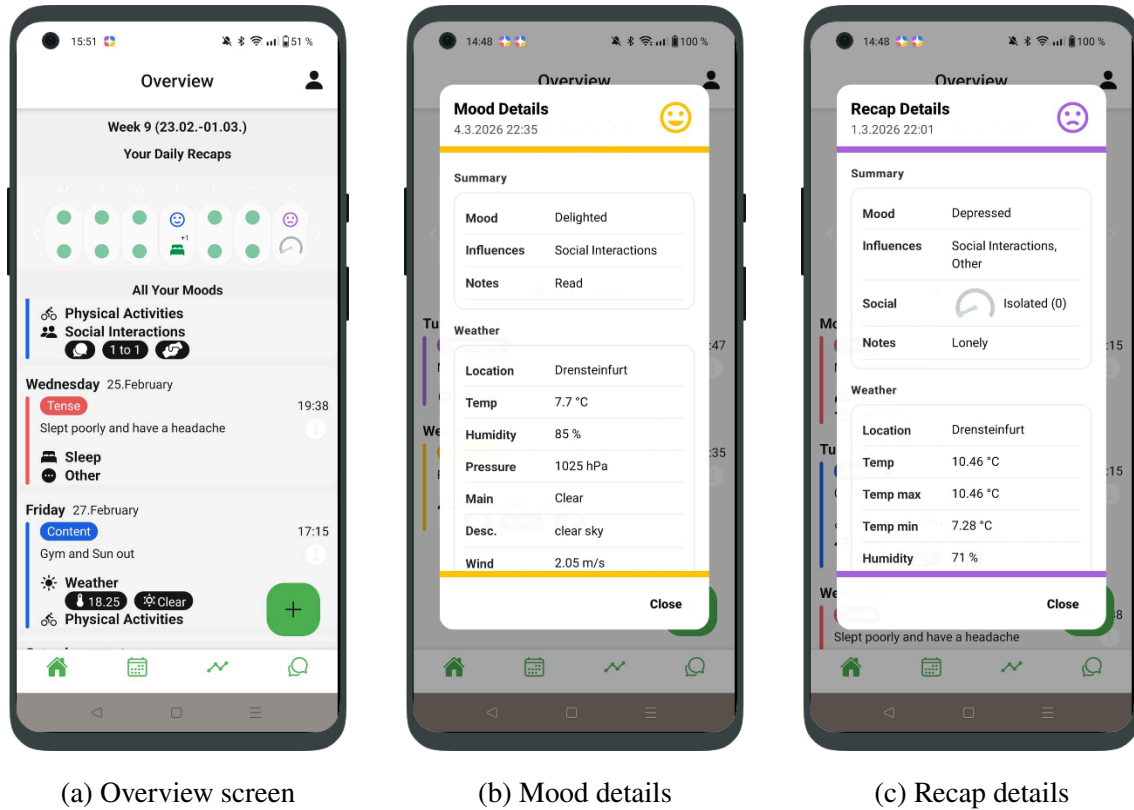
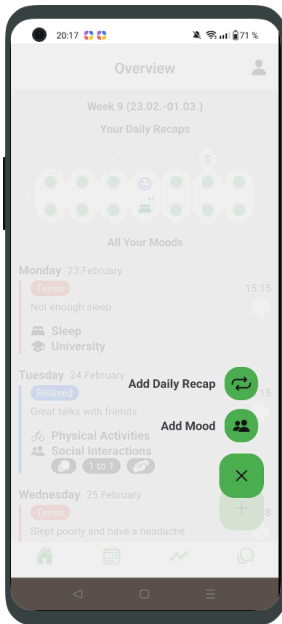


Figure 3.3.: Overview screen and details views in *MoodScope*.

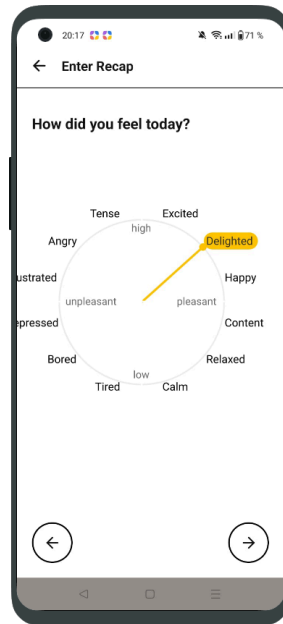
The tab bar at the bottom allows users to switch between the overview, mood entry list, daily recaps, the two analytics views, and the recommendation screen. This navigation structure follows the original prototype but was visually redesigned [4].

3.2.3. Entry Flow

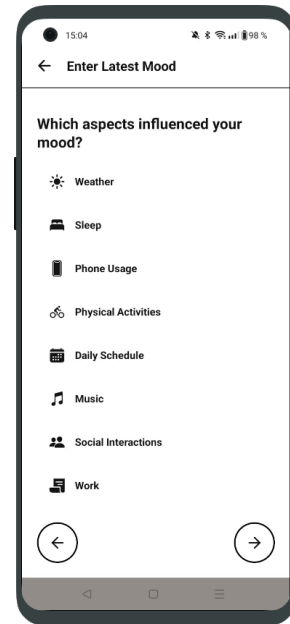
MoodScope offers two types of entries: *mood entries* for capturing a momentary emotional state, and *daily recaps* for reflecting on an entire day. Both share the same multi-step entry flow (Figure 3.4) but differ in certain fields to match their respective purpose.



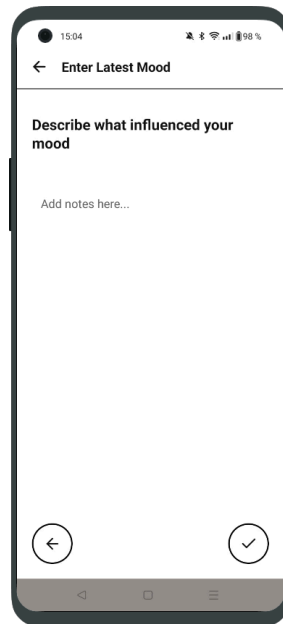
(a) Entry type selection



(b) Selecting mood



(c) Context & influences



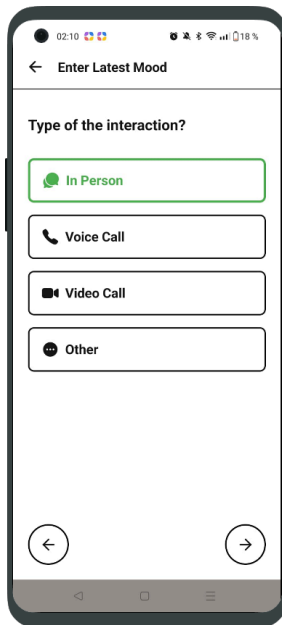
(d) Notes

Figure 3.4.: Multi-step entry flow. Users first select the entry type and mood, then specify influencing factors and context, and optionally add notes. Screen (c) is scrollable.

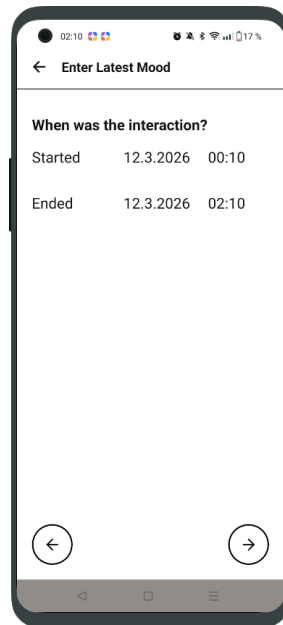
The flow begins with a type selection screen (a), where users choose between a mood entry and a daily recap, and then select a mood on the valence–arousal circumplex shown in Figure 3.4(b), which is discussed in more detail in [4]. The circumplex lights up in the colour corresponding to the hovered-on mood. Next, users can select one or more self-observed influences (for example, social interactions, sleep, or university; Figure 3.4(c)). When creating a momentary mood entry and Social Interactions is selected as an influence,

3. App Design & Implementation

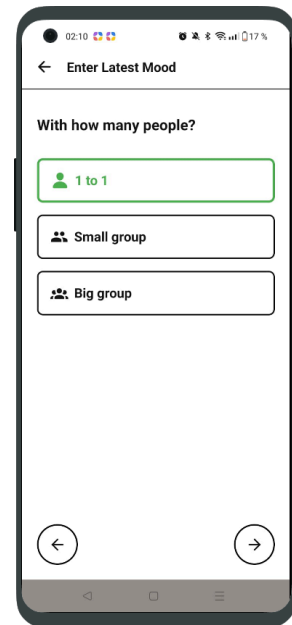
the app presents additional screens (Figure 3.5(e)–(h)) that capture structured details about the interaction type, the number of interaction partners, and the relationship to the other person or people.



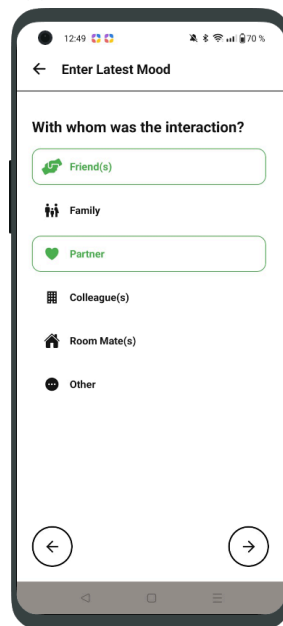
(e) Type of interaction



(f) Time of interaction



(g) Amount of people



(h) Relationship with interaction partner

Figure 3.5.: Additional screens shown when Social Interaction is selected as an influence, in order of appearance.

Daily recaps complement the mood entries by allowing users to reflect on an entire day rather than a single

situation. In addition to the shared fields, recaps include a social load slider that captures the perceived intensity of social interactions over the course of the day (Figure 3.6(a)). When creating a recap, users specify whether it refers to today or the previous day; for recaps marked as “yesterday”, the app retrieves weather and context data for the previous day via Open-Meteo, whereas “today” recaps use the current-day snapshot (Figure 3.6(b)). In the background, the app attaches snapshots of weather, health, and music context to each entry before sending it to the backend together with the self-logged data.

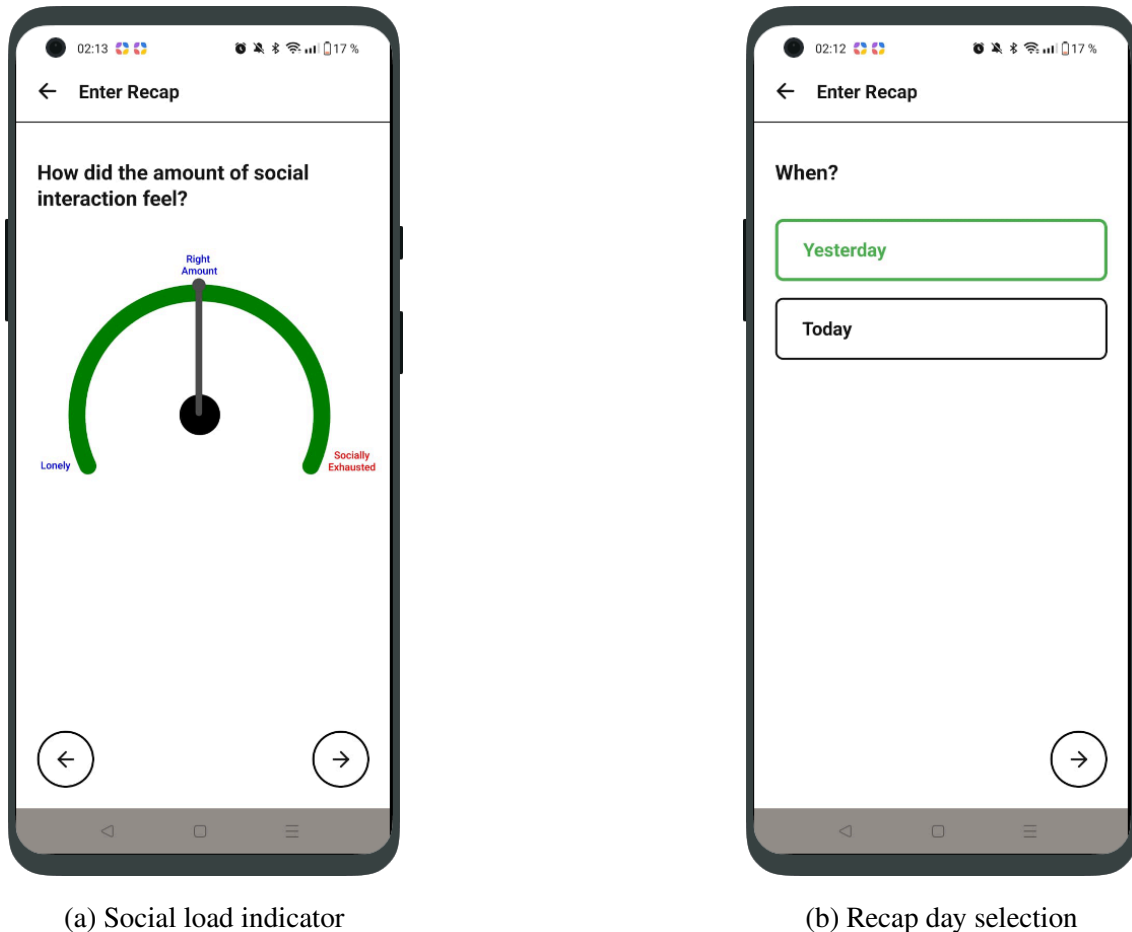


Figure 3.6.: Screens exclusive to daily recaps.

3.2.4. Analytics Views

MoodScape provides separate analytics views for mood entries and daily recaps.

The Mood Analytics screen (Figure 3.7) presents the *Mood Bar* and *Mood History* for the month shown at the top of the screen. Arrows to the left and right of the date label allow users to navigate between months. The Mood Bar shows the overall distribution of moods in the selected month, both as absolute counts next to each icon and as percentages of the total number of entries. Below, the Mood History displays the distribution of moods across individual days, allowing users to see which moods were tracked on which day and to recognise longer-term trends.

3. App Design & Implementation

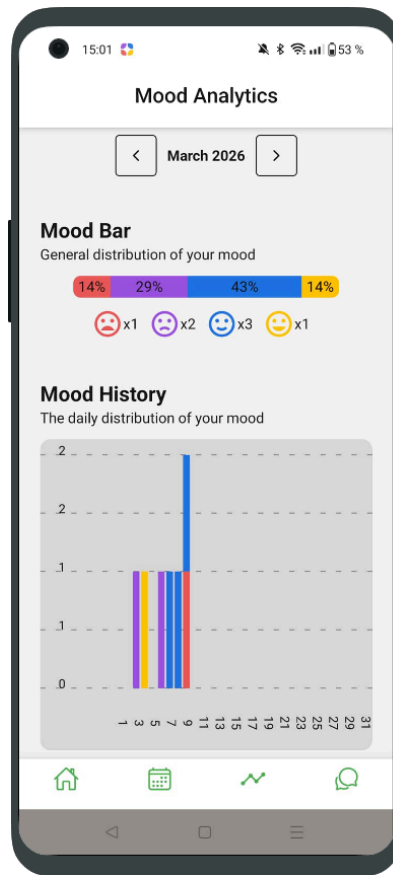
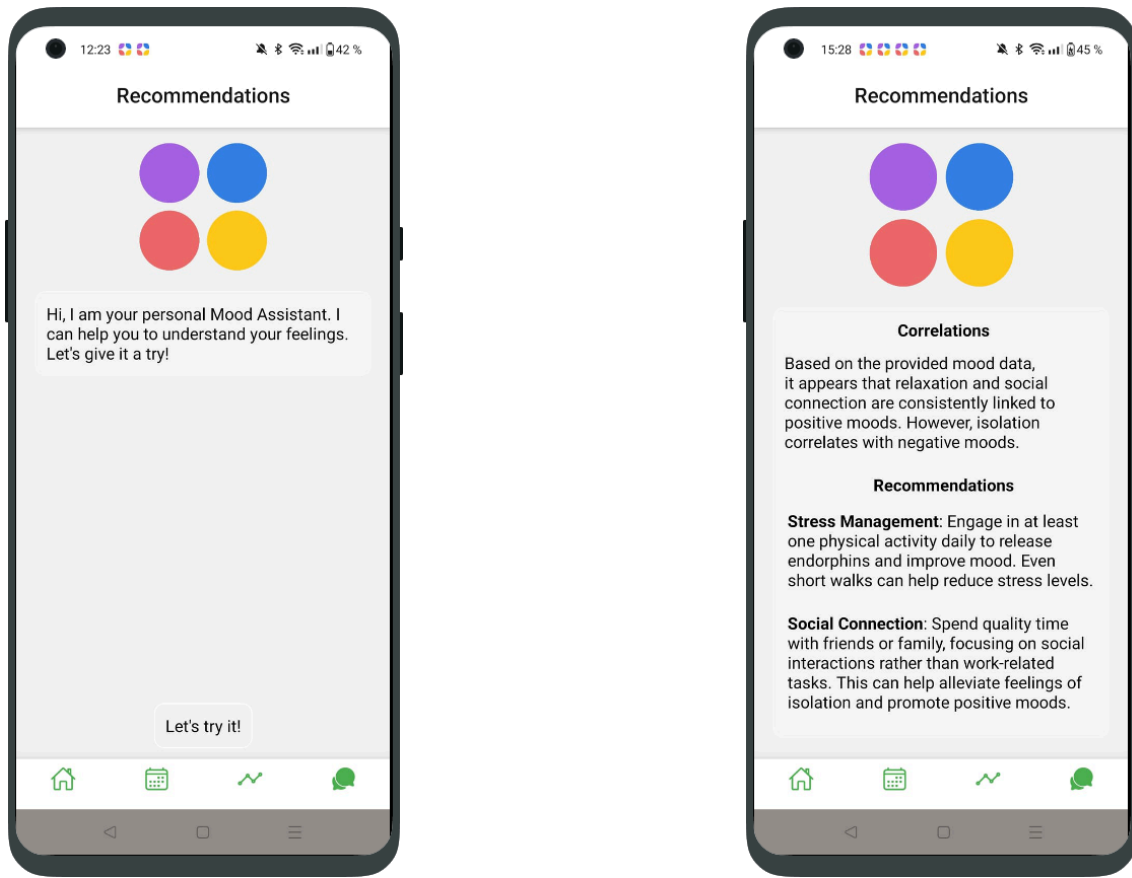


Figure 3.7.: Mood analytics view showing monthly trends and distributions of mood labels.

The Recap Analytics screen (Figure 3.8) displays the current month at the top and visualises daily recaps in a calendar view, with days annotated by the corresponding mood icon. Users can navigate to previous months using arrows next to the month label. Below the calendar, a *Recap History* line chart shows the distribution of recap moods over the month, with days on the x-axis and the four recap mood categories on the y-axis. This makes it possible to see intuitively when the trend moves towards more negative moods (downwards) or more positive moods (upwards), following the design of the original prototype [4].

3. App Design & Implementation



(a) Recommendation introduction

(b) Recommendation list

Figure 3.9.: Recommendation screen in MoodScape (tailored condition shown).

Users can open the recommendation screen from notifications, that are sent when the App is opened, or from the overview to request a new recommendation and scroll through the list of suggestions. Additional screenshots, including an example baseline recommendation screen, are shown in Appendix D.7.

3.3. Backend

The backend of MoodScape was newly implemented in Python using FastAPI and provides all REST endpoints consumed by the Android app. It is deployed on a university-managed server using Uvicorn as the ASGI server, manages the SQLite database via SQLAlchemy, stores mood entries and daily recaps as JSON payloads, and orchestrates communication with the locally hosted Ollama model that generates both baseline and tailored recommendations.

3.3.1. Technology Stack and Structure

The project follows a modular structure with separate modules for database access, data models, and routing :

- `main.py` initialises the FastAPI application, configures CORS, creates the database tables, and registers the routers for patient management and recommendation generation. During startup, it sends a small warm-up request to the Ollama server so that the first real recommendation request does not suffer from excessive cold-start latency.
- `database.py` defines the SQLite connection (`app.db`) and exposes a dependency that yields SQLAlchemy sessions to the route handlers.
- `models.py` and `schemas.py` contain the SQLAlchemy models (for example `User`, `Entry`, `Recommendation`) and the corresponding Pydantic request and response schemas.
- `routers/patient.py` groups all endpoints related to user management, permissions, surveys, and mood/recap entries.
- `routers/recommendation.py` implements the endpoints that call the local Ollama instance and post-process its outputs for both the baseline and tailored recommendation conditions.

CORS is configured to allow the Android client to access the API and send authentication headers, while all services remain within the university-managed infrastructure.

3.3.2. User and Entry Management

The patient router implements the basic application logic for study participants. On the first login, the app sends a pseudonymous participant code to `/patient/login`. The backend either looks up the existing user or creates a new one, assigns them to one of the two crossover groups, which determines the order in which the participant experiences the baseline and tailored conditions across the two study weeks, and generates a random access token that must be included in subsequent requests via the `Authorization` header. A shared helper function `get_current_user` validates this token for all other endpoints and returns a 401 error if it is missing or invalid.

Mood entries and recaps are stored in a generic `Entry` table. Endpoints such as `/patient/addEntry` and `/patient/recapData` receive structured JSON objects from the frontend and persist them in the database, marking recaps via a Boolean flag. Retrieval endpoints (`/patient/getAllMoodEntries`, `/patient/getAllRecaps`) read all entries for the current user, sort them chronologically, and return the original JSON enriched with a database identifier. Finally, `/patient/addRecommendation` is used to log both the input payload and the model output for each generated recommendation together with metadata such as generation time and condition (see Appendix D.4).

Table 3.2 summarises the most important backend endpoints.

Table 3.2.: Core backend endpoints for user and entry management

Endpoint	Method	Description
<code>/patient/login</code>	POST	Login via participant code, user creation and condition assignment
<code>/patient/addPermissions</code>	POST	Store granted permissions as JSON for the current user
<code>/patient/addEntry</code>	POST	Store a new mood entry for the current user
<code>/patient/getAllMoodEntries</code>	GET	Fetch all mood entries in chronological order
<code>/patient/recapData</code>	POST	Store daily recap entries
<code>/patient/getAllRecaps</code>	GET	Retrieve all recap entries
<code>/patient/addRecommendation</code>	POST	Log recommendation input and output for analysis

3. App Design & Implementation

3.3.3. Recommendation Endpoints and Ollama Integration

The recommendation router encapsulates the interaction with the local Ollama server and exposes two main endpoints:

- `/patient/recommendation` for tailored recommendations based on the user's recent mood and recap history.
- `/patient/baselineRecommendation` for generic wellbeing suggestions that do not use individual mood data.

For the tailored endpoint, the frontend sends a JSON-encoded string containing arrays of mood entries and recaps. The backend first parses this string and converts any nested JSON strings into Python dictionaries. It then calls a helper function `clean_entries_for_ollama`, which reduces the payload to a compact, LLM-friendly representation: only the ten most recent mood entries and ten most recent recaps are kept, and each entry is filtered down to essential fields such as mood label, self-observed influences, a truncated free-text note, optional social information, and aggregated metrics for steps, sleep duration, temperature, and weather description. Music information is also reduced to a small set of frequently listened genres and artists. The result is a concise JSON object that is appended to the tailored system prompt and sent to the local `Llama3.2:1b` model via Ollama's HTTP API (see Appendix D.3 for a full example payload). In addition to the prompt-level instructions, the Ollama API call uses the `format` parameter to enforce JSON output at the inference level, reducing the likelihood of malformed responses. This JSON schema for the model requires an introduction string and exactly three suggestions, each with a `type` and an `explanation` field. After receiving a response from Ollama, the backend:

- parses the response, including a fallback path that strips Markdown code fences if the content is returned as a formatted code block rather than plain JSON,
- validates the presence of the required keys and creates simple fallback suggestions if individual items are missing,
- normalises the suggestions with a helper function so that each suggestion has a consistent shape, and
- adds a `_meta` object containing the model name, the generation time, and the number of mood and recap entries used as input.

The baseline endpoint follows a similar pattern but does not receive or use individual mood data. Instead, the frontend passes up to three requested categories and a pool of generic example activities (see Appendix D.6). The backend trims this example pool to keep the prompt within a safe length, embeds it together with the categories into a simpler baseline prompt, and calls the same local model using the same JSON output schema. The returned suggestions are normalised in the same way and tagged with an `is_baseline` flag in the metadata so that baseline and tailored recommendations can be distinguished in the subsequent analysis more easily. Examples for parsed JSON outputs of both recommendation conditions and the enforced Schema can be found in Appendix D.

All external data collection (for example weather, health, and Spotify listening history) happens on the client side and is only sent to the backend in aggregated form as part of the mood entries and recaps. The backend itself only communicates with the local SQLite database and the local Ollama instance, which helps to ensure that sensitive mood and context data remain within the university's infrastructure and are not transmitted to commercial cloud providers.

4. Study Design

This chapter presents the apparatus, procedure, measurements and participants of the exploratory field study with a small sample ($n = 10$) conducted, evaluating MoodScape's recommendations, usability and user experience.

4.1. Apparatus

To conduct this study, three main components were used: online questionnaires deployed with Qualtrics, the Android application "MoodScape", and a backend running on a university-supplied server. Qualtrics was chosen because of its flexibility and straightforward data export.

Context data were gathered via external APIs: OpenWeather and Open-Meteo for weather data, Last.fm for music data (recently listened songs, artists and genres, listening duration), and the Health Connect API to access health and activity data from compatible apps. All three integrations were strictly optional; participants could use MoodScape without granting location access, connecting a Last.fm account, or enabling Health Connect. In such cases, the corresponding metrics were simply omitted from the recommendation payload, and recommendations were generated based on the remaining mood and context information.

4.2. Procedure

The quantitative study followed a 2×2 AB/BA crossover design [31, 37]. Participants were divided into two groups: Group 1 received tailored recommendations in week 1 and baseline recommendations in week 2, whereas Group 2 received the inverse order. Groups were assigned automatically based on the order of logging into the app for the first time.

Before participation, all users received an information sheet and consent form, followed by a participation code and an initial online questionnaire assessing demographics, prior experience with mood tracking, use of context-data apps, and perceived influences on their mood. After providing consent, participants installed the MoodScape APK and were instructed to use the app for two weeks, recording at least one mood entry or recap per day (with no upper limit) and viewing recommendations as they wished.

The field study ran from 23.02.2026 to 09.03.2026. On day 8 (after week 1) participants completed a second questionnaire about the first week's correlations and recommendations. Immediately after week 1 the recommendation condition switched for both groups, and participants continued using the app under the alternate condition. On day 15 (after week 2) they completed a third questionnaire about the second week's correlations and recommendations as well as their overall experience with the app.

4.3. Measurements

The study collected three groups of measures: perceived recommendation quality, engagement with recommendations, and overall user experience and usability of the app.

4. Study Design

4.3.1. Perceived Recommendation quality and correlations

Perceived quality of the system's output was assessed in the mid- and end-of-study questionnaires. For each week, participants rated:

- overall satisfaction with the information about correlations between contextual factors and mood (5-point Likert scale),
- perceived personalisation of the correlation information (5-point Likert scale),
- the extent to which the information helped them to understand their mood (5-point Likert scale).

In addition, participants indicated whether the system had discovered correlations about their mood that they had not noticed before and, if so, whether these new correlations made sense to them (categorical questions). The same structure was used for the recommendations themselves: participants rated overall satisfaction with the recommendations, perceived personalisation, and perceived helpfulness for improving mood on 5-point Likert scales.

4.3.2. Engagement with recommendations

Behavioural engagement with the recommendations was measured by items asking how often participants tried the recommended actions (5-point Likert scale from “never” to “always”). For participants who reported that they never or rarely tried the recommendations, a multiple-choice item captured reasons for low adherence (e.g. lack of time, forgetting, lack of interest, doubts about effectiveness, difficulty integrating into daily routine, missing resources, lack of results, lack of understanding), with an additional free-text option for other reasons. These items were included in both weekly questionnaires to compare engagement between the two recommendation conditions.

4.3.3. User experience and usability

Overall experience with MoodScope after the two-week usage period was assessed with two established questionnaires. User experience was measured using the User Experience Questionnaire (UEQ). The UEQ measures user experience along six dimensions; three capture pragmatic quality (perspicuity, efficiency, dependability) and two capture hedonic quality (stimulation, novelty), plus an overall attractiveness dimension [28, 42]. Perceived usability was measured using the System Usability Scale (SUS), a 10-item questionnaire with 5-point Likert items that yields a single usability score on a 0–100 scale [8, 20]. In addition, participants' reflective engagement with the app was assessed using the nine items of the Technology-Supported Reflection Inventory (TSRI) [5], a validated scale for evaluating how effectively an interactive system supports reflection (see Section 2.5). The TSRI measures three dimensions on a 7-point Likert scale: *Insight* (whether the system prompted personal change), *Exploration* (whether the system supported engagement with one's own data), and *Comparison* (whether the system facilitated social reflection). The full item wording for all questionnaires is provided in Appendix A.

4.4. Participants

The study was conducted with ten invited participants, all of whom took part in the field study. Of these ten participants, seven were male and three female. The mean age was 30.6 years (SD = 14.02), with two notable outliers aged 55 and 59 pulling the mean upwards; without these, the mean age would be 23.6 years.

Half of the participants reported being university students, 40% were employed, and one participant reported being unemployed. Eighty percent considered themselves to be self-reflective. Of those, 63% reported that

they mostly enjoy self-reflection, 25% that they always do, and 13% that their enjoyment depends on the situation and topic.

Regarding prior experience with mood tracking, 60% reported that they had never tracked their mood before, and 40% reported that they had. Of those with prior experience, 25% had used a different smartphone app and 75% had used a paper diary. Seventy-five percent of participants with prior mood tracking experience stated that it had helped them understand their emotions better, while 25% reported that it had not.

Eighty percent of participants reported regular use of a music app (Spotify, YouTube Music, SoundCloud, Streamcloud or Amazon Music). Seventy percent reported that they do not regularly use health or fitness apps; the remaining participants listed Strava, MyFitnessPal and Samsung Health as examples.

Participants were also asked which factors they believe influence their mood.

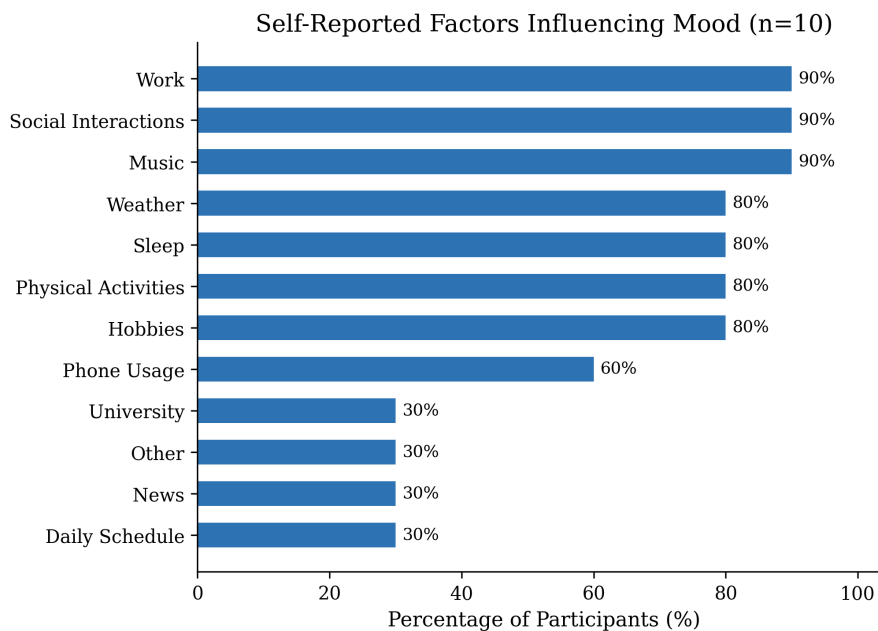


Figure 4.1.: Self-reported factors influencing mood among study participants (n=10)

As shown in Figure 4.1, 9 of the 10 participants indicated that work, social interactions and music influence their mood, and 80% reported that weather, sleep, physical activities and hobbies influence their mood. Sixty percent attributed influences on their mood to phone usage, and 30% considered university, news and daily schedule to be influential.

5. Results

This chapter presents the findings from the two-week field study. Given the small sample size ($n = 10$), all quantitative results are reported descriptively and should be interpreted as exploratory rather than confirmatory. The data were collected across both conditions of the AB/BA crossover design described in Chapter 4: each participant experienced both the baseline and tailored recommendation condition, yielding 10 ratings per condition for each measure.

5.1. Study Compliance and Data Overview

Before examining the questionnaire results, it is important to consider how actively participants engaged with the app during the study period. Over the two-week field study (February 23 to March 9, 2026), participants created a total of 177 mood entries, including both real-time entries and retrospective recaps. The number of entries per participant varied considerably, ranging from 4 (P06) to 38 (P07), with a mean of $M = 17.7$ ($SD = 10.0$, $Mdn = 17.0$).

Table 5.1.: Mood entries per participant and study week. “Cond.” indicates whether the participant received tailored (T) or baseline (B) recommendations in that week. Active days denotes the number of distinct days with at least one mood entry.

Participant	Group	Week 1		Week 2		Total entries	Active days
		Cond.	Entries	Cond.	Entries		
P01	1	T	8	B	7	15	10
P02	2	B	12	T	7	19	11
P03	1	T	10	B	11	21	14
P04	2	B	3	T	2	5	5
P05	2	B	11	T	9	20	12
P06	1	T	4	B	0	4	2
P07	1	T	17	B	21	38	10
P08	1	T	15	B	12	27	14
P09	2	B	5	T	8	13	8
P10	2	B	7	T	8	15	13
Total / Mean			92		85	177	9.9

5. Results

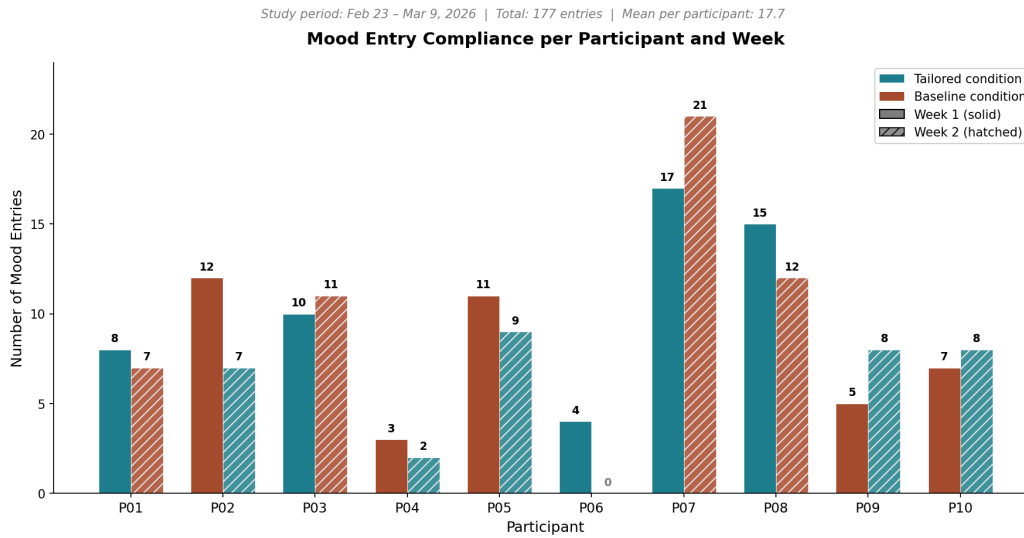


Figure 5.1.: Mood entries per participant, split by study week. Colour indicates the recommendation condition (teal = tailored, rust = baseline); hatching distinguishes Week 2 from Week 1. P06 did not create any entries in Week 2.

Figure 5.1 and Table 5.1 reveal substantial differences in engagement across participants. P07 and P08 were the most active users, frequently logging multiple entries per day, whereas P04 and P06 showed minimal engagement. Notably, P06 did not create any mood entries during Week 2 and logged only four entries on two active days in total, which limits the interpretability of this participant’s questionnaire responses. P04 similarly logged only five entries over five days across both weeks.

Aggregated by condition rather than by week, participants created a comparable number of entries during the baseline ($M = 8.9$, $SD = 5.9$) and tailored ($M = 8.8$, $SD = 4.5$) condition periods, indicating that engagement levels were not systematically influenced by the recommendation type. Entry counts were slightly higher in Week 1 (92) than in Week 2 (85), which may reflect a mild novelty-wears-off effect, although the difference is small and unevenly distributed across participants.

5.2. Perceived Quality of Correlation Insights

Participants rated their satisfaction with the correlation insights, the perceived personalisation, and whether the information helped them understand their mood, on a 5-point Likert scale (1 = not at all, 5 = completely).

Table 5.2.: Perceived quality of correlation insights by condition ($n = 10$ per condition)

Item	Baseline		Tailored	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Satisfaction with correlations	2.60	0.84	2.60	0.84
Personalisation of correlations	2.10	0.99	2.80	0.92
Helped understand mood	2.80	0.79	2.90	0.99

As shown in Table 5.2 and Figure 5.2, both conditions received moderate ratings across all three items. Overall satisfaction with the correlation insights was identical between conditions ($M = 2.60$, $SD = 0.84$). Notably, the tailored condition was perceived as more personalised ($M = 2.80$, $SD = 0.92$) than the baseline ($M = 2.10$, $SD = 0.99$), and participants reported that the tailored information helped them understand their mood slightly more ($M = 2.90$, $SD = 0.99$ vs. $M = 2.80$, $SD = 0.79$). However, given the small sample size and overlapping standard deviations, these differences should be interpreted with caution.

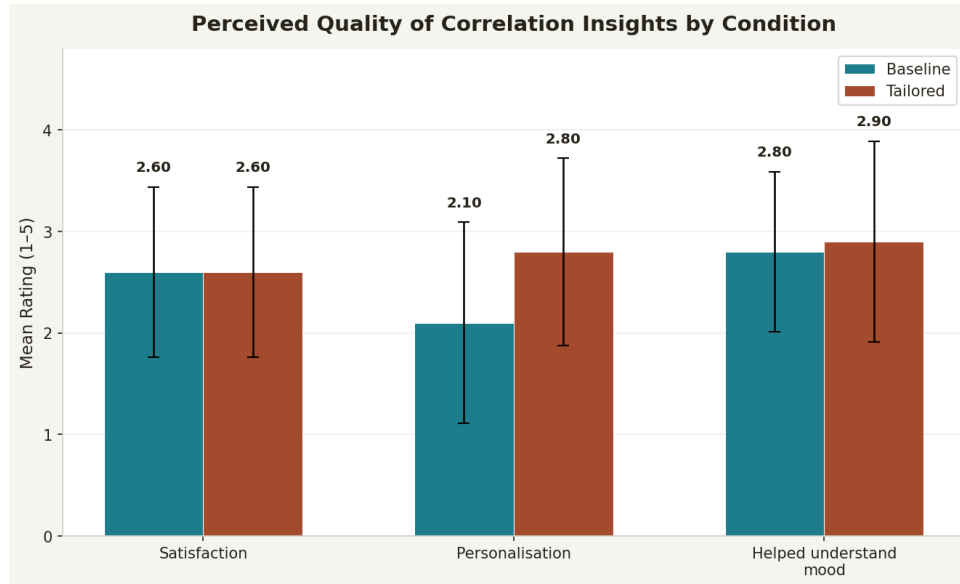


Figure 5.2.: Mean ratings for perceived quality of correlation insights by condition. Error bars represent one standard deviation.

When asked whether the system had discovered correlations they had not noticed before, eight out of ten participants in the baseline condition and six out of ten in the tailored condition answered *no*. The Other people answered *yes*, *sometimes*, with all of them then stating, that those, when newly discovered, made sense.

5.3. Perceived Quality of Recommendations

Participants rated their satisfaction with the recommendations, perceived personalisation, how frequently they tried the recommended actions, and whether the suggestions helped improve their mood, again on a 5-point Likert scale.

Table 5.3.: Perceived quality of recommendations by condition ($n = 10$ per condition)

Item	Baseline		Tailored	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Satisfaction with recommendations	2.40	0.70	2.90	0.88
Personalisation of recommendations	2.00	0.82	2.80	0.92
Frequency of trying recommendations	2.60	0.84	2.80	1.23
Helped improve mood	2.00	0.94	2.40	1.43

5. Results

Table 5.3 and Figure 5.3 show that the tailored condition received higher ratings than the baseline on all four items. The most pronounced difference was observed for perceived personalisation ($M = 2.80$, $SD = 0.92$ vs. $M = 2.00$, $SD = 0.82$), followed by overall satisfaction ($M = 2.90$, $SD = 0.88$ vs. $M = 2.40$, $SD = 0.70$) and perceived mood improvement ($M = 2.40$, $SD = 1.43$ vs. $M = 2.00$, $SD = 0.94$). The frequency of trying the recommendations was similar between conditions ($M = 2.80$ vs. $M = 2.60$). While these differences consistently favour the tailored condition, the overlapping standard deviations and small sample size warrant cautious interpretation.

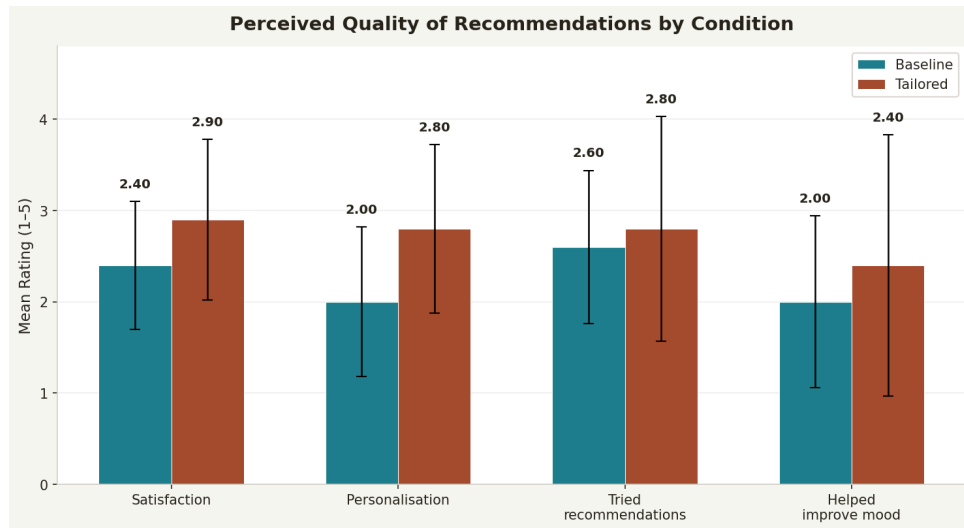


Figure 5.3. Mean ratings for perceived recommendation quality by condition. Error bars represent one standard deviation.

The tailored condition showed notably higher variance across most items (e.g., $SD = 1.43$ for mood improvement vs. $SD = 0.94$ under baseline), suggesting more heterogeneous responses. This is further illustrated in Figure 5.4, which displays each participant's individual ratings across both conditions. Participants P03 and P04 consistently gave higher ratings in both conditions, while P07 and P08 rated both conditions low across most items.

5.3. Perceived Quality of Recommendations

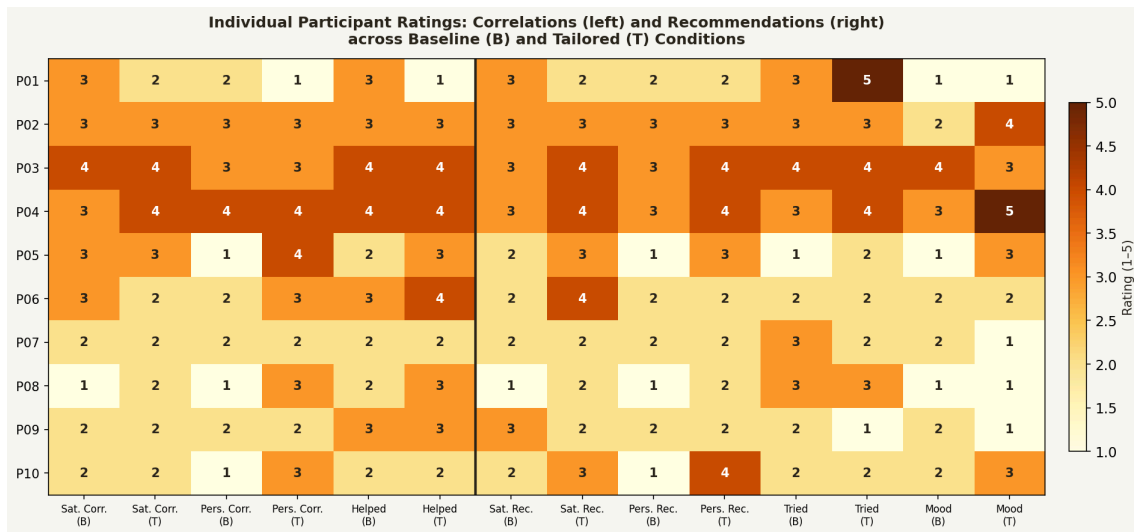


Figure 5.4.: Individual participant ratings for correlation insights (left) and recommendations (right) across baseline (B) and tailored (T) conditions. Higher values indicate more positive ratings on a 1–5 scale.

Participants who reported rarely or never trying the recommendations were asked to select the reasons (multiple selections allowed). Table 5.4 breaks down the cited reasons by condition. Overall, nine participants provided Q6 responses across both weeks (four under baseline, five under tailored), yielding 21 reason mentions in total.

Table 5.4.: Reasons for not following recommendations by condition (multiple selections allowed)

Reason	Baseline	Tailored	Total
Doubt about Effectiveness	4	2	6
Lack of Interest	2	3	5
Lack of Time	1	3	4
Forgotten	2	2	4
Difficulty Understanding	1	0	1
No Positive Results	0	1	1
Accessibility Issues	0	0	0
Difficulty Performing	0	0	0
Other	0	0	0
Total mentions	10	11	21

The total number of barrier mentions was nearly balanced between conditions (10 baseline vs. 11 tailored), with four and five respondents respectively. *Doubt about Effectiveness* was the most frequently cited reason overall (6 mentions) and was cited more often under the baseline condition (4 vs. 2), whereas *Lack of Interest* and *Lack of Time* were cited slightly more often under the tailored condition. Given the small number of respondents and total mentions, these differences are likely attributable to individual variation rather than systematic condition effects.

5.4. System Usability

System usability was assessed using the System Usability Scale (SUS) [8] administered in the end survey after both weeks. Individual SUS scores ranged from 47.5 to 90.0, with a mean of $M = 74.75$ ($SD = 14.26$, $Mdn = 78.8$).

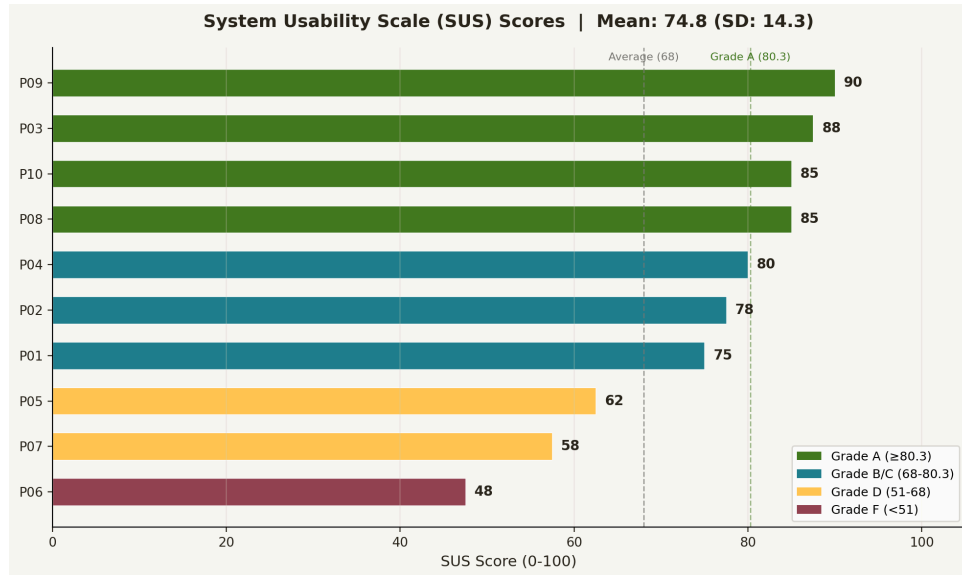


Figure 5.5.: Individual SUS scores per participant, colour-coded by adjective rating. The dashed lines indicate the empirical average (68) and Grade A threshold (80.3) [40]. Mean and SD rounded up.

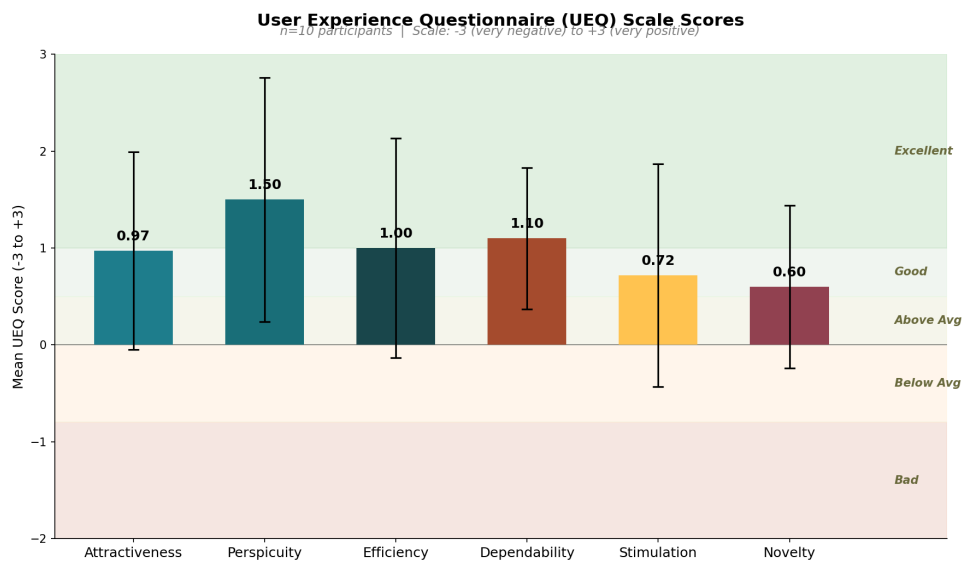
According to commonly used SUS benchmarks, scores around 68 are considered to reflect “average” usability, while scores above roughly 80 fall into an “excellent” range and correspond to A grades in Lewis et al.’s letter-grade interpretation [20, 29]. As shown in Figure 5.5, four participants scored in the excellent range (≥ 80.3) after Sauro [40], three scored between 68 and 80.3, and three scored below 68. The overall mean of 74.75 therefore places MoodScape above the average benchmark and roughly in the B range on the SUS letter-grade scale.

5.5. User Experience

The User Experience Questionnaire (UEQ) [28] was administered once in the end survey. Following the standard UEQ scoring procedure [42], item responses on a 7-point semantic differential scale were transformed to a -3 to $+3$ range and aggregated into six subscales: Attractiveness, Perspicuity, Efficiency, Dependability, Stimulation, and Novelty.

Table 5.5.: UEQ scale scores ($n = 10$; scale range: -3 to $+3$)

Scale	M	SD
Attractiveness	0.97	1.02
Perspicuity	1.50	1.26
Efficiency	1.00	1.13
Dependability	1.10	0.73
Stimulation	0.72	1.15
Novelty	0.60	0.84

**Figure 5.6.:** Mean UEQ scale scores with standard deviations. Background shading indicates the UEQ benchmark categories [42].

All six scales received positive mean scores (Table 5.5, Figure 5.6). Perspicuity scored highest ($M = 1.50$, $SD = 1.26$), indicating that participants found the system easy to understand and learn. Dependability ($M = 1.10$), Efficiency ($M = 1.00$), and Attractiveness ($M = 0.97$) also received clearly positive ratings. The two hedonic quality scales scored lower: Stimulation ($M = 0.72$, $SD = 1.15$) and Novelty ($M = 0.60$, $SD = 0.84$) both fell in the neutral range of the UEQ benchmark [42]. As shown in the per-participant heatmap (Figure 5.7), P04 rated the system most positively across all scales, while P07 and P06 provided the most critical assessments, particularly for Stimulation and Attractiveness.

5. Results

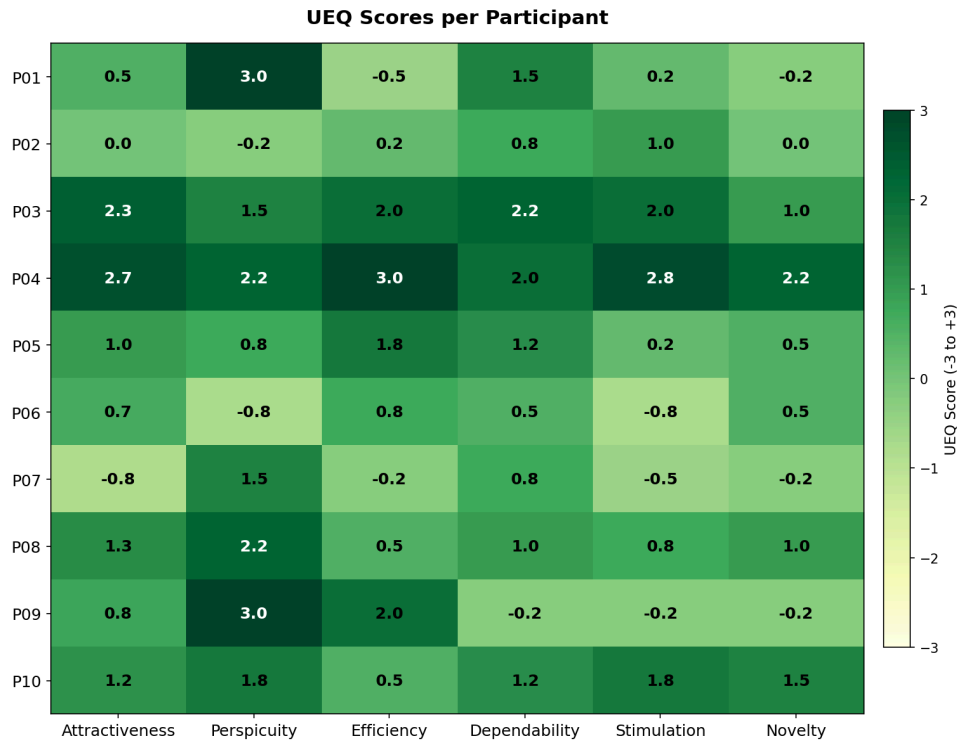


Figure 5.7.: UEQ scores per participant across all six scales. Warmer (greener) colours indicate more positive ratings.

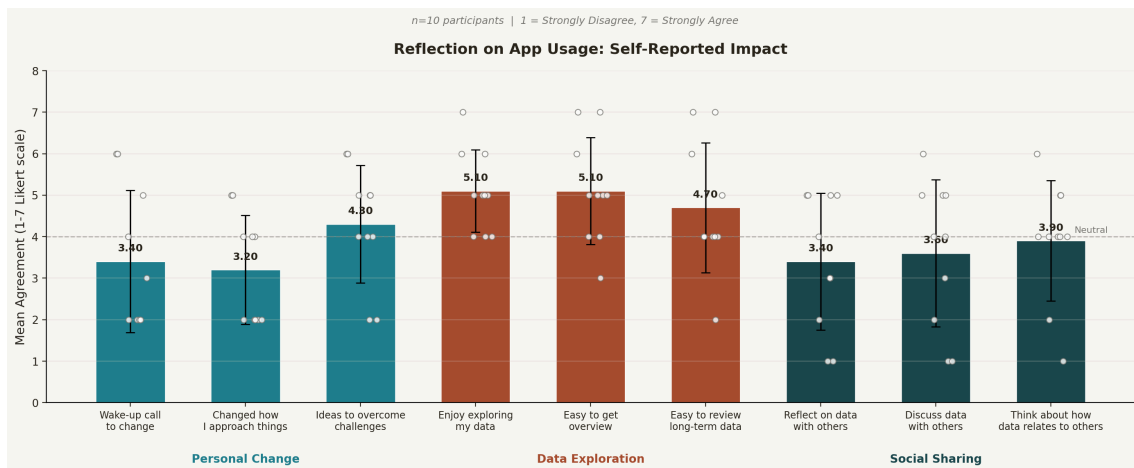
Following the distinction introduced by Hassenzahl [21], the three pragmatic quality scales (Perspicuity, Efficiency, Dependability) describe goal-directed aspects of interaction, while the two hedonic quality scales (Stimulation, Novelty) capture non-goal-directed experiential aspects. In this study, the pragmatic qualities were all rated above 0.8 (the threshold for a positive evaluation in the UEQ framework), whereas both hedonic scales fell below that threshold. This gap suggests that MoodScape was perceived as functionally useful and easy to use, though somewhat less exciting or novel.

5.6. Reflection on App Usage

In the end survey, participants rated nine statements about their overall experience using the app on a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree). These items can be grouped into three conceptual categories that align with the labels in Figure 5.8: *Insight* (personal change; items 1–3), *Exploration* (data exploration; items 4–6), and *Comparison* (social sharing; items 7–9).

Table 5.6.: Reflection on app usage ($n = 10$; 7-point Likert scale)

Category	Statement (abbreviated)	M	SD	Mdn
<i>Insight</i> (personal change)	Wake-up call to make changes	3.40	1.71	2.5
	Changed how I approach things	3.20	1.32	3.0
	Ideas on how to overcome challenges	4.30	1.42	4.5
<i>Exploration</i> (data exploration)	Enjoy exploring my data	5.10	0.99	5.0
	Easy to get an overview	5.10	1.29	5.0
	Easy to review long-term data	4.70	1.57	4.0
<i>Comparison</i> (social sharing)	Reflect on data with others	3.40	1.65	3.5
	Discuss data with others	3.60	1.78	4.0
	Think about how data relates to others	3.90	1.45	4.0

**Figure 5.8.:** Mean agreement ratings for reflection statements, grouped by category. Individual data points are shown as dots. The dashed line indicates the neutral midpoint (4).

The Exploration items (labelled as data exploration in Figure 5.8) received the highest ratings overall ($M = 4.97$ across all three items), with participants particularly agreeing that they enjoyed exploring their data ($M = 5.10$) and that the system made it easy to get an overview ($M = 5.10$). These scores are well above the neutral midpoint of 4, indicating a positive assessment of the app's data exploration capabilities.

The Insight items (corresponding to personal change) scored below or near the neutral midpoint. While participants somewhat agreed that the system gave them ideas on how to overcome challenges ($M = 4.30$), they were more neutral about whether the system had served as a wake-up call ($M = 3.40$) or changed how they approach things ($M = 3.20$). This suggests that the two-week study period may not have been sufficient for deeper behavioural changes to manifest.

The Comparison items (corresponding to social sharing) fell around the neutral midpoint ($M = 3.63$ across all three items), with moderate variability. This is consistent with the individual, self-tracking nature of the app, which did not include explicit social features.

5.7. Qualitative Feedback

In the open-ended question at the end of the study, six participants provided improvement suggestions. The most commonly mentioned themes were:

- **Missing delete/edit functionality:** Two participants (P05, P10) noted the inability to delete or modify mood entries after submission.
- **Visual design and layout:** Two participants (P06, P07) suggested improvements to the app's visual attractiveness, including colours and layout.
- **Recommendation quality:** Two participants (P07, P08) noted that the recommendations were sometimes too general or repetitive. P08 explicitly attributed this to the underlying language model: *"Presumably, a better LLM would yield better results."*
- **Richer mood input:** P07 expressed a desire for more emotion categories, intensity ratings, and more flexible time tracking for social interactions.
- **Additional data features:** P04 and P10 suggested features such as viewing which songs were listened to at specific times and reducing visual clutter in the detail views.

6. Discussion

This chapter interprets the results presented in Chapter 5 in the context of the related work reviewed in Chapter 2. It discusses the perceived quality of the system’s recommendation and correlation features, reflects on MoodScape’s role as a reflective well-being tool, considers the implications of the privacy-preserving architecture, and acknowledges the limitations of the study.

6.1. Perceived Quality of Recommendations and Correlations

The central research question of this thesis concerns whether LLM-generated, context-aware recommendations are perceived as more personalised and helpful than generic baseline recommendations. The descriptive results show a consistent pattern: the tailored condition received higher ratings than the baseline on all four recommendation items, with the largest differences observed for perceived personalisation ($M = 2.80$ vs. $M = 2.00$) and overall satisfaction ($M = 2.90$ vs. $M = 2.40$). A similar pattern emerged for the correlation insights, where the tailored condition was rated as more personalised ($M = 2.80$ vs. $M = 2.10$), while satisfaction and perceived helpfulness were nearly identical between conditions.

These findings provide tentative support for the hypothesis that context-aware, LLM-generated recommendations are perceived as more personalised than generic suggestions, even when generated by a relatively small, locally hosted model. However, the absolute rating levels remain moderate—all means fall below the midpoint of the 5-point scale—indicating that neither condition fully met participants’ expectations. This aligns with prior work on mood-tracking applications, which has shown that users generally appreciate receiving support and suggestions but often find that existing apps provide limited, generic guidance and insufficiently actionable feedback [11, 43]. Participant P08 explicitly noted in the qualitative feedback that “presumably, a better LLM would yield better results,” pointing to the limitations of the Ollama-hosted model used in this study. The quality of the language model’s output thus remains a critical bottleneck, even though the personalisation approach itself appears to be perceived positively.

The higher variance observed in the tailored condition across most items (e.g., $SD = 1.43$ vs. $SD = 0.94$ for mood improvement) is also worth noting. This heterogeneity suggests that the tailored recommendations resonated well with some participants but not with others, possibly depending on individual differences in how much personal context was available in the mood data or on participants’ general receptiveness to AI-generated suggestions. The per-participant heatmap (Figure 5.4) illustrates this clearly: participants P03 and P04 gave consistently positive ratings across both conditions, while P07 and P08 rated both conditions low.

The compliance data reported in Section 5.1 add an important dimension to this picture. Engagement levels varied dramatically—from just 4 entries (P06) and 5 entries (P04) to 38 entries (P07) and 27 entries (P08)—yet the relationship between engagement and perceived quality was not straightforward. P04 logged very few entries but gave among the most positive ratings across all items, whereas P07 and P08 were highly active users who nevertheless rated both conditions low. One possible interpretation is that participants with more extensive mood data may have formed higher expectations for what personalised recommendations should offer, leading to greater disappointment when the small LLM’s output fell short. Conversely, less engaged participants may have had fewer reference points and thus judged the recommendations less critically. The case of P06 is particularly noteworthy: this participant logged zero entries during Week 2 (the baseline condition period) yet still provided questionnaire responses for that week, raising questions

6. Discussion

about the validity of these ratings. Although excluding individual participants from a sample of ten would be statistically problematic, the compliance data highlight that not all questionnaire responses rest on equally solid foundations of actual app usage.

Importantly, the aggregated compliance data show no systematic condition bias: participants created a comparable number of entries during the baseline ($M = 8.9$) and tailored ($M = 8.8$) condition periods, indicating that the recommendation type did not influence overall engagement levels. This symmetry strengthens the interpretation that the observed differences in perceived quality reflect genuine differences in how the two types of recommendations were perceived, rather than artefacts of differential engagement. This pattern is consistent with findings from related work on personalised digital health interventions, where individual differences in engagement and perceived usefulness are commonly reported [11].

To explore these differences more formally, Wilcoxon signed-rank tests were conducted as supplementary analyses (see Appendix D.1). Of the seven comparisons, the “Personalisation of recommendations” item reached statistical significance ($W = 0.0$, $p = .039$, $r = .65$), with a large effect size favouring the tailored condition. While this result must be treated with extreme caution due to the small sample size and the absence of multiple-comparison correction, it provides preliminary quantitative support for the descriptive finding that participants perceived the tailored recommendations as more personalised. None of the remaining six comparisons reached statistical significance at $\alpha = .05$, which is unsurprising given the very low statistical power of paired non-parametric tests with $n = 10$. Overall, the absence of significant differences on most items does not imply the absence of effects; rather, it reflects the inherent power limitations of small-sample exploratory field studies.

6.2. Usability and User Experience

The System Usability Scale score of $M = 74.75$ ($Mdn = 78.8$) places MoodScape above the commonly cited benchmark of 68 for average usability, whose suitability for digital health applications was examined by Hyzy et al. [23]. In terms of SUS interpretation frameworks, this corresponds to an above-average score roughly in the B range of the letter-grade and adjective ratings proposed by Lewis and colleagues [20, 29]. This is an encouraging result for a research prototype, particularly given that the app was developed in the context of a Bachelor’s thesis and ran on a university-managed server infrastructure rather than a polished commercial backend. Combined with the fact that five out of ten participants scored in the excellent range (≥ 80.3) proposed by Sauro. [40], this suggests that the core interaction flow—logging mood, viewing correlations, and reading recommendations—was perceived as straightforward by the majority of users.

The UEQ results complement this picture. All three pragmatic quality scales—Perspicuity ($M = 1.50$), Dependability ($M = 1.10$), and Efficiency ($M = 1.00$)—exceeded the 0.8 threshold that indicates a positive evaluation in the UEQ framework, while both hedonic scales fell below it: Stimulation ($M = 0.72$) and Novelty ($M = 0.60$). This pattern is consistent with Hassenzahl’s model of user experience [21]: systems that are clearly functional and easy to use may still lack the engaging, novel qualities that drive long-term adoption. The particularly low Novelty score suggests that participants did not perceive MoodScape as especially innovative compared to other applications, which is plausible for a mood-tracking app built with familiar UI paradigms. The low Stimulation score aligns with the qualitative feedback from participants P06 and P07, who suggested improvements to the visual design. Prior work on mental-health and mood-tracking apps has identified visual appeal and perceived novelty as important factors for sustained engagement [11], suggesting that future iterations of MoodScape should invest in more polished visual design and novel interaction patterns to increase hedonic appeal.

An interesting observation from the per-participant UEQ heatmap (Figure 5.7) is the considerable variability across participants. P04 rated the system very positively across all dimensions, while P07 and P06 were more critical. This inter-individual makes sense in small-sample UEQ evaluations and highlights the importance of combining standardised questionnaires with qualitative data to understand individual user experiences more deeply.

6.3. MoodScape as a Reflective Well-Being App

The title of this thesis characterises MoodScape as a *reflective* well-being app, and the TSRI results from the end survey (Section 5.6) provide an opportunity to examine whether this characterisation is warranted against the theoretical background reviewed in Section 2.5. The Exploration subscale received the highest ratings (all above the neutral midpoint of 4), with participants particularly agreeing that they enjoyed exploring their data ($M = 5.10$) and that the system made it easy to get an overview ($M = 5.10$). This suggests that MoodScape’s analytics features—the correlation insights, mood visualisations, and contextual summaries—successfully supported what the TSRI framework describes as *Exploration*: the process of actively engaging with one’s own tracked data to gain new insights [5].

However, the Insight subscale scored below the neutral midpoint. Participants were relatively neutral about whether the app served as a “wake-up call” ($M = 3.40$) or changed how they approach things ($M = 3.20$). This pattern can be interpreted through Fleck and Fitzpatrick’s levels of reflection [18]: MoodScape’s data visualisations and correlation features appear to support lower levels of reflection well (R0–R2: revisiting, describing, and exploring data), but the transition to transformative reflection (R3)—where users reorganise their perspective and change behaviour—did not manifest within two weeks. The compliance data (Section 5.1) add context to this finding: the wide range of engagement levels (4–38 entries across participants) means that some users accumulated substantially less data to reflect upon. Participants with very few entries—such as P06 (4 entries) and P04 (5 entries)—had limited material for meaningful data exploration or pattern discovery, which may have contributed to the moderate Insight scores. This is also consistent with Schön’s distinction between *reflection-in-action* (immediate awareness during data exploration) and *reflection-on-action* (longer-term deliberate review leading to behavioural change) [41]: a two-week study period may be sufficient to trigger the former but too short for the latter. Prior work on mood-tracking apps has similarly found that users often adopt regular logging and review routines to increase self-awareness, but that translating these insights into concrete behaviour change is challenging and only weakly supported by most existing apps [11, 43].

Notably, the item “Ideas on how to overcome challenges” scored highest within the Insight subscale ($M = 4.30$), which suggests that the recommendation feature—despite its moderate satisfaction ratings—did contribute to participants’ sense that the system offered actionable guidance. In the TSRI framework, this item captures whether the system helps users identify concrete steps for improvement [5]. The fact that this item scored above the neutral midpoint while the other two Insight items did not, could suggest that MoodScape’s recommendations served as a partial bridge between data exploration and transformative reflection: they provided ideas, but the study period was too short for these ideas to manifest as actual behavioural change. In this sense, MoodScape occupies an intermediate position in the spectrum touched upon in Section 2.1: more than a passive tracking tool, but not yet a fully interactive coaching system.

The Comparison subscale fell around the neutral midpoint ($M = 3.63$), which is unsurprising given that MoodScape was designed as an individual self-tracking tool without explicit social features. Nevertheless, some participants indicated that they discussed their data with others ($M = 3.60$) or thought about how their data relates to other people ($M = 3.90$). The TSRI literature notes that social aspects of reflection—such as conversations about tracked data—can deepen individual sense-making [5]. This might suggest that even without built-in sharing mechanisms, mood-tracking data can prompt social reflection—an aspect that could be explored in future designs, for instance through optional sharing of anonymised trends or comparison with aggregate data.

Overall, interpreting the results through the TSRI framework and Fleck and Fitzpatrick’s levels of reflection, MoodScape goes beyond a conventional mood-tracking app by providing several reflection-support features. The system is strongest in facilitating Exploration (R1–R2 in Fleck and Fitzpatrick’s terms) and offering initial ideas for improvement (a first step towards R3), but more limited in prompting deep personal change within the timeframe of this study. As the personal informatics literature has emphasised, reflection does not follow automatically from data access but must be actively encouraged [14]. Truly reflective well-being

6. Discussion

support would likely require richer scaffolding, such as more interactive guidance, goal-setting features, or longer-term engagement mechanisms—directions that future work could explore.

6.4. Privacy and the Local LLM Architecture

A central design motivation for MoodScape, as discussed in Sections 2.2 and 2.3, was to avoid sending sensitive mood and health data to external cloud-based AI services. The architecture running Ollama on a university-managed server represents a pragmatic compromise between fully on-device processing and reliance on commercial API providers: the data never leave the institutional network, the university maintains full control over the server infrastructure, and no third-party provider receives access to participants' mood logs or health data.

From a GDPR perspective, this design choice aligns with the principles of data minimisation (Art. 5 GDPR) and data protection by design and by default (Art. 25 GDPR), as discussed in Section 2.3. Unlike commercial mental-health apps that have been shown to share user data with advertising networks and third-party analytics services [22, 25, 33], MoodScape's architecture ensures that sensitive mood data remain within a controlled environment. The trade-off, however, is clear: the locally hosted Ollama model (a relatively small open-source LLM) cannot match the output quality of larger commercial models such as GPT-4 or Claude. This trade-off was directly reflected in the previously mentioned participant P08's feedback and likely constrained the magnitude of the observed advantage of the tailored condition: although participants rated the tailored recommendations higher on all four items, the absolute satisfaction levels remained moderate.

This tension between privacy and model quality is an important consideration for future work. As open-source language models continue to improve in quality and efficiency, the gap between locally hosted and cloud-based solutions may narrow. Furthermore, emerging techniques such as on-device inference [47] on modern smartphones could eventually enable fully local processing without the need for a dedicated server, further strengthening the privacy guarantees while potentially improving recommendation quality through access to larger models.

6.5. Reasons for Not Following Recommendations

The analysis of reasons for not trying the recommendations (Section 5.3) revealed that *Doubt about Effectiveness* was the most frequently cited barrier (6 mentions), followed by *Lack of Interest* (5 mentions), *Lack of Time* and *Forgotten* (4 mentions each). These findings echo common barriers reported in the broader literature on digital health interventions and self-management tools [7, 30].

When broken down by condition (Table 5.4), the total number of barrier mentions was nearly balanced between conditions (10 baseline vs. 11 tailored). Notably, *Doubt about Effectiveness* was cited more often under the baseline condition (4 vs. 2), whereas *Lack of Time* was cited more often under the tailored condition (3 vs. 1). This relatively balanced distribution suggests that the barriers to following recommendations were not substantially different between conditions. The higher baseline count for *Doubt about Effectiveness* may reflect the fact that generic recommendations, by their nature, are less tied to the individual's specific context and may therefore appear less convincing. However, given the small number of respondents (four under baseline, five under tailored) and total mentions, these differences are more likely attributable to individual variation.

Future iterations of MoodScape could address these barriers by implementing more interactive recommendation delivery (e.g., letting users indicate preferences or rate suggestions), by providing evidence or explanations for why specific actions might help, or by introducing gentle nudging mechanisms that reduce the cognitive effort required to try a recommendation.

6.6. Limitations

Several limitations should be considered when interpreting the findings of this study.

Sample size and statistical power. The most significant limitation is the small sample size of $n = 10$ participants. While this is not unusual for exploratory field studies with research prototypes in HCI, it severely limits the ability to detect meaningful differences between conditions. As the supplementary Wilcoxon tests (Appendix D.1) complement, the statistical power for paired comparisons with ten observations is very low. The descriptive differences reported in this thesis should therefore be understood as preliminary observations that require validation in larger-scale studies.

Study duration. The two-week study period, with one week per condition, may have been too short for participants to fully engage with the recommendation features and observe potential effects on their mood or behaviour. Prior work on mood-tracking and technology-supported reflection suggests that while users can quickly start logging and reflecting on their data, achieving more substantial or sustained behaviour change usually requires ongoing support and longer-term use [5, 11]. A longer study could allow for more sustained engagement and potentially larger differences between conditions.

Model quality. The use of Ollama with a relatively small open-source model on a university server constrained the quality of the generated recommendations. A more capable model might have produced more nuanced, contextually appropriate, and ultimately more satisfying suggestions. However, using a larger model would either require more powerful hardware or reliance on external API providers, conflicting with the privacy-by-design motivation of the project.

Limited contextual data in the tailored condition. The quality of tailored recommendations depended on the availability of contextual signals, such as Health Connect metrics or Last.fm listening history. For participants who had not granted permissions, linked accounts, or accumulated sufficient activity data, the model effectively received a reduced payload, which may have led to less differentiated or more general recommendations—particularly during the early stages of the study.

Carry-over effects. Although the AB/BA crossover design helps control for individual differences by having each participant serve as their own control, it also introduces the possibility of carry-over effects, as described by Lim et al. [31]: the experience of one condition may influence the perception of the subsequent condition. For example, participants who received tailored recommendations in Week 1 might have found the baseline recommendations in Week 2 noticeably less personalised, or vice versa. The small sample size makes it difficult to systematically assess such order effects.

Participant population. All ten participants were recruited through the researcher’s personal network, including some university students. This convenience sampling limits the generalisability of the findings, as participants may differ from the broader population in their technology literacy, openness to self-tracking, and patterns of mood variation.

6. Discussion

Uneven engagement. As the compliance analysis (Section 5.1) revealed, engagement levels varied dramatically across participants, with some logging fewer than five mood entries over the entire study period. In particular, P06 created zero entries during Week 2, yet provided questionnaire responses for both weeks. This means that some condition ratings—especially for low-engagement participants—are based on very limited actual exposure to the app’s recommendation features. A minimum-engagement threshold was not imposed in this study to preserve the already small sample size, but future studies should consider requiring a minimum number of entries per week for data to be included in the per-condition analysis.

7. Conclusion

This thesis presented MoodScape, a reflective well-being app that combines mood tracking with context-aware, LLM-generated recommendations while preserving user privacy through a locally hosted language model. The system was evaluated in a two-week AB/BA crossover field study with ten participants, comparing a tailored recommendation condition (generated by the Ollama model using participants' personal mood and contextual data) against a generic baseline condition.

The descriptive results revealed a consistent pattern favouring the tailored condition: participants rated the LLM-generated recommendations higher than the baseline on all four comparison items, with the most pronounced differences on perceived personalisation ($M = 2.80$ vs. $M = 2.00$) and overall satisfaction ($M = 2.90$ vs. $M = 2.40$). A supplementary Wilcoxon signed-rank test confirmed this for the personalisation item ($p = .039$, $r = .65$), providing preliminary statistical support despite the small sample. The correlation insights showed a similar, though less pronounced, pattern. However, all absolute ratings remained below the scale midpoint, and the barriers to following recommendations were nearly equally distributed across conditions. These findings suggest that the personalisation approach itself was perceived positively, but the limited output quality of the small, locally hosted LLM constrained its effectiveness: the architecture successfully preserved data privacy, yet the model's ability to generate compelling, contextually appropriate recommendations left room for improvement. In addition, the tailored condition depended on the availability of contextual signals (e.g., health and music data), which were sometimes missing due to ungranted permissions or unlinked accounts, or were initially sparse, meaning that some "tailored" recommendations were effectively based on reduced input and may have felt less distinct from the baseline.

At the same time, MoodScape demonstrated encouraging results as a usable and functional system. The SUS score of 74.75 places it above the general benchmark, and the UEQ results support that participants perceived the system as easy to understand, dependable, and efficient to use. All three pragmatic quality scales exceeded the positive-evaluation threshold, while both hedonic scales (Stimulation and Novelty) remained in the neutral range—indicating that the system's functional strengths were not yet matched by equally engaging or innovative interaction qualities. Participants particularly valued the data exploration features, rating them well above the neutral midpoint, and moderately agreed that the system gave them ideas on how to overcome challenges. However, engagement levels varied considerably across participants (4–38 mood entries), and some participants with very low engagement still provided questionnaire ratings, which limits the interpretability of individual data points. Despite this variability, the aggregated compliance data showed no systematic difference in engagement between the two conditions, supporting the validity of the condition comparison. Overall, these findings suggest that MoodScape succeeded in its goal of going beyond passive mood logging by supporting active data exploration and, to a degree, self-reflection.

The reflectivity dimension—central to the thesis title—showed a differentiated picture. Interpreted through Fleck and Fitzpatrick's levels of reflection [18] and the TSRI framework [5], MoodScape seemingly supported lower-level reflection (data exploration and revisiting, R0–R2) through its analytics and correlation features, as evidenced by high Exploration subscale scores. However, transformative reflection (R3)—manifesting as deeper personal change—did not emerge within the two-week study period. This aligns with Schön's distinction between reflection-in-action and reflection-on-action [41] and with broader work on technology-supported reflection, which emphasises that deeper, behaviour-relevant reflection tends to emerge through sustained engagement over time rather than in brief interactions [14].

The privacy-preserving architecture represents a meaningful contribution in the context of growing concerns about data protection in digital mental-health tools. By running the LLM on a university-managed server

7. Conclusion

rather than relying on commercial cloud providers, MoodScape demonstrates that it is technically feasible to offer AI-powered recommendation features without exposing sensitive mood and health data to third parties. The trade-off between privacy and model quality remains relevant: while the tailored condition was preferred, a more capable model would likely have amplified this advantage.

7.1. Outlook

Several directions for future research emerge from the findings and limitations of this study.

Larger-scale evaluation. The most immediate next step would be to replicate the study with a substantially larger sample (e.g., $n \geq 30$) to enable meaningful inferential statistical analysis and to assess whether the consistent tailored-condition advantage observed here generalises to a broader population. A longer study duration (e.g., four to six weeks) would also allow for a more thorough examination of whether tailored recommendations lead to sustained behavioural changes over time.

Improved language models. As open-source LLMs continue to improve rapidly, future iterations of MoodScape could benefit significantly from more capable models. Replacing the current Ollama model with a larger or more recent open-source alternative—while still hosting it locally—could substantially improve the quality, specificity, and coherence of the generated recommendations. Evaluating how recommendation quality scales with model size and capability would be a valuable contribution, both for MoodScape specifically and for the broader field of privacy-preserving AI in digital health.

On-device inference. When the availability of efficient small language models optimised for mobile devices increases, future work could explore running the recommendation engine directly on the user's smartphone. This would eliminate the need for a dedicated server entirely, further strengthening privacy guarantees and enabling offline functionality.

Richer reflection support. The reflection results suggest that MoodScape's current feature set supports data exploration well but falls short of prompting deeper personal change. Future designs could incorporate more interactive reflection scaffolding, such as guided journaling prompts based on detected mood patterns, goal-setting and progress-tracking features, or periodic reflection summaries that help users connect their mood data to concrete life events and behavioural patterns. Integrating validated instruments like the TSRI [5] as repeated outcome measures over time would also allow for a more rigorous assessment of whether the system truly supports reflective well-being.

Interactive recommendations. The barrier analysis revealed that lack of interest and doubt about effectiveness were common reasons for not following recommendations. Future versions could address this through more interactive recommendation delivery: allowing users to rate, dismiss, or request alternatives for individual suggestions; providing brief explanations of why a recommendation was generated; or introducing opt-in nudging mechanisms (e.g., gentle reminders to try a suggestion at a suitable time). Such features could help bridge the gap between receiving a recommendation and actually acting on it.

Social and collaborative features. Although MoodScape was designed as an individual self-tracking tool, the moderate social sharing scores suggest some appetite for reflecting on mood data with others. Future work could carefully explore optional features such as sharing anonymised mood trends with trusted contacts, comparing personal patterns against aggregate data, or enabling therapist access to summaries—always with privacy as a guiding principle and explicit user consent as a prerequisite.

Bibliography

- [1] Aisha Alansari, Hamzah Luqman. “Large Language Models Hallucination: A Comprehensive Survey”. In: *arXiv preprint arXiv:2510.06265* (2025). DOI: /10.48550/arXiv.2510.06265.
- [2] Eric P. S. Baumer. “Reflective Informatics: Conceptual Dimensions for Designing Technologies of Reflection”. In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, 2015, pp. 585–594. DOI: 10.1145/2702123.2702234.
- [3] Eric P. S. Baumer, Vera Khovanskaya, Mark Matthews, Lindsay Reynolds, Victoria Schwanda Sosik, Geri Gay. “Reviewing Reflection: On the Use of Reflection in Interactive System Design”. In: *Proceedings of the 2014 Conference on Designing Interactive Systems (DIS '14)*. ACM, 2014, pp. 93–102. DOI: 10.1145/2598510.2598598.
- [4] Magdalena Benkard. “Selbstmanagement der psychischen Gesundheit mithilfe einer App zur kontextbezogenen Stimmungsaufzeichnung”. Institute of Computer Science. Master’s thesis. Munich, Germany: Ludwig-Maximilians-Universität München, 2024.
- [5] Marit Bentvelzen, Jasmin Niess, Mikołaj P. Woźniak, Paweł W. Woźniak. “The Development and Validation of the Technology-Supported Reflection Inventory”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. arXiv:2103.16399. ACM, 2021, pp. 1–8. DOI: 10.1145/3411764.3445673.
- [6] Charlotte Blease, John Torous. “ChatGPT and mental healthcare: balancing benefits with risks of harms.” In: *BMJ Mental Health*. 26.1 (2023), e300884. DOI: 10.1136/bmjment-2023-300884.
- [7] Judith Borghouts, Elizabeth Eikey, Gloria Mark, Cinthia De Leon, Stephen M Schueller, Margaret Schneider, Nicole Stadnick, Kai Zheng, Dana Mukamel, Dana H Sorkin. “Barriers to and Facilitators of User Engagement With Digital Mental Health Interventions: Systematic Review.” In: *J Med Internet Res*. 23.3 (2021), e24387. DOI: 10.2196/24387.
- [8] John Brooke. “: A quick and dirty usability scale”. In: *Usability Eval. Ind.* 189 (Nov. 1995).
- [9] Francesca Bruni, Valentina Mancuso, Elisa Pedroli. “How to Evaluate User Experience in Digital Health? A Scoping Review of Questionnaires in Virtual Reality Applications”. In: *Frontiers in Digital Health* 7 (2025), p. 1561364. DOI: 10.3389/fdgth.2025.1561364.
- [10] Michelle Nicole Burns, Mark Begale, Jennifer Duffecy, Darren Gergle, Chris J Karr, Emily Giangrande, David C Mohr. “Harnessing Context Sensing to Develop a Mobile Intervention for Depression”. In: *J Med Internet Res* 13.3 (2011), e55. DOI: 10.2196/jmir.1838.
- [11] Clara Caldeira, Yu Chen, Lesley Chan, Vince Pham, Yunan Chen, Kai Zheng. “Mobile apps for mood tracking: an analysis of features and user reviews”. In: *AMIA Annual Symposium Proceedings 2017* (2018), pp. 495–504.
- [12] Cyber Security Cooperative Research Centre. *Mental Health Apps: Privacy Risks*. Accessed 2026-03-02. 2021. URL: <https://cybersecuritycrc.org.au/mental-health-apps-privacy-risks/>.

- [13] Muna Dubad, Farah Elahi, Steven Marwaha. “The Clinical Impacts of Mobile Mood-Monitoring in Young People With Mental Health Problems: The MeMO Study”. In: *Frontiers in Psychiatry* 12 (2021). DOI: 10.3389/fpsy.2021.687270.
- [14] Daniel A. Epstein, An Ping, James Fogarty, Sean A. Munson. “A Lived Informatics Model of Personal Informatics”. In: *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*. ACM, 2015, pp. 731–742. DOI: 10.1145/2750858.2804250.
- [15] European Data Protection Board. *Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models*. https://www.edpb.europa.eu/system/files/2024-12/edpb_opinion_202428_ai-models_en.pdf. Accessed 2026-03-02. 2024.
- [16] Yangyan Fan, Ahui Fan, Zhiping Yang, Daiming Fan. “Global burden of mental disorders in 204 countries and territories, 1990-2021: results from the global burden of disease study 2021.” In: *BMC Psychiatry* 25.1 (May 2025), p. 486. DOI: 10.1186/s12888-025-06932-y.
- [17] Andrea Fiorillo. “A roadmap for better and personalized mental health care in Europe: the priorities of the European Psychiatric Association”. In: *European Psychiatry* 68.1 (2025), e60. DOI: doi:10.1192/j.eurpsy.2025.2456.
- [18] Rowanne Fleck, Geraldine Fitzpatrick. “Reflecting on Reflection: Framing a Design Landscape”. In: *Proceedings of the 22nd Conference of the Computer-Human Interaction Special Interest Group of Australia on Computer-Human Interaction (OZCHI '10)*. ACM, 2010, pp. 216–223. DOI: 10.1145/1952222.1952269.
- [19] Thomas Germain. *Mental Health Apps Aren't All As Private As You May Think*. Accessed 2026-03-02. 2021. URL: <https://www.consumerreports.org/health/health-privacy/mental-health-apps-and-user-privacy-a7415198244/>.
- [20] Rebecca Grier, Aaron Bangor, Philip Kortum, S. Peres. “The System Usability Scale: Beyond Standard Usability Testing”. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Vol. 57. 1. 2013, pp. 187–191. DOI: 10.1177/1541931213571042.
- [21] Marc Hassenzahl, Noam Tractinsky. “User Experience – A Research Agenda”. In: *Behaviour & Information Technology* 25.2 (2006), pp. 91–97. DOI: 10.1080/01449290500330331.
- [22] Kit Huckvale, John Torous, Mark E. Larsen. “Data Sharing and Privacy of Smartphone Apps for Depression and Smoking Cessation”. In: *JAMA Network Open* 2.4 (2019), e192542. DOI: 10.1001/jamanetworkopen.2019.2542.
- [23] Maciej Hyzy, Raymond Bond, Maurice Mulvenna, Lu Bai, Alan Dix, Simon Leigh, Siobhan Hunt. “System Usability Scale Benchmarking for Digital Health Apps: Meta-analysis.” In: *JMIR Mhealth Uhealth* 10.8 (2022), e37290. DOI: 10.2196/37290.
- [24] Marcello Ienca, Gianclaudio Malgieri. “Mental Data Protection and the GDPR”. In: *Journal of Law and the Biosciences* 9.1 (2022), lsac006. DOI: 10.1093/jlb/ljac006.
- [25] Leonardo Horn Iwaya, Ali M. Babar, Awais Rashid, Chamila Wijayarathna. “On the Privacy of Mental Health Apps: An Empirical Investigation and Its Implications for App Development”. In: *Empirical Software Engineering* 28.2 (2023). DOI: 10.1007/s10664-022-10236-0.

- [26] Matthew Jörke, Yasaman S. Sefidgar, Talie Massachi, Jina Suh, Gonzalo Ramos. “Pearl: A Technology Probe for Machine-Assisted Reflection on Personal Data”. In: *Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23)*. ACM, 2023, pp. 471–488. DOI: 10.1145/3581641.3584054.
- [27] Min-Gyu Kim, Gaeun Hwang, Junhyuk Chang, Seheon Chang, Hyun Woong Roh, Rae Woong. Park. “Performance of Open-Source Large Language Models in Psychiatry: Usability Study Through Comparative Analysis of Non-English Records and English Translations”. In: *Journal of Medical Internet Research* 27 (2025), e69857. DOI: 10.2196/69857.
- [28] Bettina Laugwitz, Theo Held, Martin Schrepp. “Construction and Evaluation of a User Experience Questionnaire”. In: *HCI and Usability for Education and Work*. Ed. by Andreas Holzinger. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 63–76. ISBN: 978-3-540-89350-9.
- [29] James R. Lewis, Jeff Sauro. *Item Benchmarks for the System Usability Scale*. Accessed 2026-03-17. 2018. URL: <https://uxpajournal.org/item-benchmarks-system-usability-scale-sus/>.
- [30] Ian Li, Anind Dey, Jodi Forlizzi. “A Stage-Based Model of Personal Informatics Systems”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, 2010, pp. 557–566. DOI: 10.1145/1753326.1753409.
- [31] Chi-Yeon Lim, Junyong In. “Considerations for crossover design in clinical study”. In: *Korean Journal of Anesthesiology* 74.4 (Aug. 2021), pp. 293–299. DOI: 10.4097/kja.21165.
- [32] Tayla McCloud, Rebecca Jones, Gemma Lewis, Vaughan Bell, Elias Tsakanikos. “Effectiveness of a Mobile App Intervention for Anxiety and Depression Symptoms in University Students: Randomized Controlled Trial”. In: *JMIR mHealth and uHealth* 8.7 (2020), e15418. DOI: 10.2196/15418.
- [33] Mozilla Foundation. **Privacy Not Included: Mental Health Apps*. Accessed 2026-03-13. 2023. URL: <https://www.mozillafoundation.org/en/privacynotincluded/categories/mental-health-apps/>.
- [34] Eugenie Park, Darrell M. West. *Why Mental Health Apps Need to Take Privacy More Seriously*. Accessed 2026-03-02. 2024. URL: <https://www.brookings.edu/articles/why-mental-health-apps-need-to-take-privacy-more-seriously/>.
- [35] Lisa Parker, Vanessa Halter, Tanya Karliychuk, Quinn Grundy. “How private is your mental health app data? An empirical study of mental health app privacy policies and practices”. In: *International Journal of Law and Psychiatry* 64 (2019), pp. 198–204. ISSN: 0160-2527. DOI: 10.1016/j.ijlp.2019.04.002. URL: <https://www.sciencedirect.com/science/article/pii/S0160252718302681>.
- [36] Ashley Polhemus, Sara Simblett, Erin Dawe-Lane, Gina Gilpin, Benjamin Elliott, Sagar Jilka, Jan Novak, Ralica Ileana Nica, Gergely Temesi, Til Wykes. “Health Tracking via Mobile Apps for Depression Self-management: Qualitative Content Analysis of User Reviews”. In: *JMIR Hum Factors* 9.4 (2022), e40133. DOI: 10.2196/40133.
- [37] James F. Reed III. “AB/BA Crossover Trials - Binary Outcome”. In: *Journal of Modern Applied Statistical Methods* 5.2 (Jan. 2005), pp. 452–457. DOI: 10.22237/jmasm/1162354680.

Bibliography

- [38] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 (General Data Protection Regulation). <https://eur-lex.europa.eu/eli/reg/2016/679/oj>. Accessed 2026-03-13. 2016.
- [39] Annika Rischmüller. *Data Protection in (Mental) Health Apps*. Accessed 2026-03-02. 2024. URL: <https://www.datenschutz-notizen.de/data-protection-in-mental-health-apps-3043790/>.
- [40] Jeff Sauro. *Measuring Usability with the System Usability Scale (SUS)*. Accessed 2026-03-17. 2011. URL: <https://measuringu.com/sus/>.
- [41] Donald A. Schön. *The Reflective Practitioner: How Professionals Think in Action*. New York: Basic Books, 1983.
- [42] Martin Schrepp, Andreas Hinderks, Jörg Thomaschewski. *User Experience Questionnaire (UEQ) Handbook*. Version 2. UEQ Research Group. 2014. URL: <https://www.ueq-online.org/Material/Handbook.pdf>.
- [43] Stephen M Schueller, Martha Neary, Jocelyn Lai, Daniel A Epstein. “Understanding People’s Use of and Perspectives on Mood-Tracking Apps: Interview Study”. In: *JMIR Ment Health* 8.8 (2021), e29368. DOI: 10.2196/29368.
- [44] Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, Pranav Sandeep Dulepet, Saurav Vidyadhara, Dayeon Ki, Sweta Agrawal, Chau Pham, Gerson Kroiz, Feileen Li, Hudson Tao, Ashay Srivastava, Hevander Da Costa, Saloni Gupta, Megan L. Rogers, Inna Goncareenco, Giuseppe Sarli, Igor Galynker, Denis Peskoff, Marine Carpuat, Jules White, Shyamal Anadkat, Alexander Hoyle, Philip Resnik. “The Prompt Report: A Systematic Survey of Prompting Techniques”. In: *arXiv preprint arXiv:2406.06608* (2024). DOI: 10.48550/arXiv.2406.06608.
- [45] SecurePrivacy. *Mental Health App Data Privacy: HIPAA-GDPR Hybrid Compliance*. Accessed 2026-03-02. 2025. URL: <https://secureprivacy.ai/blog/mental-health-app-data-privacy-hipaa-gdpr-compliance>.
- [46] Narotam Singh. *On-Device Therapy: Edge AI for Private Mental Health*. Accessed 2026-03-02. 2025. URL: <https://www.linkedin.com/pulse/on-device-therapy-edge-ai-private-mental-health-narotam-singh-fbbrc>.
- [47] Xubin Wang, Zhiqing Tang, Jianxiong Guo, Tianhui Meng, Chenhao Wang, Tian Wang, Weijia Jia. “Empowering Edge Intelligence: A Comprehensive Survey on On-Device AI Models”. In: *ACM Comput. Surv.* 57.9 (2025). DOI: 10.1145/3724420.
- [48] Taylor Wessing. *GDPR Compliance for Digital Health Apps*. Accessed 2026-03-14. 2023. URL: <https://www.taylorwessing.com/en/insights-and-events/insights/2021/04/dsgvo-compliance-bei-digital-health-apps>.
- [49] Emily Widnall, Claire Ellen Grant, Tao Wang, Laura Cross, Sumithra Velupillai, Angus Roberts, Robert Stewart, Emily Simonoff, Johnny Downs. “User Perspectives of Mood-Monitoring Apps Available to Young People: Qualitative Content Analysis”. In: *JMIR Mhealth Uhealth* 8.10 (2020), e18140. DOI: 10.2196/18140.

- [50] Sarah Wittmar, Tom Frankenstein, Vincent Timm, Peter Frei, Nicolas Kurpiers, Stefan Wölwer, Axel Georg Meender Schäfer. “User Experience With a Personalized mHealth Service for Physical Activity Promotion in University Students: Mixed Methods Study”. In: *JMIR Form Res* 9 (2025), e64384. DOI: 10.2196/64384.
- [51] Enhui Yang, Feng Chen, Yang Yang, Yuansi Zhang, Hao Lin, Yhang Zu, Maoping Chu. “Global trends in depressive disorder prevalence and DALYs among young populations: a comprehensive analysis from 1990 to 2021.” In: *BMC Psychiatry* 24.1 (2024), p. 943. DOI: 10.1186/s12888-024-06419-2.

All links were last followed on 2026-06-15.

A. Full Survey Questionnaires

The following pages show the full questionnaires used in the study.

Participant Code

Enter your Participation Code:

Demographic Questions

What is your Gender?

- Male
- Female
- Non-binary / third gender
- Prefer not to say

Enter your Age:

What do you do professionally?

- (University) Student
- Employee
- Self-employed
- Unemployed
- Retired
- Other:

Mood tracking applications Indicators

Do you consider yourself a self-reflective Person?

- No
- Yes

Do you enjoy self-reflection?

- No, never
- Mostly not
- Depends on the situation and what I reflect about
- Mostly yes
- Yes, always

Have you ever tracked your mood before?

- No, I have never
- Yes, I have tracked my mood in the past
- Yes, I am currently tracking my mood
- Yes, I have tracked my mood in the past and I am currently tracking my mood

Do you feel that mood tracking has helped you understand your emotions better?

- No
- Yes

How do/did you track your mood? (Select all that apply)

- Smartphone App
- Paper Diary
- Digital Journals / Notes App
- Voice Recordings
- Other

Do you regularly use a music app?

- No

Yes

What App(s) do you use?

Do you regularly use a health or fitness App?

No

Yes

What Health App(s) do you use?

Factors influencing the Mood

Which factors do you believe have an influence on your mood? (Please select all that apply)

Social Interactions

- Music
- Phone Usage
- Weather
- Sleep
- Physical Activities
- Daily Schedule
- Work
- Hobbies
- News
- University
- Other

Powered by Qualtrics

Participant Code

The questionnaire will approximately take 5 minutes to complete.

Please enter the participation Code you received:

Correlations and Recommendations - Week 1

The app has provided you with some information about correlations between mood and contextual data, followed by some recommendations on how to improve your mood.

First, please indicate your agreement for the different statements about the correlations in the first week of using the app.

Overall, how satisfied are you with the information provided by the app about the correlations between influencing factors and your mood?

Not at all satisfied (1)

Slightly satisfied (2)

Moderately satisfied (3)

Very Satisfied (4)

Completely satisfied (5)

How personalised did you find the information?

Not at all personalised (1)

Slightly personalised (2)

Moderately personalised (3)

Very personalised (4)

Completely personalised (5)

Has the information helped you to understand your mood?

Never (1)

Rarely (2)

Sometimes (3)

Often (4)

Always (5)

Has the system discovered correlations about your mood that you hadn't noticed before?

- No
- Yes, sometimes
- Yes, always

Do you think these new correlations make sense in terms of your mood?

No Yes

Now, please indicate your agreement for the different statements about the recommendations in the first week of using the app.

Overall, how satisfied are you with the recommendations provided by the app?

Not at all satisfied (1)

Slightly satisfied (2)

Moderately satisfied (3)

Very satisfied (4)

Completely satisfied (5)

How personalised did you think the recommendations were?

Not at all personalised (1)

Slightly personalised (2)

Moderately personalised (3)

Very personalised (4)

Completely satisfied (5)

How often have you tried the app's recommendations actions?

Never (1)

Rarely (2)

Sometimes (3)

Often (4)

Always (5)

Have the suggestions helped to improve your mood?

Never (1)

Rarely (2)

Sometimes (3)

Often (4)

Always (5)



Can you explain why you have never or rarely tried the recommendations?

Lack of Time: "Haven't found the time."

Forgotten: "Forgot to try the recommendations."

- Lack of Interest: "Recommendations didn't appeal to me."
- Doubt: "I was unsure if the Recommendations would help me."
- Difficulty: "I had trouble doing the Recommendations."
- Accessibility: "I didn't have access to resources mentioned in the Recommendations."
- Results: "I tried some recommendations but didn't notice any positive results."
- Understanding: "I didn't fully understand how to implement the Recommendations."
- Other Reasons

Powered by Qualtrics

Participant Code

The questionnaire will approximately take 10 minutes to complete.

Please enter the participation Code you received:

Correlations and Recommendations - Week 2

The app has provided you with some information about correlations between mood and contextual data, followed by some recommendations on how to improve your mood.

First, please indicate your agreement for the different statements about the correlations in the second week of using the app.

Overall, how satisfied are you with the information provided by the app about the correlations between influencing factors and your mood?

Not at all satisfied (1)	Slightly satisfied (2)	Moderately satisfied (3)	Very Satisfied (4)	Completely satisfied (5)
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

How personalised did you find the information?

Not at all personalised (1)	Slightly personalised (2)	Moderately personalised (3)	Very personalised (4)	Completely personalised (5)
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Has the information helped you to understand your mood?

Never (1)	Rarely (2)	Sometimes (3)	Often (4)	Always (5)
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Has the system discovered correlations about your mood that you hadn't noticed before?

- No
- Yes, sometimes
- Yes, always

Do you think these new correlations make sense in terms of your mood?

No

Yes

Now, please indicate your agreement for the different statements about the recommendations in the second week of using the app.

Overall, how satisfied are you with the recommendations provided by the app?

Not at all satisfied (1)

Slightly satisfied (2)

Moderately satisfied (3)

Very satisfied (4)

Completely satisfied (5)

How personalised did you think the recommendations were?

Not at all personalised (1)

Slightly personalised (2)

Moderately personalised (3)

Very personalised (4)

Completely satisfied (5)

How often have you tried the app's recommendations actions?

Never (1)

Rarely (2)

Sometimes (3)

Often (4)

Always (5)

Have the suggestions helped to improve your mood?

Never (1)

Rarely (2)

Sometimes (3)

Often (4)

Always (5)



Can you explain why you have never or rarely tried the recommendations?

Lack of Time: "Haven't found the time."

Forgotten: "Forgot to try the recommendations."

- Lack of Interest: "Recommendations didn't appeal to me."
- Doubt: "I was unsure if the Recommendations would help me."
- Difficulty: "I had trouble doing the Recommendations."
- Accessibility: "I didn't have access to resources mentioned in the Recommendations."
- Results: "I tried some recommendations but didn't notice any positive results."
- Understanding: "I didn't fully understand how to implement the Recommendations."
- Other Reasons

Summary of entire app Usage

Please indicate your agreement with the individual statements as a summary of the entire app usage.

Strongly disagree Disagree Somewhat disagree Undecided Somewhat agree Agree Strongly agree

Using the system has led to a wake-up call to make changes in my life

Strongly disagree Disagree Somewhat disagree Undecided Somewhat agree Agree Strongly agree

As a result of using the system, I have changed how I approach things

Using the system gives me ideas on how to overcome challenges

I enjoy exploring my data with the system

The system makes it easy to get an overview of my personal data

The system makes it easy to review my long-term personal data

Strongly disagree Disagree Somewhat disagree Undecided Somewhat agree Agree Strongly agree

I reflect on my data in the system with others

The system helps me discuss my data with others

The system makes me think about how my personal data relates with that of others

Are there any improvements you would suggest for the app?

SUS

Please rate the following statements about the app in general on a scale from 1 (strongly disagree) to 5 (strongly agree)

	Strongly disagree (1)	(2)	(3)	(4)	Strongly agree (5)
1. I think that I would like to use this system frequently.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
2. I found the system unnecessarily complex.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
3. I thought the system was easy to use.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
4. I think that I would need the support of a technical person to be able to use this system.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
5. I found the various functions in this system were well integrated.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
6. I thought there was too much inconsistency in this system.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
7. I would imagine that most people would learn to use this system very quickly.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
8. I found the system very cumbersome to use.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

	Strongly disagree (1)	(2)	(3)	(4)	Strongly agree (5)
9. I felt very confident using the system.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
10. I needed to learn a lot of things before I could get going with this system.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

UEQ

This questionnaire consists of pairs of contrasting attributes that may apply to the app. The circles between the attributes represent gradations between the opposites. Please express your impression by ticking the circle that most closely reflects your opinion. Sometimes you may not be completely sure about your agreement with a particular attribute, or you may feel that it does not fully apply. Nevertheless, please tick one circle in every line. Decide spontaneously and do not think too long about each item to reflect your first impression.

annoying	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	enjoyable
not understandable	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	understandable
creative	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	dull
easy to learn	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	difficult to learn
valuable	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	inferior

boring	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	exciting
not interesting	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	interesting
unpredictable	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	predictable
fast	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	slow
inventive	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	conventional
obstructive	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	supportive
good	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	bad
complicated	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	easy
unlikable	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	pleasing
usual	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	leading edge
unpleasant	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	pleasant
secure	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	not secure
motivating	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	demotivating
meets expectations	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	does not meet expectations
inefficient	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	efficient
clear	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	confusing
impractical	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	practical
organized	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	cluttered
attractive	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	unattractive
friendly	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	unfriendly
conservative	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	innovative

B. Informed Consent Form

The following pages show the informed consent form that participants had to sign before taking part in the study.

Informed Consent of Participation

You are invited to participate in the field study **Design and Implementation of a Well-Being App**, initiated and conducted by Dennis Pfeiffer, and Prof. Dr. Sven Mayer. The research is supervised by Prof. Dr. Sven Mayer at TU Dortmund University under . This study is funded and financed by None.

Please note:

- Your participation is voluntary.
- The field study will last approximately 2 Weeks (5 minutes per day).
- We will record personal demographics (age, gender, and profession).
- We will record interaction data (e.g., clicks, timings) during the session.
- You will receive no compensation.
- We may publish results from this and other sessions.
- All data you provide in this field study will be published anonymized; thus, they will neither include your name nor cannot be associated with your identity.

If you have any questions about the whole informed consent process of this research or your rights as a human research subject, please contact our ethics committee office: None. You should carefully read the settings below. You may take as much time as you need to read the consent form. If you do not fully agree with something, or if your questions have not all been answered to your satisfaction, then you should not give your consent.

1. Purpose and Goal of this Research

The purpose of this study is to investigate how a mood-tracking application supported by a locally hosted AI model can help users reflect on and improve their well-being. The goal of this research is to find out whether the technology can improve user's Well-Being and understanding of it. Your participation will help us achieve this goal. The results of this research may be presented at scientific or professional meetings or published in scientific proceedings and journals.

2. Participation and Compensation

Your participation in this field study is voluntary. You will be one of approximately 10 people being surveyed for this research. You will receive no compensation. You may withdraw and discontinue participation at any time. If possible, you may refuse to answer any questions you do not want to answer or withdraw from participation at any time.

At any time and without giving any reason, you can notify us that you want to withdraw the consent given (GDPR Art. 21). In case of withdrawal, your data stored based on your consent will be deleted or anonymized where this is legally permissible (GDPR Art. 17). If deletion is impossible or only possible with unreasonable technical effort, your data will be anonymized by deleting the personal identification information. However, anonymization of your data cannot entirely exclude the possibility of subsequent tracing of information to you via other sources. Finally, once the data is anonymized, the deletion of the data is not possible anymore as we will not be able to identify which data is yours.

3. Procedure

After giving consent, you will be guided through the following steps:

You will be asked to complete a questionnaire about some demographic data and your experience with mood tracking (prior knowledge is not required). Then the app will be sent to you and you can use it. On the 8th day, you will receive an interim questionnaire asking you how you felt using the app during the first week. On the 15th day, you will receive the final questionnaire, in which you can reflect on your use of the app after the second week.

The complete procedure of this field study will last approximately 2 Weeks (5 minutes per day).

4. Risks and Benefits

There are no risks associated with this field study. Discomforts or inconveniences will be minor and are unlikely to happen. If you feel uncomfortable, you may discontinue your participation. (You will not directly benefit through participation in this field study. We hope that the settings obtained from your participation may help to bring forward the research in this field.) With this research, we will advance knowledge in this research field.

5. Data Protection and Confidentiality

The General Data Protection Regulation (GDPR) of the European Union (EU) governs that data collection process. The legal basis for processing the personal data is the consent in accordance with GDPR Art. 6 (1). The GDPR guarantees a set of right to the data subjects, including the right to access, rectification, and erasure of personal data.

- You have the right to access your personal data at any time (GDPR Art. 15).
- You have the right to correct inaccurate personal data at any time (GDPR Art. 16).
- You have the right to have your personal data deleted (GDPR Art. 17).
- You have the right to limit the processing of your personal data (GDPR Art. 18).
- You have the right to have your data transferred to others (GDPR Art. 20).
- You have the right to withdraw the consent given (GDPR Art. 21).

If you wish to exercise any of your rights, please contact the researchers.

We will record personal data (age, gender, and profession) during participation. Researchers will not identify you by your real name in any reports using settings obtained from this field study, and your confidentiality as a participant in this field study will remain secure. Data collected in this field study will be treated in compliance with the GDPR.

We will record personal demographics, and interaction logs during the field study. All data you provide in this field study will not be published and kept confidential.

Your non-anonymized data will be stored for a minimum of 10 years from the time your consent is given, unless you withdraw your consent before this period has elapsed. Your non-anonymized data will be stored in a secure location and will be accessible only to the researchers involved in this work.

Anonymized data collected can be shared publicly. Data collected that have not been made public will be deleted after the end of the research.

As with any publication or online-related activity, the risk of a breach of confidentiality is always possible. According to the GDPR, the researchers will inform the participant if a breach of confidential data is detected.

6. Identification of Investigators

If you have any questions or concerns about the research, please feel free to contact:

Dennis Pfeiffer (dennis.pfeiffer@tu-dortmund.de)

Principal Investigator: Prof. Dr. Sven Mayer (Joseph-von-Fraunhofer-Straße 25, 44227 Dortmund, sven.mayer@tu-dortmund.de.de)

7. Informed Consent and Agreement

This consent form will be retained securely and in compliance with the GDPR for no longer than necessary.

- I understand the explanation provided to me. I have been given a copy of this form. I have had all my questions answered to my satisfaction, and I voluntarily agree to participate in this field study.
- I voluntarily consent to my data being recorded and subsequently processed in line with the GDPR. I have been informed about the consequences of withdrawing my consent.

Printed Name of Subject

Signature of Subject

Location, Date

C. Additional Visualizations

C.1. Age Distribution of Participants

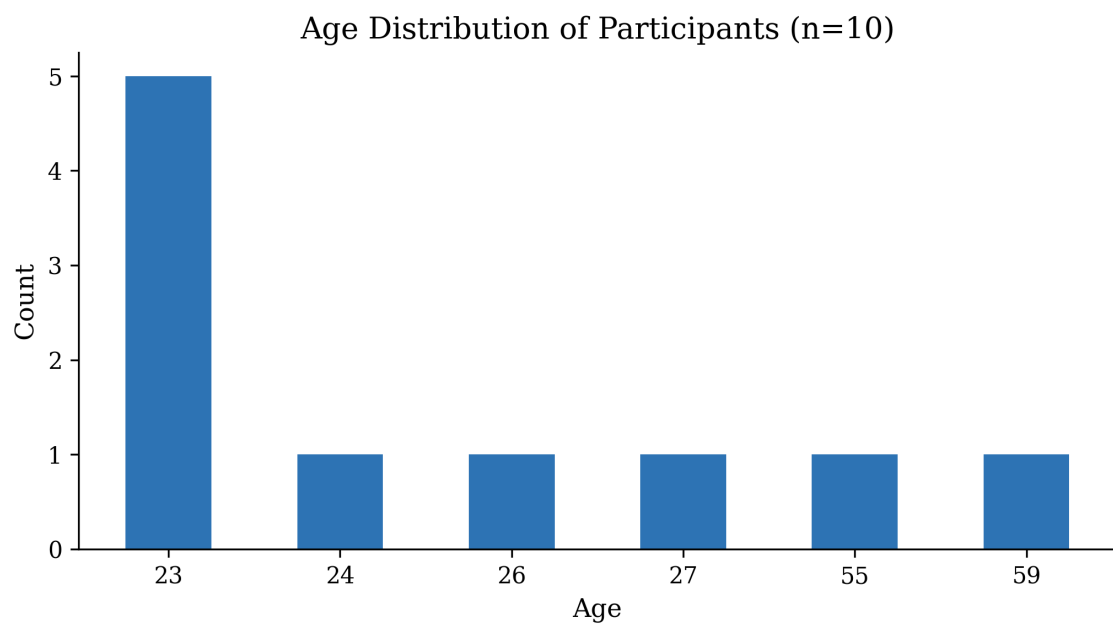


Figure C.1.: Age distribution of study participants.

C.2. Other Demographic Data

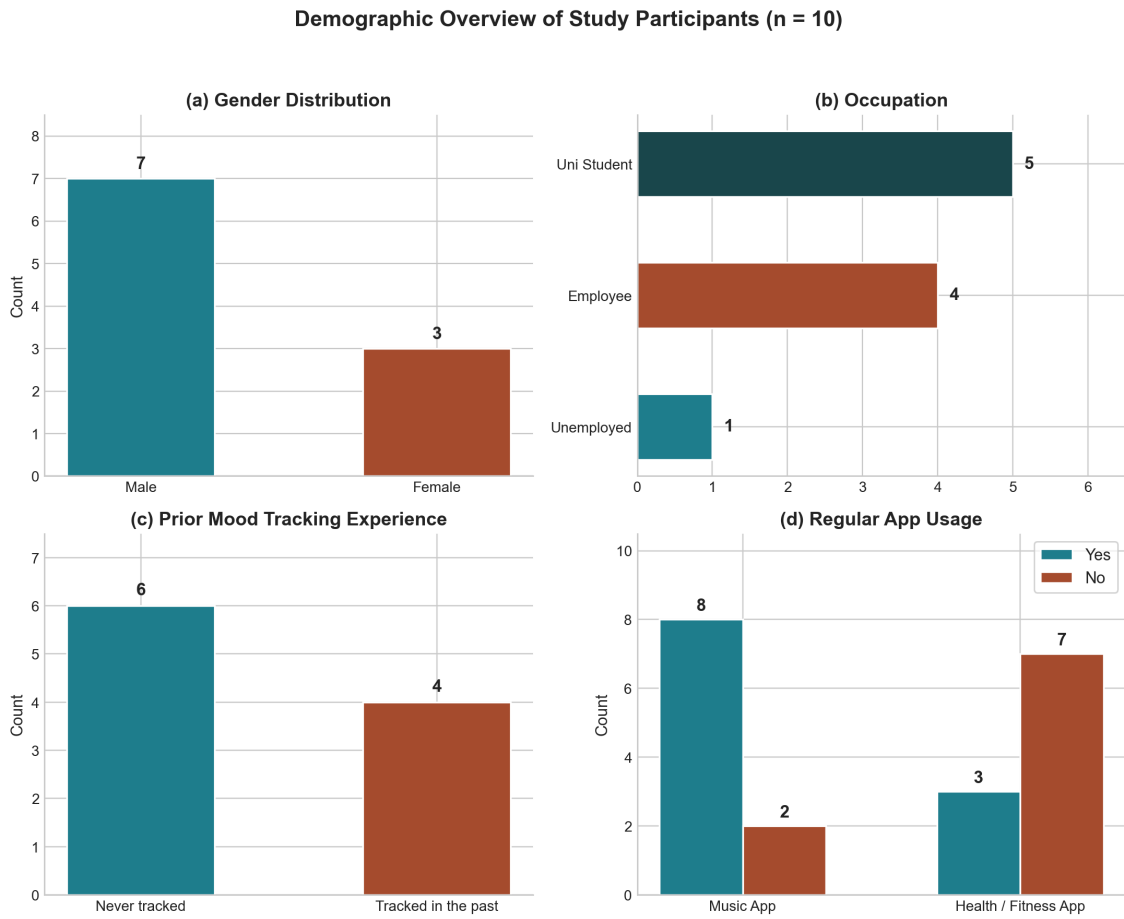


Figure C.2.: (a) Gender Distribution, (b) Occupation, (c) Prior Mood Tracking Experience and (d) Regular App Usage

D. Background

D.1. Wilcoxon Signed-Rank Tests

To explore whether the descriptive differences between the baseline and tailored conditions reached statistical significance, Wilcoxon signed-rank tests were conducted for all seven per-condition items. Given the small sample size ($n = 10$ paired observations), these tests have very low statistical power and are reported here as supplementary exploratory analyses rather than confirmatory tests.

Table D.1.: Wilcoxon signed-rank test results for baseline vs. tailored condition ($n = 10$ pairs, two-sided)

Item	M	M_B	M_T	Non-zero	W	p	r
Satisfaction (Correlations)	2.60	2.60	4	5.0	1.000	.00	
Personalisation (Correlations)	2.10	2.80	5	1.5	.102	.52	
Helped understand mood	2.80	2.90	4	4.0	.705	.12	
Satisfaction (Recommendations)	2.40	2.90	8	8.0	.132	.48	
Personalisation (Recommendations)	2.00	2.80	5	0.0	.039*	.65	
Frequency of trying recs	2.60	2.80	5	5.0	.480	.22	
Helped improve mood	2.00	2.40	7	7.5	.260	.36	

Note. M_B = baseline mean; M_T = tailored mean; Non-zero = number of non-zero differences; W = Wilcoxon test statistic; r = effect size estimate ($|Z|/\sqrt{N}$). * $p < .05$ (uncorrected).

As Table D.1 shows, one of the seven comparisons reached statistical significance at $\alpha = .05$: the “Personalisation of recommendations” item ($W = 0.0$, $p = .039$, $r = .65$), with a large effect size favouring the tailored condition. All five participants with non-zero differences rated the tailored condition higher on perceived personalisation. The “Personalisation of correlations” ($p = .102$, $r = .52$) and “Satisfaction with recommendations” ($p = .132$, $r = .48$) items showed medium-to-large effect sizes that did not reach significance, likely due to insufficient power. These results are consistent with the descriptive findings reported in Chapter 5 and suggest that the tailored condition’s advantage in perceived personalisation may represent a genuine effect that warrants validation in larger-scale studies. However, without correction for multiple comparisons, the single significant result should be interpreted with caution.

D.2. Ollama format Schema

```
OLLAMA_OUTPUT_SCHEMA = {
  "type": "object",
  "required": ["introduction", "suggestions"],
  "additionalProperties": False,
  "properties": {
```

D. Background

```
"introduction": {"type": "string"},
"suggestions": {
  "type": "object",
  "required": ["1", "2", "3"],
  "additionalProperties": False,
  "properties": {
    "1": {
      "type": "object",
      "required": ["type", "explanation"],
      "additionalProperties": False,
      "properties": {
        "type": {"type": "string"},
        "explanation": {"type": "string"}
      }
    },
    "2": {
      "type": "object",
      "required": ["type", "explanation"],
      "additionalProperties": False,
      "properties": {
        "type": {"type": "string"},
        "explanation": {"type": "string"}
      }
    },
    "3": {
      "type": "object",
      "required": ["type", "explanation"],
      "additionalProperties": False,
      "properties": {
        "type": {"type": "string"},
        "explanation": {"type": "string"}
      }
    }
  }
}
```

Listing D.1: Ollama Recommendation format schema

D.3. Cleaned Payload Example

```
{
  "moods": [
    {
      "mood": "Tense",
      "context": ["Sleep", "University"],
      "note": "Not enough sleep",
      "metrics": {
        "temp_c": 9.92,
        "weather": "Rain"
      }
    }
  ]
}
```

```
    }
  },
  {
    "mood": "Relaxed",
    "context": ["Physical Activities", "Social Interactions"],
    "note": "Great talks with friends",
    "metrics": {
      "steps": 9625,
      "temp_c": 11.59,
      "weather": "Clouds"
    },
    "social": {
      "type": "In Person",
      "amount": "1 to 1"
    }
  },
  {
    "mood": "Tense",
    "context": ["Sleep", "Other"],
    "note": "Slept poorly and have a headache",
    "metrics": {
      "steps": 243,
      "temp_c": 13.81,
      "weather": "Clouds"
    }
  },
  {
    "mood": "Content",
    "context": ["Weather", "Physical Activities"],
    "note": "Gym and Sun out",
    "metrics": {
      "steps": 6861,
      "temp_c": 18.25,
      "weather": "Clear"
    }
  },
  {
    "mood": "Happy",
    "context": ["Hobbys"],
    "note": "Cooked a delicious Meal today",
    "metrics": {
      "steps": 206,
      "temp_c": 13.81,
      "weather": "Clouds"
    }
  }
],
"recaps": [
  {
    "mood": "Calm",
    "context": ["Sleep", "Hobbys"],
    "note": "Wrote",
    "metrics": {
```

D. Background

```
    "steps": 16,
    "temp_c": 13.81,
    "weather": "Clouds"
  }
},
{
  "mood": "Depressed",
  "context": ["Social Interactions", "Other"],
  "socialLoad": "Isolated",
  "note": "Lonely",
  "metrics": {
    "steps": 84,
    "temp_c": 10.46,
    "weather": "Clouds"
  }
}
]
}
```

Listing D.2: Example of cleaned payload sent to Ollama

D.4. Example Baseline Recommendation

```
{
  "introduction": "The weather affects mood individually. Try to find out if and how it affects you and which weather conditions have a positive or negative impact on your mood. Leisure activities can improve resilience and help you deal with stress. Regular exercise also has a positive effect on physical and mental health. Try to incorporate these activities into your daily routine. Music triggers positive processes in the brain that can promote well-being and reduce stress. Find out what kind of music helps you in certain situations.",
  "suggestions": {
    "1": {
      "type": "Meditation",
      "explanation": "Start the day with a short meditation session to set a positive tone and increase focus throughout the morning. Try using a guided meditation app or simply focusing on your breath."
    },
    "2": {
      "type": "Movies",
      "explanation": "Plan a movie night at home with friends or family, where you can relax and enjoy each other's company without distractions. You can also invite acquaintances to join in on the fun."
    },
    "3": {
      "type": "Hobbies",
      "explanation": "Take some time to explore your hobbies and interests. Whether it's painting, cooking, or playing a sport, make sure you have some activities lined up to enjoy and learn something new."
    }
  },
  "_meta": {
```

```

    "generation_time_seconds": 53.54,
    "model": "llama3.2:1b",
    "is_baseline": true
  }
}

```

Listing D.3: Example Baseline recommendation JSON output

D.5. Example Tailored Recommendation

```

{
  "introduction": "You've recently experienced some tense moods, and this analysis focuses on helping you understand possible patterns and ways to support your wellbeing.",
  "suggestions": {
    "1": {
      "type": "General Wellbeing",
      "explanation": "You sometimes seem to rely on physical activities like going to the gym or cooking as a way to handle stress, which can be very helpful. At the same time, consistently short or poor sleep might be contributing to feeling tense. It could be worth experimenting with a regular bedtime, winding-down routine, and aiming for 7--8 hours of sleep to see if your tension eases."
    },
    "2": {
      "type": "Social Connection",
      "explanation": "Moments like great talks with friends appear alongside more positive or relaxed moods. In contrast, feelings of loneliness or isolation might coincide with more difficult moods. It might help to plan small, regular check-ins or activities with people you trust, so social contact feels more steady rather than only happening in peaks."
    },
    "3": {
      "type": "Self-Care",
      "explanation": "You report feeling better during activities such as cooking, going to the gym, or engaging in hobbies, which suggests that active self-care could support your mood. You could explore building a simple self-care toolkit, for example short walks, writing, or a creative hobby you enjoy. Trying one of these when you notice tension or low mood might help you feel a bit more grounded, even if the underlying situation does not change immediately."
    }
  },
  "_meta": {
    "generation_time_seconds": 58.73,
    "model": "llama3.2:1b",
    "moods_count": 5,
    "recaps_count": 2
  }
}

```

Listing D.4: Example tailored recommendation JSON output for the cleaned payload above

D.6. Guiding Recommendations sent with every Baseline Recommendation Payload

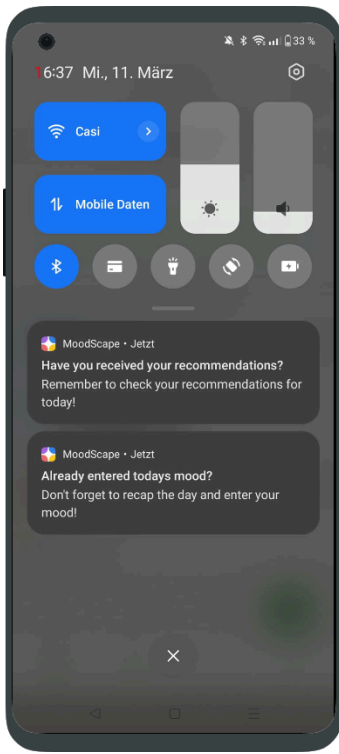
```
baselineRecommendationsENG = [  
  { text: "Meet a friend for a cup of coffee or an ice cream at home or nearby.", factors: "Social" },  
  { text: "Call a friend or family member and share a positive experience from today with each other.", factors: "Social" },  
  { text: "Check if there are any interesting events in your area that you can attend alone or with acquaintances.", factors: "Social" },  
  { text: "Take a break from the digital world and invite friends for a game night where you can play board and card games together.", factors: "Social, Analog, Games" },  
  { text: "Invite friends to a barbecue party where you can create recipe ideas together.", factors: "Social" },  
  { text: "Invite your friends and get creative together. Craft, pottery, or paint together.", factors: "Social, Hobbies" },  
  { text: "Host a movie night where you can relax and watch movies or series with friends/colleagues/family.", factors: "Social" },  
  { text: "Take a relaxing walk outside. If you have a forest or park nearby, these are great places for it.", factors: "Sport, Nature" },  
  { text: "Do a short fitness session with light exercises and listen to a suitable song for each exercise.", factors: "Sport, Music" },  
  { text: "Listen to some of your favorite songs and dance to them.", factors: "Sport, Music" },  
  { text: "Take your bike and explore nature on a relaxed bike tour.", factors: "Sport, Nature" },  
  { text: "Check if there is a course nearby where you can practice a sport with others that interests you.", factors: "Sport, Social" },  
  { text: "Look online for videos that guide you through fitness exercises.", factors: "Sport" },  
  { text: "Incorporate small workouts into your daily routine, e.g., heel raises while brushing your teeth, sit-ups while unloading the dishwasher.", factors: "Sport" },  
  { text: "Prepare small snacks and have a picnic in the park.", factors: "Social, Nature" },  
  { text: "Go for a swim. If the weather permits, preferably in an outdoor pool.", factors: "Sport, Nature, Social" },  
  { text: "Go out into nature and pay attention to your surroundings. You can collect small treasures from nature and take them home as a memory.", factors: "Nature" },  
  { text: "Experiment with cooking or baking and create your own recipes. Show them to your friends and family.", factors: "Hobbies, Social" },  
  { text: "Take a moment of mindfulness where you focus only on yourself.", factors: "Meditation" },  
  { text: "Go outside and look for unusual objects that you can photograph.", factors: "Nature" },  
  { text: "Take a moment to read a book or magazine that interests you.", factors: "Hobbies" },  
  { text: "Try to invent your own little song by humming a melody and thinking of lyrics.", factors: "Music" },  
  { text: "Play a mixed playlist and spontaneously paint a picture for each song.", factors: "Music, Hobbies" },  
  { text: "Host a disco night and invite friends.", factors: "Social, Music" },  
  { text: "Listen to calm music or relaxing sounds and have a short meditation session.", factors: "Music, Meditation" },  
]
```

D.6. Guiding Recommendations sent with every Baseline Recommendation Payload

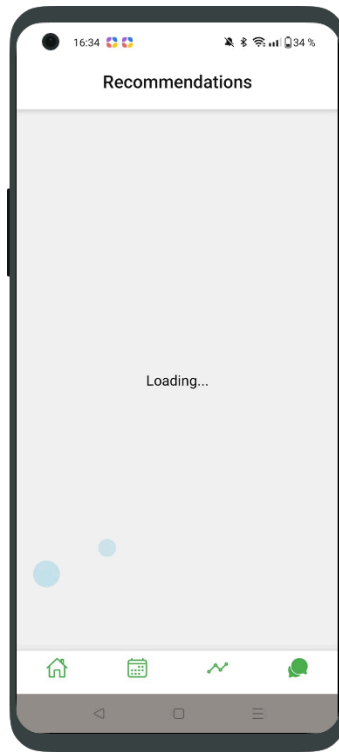
```
{ text: "Look for yoga videos that you can do with friends.", factors: "Social, Meditation"
},
{ text: "Find videos online that guide you through meditation.", factors: "Meditation" },
{ text: "Put your phone away and consciously take a moment away from the digital world.",
factors: "Meditation, Phone" },
{ text: "Make a plan for your upcoming tasks to go into the next days structured.", factors:
"Planning" },
{ text: "Try not to read any news for a day.", factors: "Phone, News" },
{ text: "Take a mental journey to a place where you would like to be right now and imagine
the environment as detailed as possible, including smells, colors, and sounds.", factors: "
Meditation" }
];
```

Listing D.5: Guiding Baseline Recommendation Examples

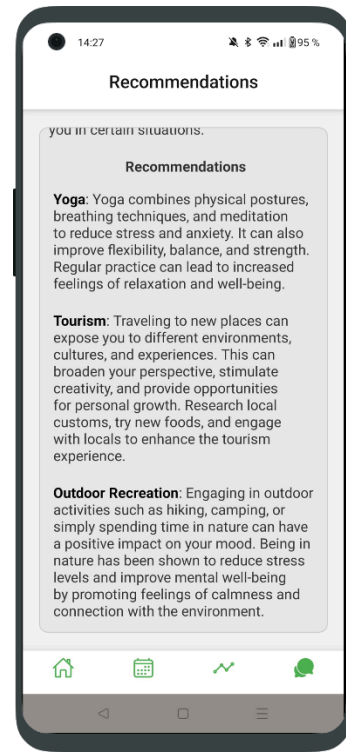
D.7. Additional *MoodScape* Screenshots



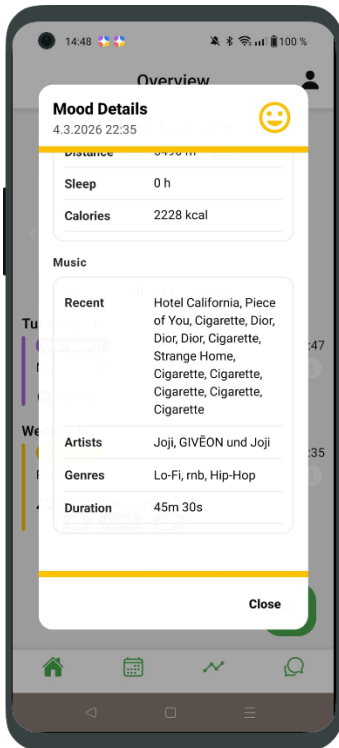
(a) Notifications sent out by *MoodScape*, when it is opened



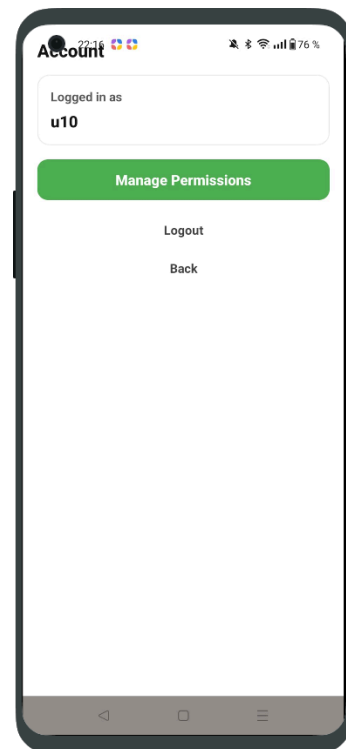
(b) Recommendation Loading Screen



(c) Example Screen for Base-line Recommendation



(d) Mood Detail Screen scrolled down



(e) Account screen 97

Figure D.1.: Additional *MoodScape* screens.

Used Aids

During the preparation of this work, the author used Perplexity, Claude in order to: Grammar and spelling check, Paraphrase and reword, Drafting content, Generate images, Content enhancement, Plagiarism detection. Perplexity was used to generate python figures in the Results Chapter; Claude Code was used to support parts of the Android app implementation. After using this tool/service, the author reviewed and edited the content as needed and take(s) full responsibility for the thesis's content.

place, date, signature