Narrating Fitness: Leveraging Large Language Models for Reflective Fitness Tracker Data Interpretation

Narrating Fitness: Leveraging Large Language Models for Reflective Fitness Tracker Data Interpretation

Konstantin R. Strömel
kstroemel@uos.de
Osnabrück University, Institute of Cognitive Science
Osnabrück, Germany

Stanislas Henry
shenry005@bordeaux-inp.fr
ENSEIRB-MATMECA Bordeaux
Bordeaux, France

Tim Johansson
tijohans@chalmers.se
Chalmers University of Technology
Gothenburg, Sweden

Jasmin Niess
jasminni@ifi.uio.no
University of Oslo
Oslo, Norway

Paweł W. Woźniak
pawelw@chalmers.se
Chalmers University of Technology
Gothenburg, Sweden
TU Wien
Vienna, Austria

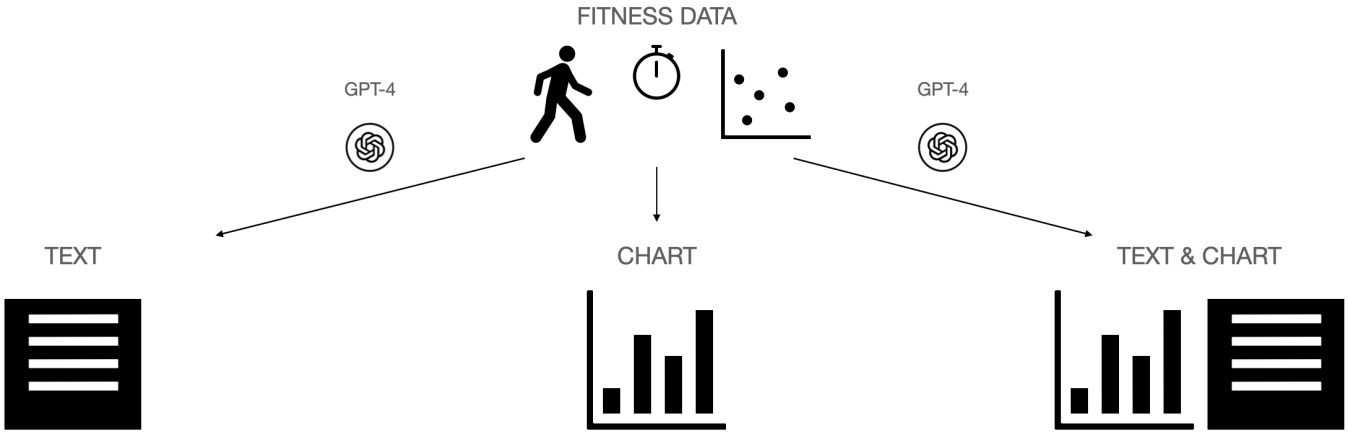


Figure 1: In this study, we examine how different data representations support reflection in personal informatics. We compare text generated with a Large Language Model, a standard chart and the combination of both. We found that the text fostered reflection and engagement.

ABSTRACT

While fitness trackers generate and present quantitative data, past research suggests that users often conceptualise their wellbeing in qualitative terms. This discrepancy between numeric data and personal wellbeing perception may limit the effectiveness of personal informatics tools in encouraging meaningful engagement with one's wellbeing. In this work, we aim to bridge the gap between raw numeric metrics and users' qualitative perceptions of wellbeing. In an online survey with $n = 273$ participants, we used step data from fitness trackers and compared three presentation formats: standard charts, qualitative descriptions generated by an LLM (Large Language Model), and a combination of both. Our findings reveal that users experienced more reflection, focused attention and reward when presented with the generated qualitative data compared to the standard charts alone. Our work demonstrates how automatically generated data descriptions can effectively complement numeric fitness data, fostering a richer, more reflective engagement with personal wellbeing information.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**.

KEYWORDS

personal informatics, generative AI, fitness trackers, reflection

ACM Reference Format:
Konstantin R. Strömel, Stanislas Henry, Tim Johansson, Jasmin Niess, and Paweł W. Woźniak. 2024. Narrating Fitness: Leveraging Large Language Models for Reflective Fitness Tracker Data Interpretation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24),*

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642032

May 11–16, 2024, Honolulu, HI, USA. ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3613904.3642032

1 INTRODUCTION

In a world that is increasingly quantified [42], Personal Informatics (PI) offers an alluring promise: to give us actionable insights into our lives, from our health to our habits [21]. However, amid this data-driven revolution, the qualitative aspects of wellbeing are often overlooked [10]. For many, reflecting on one's wellbeing is not merely an academic exercise but a gateway to genuine understanding and meaningful change [7]. Yet, conventional PI systems remain overwhelmingly goal- and performance-oriented, reducing the multi-dimensional facets of our lives to mere numbers [21]. This quantitative focus inadvertently sidelines holistic wellbeing, an aspect that should be at the heart of PI [57]. Consequently, it remains a challenge for Human-Computer Interaction (HCI) to study how we can design PI systems that focus less on numbers and encourage reflection without a focus on quantifying performance. This challenge is particularly timely as recent technical developments offer an accessible tool for generating qualitative feedback—Large Language Models (LLMs). LLMs offer new potential for textual feedback in PI systems, surpassing scripted text with their adaptability and ability to generate novel, dynamic narratives.

While a considerable body of work in PI and Artificial Intelligence (AI) has primarily focused on predictive and recommendation systems [18, 41], our approach takes a different path by emphasising the role of reflection in understanding personal fitness data. Traditional PI and AI interventions aim to reduce the cognitive load of individuals and provide prescriptive insights for immediate or future actions [24]. This trend is also recently visible in commercial systems, such as dedicated health recommendation models[1]. Such solutions are affected by a risk of users losing agency. Further, many users are reluctant to acknowledge recommendations from a PI system unless they are perfectly tailored to their life context [13]. In contrast to past work, our study explores how AI can be used to foster tailor-made reflection for wellbeing. We leverage Large Language Models (LLM) to create qualitative output that fosters reflection rather than just offer recommendations or predictions. This method aims to facilitate a nuanced understanding of one's fitness data, allowing for a more informed interaction with one's health metrics. Our work is driven by the following research question: *How can we use LLMs to encourage engagement with and facilitate reflection on fitness tracker data?*.

To that end, we first conducted interviews with $n = 10$ participants to explore attitudes towards AI-generated wellbeing narratives. Armed with these insights, we designed and implemented a dedicated study platform where users could donate their fitness data (steps) for an online experiment. Our experiment, with $n = 273$ participants, used three conditions, which corresponded to three types of tracker data representation: data presented as text-only, as a standard chart, or as a combination of text and chart. We measured the reflection and user experience fostered under the three conditions. Our findings suggest that text-based descriptions fostered deeper reflection through comparison, commanded more focused attention from the users, and made the users feel that the act of viewing personal data was worth the effort.

We offer three contributions to the HCI community: (1) an interview pre-study that uncovers how users feel about transforming fitness data into stories and their receptivity to automatically generated descriptions; (2) an online experiment conducted on a dedicated platform to assess how these AI-generated narratives influence reflection on and engagement with the users' personal data; and (3) insights on the potential of generative AI to improve the quality of reflection in PI systems. The rest of this paper is structured as follows. We begin by surveying relevant literature in areas such as reflection in PI, the role of storytelling in fitness tracking, and the applications of generative AI in HCI. We then detail our methodological approach, including our interview pre-study and online experiment, before discussing the outcomes and their implications. We conclude by outlining avenues for future research that can build upon our findings.

2 RELATED WORK

Here, we first contextualise our inquiry within work on personal informatics. We then review past research efforts in fostering data-driven reflection and show how past work on using narratives for wellbeing inspires our investigation.

2.1 Personal Informatics and its Qualitative Aspect

Understanding the intricacies of PI is an established research pursuit in HCI [21]. The field has built a number of models of how a user progresses through a PI experience. Li et al.'s [37] model, later extended by Epstein et al. [22], postulates that PI is an ever-evolving practice where users are often tasked with making informed decisions regarding their routines and behaviours. Niess and Wozniak [47] showed that goal setting is a key activity in which users aim to translate internal, qualitative goals (e.g. a healthier lifestyle) into targets that can be implemented with tracking technology (e.g. a step goal). Agapie et al. [2] extended this approach by charting how users related goals of different complexity to their varying needs in mental health contexts. A common feature of most models of PI in HCI, is the need to understand the connection between the number-driven digital world of tracking technologies and the qualitative world of needs and feelings associated with the users' wellbeing. Such a connection facilitates reflection, which is seen as a key step in personal informatics [7]. Providing users with the means to navigate that connection is seen as a key design goal for PI systems [47, 54]. In this work, we aim to address this by investigating how qualitative feedback can contribute to helping users translate numeric tracking data into reflective insight.

Yet, the role of the qualitative aspect of PI is not limited to reflection. Multiple works have shown that one's PI journey is inherently complex and there is variety in how users navigate their tracking. For instance, Rooksby et al. [57] highlighted distinct tracking styles that represent varied data-driven preferences of users. Tang and Kay [65] studied data habits of long-term tracker users, discovering that users value insights on their compliance and can introspect on their tracker usage patterns by relating them to life events. This was echoed in Elsden et al.'s [20] work about reviewing one's

[1] https://tryterra.co/products/odinai

past tracking data where participants reported creating meaning in discovering how different data points described their subjective perception of past events. These examples show that, despite the large diversity among users who engage in PI, the lived experience of the tracked activities is primarily qualitative. These findings motivate our work, which aims to use computational tools to build better bridges between the quantitative data which trackers generate and the qualitative terms in which users think of wellbeing.

2.2 Fostering Reflection in Personal Informatics

Reflection is regarded as key element in the success of PI systems [6, 10]. This has resulted in a number of research efforts focused on the design of system which aim to provide users with insight into their wellbeing by analysing their personal data. Yet, past critiques of personal informatics systems have pointed out their passive approach to prompting reflection [14, 34]. Baumer [5] highlighted that these platforms often presume that merely presenting users with past data visualisations will automatically spur reflection. This notion contradicts theories of reflection that underscore the need to actively foster reflective practices, as it does not arise unprompted [62]. To explore the reflection component of the personal informatics journey, Bentvelzen et al. [7] developed the *Technology-Mediated Reflection Model (TMRM)*, which emphasises the need to understand PI as an activity where users need constant improvement and support for changing perspectives. These findings show a constant need for creating new systems that support reflection by offering the users new ways to analyse their data. Here, we investigate how tools that generate narratives can effectively support reflection.

Past research explored many design directions in fostering reflection. One strong theme was physicalisation. Systems like LOOP [59] or SweatAtoms [30] used physical artefacts that prompted users to think about their data. Another trend was supporting journaling as an in-depth PI practice. Ayobi et al. [4] showed how users reflected on data in analogue bullet journals. Journaling is also a commonly supported reflection method in commercial systems [10]. As HCI studied more and more systems for reflection, results also showed possible negative aspects of such solutions. Niess et al. [46] reported that commonly used fitness tracker visualisations may trigger negative thought cycles, i.e. rumination. Eikey et al. [19] further studied this phenomenon in the context of diet tracking and found that users were in danger of rumination. As a consequence, avoiding negative thought cycles is a key design concern for systems which support reflection. This paper is interestingly different from past work in supporting reflection as it uses LLM-powered narratives to engage users with their data. Further, recognising recent studies on the negative aspects of PI, we investigate how reflective feedback can be designed to avoid rumination.

2.3 Narratives for Personal Wellbeing

There is past evidence that narratives can be effective in engaging users in reflective experiences [58]. Recent studies looked into the relationships between humans and conversational agents, aiming to foster self-reflection and promote wellbeing. E-coaching using scripted chatbots showed initial promise [11], yet these approaches were primarily focused on behaviour change and not reflection. Skjuve et al. [61] examined the development and impact of human-chatbot relationships. Their study, driven by the Social Penetration Theory, identified that initial engagements with chatbots, such as Replika, are superficial but evolve into deeper connections over time. They emphasised the significance of chatbot characteristics, like being accepting and non-judgmental, in the evolution of these relationships. Another study by Lee et al. [36] focused on the potential of chatbots to support deep self-disclosure among users. By leveraging the principle of reciprocity in human-machine dialogue, they demonstrated that chatbots can promote sustained self-disclosure and enhance user intimacy and enjoyment. While these results show that chatbots can be effective tools in encouraging thinking about one's data, starting a conversation with a chatbot is a different experience from current common forms of PI systems where users are presented with their data. Our work is inspired by work on chatbots and investigates if a narrative approach to data can be used in everyday fitness tracker apps or dashboards.

Another strain of work investigating reflective systems for emotional wellbeing has been gaining momentum. Kocielnik et al. [34] introduced the Reflection Companion, a mobile system that supports user reflection on personal sensed data, specifically on physical activity. Their system utilised mini-dialogues to provoke reflection, leading to heightened user motivation and behavioural change. On the other hand, Hollis et al. [26] delved into the implications of presenting and reflecting on mood data. Their interventions revealed that forecasting future moods, based on past data, can inspire users to take preventive actions against anticipated negative moods. Similarly, Desai et al. [18] developed a smartphone app called GlucOracle for individuals with type 2 diabetes. This app utilised self-tracking data to produce personalised forecasts for post-meal glucose levels, aiding users in making informed health choices. Murnane et al. [44] studied the design of data-driven narratives as a means to motivate healthy behaviour, emphasising the power of stories over conventional quantitative data representations. Their exploration of the WhoIsZuki application demonstrates the potential and efficacy of qualitative feedback mechanisms in promoting active lifestyles, which inspired us to compare different modes of fitness data presentation and their impact on user experience. Past research shows that AI-based predictions, narratives or recommendations may contribute to the PI experience. This is reflected in recent market trends such as the launch of Odin AI[2], an AI tool designed specifically for health recommendations based on fitness tracker data. Recently, Whoop launched a feature in collaboration with OpenAI where one can chat with a personal chatbot coach to discuss one's fitness data[3]. Our work seeks to answer this trend and empirically study if and how personalised narratives can contribute to the fitness tracking experience. To increase the ecological validity of our work, we focus on data from common consumer-grade fitness trackers.

3 METHOD

Our research began with preliminary interviews to understand initial attitudes towards narratives of fitness data. Alongside this, we

[2]https://tryterra.co/products/odinai
[3]https://www.whoop.com/us/en/thelocker/introducing-whoop-coach-powered-by-openai/

consulted existing literature, ensuring a well-informed foundation for designing a tailored GPT-4 prompt. The central component of our study was an interactive platform, developed to let users donate their step data and craft narratives around it. The platform's design emphasised data integrity. We focused on three primary modes of data presentation: text-only, chart-only, and a combination of both. The goal was to determine which mode supported reflection and engaged users the most. Evaluations utilised both quantitative metrics, assessing aspects like comprehension and preference, and qualitative data, such as an open-ended question to capture nuanced feedback. The combined insights from these evaluations formed the basis of our recommendations for future system designs. Figure 2 shows our research process. Next, we provide the details of each step in our study.

4 PRE-STUDY: INTERVIEWS

While related work shows the potential of narratives for fostering reflection in PI, the theoretical basis for generating narratives for tracking with LLMs is largely missing. Thus, to establish a starting point for our inquiry, we needed to expose users to LLM-generated fitness narratives. To this end, we conducted a series of exploratory interviews to understand the nuances of how users assess, understand, and talk about their personal fitness data. This step allowed us to frame our inquiry in line with user perceptions of qualitative feedback in PI. The interviews offered an opportunity to ascertain users' needs and expectations for the qualitative descriptions generated by LLMs. Through participant feedback on initial examples, we gained important insights into the tone, content, and general user experience expected from AI-generated narratives. The examples which we used were most likely erroneous—we could not provide any requirements to the model before establishing them—and served primarily as an invitation to a discussion. Information from the interviews served to inform our inquiry into how LLM prompts for qualitative feedback for fitness data could be designed. Given the controversial nature of users' acceptance of recommendations from PI systems in previous studies [26, 41], it was a key challenge to gauge users' willingness to engage in an LLM-based exploration of their fitness data.

4.1 Participants

We recruited $n = 10$ participants, aged $M = 27.7$, $SD = 3.8$. Seven participants identified as female and three as male. See table 1 for detailed information about each of the participants. We used social networks and snowball sampling to recruit a sample of users with diverse patterns of use of tracking technology. Interviewees ranged from those reporting no fitness tracking at all, over daily tracker users to participant, P07, who used two different tracking devices simultaneously. We recruited participants from diverse backgrounds, including arts (P03), engineering (P06), healthcare (P09) and computing (P10). This choice of participants enabled us to gain insights from users with different attitudes to tracking and different PI journeys.

4.2 Interview Protocol

In our interviews, we focused on four key areas. First, we asked each participant about their current fitness tracker usage to understand how often and for what activities they use their devices. Second, we delved into how the participant reviews and interprets their tracked data. Third, we invited the participant to tell their own wellbeing and fitness tracking story based on their data from the last week, which provided insights into how they understand and make sense of the information. Lastly, we gauged each participant's reactions to example stories we presented. This not only provided a tangible basis for the ensuing conversation but also ensured a personalised touch to every narrative. Such an approach ensured participants drew directly from their data, providing us with firsthand insight into their thought processes, the significance they attach to different metrics, and the narratives they constructed around their physical wellbeing. This approach helped us better understand the preferences and attitudes of the participants towards the narratives generated by our system and identify initial requirements for LLM-generated fitness data descriptions. Below we present one example of the automatically generated descriptions presented to the participants, the full stimulus is available in auxiliary material. We note that this example was intended as discussion starter, which we generated using a naive approach—by simply providing a Fitbit export of one week's worth of data into GPT-4 and asking the model to narrate it:

> *Throughout the week, you had a mix of active and less active days. On your most active day, you covered a remarkable distance, climbed numerous floors, and engaged in a significant amount of very active minutes. On the other hand, your least active day saw a shorter distance covered, fewer floors climbed, and less time spent in very active minutes. Overall, your week showcased a balance between sedentary time and various levels of physical activity. It's clear that you've made an effort to stay active and maintain a healthy lifestyle. (This quote is automatically generated text. It also does not represent narrative feedback generated according to the principles established in this paper.)*

Table 1: An overview of participants in the interview pre-study. We recruited participants with diverse tracker habits.

PID	Age	Gender	Occupation
P01	26	Female	Student
P02	33	Female	Academic teacher & Photographer
P03	24	Female	Student
P04	33	Male	PhD Candidate
P05	25	Male	PhD Candidate
P06	29	Male	Renewable Energy Consultant
P07	24	Female	Student
P08	27	Female	Researcher
P09	23	Female	Occupational Therapist
P10	33	Female	PhD Candidate

4.3 Data Analysis

Interviews were transcribed verbatim, resulting in a data set spanning a recording duration of 6 hours and 38 minutes. Two authors analysed the data using open coding in line with Blandford et

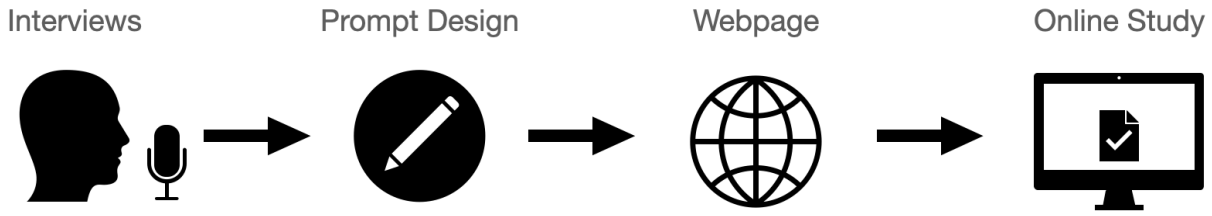


Figure 2: An overview of the research activities in this paper. We first conducted an interview pre-study. The results of the interviews informed our design for the prompt which generated narrative descriptions of fitness data. We then used the prompt to implement a web page which generated narratives based on the users' tracker data. The web page allowed us to conduct an online study of attitudes towards different data representations for tracker data: TEXT, CHART, and TEXT+CHART.

al. [12]. This initial phase of analysis involved two authors independently open-coding the data. Each author examined the transcripts meticulously, assigning preliminary codes to segments of the data without preconceived categories, ensuring an inductive approach to building themes. Subsequently, these authors met to discuss their independently derived codes, engaging in an interactive process to refine and consolidate the code groups. This collaborative approach facilitated the integration of diverse perspectives, enriching the analytical depth of the study. As a second step, two researchers employed affinity diagramming to identify patterns in the data. Through this iterative process, we constructed three themes which describe the content of the interviews: ABSTRACTION, TONE, and ROLE OF AI.

4.4 Findings

In the following section, we present the findings of the interviews with a focus on the comments about the example texts.

4.4.1 Abstraction. This theme highlights the distinct generalisations and lack of specificity in the language model's output. The balance between users' desire for personalised narratives and the model's inclination towards performance was a key aspect pondered by the participants:

> *What is also quite interesting to me is that this message also says: 'You climbed a decent number of floors, which is fantastic for your leg muscles and cardiovascular health.' If I would see that message I would kind of be like: Okay, but isn't like almost every single exercise good for your vascular health, you know? So I think in that sense, it's a little bit too general for me, because I'm kind of like okay, yeah, that's common sense. (P07)*

Several participants expressed discomfort with this ambiguity. They articulated a preference for narratives that provide more grounded insights, directly derived from their fitness data. Their feedback often centred on the AI's generalised approach and its inadequacy in providing tangible data interpretations. As expressed by Participant P06:

> *This is too vague. It's so general, that you can say that about any week. I'm missing the numerics that it says you were active on 2 out of 7 days and had 5 lazy days. (P06)*

Participants frequently expressed a desire for the narrative to incorporate highly specific context, delving into details often beyond the reach of conventional fitness tracker technologies. This underscores a gap between user expectations and the current capabilities of such devices in providing deeply contextualised insights:

> *But I think it's kind of low on information because it describes the most active day and then the least active day, just as the opposite of that. (...) Maybe it would be, (...) name what actually was done. Did this person go on a bike ride or run? Or, where did this person do it? Maybe the app can see: okay, you went to a forest for your run or you ran in the city? (...) a bit more nuanced information would be interesting. I feel like this doesn't tell that much. (P08)*

4.4.2 Tone. This theme describes the nature and manner of communication exhibited by the language model in conveying personal narratives. Participants' feedback indicated that the model's tone was perceived as patronising, superficial, and lacking in genuine human emotion. Participants highlighted their discomfort with the tone. P08, for instance, felt that the narrative was superficial and unnecessarily affirming, comparing its tone to that of addressing a child:

> *The tone... I find [it] a bit patronising. It sounds like it's someone speaking to a child or a dog. Being like: 'Great job, well done!' or 'It's okay, if you have an easy day' and 'listen to your body'. It feels a bit superficial, (...) patronising to me. If I would have that in an app, I would be annoyed by it. And I would be like, just give the information I could interpret myself and I don't need this blah blah around it. (P08)*

Participants saw that the narrative could potentially present an overly optimistic version of reality. P04's feedback underscored the sentiment of doubt due to the predominantly positive tone, suggesting the narrative could potentially fail to mention less favourable outcomes:

> *(...) it sounds very positive and good. But that would make me question, is it sweet talking the negative things? Like 'Oh, you didn't do something, but that's good because you took a break and so on.' (P04)*

Participants also expressed concerns regarding the inherent 'robotic' feel of the AI-generated content. Despite being structurally sound, the narrative felt devoid of human touch. It is evident from the feedback that participants desired a balance: the objective analysis of a machine, but with the warmth and nuance typical of human communication. P6's feedback echoes this sentiment:

I would prefer if it sounds like a friend says this to me and not a machine. (P06)

4.4.3 Role of AI. Here, we describe how participants saw AI-generated advice as part of how they cared about their wellbeing, i.e. the experience of a narrative as an additional source of information for wellbeing. An overarching sentiment was the sense of unfamiliarity and lack of personality. Furthermore, as captured by P05, the AI's conclusions, while data-driven, do not necessarily align with an individual's self-defined goals or personal benchmarks, which illustrates that the role of the AI and which stance participants expect it to take is still somewhat unclear:

I would say that transforming digital data into words descriptions, especially written as in this sample bears a risk of this feeling of being judged by a computer. Because as it is written here, it doesn't refer to any of my self-defined goals. It doesn't compare me to any objective reference point. For example, like, you engage in a significant amount of very active minutes, which can mean anything. Also, who says this is significant? (P05)

Participants also mentioned the challenge of trust and transparency with AI. Participants who were familiar with text generation were concerned about the black box nature of LLMs. Thus, the narratives' conclusions and recommendations came without the nuances of human context:

It's sweet and maybe some people respond positively to that, but I guess others would feel slightly distrustful. It just feels like someone is making a judgement call on you based on information that is at best incomplete. (P01)

4.5 Utilising the Interviews to Inform the Online Study

Informed by our analysis, we aimed to tailor our upcoming experiment to focus on the generation of a narrative as an innovative form of data representation. Our intent was to develop narratives that would be applicable, hold potential relevance, and ensure no harm was directed towards participants. We implemented several guiding principles based on the feedback, grounded in the themes identified.

Scope of Feedback. Recognising the constraints highlighted under the Abstraction theme, we accepted that the data available from current fitness trackers would guide our narratives to operate predominantly at a general level. This decision matched the participant's feedback who, while desiring specificity, understood the limitations of the technology at hand. Here, we recognise that a lack of context is a known limitation of fitness trackers [21], which cannot be yet addressed by LLMs.

Positive Framing. The Tone theme indicated that the LLM's handling of negative feedback was not always received well. With a keen focus on avoiding potential harm, especially around triggering rumination (which was recently identified as a key risk when presenting tracking data [19, 46]), we emphasised a neutral and constructive narrative tone.

Clarity and Directness. The feedback from the The Role of AI theme underscored the need for clarity. With this in mind, we aimed to ensure our narrative feedback was direct, eliminating any ambiguities or vague language that might introduce confusion or misinterpretation. This also ensures avoiding the feeling of patronising and lack of agency which are recognised negative aspects of trackers [60, 63] that were also emphasised in our interviews.

Using these principles derived from the themes, our goal was to present narrative data representation that resonated with users. These findings guided the online study, ensuring a user-centred and informed approach. The principles derived here serve as a necessary prerequisite to our online study.

5 ONLINE STUDY

Here, we describe the details of our online study which investigated if automatically-generated narratives can enhance the fitness tracking experience. We report on our study design, how we built a custom study platform and the results of our investigation.

5.1 Study Design

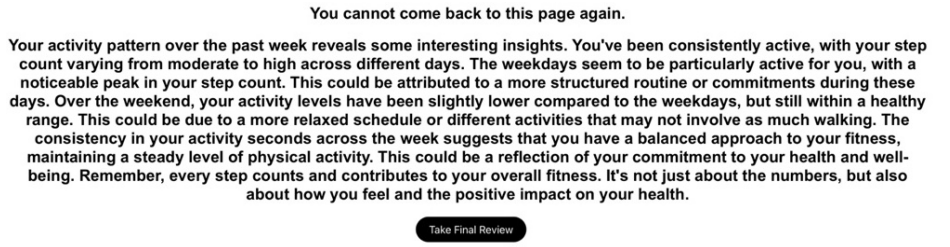
We conducted a between-subjects study design with three conditions, i.e. three different ways of representing the participants' fitness data, obtained from their fitness trackers. We used this study design for ecological validity—the participants communicated their impressions of a representation of their own data.

5.1.1 Conditions. In our study, participants were presented with one of three dynamically generated representations of their fitness data, which correspond to our three experimental conditions. All three conditions presented step data from the last seven days in the life of the participant (the day of the study was excluded). We chose step data as it is the most studied fitness metric in PI studies [21] due to its prevalence in commercial systems. Further, steps offer a unitless measure which can be easily interpreted by users [8, 22]. This implies that using conditions based on steps allowed recruiting a broad range of participants (i.e. fitness tracker users who track steps) and increased the likelihood of the users being familiar with the metric. Figure 3 shows the conditions in the online study.

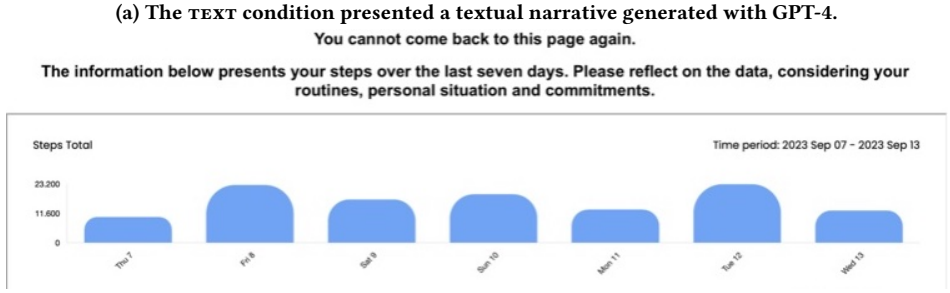
In the text condition, participants saw only an LLM-generated paragraph which narrated the last seven days of their step activity. The chart condition featured a standard bar chart displaying steps data per day. This was designed to mimic most commercial solutions for fitness trackers where a week of step data is presented in a bar chart, with one bar per day. By choosing a solution prevalent in consumer systems, we aimed to establish a strong baseline [27] for the study. The chart featured no additional annotations, only step data was present. In the text+chart condition, the participants were presented with both data representations simultaneously. Importantly, the data presented (as input to text generation or chart) across the three conditions was the same across the three conditions and focused on steps.

5.1.2 Measures. We investigated the effect of the data representation for steps used on how the users perceive the feedback in a PI system. To that end, we measured how the three conditions performed in terms of scales used to quantify design goals in PI.

We used the Technology-Supported Reflection Inventory (TSRI, [9]) to determine if the conditions supported data-driven reflection at different levels. The TSRI is a measure designed to assess how well



(a) The TEXT condition presented a textual narrative generated with GPT-4.



(b) In CHART, we showed a standard steps bar chart, generated using Terra API, to the participants.

Figure 3: The conditions in our study. In the TEXT+CHART, the text was presented above the chart.

an interactive system facilitates reflection. Utilising a seven-point Likert scale, it uses three primary dimensions: *Insight, TSRI-I*, which evaluates the system's ability to offer meaningful understanding; *Exploration, TSRI-E*, focusing on the ease and enjoyment of navigating personal data; and *Comparison, TSRI-C*, gauging the capacity for users to contrast their experiences with others. Collectively, these dimensions provide a comprehensive view of an interactive system's efficacy in fostering reflection. Bentvelzen et al. [9] note that the way in which a system supports reflection can be partly determined by the user's trait reflection. In line with the suggestions of Bentvelzen et al., we also measured trait reflection using the reflection-rumination questionnaire (RRQ) scale [67] (we refer to this score as the RRQ score in the remainder of this paper).

Next, we measured how the different data representations engaged the users with the User Engagement Scale—Short Form (UES-SF [51]). UES-SF is a tool designed to measure user engagement in digital interfaces. It evaluates how deeply users are immersed in their interactions, a dimension termed as Focused Attention (UES-FA). Further, the scale assesses the Perceived Usability (UES-PU), considering both the ease of interaction and any resultant negative emotions. Aesthetic Appeal (UES-AE) quantifies visual allure and design aesthetics of the interface, while the Reward Factor (UES-RW) measures value users derive from the interaction and the curiosity fostered by the system. For our study, a key aspect of UES-SF is that its subscales were previously used to study wellbeing [68] and storytelling [49].

5.2 Procedure

The study procedure is illustrated in Figure 4. Upon logging into Prolific[4], participants were informed about the prerequisites for the study. They were then directed to the study platform where they gave their consent in order to participate. Following this, participants were prompted to answer the RRQ. After this, the next step required participants to securely log in with their fitness data provider and provide consent to share only seven days' worth of step data for the purpose of this study. Once consent was given, data was retrieved while the participants completed the RRQ.

Subsequently, participants were presented with a data representation based on one of three conditions: TEXT, CHART or TEXT and CHART. Assignment to a condition was balanced and based on order of participation. After engaging with the data representation, participants were asked to complete both the TSRI and UES-SF questionnaires. Finally, they responded to an open-ended question, enquiring about their impressions of the presented data and any observations they made: 'What is your impression of the data presented? What did you observe?'.

To ensure smooth communication and address any concerns, a contact button was prominently displayed throughout the study, allowing participants to communicate directly with the researchers.

5.3 Apparatus

We executed this study as custom online survey running on an in-house server with Apache [5]. This ensured compliance with local data regulations. Terra API [6] was used to obtain step data from various different fitness providers. Terra was chosen as the tool for receiving participant data because it offered a wide compatibility in terms of devices and a uniform data format across different data providers. Terra also offered secure, encrypted communication. However, due to specific limitations, devices like the Apple Watch and Samsung devices were excluded because of additional verification requirements by Apple Health Kit and Samsung Health.

[4]https://app.prolific.co/

[5]https://httpd.apache.org/

[6]https://tryterra.co/

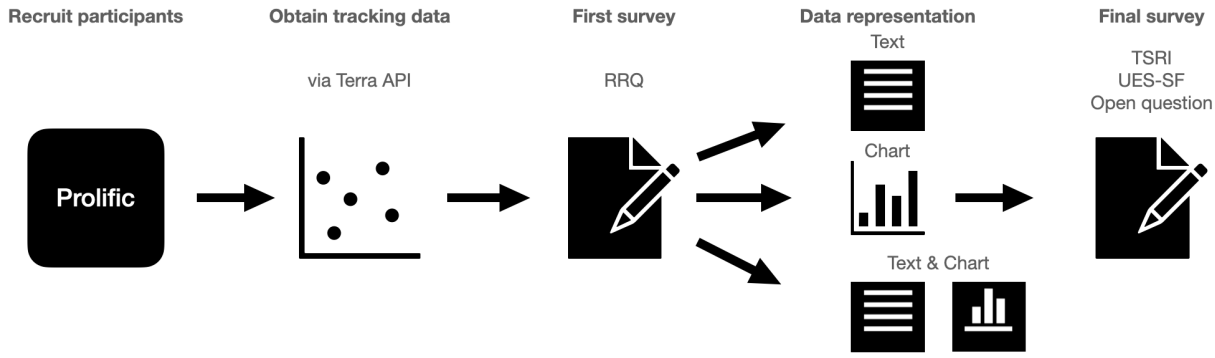


Figure 4: The online study from a user's point of view. Participants were recruited via Prolific and donated their fitness data using Terra. They then completed the study on a dedicated web page.

Huawei devices were also not used owing to prevailing national restrictions. Additionally, any fitness trackers that did not specifically measure steps were not supported in our study. Figure 5 shows how users logged into Terra API and completed questionnaires on the study webpage.

The data reception was managed via a webhook implemented in Node.js. Concurrently, a Node.js server[7] was responsible for displaying the study interface and collating participant responses. All acquired data, encompassing questionnaire responses, step metrics, and generated prompts, were securely archived in MongoDB. No identifiable participant data was retained. For data representation, we employed Terra API to generate charts. Given the diverse origins of data due to varied provider apps, participants might have been familiar with different visual presentations of their fitness data. Using a standardised chart from Terra API offered a neutral option with which participants were unlikely to be familiar, yet resembling know solutions. Lastly, the prompts provided to participants were dynamically generated using the Large Language Model GPT-4 [50] from OpenAI. The study was conducted in August 2023 using the model version `gpt-4-0613` via the API.

5.3.1 Prompt Engineering. Textual prompts are needed to effectively use pre-trained language models. A prompt is a set of instructions that provides guidance for the LLM to create the desired output. Prompt engineering refers to the iterative process of formulating and testing different prompts to find the most appropriate one, which allows a LLM to consistently solve the task at hand [39]. To implement textual data representations in a manner which could be studied, designing a robust LLM prompt was key. We used insight from the interview pre-study and literature to build our prompt. Utilising a zero-shot learning approach, our narrative design lacked explicit example narratives as training samples [17]. This was the case as we were not aware of any prior work on generating text from fitness data using LLMs and the scope of its data explanation capabilities. As such, we could not identify possible desired examples of fitness narratives, making a few-shot approach not feasible. Consequently, we provided the LLM with a textual description of the requirements, emphasising the steps, and features of acceptable narratives in line with the interview results.

[7] https://serverjs.io/

The prompt specifically instructs the LLM to offer a 'high-level overview' of the fitness data. By not getting into the specifics and asking to ignore days with very low step counts, it is inherently ensuring that the feedback operates predominantly at a general level, focusing on description and not recommendation (SCOPE OF FEEDBACK). We emphasise maintaining a 'neutral tone' and ensuring the feedback does not sound 'patronising', reflecting the principles in the TONE theme. There is a clear directive against negative feedback, aligning with the desire to prevent triggering rumination, a recognised risk in presenting tracking data [46]. We explicitly state that the story should be 'short and concise'; and instruct the model to 'stay objective'. The feedback should be direct, making sure there's no ambiguity, and eliminating any chance for misinterpretation. This is aligned with the ROLE OF AI theme feedback, ensuring the narrative is straightforward and does not introduce a feeling of lack of agency (CLARITY AND DIRECTNESS). In an iterative process, we designed a prompt that did not produce strong recommendations nor negative comments about user activity. The prompt was then informally piloted with three fitness tracker users and further checked by one experienced PI researcher external to the prompt engineering process. To further mitigate the risk of producing negative feedback due to the inherent unpredictability of LLMs, we ensured that participants who submitted data with very low activity levels were excluded from the study, i.e. they were not presented with any data representation. We also retrospectively reviewed all generated narratives in the survey and did not identify any potentially harmful texts. The generated texts are provided in the auxiliary material. Additionally, GPT-4 requires setting the temperature parameter t, which we set at 0.5, aiming for minimal variation while optimising control over the narrative output, according to the model specifications [50].

> *You are a personal assistant that receives fitness data from an individual. Your task is to tell a story based on it. This story should be a personalised high level overview of the data. Please focus on the steps. Use a neutral tone and stay objective. Do not sound patronising. Do not use any numbers. The result should not be performance based. Please give references from the data to support your statement. Ignore days with very low step counts (under 1000 steps). Stay short and concise, please limit your answer to 200 words. Your text*

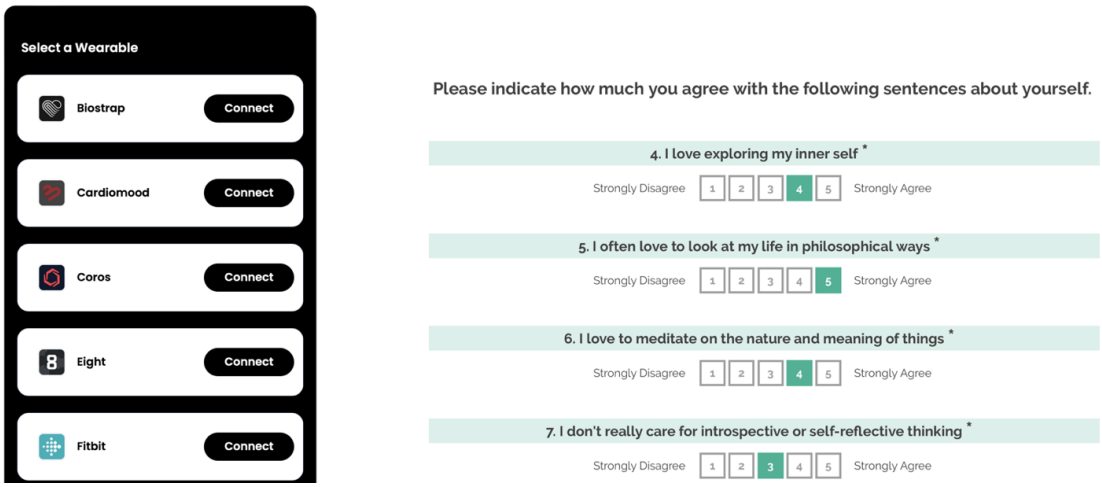


Figure 5: Elements of the study webpage. The login dialog allowing the user to choose their fitness data provider (left) and part of a questionnaire form (right). All webpage components were developed specifically for our study.

> *should provoke user reflection about the data considering routines, personal situation and commitments, but do not ask questions. The activity data is provided in the following format: [...]*

We used this prompt to generate all textual feedback in our online study. Data was provided to the model in a Python-style dictionary format injected into the prompt. All generated texts as well as the full prompt text and message objects for the API calls are available in the auxiliary material. Below is an illustrative example of a narrative generated by the GPT-4 for our survey:

> *Your fitness data shows a pattern of consistent activity throughout the week. On Thursday and Friday, your steps were similar, suggesting a routine or commitment on these days. Over the weekend, there was a significant increase in steps, with Sunday being the most active day. This could imply that you are using your free time for longer walks or more intense activities. The start of the week saw a decrease in steps, but still remained higher than Thursday and Friday. This suggests that despite the commitments of the working week, you are still prioritising physical activity. Wednesday's data indicates a slight decrease in steps compared to the beginning of the week. This could be due to various factors such as fatigue or changes in daily schedule. Your data reflects a balance between routine commitments and physical activity, with a noticeable increase in steps during the weekend. (This quote is automatically generated text.)*

5.4 Participants

We recruited a total of $n = 273$ participants through the online research recruitment service, Prolific. The mean age of participants was $M = 43.16$, $SD = 12.75$. The participant genders were distributed as follows: 154 participants were female, 117 male, one non-binary and one preferred not to disclose their gender. The participants resided primarily in the United Kingdom and the majority of them was actively employed. Detailed demographic data is available in the auxiliary material. Each participant received compensation for their involvement at a rate of £9 per hour, with a median study completion time of $7:22 min$, according to Prolific. Prior to opening the prolific study, participants were informed that they were required to be adult active users of a fitness tracking technology supported by the study platform. Further, we required that the participants donated step data from the last seven days preceding the day of the study. This way, we assured that study participants were familiar with fitness trackers and the data represented in the study described their real-life activity levels.

We made several exclusions to ensure data quality and relevance from the raw data gathered from Prolific, which originally featured $n = 324$ participants. In total, we excluded 51 participants due to the absence of data and instances where participants did not show activity on at least five out of the seven study days. Such data could have resulted in negative feedback, even though we were not able to achieve such during prompt piloting. Further participants were excluded as they used a device unsupported by the study platform. Another group of excluded participants was affected by poor server performance of their fitness data provider or the study server. This corresponds to a drop out rate of 16%, which is acceptable according to standard [45]. These participants were compensated despite their data not being used in the study.

6 ONLINE STUDY RESULTS

Here, we present the results of the study. First, we report on the metric scores. Then, we describe the qualitative results. Our study is driven by the question: *What are the differences between presenting PI data as standard chart and LLM-generated text in terms of supporting reflection and user experience?*

6.1 Quantitative Results

We first report the quantitative results of our study. We conducted one-way ANOVA-type procedures for all the measures in the study.

Table 2: Mean values and standard errors for TSRI (General measure of reflection) and its subscales (TSRI-I—Insight, TSRI-E—Exploration, and TSRI-C—Comparison) across the three conditions along with ANCOVA results. We found a significant effect for TSRI-C. $*$ denotes a significant difference between TEXT and CHART in a Tukey HSD test, at $p = .03$.

	TSRI		TSRI-I		TSRI-E		TSRI-C	
Condition	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SE*
TEXT	28.48	7.47	8.00	3.25	11.70	2.36	8.78$*$	3.35
CHART	26.31	7.16	7.60	3.37	11.14	2.68	7.57$*$	2.83
CHART+TEXT	28.24	8.54	8.03	3.65	11.69	2.78	8.51	3.32
ANCOVA	$F_{2,270} = 2.19$, $p = .11$		$F_{2,270} = 0.43$, $p = .64$		$F_{2,270} = 1.40$, $p = .25$		$F_{2,270} = 3.78$, $p = .03$	

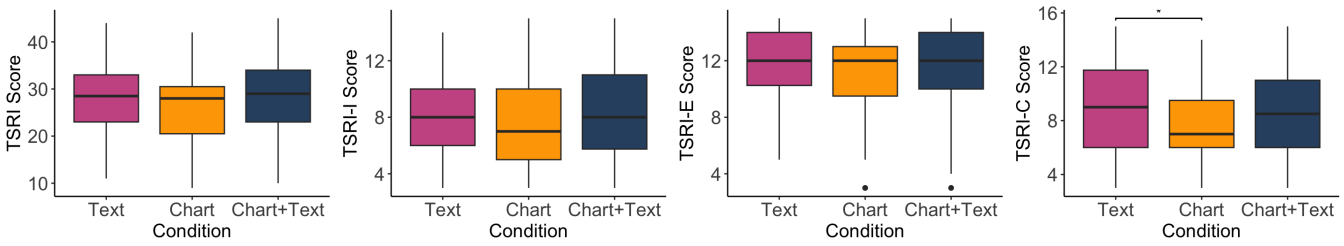


Figure 6: TSRI and subscale scores across the three conditions. We found a significant effect of condition on TSRI-C (Comparison).

Different exclusion rates across the conditions resulted in an unbalanced design, yet single-factor ANOVA tests are resistant to moderate differences in the amount of data in cells [43]. Whenever we report p-values, they are Bonferroni-corrected. We used an alpha level of .05 for all statistical tests. The raw data sheet for analysis is available in auxiliary material.

TEXT scored significantly higher than CHART in terms of comparison. We conducted a one-way ANCOVA to investigate the effect of data representations on the TSRI score and its subscales, with the RRQ score as covariate. The covariate effect was significant in all cases. We did not find a significant effect on the overall scale, but a significant effect for TSRI-C–Comparison was observed. Post-hoc comparisons using Tukey HSD showed a significant difference in TSRI-C for the condition pair TEXT and CHART, at $p = .03$. Table 2 shows the results of all test and Figure 6 visualises the results.

TEXT received significantly higher focused attention and reward scores than CHART. Similarly, we analysed the results in terms of UES-SF. We conducted one-way ANOVAs to study the effect of the data representation presented on UES-SF and its subscales. While we did not find a significant effect on the total scale score, we found significant effects for the FA and RW subscales as shown in Table 3. Post-hoc testing with Tukey HSD showed that there were significant differences between TEXT and CHART in terms of UES-FA and UES-RW, at $p = .02$ and $p = .01$, respectively. Figure 7 illustrates the results.

6.2 Answers to the Open Question

Each participant was asked to write at least one sentence about their impression of the data representation. Feedback generated in this way was then imported into Atlas.ti for qualitative analysis. In line with the pragmatic approach discussed by Blandford et al. [12], two researchers open-coded the full data set, applying the same process as for the pre-study, see Section 4.3. Through iterative discussions, we constructed three themes which illustrate how the users reacted to the generated fitness feedback: INTERPRETING FEEDBACK, TRACKING GOALS, and COMPARISON. Next, we present the three themes, focusing on understanding the inclusion of generated text in the data representation. We illustrate our findings with quotes from the data combined with the respective conditions to which the participants were subjected in brackets. The full set of comments is available in the auxiliary material.

6.2.1 Interpreting Feedback. Participants communicated how they interpreted the different forms of feedback presented in the study. In terms of sensemaking, the participants primarily discussed the recognition of patterns and how external factors influenced their activity. They reported that the narratives assisted users in identifying trends or unusual occurrences:

> *[I have been] More inconsistent than I thought, (...), I just try and do the 11,000 [steps] a day and that's it really, it's not a life changing thing if I fail, but I generally hit target. (TEXT)*

Often, participants noticed that the system did not have access to enough contextual information to offer a reasonable interpretation of the data. Participants who were affected by temporary medical conditions often found step data less meaningful, rendering the feedback less appropriate for them:

> *Unfortunately I have recently fractured a bone so I am limited in the amount I can do. When in good health I am obsessed with the number of steps I do and so the results would have been very interesting. (TEXT)*

Other participants expected that the feedback would include more specific information about their activity. Experienced users of

Table 3: Mean values and standard errors for User Enaggement Scale—Short Form (UES-SF) and its subscales (UES-AE—Aesthetic Appeal, UES-FA—Focused Attention, UES-RW—Reward, and UES-PU—Perceived Usability) for the three conditions. ANOVA results are shown in the last row. $*$ shows significant pairwise comparisons between TEXT and CHART at $p < .05$.

	UES-SF		UES-AE		UES-FA		UES-RW		UES-PU	
Condition	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
TEXT	13.55	2.63	3.04	0.92	2.38$*$	0.99	3.64$*$	1.00	4.49	0.67
CHART	12.83	2.68	3.12	0.97	2.01$*$	0.85	3.26$*$	1.05	4.45	0.65
CHART+TEXT	13.28	2.76	3.08	0.96	2.22	0.95	3.45	1.07	4.53	0.62
ANOVA	$F_{2,270} = 1.68$, $p = .19$		$F_{2,270} = 0.15$, $p = .87$		$F_{2,270} = 3.69$, $p = .02$		$F_{2,270} = 3.16$, $p = .04$		$F_{2,270} = 0.36$, $p = .70$	

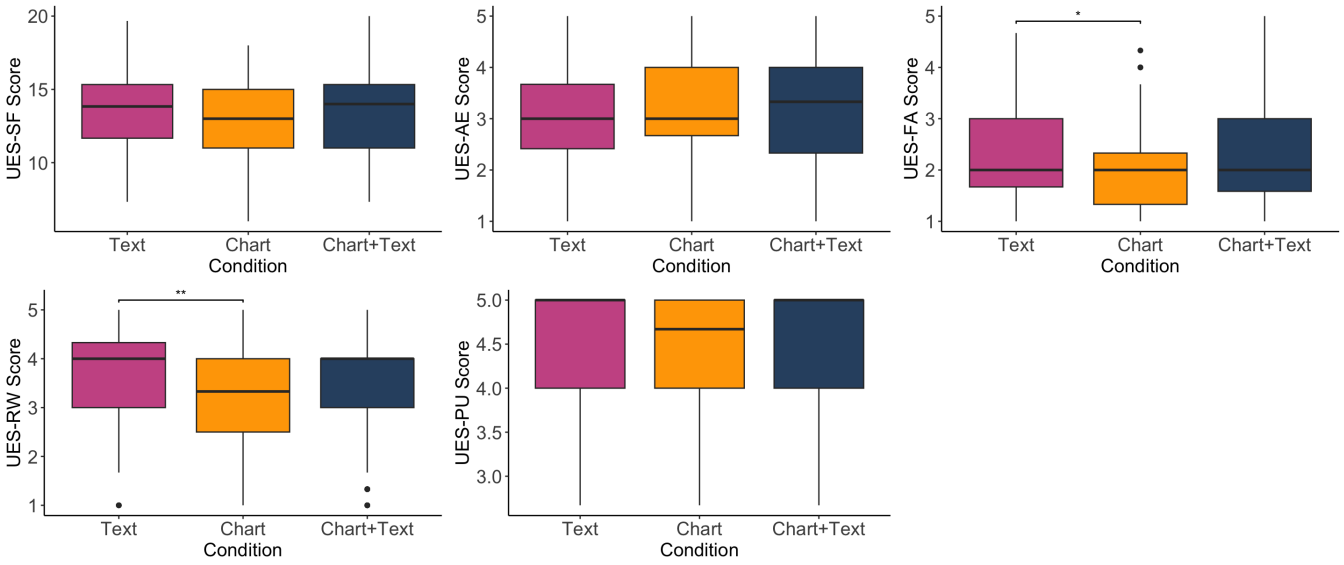


Figure 7: UES-SF and subscale scores in the three conditions. Significant effects were found for UES-FA (Focused Attention) and UES-RW (Reward).

fitness tracking often expected an analysis that would go beyond the scope of the data gathered by their tracker beyond the scope of what was used in the study:

> *It did not seem to take into account other data that day, so whilst my Thursday was a lower step count it had over 90 minutes rowing. All that I saw could have been obtained from my Garmin app however having further insights would make it useful. (CHART+TEXT)*

A few participants questioned the accuracy of the data, indicating a potential trust issue with AI-generated narratives or the data source itself. Some users supposed that there might have been technical issues:

> *I don't think the data presented was real/accurate. I have achieved 10,000 steps most days in the last week. (CHART)*

6.2.2 Tracking Goals. This theme describes how participants related the presented feedback to their goals. Despite our prompt not mentioning goals, users would still relate the presented feedback to their step targets. Many participants found that the generated narratives provided them with insights into their activity habits and triggered reflection on their goals:

> *I love how it knew exactly how something could have caused me not to hit my targets. That's mind blowing! I want to use this every day. (TEXT)*

Some participants reported that the feedback made them think more critically about their activity patterns and prompted them to set actionable goals. One participant discussed how they altered between reaching and failing to reach fitness goals on different days:

> *The data showed me that although I sometimes hit my step goal, there are many days where I do not. This caused me to think about days where I do vs. don't hit my steps and what I could do to hit that goal every day. (CHART+TEXT)*

For some users, the visualisation combined with the text prompted a realisation about their activity trends, fostering a renewed sense of commitment to their goals. This participant expressed surprise at the fact that they did not meet their goal, despite the feedback not featuring goals:

> *Honestly quite shocked I hadn't achieved my step count on most of the days this week. It's interesting seeing the data presented like this and any patterns that have developed. (CHART+TEXT)*

6.2.3 Comparison. Here, we describe how participants reported comparing their activity levels featured in the data representation to different reference points. Participants found value in the text presentation format, particularly in how it aided their comparison across the different days. Many reported that they perceived an ease of understanding:

> *It was nicely presented, easy to see and understand. I could observe how daily targets were met and then compare the results with what I did those days. (CHART+TEXT)*

However, some participants, who already had established tools or methods for viewing and interpreting their data, were disappointed with the limited depth of the feedback. They entertained the idea of the feedback being part of a larger tracking ecology with personalised data curation practices:

> *It was a different way to view the data, although Garmin is very good at showing my stats and I also export it to Google and look at trends there too (TEXT)*

Another recurring topic was relating the presented feedback to routines in order to compare the activity levels on different days. In a comparative mindset, participants drew connections between their activity levels and regular routines. Working from home was often mentioned as affecting daily step counts:

> *I can see that working from home and a busy weekday impacts my activity level. It was clear that I'm more active on the weekend. (CHART+TEXT)*

7 DISCUSSION

Here, we contextualise the results of our study and outline how they inform the design of future PI systems. We further discuss the implications of the possible use of AI-generated narratives in PI and consider the limitations of our work. By juxtaposing open question comments with quantitative metrics, we aim to understand the user experience of the different data representations. Further, we present recommendations for future PI systems based on reading our results through related work.

7.1 Attention Captivation by Text Descriptions

Our findings indicate that text descriptions capture the attention of users, as demonstrated by the UES-FA scores. Such narrative forms can be instrumental in guiding users' focus towards specific aspects of personal informatics, potentially fostering reflection cf. [23]. However, it is worth noting the potential influence of a novelty effect and how such attention direction will be affected by the TONE of the narrative. Our analysis showed no significant performance difference between TEXT and TEXT+CHART. There may be a potential novelty effect, where users might be primed to prioritise text over charts in combined formats. Further research is needed to explore how this potential novelty effect influences the interaction between text and visuals in user engagement. Our findings confirm past results that directing attention to oneself in PI is possible through engaging in a narrative [56]. We extend past findings by showing that such a narrative can be automatically generated. In particular, our data suggests an increase in user engagement metrics when data representations are personalised and dynamic, suggesting that the effectiveness of these narratives is closely tied to their relevance and adaptability to individual user profiles.

Recommendation for future systems: Future systems might leverage AI-generated text to direct attention towards certain metrics or areas, but they should be mindful of the diminishing effects over time and consider periodically updating or varying the narrative content to maintain user interest. Future work should focus on creating personalised narratives.

7.2 Text and the Comparative Mindset

The narrative style used in our study seems to stimulate a comparative mindset in users, as evident from the TSRI-C scores we obtained and qualitative feedback (COMPARISON). While this might enhance a reflective stance, it also poses risks of fostering competitiveness or negative thought cycles, similarly to visual approaches [46]. While comparison is often necessary for the sensemaking process to fully understand personal informatics data [28] and can foster positive social interactions [53, 66], excessive comparison can foster an overly critical mindset [38] or competition [55]. This is particularly relevant for inclusive PI systems as only a part of PI users is motivated by competition [33] and some users prefer a tracking experience without goals or comparisons [48]. Focusing on comparison may also lead to a negative experience for underrepresented groups of users [64]. For future research in PI, our work shows that using text feedback is affected by the same potential negative effects as using visuals. **Recommendation for future systems:** Systems aiming to employ narratives should ensure that comparisons are contextualised, potentially by providing benchmarks or by promoting positive and constructive comparative metrics rather than purely competitive ones. Thus, similarly to our system, future PI solutions that use generated text must avoid setting a solely comparative mindset.

7.3 Text for Focusing on Analysis and Reward

Our work provided results which suggest that the textual descriptions contribute to an enriched atmosphere of analysis, making the overall user experience rewarding, as captured by the UES-RW. This implies that AI-generated narratives have the potential to not just present data but also create an experience focused on analysis and sensemaking. The UES-RW scores indicate that the users perceived the TEXT condition as worth their time. In our study, we explicitly ensured that textual feedback did not feature numbers, sacrificing accuracy for richness of description in order to focus on the qualitative experience of PI. This is in line with work by Niess and Wozniak [47] which showed how most users frame their fitness goal qualitatively. Thus, rich textual descriptions can contribute to an experience of meaning in PI [47]. As we observed in INTERPRETING FEEDBACK, some users are likely to ascribe interpretive power to textual feedback, expecting information beyond what is possible in conventional trackers. This implies that the boundary between measurement and interpretation, already present in current systems [8], where some metrics are sensor measurements and some are algorithmic predictions will be even harder to navigate.

Designers of future PI systems will be faced with the requirement to transparently communicate how rich feedback is generated and how precise it is.

Recommendation for future systems: AI narratives in PI can be used to enhance engagement, but there is a need to inform the user on how the narratives relate to their PI data. Through generated text, PI systems can offer a qualitative experience which supports qualitative goals at the potential cost of accuracy.

7.4 Text as a Complementary Tool

Both our interviews and online study underscore the idea that text should not operate as the sole data representation medium. As we did not observe differences between the TEXT and TEXT+CHART conditions, further research is needed to explore the interplay of using multiple data representations simultaneously. Instead, narratives should ideally supplement visuals, which have been shown to be particularly effective when the primary design objective is to guide users' attention [25]. Our data suggest that users are accustomed to the standard commercial data representations and charts are likely to continue being the main communication form of personal informatics [28]. Yet, our results show that a single interaction with generated text can offer tangible benefits. Past work has shown that the data representations with which users engage undergo constant evolution, which facilitates a constant change of perspective [7]. Results in INTERPRETING FEEDBACK show that generated text sets a lens that may make users notice particular facts or patterns. While past work has shown that this is possible with brief pre-scripted text [15], our results illustrate that the strategy is also effective for richer text descriptions. Consequently, automatically generated text can be one more effective tool in the PI toolbox. **Recommendation for future systems:** Hybrid interfaces that seamlessly integrate both visual and narrative elements, allowing users to switch or combine views based on their preference, can enhance user comprehension and engagement in PI. Narratives can be used to offer alternative perspectives and draw attention to particular data points, which is a desired PI experience.

7.5 Challenges with Text Generation

Here, we reflect on the design process of the tool used in this paper. The process of text generation, while promising, is fraught with both practical and ethical challenges. As gleaned from our interviews (ROLE OF AI), there is a pressing need to address these concerns to ensure that the generated narratives are not just effective but also ethically sound. A key issue that emerges from our work is preventing the LLM from generating data representations which could potentially trigger negative reactions in users, i.e. limit the possibility of inducing rumination [19, 46]. In our study, we assured that negative framing [29] was not used by following a specific prompt design. Looking at the narratives generated in the study, we created feedback that was potentially incorrect, but never negative. Yet, this design necessitated that the text produced was highly abstract, which may be detrimental to the PI experience, cf. ABSTRACTION. Further, applications of LLMs in PI are currently affected by concerns similar to applying them in health [32]: the models have a tendency to hallucinate due to a lack of relevant world knowledge [70]. Counterfactual feedback could induce a highly negative experience of PI. This appears to be particularly challenging as our work shows that text can create a sense of reward (UES-RW). As users appear to value such feedback, it remains a design challenge to assure its high quality. **Recommendation for future systems:** Engaging with ethical AI guidelines and possibly integrating user feedback loops to refine and correct narrative outputs can enhance the trustworthiness and effectiveness of the generated text. When designing for health interventions, which is not the goal in this paper, it is crucial to limit the scope of the LLM and introduce strict means of mitigating possible harm.

7.6 Ways Forward

The exploration of AI-generated narratives in the realm of personal informatics opens numerous possibilities. One promising avenue is the enhancement of system interactivity, remembering that most LLMs, including the one used in our study, are designed to support conversations. By enabling users to question and interact with the generated text, there is potential for deeper, more personalised feedback. Such an interactive feature would make the data more engaging and dynamic, allowing users to delve into specific areas of interest and additionally provide more information on their personal motivation, goals and preferences. This additional context information for the LLM could address the limitations pointed out in the interviews (ABSTRACTION and SCOPE OF FEEDBACK). Past work has shown that users desire a certain level of abstraction in how they view their data [47], which LLMs can provide. Designers, however, must be weary to balance abstraction and accuracy.

Conversations with AI models are affected by explainability issues [35]. Our work provides initial empirical evidence that an LLM-powered conversation about one's own data could benefit PI systems. Yet, our experience from this work indicates that this comes with a number of caveats. First, such a solution would make PI systems resemble health-oriented chatbots, which, arguably, enjoyed limited success in practice and research [1]. Alternatively, the LLM could assume the role of a coach, as suggested by earlier research [11], yet this would be in opposition to the need to avoid patronising, as seen in our interviews. Second, we note in the initial stages of our prompt design, the GPT-4 model was inclined towards providing feedback focused on performance, resembling narratives typical of persuasive technology. We needed to explicitly contain this in the prompt used in our study. While these problems may be specific to the training set of that model, there is a probability that general LLMs show an inclination towards persuasion. Thus, using AI-generated narratives may involve the designer of the PI system in a number of issues for which persuasive technology has been criticised in the HCI community [69]. Consequently, it remains an open question if and how providing more context to or involving the user in a dialogue with an LLM can benefit the PI experience.

Furthermore, LLMs appear to be a promising tool to support sensemaking across metrics. As our study constituted the first, to our knowledge, exploration of using LLMs with fitness data, we focused on a single metric. Yet, future systems should build narratives based on a broad spectrum of user-specific data points. This notion is key in our current understanding of PI—diverse data sources form the bedrock of reflection [22, 37, 47]. In our exploration, tailoring content around user-specific information based on tracker

data surfaced as a potential enhancement. The move towards a more individualised narrative could be paramount in bridging the gap between raw data and actionable insights [41]. However, such perspectives are also affected by potential challenges. Personalising feedback also introduces a myriad of ethical concerns around data privacy and potential biases in AI interpretations [16, 31]. Coupling this with the integration of personal goals and the complexity of goal setting (which appears to be unavoidable; see TRACKING GOALS) within the AI narratives can add another layer of complexity, cf [2, 47]. As a result, PI systems may generate feedback that is significantly beyond the control of the designer. Alternatively, it may expose the PI design space to potentially unethical parties by facilitating and driving users towards a particular, prescriptive way to interpret data.

The choice of using an LLM over scripted text in personal informatics systems offers new potential due to LLMs' adaptability and ability to generate novel content. Scripted text, while consistent, can quickly become repetitive and lose its appeal, leading to reduced user engagement over time. In contrast, LLMs offer dynamic and evolving narratives that can adapt to changing user data, interactions, and contexts, continually introducing fresh elements of interest. This capability to perpetually generate new narrative content helps maintain user interest and engagement. Yet, in our study, we cannot separate the novelty effect of the content of the narrative and the text modality per se. While we show that LLM-generated text offers new potential for PI, future research should study its optimal use.

Finally, we note that our results suggest that text may be simply used as an extension for conventional visual representations. However, we note that this may also generate additional complexity. The juxtaposition of AI-generated narratives with traditional visuals creates new challenges for the interface as it will need to be more complex. With new LLM-based tools, e.g. Whoop Coach, appearing on the consumer market, we will need to understand how narratives become part of a larger ecology of data representations in PI. To strike the right balance between data representations, designers will need to consider cognitive load, coherence, and potential biases that each representation might introduce. Thus, given that our work suggests that AI-generated narratives can enhance PI experiences, it is a challenge for HCI to find the right place for narratives in PI interfaces.

7.7 Limitations

We note that this initial exploration of using LLMs to enhance PI experiences is prone to certain limitations. First, our study was exclusively conducted online. Physical contexts, such as real-world environmental settings and the dynamics of face-to-face interactions, might influence users' perceptions of PI as shown in many studies of physical artefacts for tracker data, e.g. [59]. Further, we faced technological limitations, as certain popular fitness trackers like the Apple Watch, Samsung and Huawei devices were not supported. This exclusion might have skewed our participant demographic to users of specific platforms, potentially not capturing the entirety of personal informatics users' experiences and insights.

Another inherent limitation revolves around the actual usage of fitness trackers. The varied rates at which participants wore their trackers on different days might influence the data's reliability. This variability can be attributed to several factors, including battery life, daily routines, or simply forgetfulness [65]. While we only included participants who recorded steps in at least five of the last seven days, we had no control over tracker usage patterns during those days. We see this challenge as inherent to studies in PI, which use real-life data. Our sample size is likely to have compensated for such an effect. Another limitation pertains to the scales which we used for gauging participant engagement and reflection. Behavioural measures, such as eye-tracking, might have offered different insights. However, the completion times observed on Prolific do provide an indication that participants review the content with care. Lastly, our use of GPT-4 introduces its own set of challenges. As with any language model, GPT-4 possesses inherent biases and limitations. The opaque nature of these models implies that we cannot have full control over their outputs or ensure exact replication in future studies. Since the conclusion of our study, GPT has been updated and its output may change. Newer versions of the model feature a 'seed' parameter which increases replicability [3]. Furthermore, fine-tuning an LLM allows designing a task-specific adaptation to improve the performance in a certain domain. We theorise that a fine-tuned model would have been more effective than the solution presented in this paper. The OpenAI API now allows fine-tuning, but only for the older GPT-3.5 Turbo model [52]. Yet, both of these features were not available at the time of the study. Our work shows the potential benefits of text-based narratives in PI. These benefits are likely to be more profound as better models for narrating PI are developed. Long-term use of such systems will most likely require PI-specific personalised models where feedback evolves over time in a way specific to individual users. We note that our results are partly dependent upon the specifics of this version of the model. Future work can address this by building more specialised models, designed for generating data narratives.

Further, we acknowledge that our study is embedded in HCI work on PI, which primarily targets WEIRD (Western, Educated, Industrialised, Rich, and Democratic [64]) participants. Thus our study is also affected by cultural bias, especially as Prolific samples also exhibit WEIRD bias [40]. Similarly, most of the participants recruited for our interview study possessed or studied towards a university degree. We acknowledge that our insights may be relevant only to users from Western cultures, who can afford PI devices. We make the anonymised demographic data for the survey available to contribute to the discourse on how to build PI systems for a broader range of users.

8 CONCLUSION

In this work, we explored how LLM-generated qualitative descriptions influenced the manner in which users engaged with their fitness data. To this end, we conducted an interview pre-study and an online experiment, which used real-life fitness tracker data. The textual presentation augmented the levels of comparison-based reflection among participants. Further, presenting fitness data as text captured users' attention more effectively and made the interaction more rewarding for participants. Our results show the potential of integrating AI-driven narratives to bolster user engagement with personal informatics tools. Based on our findings, we recommend

that personal informatics systems utilise LLM-generated narratives to direct and sustain user attention, ensure comparisons are presented in a positive manner, enrich the user experience with a blend of visual and textual representations, and approach text generation with an emphasis on ethical considerations. We hope that our work inspires future research about how narrative-driven methods can contribute to building better PI systems which benefit user wellbeing.

ACKNOWLEDGMENTS

This work was supported by the Swedish Research Council, award number 2022-03196. Konstantin R. Strömel was supported by the German Academic Exchange Service (DAAD) as part of the RISE-Worldwide programme. Paweł W. Woźniak thanks the University of St. Gallen for providing a writing retreat that enabled focusing on this work.

REFERENCES

[1] Alaa A Abd-Alrazaq, Mohannad Alajlani, Nashva Ali, Kerstin Denecke, Bridgette M Bewick, and Mowafa Househ. 2021. Perceptions and opinions of patients about mental health chatbots: scoping review. *Journal of medical Internet research* 23, 1 (2021), e17828.

[2] Elena Agapie, Patricia A Areán, Gary Hsieh, and Sean A Munson. 2022. A Longitudinal Goal Setting Model for Addressing Complex Personal Problems in Mental Health. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–28.

[3] Shyamal Anadkat. 2023. *How to make your completions outputs consistent with the new seed parameter.* Retrieved December 12, 2023 from https://cookbook.openai.com/examples/deterministic_outputs_with_the_seed_parameter

[4] Amid Ayobi, Tobias Sonne, Paul Marshall, and Anna L. Cox. 2018. Flexible and Mindful Self-Tracking. (2018), 1–14. https://doi.org/10.1145/3173574.3173602

[5] Eric P.S. Baumer, Vera Khovanskaya, Mark Matthews, Lindsay Reynolds, Victoria Schwanda Sosik, and Geri Gay. 2014. Reviewing reflection: On the use of reflection in interactive system design. *Proceedings of the Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques, DIS* (2014), 93–102. https://doi.org/10.1145/2598510.2598598

[6] Eric P S Baumer. 2015. 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)*. Association for Computing Machinery, New York, NY, USA, 585–594. https://doi.org/10.1145/2702123.2702234

[7] Marit Bentvelzen, Jasmin Niess, and Pawel W. Wozniak. 2021. The technology-mediated reflection model: Barriers and assistance in data-driven reflection. In *Conference on Human Factors in Computing Systems - Proceedings (CHI '21)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411764.3445505

[8] Marit Bentvelzen, Jasmin Niess, and Paweł W Woźniak. 2023. Designing Reflective Derived Metrics for Fitness Trackers. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 4 (2023), 1–19.

[9] Marit Bentvelzen, Jasmin Niess, Mikołaj P. Woźniak, and Paweł W. Woźniak. 2021. The Development and Validation of the Technology-Supported Reflection Inventory. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3411764.3445673

[10] Marit Bentvelzen, Paweł W Woźniak, Pia SF Herbes, Evropi Stefanidi, and Jasmin Niess. 2022. Revisiting reflection in hci: Four design resources for technologies that support reflection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 1 (2022), 1–27.

[11] Robbert Jan Beun, Siska Fitrianie, Fiemke Griffioen-Both, Sandor Spruit, Corine Horsch, Jaap Lancee, and Willem-Paul Brinkman. 2017. Talk and Tools: the best of both worlds in mobile user interfaces for E-coaching. *Personal and ubiquitous computing* 21 (2017), 661–674.

[12] Ann Blandford, Dominic Furniss, and Stephann Makri. 2016. Qualitative HCI Research: Going Behind the Scenes. *Synthesis Lectures on Human-Centered Informatics* 9 (2016), 1–115. https://doi.org/10.2200/S00706ED1V01Y201602HCI034

[13] Federica Cena, Amon Rapp, Silvia Likavec, and Alessandro Marcengo. 2018. Envisioning the future of personalization through personal informatics: A user study. *International Journal of Mobile Human Computer Interaction (IJMHCI)* 10, 1 (2018), 52–66.

[14] Eun Kyoung Choe, Bongshin Lee, Haining Zhu, and Nathalie Henry Riche. 2017. Understanding self-reflection: How people reflect on personal data through visual data exploration. In *ACM International Conference Proceeding Series (Pervasive-Health '17)*. Association for Computing Machinery, New York, NY, USA, 173–182. https://doi.org/10.1145/3154862.3154881

[15] Sunny Consolvo, Predrag Klasnja, David W. McDonald, and James A. Landay. 2014. Designing for Healthy Lifestyles: Design Considerations for Mobile Technologies to Encourage Consumer Health and Wellness. *Found. Trends Hum.-Comput. Interact.* 6, 3–4 (apr 2014), 167–315. https://doi.org/10.1561/1100000040

[16] Louis Anthony Cox Jr. 2023. Pushing Back on AI: A Dialogue with ChatGPT on Causal Inference in Epidemiology. In *AI-ML for Decision and Risk Analysis: Challenges and Opportunities for Normative Decision Theory*. Springer, 407–423.

[17] Hai Dang, Lukas Mecke, Florian Lehmann, Sven Goller, and Daniel Buschek. 2022. How to Prompt? Opportunities and Challenges of Zero- and Few-Shot Learning for Human-AI Interaction in Creative Applications of Generative Models. arXiv:2209.01390 [cs.HC]

[18] Pooja M. Desai, Elliot G. Mitchell, Maria L. Hwang, Matthew E. Levine, David J. Albers, and Lena Mamykina. 2019. Personal Health Oracle: Explorations of Personalized Predictions in Diabetes Self-Management. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300600

[19] Elizabeth Victoria Eikey, Clara Marques Caldeira, Mayara Costa Figueiredo, Yunan Chen, Jessica L Borelli, Melissa Mazmanian, and Kai Zheng. 2021. Beyond self-reflection: introducing the concept of rumination in personal informatics. *Personal and Ubiquitous Computing* 25, 3 (2021), 601–616.

[20] Chris Elsden, David S Kirk, and Abigail C Durrant. 2016. A quantified past: Toward design for remembering with personal informatics. *Human–Computer Interaction* 31, 6 (2016), 518–557.

[21] Daniel A Epstein, Clara Caldeira, Mayara Costa Figueiredo, Xi Lu, Lucas M Silva, Lucretia Williams, Jong Ho Lee, Qingyang Li, Simran Ahuja, Qiuer Chen, Payam Dowlatyari, Craig Hilby, Sazeda Sultana, Elizabeth V Eikey, and Yunan Chen. 2020. Mapping and Taking Stock of the Personal Informatics Literature. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 4 (12 2020). https://doi.org/10.1145/3432231

[22] Daniel A. Epstein, An Ping, James Fogarty, and Sean A. Munson. 2015. A lived informatics model of personal informatics. *UbiComp 2015 - Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (2015), 731–742. https://doi.org/10.1145/2750858.2804250

[23] Eivind Flobak, Oda Elise Nordberg, Frode Guribye, Tine Nordgreen, and Ragnhild Johanne Tveit Sekse. 2021. "This is the Story of Me": Designing Audiovisual Narratives to Support Reflection on Cancer Journeys. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference* (Virtual Event, USA) *(DIS '21)*. Association for Computing Machinery, New York, NY, USA, 1031–1045. https://doi.org/10.1145/3461778.3462005

[24] Elliot G. Mitchell, Elizabeth M. Heitkemper, Marissa Burgermaster, Matthew E. Levine, Yishen Miao, Maria L. Hwang, Pooja M. Desai, Andrea Cassells, Jonathan N. Tobin, Esteban G. Tabak, David J. Albers, Arlene M. Smaldone, and Lena Mamykina. 2021. From Reflection to Action: Combining Machine Learning with Expert Knowledge for Nutrition Goal Recommendations. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–17. https://doi.org/10.1145/3411764.3445555

[25] Rúben Gouveia, Evangelos Karapanos, and Marc Hassenzahl. 2015. How Do We Engage with Activity Trackers? A Longitudinal Study of Habito. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Osaka, Japan) *(UbiComp '15)*. Association for Computing Machinery, New York, NY, USA, 1305–1316. https://doi.org/10.1145/2750858.2804290

[26] Victoria Hollis, Artie Konrad, Aaron Springer, Matthew Antoun, Christopher Antoun, Rob Martin, and Steve Whittaker. 2017. What Does All This Data Mean for My Future Mood? Actionable Analytics and Targeted Reflection for Emotional Well-Being. *Human–Computer Interaction* 32, 5-6 (Nov. 2017), 208–267. https://doi.org/10.1080/07370024.2016.1277724 Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/07370024.2016.1277724.

[27] Kasper Hornbæk et al. 2013. Some whys and hows of experiments in human–computer interaction. *Foundations and Trends® in Human–Computer Interaction* 5, 4 (2013), 299–373.

[28] Dandan Huang, Melanie Tory, Bon Adriel Aseniero, Lyn Bartram, Scott Bateman, Sheelagh Carpendale, Anthony Tang, and Robert Woodbury. 2014. Personal visualization and personal visual analytics. *IEEE Transactions on Visualization and Computer Graphics* 21, 3 (2014), 420–433.

[29] Punam Anand Keller, Isaac M Lipkus, and Barbara K Rimer. 2003. Affect, framing, and persuasion. *Journal of Marketing Research* 40, 1 (2003), 54–64.

[30] Rohit Ashok Khot, Jeewon Lee, Larissa Hjorth, and Florian 'Floyd' Mueller. 2014. SweatAtoms: Understanding Physical Activity Through Material Artifacts. *CHI '14 Extended Abstracts on Human Factors in Computing Systems* (2014), 173–174. https://doi.org/10.1145/2559206.2579479

[31] Vera Khovanskaya, Eric PS Baumer, Dan Cosley, Stephen Voida, and Geri Gay. 2013. " Everybody knows what you're doing" a critical design approach to personal informatics. In *Proceedings of the SIGCHI Conference on Human Factors*

in Computing Systems. 3403–3412.

[32] Jens Kleesiek, Yonghui Wu, Gregor Stiglic, Jan Egger, and Jiang Bian. 2023. An opinion on ChatGPT in health care—written by humans only. , 701–703 pages.

[33] Kristina Knaving, Paweł Woźniak, Morten Fjeld, and Staffan Björk. 2015. Flow is not enough: Understanding the needs of advanced amateur runners to design motivation technology. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 2013–2022.

[34] Rafal Kocielnik, Lillian Xiao, Daniel Avrahami, and Gary Hsieh. 2018. Reflection Companion: A Conversational System for Engaging Users in Reflection on Physical Activity. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 2 (7 2018). https://doi.org/10.1145/3214273

[35] Himabindu Lakkaraju, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. 2022. Rethinking Explainability as a Dialogue: A Practitioner's Perspective. arXiv:2202.01875 [cs.LG]

[36] Yi-Chieh Lee, Naomi Yamashita, Yun Huang, and Wai Fu. 2020. " I hear you, I feel you": encouraging deep self-disclosure through a chatbot. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–12.

[37] Ian Li, Anind Dey, and Jodi Forlizzi. 2010. A Stage-Based Model of Personal Informatics Systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. Association for Computing Machinery, New York, NY, USA, 557–566. https://doi.org/10.1145/1753326.1753409

[38] James J Lin, Lena Mamykina, Silvia Lindtner, Gregory Delajoux, and Henry B Strub. 2006. Fish'n'Steps: Encouraging physical activity with an interactive computer game. In *UbiComp 2006: Ubiquitous Computing: 8th International Conference, UbiComp 2006 Orange County, CA, USA, September 17-21, 2006 Proceedings 8*. Springer, 261–278.

[39] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. arXiv:2107.13586 [cs.CL]

[40] Prolific Academic Ltd. [n. d.]. https://researcher-help.prolific.com/hc/en-gb/articles/360009501473-What-are-the-advantages-and-limitations-of-an-online-sample-

[41] Lena Mamykina, Daniel A. Epstein, Predrag Klasnja, Donna Sprujt-Metz, Jochen Meyer, Mary Czerwinski, Tim Althoff, Eun Kyoung Choe, Munmun De Choudhury, and Brian Lim. 2022. Grand Challenges for Personal Informatics and AI. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems (CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3491101.3503718

[42] Andrea Mennicken and Wendy Nelson Espeland. 2019. What's new with numbers? Sociological approaches to the study of quantification. *Annual Review of Sociology* 45 (2019), 223–245.

[43] George A. Milliken and Dallas E. Johnson. 2009. *Analysis of Messy Data Volume 1*. Chapman and Hall/CRC. https://doi.org/10.1201/ebk1584883340

[44] Elizabeth L Murnane, Xin Jiang, Anna Kong, Michelle Park, Weili Shi, Connor Soohoo, Luke Vink, Iris Xia, Xin Yu, John Yang-Sammataro, Grace Young, Jenny Zhi, Paula Moya, and James A Landay. 2020. Designing Ambient Narrative-Based Interfaces to Reflect and Motivate Physical Activity. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376478

[45] Lung National Heart, Blood Institute, et al. 2014. Quality assessment of controlled intervention studies. *USA: National Institutes of Health, Department of Health and Human Services* (2014).

[46] Jasmin Niess, Kristina Knaving, Alina Kolb, and Paweł W. Woźniak. 2020. Exploring Fitness Tracker Visualisations to Avoid Rumination. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services* (Oldenburg, Germany) *(MobileHCI '20)*. Association for Computing Machinery, New York, NY, USA, Article 6, 11 pages. https://doi.org/10.1145/3379503.3405662

[47] Jasmin Niess and Paweł W. Woźniak. 2018. Supporting meaningful personal fitness: The tracker goal Evolution Model. *Conference on Human Factors in Computing Systems - Proceedings* 2018-April (2018), 1–12. https://doi.org/10.1145/3173574.3173745

[48] Jasmin Niess, Paweł W Woźniak, Yomna Abdelrahman, Passant ElAgroudy, Yasmeen Abdrabou, Caroline Eckerth, Sarah Diefenbach, and Kristina Knaving. 2021. 'I Don't Need a Goal': Attitudes and Practices in Fitness Tracking beyond WEIRD User Groups. In *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction*. 1–14.

[49] Heather L O'Brien. 2017. Antecedents and learning outcomes of online news engagement. *Journal of the Association for Information Science and Technology* 68, 12 (2017), 2809–2820.

[50] R OpenAI. 2023. GPT-4 technical report. *arXiv* (2023), 2303–08774.

[51] Heather L. O'Brien, Paul Cairns, and Mark Hall. 2018. A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *International Journal of Human-Computer Studies* 112 (April 2018), 28–39. https://doi.org/10.1016/j.ijhcs.2018.01.004

[52] Andrew Peng, Michael Wu, John Allard, Logan Kilpatrick, and Steven Heidel. 2023. *GPT-3.5 Turbo fine-tuning and API updates*. https://openai.com/blog/gpt-3-5-turbo-fine-tuning-and-api-updates

[53] Laura R Pina, Sang-Wha Sien, Teresa Ward, Jason C Yip, Sean A Munson, James Fogarty, and Julie A Kientz. 2017. From personal informatics to family informatics: Understanding family practices around health monitoring. In *Proceedings of the 2017 acm conference on computer supported cooperative work and social computing*. 2300–2315.

[54] Amon Rapp and Arianna Boldi. 2023. Exploring the Lived Experience of Behavior Change Technologies: Towards an Existential Model of Behavior Change for HCI. *ACM Trans. Comput.-Hum. Interact.* (jun 2023). https://doi.org/10.1145/3603497 Just Accepted.

[55] Amon Rapp and Lia Tirabeni. 2020. Self-tracking while doing sport: Comfort, motivation, attention and lifestyle of athletes using personal informatics tools. *International Journal of Human-Computer Studies* 140 (2020), 102434.

[56] Amon Rapp and Maurizio Tirassa. 2017. Know thyself: a theory of the self for personal informatics. *Human–Computer Interaction* 32, 5-6 (2017), 335–380.

[57] John Rooksby, Mattias Rost, Alistair Morrison, and Matthew Chalmers. 2014. Personal tracking as lived informatics. *Conference on Human Factors in Computing Systems - Proceedings* (2014), 1163–1172. https://doi.org/10.1145/2556288.2557039

[58] Herman Saksono and Andrea G. Parker. 2017. Reflective informatics through family storytelling: Self-discovering physical activity predictors. *Conference on Human Factors in Computing Systems - Proceedings* 2017-May (2017), 5232–5244. https://doi.org/10.1145/3025453.3025651

[59] Kim Sauvé, Steven Houben, Nicolai Marquardt, Saskia Bakker, Bart Hengeveld, Sarah Gallacher, and Yvonne Rogers. 2017. LOOP: A Physical Artifact to Facilitate Seamless Interaction with Personal Data in Everyday Life. In *Proceedings of the 2017 ACM Conference Companion Publication on Designing Interactive Systems (DIS '17 Companion)*. Association for Computing Machinery, New York, NY, USA, 285–288. https://doi.org/10.1145/3064857.3079175

[60] Hanna Schneider. 2017. Adapting at run-time: Exploring the design space of personalized fitness coaches. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces Companion*. 173–176.

[61] Marita Skjuve, Asbjørn Følstad, Knut Inge Fostervold, and Petter Bae Brandtzaeg. 2021. My chatbot companion-a study of human-chatbot relationships. *International Journal of Human-Computer Studies* 149 (2021), 102601.

[62] Petr Slovak, Chris Frauenberger, and Geraldine Fitzpatrick. 2017. Reflective practicum: A framework of sensitising concepts to design for transformative reflection. *Conference on Human Factors in Computing Systems - Proceedings* 2017-May (2017), 2696–2707. https://doi.org/10.1145/3025453.3025516

[63] Katta Spiel. 2019. Body-positive computing as a means to counteract normative biases in fitness trackers. *XRDS: Crossroads, The ACM Magazine for Students* 25, 4 (2019), 34–37.

[64] Katta Spiel, Fares Kayali, Louise Horvath, Michael Penkler, Sabine Harrer, Miguel Sicart, and Jessica Hammer. 2018. Fitter, Happier, More Productive?: The Normative Ontology of Fitness Trackers. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems (CHI EA '18)*. ACM, New York, NY, USA, alt08:1–alt08:10. https://doi.org/10.1145/3170427.3188401

[65] Lie Ming Tang, Jochen Meyer, Daniel A Epstein, Kevin Bragg, Lina Engelen, Adrian Bauman, and Judy Kay. 2018. Defining adherence: making sense of physical activity tracker data. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 2, 1 (2018), 1–22.

[66] Tammy Toscos, Anne Faber, Shunying An, and Mona Praful Gandhi. 2006. Chick clique: persuasive technology to motivate teenage girls to exercise. In *CHI'06 extended abstracts on Human factors in computing systems*. 1873–1878.

[67] Paul D. Trapnell and Jennifer D. Campbell. 1999. Private self-consciousness and the five-factor model of personality: Distinguishing rumination from reflection. *Journal of Personality and Social Psychology* 76, 2 (1999), 284–304. https://doi.org/10.1037/0022-3514.76.2.284 Place: US Publisher: American Psychological Association.

[68] Nadine Wagener, Marit Bentvelzen, Bastian Dänekas, Paweł W Woźniak, and Jasmin Niess. 2023. VeatherReflect: Employing Weather as Qualitative Representation of Stress Data in Virtual Reality. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. 446–458.

[69] Fahri Yetim. 2013. Critical perspective on persuasive technology reconsidered. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 3327–3330.

[70] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. arXiv:2309.01219 [cs.CL] https://doi.org/10.48550/arXiv.2309.01219