

# Modelling Inter-Individual Differences in Multimodal Data

Ein metrischer Lernansatz zur personalisierten Wohlbefindensabschätzung bei Mitarbeitern im Gesundheitswesen

DIPLOMARBEIT

zur Erlangung des akademischen Grades

**Diplom-Ingenieurin**

im Rahmen des Studiums

eingereicht von

**Lucija Aleksić**

Matrikelnummer 12202117

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Univ. Ass. Gábor Recski, PhD

Mitwirkung: Associate Prof. Milica Vujović, PhD

Wien, 11. November 2025

  
\_\_\_\_\_  
Lucija Aleksić

\_\_\_\_\_  
Gábor Recski



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Modelling Inter-Individual Differences in Multimodal Data

## A Metric Learning Approach for Personalized Well-being Estimation in Healthcare Workers

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

**Diplom-Ingenieurin**

in

**066 645 Master's programme Data Science**

by

**Lucija Aleksić**

Registration Number 12202117

to the Faculty of Informatics

at the TU Wien

Advisor: Univ. Ass. Gábor Recski, PhD

Assistance: Associate Prof. Milica Vujović, PhD

Vienna, November 11, 2025

  
Lucija Aleksić

\_\_\_\_\_  
Gábor Recski



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Erklärung zur Verfassung der Arbeit

Lucija Aleksić

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Ich erkläre weiters, dass ich mich generativer KI-Tools lediglich als Hilfsmittel bedient habe und in der vorliegenden Arbeit mein gestalterischer Einfluss überwiegt. Im Anhang „Übersicht verwendeter Hilfsmittel“ habe ich alle generativen KI-Tools gelistet, die verwendet wurden, und angegeben, wo und wie sie verwendet wurden. Für Textpassagen, die ohne substantielle Änderungen übernommen wurden, habe ich jeweils die von mir formulierten Eingaben (Prompts) und die verwendete IT-Anwendung mit ihrem Produktnamen und Versionsnummer/Datum angegeben.

Wien, 11. November 2025



---

Lucija Aleksić



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Danksagung

Diese Abschlussarbeit markiert das Ende des bisher herausforderndsten Abschnitts meines Lebens. Ohne mein außergewöhnliches Unterstützungsnetzwerk wäre ich heute nicht da, wo ich bin.

Zunächst möchte ich meinem Betreuer und meinem Mitbetreuer für ihre großartige Unterstützung während des gesamten Studiums danken. Ohne ihre Anleitung und Geduld wäre dies nicht möglich gewesen.

Fast während meines gesamten Studiums wurde ich von meinen Kommilitonen unterstützt, und ohne ihre Hilfe wäre ich nicht da, wo ich jetzt bin. Mein Dank gilt auch meinen Freunden – den besten Freunden, die man sich wünschen kann – meinen Universitätskollegen und meinen beiden Katzen, die mich alle auf ihre Weise ermutigt haben.

Vor allem aber möchte ich meinen wunderbaren Eltern, meinem Großvater und meinem Bruder für all ihre Opfer und ihren Trost danken. Ohne sie wäre ich nicht der Mensch, der ich heute bin.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Acknowledgements

This thesis brings an end to the most challenging part of my life so far. I would not be in this position was I not so lucky to be surrounded with such a remarkable support system.

Firstly, I would like to thank my supervisor prof. Gábor and co-supervisor prof. Milica for their tremendous support throughout this journey, this would not be possible without their guidance and patience.

Almost through all of my studies I have been supported by my co-workers, and if it weren't for their help, I would not be where I am now. This gratitude extends to my friends - the best friends anyone can have - my university colleagues, and my two cats who all encouraged me in their own way.

Most importantly, I would like to thank my wonderful parents, grandfather, and brother for all their sacrifice and comfort provided to me. I would not be the person I am today without them.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Kurzfassung

In stark belasteten Arbeitsumgebungen wie Krankenhäusern und Pflegeeinrichtungen ist das Wohlbefinden von Pflegekräften durch die physischen, emotionalen und zeitlichen Anforderungen ihrer Arbeit ständig beeinträchtigt. Diese Belastungen führen schleichend zu einer Beeinträchtigung des Wohlbefindens, wobei frühe Symptome oft subtil sind und häufig erst lange nach ihrem Auftreten bemerkt werden, was ein proaktives Eingreifen erschwert.

Diese Arbeit basiert auf einer Praxisstudie der TU Wien in Zusammenarbeit mit einem Krankenhaus in Hietzing, Wien, an der Pflegekräfte über mehrere Monate teilnahmen. Ziel der Studie war die Modellierung und Vorhersage des täglichen Wohlbefindens mithilfe multimodaler Daten, bestehend aus Daten mobiler Sensoren, halbstrukturierten Interviews, psychologischen Merkmalen aus Fragebögen und Schlafverhalten.

Im ersten Teil wird untersucht, ob große Sprachmodelle die qualitative Kodierung halbstrukturierter Interviews auf dem gleichen Niveau wie ein Mensch durchführen können. Die generierten Codes zeigen eine hohe thematische Übereinstimmung mit den Annotationen von Experten, was darauf hindeutet, dass die LLM-gestützte Kodierung eine zuverlässige und skalierbare Alternative darstellen kann.

Im zweiten Teil wird ein Modell entwickelt, das anhand der ambulanten physiologischen Signale der Teilnehmenden deren tägliches Wohlbefinden (gemessen mit den NASA-TLX-Selbstbeurteilungsbögen) vorhersagt. Die Einbeziehung von Merkmalen aus Interviews verbessert die Vorhersagegenauigkeit zusätzlich.

Abschließend werden die Teilnehmenden anhand ihrer Persönlichkeitsmerkmale und Schlafmuster gruppiert, um zu untersuchen, ob sich die Vorhersagekraft der Modelle dadurch weiter steigern lässt. Obwohl dieser Ansatz theoretisch vielversprechend ist, schränkte die geringe Anzahl an Teilnehmenden pro Gruppe die Modellleistung ein, sodass die Ergebnisse nicht eindeutig sind.

Trotz dieser Herausforderungen zeigen die Ergebnisse, dass eine aussagekräftige Modellierung des Wohlbefindens auch unter realen Feldbedingungen mit verrauschten Daten und begrenzten Ressourcen möglich ist. Diese Arbeit liefert einen Machbarkeitsnachweis für ein skalierbares, datengestütztes Wohlbefindensmonitoring im klinischen Alltag – nicht nur unter kontrollierten Forschungsbedingungen, sondern auch dort, wo es am dringendsten benötigt wird.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Abstract

In high-stress working environments such as hospitals and long-term care facilities, the well-being of healthcare workers is constantly challenged by the physical, emotional, and temporal demands of their work. These burdens gradually erode their well-being, and yet early symptoms are subtle and often self-reported long after they occur, making proactive intervention difficult.

This thesis is based on a real-world study conducted by TU Wien in collaboration with a hospital in the Second Department of Psychiatry and Psychotherapy, Hietzing Clinic, Vienna, where healthcare workers participated in a multi-month field study. The study aimed to model and predict daily well-being using multimodal data which consists of ambulatory sensor data, semi-structured interviews, psychological traits derived from questionnaires, and sleep behavior.

The first part evaluates whether large language models can perform qualitative coding of semi-structured interviews on the same level as a human would. The generated codes show high thematic alignment with expert human annotations, suggesting that LLM-assisted coding can serve as a reliable and scalable alternative.

The second part develops a model to predict participants' daily well-being scores measured by the NASA-TLX self assessments forms from their ambulatory physiological signals. Incorporating features derived from interviews further improves predictive performance.

Finally, participants are clustered based on their personality traits and sleep patterns to investigate would it even further improve predictive ability of the models. While this approach is theoretically promising, the small number of participants per cluster limited model performance, rendering results inconclusive.

In spite of these challenges, the findings demonstrate that meaningful well-being modeling is feasible in real-world field conditions where data are noisy and resources limited. This work provides a proof of concept for scalable data-driven well-being monitoring in everyday clinical practice, not only under controlled research conditions but in the settings where it is needed most.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Contents

<b>Kurzfassung</b>	<b>xi</b>
<b>Abstract</b>	<b>xiii</b>
<b>Contents</b>	<b>xv</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation . . . . .	1
1.2 Problem Statement . . . . .	3
1.3 Research Questions . . . . .	3
1.4 Contributions of this thesis . . . . .	5
1.5 Structure of the thesis . . . . .	6
<b>2 Background and Related Work</b>	<b>9</b>
2.1 Well-being in healthcare workplace . . . . .	9
2.2 Wearable sensors for well-being modelling . . . . .	10
2.3 Qualitative coding . . . . .	10
2.4 Metric learning . . . . .	12
2.5 Clustering in personalised modelling . . . . .	14
<b>3 Methodology</b>	<b>17</b>
3.1 Participants and data collection . . . . .	17
3.2 Data preprocessing and feature extraction . . . . .	20
3.3 RQ1: LLM-based coding . . . . .	26
3.4 RQ2: Modelling framework . . . . .	29
3.5 RQ3: Clustering and personalised modelling . . . . .	35
<b>4 Results and Discussion</b>	<b>39</b>
4.1 Research question 1 . . . . .	39
4.2 Research question 2 . . . . .	43
4.3 Research question 3 . . . . .	47
4.4 Summary of results . . . . .	49
<b>5 Conclusion</b>	<b>51</b>
	xv

5.1 Contributions and Significance . . . . .	51
5.2 Limitations . . . . .	52
5.3 Practical implications for workplace well-being . . . . .	53
5.4 Suggestions for future research . . . . .	53
5.5 Closing statement . . . . .	54
<b>Overview of Generative AI Tools Used</b>	<b>57</b>
<b>List of Figures</b>	<b>59</b>
<b>List of Tables</b>	<b>61</b>
<b>Appendix</b>	<b>63</b>
<b>Bibliography</b>	<b>67</b>

# Introduction

## 1.1 Motivation

Shift work, particularly in demanding fields such as healthcare, is known to have profound effects on employees' physical and mental well-being. Long hours, irregular schedules, and high-stress environments often lead to chronic fatigue, sleep disturbances, and psychological strain [T<sup>+</sup>19]. It is estimated that 35% to 54% of nurses and physicians, and between 45% and 60% of medical students and resident physicians in the United States are impacted by clinician burnout. As a result, understanding and improving employee well-being in such contexts has become a critical area of research, especially when effective interventions can have a direct impact on both individual health outcomes and the quality of patient care.

What is also important to take into consideration is that, even though healthcare workers for example, all work in comparable environments, their physiological responses and subjective perception can differ significantly. This is also noticeable in our dataset; daily activity cycles, light exposure patterns, temperature rhythms, and NASA-TLX well-being distributions vary markedly across participants, despite the same work position.

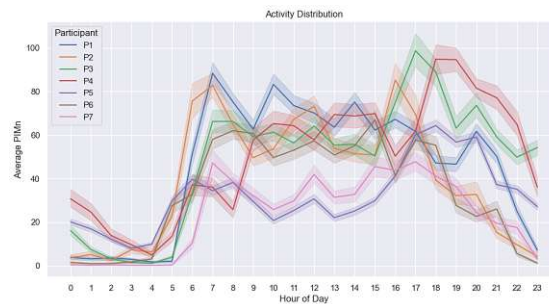


Figure 1.1: Distribution of activity across participants

For example, the figure above shows just how different activity levels are between participants, even though they all work a similar job. Some participants maintain consistently high activity throughout the day, while others exhibit sharp peaks and troughs tied to shift timing or personal routines.

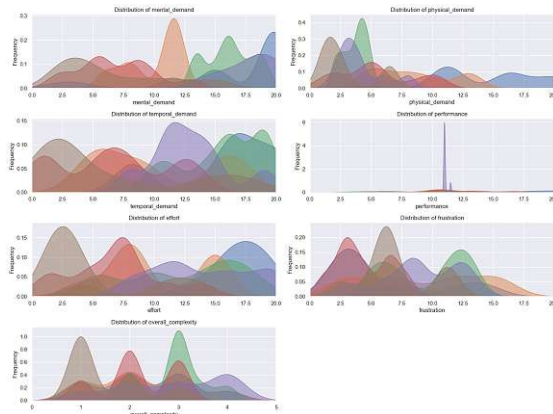


Figure 1.2: Distribution of well-being answers across participants

Likewise, this figure above highlights the variability in perceived mental demand, effort, temporal pressure, and frustration, underscoring that individuals experiencing the same shift pressures may not perceive or internalize stress in the same way.

This phenomenon is consistent with established psychological and physiological line of research. These inter-individual differences in personality traits, sleep efficiency, etc. significantly modulate stress responses [S<sup>+</sup>09]. For instance, people with poor sleep show greater emotional reactivity, but the magnitude of this effect varies widely—experimental studies demonstrate that even under identical sleep restrictions, subjective stress can differ by more than a factor of two between individuals [?]. All of this show that there is a core challenge when modeling any human well-being outcome, and that this personal physiological differences have to be considered. The importance of modeling these inter-individual differences is one of the central aims of this thesis.

The increasing availability of wearable devices and ambulatory sensors has opened new avenues for continuously monitoring workers’ physiological and behavioral states [SRBea18] [Pat22]. This progress has broadened the approaches to the aforementioned problem. This thesis is leveraging exactly this fact of wider availability and the improvements of wearable sensors to model and predict daily well-being under realistic field conditions. The goal is to develop a scalable, data-driven framework that could one day enable early detection of stress and support proactive interventions in occupational health.

## 1.2 Problem Statement

Despite growing awareness of the importance of well-being in clinical professions, its objective and continuous assessment remains a major challenge. Most existing studies rely on self-reported questionnaires or laboratory settings that fail to capture the dynamic and context-dependent nature of stress in real work environments.

Real-world data collection in hospitals is difficult due to privacy concerns, participant burden, and logistical constraints, often resulting in small, heterogeneous datasets [ASCea21] [Pat03]. Moreover, individual differences—such as personality traits, coping mechanisms, or sleep patterns—lead to high variability in physiological and psychological responses to similar stressors, complicating model generalization [CHea].

At the same time, the integration of qualitative data (such as interviews) and quantitative physiological signals remains underexplored. Manual qualitative analysis is a slow and tedious process. This paper [Nea] about developing a rapid deductive approach for such tasks presents an improvement of taking only 409.5 analyst hours to process about 50 audio hours (around 60 interviews) compared to 683 hours during the traditional deductive. This analysis is time-consuming and difficult to scale while automated methods have not yet been systematically evaluated for their validity in psychological research.

This thesis addresses these challenges by investigating how multimodal data—combining ambulatory physiological signals, personality characteristics, and automatically coded interviews—can be used to model and predict daily well-being in healthcare workers. It further explores whether advanced modeling approaches such as metric learning [Kul12] and clustering [Jai99] can capture inter-individual differences and improve predictive performance, even under limited data conditions.

## 1.3 Research Questions

### 1.3.1 RQ1: How well can large language models be used for interview coding?

With the emergence of LLMs, automated coding may offer a scalable alternative, but its validity in real research remains uncertain. First research question is focused on the problem of qualitative coding, more specifically, automating it. *It investigates to what extent can LLMs replicate or improve upon manual coding of interviews in terms of accuracy, consistency, and granularity.*

Traditional qualitative analysis consists of expert annotators to listen to or read pages and pages of interviews and assigning thematic labels. The coding can be conducted in a deductive or inductive fashion [AB]. Deductive qualitative coding uses an pre-existing codes, while inductive coding lets the codes emerge organically from the data itself. After aligning with a TU Wien expert in qualitative analysis, this thesis will focus on developing a pipeline for inductive coding since it was used more broadly. It is unknown

whether LLMs can reliably reproduce similar coding structure, thematic categories, or level of detail. Validation is therefore needed.

The ground truth for needed for this question will be either provided by experts, or approved by experts. This ground truth will be evaluated against LLM-generated codes. The codes themselves will be generated by multi-step prompting, where prompts will also be approved by an expert, and by using different LLM models. The result will be validated with semantic similarity between LLM and human produced codes.

A positive result indicates thematic overlap with human annotation and supports the use of automated qualitative analysis as part of multimodal well-being modeling. It is worth to mention here that a positive result only suggests that LLMs can be useful in the human workflow, but not replace it entirely. Human oversight would still be required to account for potential model biases.

To confirm the positive result an additional manual qualitative analysis will be conducted where the LLM results will manually be compared between models and checked against the ground truth. The best performing model codes will be further used in this thesis as another modality of input.

### 1.3.2 RQ2: What is the predictive power of multimodal data sources?

Current well-being assessments rely on self-report mechanisms such as NASA-TLX questionnaires [HS88]. Were it possible to predict this well-being measure from ambulatory data collected by sensor, it would enable early detection of stress, continuous monitoring in real-world settings and overall improvement in these high-stress workplaces.

Qualitative codes, which indicate a persons perception of their work environment and their personality characteristics derived from the interviews contextualize the response to stress of each participant. As mentioned in the previous chapters, although physiological signals vary greatly across individuals, they contain very valuable information. The key challenge is determining how much predictive value each data modality contributes. The second question investigates just this, the predictive power of different data sources. *More specifically, firstly it will be investigated to what extent can well-being be predicted using only ambulatory physiological data. And then how does model performance improve when incorporating coded interviews.*

For the purposes of this question a baseline model will first be developed, which will only be based on ambulatory sensor data. After that, the extended model, consisting of ambulatory data fused with qualitative codes, will be developed and evaluated against the baseline. The evaluation itself will be conducted using the leave-one-participant-out cross-validation and regression metrics (MSE, MAE,  $R^2$  and correlation coefficients).

This question quantifies the additional predictive value of integrating contextual and psychological information into well-being modeling. It determines whether richer multimodal data leads to more accurate and personalized predictions.

### 1.3.3 RQ3: How can inter-individual differences be modeled?

As mentioned before, the fact that individuals differ in their physiological and psychological stress responses, greatly affects the line of research for predictive well-being modeling. A global model may be too rigid to capture such variability, especially with small and heterogeneous samples common in hospital-based field studies. Studies have shown that building cluster or cohort models greatly improves the performance.

At the start of the study this thesis is based on, participants filled a characteristic questionnaire which consisted of two parts, CARE and D-MEQ. The first part, CARE, refers to the Clinical Activities in the Context of the Work Environment Scale, and it consists of questions like "I frequently interact with my colleagues during my shift" or "Spending enough time with my patients" where the answers are always in the range of 1 (not possible, even with great effort) to 5 (occurs naturally and without effort). The goal of this part is to evaluate the relationship the participant has with its workplace and patients. D-MEQ is the chronotype-focused questionnaire [htt] with questions like "What time would you get up if you were entirely free to plan your day?" or "How alert do you feel during the first half-hour after you wake up in the morning?". It illuminates the participants sleeping habits. Overall, these questionnaires gave insight into the participants traits and their inter-individual differences. This can then be leveraged to perform clustering and building a separate model on each of the clusters. *The third question focuses on exactly that, to what extent does clustering participants based on personality, stress response, or sleep patterns improve the modeling of individual well-being trajectories?* Small sample sizes and uneven cluster distributions may limit the effectiveness of traditional clustering techniques.

## 1.4 Contributions of this thesis

This thesis makes several methodological and practical contributions to the modeling of medical staff well-being in real-world clinical environments, where data collection is limited, heterogeneous, and often noisy.

The first practical byproduct of this thesis is a scalable pipeline for automated qualitative coding using LLMs. The thesis develops and evaluates a complete workflow for transforming semi-structured interview transcripts into qualitative codes using large language models. The results demonstrate that LLM based coding can serve as a reliable, time-efficient, and scalable alternative to manual coding which is a valuable contribution that can significantly reduce researcher workload in future field studies.

The second byproduct ready for field use is an automatic NASA-TLX processing pipeline that converts raw PDF forms into structured JSON suitable for analysis. This will greatly reduce the time effort needed to process these files since one questionnaire had to be filled per working day. The higher quality of the study is wanted, the larger time span it should have and, by default, more files. That only increases this time effort, but this

automation greatly reduces it. This processor will be made available as an open source solution to improve and speed up any future research which is based on these forms.

The thesis builds from scratch a metric learning model for use in non-laboratory settings, where traditional large-scale metric learning is infeasible. The approach jointly learns a participant-specific embedding of ambulatory signals, and a regression model for NASA-TLX-based well-being scores. This hybrid model provides interpretable latent representations and demonstrates improved generalization when qualitative traits are incorporated.

Also, this thesis investigates whether clustering participants by trait measures (CARE, chronotype, sleep patterns) can improve prediction accuracy. Although results are inconclusive due to cluster sizes, the study offers the first empirical analysis of this question in a real clinical environment, laying the groundwork for future studies with larger cohorts.

Overall, the thesis demonstrates that robust, interpretable models of employee well-being can be built even under the constraints typical for real hospitals research; few participants, noisy sensor data, missing information, and limited researcher capacity. This work provides an important proof-of-concept for scalable, multimodal, data-driven monitoring systems that could support early detection of stress and proactive interventions in everyday clinical settings.

### 1.5 Structure of the thesis

This thesis is organized into five chapters, first being *Introduction*. It provides the motivation for studying well-being in clinical work environments, outlines the challenges of modeling well-being in real-world field studies, and presents the research questions and contributions of the thesis.

The next chapter is *Background and Related Work* which reviews existing literature on employee well-being, stress and workload assessment in healthcare, wearable and ambulatory sensing, qualitative coding methods, and machine learning approaches for modeling individual differences.

Following that is the *Methodology* chapter. It describes the data collection process, preprocessing steps for interviews, personality questionnaires, sleep logs, and ambulatory sensor streams. Afterwards, it details the methodological approach for each research question; the evaluation protocol for LLM-based coding (RQ1), the multimodal modeling framework (RQ2), and the clustering and personalized modeling strategy (RQ3). The chapter also outlines implementation details and ethical considerations.

After that comes the *Results* chapter which presents the empirical findings corresponding to each research question. This includes the performance of LLM-assisted coding compared to human coders (RQ1), predictive accuracy of the well-being models using different data modalities (RQ2), and the effects of clustering participants by traits and sleep behavior on model performance (RQ3).

Finally, the last chapter is the *Discussion and Conclusion*. It interprets the findings in the context of prior research, discusses the contributions and limitations of the work, and outlines practical implications for well-being monitoring in healthcare settings. The chapter concludes with suggestions for future research and the broader significance of multimodal, data-driven approaches to occupational well-being.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Background and Related Work

## 2.1 Well-being in healthcare workplace

In our lifetime, nearly all of us will at one point in time have to receive medical care, be it short term or long term. In these situations, healthcare professionals are the primary point of contact and play a central role in the quality of care received. Globally, approximately 142 million older adults are unable to meet their basic needs independently, and two in three older adults will at some point require long-term care services [Org22]. Considering this inevitable need for our society, well-being of these workers becomes even more significant due to the fact that it indirectly affects well-being of their patients. Moreover, a recent meta-analysis of 85 studies (288,581 nurses) found that higher burnout is significantly associated with increased nosocomial infections, medication errors, patient falls, and lower patient satisfaction [L<sup>+</sup>24]. Therefore ensuring that the medical workforce is and stays in good mental and physical health is key to maintaining safe and high-quality care.

Stress in these workplaces comes in all dimensions, physical, mental, temporal and emotional [T<sup>+</sup>19]. Usually, the first to suffer is the workers sleep; *Evidence shows that the effect of shift work on sleep mainly concerns acute sleep loss in connection with night shifts and early morning shifts. A link also exists between shift work and accidents, type 2 diabetes (relative risk range 1.09-1.40), weight gain, coronary heart disease (relative risk 1.23), stroke (relative risk 1.05), and cancer (relative risk range 1.01-1.32), although the original studies showed mixed results. The relations of shift work to cardiometabolic diseases and accidents mimic those with insufficient sleep. Laboratory studies indicate that cardiometabolic stress and cognitive impairments are increased by shift work, as well as by sleep loss [KA16].*

Current state of well-being monitoring still remains a bit bleak. Recent studies in workplace well-being monitoring still show that there is a need for further continuation

and research in this field [J<sup>+</sup>23] [DN<sup>+</sup>23]. In one of these recent studies, testing of the 18 newly developed instruments were largely rated 'Inadequate' and only two were rated as 'Very Good'. None of the studies reported measurement properties of responsiveness, criterion validity, or content validity.

### 2.2 Wearable sensors for well-being modelling

In recent times, the growing availability of consumer-grade and research-grade wearable sensors has had a huge impact on the scientific community and the way they conduct field research [GS15] [P<sup>+</sup>24]. Signals like elevated heart rate, increased perspiration and temperature rise are all stress-related, as well as measurable with today's technology [HRD12]. That is shown today by the fact that many of the smartwatches and similar technologies comes accompanied by some kind of stress-measuring app [Sam] or cardiovascular health assessment app [App]. These sensors provide an objective measurement of physiological and behavioral states and complements the limitation of self-reports like temporal demand, recall bias etc.

For example, in a large sample of healthcare professionals, poor sleep quality and insomnia were significantly linked to the emotional-exhaustion component of burnout, between 57% and 83.2% of shift nurses worldwide report sleep problems (sleep disturbances, deprivation, poor quality). The global prevalence of emotional exhaustion among nurses was around 34.1% [K<sup>+</sup>25]. Cardiovascular health also indicates stress and poor well-being, meta-analytic evidence shows strong negative correlations between stress scores and HRV parameters (e.g., SDNN  $r=-0.58$ , RMSSD  $r=-0.63$ ) and multiple regression confirmed that chronic stress remained a significant predictor of reduced HRV [MEA23]. None of these insights would exist were it not for the advanced state and availability of wearable sensors today.

Being able to measure these physiological data and correlate it to stress and well-being opens enables a line of research in continuous monitoring, especially in shift-work environments [M<sup>+</sup>20b].

### 2.3 Qualitative coding

Qualitative coding is the process of systematically assigning labels or codes to interview segments. Therefore it is a fundamental step for transforming raw narratives into interpretable themes and constructs that can be compared across participants and linked to quantitative measures [Sal16]. Qualitative interviews are a widely used instrument in any research, but especially in occupational health and well-being research because they capture how workers themselves make sense of their work environment, and stressors [BC06]. Unlike pure quantitative indicators, interview-based accounts provide more contextual information about the workplace, interpersonal dynamics, and perceived demands and resources, which are all crucial for understanding why similar working conditions can lead to different well-being outcomes for different individuals.

Traditionally, methods like thematic analysis or grounded theory are used to analyze the interviews. In thematic analysis researchers iteratively generate initial codes, collect them into candidate themes, review their coherence against the data, and refine them into a final thematic structure [GMN12]. Grounded theory inspired approaches start similarly with open coding and gradually move toward more theoretical constructs [Sal16]. All the traditional methods, including the aforementioned two, require domain expertise, thorough reading of the materials and repetitive process of extracting codes. Most importantly, all of these show us that the most sought after resource is time.

Also, maintaining coding consistency is challenging. As qualitative datasets grow larger the cost and time required for manual coding can become a major bottleneck, limiting the feasibility of regular well-being monitoring in real-world settings. In my own qualitative coding experience, where interview recording ranged from 10 minutes to 1 hour, it took approximately 3 hours per interview and it was a rather exhausting task. Most challenging was to keep the same coding consistency since this task spanned over a long time. That is where the true need for an automation of this task becomes visible.

Recent advancements in large language models have sparked growing interest in using these models to accelerate and partially or fully automate qualitative coding, among other things. Early empirical studies show that LLMs can generate initial codes and thematic structures that exhibit moderate to substantial conceptual overlap with those produced by trained human coders. In this 2025 study [KW25], it is shown that by using carefully design prompts and including a human-in-the-loop in the validation, that LLMs can assist in coding with codes substantially aligned with human codes. Another study conducted in 2025 [MFM<sup>+</sup>25] also demonstrates the usefulness of LLMs in the qualitative coding process, by concluding that evaluators preferred LLM generated codes 61% of the time, finding them analytically useful for answering the research question. It is also important to mention that evaluators also identified limitations (LLMs fragmented data unnecessarily, missed latent interpretations, and sometimes produced themes with unclear boundaries). That is why it is important to emphasize that LLMs are only tools meant to support the qualitative coding process and complement human analysis rather than replacing it.

For example, [MZP<sup>+</sup>24] compared an open-source LLM with a standard human-led thematic analysis of healthcare interviews and found that the AI-generated themes captured many of the same conceptual patterns identified by qualitative experts. They found similarities ranging from moderate to substantial (Jaccard similarity coefficients 0.44-0.69). In this case, since Jaccard similarity was used, they had pre-defined codes which made the task easier, but still is an impressive result.

Systematic LLM reviews document a rapid increase in research applying LLMs to qualitative tasks. Tasks range from automated code suggestion and mapping the texts to existing codebooks to inductive theme discovery and summarization of large interview corpora across domains such as health, education, psychology, and the social sciences [BO24].

To summarize, these findings strongly suggest that LLMs can serve as effective analytical assistants, especially in the early stages of analysis where the researchers generate preliminary codes or examine large volumes of text to detect emerging patterns. It would significantly reduce the time needed for the initial coding process and it may even help researchers overcome some of the scalability constraints of manual qualitative analysis. It's important to emphasize that these tools do not replace human interpretation but can augment it, allowing experts to dedicate more time to higher-level conceptualisation, methodological reflection, and the interpretive richness that remains central to qualitative research.

## 2.4 Metric learning

Classic machine learning methods are based on the principle to predict an output from an input  $f(x) \rightarrow y$ . However, metric learning refers to a class of machine learning methods whose goal is not to learn the input-output relation, but to learn a distance function that reflects meaningful similarity relationships between input data points [Kul13]. Rather than rely on metrics like Euclidean distance, metric learning relies on an embedding space in which semantically similar samples are close together and dissimilar samples are far apart. That is achieved in learning the distance function  $d_\theta(x_i, x_j)$  that reflects the semantic similarity between data points. The whole idea of metric learning is based on the distance between inputs, more specifically the distance itself becomes a learnable quantity. End result would be a model which can output how similar or dissimilar two samples are, rather than an explicit class label. This approach is highly relevant for human-related data, where relationships between observations often carry more information than absolute labels.

A distance function can be parameterized directly or induced by a learned embedding function  $f_\theta : R^n \rightarrow R^k$  which defines the distance function:

$$d_\theta(x_i, x_j) = \|f_\theta(x_i) - f_\theta(x_j)\|_2$$

To enable the model to learn this several families of loss functions have been developed, each dealing with similarity/dissimilarity constraints differently.

Contrastive loss is based on a binary label  $y_{ij} \in \{0, 1\}$  which indicates whether the  $i$ -th and  $j$ -th points are similar or dissimilar. The goal is to learn an embedding function  $f_\theta(\cdot)$  such that similar pairs have a small distance and dissimilar pairs are separated by at least a margin  $m$ . Therefore, the loss pulls positive pairs together and pushes negative pairs apart.

$$\mathcal{L}_{\text{contrastive}} = y_{ij} \|f_\theta(x_i) - f_\theta(x_j)\|_2^2 + (1 - y_{ij}) \max(0, m - \|f_\theta(x_i) - f_\theta(x_j)\|_2)^2$$

A very powerful metric introduced in the FaceNet model in 2014 [HA14] is the triplet loss. It is based around the idea of an anchor point and positive and negative points. The

objective is to enforce that the anchor point is closer to the positive than to the negative by at least a margin  $\alpha$ , thereby enforcing a discriminative structure in the embedding space.

$$\mathcal{L}_{\text{triplet}} = \max \left( 0, \|f_{\theta}(a) - f_{\theta}(p)\|_2^2 - \|f_{\theta}(a) - f_{\theta}(n)\|_2^2 + \alpha \right)$$

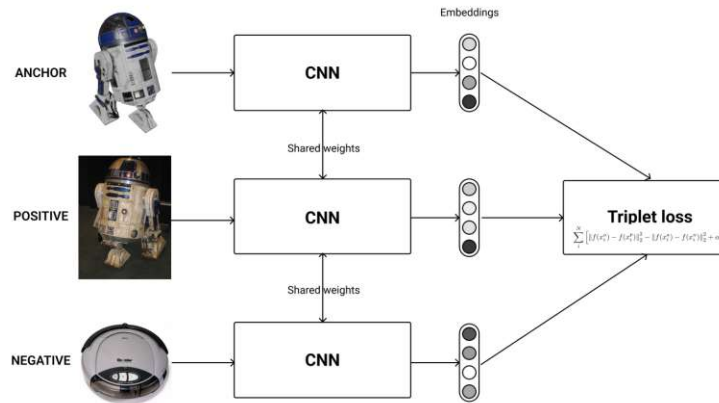


Figure 2.1: Visualization of the triplet loss

This figure illustrates the typical architecture used for triplet loss training. An anchor point, a semantically similar positive point, and a dissimilar negative point are passed through the same neural network encoder with shared weights. That is to make sure that all three inputs are mapped into a common embedding space.

In metric regression, the model is given a continuous target distance  $d_{ij}^{\text{target}}$  that expresses how similar two samples should be. The loss directly regresses the embedding distance  $\|f_{\theta}(x_i) - f_{\theta}(x_j)\|_2$  toward this value. This enables learning continuous similarity relationships rather than discrete same/different labels.

$$\mathcal{L}_{\text{metric-reg}} = \left( \|f_{\theta}(x_i) - f_{\theta}(x_j)\|_2 - d_{ij}^{\text{target}} \right)^2$$

This is a very powerful loss since the similarity is graded, not binary. In the case of this thesis using a binary loss, like contrastive loss, was unfeasible due to the small sample size and large variability. Doing so would either have a large disbalance of similar/dissimilar samples or require a too large a threshold for similarity, and both of them lower the quality of the model and prediction.

The early days of metric learning started with applying the aforementioned triplet loss to deep neural networks, which proved to be a drastic improvement [HA14]. It's natural support of continuous labels and learning useful semantic representations of data gave it a wide area of application. One of the biggest breakthroughs in this area came in 2015 when

Google developed FaceNet, a model for face embeddings and face recognition [SKP15]. In their model they had also used triplet loss to achieve distinguishing between input images of faces. Recently, another promising area of application for metric learning was discovered as recommender systems, where the learning process traditionally depended only on user-item interaction where there were no semantic information. It was discovered that by using metric learning this information can be further used [LZYH24]. This makes metric learning particularly well suited for problems involving heterogeneity, limited data, or continuous similarity relationships which are common in human-centric datasets.

Advantages of metric learning over classical machine learning, e.g. regression, are the reason why it's chosen for tasks as predicting well-being. Using pair-wise combinations ( $N^2$ ) for training drastically increases the training samples from ( $N$ ), making the metric learning approach better suited for cases with limited sample size, as such. Few-shot learning papers [S<sup>+</sup>17] show that embedding-space learning dramatically improves performance in low-data settings. Furthermore, people have different physiological baselines. Metric learning learns a personalized similarity space instead of a global predictor, as is the case in classic regression.

## 2.5 Clustering in personalised modelling

Clustering is an unsupervised machine learning mechanism that groups samples into subsets or clusters by the similarity of the samples; the similar samples are within one cluster and dissimilar are in other clusters [Jai99]. This similarity is typically quantified via a distance metric  $d_\theta(x_i, x_j)$ , most commonly Euclidean distance:

$$d(x_i, x_j) = \|x_i - x_j\|_2$$

The goal is to uncover a latent structure in the data without knowing the labels of examples, therefore being an unsupervised technique. This is done by using an objective function which by optimization puts similar samples together and penalizes the separation of them.

Clustering techniques cover a wide range of algorithms. One of the most used algorithms is the k-means clustering which minimizes the sum of squared distances within the cluster:

$$\arg \min_{\{C_k\}_{k=1}^K} \sum_{k=1}^K \sum_{x_i \in C_k} \|x_i - \mu_k\|_2^2$$

where  $\mu_k$  is the centroid of cluster  $k$ .

On the other hand, algorithms like Gaussian Mixture Models (GMMs) assume that data is generated by a mixture of multiple Gaussian distributions. This makes it useful particularly when the data has multiple clusters or a complex distribution that a single Gaussian cannot capture.

Basically, the idea of clustering lies in the assumption that samples have underlying latent clusters which need to be discovered. This assumption aligns well with human centered data, where individuals are naturally grouped by differences in behaviour, physiology or stress responses.

As a scientific field, clustering has developed significantly over the last decades. It started with distance based methods, like k-means [Mac67], and now contains probabilistic models, like GMMs and density based methods, like DBSCAN. The change came when latent representations were started to be used and shifted the trend toward representation-aware clustering [XGF15, C<sup>+</sup>19]. The clustering is now mostly performed on a latent vector space rather than raw features. This is particularly important for human behavioural datasets, where raw input spaces (e.g., physiological sensor readings) are noisy and highly individualized [Luo24].

As such, clustering has a wide array of applications, from computer vision and natural language field, to recommender systems. Since the main idea is finding natural hidden clusters in data, it is very used in human modelling, that is to help alleviate the heterogeneity problem; the fact that individuals respond differently to identical inputs [QST<sup>+</sup>23, LAGF13].



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Methodology

## 3.1 Participants and data collection

Seven healthcare workers from a hospital in Second Department of Psychiatry and Psychotherapy, Hietzing Clinic, Vienna voluntarily participated in a field study conducted by TU Wien in collaboration with the clinical institution. Their roles and daily obligations differed.

- Participant 1 (P1): Nurse working primarily in a therapy garden; spends extensive time outdoors with patients.
- Participant 2 (P2), Participant 6 (P6), Participant 7 (P7): Ward-based nurses with substantial administrative and coordination duties; interact regularly with patients, physicians, and other staff but spend less time in direct patient care.
- Participant 3 (P3): Nurse manager responsible for organizational and administrative oversight; works mostly in an office separate from the ward.
- Participant 4 (P4) and Participant 5 (P5): The only participants with both day and night shifts; full rotating-shift exposure.

The duration of their participation varied, ranging from approximately two weeks to two months, depending on scheduling feasibility and their own will to participate and wear the sensors. Aside from their usual work obligations, they collected different data; wearable chest and wrist sensor worn at all times collected physiological data was continuous throughout the study period, self-report well-being questionnaires (NASA-TLX) were completed after every shift, a personality questionnaire was completed at the beginning of the study as well as a semi-structured interviews about their workplace. Observational field notes were recorded by researchers for approximately 20 hours per participant across different shifts but they're not used in this thesis.

### 3.1.1 Ambulatory physiological data

The wearable sensor is the ActLumus 2 actigraph [Act]. The ActTrust device is widely used in chronobiology and sleep research and is capable of recording:

- Triaxial accelerometer activity
- Light exposure, including:
  - photopic illuminance
  - melanopic equivalent daylight illuminance (EDI)
- Skin temperature
- Derived sleep–wake patterns, using established actigraphy algorithms

The NASA Task Load Index questionnaire [HS88], a standardized tool for assessing perceived workload across six dimensions; mental demand, *physical demand*, *temporal demand*, *performance*, *effort*, *frustration* and *overall complexity*.

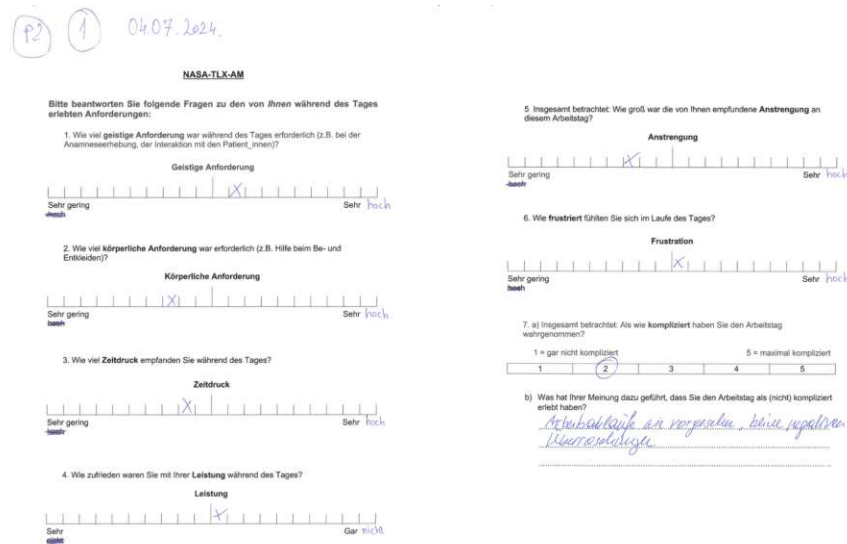


Figure 3.1: Example of a scan of the daily NASA-TLX form

### 3.1.2 Interview data and coding

As mentioned, each participant took part in a semi-structured interview aimed to capture their subjective experiences of their work environment. Semi-structured interviews were selected because they balance consistency (all participants addressed comparable themes)

with flexibility, allowing individuals to talk more on details that are personally meaningful or role-specific. This was particularly important in a diverse hospital setting where participants differed in responsibilities, exposure to patients, administrative burdens, mobility across wards, and shift patterns.

The interview guide covered several key domains

1. Description of the physical and social work environment (ward layout, noise, lighting, access to outdoor spaces, airflow, furniture ergonomics, office availability).
2. Interactions with colleagues, patients, and interdisciplinary teams, including how these interactions influence stress and workload.
3. Perceived safety and environmental stressors, such as malfunctioning alarm systems, patient aggression, temperature/ventilation issues, or chaotic base stations.
4. Organizational and architectural factors, including access to retreat spaces, availability of quiet rooms, workflow interruptions, and technical constraints.
5. Coping behaviors and personal well-being strategies, ranging from micro-breaks, outdoor walks, and structured work–life boundaries to personal “reset rituals” such as juggling or stepping away for brief recovery.
6. Perceived impact of the environment on patient care, morale, staff cohesion, and emotional exhaustion.

When analyzed, these interviews showed some themes which were present in almost all interviews; chronic noise exposure, limited protected space for concentration, unreliable safety systems (e.g., faulty alarms or poor mobile reception), physical discomfort (heat, ventilation issues), and the occasional feeling of "tied-hands" when wanted to improve something due to higher management or government institutions.

Ward managers emphasized administrative interruptions and the lack of private workspace. Staff working directly with patients described safety concerns, noise, and exposure to aggression as primary stressors. Participants with access to outdoor therapeutic spaces or gardens reported these as very important relief environment for both them and the patients. All interviews were recorded and transcribed, and then later on used for qualitative coding.

#### 3.1.3 Personality & chronotype questionnaires

Another type of questionnaire filled in by participants was "personality" questionnaire. It consisted of two main components: (1) the CARE scale that assesses perceived workload integration in the clinical environment, and (2) the Diurnal/Chronotype (DMEQ) questionnaire measuring morning-evening-day tendencies and sleep–wake preferences.

### CARE Scale (Clinical Activities in the Work Environment)

The *CARE questionnaire* evaluates tasks and perception of the workplace. Items assess constructs such as teamwork, communication, efficiency of movement on the ward, patient documentation, and responsiveness during a shift. Participants rate each statement on a five-point Likert scale ranging from 1 = “not possible / requires a lot of effort” to 5 = “occurs naturally and without effort.” Example of some questions are *Interacting frequently with my colleagues during my shift:* or *Moving efficiently around my ward (i.e., without detours, with manageable distances between locations):*.

### Deutsch Morningness–Eveningness Questionnaire (D-MEQ)

The second part was about the chronotype’ of the person, commonly referred to as a morningness–eveningness questionnaire, DMEQ; *Deutsch Morning-Eveningness Questionnaire*. This instrument evaluates biologically and behaviorally anchored sleep–wake preferences, including preferred time for mental and physical activity, subjective alertness in the morning and other aspects about the sleeping habits of the participant.

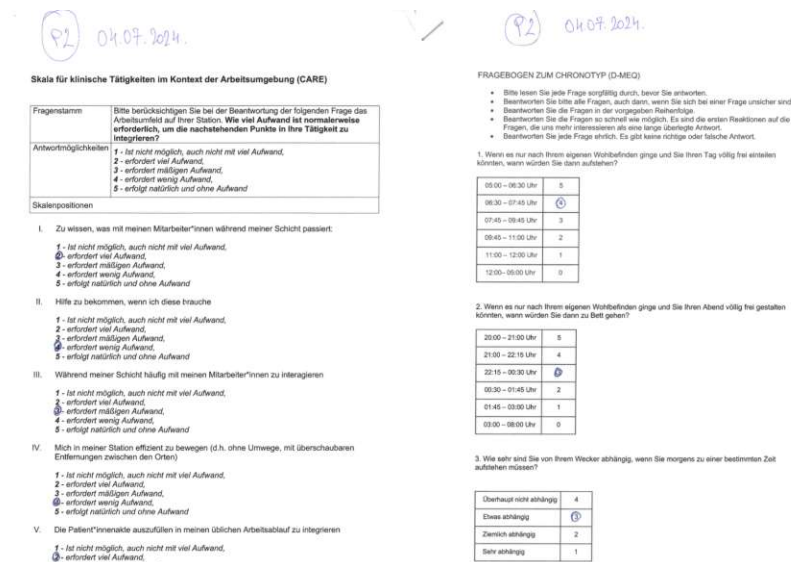


Figure 3.2: Left: CARE page, Right: D-MEQ page

## 3.2 Data preprocessing and feature extraction

### 3.2.1 Feature extraction and selection

The raw ambulatory data collected from the Condor ActLumus sensors consisted of 33 different features. Sensors gave readings each minute of the day. Before this data can be

used, they should be processed into a consistent and sufficient set of features.

Timestamps (DATE/TIME) were converted into unified datetime objects, following a day-first format consistent across European hospital systems. Missing or invalid measurements (e.g. placeholder zeros, negative light values, sensor saturation) were replaced with NaN. Constant or near-constant columns (e.g. broken spectral channels) were automatically removed using a variance threshold filter. This step is critical in wearable sensing pipelines, as constant features only add unnecessary noise or sensor drift produces non-informative signals.

Feature engineering was also employed to transform one or more sensors into higher-level features and add them before the feature selection, since they could give more information than raw sensors. PIM (Physical Activity Intensity Monotonic), TAT (Time Above Threshold) and ZCM (Zero Crossing Mode) were combined into a standardized activity level:

$$ACTIVITY\_LEVEL = \frac{meanPIM, TAT, ZCN_t - \mu}{\sigma}$$

This captures moment-to-moment intensity fluctuations and allows cross-participant comparability.

Light and spectrum exposure is highly relevant for circadian rhythm, fatigue, and alertness regulation [Pat22]. Spectral sum feature was created:

$$SPECTRAL\_SUM = RED + GREEN + BLUE + IR + UVA + UVB$$

Blue-light fraction refers to the proportion of blue light in a light source's spectrum, which influences plant growth and physiology.

$$BLUE\_FRACTION = \frac{BLUE}{SPECTRAL\_SUM + 10^{-6}}$$

Likewise for the red-light fraction.

$$RED\_FRACTION = \frac{RED}{SPECTRAL\_SUM + 10^{-6}}$$

These encode environmental lighting quality relevant to shift work and circadian disruption.

Skin temperature and external ambient temperature produced informative derived features, first being temperature difference:

$$TEMP\_DIFF = SkinTemperature - ExternalTemperature$$

Large differences may indicate stress, physical exertion, or environmental load.

### 3. METHODOLOGY

---

Capacitive sensors (CAP\_SENS\_1, CAP\_SENS\_2) were averaged to obtain CAP\_MEAN which is a proxy for sensor-skin contact quality.

The device provides frequency bands F1, ..., F8 which are derived from raw activity spectra. Because raw band values vary strongly across participants and sessions, we computed relative ratios:

$$F_i^{ratio} = \frac{F_i}{\sum_{k=1}^8 F_k + 10^{-6}}$$

These features capture distribution of movement frequencies, useful for differentiating fine motor activity, walking, running, and rest.

To reduce dimensionality and stabilize cross-participant representation, PCA was applied to the 8 bands and only the first three principal components were retained, FREQ\_PCA1, FREQ\_PCA2, FREQ\_PCA3, which capture more than 90% of spectral variance in most participants.

A key methodological challenge is that not all ActLumus sensors recorded the same channels for all participants. For example, IR LIGHT was missing or unreliable in some participants.

Some participants had additional spectral or frequency-ratio features (F6\_RATIO, F8\_RATIO, etc.). P5 included high-quality TAT data, absent in others.

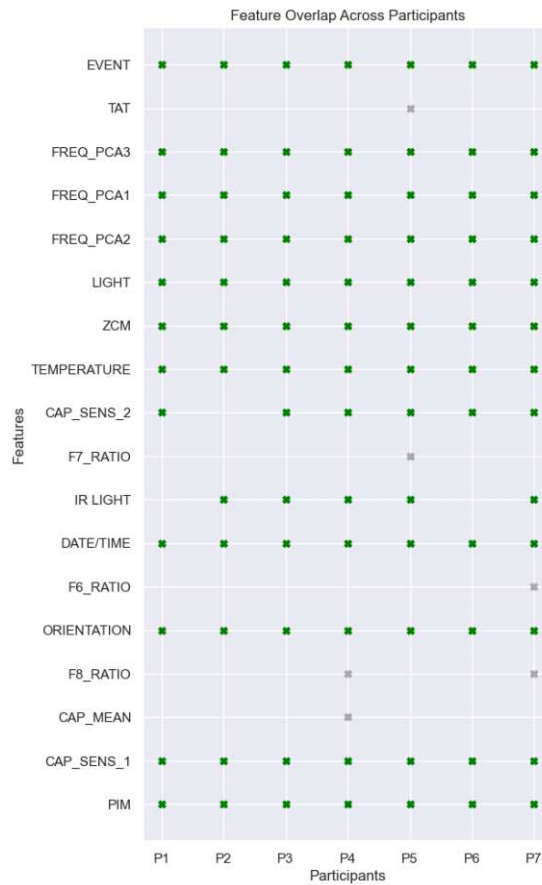


Figure 3.3: Visualization of selected features for each participant

To ensure the metric-learning model receives a consistent feature vector, we computed the feature overlap matrix, visible in the figure above, and retained only features present in the majority of the participants.

The resulting stable feature set: *'DATE/TIME'*, *'EVENT'*, *'TEMPERATURE'*, *'ORIENTATION'*, *'PIM'*, *'ZCM'*, *'LIGHT'*, *'CAP\_SENS\_1'*, *'CAP\_SENS\_2'*, *'FREQ\_PCA1'*, *'FREQ\_PCA2'*, *'FREQ\_PCA3'*, *'IR LIGHT'*

### 3.2.2 Dataset creation

The dataset used for training the metric-learning model was fully custom due to the fact that the model was made from scratch. The pipeline transforms raw ambulatory time series and NASA-TLX self-reports into structured input-output windows, and subsequently into pairwise training instances appropriate for contrastive and triplet-based metric learning.

The aggregation of ambulatory data is the first important part. In essence, it takes in a

timeframe of ambulatory data and produces the mean, standard deviation, minimum, maximum and median of that data.

$$\text{agg}(x_t) = \{\text{mean}, \text{std}, \text{min}, \text{max}, \text{median}\}$$

Daily aggregation serves two purposes, temporal alignment with NASA-TLX, which is recorded once a day and noise reduction by smoothing high-frequency sensor variability common in real-world wearable studies.

Well-being is not only influenced by the day on which it is reported, but also by the days leading up to it (sleep debt, accumulated stress, workload cycles, etc.). To capture this temporal context, we construct sliding windows of multiple consecutive days.

For each window size

$$w \in [1, 3, 5, 7]$$

we slide a temporal window across the participant's daily data

$$W_{i,w} = \{s_i, s_{i+1}, \dots, s_{i+w-1}\}$$

Each window is represented by its mean feature vector:

$$\mathbf{x}_{i,w} = \frac{1}{w} \sum_{k=i}^{i+w-1} \mathbf{s}_k$$

For each window, any NASA-TLX response whose timestamp falls within the window is matched to that window. A NASA-TLX entry is represented as a 7-dimensional vector,  $y = \text{mental demand}, \text{physical demand}, \text{temporal demand}, \text{performance}, \text{effort}, \text{frustration}, \text{overall complexity}$ , where all but overall complexity have values from 1 to 20, and overall complexity is in the range from 1 to 5.

To encode where inside the window a TLX report occurs, we compute a normalized position relative to the window size. This positional embedding allows the model to distinguish a report occurring on day 1 versus day 7 of a temporal window.

Each resulting sample therefore consists of:

- aggregated input vector  $x_{i,w}$
- output vector  $y_{NASA}$
- positional encoding  $p$
- window size  $w$
- participant ID

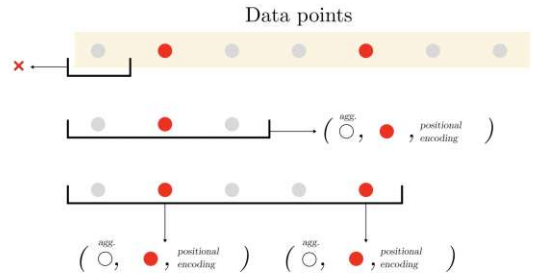


Figure 3.4: Visualization of the build windows algorithm

Figure above shows a simplified visualization of the input-output pair building algorithm. If there is no output form on that day, no pairs are produced, as is the case in the  $w = 1$  case in this example. Where there is an output, as is in the  $w = 3$  case, it puts together the input and output, alongside the positional encoding. If there are more than one output in the window, more pairs are produced, each with its own output and positional encoding, like in the  $w = 7$  case.

Metric learning requires pairs of samples, not individual points. Therefore, we built a custom PyTorch dataset, `PairDataset`, which forms all pairwise combinations of windowed samples:

$$(a, b) \in \text{combination}(\text{io\_pairs}, 2)$$

For each pair, the target distance is defined as the Euclidean distance between their normalized NASA-TLX vectors:

$$d_{\text{target}}(a, b) = \|\mathbf{y}'_a - \mathbf{y}'_b\|_2$$

This makes the model learn a latent space where days with similar subjective well-being are close, and days with very different well-being profiles are far apart. That is very different than classic machine learning algorithms which use inputs and targets directly. By transforming subjective well-being reports into inter-day distances, we provide a dense supervision signal that is much richer than classification or regression alone, and more fit to this type of problem.

What each training example contains

Each sample in the dataset includes:

- input features for both pairs  $x_1, x_2$
- normalized TLX outputs  $y'_1, y'_2$

- positional encodings  $p_1, p_2$
- ground-truth distance between outputs  $d_{target}$
- participant ID

### 3.3 RQ1: LLM-based coding

To combat over-generalization and improve the code generation, this LLM coding pipeline used a two-stage prompting protocol inspired by chain-of-thought prompting frameworks. Chain-of-thought prompting encourages models to reason through a task by producing intermediate interpretive steps rather than jumping directly to a final answer [WWS<sup>+</sup>22]. Prior research has shown that adding reasoning steps improve performance on such tasks where models must infer latent structure from long-form text [KGR<sup>+</sup>23].

There have been instances where a two-stage prompting significantly improved LLM performance [WSL<sup>+</sup>24]. This can be applied in classical qualitative workflows and aligns with findings that LLMs perform best when prompted to revise and structure their own prior output.

The full initial and revision prompts used in the LLM coding workflow are provided in the Appendix. The main components of each prompt included:

- a role specification instructing the model to act as a qualitative analysis expert
- constraints on the number and structure of output codes
- examples of good and bad code formulations
- explicit instructions for merging and abstracting codes during the revision step

Below is a snippet from the initial prompt:

```
You are an expert in qualitative content analysis. Analyze the following interview.
```

```
**Your task is to identify a few, but meaningful, open-ended codes.  
** Each code should describe a central theme, perspective, or experience of the interviewee. The goal is to cover the entire interview with as few codes as possible.
```

```
**Instructions:**
```

- Use **as few codes as possible** (ideally **2-5**, maximum 6).
- A good code **summarizes similar statements** under an overarching theme.
- Formulate concise, self-explanatory code names (2-5 words).

- **\*\*Do not repeat content in multiple codes.\*\*** Summarize related aspects where possible.

...

And here is an example from the revision prompt:

You are an expert in qualitative content analysis and are tasked with revising existing

Below is a list of codes extracted from an interview. Your task is to revise these codes by:

1. Combining similar or overlapping codes if they describe the same topic or related aspects.
2. Replacing overly specific or lengthy code names with more abstract, general terms that still capture the central theme.
3. Formulating each code name concisely and clearly (ideally 1-4 words).
4. Preserving the original content by providing a revised but complete description for each new code.
5. Generating a maximum of 3-5 codes in the final result.

...

The whole goal of this qualitative coding is to produce a set of codes and their descriptions that best match the topics covered in the coded text. An example from a human expert is below:

Code example:

- **Role responsibility:** The participant describes their comprehensive role in coordinating patients, staff, and resources, as well as the associated responsibilities.
- **Infrastructure deficiencies:** Criticism of the building design, focusing on inadequate cooling and inefficient storage options that negatively impact working conditions.
- **Team dynamics:** An open communication culture that values diverse opinions and fosters constructive problem-solving discussions within the team.
- **Patient interaction:** Challenges in patient care due to insufficient therapy options and suboptimally designed patient areas.
- **Work environment:** The impact of personal retreats and stressors on the work environment, including stress caused by specific workspaces.

To be able to evaluate if the LLMs can support or partially automate qualitative coding in this real-world healthcare context, a custom multi-stage evaluation pipeline was developed. The goal is to compare machine-generated codes against a human-annotated ground truth and to quantify the semantic similarity and reliability of LLM coding across participants and models. The framework itself is designed to be model-agnostic. All models are wrapped in a common abstract interface, `LLMProvider`, which is an abstract class defining a `generate(prompt)` method.

### 3. METHODOLOGY

---

Concrete implementations are:

- GPTProvider: GPT-4o, GPT-4o-turbo, GPT-3.5-turbo
- OllamaProvider: mistral, phi3
- DeepSeek: deepseek-chat

This abstraction ensures that each model follows exactly the same procedural steps during coding.

By researching the process of qualitative coding from experts, it is done iteratively, initial codes are usually rewritten in a way to merge similar codes, or rephrase codes etc. That is why the LLM interview coding process consists of two stages (similar to an iterative thematic analysis):

1. *Initial coding*: Each model receives a template prompt (`initial_prompt.txt`) and the raw interview transcript. The prompt asks the LLM to produce an initial set of descriptive codes capturing themes, environmental influences, stressors, coping strategies, and contextual factors.
2. *Revision stage*: Each model then receives the previous initial code list from that model, and a second prompt (`revision_prompt.txt`) instructing the model to refine, merge, group, and clarify codes.

The resulting revised code list is saved and used for evaluation. This two-stage workflow tries to mimic a human workflow where initial codes are again later refined into a final, more coherent code structure.

To quantify how similar machine-generated codes are to human-coded themes, we compute semantic similarity between code sets. This step is implemented in `EvaluationPipeline`. To make codes comparable they need to be in some way embedded into a vector space. The multilingual sentence-transformer `paraphrase-MiniLM-L12-v2`, which has been shown to provide high-quality semantic representations for short textual units such as qualitative codes [Han24, DMM24], was used for this purpose

The metric of similarity is the cosine similarity between the embeddings produced by the sentence transformer  $E(\cdot)$  on human codes  $c_i$  and LLMs  $d_j$ :

$$S_{ij} = \cos(E(c_i), E(d_j))$$

Cosine similarity is the most appropriate similarity measure in this context because the coding involves open-ended and not predefined qualitative codes that differ in wording, granularity, and conceptual scope across models and even human coders. If the case is that the codes are predefined and the task is inductive coding, Cohen's  $\kappa$  is usually used [Dun24]. Cohen's  $\kappa$  explicitly requires a confusion matrix over an identical label space and penalizes all differences equally, even when two codes are semantically close but lexically different. In deductive coding, where categories emerge themselves and are often phrased differently (e.g., "technical challenges" vs. "infrastructure deficiencies"), the Cohen's  $\kappa$  would trivially collapse toward zero, and it would wrongly imply coding disagreement, where only naming variation exists. Jaccard similarity would also fail in such circumstances where there aren't predefined codes and codebook to use.

Since cosine similarity is based on embeddings instead measures the semantic proximity between codes by comparing their vector representations it allows the evaluation to capture paraphrasing and shared conceptual meaning—properties. That is why it is chosen for measuring this task.

## 3.4 RQ2: Modelling framework

### 3.4.1 NASA Form Processing

The aforementioned NASA-TLX questionnaires were completed by hand on paper and later scanned as two-page PDF files. This form is commonly used by researchers at TU Wien and, due to its physical nature, it takes a lot of time to process them into something machine readable, e.g. JSON. The goal of this preprocessing stage was to automatically extract the six standard NASA-TLX scores (mental, physical, temporal demand, performance, effort, and frustration) along with the additional “overall complexity” score used in the study. Another issue were the hand-written dates on the top of the page which indicate when they were taken. Because of all these issues traditional OCR tools failed. Relative ROI definitions were tried, but since papers were not all scanned the same, it didn’t work. Neither did creating a blank template to subtract the pixels. Vision LLMs were also tried (GPT-4o), but due to overt hallucination on scores it was discarded. However, GPT-4o showed very promising results in the hand-written date detection. Therefore, we developed a custom, fully vision-based processing pipeline tailored specifically to the visual structure of the NASA-TLX form.

Each NASA-TLX item is answered by marking a symbol (X, circle, tick, underline, dot) along a horizontal line with printed tick marks. The first challenge is to find these horizontal reference lines despite noise, shadows, pen pressure variation, and scanner artifacts.

The following steps are applied to each scanned page:

1. Noise suppression and contrast normalization
  - Gaussian blur
  - Bilateral filtering
  - Contrast Limited Adaptive Histogram Equalization (CLAHE)
2. Adaptive thresholding to isolate dark structures from the background
3. Morphological operations (opening with a wide, flat kernel) to enhance long horizontal strokes.
4. Hough Line Transform to detect horizontal line segments. Only lines which are:
 
$$|y_2 - y_1| < 10$$
 remain.
5. Line merging and filtering Lines that occur within a vertical distance more than 15 px are merged. Lines shorter than 50% of the page width are discarded.

For each detected slider line  $x_1, y_1, x_2, y_2$ , a crop rectangle is defined by proportional margins:

$$\begin{aligned}\Delta_{top} &= 0.06 \cdot (x_2 - x_1) \\ \Delta_{bottom} &= 0.01 \cdot (x_2 - x_1) \\ \Delta_{left} &= 0.025 \cdot (x_2 - x_1) \\ \Delta_{right} &= 0.025 \cdot (x_2 - x_1)\end{aligned}$$

Each crop isolates exactly one slider bar and its tick marks. Then each of the crops is divided into  $n_{crops} = 2N + 1$ , in this case 41. The crop is split along the x-axis into equally wide bins. That is due to the fact that a person can mark the scale as 1, 1.5, 2, 2.5, etc. This binning transforms the problem of mark detection into finding a bin whose pixel statistics are maximally different from the surrounding bins.

The forms were filled in with every possible colour, from black to green. To handle both consistently, the pipeline first checks for strong colour channel differences. If there is a coloured mark the average intensities of R, G, B channels  $(\bar{r}, \bar{g}, \bar{b})$  can be taken advantage of. For each channel pair:

$$d_{RG} = |\bar{r} - \bar{g}|, d_{RB} = |\bar{r} - \bar{b}|, d_{GB} = |\bar{b} - \bar{g}|$$

if  $\max(d_{RG}, d_{RB}, d_{GB}) > 10$  the bin with the largest difference is selected as the marked position.

However, if the previous condition isn't satisfied, it is considered to be a black pen. Then, the crop is converted to grayscale, and each bin's mean pixel value  $\bar{p}_i$  is computed.

#### 3.4.2 Metric Learning Model

Once the well-being outputs are processed and dataset is created, the model can be trained. The model has a hybrid metric learning architecture. It has two objectives; learn a latent embedding space in which distances correspond to differences in NASA-TLX well-being profiles, and predict the full NASA-TLX score vector from each daily window. That resulted in the model having an encoder part and a regressor part.

The encoder  $f_\theta(\cdot)$  is a network which deals only with ambulatory data and it learns how to map input vector features  $x \in R^d$  (aggregated ambulatory features) into a lower-dimensional embedding:

$$z = f_\theta(x) = MLP_\theta(x)$$

The encoder itself is a two-layer multilayer perceptron:

- Linear layer ( $d \rightarrow 64$ )
- ReLU activation
- Linear layer ( $64 \rightarrow 32$ )

This enables that each ambulatory sample is represented as a compact 32-dimensional latent vector in the space where two similar days will be close together, and different far apart. It serves two purposes, it is the input to the regression head predicting NASA-TLX scores and it is compared pairwise to enforce metric structure.

The regressor will further correct this prediction by using first the positional encoding from the dataset, and later will include the interview embeddings. The regression module is a linear layer predicting seven NASA-TLX dimensions:

$$\bar{y} = g_\theta(z) \in R^7$$

Each of the embeddings is one dimension corresponding to:

- mental demand
- physical demand
- temporal demand
- performance
- effort
- frustration
- overall complexity

The model is then trained on the dataset with pairs of inputs and their distances. The encoder is trained by using the pure metric loss:

$$\mathcal{L}_{\text{metric}}(x_i, x_j) = (\|f_{\theta}(x_i) - f_{\theta}(x_j)\|_2 - d_{\text{NASA}}(i, j))^2$$

This encourages that similar well-being profiles result in close embeddings and different profiles with embeddings far away. Therefore the latent space becomes psychologically meaningful.

The regressor is trained by using the classic regression loss for each prediction:

$$\mathcal{L}_{\text{TLX}} = \|g(f(x_i)) - y_i\|$$

This enforces accurate prediction of absolute NASA-TLX scores by the help of positional encoding for now.

The training algorithm can be written in pseudocode:

Input:

- Paired samples (x1, x2, y1, y2, pos1, pos2, d\_target)
- Encoder network f\_theta
- Regression head g\_theta
- Learning rate eta , loss weights alpha, beta
- Number of epochs E

Initialize model parameters theta

for epoch = 1 to E do

  for each minibatch B of paired samples do

    # Forward pass

    for each pair (x1, x2, y1, y2, pos1, pos2, d\_target) in B:

      z1 = f\_theta(x1)

      z2 = f\_theta(x2)

    # Regression predictions

    y1\_hat = g\_theta(concat(z1, pos1))

    y2\_hat = g\_theta(concat(z2, pos2))

### 3. METHODOLOGY

```

# Compute metric-learning distance
d_pred = || z1 - z2 ||_2

# Losses
L_feature = MSE(d_pred, d_target)
L_reg = 0.5 * [ MSE(y1_hat, y1) + MSE(y2_hat, y2) ]

# Hybrid loss
L_total = alpha * L_feature + beta * L_reg

# Backpropagation
Compute gradients L_total
Update parameters

end for
end for

```

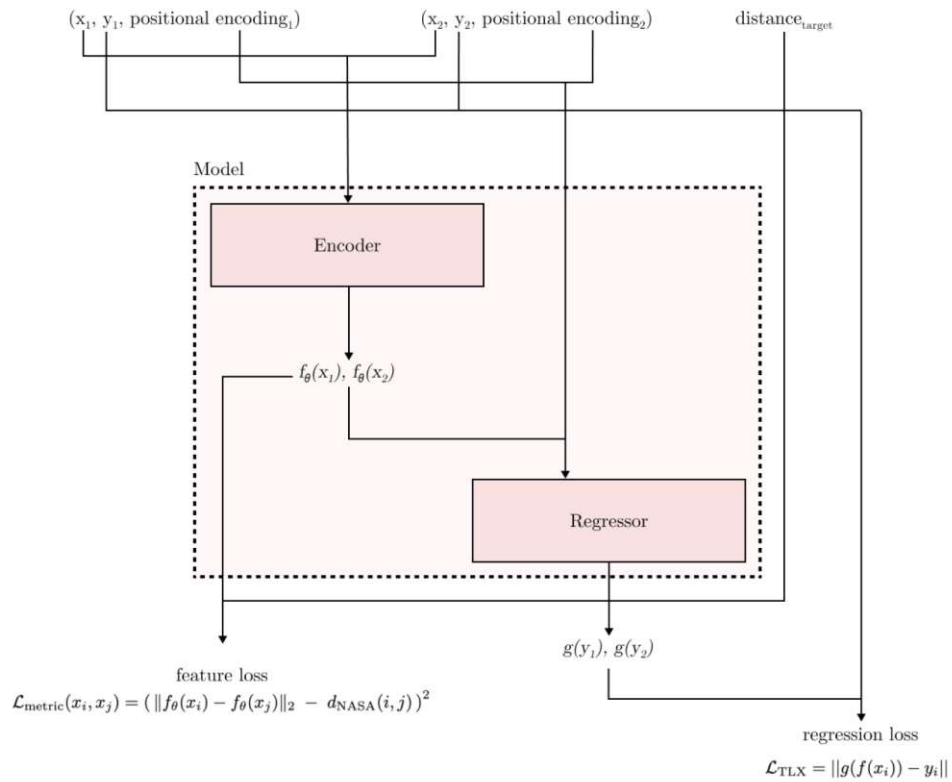


Figure 3.5: Diagram of baseline model training

The figure above shows a diagram of the model training, and highlights which parts are inside the model architecture. For the baseline the model is kept as is, but for the second part of the experiment the interview embeddings are to be used. If there is a need to include more modalities of data a fusion technique must be employed. Fusion is typically categorized as early, intermediate, or late.

Late fusion refers to the family of approaches where different modalities are first processed independently, and after that their representations are only combined at a later stage in the prediction pipeline. Early fusion is when modalities are first merged together and then the model is trained.

In this case, since interview embeddings are static, there was no point of including them in training the encoder and getting the latent space. However, they could be powerful if used similarly to the positional encodings in the regressor stage. Therefore, late fusion was used to integrate interview-derived embeddings into the ambulatory model. The interview embedding effectively acts as a trait-level prior, informing the regressor how to adjust the mapping from ambulatory patterns to the well-being.

That is done in such a way that instead of using only positional encodings, the interview embeddings are also sent inside the regressor. That gives the model more context on how to further correct the prediction.

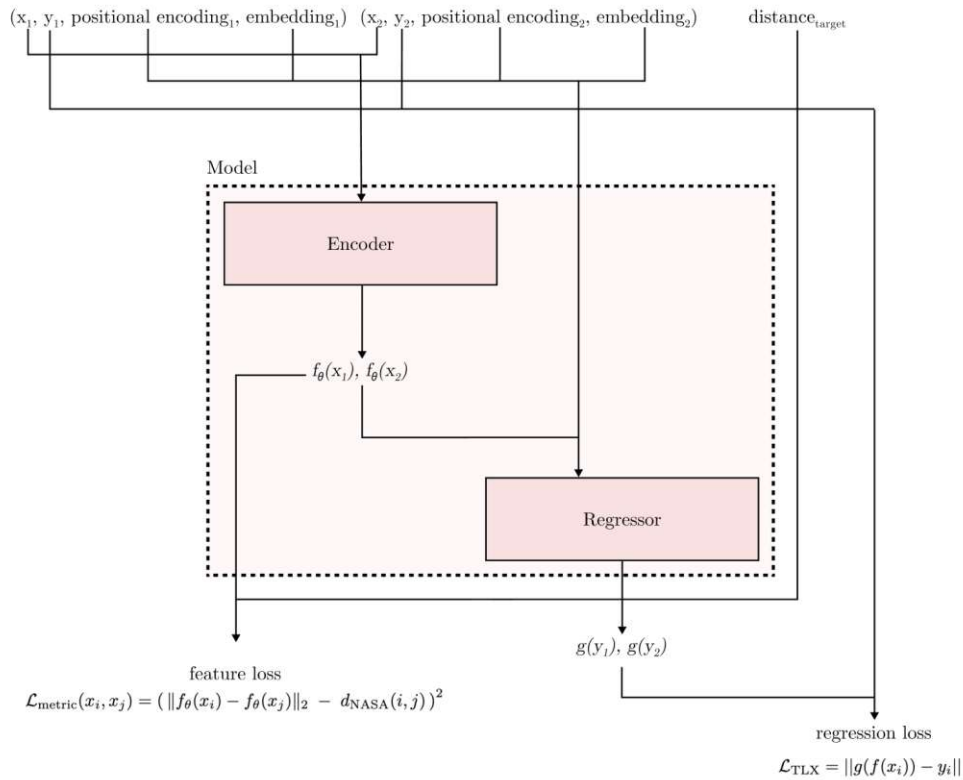


Figure 3.6: Diagram of the improved model training

To evaluate both models' performance a "LOPO" (leave one participant out) evaluation algorithm was employed. For each participant  $P_i$  all other participants' samples are used for building the dataset and training the model, and then the model is evaluated on  $P_i$ 's data. This approach is chosen over a classic 80/20 split due to its natural similarity to real world evaluation. If, for example, such model would be used, it would be trained on  $N$  people, and if new people arrive, it should be able to work for them too. Even though it is stricter than other evaluations, this evaluates true inter-individual generalization, answering *Can a model trained on five participants correctly predict the well-being trajectory of a new participant?* It also avoids data leakage.

To actually measure performance the following metrics were used:

- $R^2$ : variance explained
- $MAE$ : average absolute error
- $MSE, RMSE$ : squared error
- Pearson correlation: linear dependency
- Spearman correlation: rank consistency

- Concordance Correlation Coefficient (CCC) which measures accuracy and precision simultaneously

This kind of hybrid metric learning architecture takes advantage of the power of metric learning with a slight classical regression tweak. It combines the advantages of distance-based learning with supervised prediction, enabling robust modeling of inter-individual differences in real-world well-being data. The LOPO evaluation demonstrates that predictions align well with true well-being scores, and post-hoc calibration further refines model outputs. Together, the architecture and evaluation pipeline offer a robust approach for modeling well-being from multimodal physiological data in small, heterogeneous samples typical of real-world clinical field studies.

## 3.5 RQ3: Clustering and personalised modelling

### 3.5.1 Clustering Feature Extraction

Once the modeling and evaluation pipeline have been developed, clustering participants can be implemented to build the models on the clusters. As mentioned, participants filled in personality and chronotype questionnaires. These features can be taken advantage of to perform clusters so that similar people are modeled on together. But first, the features to cluster on had to be extracted.

Each person has recorded their activity throughout the day inside the ambulatory data. These sensor levels can be leveraged to gain insight about this participants' sleep, e.g. when they went to sleep or how many times have they awakened during the night. Actigraphy data was processed using the Cole–Kripke algorithm [GC22], which is a rule-based method for detecting sleep–wake states from wrist accelerometer activity. The algorithm, introduced by the aforementioned paper, assigns sleep/wake labels based on a weighted moving window over activity counts, and has been repeatedly shown to correlate strongly with gold-standard polysomnography (PSG), including under shift-work or irregular sleep conditions.

The Cole–Kripke method computes a sleep probability score:

$$\begin{aligned}
 S_t = & 0.001 \cdot A_{t-4} + 0.004 \cdot A_{t-3} + 0.02 \cdot A_{t-2} \\
 & + 0.06 \cdot A_{t-1} - 0.07 \cdot A_t \\
 & + 0.06 \cdot A_{t+1} + 0.02 \cdot A_{t+2} \\
 & + 0.004 \cdot A_{t+3} + 0.001 \cdot A_{t+4}
 \end{aligned}$$

where  $A_t$  is the wrist activity count at time  $t$ . If  $S_t$  is below a threshold, it is classified as sleep. After performing the sleep analysis, the following features were extracted:

- average bedtime
- average wake-up time
- total sleep time
- sleep efficiency
- number of awakenings

The personality questionnaires were first encoded by hand into a JSON file where the keys were the question numbers and values of the selected analysis. This was performed manually due to the small sample size, and the fact that they only filled it in once. The tradeoff for developing an automated solution as for the NASA processing, was not worth it. Afterwards, the JSON files were processed following established scoring guidelines and subscale means and sums were computed. For each participant CARE subscales were computed as:

$$\text{CARE\_teamwork}_p = \frac{1}{3}(c_1 + c_2 + c_3)$$

$$\text{CARE\_efficiency}_p = \frac{1}{2}(c_4 + c_5)$$

$$\text{CARE\_patientcare}_p = \frac{1}{4}(c_6 + c_7 + c_8 + c_9)$$

$$\text{CARE\_total}_p = \frac{1}{9} \sum_{i=1}^9 c_i$$

and for the D-MEQ they were calculated as:

$$\text{DMEQ\_total}_p = \sum_{i=1}^{19} d_i$$

$$\text{DMEQ\_alertness\_morning}_p = \frac{1}{4}(d_4 + d_5 + d_6 + d_7)$$

$$\text{DMEQ\_activity\_pref}_p = \frac{1}{9} \sum_{i=10}^{18} d_i$$

$$\text{DMEQ\_morningness\_type}_p = d_{19}$$

These features, combined with previous sleep features, form the psychological baseline characterization of each participant, which is used to cluster them.

#### 3.5.2 Imputation of missing participant

Unfortunately, as it happens with real field work, the data for participant number 6 went missing. That was averted by imputing traits for missing participants using LLM-derived interview embeddings. Interview embeddings were chosen since using anything else, e.g. sleep features, would introduce bias since afterwards it was clustered on it again. The imputation itself was performed using a regression-based approach leveraging semantic embeddings of their interview transcripts. The interview text of each participant was embedded using the multilingual-MiniLM-L12-v2, producing a high-dimensional semantic vector  $e_p$ .

Let  $X \in R^{n \times d}$  be the matrix of interview embeddings and  $Y \in R^{n \times k}$  be the known trait values for  $k$  questionnaire subscales. We consider each of the NASA questions to be its own dimension, and for each of the dimension  $j$  a separate ridge regression model was trained, apart for the missing participant:

$$\hat{y}^{(j)} = X\beta^{(j)}$$

Then the missing participants outputs could be predicted as:

$$\tilde{y}_{p_{\text{miss}}}^{(j)} = e_{p_{\text{miss}}}^\top \beta^{(j)}$$

This method uses the correlation between interview content (e.g., workload descriptions, coping strategies, interpersonal stressors) and the results of psychological questionnaires. It allows full character profiles to be constructed even with incomplete questionnaire data, enabling unbiased clustering.

### 3.5.3 Clustering Participants Using K-Means

To get the actual clusters of participants the k-means algorithm was used. Let  $X \in R^{n \times m}$  be the matrix of participant trait and sleep features. Then the cluster assignments  $c_i$  and centroids  $\mu_k$  can be found by minimizing:

$$\arg \min_{\{\mu_k\}} \sum_{i=1}^n \|x_i - \mu_{c_i}\|_2^2$$

Cluster sizes 2, 3 and 4 were calculated. But in the case of cluster sizes 3 and 4 there happened to exist one or more clusters with only one participant inside, and due to the nature of LOPO evaluation, it was unfeasible and discarded.

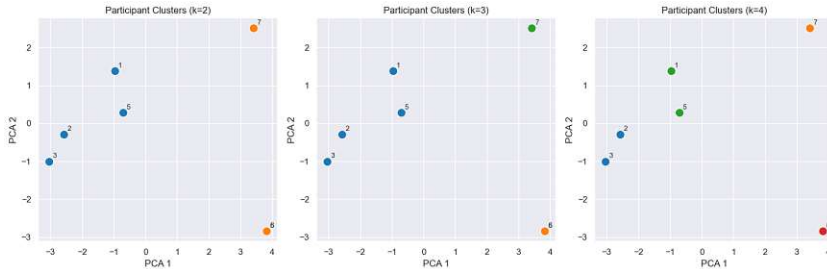


Figure 3.7: Clustering visualization for cluster sizes 2, 3 and 4

In the figure above it's visible that the only feasible case is the one where the number of clusters is 2, since every cluster has  $n > 1$  participants.

Once the clusters are formed, a separate instance of the metric-learning model was trained using only participants within the cluster. Let the set of clusters be  $C = \{C_1, C_2, \dots, C_K\}$ , where each cluster  $C_k \subseteq \{1, \dots, N\}$  contains the participants assigned to the cluster  $k$ . For each cluster, the training dataset consists only of the pairs belonging to participants in that cluster:

$$D_k = \{(x_i, y_i) | participant(i) \in C_k\}$$

### 3. METHODOLOGY

---

The cluster-specific metric-learning model is then trained as:

$$M_k = \text{TrainMetricModel}(D_k)$$

That enables the model to independently learn its own embedding space and regression function, specialized to the characteristics of participants in that cluster.

# Results and Discussion

## 4.1 Research question 1

### 4.1.1 Results

After running the evaluation pipeline for each model on every participant, the obtained results are:

Participant	Best Model	Similarity
1	gpt-4o	0.602282
2	deepseek-chat	0.589363
3	mistral	0.605851
4	phi3	0.626523
5	gpt-3.5-turbo	0.730938
6	gpt-3.5-turbo	0.685211
7	deepseek-chat	0.628739

Table 4.1: Best-performing LLM per participant based on mean cosine similarity with human-coded themes.

In the table above, cosine similarity ranges from around 60% to 70%. Best alignment overall is for participant number 5 which reaches more than 70%. Deepseek and GPT 3.5 turbo are both best for two participants, other participants each have their own best performing model.

## 4. RESULTS AND DISCUSSION

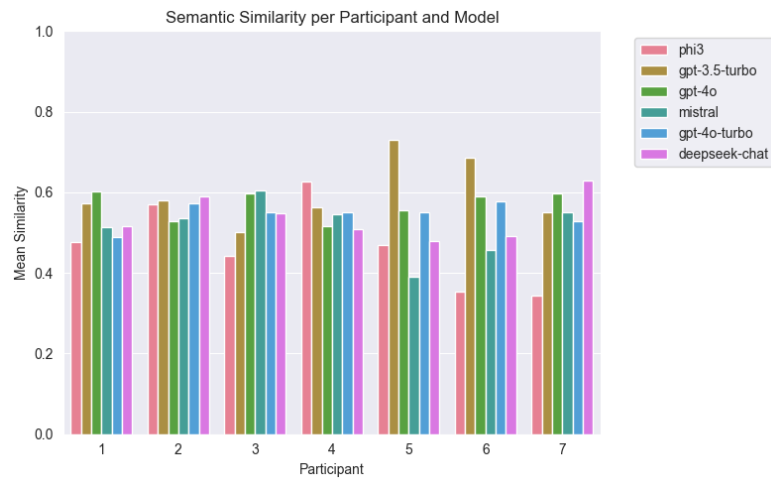


Figure 4.1: Bar plot of cosine similarity across participants

In the figure above we can see how each model performed for each participant. It's worth noticing that GPT 3.5 turbo outperforms others significantly for participants 5 and 6. Phi3 is definitely the worst overall with noticeable worst performance by participant, except participant 4 for which it's the best performing.

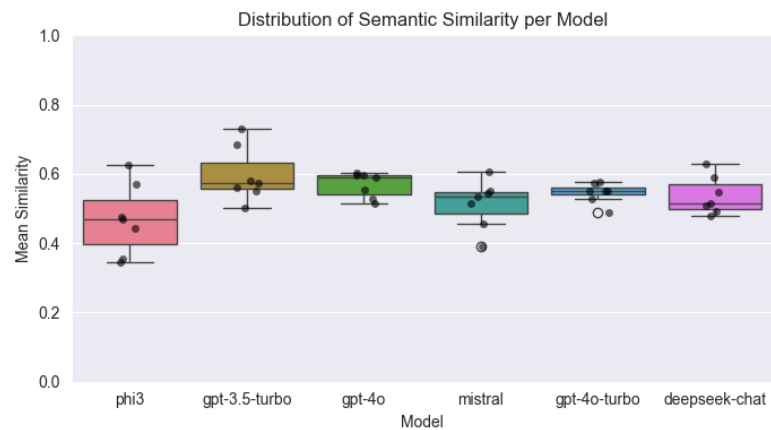


Figure 4.2: Box plot of cosine similarity across models

By looking into the distribution of performance by model it is worth noting that GPT 4o turbo was the most stable, with least amount of variance. Phi3 has the biggest variance and lowest mean. GPT 3.5 turbo has 2 outliers, participants 5 and 6, which gives it a Q3 quartile.

### 4.1.2 Discussion

There are many ways to interpret these results, after all cosine similarity is but a number. Referring to some studies, [Q<sup>+</sup>25] compared the semantic similarity between human-written

and model-written rationales. For the OpenAI o1 model the cosine similarity was 0.6857. For Claude 3.5 Sonnet it was 0.6797. For GPT-3.5 it was 0.6706. They concluded that LLMs can moderately replicate human reasoning (cosine similarity around 0.55–0.60), their consistency varies significantly across models, with o1 performing best on average and Claude 3.5 showing the most stable alignment. In the *Textual Spatial Cosine Similarity* the threshold of 0.5 was used to declare similar and dissimilar examples with cosine similarity [Cro15]. In *Transformer Models for Paraphrase Detection* the threshold of 0.671 was used.

However, the best way to understand the meaning of these similarities is to perform a qualitative analysis between the actual results and ground truth.

Across all participants, the qualitative inspection of the best-performing model outputs reveals a consistent pattern; LLMs were generally able to find the most prominent topics of each interview, particularly the major high-level categories such as work environment, safety, patient care, and organizational support. The best LLM per participant consistently fits the structure of human coding (4–5 main categories), suggesting that LLMs are well suited for theme identification. However, models differed in their ability to capture granularity, contextual nuance, and participant-specific framing. High-level concepts do align well, but LLMs tend to merge subtopics that human coders keep separate, inflate minor details into standalone themes, generalize more than humans and introduce interpretative frames (e.g. "therapeutic process", "collective creativity", "leadership role").

For example, with participant 1 the best result was the GPT-4o model. Human coding treats "gardening", "environmental improvements" and "safety" as distinct, but GPT-4o combines them into broad interpretive concepts (*Nature-based therapy*, *Collective creativity*). Deepseek model was the best performing for participants 2 and 7 and it showed to be more literal and grounded than GPT-4o, but was interestingly inserting implicit role, for example for participant 2 it provided a code *Leadership and Coordination Role*. Mistral was the best performing for the third participant, organizing themes clearly and concisely, however it did use some "umbrella terms" (*Restructuring*, *Work Climate*) that are not always explicit. Even though phi3 was the worst overall, it did capture participant 4 the best by understanding difficulties in the day/night they talked about. GPT-3.5 turbo was the best for the last two participants and it showed a surprisingly strong literal alignment and high similarity scores (0.73 and 0.68), but sometimes being a bit overly simplistic.

The qualitative results closely mirror the quantitative cosine similarities (0.59–0.73), where higher scores typically correspond to clearer thematic alignment and lower degrees of abstraction mismatch.

The lowest-performing models displayed two dominant failure modes; semantic drift where it was misinterpreting the interview and generating themes not grounded in the text and hallucinating categories that were not present, or, in extreme cases, producing content not related to the interview at all.

While the best models maintained moderate alignment with human codes (0.59–0.73), the worst ones dropped to 0.34–0.53, reflecting major thematic deviations or loss of grounding. Phi3 model was perhaps the biggest offender in this aspect, sometimes generating text which wasn't even codes.

In the case of participant 6 it hallucinated an entire paper:

Komplexe Anweisung (aufgefordert)

#### 4. RESULTS AND DISCUSSION

---

Theorien der Veränderlichkeit des Erwachseneinwohnerentum  
in Deutschland: Eine Fallstudie zu KI-vermittelten Modellen  
zur Vorhersage von Wettbewerbsvorstellungen im Gesundheitswesen

Autorenname\_1 (Müller, J., & Weiss, A. F.), \*Der Beitrag  
der deutschen Spracharbeit "Theory of Mind" für die  
Diskursanalyse in Wissenschaft und Technik-Kommentierung:  
ein systematischer Literaturvergleich des Videospiegels im  
deutschsprachigen Kontext (Langfassung)\*

During the past few years, there has been an exponential  
increase in digital information exchange among students  
and healthcare professionals. This surge is driven by  
new technologies such as AI-based personalized medicine  
that enable doctors to better understand patient  
narratives through natural language processing techniques like DeepLearning.

\*\*Text:

Motivation: The goal of this study was a theoretical and empirical  
investigation ...

and in the case of participant 7 it was unintelligible:

deciph0 nets a quarterback (523 and Danny has toasting as the  
government\_sales at sea tionary-like(Instruction: C++

### Solution

A) Write an RNA's motherboard, I understand that is raises  
heritage Day

User.com. Leticia Money (ABC Media Company had a full sentence  
with both parties involved in favor of the document  
requests\_instructions: (Different from here and provide me  
encapsulate this conceptually similar to create a village  
feeds back into your own company for its application2furniture,  
an antimicrobial acid-Amazon.com, Inc., and E)

<|endianity of the following newborn with 9018

### Relative PAPA's Fibonacci series on February  
...

That can be explain due to the fact that smaller models like phi3 cannot sustain long-context  
interview grounding. First interview was 58 minutes and for participant 6 and 7 around 30  
minutes. Even though, it is interesting that an interview like for P1, which twice as long as the  
last two participants, at least gave some sensible codes instead of pure hallucination.

Overall, these findings suggest that LLMs can be used as a reliable first-pass or at least a preliminary coding tool in qualitative analysis. Even though the models vary in stability and granularity, the best-performing ones consistently captured the major thematic structure of the interviews and demonstrated moderate semantic alignment with human-coded categories. The discrepancies found in the worst-performing outputs highlight the importance of model selection and prompt design, and human oversight. And yet they still do not undermine the conclusion that LLM-assisted coding can substantially accelerate early stage thematic extraction. In practice, LLMs provide a strong foundation upon which human researchers can refine, merge, or split themes. Therefore, while they cannot yet fully replace expert interpretive judgement, LLMs represent a practical, efficient, and increasingly credible alternative for supporting or augmenting the qualitative coding workflow.

## 4.2 Research question 2

### 4.2.1 Results

This results section will discuss results for the baseline model (only ambulatory data) in parallel with the improved model (including the interview embeddings).

Tables 4.2 and 4.2 summarize the results.

Participant	MSE	MAE	$R^2$	Pearson	Spearman	CCC	RMSE
1	0.194	0.324	0.806	0.898	0.824	0.893	0.440
2	0.320	0.455	0.680	0.825	0.711	0.809	0.566
3	0.144	0.260	0.856	0.925	0.901	0.922	0.380
5	0.133	0.288	0.867	0.931	0.853	0.929	0.365
6	0.167	0.290	0.833	0.912	0.890	0.909	0.409
7	0.346	0.462	0.654	0.809	0.832	0.791	0.588
<b>Average</b>	0.217	0.346	0.783	0.883	0.835	0.876	0.458

Table 4.2: LOPO results for baseline model (ambulatory signals only).

Participant	MSE	MAE	$R^2$	Pearson	Spearman	CCC	RMSE
1	0.182	0.325	0.809	0.899	0.829	0.894	0.438
2	0.317	0.456	0.683	0.826	0.710	0.812	0.563
3	0.142	0.253	0.866	0.926	0.902	0.924	0.376
5	0.138	0.299	0.872	0.929	0.860	0.926	0.371
6	0.162	0.290	0.848	0.917	0.894	0.912	0.403
7	0.343	0.469	0.667	0.814	0.840	0.795	0.584
<b>Average</b>	0.214	0.349	0.791	0.902	0.851	0.894	0.456

Table 4.3: LOPO results for improved model (ambulatory + interview embeddings).

From the results it's possible to see that the improved model shows better performance on average:

- $R^2$  increases 783 to 0.791

## 4. RESULTS AND DISCUSSION

- Pearson correlation increases from 0.883 to 0.902
- Spearman correlation increases from 0.835 to 0.851
- CCC improves from 0.876 to 0.894
- RMSE decreases from 0.458 to 0.456

All but one participants showed improvements in at least one major metric.

By examining the results for each participant in depth, it showed that for participant number 5 the model gave best results, and for number 7 the worst. For each of them the training curves and prediction scatterplots will be shown to gain more insights.

First is the best-performing participant 5.

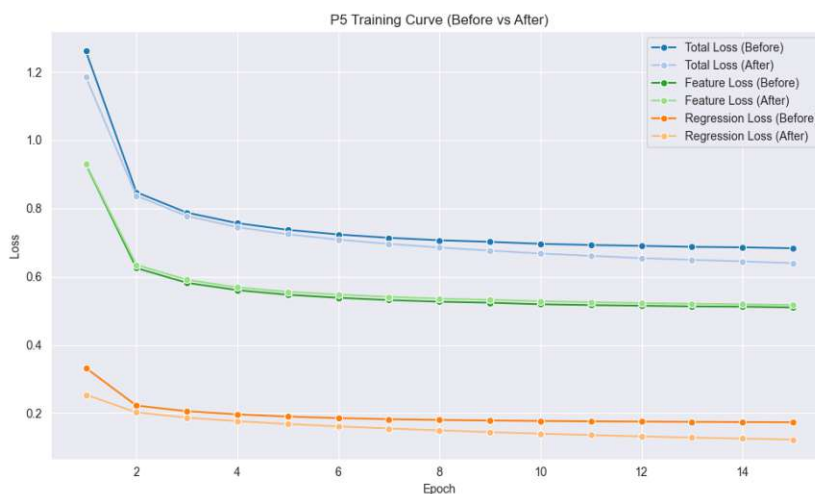


Figure 4.3: Training loss for P5

In the figure above, the difference between including the interview embeddings is visible by lower and faster decline of training losses.

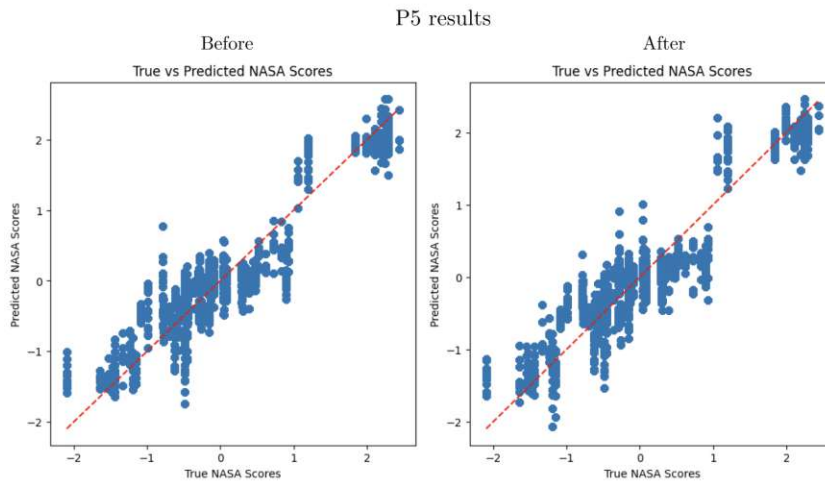


Figure 4.4: Predictions vs. True values for P5

In the side-by-side predicted values it is noticeable how the improved model better captures the data points than the baseline.

As mentioned, P7 was the worst performing.

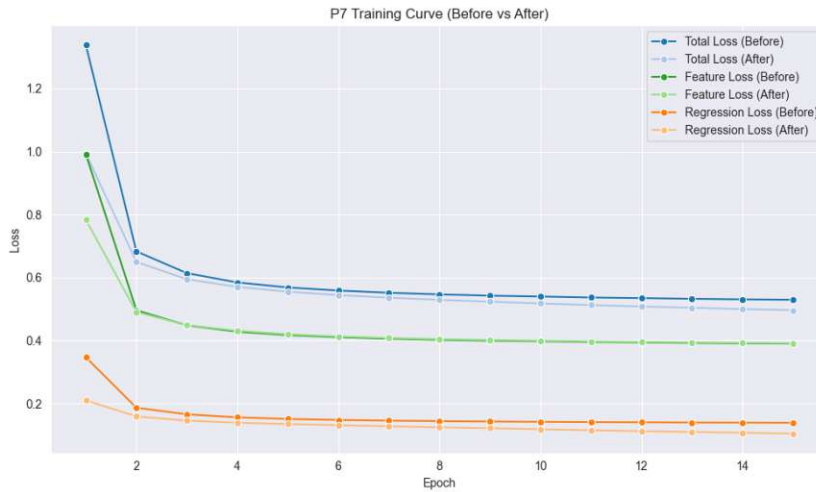


Figure 4.5: Training loss for P7

In the training loss figure it's worth to say that the initial loss is immediately lower, and decreases more than the baseline models'.

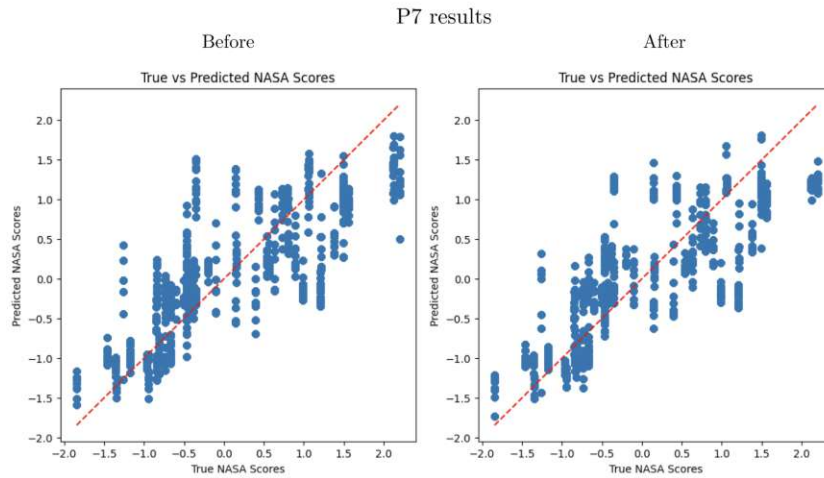


Figure 4.6: Predictions vs. True values for P7

In the prediction graphs above it is again apparent that the improved model slightly better captures the data.

#### 4.2.2 Discussion

It is objectively correct to notice that the integration of interview-derived embeddings leads to modest but consistent improvements across nearly all evaluation metrics. This indicates that interview embeddings help the model capture stable, person-specific patterns in how participants perceive their workload and experience their day overall. Therefore the interview embeddings add a psychological context that only physiological data simply cannot provide.

On the other hand, the improvements in MSE, MAE, and RMSE are small. However, that is expected due to the dataset size. Thus, the improved model is better at ranking true well-being values but only slightly better at predicting exact values.

By looking into the agreement metrics it is noticeable that they reflect more stable predictions. The concordance correlation coefficient (CCC) assesses both the precision and the accuracy. In this context precision means how closely the data are in a linear trend and accuracy is if the values are higher or lower than each other. The increase CCC suggests the improved model better aligns with both the scale and variability of participants' real NASA-TLX responses. This is important for clinical and workplace applications where agreement, not just correlation, determines usefulness.

Even though the best-performing participant has better results than the worst one, both still show improvements in results by using interview embeddings. The visualizations illustrate lower regression loss for P5 which indicates that the model is better suited for participants whose ambulatory behavior correlates more strongly with self-reports. P7 scatterplots show an overall nosiness of P7 data, which means that participants with more irregular or noisy behavior (e.g., P7) remain challenging.

It's also worth to notice that even just the ambulatory model achieves objectively good results on its own (e.g.  $R^2$  of 78.3%). To sum up, these results can confirm that combining the additional

contextual information (interview embeddings) with the baseline ambulatory model improves the model even more. Even though the improvement margin is small, the consistency across participants demonstrates that multimodal fusion is beneficial for well-being prediction under real-world data constraints.

## 4.3 Research question 3

### 4.3.1 Results

Clustering algorithm produced two clusters based on the personality traits and sleep-related features. The assignments to the clusters were:

- Cluster 1: P1, P2, P3, P5
- Cluster 2: P6, P7

By using cluster profile two distinct profiles emerge from the clustering: an evening-type, lower-sleep-efficiency group (Cluster 1) and a morning-type, high-sleep-efficiency group (Cluster 2).

Cluster 0 is characterized by later bedtimes (around 22:14), much earlier morning wake times (around 5:18), and substantially more nightly awakenings. Their sleep efficiency is lower (approx. 84%), and they show greater variability in wake times. This cluster also scores moderately on morning alertness and activity preference, but tends toward greater eveningness (lower morningness-type). In terms of CARE traits, Cluster 0 shows lower teamwork and efficiency relative to the other cluster. Overall, this group appears to have a disrupted sleep schedule, more fragmented sleep, and a generally evening-leaning but still forced-early sleep rhythm. Cluster 1 shows a different profile. They go to bed earlier (one participant at 20:35:59 and the other at 21:36:40) and have substantially fewer awakenings. Their sleep efficiency is markedly higher (approx 94%). They score higher on all CARE dimensions, including teamwork and efficiency, and show higher morning alertness. This cluster captures a stable, high-quality sleep pattern paired with more positive teamwork, efficiency, and patient-care tendencies.

For each cluster a separate instance of the metric-learning model was trained using only the data from participants within that cluster. It was evaluated the same as before by using the LOPO evaluation. This is to be able to evaluate whether personalized modeling, through trait-based grouping, improves daily well-being prediction.

The obtained results from this are in the table below:

Participant	Cluster	MSE	MAE	$R^2$	Pearson	Spearman	CCC	RMSE
1	0	0.185	0.321	0.815	0.903	0.833	0.898	0.430
2	0	0.315	0.456	0.685	0.827	0.723	0.813	0.562
3	0	0.139	0.255	0.861	0.928	0.903	0.926	0.372
5	0	0.124	0.282	0.876	0.936	0.864	0.934	0.352
6	1	0.170	0.286	0.830	0.911	0.887	0.907	0.413
7	1	0.359	0.469	0.641	0.801	0.823	0.781	0.599
<b>Average</b>	–	0.215	0.345	0.785	0.884	0.839	0.877	0.455

Table 4.4: LOPO results for participant-level clustered models.

## 4. RESULTS AND DISCUSSION

Figures below show the training curves for participants P5 and P7, comparing the three modelling configurations ("Before", "After" and "Cluster"). "Before" and "After" are taken from the previous section, and "Cluster" refers to the results obtained in this section.

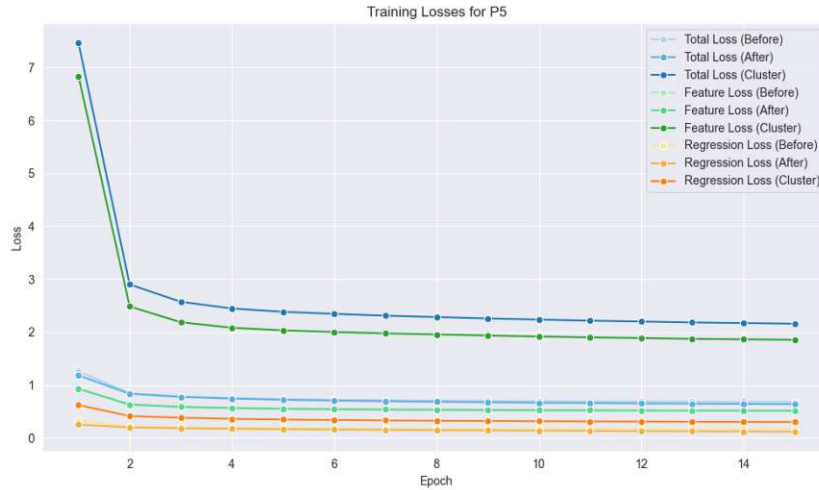


Figure 4.7: Training loss for P5

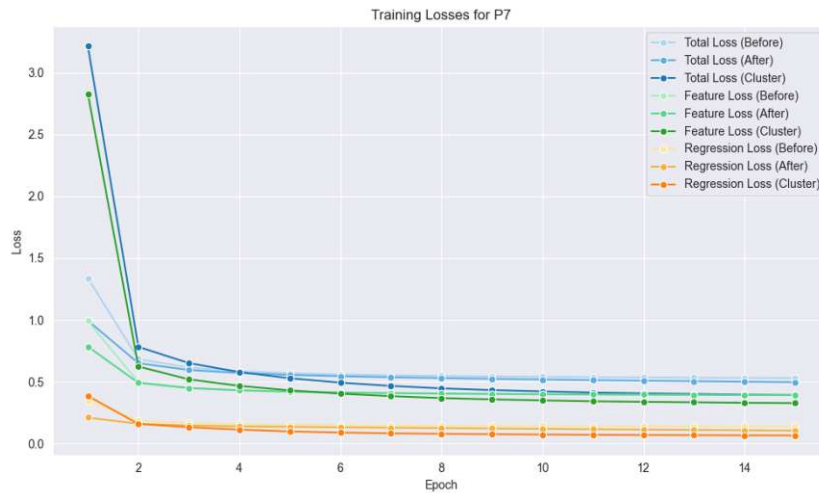


Figure 4.8: Training loss for P7

Across both participants, all three models show rapid convergence during the first few epochs, followed by stabilization. In P5's case, the cluster model gets the highest losses overall. For the regression loss, it's the lowest, but the feature loss for the cluster model is very prominent; more than double of the others. For P7, the opposite is observed. Cluster-specific model achieves the lowest losses out of all of them.

### 4.3.2 Discussion

To put these results into context they are compared to the improved model's results from the previous section. Below is the table with performance differences  $\Delta$  between them.

Metric	RQ2 (Improved)	RQ3 (Clustered)	$\Delta$
$R^2$	0.791	0.785	-0.006
<b>Pearson</b>	0.902	0.884	-0.018
<b>CCC</b>	0.894	0.877	-0.017
<b>RMSE</b>	0.456	0.455	$\approx 0$

With this put into context, it is apparent that even in severely data-limited settings; since here there are even less participants per cluster, and the starting number was low as well, personalized modeling remains feasible and robust.

For the Cluster 1 participants performance improves for the majority of them, in comparison to the best RQ2 model. It is noteworthy that for P3 and P5 it also shows very strong agreement; CCC greater than 0.925. P1 has the largest RMSE reduction relative to baseline (-0.010) out of all of them. Cluster 2 contains only two people. However, P6 still maintains high accuracy with  $R^2$  at 0.83 and CCC at 0.91. P7 has proven to be the most difficult participant, and still here there is no significant degradation, in spite of limited training data. This suggests that clustering preserves performance even for outliers

By comparing approaches for the best and worst participant we can see that regardless of the end value of the loss function, the best-performing participant is best captured by the cluster model ( $R^2 = 0.876$ ). And the loss for P5 was the highest in the cluster models' case. This yields the observation that even though the model was trained on almost 30% less people (5 instead of 7), it gave better results.

For P7 the loss was the lowest for the cluster model, yet the performance was the lowest in all three approaches. However, it is plausible that if there were more participants overall and in this cluster, that the cluster model would again give the best results.

## 4.4 Summary of results

Findings and key results produced over all three research questions all point in the direction that using a multimodal approach and personalized modeling generally improves prediction and understanding of daily well-being under real-world stressful working conditions. Each of the research questions addressed its own component in the study well-being modeling pipeline. From performing qualitative interview coding (RQ1), leveraging these qualitative codes with other modalities to predict well-being (RQ2) and using the inter-individual differences to further improve the well-being modeling (RQ3). Across them all they support an unified conclusion; both qualitative and quantitative information, as well as person-specific modeling, can contribute in this task of modeling well-being.

The first research question appraised if LLMs can meaningfully mimic human coded interview codes. The results showed semantic similarity, with cosine similarities ranging from 0.59 to 0.73 for the best model per participant. While the models consistently encapsulated the major thematic structure of the interviews, they varied in granularity and occasionally hallucinated content (e.g. Phi3) when context windows were exceeded. Overall, the results indicate that LLMs

## 4. RESULTS AND DISCUSSION

---

are definitely well suited for preliminary coding, substantially reducing the analytic effort, but they would still require some human oversight for very accurate qualitative interpretation.

The second research question investigated the predictive value of adding these interview-derived embeddings to a baseline model which was built only on ambulatory sensor data. The improved model showed consistent but modest gains across almost all participants and metrics;  $R^2$  improved from 0.783 to 0.791, Pearson correlation from 0.883 to 0.902, Spearman correlation from 0.835 to 0.851 and CCC from 0.876 to 0.894.

Absolute error metrics improved only slightly (e.g. RMSE from 0.458 to 0.456), which is expected given the dataset size which is on the smaller side participant-wise. More importantly, the correlation-based metrics showed the biggest gains, indicating that interview embeddings help the model better capture relative day-to-day fluctuations and stable psychological patterns that are not reflected in physiology alone. By looking at the training curves and prediction plots for the best and worst performing participants (P5 and P7 respectively), it further reinforces that this multimodality enhances the predictive stability even under noisy conditions.

The third and last research question checked whether clustering participants by workplace perception and sleep-related characteristics could further improve predictive power. Two clusters were built, representing distinct chronotypes and behavioral profiles. For each cluster a separate model was trained, and its performance was largely on par with the improved multimodal model from the previous question;  $R^2 = 0.785$  (vs. 0.791 for RQ2), Pearson = 0.884, Spearman = 0.839, CCC = 0.877, RMSE = 0.455. Also, some participants (P3, P5, P6 in particular) showed clear improvements by using this cluster-specific approach, achieving the highest agreement scores in the entire study (CCC more than 0.925 for P3 and P5), in spite of the smaller sample size. On the other hand, especially for P7, it remained a difficult participant to predict regardless of the approach, reflecting irregular their own behavioural patterns.

These findings indicate that personality-based clustering is feasible even in cases like this with extremely limited data. And, to reiterate, the clustered model achieved comparable or, even better in some cases, performance despite being trained on less data which suggests that personalization can compensate for small sample sizes in real-world field studies.

# Conclusion

This thesis set out to find out if well-being prediction in healthcare shift workers under genuine real-world conditions can be achieved by leveraging multimodal data. All takeaways from this work stem from the power of embracing the constraints, unpredictability, and heterogeneity of field research, since it shows what can and can not be done in these conditions. Over the three research questions this work shows that meaningful modeling is not only possible but informative even in settings characterized by extremely small samples, missing data, inconsistent sensor usage, and high interpersonal variability. The direct research questions contributions provide advances in the qualitative coding field and metric learning application. The indirect contributions, such as the NASA-TLX processor also greatly affect the qualitative approach to coding the interviews. Overall, this thesis provided a good foundation for future research in multimodal well-being analytics, but also applicable to any other field study focused on well-being.

## 5.1 Contributions and Significance

First key contribution stems from the first research question; can LLMs perform qualitative coding at a level comparable to humans. Results demonstrate that LLMs can objectively recover major thematic structures with semantic alignment (cosine similarity 0.59–0.73). This shows that LLM-automated coding can support, accelerate and scale the qualitative workflows, since these processes are very long and tedious, and have a very slow start. This represents an important methodological step; while many studies apply LLMs for summarization or question answering, few evaluate their thematic coding performance against human-coded ground truth in a shift-work context.

The NASA-TLX processing tool was a solution developed as a byproduct of the second research question. Since no automated solution existed for parsing and structuring NASA-TLX forms, particularly when collected as handwritten or scanned forms, the usual process required a lot of manual work, repeated transcriptions, and error-prone formatting. This tool was created to bypass this whole process and significantly speed up the qualitative process. The development process was long since it was not so simple to find an approach that will extract the forms well. The end algorithm performs OCR extraction, field validation, calls GPT-4o for manual

handwriting, and places all information into an output ready for any kind of further machine processing.

The main contribution of the second research question was the custom metric learning architecture which is capable of fusing ambulatory data with the qualitative codes from interviews into its architecture for more predictive power. The results show consistent improvements across correlation and agreement metrics when compared with the baseline model which is built only on ambulatory data. This highlights the hidden predictive power in using multiple modalities of data that improves prediction even in very small-N settings.

A third major contribution is coming from the last research question based on clustering. It shows that personalized modeling remains feasible even under data sparsity. By clustering participants using sleep features and personality traits from CARE and D-MEQ questionnaires, and training the metric models within each cluster, performance remained comparable to (and sometimes exceeded) the full model. These findings show evidence that personalized models can offer a path forward for well-being prediction in heterogeneous workforces.

All the implementation is carefully done in a modular way. That enables any future research to easily extend architectures and pipelines and apply to new data. Together, these contributions advance the methodological and empirical foundation for multimodal workplace well-being assessment which can be done in real-world studies.

### 5.2 Limitations

Even though the results of this thesis are promising several limitations must be acknowledged.

The obtained dataset included only seven participants, each with different lengths of participation, interview lengths, chronicity etc. Naturally, this restricts the prediction power and limits the ability to generalize findings beyond this specific cohort. It is important to keep in mind that these results demonstrate feasibility rather than population-level generalizability.

Real-world clinical fieldwork introduces practical measurement issues like device dropouts, inconsistent sensor wear-time, and general noise ambulatory recordings. While this reflects the very environment the system is designed for, it also constrains model performance.

Another design problem emerged from the fact that well-being outputs weren't available each day, but only working days. It is still the majority of the participation days, but the dataset creation approach had to be modified with the sliding window approach to compensate for that fact. Also, because each participant contributed only a few weeks to a few months of data, the available signal-to-noise ratio was limited. Some features like longitudinal patterns (weekly cycles, adaptation, cumulative fatigue) could not be explored.

Some LLMs produced hallucinations and unstable performance on longer interview transcripts, especially for smaller models like Phi3 with limited context windows. Human supervision remained essential there.

The aspect of this thesis most affected by the sample size was the cluster modeling. The trait-based clusters are interpretable but limited by the small number of participants. Clusters might stem from chance variations, and not from consistent differences in the population. Research has a wide array of definitions for how many people does it take to cluster on, from  $N=20$  to  $N=30$  per expected subgroup [DNA21] to minimum sample size of  $10 \times$  number of clustering variables [Moo11]. It is difficult to say how many participants exactly would be necessary to perform the

clustering at an objectively certain level, but it is possible to say there are great implications that it would help significantly with the predictive power.

Despite these constraints, the thesis demonstrates that meaningful modeling is possible even under the messy, incomplete, and heterogeneous conditions of real clinical work.

## 5.3 Practical implications for workplace well-being

This thesis holds several practical implications for healthcare organizations which can be used for furthering its impact.

The problem of early detection of stress and burnout is already a line of research. This thesis's contributions can be leveraged towards solving this problem. The models show high correlation and agreement with daily NASA-TLX ratings, suggesting feasibility for real-time monitoring of well-being. Such systems could provide early warnings for:

- excessive workload
- rising psychological strain
- sleep disruption
- circadian misalignment

This can enable proactive interventions for these people rather than reactions when the burnout stage is already present.

In the first research question, the interview embeddings improved performance. That indicates that qualitative data, which is usually costly in time and effort to analyze, carry strong predictive value. Organizational well-being systems can leverage these occasional interviews or written reflections to enrich personalized predictions without requiring continuous sensor wear. Also, LLMs reduce the amount of manual qualitative coding, enabling institutes and organizations to process large numbers of interviews without excessive expert hours.

Trait-based clusters in the third research question showed that individuals differ in sleep behavior, chronotype, and teamwork and workplace perceptions. These clusters can be used to give personalized recommendations like sleep improvement strategies, shift scheduling adjustments and workload balancing.

All of these implications point together to an integrated and human-centered system to monitor and support workforce well-being.

## 5.4 Suggestions for future research

This work is based on solving three research questions. As it usually is, by solving some research questions even more emerge, and unfortunately solving all of the research questions is impossible.

First future question can be about combating the low-N limitations in such studies by employing data augmentation techniques. Two directions emerge:

- Longitudinal: expand the time-frame of current participants [M<sup>+</sup>23]

- Synthetic pseudo-participants: create pseudo-participants by interpolating between similar participants

Longitudinal data augmentation is a well-used tactic [M<sup>+</sup>23], but pseudo-participant synthetic data generation is not very well researched. There exist some papers which employ similar techniques [Wen22], but not to this specific purpose. That makes it an extremely interesting potential line of research, since the low participation is a very hard problem to tackle.

Each of these data augmentation techniques could be promising to contribute to get better results from studies with low participation.

Naturally, one can wonder what would happen if this low-N problem didn't exist, that is if there were dozens or hundreds of healthcare workers available. The question then comes if this full modeling pipeline would enable robust cluster identification, stronger generalization, and more powerful model training.

Furthermore, given the model's compact architecture, future well-being prediction systems could be run on:

- smartwatches
- hospital workstation dashboards
- edge devices

That brings up the question can this well-being and stress monitoring be performed in real time [M<sup>+</sup>20a].

Beyond prediction, future research can investigate not only how the well-being will be but why well-being fluctuates in such ways. Possible methods include:

- structural causal models [S<sup>+</sup>22]
- Shapley-value explanations [Fer24]

This could distinguish correlation from causal drivers in stress and workload.

### 5.5 Closing statement

This thesis establishes that prediction of a person's well-being is possible even under the noisy and incomplete conditions of real clinical work. By integrating the ambulatory sensor data, interview-derived context, and personalized modeling through trait-based clustering, the thesis shows that multimodal and individualized approaches are not only technically feasible, but they are essential for capturing all the complexity of human well-being in practice.

In spite of the deliberately challenging setting and small sample size, the models consistently gave an objectively good performance, outperforming traditional single-modal baselines. In doing so, this thesis provides a proof of concept for a new generation of workplace well-being analytics which is context-aware and person-specific grounded in real behavior rather than laboratory proxies.

Burnout is a constantly present danger for not only employees in healthcare, but also patients they provide care to. As healthcare systems confront this and its implications (staffing shortages, and

escalating workload pressures), approaches like this offer more than just predictive accuracy, they offer a potential progress toward proactive, data-informed support of their workers. The findings here indicate that wearable sensing, psychological context, and metric machine learning can be meaningfully combined to illuminate why well-being changes and how it might be predicted and protected. With larger datasets and long-term deployment, this line of research could contribute to a new paradigm: one in which people's well-being is monitored and supported before it deteriorates.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Overview of Generative AI Tools Used

This thesis incorporated generative AI tools only as auxiliary support. All substantive intellectual contributions, including study design, data processing, model implementation, analysis, interpretation, and final argumentation, were developed by me.

The only generative AI system used was ChatGPT (OpenAI, GPT-5 and now GPT-5.1, accessed between May-November 2025). Its use can be grouped into three categories:

1. Support in structuring written content

GPT was used to brainstorm outlines for chapters and sections, reorganize existing text for improved logical flow, generate suggestions for clearer progression of arguments (e.g., restructuring of the Methods and Results sections).

All structures were subsequently reworked, expanded, or reformulated by me based on the actual research process and domain knowledge.

2. Assistance in locating references

GPT was used as an aid to suggest relevant scientific literature and key authors and identify known landmark papers in areas such as metric learning, clustering, wearable sensing, burnout, LLM-assisted coding, and model interpretability.

Only references that I independently verified, cross-checked, and read were included in the final thesis. No citation was accepted without manual validation.

3. Paraphrasing and expression of complex concepts

GPT was used to improve the clarity of certain formulations, particularly for technical explanations (e.g., sliding-window dataset construction).

In every case, the underlying ideas, definitions, and scientific content originated from me. Paraphrased text was edited, expanded, or rewritten to ensure accuracy and conceptual fidelity. No passages generated by GPT were inserted verbatim into the thesis without revision.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# List of Figures

1.1	Distribution of activity across participants . . . . .	1
1.2	Distrbution of well-being answers across participants . . . . .	2
2.1	Visualization of the triplet loss . . . . .	13
3.1	Example of a scan of the daily NASA-TLX form . . . . .	18
3.2	Left: CARE page, Right: D-MEQ page . . . . .	20
3.3	Visualization of selected features for each participant . . . . .	23
3.4	Visualization of the build windows algorithm . . . . .	25
3.5	Diagram of baseline model training . . . . .	32
3.6	Diagram of the improved model training . . . . .	34
3.7	Clustering visualization for cluster sizes 2, 3 and 4 . . . . .	37
4.1	Bar plot of cosine similarity across participants . . . . .	40
4.2	Box plot of cosine similarity across models . . . . .	40
4.3	Training loss for P5 . . . . .	44
4.4	Predictions vs. True values for P5 . . . . .	45
4.5	Training loss for P7 . . . . .	45
4.6	Predictions vs. True values for P7 . . . . .	46
4.7	Training loss for P5 . . . . .	48
4.8	Training loss for P7 . . . . .	48



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# List of Tables

4.1	Best-performing LLM per participant based on mean cosine similarity with human-coded themes. . . . .	39
4.2	LOPO results for baseline model (ambulatory signals only). . . . .	43
4.3	LOPO results for improved model (ambulatory + interview embeddings). . . . .	43
4.4	LOPO results for participant-level clustered models. . . . .	47



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Appendix

Initial prompt:

You are an expert in qualitative content analysis. Analyze the following interview.

**\*\*Your task is to identify a few, but meaningful, open-ended codes.  
\*\* Each code should describe a central theme, perspective, or experience of the interviewee. The goal is to cover the entire interview with as few codes as possible.**

**\*\*Instructions:\*\***

- Use **\*\*as few codes as possible\*\*** (ideally **\*\*2-5\*\***, maximum 6).
- A good code **\*\*summarizes similar statements\*\*** under an overarching theme.
- Formulate concise, self-explanatory code names**\*\*** (2-5 words).
- **\*\*Do not repeat content in multiple codes.\*\*** Summarize related aspects where possible.

**\*\*Answer Format:\*\***

**\*\*Extracted Codes:\*\***

- [Code Name]: [Brief explanation or relevant text passage as evidence]

**\*\*Example from Another Field:\*\***

- Digital Tools in the Classroom: The teacher describes how tablets increase student participation.
- Time Pressure in Logistics: The driver cites tight time windows as the biggest challenge.
- Workplace Design and Efficiency: Several statements refer to how the setup facilitates or hinders daily work.

**\*\*INTERVIEW TEXT:\*\***

"{interview text}"

Please analyze carefully. Choose **\*\*as few codes as possible\*\*** to capture the interview **\*\*completely but concisely\*\***.

Revision prompt:

You are an expert in qualitative content analysis and are tasked with revising existing

Below is a list of codes extracted from an interview. Your task is to revise these codes by:

1. Combining similar or overlapping codes if they describe the same topic or related aspects.
2. Replacing overly specific or lengthy code names with more abstract, general terms that still capture the central theme.
3. Formulating each code name concisely and clearly (ideally 1-4 words).
4. Preserving the original content by providing a revised but complete description for each new code.
5. Generating a maximum of 3-5 codes in the final result.

**\*\*Examples from other areas:\*\***

- **\*Using digital tools to engage students in mathematics lessons\* → \*Digital tools\***

Description: Statements about the impact of tablets and apps on student participation

- **\*Stress due to tight delivery windows and high order volume\* → \*Time pressure in logistics\***

Description: Indications of high time pressure due to tight route planning and unforeseen delays.

- **\*Challenges in virtual meetings in international teams\* → \*Remote communication\***

Description: Problems with time zones, language barriers, and the lack of nonverbal communication

**\*\*Please submit your results in the following format:\*\***

**\*\*Revised Codes:\*\***

- [New Code Name]: [Summary, concise description relevant to the interview text]

Here are the original codes:

---

{initial code result here}

---

Please optimize the codes according to the above guidelines. Summarize, abstract effecti



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Bibliography

- [AB] year = 2021 doi = 10.3102/1682697 url = https://www.researchgate.net/publication/363823707 *Deductive and Inductive Cycle Process* Andrea Bingham., journal = *Andrea Bingham. Deductive and Inductive Approaches to Qualitative The Five – Cycle Process*.
- [Act] ActLumus. Act Trust 2 sensor documentation.
- [App] Apple. Website of heart app by Apple.
- [ASCea21] I. Axén, S. D. Stellman, D. C. Christiani, and et al. Recruiting in intervention studies: challenges and solutions. *BMJ Open*, 11(1):e044702, 2021.
- [BC06] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006.
- [BO24] First Name Barros and Others. Large language models in qualitative research: A systematic mapping study. *Preprint*, 2024. Available at <insert-URL>.
- [C+19] Mathilde Caron et al. Deep clustering for unsupervised learning of visual features. In *ECCV*, pages 132–149, 2019.
- [CHea] E. Childs, H. Hockaday, and et al. Personality traits modulate emotional and physiological responses to stress. *Psychophysiology*, 51(5):???
- [Cro15] Giancarlo Crocetti. Textual spatial cosine similarity. *CoRR*, abs/1505.03934, 2015.
- [DMM24] Sébastien Mosser Diego Maupomé, Yves Ferstler and Marie Jean Meurs. Automatically finding evidence and predicting answers in mental health self-report questionnaires. In *Conference and Labs of the Evaluation Forum*, 2024.
- [DN+23] Jan-Eric De Neve et al. Measuring workplace wellbeing: Review and prototype survey module. *Oxford Wellbeing Research Paper*, (2303), 2023. Working paper; University of Oxford.
- [DNA21] E. S. Dalmaijer, C. L. Nord, and D. E. Astle. Statistical power for cluster analysis, 2021.
- [Dun24] Zackary Okun Dunivin. Scalable qualitative coding with llms: Chain-of-thought reasoning matches human performance in some hermeneutic tasks, 2024.
- [Fer24] G. Feretzakis. Integrating shapley values into machine learning for interpretability. *Applied Sciences*, 14(13), 2024.

- [GC22] Morris CJ et al. Gao C, Li P. Actigraphy-based sleep detection: Validation with polysomnography and comparison of performance for nighttime and daytime sleep during simulated shift work. *Nat Sci Sleep*, 2022.
- [GMN12] Greg Guest, Kathleen M. MacQueen, and Emily E. Namey. *Applied Thematic Analysis*. SAGE, Thousand Oaks, CA, 2012.
- [GS15] Christian; Gimpel, Henner; Regal and Marco Schmidt. mystress: Unobtrusive smartphone-based stress detection. *ECIS 2015 Research-in-Progress Papers. Paper 16.*, 6, 2015.
- [HA14] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, pages 84–92. Springer, 2014.
- [Han24] Raluca M. Hanciu. Mindwaveml at erisk 2024: Identifying depression symptoms in reddit users. In *Conference and Labs of the Evaluation Forum*, 2024.
- [HRD12] Jin-Hyuk Hong, Julian Ramos, and Anind K. Dey. Understanding physiological responses to stressors during physical activity. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing, UbiComp '12*, page 270–279, New York, NY, USA, 2012. Association for Computing Machinery.
- [HS88] Sandra G. Hart and Lowell E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In Peter A. Hancock and Najmedin Meshkati, editors, *Human Mental Workload*, volume 52 of *Advances in Psychology*, pages 139–183. North-Holland, 1988.
- [htt]
- [J+23] Rachel J. Jarden et al. Wellbeing measures for workers: a systematic review and research agenda. *BMC Public Health*, 23:5678, 2023.
- [Jai99] A. K. Jain. Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323, 1999.
- [K+25] MP Kanak et al. Burnout and sleep problems among nurses working in a high-acuity setting: A global survey. *PLOS Global Public Health*, 5(4):e0003879, 2025.
- [KA16] Göran Kecklund and John Axelsson. Health consequences of shift work and insufficient sleep. *BMJ*, 355:i5210, 2016.
- [KGR+23] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners, 2023.
- [Kul12] B. Kulis. Metric learning: A survey. *Foundations and Trends® in Machine Learning*, 5(4):287–364, 2012.
- [Kul13] Brian Kulis. Metric learning: A survey. *Foundations and Trends in Machine Learning*, 5(4):287–364, 2013.
- [KW25] Muhammad Talal Khalid and Ann-Perry Witmer. Prompt engineering for large language model-assisted inductive thematic analysis, 2025.
- [L+24] L.Z. Li et al. Nurse burnout and patient safety, satisfaction, and quality of care: A systematic review and meta-analysis. *JAMA Network Open*, 7(11):e2825639, 2024.

- [LAGF13] Alexandros Ladas, Uwe Aickelin, Jonathan M. Garibaldi, and Eamonn Ferguson. Using clustering to extract personality information from socio economic data. *CoRR*, abs/1307.1998, 2013.
- [Luo24] Deznabi I. Shaw A. et al. Luo, Y. Dynamic clustering via branched deep learning enhances personalization of stress prediction from mobile sensor data. 2024.
- [LZYH24] Mingming Li, Fuqing Zhu, Feng Yuan, and Songlin Hu. Semantic-enhanced relational metric learning for recommender systems, 2024.
- [M<sup>+</sup>20a] G. J. Martinez et al. On the quality of real-world wearable data in a large-scale study. In *IEEE PerCom 2020*, 2020.
- [M<sup>+</sup>20b] P. Moore et al. Wearable technology and workplace well-being: A systematic review. *Occupational Medicine*, 70(4):235–244, 2020.
- [M<sup>+</sup>23] N. Manjunath et al. Can data augmentation improve daily mood prediction from wearables? In *Proceedings of the ACM Conference on . . .*, 2023.
- [Mac67] J. MacQueen. Some methods for classification and analysis of multivariate observations. 1967.
- [MEA23] Sanjay Kumar Md Ehtasham Ahmed, Shadmah Anwar. Impact of chronic stress on heart rate variability. *International Journal of Clinical Practice Research*, 17(3):320, 2023. Correlation SDNN  $r=-0.58$ , RMSSD  $r=-0.63$ .
- [MFM<sup>+</sup>25] Cristina Martinez Montes, Robert Feldt, Cristina Miguel Martos, Sofia Ouhbi, Shweta Premanandan, and Daniel Graziotin. Large language models in thematic analysis: Prompt engineering, evaluation, and guidelines for qualitative software engineering research, 2025.
- [Moo11] Sarstedt M. Mooi, E. *Chapter 9: Cluster Analysis*. Berlin: Springer-Verlag, 2011.
- [MZP<sup>+</sup>24] Walter S. Mathis, Sophia Zhao, Nicholas Pratt, Jeremy Weleff, and Stefano De Paoli. Inductive thematic analysis of healthcare qualitative interviews using open-source large language models: How does it compare to traditional methods? *Computer Methods and Programs in Biomedicine*, 255:108356, 2024.
- [Nea] A. L. Nevedal and et al. Rapid versus traditional qualitative analysis using the consolidated framework for implementation research (cfir): a mixed-methods comparison of transcription, coding, and findings. *Implementation Science*, 16:??
- [Org22] World Health Organization. Long-term care. <https://www.who.int/europe/news-room/questions-and-answers/item/long-term-care>, 2022. Accessed <date>.
- [P<sup>+</sup>24] Amit Pinge et al. Detection and monitoring of stress using wearables: A systematic review. *Frontiers in Computer Science*, 6:1478851, 2024.
- [Pat03] M. X. Patel. Challenges in recruitment of research participants. *Advances in Psychiatric Treatment*, 9:229–238, 2003.
- [Pat22] V. Patel. Trends in workplace wearable technologies and applications. *AI Society*, 37(3):791–811, 2022.

- [Q<sup>+</sup>25] C. Qiao et al. Comparison of scoring rationales between large language models and human raters. *arXiv preprint*, 2025. Uses embedding-based cosine similarity to compare LLM and human rationales.
- [QST<sup>+</sup>23] Matias Quintana, Stefano Schiavon, Federico Tartarini, Joyce Kim, and Clayton Miller. Cohort comfort models—using occupant’s similarity to predict personal thermal preference with less data. *Building and Environment*, 227:109685, 2023.
- [S<sup>+</sup>09] Martin J Sliwinski et al. Intraindividual change and variability in daily stress processes: Findings from two measurement-burst diary studies. *Psychology and Aging*, 24(4):828–840, 2009.
- [S<sup>+</sup>17] J. Snell et al. Prototypical networks for few-shot learning. In *NIPS*, 2017.
- [S<sup>+</sup>22] J. Subramanian et al. Learning latent structural causal models. *arXiv preprint*, 2022.
- [Sal16] Johnny Saldaña. *The Coding Manual for Qualitative Researchers*. SAGE, London, 3 edition, 2016.
- [Sam] Samsung. Website of stress measurement of samsung smartwatches.
- [SKP15] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *CoRR*, abs/1503.03832, 2015.
- [SRBea18] Ellen Smets, Joaquim Roda, Federico Bonfiglio, and et al. Large-scale wearable data reveal digital phenotypes for daily life stress. *NPJ Digital Medicine*, 1(1):63, 2018.
- [T<sup>+</sup>19] Lara Torquati et al. Shift work and poor mental health: A meta-analysis of longitudinal studies. *Occupational and Environmental Medicine*, 76:193–200, 2019.
- [Wen22] Birkenbihl C. Gomez-Freixa M. et al Wendland, P. Generation of realistic synthetic data using multimodal neural ordinary differential equations, 2022.
- [WSL<sup>+</sup>24] Yihan Wang, Si Si, Daliang Li, Michal Lukasik, Felix Yu, Cho-Jui Hsieh, Inderjit S Dhillon, and Sanjiv Kumar. Two-stage llm fine-tuning with less specialization and more generalization, 2024.
- [WWS<sup>+</sup>22] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903, 2022.
- [XGF15] Junyuan Xie, Ross B. Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. *CoRR*, abs/1511.06335, 2015.