

JDR-LLM: Inferring Job Demands and Resources of PhD Students Using Multimodal LLM-Based Framework

MOHAMMED SAFERI RAHMAN, Florida State University, USA

TE-YEN WU, Florida State University, USA

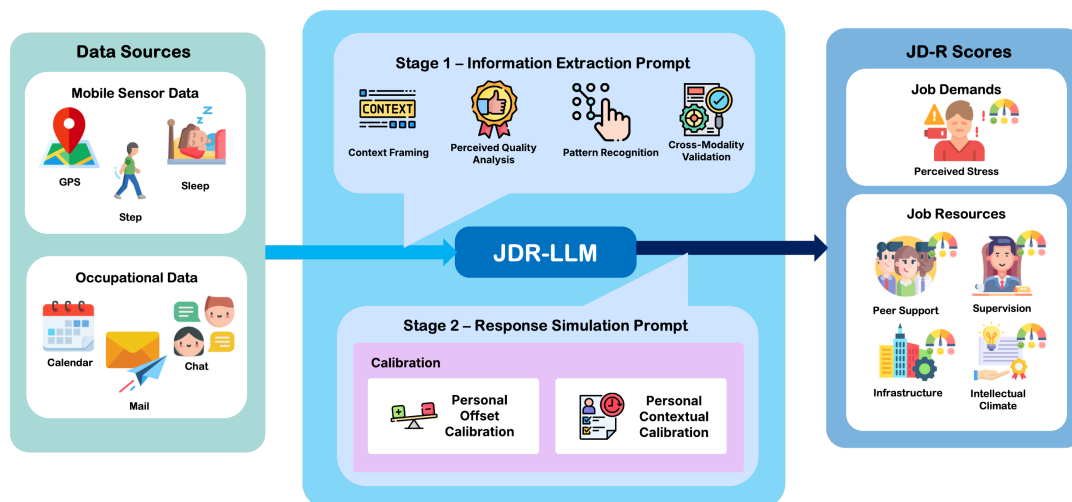


Fig. 1. In this paper, we present an LLM-based framework for inferring Job Demands and Resources (JD-R) of PhD students using multimodal data. The system combines mobile sensor and occupational data through a two-stage prompting design with personalized calibration to estimate perceived stress and academic support.

PhD students frequently face high job demands and limited resources, increasing their risk of mental health challenges. While the Job Demands–Resources (JD-R) model provides a useful framework for assessing well-being, it traditionally relies on self-reports. In this work, we propose a multimodal approach that uses large language models (LLMs) to infer JD-R factors from mobile sensor (GPS, steps, sleep) and occupational data (calendars, emails, and chat logs). We conducted a four-week study with 14 PhD students and introduced a *two-stage prompting framework*: first extracting relevant context, then simulating Likert-scale responses to JD-R questions. We explore two calibration methods, personal offset calibration (POC) and personal contextual calibration (PCC) to improve personalization. Our approach significantly outperforms direct prompting, achieving MAEs of 0.3815 and 0.3441 for job demand and resource scores, respectively. Ablation studies further highlight the value of combining sensing and textual data for passive, continuous well-being assessment.

CCS Concepts: • **Human-centered computing** → **Ubiquitous and mobile computing**.

Authors' Contact Information: Mohammed Saferi Rahman, mr21co@fsu.edu, Florida State University, Tallahassee, Florida, USA; Te-Yen Wu, teyen.wu@fsu.edu, Florida State University, Tallahassee, Florida, USA.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

Manuscript submitted to ACM

Manuscript submitted to ACM

Additional Key Words and Phrases: Job Demands and Resources, PhD Student, Mental Health, Large Language Model, Two Stage Prompting, Personalized Calibration

ACM Reference Format:

Mohammed Saferi Rahman and Te-Yen Wu. 2025. JDR-LLM: Inferring Job Demands and Resources of PhD Students Using Multimodal LLM-Based Framework. 1, 1 (September 2025), 36 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 INTRODUCTION

Due to the high workload and constant performance pressure, the academic journey of PhD students is often accompanied by significant mental health challenges. Research has shown that 32% of PhD students are at risk of developing mental health disorders [36]. To better understand these challenges, researchers have applied the Job Demands-Resources (JD-R) model, originally from occupational psychology, to classify factors as job demands (e.g., academic pressures, supervisor expectations) or job resources (e.g., peer support, supervision) [13]. By analyzing these factors, the JD-R model provides insight into burnout, engagement, and overall performance outcomes for graduate students [5, 6, 24, 32, 35].

While this has been valuable in psychology, it has relied on periodic self-reports, typically collected on a monthly basis. This limits the detection of short-term fluctuations and may delay interventions. Increasing the frequency of self-reports could address this issue but often suffers from low user compliance [63, 74]. Fortunately, advancement in mobile sensing and large language models has shown promise in addressing this problem by automating well-being assessment [9, 47, 61, 68, 69, 71, 75, 79]. For example, StudentLife used smartphone sensor data to track undergraduate mental well-being [68] while Mental-LLM employed large language models to parse online social media data to predict individual’s mental health [75]. Building on these advancements, there is an opportunity to enhance the JD-R model by integrating continuous and automated monitoring, enabling more frequent and timely assessments of PhD students’ occupational well-being.

However, applying these techniques to analyze PhD students’ occupational well-being within the JD-R model presents a knowledge gap. Most existing approaches mainly focus on predicting mental health outcomes, rather than examining the underlying job demands and resources that drive engagement and burnout, offering limited insight into specific workplace factors affecting PhD student well-being. Furthermore, most studies have ignored the potential of occupational digital interactions [8, 10, 37, 42, 55, 72, 78], such as emails and academic discussions, which institutions typically access for compliance purposes but can also serve as valuable indicators of academic pressures, supervisor expectations, and peer support. This oversight restricts the system’s ability to conduct in-depth analyses of job demands and available resources for the JD-R model. Additionally, the integration of LLM-based text analysis with mobile sensing for well-being prediction also remains under-explored, resulting in the absence of a state-of-the-art system capable of capturing the complex interplay between behavioral and linguistic indicators.

In this paper, we examine an LLM-based approach to analyze mobile sensor data and occupational data for assessing the well-being of PhD students within the framework of the JD-R model. We conduct a four-week user study involving 14 PhD students across different domains during the end of a semester, a period typically associated with heightened stress and workload. During this period, we collect a range of mobile sensor data, including GPS location, step count, and sleep duration, along with institutional and occupational data such as work calendars, emails, and digital communication logs (e.g., Microsoft Teams messages). We apply general-purpose LLMs to simulate student responses to JD-R model-based questionnaires using this multimodal dataset, aiming to capture week-to-week fluctuations in perceived stress (job demands) and academic support (job resources).

However, direct prompting lead to inconsistent and unreliable results due to the unstructured and imbalanced nature of the multimodal inputs. To address this, we explore structured prompting techniques that guide the LLM in processing and reasoning across multimodal inputs. We design a two-stage prompting framework, where the first stage extracts and summarizes relevant behavioral, emotional, and contextual information from both sensor and textual data. This extracted summary then serves as the basis for the second-stage prompt, which simulates a Likert-scale response to each JD-R survey question. To further mitigate individual bias, we also investigate different calibration methods to adjust the simulated scores based on prior one-time self-report. Our results show that our approach improves alignment with self-reported scores compared to direct prompting, achieving mean absolute errors (MAE) of 0.3815 and 0.3414 with Spearman correlations of 0.91 and 0.85 for job demands and resources, respectively. We further analyze the effect of different calibration strategies and LLM variants on inference accuracy, and conduct an ablation study to assess the contribution of each modality, highlighting the relative importance of sensor versus textual inputs in the overall inference process.

In summary, our work makes the following contributions:

- We conducted a 4-week user study with 14 PhD students, collecting self-assessments of JD-R model factors alongside mobile sensor data and occupational data for analysis.
- We explore prompting methods for integrating mobile sensing data and occupational textual data within large language models (LLMs) to infer the scores of JD-R model factors.
- We present an evaluation of the inferred scores, comparing them with self-reported scores to assess accuracy, providing insights into the effectiveness and applicability of an LLM-based framework to the JD-R model.

2 RELATED WORK

This section examines existing research in three domains: the JD-R model in the academic context, mobile sensing data in mental health and performance assessment, and large language models in mental health support.

2.1 The Job Demands-Resources Model in the Academic Context

The Job Demands-Resources (JD-R) model is a framework to understand burnout and engagement in occupational settings [35]. It analyzes how workplace resources, such as supervisor support and role clarity, can balance or compensate for job demands, such as workload and emotional stress, to promote employee well-being and performance [8]. When job demands outweigh available resources, employees may experience stress and burnout [13], whereas sufficient resources can foster motivation and engagement. The JD-R model has been applied across various work environments, serving as a foundation for designing interventions that enhance employee resilience and productivity.

Recognizing the structural similarities between work and educational environments, researchers have successfully adapted the JD-R model to academic contexts. Studies in the educational domain highlight how variations in study resources significantly impact student outcomes and well-being [5], with specific research on doctoral candidates emphasizing the need to balance scholarly demands with adequate support systems [32]. For instance, psychological studies have identified key organizational and interpersonal factors, such as academic infrastructure, institutional intellectual climate, peer support, and supervision as critical influences on emotional exhaustion and research commitment [4, 16]. Building on these findings, our work explores the feasibility of leveraging large language models to analyze doctoral students' professional digital interactions and mobile sensing data, enabling continuous monitoring of fluctuations in job demands and resources, aiming to provide insights for universities and HCI researchers to design and implement further interventions that improve student well-being and academic performance.

2.2 Mobile Sensing in Mental Health and Performance Assessment

Since mobile computing has been widely adopted in recent years, researchers have started to investigate how its sensor data, such as GPS-derived mobility patterns, screen time, call/text logs, and accelerometer data, correlate with individuals' mental health [11, 46, 56, 68, 80] and performance [57, 68]. StudentLife is one pioneering study in this field, leveraging smartphones' sensor data such as Bluetooth colocation and accelerometer data to infer behavior changes like irregular sleep, decreased social interactions, and prolonged stationary periods, linking them to mental health (e.g., depression) and academic performance (GPA) [68]. Several studies analyze step count data collected from smartphone accelerometers and found that lower daily step counts are significantly associated with higher depressive symptom severity, suggesting a strong link between physical activity levels and mental well-being [7, 17, 59, 60]. Recent advancements have further explored multimodal data sources, such as smartphone communication logs [33], audio recordings [12, 33], ambient light exposure patterns [64], and wearable-derived heart rate [33, 57] to predict depression or anxiety. To broaden the scope of mobile sensing in mental health assessment, a four-year longitudinal study is also conducted, examining smartphone usage behaviors and observing that fragmented usage patterns, such as frequent unlocking during lectures linked to increased anxiety levels [76].

Additionally, significant research has also utilized mobile sensor data in occupational environments to understand workers' well-being and performance [1, 41, 43, 44, 58]. For example, Mirjafari et al. deploy Bluetooth beacons in workplaces and analyze their data alongside smartphone and wearable sensor inputs from hundreds of information workers to differentiate between high and low job performers [43]. Aside from mobile sensor data, research has also integrated occupational digital footprints, such as email usage time and frequency, to predict stress levels [41] or mood [38]. Beyond detection, mobile sensing has also enabled targeted interventions. For instance, a four-week study has demonstrated that delivering just-in-time "micro-interventions" based on sensed data significantly reduced employees' short-term stress levels [26].

Building on these advancements, our work integrates mobile sensor data (GPS locations, step counts, sleep patterns) with occupational digital interactions (email, chat, calendar events) into an LLM-based framework. This approach enables a nuanced analysis of job demands and resources for doctoral students by extracting content-driven insights from digital communications and linking them with behavioral trends derived from mobile sensing.

2.3 Large Language Models in Mental Health Support

With significant advancements in large language models (LLMs), extensive research has shown LLMs' applicability to identifying and intervening against mental health issues such as stress, anxiety, and depression from text-based data [23, 50, 75]. For instance, Mental-LLM achieves state-of-the-art performance in detecting anxiety and depression markers from online forums, leveraging domain-specific fine-tuning to outperform general-purpose models like GPT-3.5 [75]. Additionally, LLM-powered chatbots, such as Woebot and Wysa, have been developed to provide personalized cognitive behavioral therapy interventions, leading to a 23% reduction in depressive symptoms in clinical trials [18, 28, 50].

Beyond text-based analysis, researchers have also explored integrating multimodal data sources with LLMs to enhance mental health support [47]. For example, MindScape has combined smartphone-derived features, such as screen time, sleep patterns, and physical activities, with LLMs to generate contextually relevant prompts that encourage reflective journaling and mental health tracking [47]. Additionally, some studies even apply LLMs directly to raw mobile sensor data, including heart rate and accelerometer readings, to predict mental health states [14, 19, 49, 79]. This

eliminates the need for traditional feature engineering [15, 31, 39, 67, 70], allowing LLMs to analyze sensor inputs and assess mental health conditions, such as stress levels and PHQ scores [54].

Despite these advancements, a key gap persists in leveraging LLMs to integrate semantic digital content with behavioral mobile sensor data for mental health assessment, especially in occupational settings. Specifically, existing studies often focus on surface-level digital traces, such as interaction frequency and duration [41, 45, 48], while overlooking the rich content of occupational data, such as emails, chats, and schedules, which are accessible to institutions but remain largely underutilized. To address this gap, we conduct a user study involving PhD students to collect and analyze their occupational and behavioral data. We employ structured prompting techniques to enable LLMs to process and synthesize lengthy and multimodal datasets, extracting relevant insights from both digital communication records and mobile sensor data, facilitating more precise assessments of job demands and resources in academic environments.

3 METHODS

In this section, we describe our method to structure prompts of LLMs for analyzing the combination of textual and mobile sensor data to infer factors in the JD-R model among graduate students. To ground our approach, we focus on five key factors identified by Han et al. [24]. These factors have been validated through a large-scale study involving hundreds of graduate students, demonstrating strong relationship with students' academic engagement and emotional well-being. We intentionally retain these original factors without modification or further subdivision to ensure consistency with established constructs, preserving the interpretability of model outputs in the context of existing psychological frameworks. In the following, we briefly explain the factors.

- **Perceived Stress** – A measure of job demands that reflects the extent to which individuals feel overwhelmed or stressed by their job responsibilities. Note that this measure includes various questions related to workload, time pressure, and emotional demands to assess the psychological burden experienced in academic settings. These items are collectively categorized into a single factor by prior work [24] due to the inherently multifaceted nature of academic responsibilities.
- **Peer Support** – A component of job resources, capturing the emotional and practical support provided by fellow graduate students.
- **Supervision** – A job resource factor that evaluates the quality, consistency, and helpfulness of academic guidance and feedback from supervisors.
- **Infrastructure** – A job resource factor that assesses the availability and adequacy of academic resources, including access to research tools, materials, and workspaces.
- **Intellectual Climate** – A job resource factor measuring the degree to which the academic environment is perceived as intellectually stimulating and supportive of open scholarly exchange.

The specific question sets used to assess each factor are detailed in Appendix A. For job demands (e.g., perceived stress), each question is measured on a 5-point Likert scale, ranging from 1 (never) to 5 (very often). Similarly, for job resources (e.g., peer support, supervision, infrastructure, and intellectual climate), each question is measured on a 5-point Likert scale, ranging from 1 (strongly disagree) to 5 (strongly agree). An overall score for each factor is calculated by averaging the responses to its associated questions. To derive standardized metrics, the job demand score is represented by the score for perceived stress, while the job resource score is calculated as the average of the scores for peer support, supervision, infrastructure, and intellectual climate.

Next, based on these validated questions, we designed our prompts to guide the LLM in retrieving relevant information from the occupational data and mobile sensor data. The prompts are structured to reflect the original survey question, enabling the LLM to reason about the content and context of the data and generate synthesized answers that approximate how participants might have answered the original questions.

3.1 Prompt Design

To enable a pre-trained general-purpose LLM to accurately simulate a response to each question, we design our prompts with a two-stage structured format. The two-stage structure include an information extraction prompt, which guides the model to identify and extract relevant information from lengthy textual inputs and mobile sensor data, and a response simulation prompt, which instructs the LLM to simulate a Likert-scale response (1 to 5) to a specific question based on the extracted information. This structured prompting approach empowers the model to process multimodal inputs, isolate the most relevant evidence, and produce outputs aligned with the validated constructs of the JD-R model.

3.1.1 Information Extraction Prompt. While general-purpose LLMs have demonstrated impressive language understanding and reasoning capabilities without requiring domain-specific fine-tuning, we found that directly prompting LLMs to simulate a response often results in inconsistent outputs. The model tends to focus narrowly on isolated events and overlooks broader patterns or contextual nuances. Additionally, due to the imbalance between lengthy textual inputs and relatively brief mobile sensor data, LLMs frequently underutilize the available sensor data.

To address these issues and enable the LLM to consider more comprehensive and balanced information when simulating a response, we designed the first-stage prompts to extract and summarize relevant information from both the textual and mobile sensor data for each question. This step helps the LLM distill meaningful behavioral, emotional, and contextual signals that may be buried in lengthy input or overlooked due to modality imbalance. Specifically, each information extraction prompt is designed to target one of the following four objectives:

Context Framing: Prompts in this category are designed to identify when relevant events occurred by establishing a clear temporal and contextual boundary. Without explicit anchoring of time or context, LLMs might conflate multiple occurrences or miss references entirely. By specifying the boundaries (e.g. "during the user's lab sessions"), we reduce ambiguity and help the LLM align textual and sensor data to a coherent timeframe.

Perceived Quality Analysis: These prompts aim to extract insights about the subjective quality or emotional tone of an event. Although objective data (such as location logs) can indicate whether something happened, it does not convey the user's satisfaction, frustration, or emotional reaction. For instance, when analyzing supervision, a prompt might ask the model to highlight any user statements describing how helpful or unhelpful the supervisor's feedback felt. This approach surfaces the user's qualitative perception rather than just the occurrence of an event.

Pattern Recognition: This type of prompt determines whether an observed event is an isolated incident or recurs over time, thus capturing behavioral consistency or cyclical triggers. For instance, if a user frequently describes feeling stressed, a pattern-recognition prompt would explore whether these stress episodes consistently occur in the same context (e.g., every day following a supervisor meeting, or right before a weekly lab session), thereby revealing whether certain events or time windows systematically provoke stress.

Cross-modality Validation: Finally, these prompts encourage the model to compare and synthesize information across text and sensor data, ensuring that extracted evidence is strengthened by multiple modalities. For example, if the user's textual data suggests they feel overloaded by research deadlines, a cross-modality prompt would instruct the

model to check late-night location logs or sleep records for signs of extended lab stays, insufficient rest, or a reduction in physical activity.

We formalize information extraction prompts using the following template:

$$\text{Prompt}_{IE} = \text{TextData} + \text{SensorData} + \text{Prompt}_{Info-Q} + \text{OutputConstraint}_{IE} \quad (1)$$

where *TextData* refers to institutional occupational data such as work emails, chat conversations, and calendar schedules; *SensorData* consists of mobile sensing outputs including GPS traces, step counts, and sleep logs; *Prompt_{Info-Q}* poses specific subquestions for the LLM to answer; and *OutputConstraint_{IE}* limits the response format (e.g., “Return the information within 250 tokens”). Table 1 presents examples of information extraction prompts designed for a single JD-R question, along with the corresponding outputs generated by an LLM (GPT-5). Appendix B provides the complete set of information extraction prompts for all JD-R questions. Appendix C demonstrates more examples to further illustrate the extracted outputs across different contexts.

Table 1. Example of one JD-R question from the Perceived Stress scale, branched into four information extraction prompts to summarize and contextualize relevant textual and behavioral data

JD-R Question	“In the last week, how often have you found that you could not cope with all the things that you had to do?”
Context Framing	<p>Prompt: “From the textual data, find or infer any instances where the user appears overwhelmed, overworked, or unable to handle tasks. Specify which deadlines, responsibilities, or time blocks are linked to these statements.”</p> <p>Response: “From the data provided, [STUDENT NAME] seems to experience periods of being overwhelmed or overworked. On [DATE 1], they convey being ‘very occupied with course works with many presentations and reports to prepare’ to [PROFESSOR NAME], highlighting a time block where academic obligations may be taxing. Their calendar corroborates the busy schedule, showcasing a dense timeline filled with work-related events and minimal leisure. Specifically, the pattern of extended work time on [DATE 1] (spanning from [TIME] AM to [TIME] AM across multiple sessions), coupled with limited breaks, suggests an extensive workload. Moreover, on [DATE 2], after field sampling until [TIME] PM, they continued with work, showing a packed schedule and strains linked to a major project. These reflections and their corresponding schedule details indicate a challenging balance between coursework and research commitments, where maintaining well-being alongside academic duties appears challenging.”</p>
Perceived Quality Analysis	<p>Prompt: “Extract how the user describes the severity of not coping (e.g., ‘barely managing,’ ‘completely swamped’) and highlight any emotional tone (frustration, panic, resignation).”</p> <p>Response: “Based on the chat and schedule data for [STUDENT NAME], they describe feeling ‘very occupied’ with coursework—indicating high academic engagement and potential stress due to multiple presentations and reports. However, there is no explicit mention of being overwhelmed or struggling to cope (e.g., terms like ‘barely managing’ or ‘completely swamped’ are absent). The emotional tone in their messages remains composed and professional, as they politely request schedule adjustments and course recommendations. Their communication style reflects a disciplined approach to handling academic demands, despite a notably busy schedule.”</p>

Table 1. (Continued)

JD-R Question	“In the last week, how often have you found that you could not cope with all the things that you had to do?”
Pattern Recognition	<p>Prompt: “Identify or infer each instance in the user’s text where they indicate being unable to cope. Check whether these instances recur at similar intervals or revolve around common triggers (e.g., project deadlines). Summarize any patterns or recurring factors that emerge.”</p> <p>Response: “In the chat and schedule logs, [STUDENT NAME] expresses feeling occupied due to coursework, presentations, and reports on [DATE 1]. This aligns with a densely packed schedule, primarily consisting of academic work with limited rest periods. The pattern of high workload continues across several days [DATES], including fieldwork, lab responsibilities, and tight research commitments. A discussion of a detailed work plan on [DATE 2] further highlights concurrent management of intense academic and research projects. The overall schedule reflects recurring academic demands and project-related stressors. While there are occasional short breaks and some leisure activities, they may be insufficient to counterbalance the ongoing workload pressures.”</p>
Cross-modality Validation	<p>Prompt: “Compare the user’s explicit or inferred references to feeling overwhelmed with sensor data (e.g., late-night location logs, reduced step counts). Determine whether these reported or inferred episodes coincide with extended work hours or other observable signs of overload in the data.”</p> <p>Response: “[STUDENT NAME] shows signs of being overwhelmed, explicitly stating their workload involving presentations and reports in a chat message on [DATE 1]. Calendar data confirms extended work hours from [TIME] AM to after midnight on that date, aligning with the expressed busyness. Similar patterns were observed on [DATE 2] and [DATE 3], including work on weekends. Mobile sensing data reveals low physical activity, particularly on [DATE 4] and [DATE 5], indicating long sedentary periods during work hours and not enough sleep at night. This combination of prolonged academic engagement, reduced physical movement and less sleep suggests episodes of work-related strain, consistent with the participant’s self-described feelings of being overwhelmed.”</p>

3.1.2 Response Simulation Prompt. After generating contextual summaries with information extraction prompts, we aggregate them to form a comprehensive view of the participant’s situation. This summary is then used in a response simulation prompt, which aims to guide the LLM to simulate how the participant might respond to the corresponding JD-R question. Formally, we structure this prompt as follows:

$$\text{Prompt}_{RS} = \text{Responses}_{IE}[n] + \text{Instruction}_{RS} + \text{Prompt}_{JD-R-Q} + \text{OutputConstraint}_{RS} \quad (2)$$

where $\text{Responses}_{IE}[n]$ represents a concatenation of all answers from the information extraction prompts, and Prompt_{JD-R-Q} is the original JD-R item (e.g., “In the last week, how often have you found that you could not cope with all the things that you had to do?”). In Instruction_{RS} by instructing the LLM to “assume yourself as the student”, we encourage it to synthesize the extracted information and produce a final response that mirrors the user’s perspective. This part also incorporates calibration context, either as a scalar offset or embedded examples, as detailed in (Section 3.2). And $\text{OutputConstraint}_{IE}$ limits the response format (e.g., “Respond in Likert scale integers.”)

In this way, the *Response Simulation Prompt* leverages the richer, multi-faceted summaries from the *Information Extraction Prompt* to yield more nuanced, context-aware JD-R survey answers. Note that within the JD-R framework, the primary focus is on the overall score for each factor rather than individual question responses. To infer the overall

score for factor, we average the simulated responses generated by the LLM for all questions associated with a given factor.

3.2 Calibration

To ensure that model-simulated responses consider individual differences and align with users' self-perceived experiences, personalization is necessary through calibration against self-reported scores, which account for individual bias. To achieve this, we explore two calibration strategies, aiming to adjust the simulated responses to better reflect each participant's baseline perception and improve the accuracy of inferred JD-R scores.

Personal Offset Calibration (POC). One straightforward approach to mitigating systemic biases in simulated scores is applying an offset based on the discrepancy between prior ground-truth responses and model-simulated responses for each question. Specifically, we compute the difference between a participant's self-reported Likert-scale rating and the corresponding model-simulated scale rating for each question, treating this difference as an individual offset. This offset is then applied to future simulated responses for that participant, calibrating the model's outputs to better reflect the participant's subjective baseline. This method helps correct systematic estimation, enabling lightweight personalization of the inferred well-being scores.

Personal Contextual Calibration (PCC). The other, more sophisticated approach is using LLMs to contextually calibrate their simulated responses. In this strategy, we provide the model not only with the current summary information but also with prior examples that include previously generated summaries, the corresponding model-simulated responses, and the actual self-reported responses. By learning from these past discrepancies in context, the LLM can adjust its reasoning and better align future predictions with the participant's typical reporting behavior. This contextual calibration allows the model to account for individual response tendencies and improves the consistency and accuracy of simulated scores over time.

4 IMPLEMENTATION

This section details the implementation of our method, including the data sources collected and the large language models (LLMs) employed.

4.1 Data Sources

We integrate two primary data streams: *Institutional Occupational Data* and *Mobile Sensor Data*.

4.1.1 Institutional Occupational Data (Work Chat, Emails, and Calendars). Major enterprises and universities conduct their internal communication and scheduling through enterprise platforms such as Microsoft 365 and Google Workspace. In accordance with institutional policy and data-retention regulations, system administrators already hold authorized access to these communications for compliance and auditing purposes [22, 29, 62]. Building on this existing infrastructure, we can extract both interaction metadata, message counts, reply latencies, meeting frequencies, and de-identified textual content from work e-mails, chats, and calendar descriptions. This institutional occupational data reveals each participant's workload patterns, collaboration networks, and scheduling constraints, allowing the LLM to recognize contextual signals of demands and resources (e.g., recurring deadlines, overlapping commitments, supervisor feedback). In addition to that, calendar entries further provide explicit time blocks devoted to academic tasks, personal events, and discretionary downtime, supplying a fine-grained temporal scaffold for downstream modeling.

4.1.2 Mobile Sensor Data (GPS, Step Counts, and Sleep). We supplement the occupational data with smartphone sensing logs that capture fine-grained behavioral patterns. Specifically, we collect GPS traces to determine whether participants are on campus, serving as a proxy for academic engagement or physical presence in work-related environments. In addition, we record daily step counts to gauge physical activity levels, and collect sleep data to assess rest and recovery. These data sources allow the LLM to support derived stress or workload (e.g., late-night tasks) with observed physical and spatial patterns (e.g., consistent late-night lab visits, reduced activity). While other types of sensor data may be available (e.g., heart rate), our selected modalities are representative relevant to well-being, and can be collected using only a smartphone.

4.2 Data Processing

We implement a multi-step data processing pipeline to ensure privacy as well as temporal and semantic alignment of the collected multimodal data.

4.2.1 Data Anonymization. Ethical handling of participants' personal data is paramount. Thus, all identifiable content (names, email addresses, exact GPS coordinates) was obfuscated or abstracted. Occupational data (such as chat logs, emails, and calendar entries) were pseudonymized by replacing names with role based placeholders (e.g., [Advisor]). For mobile sensor data, particularly GPS traces, were transformed via reverse geocoding and place abstraction (e.g., "Home", "Lab", "University Library", etc.) using Google's Reverse Geocoding [21] and Nearby Search [20] APIs. This approach preserves location context without revealing precise coordinates, in line with AutoLife's privacy sensitive design [73].

4.2.2 Temporal Alignment. To support coherent day-to-week inference, all timestamps were normalized to the local time zone, maintaining participants' actual daily rhythms. Occupational data (chats, emails, calendar) were chronologically sorted and aggregated into per day records. Mobile sensor streams (location traces, step counts, inferred sleep/wake cycles) were also aggregated daily. This ensures that each day's multimodal footprint aligns with weekly JD-R self-reports, matching the temporal granularity needed for accurate results.

4.2.3 Semantic Alignment. Given the heterogeneity of collected data, we implemented modality-specific semantic normalization to enhance interpretability, reduce noise, and minimize token cost for LLM-based inference. For chat logs, we filtered out system generated messages and retained only conversational exchanges involving peers or advisors, prioritizing interactions likely to reflect meaningful social or emotional contexts. For emails, HTML tags were stripped, boilerplate and signature content removed, and entries deduplicated using a composite key of sender, subject, timestamp, and message prefix to avoid redundancy and reduce input length. For calendar events, we extracted structured metadata including title, time, location, and attendees, while applying special handling for recurring and all-day events to preserve semantic context. All sender, receiver names in chat logs and emails, and attendee names in calendar events were pseudonymized. For mobile sensor data, we followed the AutoLife framework [73] to enrich raw streams with contextual meaning. GPS coordinates were transformed via reverse geocoding and place abstraction. Additionally, activity metrics such as daily step counts and motion intensity were aggregated into physical activity summaries, while sleep-related signals were processed to estimate bedtime, wake time, and perceived sleep quality. These transformations distilled the raw multimodal data into coherent, interpretable semantic tokens grouped by day, enabling contextually grounded LLM processing.

The resulting weekly JSON structures for each modality are provided as illustrative pseudocode in Appendix E.

4.3 Models

We evaluate three proprietary OpenAI models *GPT-5*, *GPT-4o*, and *GPT-4.1*, together with the open-source *Llama-4-Scout* [2, 27, 51, 52]. We include OpenAI’s most capable model to date, GPT-5 as the upper-bound baseline for reasoning quality and reliability in complex, multi-step inference [52]. *GPT-4o* is selected as our high-throughput multimodal model. It delivers high performance with substantially lower latency and cost. *GPT-4.1* is chosen specifically for its extreme long-context capacity. Finally, we incorporate *Llama-4-Scout* to enable open-source comparability, reproducibility, and a local deployment option.

- **GPT-5:** This most recently released model is closed-source and available through API provided by OpenAI. It preserves the 400k-token context window and a 128k token maximum output with reasoning token support, providing noticeably richer long-range reasoning and dialogue coherence at the cost of increased latency. We picked the `gpt-5-2025-08-07` API, one of the most versatile, high-intelligence flagship models of OpenAI.
- **GPT-4o:** A high-performance model from OpenAI’s GPT-4 family, GPT-4o is designed for seamless multimodal processing and low-latency interactions. It features a 128k token context window and supports up to 16,384 output tokens, striking a balance between capability and computational efficiency. Unlike GPT-5, it is optimized for real-time applications and remains a preferred choice for tasks requiring fast, contextually aware reasoning without the overhead of extreme context depth. We use the `gpt-4o-2024-08-06` API in our experiments.
- **GPT-4.1:** A technically ambitious release from OpenAI, GPT-4.1 features an unprecedented 1,047,576-token context window and a 32,768-token maximum output. It is architected to accommodate extremely large input artifacts. Despite its substantially larger receptive field, architectural refinements yield a runtime profile closer to GPT-4o than parameter count alone would suggest. We picked the `gpt-4.1-2025-04-14` API.
- **Llama-4-Scout:** We include an open-source baseline using the instruction-tuned `llama4:17b-scout-16e-instruct-q4_K_M` model, a 17B-parameter model from Meta’s Llama 4 family quantized to 4-bit for memory-efficient inference. This variant enables local deployment while retaining reasonable reasoning ability. All evaluations were conducted on a workstation equipped with 128GB RAM and an NVIDIA RTX A6000 GPU (48GB VRAM). Although it lacks the refined alignment and contextual optimization of commercial models, it provides a transparent, reproducible alternative for multimodal analysis tasks.

Each variant is applied to both the information-extraction and response-simulation stages, allowing us to observe how model capacity affects downstream behavior in the context of extracting key information from diverse textual modalities and mobile sensor data, and interpreting job demands and resources from such multimodal inputs.

5 FIELD STUDY

Due to the sensitivity of institutional data, no open-source datasets are available for our analysis. To address this, we conducted a four-week user study involving 14 PhD students during the final month of the academic semester. While four weeks may seem limited in duration, this period shows variability in job demands and resources due to course project deadlines, final presentations, and academic evaluations. This fluctuation is evidenced by the self-reported data presented in Section 5.4.1, which highlights the variations in perceived stress levels and available academic support throughout the period. In this section, we describe the study procedure, recruitment, data collection, and results.

5.1 Participants

We initially recruited 17 PhD students (9 male, 8 female; ages 24–34) from Florida State University, representing a range of disciplines including Biological Science, Psychology, Electrical and Computer Engineering, Civil and Environmental Engineering, Industrial and Manufacturing Engineering, Earth, Ocean and Atmospheric Science, Biochemistry and Chemical Engineering. One participant (male) withdrew midway through the study, and two participants submitted incomplete data (1 male, 1 female), resulting in 14 participants (7 male, 7 female) with complete datasets included in the final analysis. Appendix D provides more details about the demographic characteristics of participants. The study was approved by the Institutional Review Board (IRB), and all participants provided informed consent for collected data. Participants were compensated with a \$50 gift card for their time and contribution to the study.

5.2 Apparatus

To support multimodal data collection, we developed a custom smartphone application that was installed on each participant’s device. The app passively recorded GPS location, step counts, and sleep patterns, and securely synchronized this data to a central server in real time. In parallel, we leveraged the Microsoft Graph API to access participants’ occupational data. Upon user authentication, we retrieved calendar events, email metadata, and digital communication logs (e.g., Microsoft Teams), providing a view of their academic and work-related activities.

5.3 Study Procedure

Participants were enrolled in a four-week longitudinal study during the final month of the semester. Upon onboarding, each participant installed our custom data collection app on their smartphones. Throughout the study, participants continued with their regular academic activities without any intervention. At the end of each week, the app delivered a prompt asking participants to complete the JD-R questionnaire, which included items measuring perceived stress (representing job demand) and four dimensions of academic support, peer support, supervision, infrastructure, and intellectual climate, representing job resources (see Appendix A). These weekly self-reports served as ground-truth labels to allow us to compare simulated scores with participants’ actual scores over time in the JD-R framework.

5.4 Results

We first demonstrate the comparison between self-reported and LLM-inferred job demand and resources. We then summarize our experiment with different calibration methods, prompt designs, and models. We finally discuss the ablation study to show the importance of both modalities.

5.4.1 Self-reported vs. LLM-inferred Job Demand and Resources. We present the self-reported and LLM-inferred scores for job demands and resources in Figure 2. From the self-reported scores, we observe the average absolute week-to-week changes of 0.60, 0.36, 0.62, 0.50, and 0.53 across participants for perceived stress, peer support, supervision, infrastructure, and intellectual climate, respectively. Notably, perceived stress (0.60) and supervision (0.62) showed the largest fluctuations, reflecting the dynamic nature of job demands and advisor–student relationships during the end of the semester.

Building on this foundation, we derived LLM-inferred scores using the GPT-5 model from the second week onward, as the first week’s data were reserved for calibration. Also, because only the first week data were used for calibration, predictions for subsequent weeks were generated independently of one another, making each week a separate evaluation instance. This strengthens the validity of our results by ensuring that the model’s inferences do not rely on longitudinal

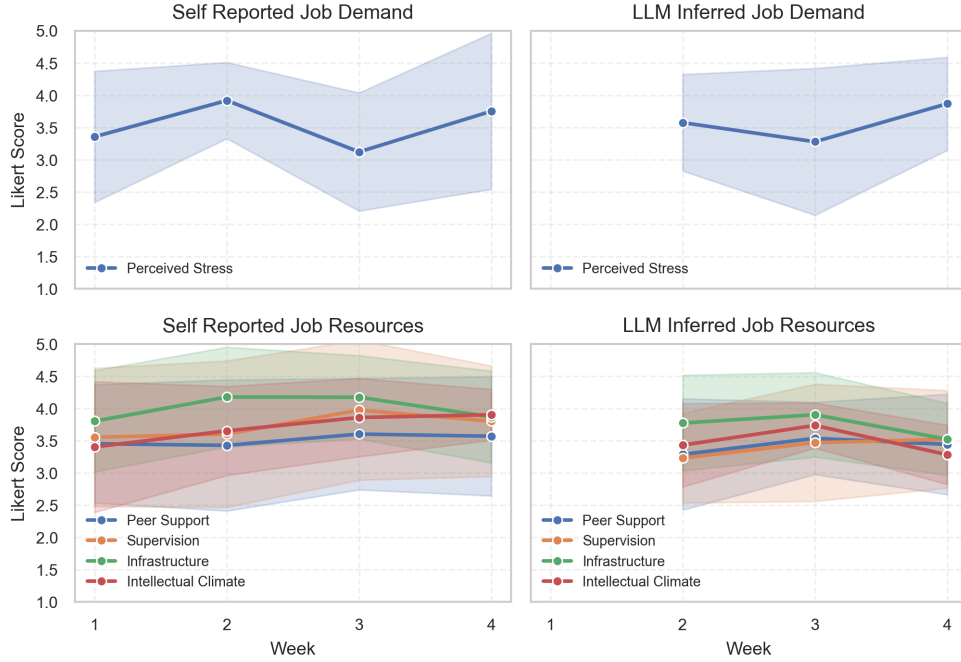


Fig. 2. Weekly trends in JD-R indicators. The top row shows self-reported scores for Job Demand (Perceived Stress) and Job Resources (Peer Support, Supervision, Infrastructure, Intellectual Climate), while the bottom row shows corresponding LLM-inferred scores generated using the 2-step PCC method on GPT-5. Lines show mean Likert scores; shaded areas represent variability (standard deviation) across participants. Data is drawn from self-reported responses of PhD students during the final month of the semester.

trends beyond the calibration point. While the long-term validity of the results is still unknown, it remains valid within at least a one-month window, enabling us to capture short-term fluctuations from monthly self-reports.

In general, the LLM-inferred scores show strong alignment with participants' self-reported values, with Spearman correlations ρ of 0.91, 0.88, 0.90, 0.80, and 0.82 for perceived stress, peer support, supervision, infrastructure, and intellectual climate, respectively (all $p \ll 0.001$). These high correlations demonstrate that the LLM is able to capture relative fluctuations in job demands and resources over time. For job demand, the LLM-inferred scores closely followed self-reported values, both indicating a general increase in stress among PhD students across the study period. A noticeable dip was observed in the third week in both measures, coinciding with a holiday break that temporarily reduced academic workload. On the job resource side, LLM-inferred scores for peer support, supervision, and infrastructure similarly tracked self-reported values, reflecting consistent trends across these factors. The one exception is the intellectual climate factor. While self-reported scores increases in the final week, the LLM-inferred score drops. Upon closer inspection of the LLM-generated data, we observed a reduction in departmental and university-wide academic events during the final week, which influenced the LLM's interpretation and led to a lower inferred score for intellectual climate.

To quantitatively assess how well the LLM-inferred scores predict participants' self-reported responses, we also compute the Mean Absolute Error (MAE) between the two. For each JD-R factor, we first calculate a composite score by averaging responses to the associated questions. We then compute the MAE for each participant by averaging the absolute differences between the inferred and self-reported composite scores across all available weekly time points.

Table 2. Mean Absolute Error (MAE) and rank correlation (Spearman’s ρ) across JD-R factors, including aggregated averages.

	Job Demand	Job Resources				Overall
	Perceived Stress	Peer Support	Supervision	Infrastructure	Int. Climate	Avg.
MAE ↓	0.3815	0.3529	0.3647	0.2882	0.3706	0.3441
Spearman’s ρ ↑	0.910	0.881	0.900	0.805	0.819	0.851

All $p \ll 0.001$.

These participant-level MAEs are then averaged across all participants to evaluate overall model performance for each JD-R factor. Using this approach, our method achieved a low MAE of 0.3815 for job demand and 0.3441 for job resources, demonstrating the alignment between the LLM-generated inferences and participants’ self-reported well-being scores. The detailed MAE for each JD-R factor is presented in Table 2.

5.4.2 Effect of Calibration on LLM-inferred JD-R Scores. To further understand the influence of calibration methods on inferred scores, we compare three strategies: no calibration, personal offset calibration (POC), personal contextual calibration (PCC). For both POC and PCC, we use data from the first week to calibrate simulated responses. All tasks are performed using GPT-5. The results are shown in Figure 3.

PCC consistently achieved lower MAE than *POC* across all JD-R categories, including job demand (0.3815 vs. 0.5391) and job resources (0.3441 vs. 0.5265), resulting in a substantially lower overall MAE (0.4148 vs. 0.6521). Paired *t*-tests confirmed that these improvements were statistically significant for peer support, infrastructure, intellectual climate, and overall MAE (all $p < 0.05$), while differences in perceived stress and supervision did not reach significance. These findings highlight *PCC* as a more reliable calibration method, with *POC* serving as a simpler alternative when token budgets are constrained.

In contrast, the *No Calibration* condition resulted in a substantially higher MAE of 1.2789, highlighting the importance of personalized calibration when inferring well-being scores. Statistical analysis confirmed that both *PCC* and *PCO* significantly outperformed the no calibration approach for job demand and job resource factors (all *p*-values < 0.05).

5.4.3 Direct vs. Structured Prompting for Inferred JD-R Scores. To understand the impact of our prompt design on inferred scores, we compare our structured prompting approach with the baseline direct prompting method. In the direct prompting approach, only the final response simulation prompt is used, with all processed occupational and mobile sensor data as input. This method omits any intermediate information extraction or structured guidance, relying solely on the LLM’s internal reasoning capabilities to identify relevant patterns and generate Likert-scale responses. Because no intermediate information extraction is performed, only POC can be applied to this method.

By comparing these two approaches, we aim to assess whether explicitly guiding the model through structured information extraction leads to more accurate and consistent simulated responses. Both approaches are based on GPT-5 with POC for fair comparison. The results are presented in Figure 4. We found that the structured prompting approach significantly outperforms the direct prompting method across all JD-R factors (all *p*-values < 0.05). Specifically, the structured prompting method achieves substantially lower MAE scores for both job demand (0.5391 vs. 1.0297) and job resource (0.5265 vs. 0.9184), yielding a markedly lower overall MAE (0.529 vs. 0.941). The error distributions in Figure 4 are notably tighter under structured prompting, with consistently lower medians and reduced spread across most JD-R categories. Direct prompting, by contrast, produces wider inter-quartile ranges and higher variability, especially in perceived stress and supervision, where errors are both larger and more dispersed. These findings underscore the

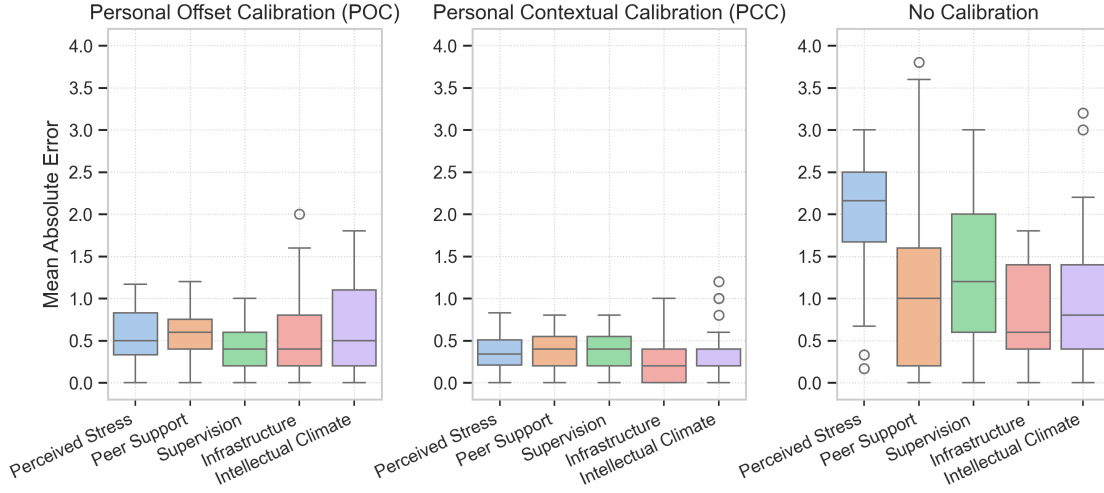


Fig. 3. Mean Absolute Error (MAE) distributions for five JD-R factors across three GPT-5 scoring strategies, all employing a 2-step prompting method. The strategies compared are: Personal Offset Calibration (POC), Personal Contextual Calibration (PCC), and No Calibration. Each boxplot summarizes per-participant MAEs, with outliers shown as dots.

importance of using the two-stage prompt design for enhancing inference and improving alignment with self-reported JD-R scores.

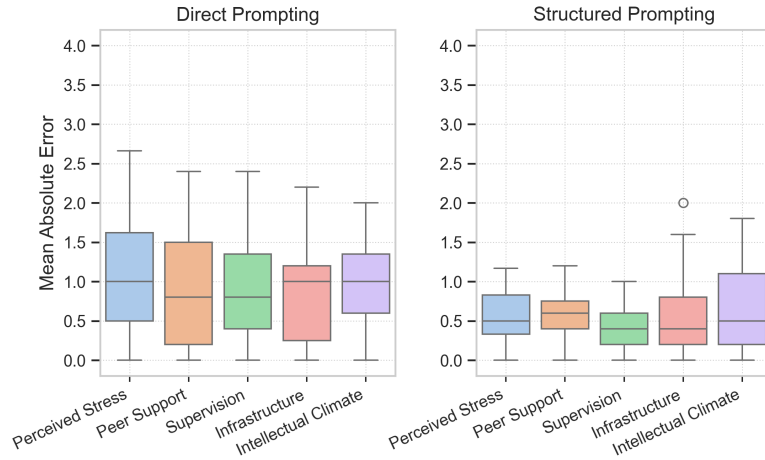


Fig. 4. Comparison of MAE across JD-R categories between direct prompting and structured prompting using GPT-5, both with Personal Offset Calibration (POC).

Table 3. Mean Absolute Error (MAE) across JD-R dimensions for two scoring/calibration strategies (2-Step RS_{POC}, 2-Step RS_{PCC}) and multiple LLMs. Subscripts denote Spearman’s ρ .

Method	Model	Job Demand	Job Resources				Overall
		Perceived Stress	Peer Support	Supervision	Infrastructure	Int. Climate	
2-Step RS_{POC}	<i>GPT-4.1</i>	0.7991 _{$\rho=0.26$}	0.7765 _{$\rho=0.51$}	0.8059 _{$\rho=0.42$}	0.5706 _{$\rho=0.48$}	0.6294 _{$\rho=0.46$}	0.7163 _{$\rho=0.53$}
	<i>GPT-4o</i>	0.6762 _{$\rho=0.42$}	0.5353 _{$\rho=0.73$}	0.5176 _{$\rho=0.65$}	0.5882 _{$\rho=0.48$}	0.6294 _{$\rho=0.47$}	0.5894 _{$\rho=0.64$}
	<i>GPT-5</i>	0.5391 _{$\rho=0.56$}	0.5471 _{$\rho=0.71$}	0.4176 _{$\rho=0.78$}	0.5353 _{$\rho=0.63$}	0.6059 _{$\rho=0.44$}	0.5290 _{$\rho=0.69$}
	<i>Llama-4</i>	0.9184 _{$\rho=0.46$}	0.7297 _{$\rho=0.68$}	0.6270 _{$\rho=0.75$}	0.5368 _{$\rho=0.72$}	0.7514 _{$\rho=0.41$}	0.7058 _{$\rho=0.68$}
2-Step RS_{PCC}	<i>GPT-4.1</i>	1.1750 _{$\rho=0.45$}	0.8222 _{$\rho=0.30$}	0.9222 _{$\rho=0.23$}	0.5167 _{$\rho=0.65$}	0.7056 _{$\rho=0.18$}	0.8283 _{$\rho=0.49$}
	<i>GPT-4o</i>	0.5881 _{$\rho=0.66$}	0.4889 _{$\rho=0.73$}	0.5222 _{$\rho=0.77$}	0.3778 _{$\rho=0.74$}	0.5833 _{$\rho=0.48$}	0.5121 _{$\rho=0.74$}
	<i>GPT-5</i>	0.3815 _{$\rho=0.91$}	0.3529 _{$\rho=0.88$}	0.3647 _{$\rho=0.90$}	0.2882 _{$\rho=0.81$}	0.3706 _{$\rho=0.82$}	0.3516 _{$\rho=0.88$}
	<i>Llama-4</i>	0.6524 _{$\rho=0.59$}	0.8162 _{$\rho=-0.12$}	0.9622 _{$\rho=0.10$}	0.8108 _{$\rho=0.06$}	0.7135 _{$\rho=0.23$}	0.7910 _{$\rho=0.32$}

5.4.4 Simulated JD-R Scores Across Different LLMs. To understand the influence of model choice on JD-R simulation, we compare GPT-4.1, GPT-4o, GPT-5, and the open-source Llama-4-Scout, across two calibration strategies: *POC* (Personal Offset Calibration) and *PCC* (Personal Contextual Calibration). The results are summarized in Table 3.

Across both calibration methods, GPT-5 consistently yields the best performance. Under *PCC*, GPT-5 achieves the lowest Mean Absolute Error (MAE) for both job demand and job resources. GPT-4o, with a 128k context window and tuned for real-time use, also performs competitively. Under *PCC*, it reaches an overall MAE of 0.5121 and Spearman correlation $\rho = 0.74$, indicating consistent and relatively reliable performance, especially considering its lower compute cost compared to GPT-5. In contrast, GPT-4.1, despite offering the largest context window, underperforms relative to GPT-4o and GPT-5. It yields the highest overall MAE (0.7163 under *POC* and 0.8283 under *PCC*), and its correlations are weaker (0.53 under *POC* and 0.49 under *PCC*). This indicates that sheer context length does not guarantee improved modeling of multimodal behavioral patterns. Finally, the open-source Llama-4-Scout, serving as a budget and reproducible alternative, perform slightly better than GPT-4.1. It achieves overall MAE of 0.7058 and correlation of 0.68 under *POC*. However, under *PCC*, its performance deteriorates with very weak correlation and high MAE. This suggests that the model struggles to utilize extended context when guided by structured calibration. Consequently, when using Llama-4-Scout as a low-cost option, we recommend applying it under *POC* rather than *PCC*.

5.4.5 Ablation Study: Contribution of Each Modality. To understand the contribution of each data modality, we conducted an ablation study by evaluating the performance of our model using only occupational data, only mobile sensor data, and both combined (multimodal). In all conditions, we used GPT-5 as the underlying model with personalized offset calibration (*POC*). We excluded any prompt that explicitly required multimodal input, ensuring that the remaining prompts could be reasonably inferred from a single modality.

The results are shown in Table 4. As expected, no single modality outperforms the multimodal approach. Mobile sensor data alone performs the weakest, with the highest overall MAE (0.8566) and low correlation ($\rho = 0.45$). In contrast, occupational data performs relatively well, particularly for *Infrastructure* (MAE = 0.5789, $\rho = 0.49$) and *Intellectual Climate* (MAE = 0.6789, $\rho = 0.55$), suggesting that structured organizational inputs can reasonably capture institutional resources and intellectual aspects of the academic environment. However, when complemented with mobile sensor data, performance improves notably for *Perceived Stress* (MAE from 0.7668 to 0.5391, ρ from 0.42 to 0.56), *Peer Support* (MAE from 0.6158 to 0.5471, ρ from 0.56 to 0.71), and *Supervision* (MAE from 0.6368 to 0.4176, ρ from 0.69 to 0.78). This highlights the critical role of behavioral and contextual signals in modeling interpersonal and individual-level

Table 4. Mean Absolute Error (MAE) across JD-R dimensions using different Data Modalities. All methods utilize the 2-step prompting with *POC* on GPT-5. Subscripts denote Spearman's ρ .

Data Type	Job Demand	Job Resources				Overall
	Perceived Stress	Peer Support	Supervision	Infrastructure	Int. Climate	
Occupational Data	0.7668 _{$\rho=0.42$}	0.6158 _{$\rho=0.56$}	0.6368 _{$\rho=0.69$}	0.5789 _{$\rho=0.64$}	0.6789 _{$\rho=0.55$}	0.6555 _{$\rho=0.66$}
Mobile Sensor Data	0.9370 _{$\rho=0.17$}	0.7676 _{$\rho=0.52$}	0.8703 _{$\rho=0.40$}	0.8162 _{$\rho=0.21$}	0.8919 _{$\rho=0.33$}	0.8566 _{$\rho=0.45$}
Multimodal Data	0.5391 _{$\rho=0.56$}	0.5471 _{$\rho=0.71$}	0.4176 _{$\rho=0.78$}	0.5353 _{$\rho=0.63$}	0.6059 _{$\rho=0.44$}	0.5290 _{$\rho=0.69$}

experiences, demonstrating that the integration of both data modalities yields a more accurate inference of job demands and resources.

6 DISCUSSION

In this section, we discuss how the inferred JD-R scores can be applied within the JD-R framework to better understand burnout and engagement among PhD students in academic settings. We then explore how these scores can inform strategies aimed at enhancing well-being and improving performance.

6.1 Using LLM-inferred JD-R Scores to Understand Burnout and Engagement

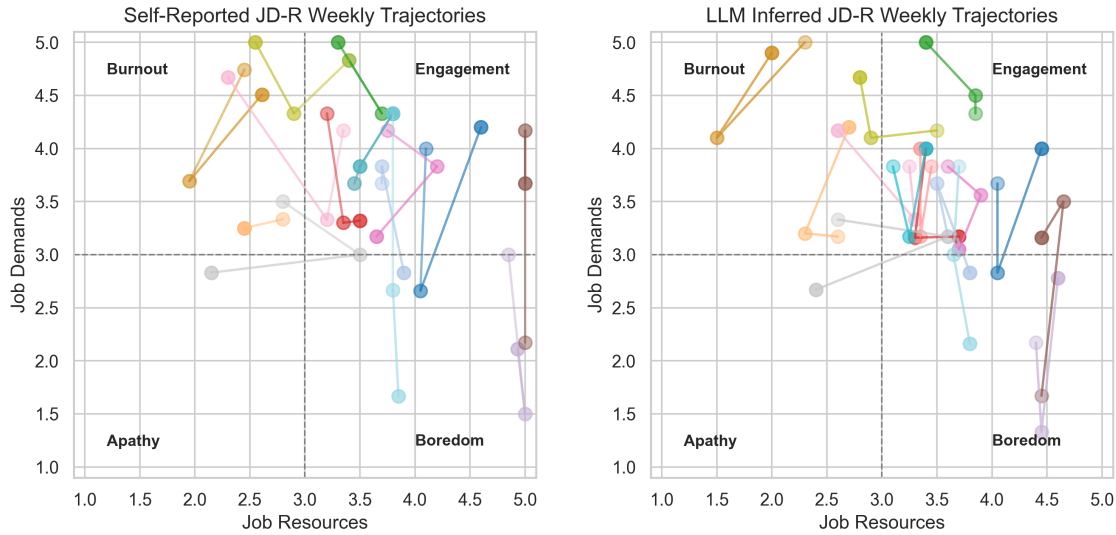


Fig. 5. Weekly JD-R Trajectories Across Participants. The left panel shows self-reported JD-R scores (Job Demands vs. Job Resources) across four weeks, while the right panel presents the corresponding LLM-inferred JD-R scores over the last three weeks. Each trajectory line represents an individual participant, with color intensity indicating progression over weeks. Quadrants are based on the Job Demands–Resources (JD-R) model: Burnout (high demand, low resource), Engagement (high demand, high resource), Apathy (low demand, low resource), and Boredom (low demand, high resource).

According to the JD-R framework literature, overall job demand and resource scores can be further interpreted to understand burnout and engagement levels [3]. A common approach is using the midpoint of the scale (e.g., 3 in our case) as a threshold to classify job demand and job resource levels as either high or low [53, 65, 77]. This creates a two-dimensional space with four quadrants: Burnout (high demand, low resource), Engagement (high demand, high resource), Apathy (low demand, low resource), and Boredom (low demand, high resource). These quadrants provide a meaningful lens for understanding different states of academic well-being among PhD students.

Based on this principle, we plot each participant’s weekly trajectory within the four-quadrant JD-R space using self-reported and LLM-inferred scores, shown in Figure 5 (left and right). In the self-reported data, one participant remained in the boredom quadrant throughout the study, two participants were consistently located in the burnout quadrant, and four participants were generally situated in the engagement quadrant. The remaining seven participants transitioned between quadrants: four moved between engagement and boredom, while three shifted between engagement and burnout. These trajectories reflected the varying academic workload and resource availability as the semester approached its conclusion. Notably, the LLM-inferred trajectories replicate these patterns with high fidelity. While minor path differences are observed, they do not alter quadrant classifications. This consistency indicates that our framework captures week-to-week fluctuations in job demands and resources, providing a meaningful approximation of burnout, engagement, and boredom dynamics among PhD students over time.

6.2 Personalized Interventions

By examining each JD-R factor, we can uncover personalized insights that inform targeted interventions. For instance, among participants in the burnout quadrant, identifying which job resources are lacking can guide support strategies. In our data, one participant exhibited consistently modest inferred scores in Supervision and Peer Support, hovering around the midpoint (2.6–3.0). Although stable, these levels may be insufficient to offset high job demands. Such patterns suggest the need for interventions like strengthened advisor engagement, more frequent check-ins, increased access to peer mentoring, or greater participation in collaborative departmental activities.

For the participant in the boredom quadrant, we observed Perceived Stress scores were notably low across Weeks 3 to 4, declining from 3.0 to 1.5, and rising slightly to 2.11, all well below the midpoint. While low stress might seem beneficial, persistent underload can lead to disengagement. In such cases, thoughtfully increasing job demand through meaningful academic challenges, structured goal-setting, or involvement in stimulating research activities may help re-engage the student, enhancing both motivation and performance. These personalized trajectories demonstrate how LLM-inferred JD-R scores can support ongoing, adaptive interventions tailored to each PhD student.

6.3 Ethical Consideration

Due to the collection and analysis of occupational data such as emails, chats, and calendars, we acknowledge that there may be ethical concerns related to workplace surveillance, particularly the potential misuse for performance evaluation or disciplinary actions. To mitigate risks, strict safeguards should be implemented. For example, supervisors would only receive high-level aggregated trends or suggestion, with individual-level data obfuscated or anonymized. This could prevent micromanagement or punitive use of personal data.

Importantly, our approach can also function as a stand-alone personal assistant. In this mode, all analysis remains private to the individual, empowering students to reflect on their own well-being and adopt strategies to improve work-life balance, without sharing sensitive information with institutional actors. Ultimately, we advocate for responsible deployment guided by transparency, informed consent, and a clear separation between support and evaluation functions.

7 LIMITATION AND FUTURE WORK

Sample Size and Study Period. Our study involved 14 PhD students over four weeks, providing preliminary validation of our LLM-based framework but limiting generalizability. Week-level inspection revealed that LLM-inferred trajectories can be sensitive to contextual availability: in the final week, fewer academic events or sparse interpersonal references led the model to underestimate intellectual climate and peer support, even when self-reports suggested otherwise. Addressing this limitation will require extended study periods and uncertainty-aware methods to mitigate content scarcity. Moving forward, we plan to scale to a broader cohort of PhD students and extend the study to a semester or academic year to better capture longitudinal patterns in PhD student well-being. However, note that similarly limited sample sizes have been used in recent studies on LLM-based applications, including ExploreLLM with 8 participants [40], GPT-4-generated psychological reports with 10 participants [30], and VeriPlan with 12 participants [34]. These precedents highlight the recognized value of small studies in uncovering early insights into emerging technologies.

Input Modality. Our current implementation focuses on a set of mobile sensor data (GPS, step count, and sleep) and occupational data (calendar events, emails, and chat messages). Although these modalities offer useful behavioral and contextual signals, incorporating additional data sources such as heart rate from wearable devices and transcripts from online meetings may enhance the granularity and sensitivity of well-being inference. These richer data streams could provide deeper insights into physiological stress responses and social interaction quality, further improving the accuracy of inferred JD-R factors.

Fine-tuning LLMs. This study relies on general-purpose LLMs (specifically GPT-5) with structured prompting but without fine-tuning on domain-specific data. While prompting strategies can yield decent results, fine-tuning LLMs on academic well-being data or incorporating task-specific adapters may further improve accuracy and personalization. Future work will explore hybrid architectures that combine fine-tuning with prompt engineering for multimodal well-being modeling.

JD-R Factors. Our framework currently focuses on five core JD-R factors (Perceived Stress, Peer Support, Supervision, Infrastructure, and Intellectual Climate) based on prior work [24]. However, our job demand representation is limited to a single factor (Perceived Stress) while it is multifaceted. In the future, we plan to incorporate additional demand dimensions, such as emotional workload scales [66] or cognitive strain scales [25], could enable more personalized and context-aware well-being interventions.

8 CONCLUSION

In this paper, we present an LLM-based framework for inferring PhD students' Job Demands and Resources (JD-R) using multimodal data, including mobile sensor logs and occupational communications such as emails. Through a four-week user study, we demonstrate the feasibility of simulating JD-R responses via a structured two-stage prompting approach. Our experiments further compare different calibration strategies, prompt designs, and model variants. Results show that structured prompting significantly outperforms direct prompting, and both Personal Offset Calibration (POC) and Personal Contextual Calibration (PCC) enhance alignment with self-reported well-being scores. Among the models evaluated, GPT-5 consistently achieves the highest accuracy. We also find that combining mobile sensing and occupational data offers complementary benefits, improving predictive performance compared to single-modality inputs. Importantly, week-by-week trajectory analysis reveals that LLM-inferred JD-R scores capture fluctuations in well-being and map onto theoretical burnout–engagement quadrants. These insights highlight the potential of LLMs for passive, personalized monitoring of occupational mental health in academic environments.

References

- [1] Daniel A. Adler, Vincent W.-S. Tseng, Gengmo Qi, Joseph Scarpa, Srijan Sen, and Tanzeem Choudhury. 2021. Identifying Mobile Sensing Indicators of Stress-Resilience. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 2, Article 51 (June 2021), 32 pages. doi:10.1145/3463528
- [2] AI at Meta. 2025. The llama 4 herd: The beginning of a new era of natively multimodal AI Innovation. <https://ai.meta.com/blog/llama-4-multimodal-intelligence/>
- [3] Arnold B Bakker and Evangelia Demerouti. 2007. The job demands-resources model: State of the art. *Journal of managerial psychology* 22, 3 (2007), 309–328.
- [4] Arnold B Bakker and Karina Mostert. 2024. Study demands–resources theory: Understanding student well-being in higher education. *Educational Psychology Review* 36, 3 (2024), 92.
- [5] A. B. Bakker, A. I. S. Vergel, and J. Kuntze. 2014. Student engagement and performance: a weekly diary study on the role of openness. *Motivation and Emotion* 39 (2014), 49–62. Issue 1. doi:10.1007/s11031-014-9422-5
- [6] Diego Bellini, Barbara Barbieri, Massimiliano Barattucci, Maria Lidia Mascia, and Tiziana Ramaci. 2022. The role of a restorative resource in the academic context in improving intrinsic and extrinsic motivation and flow within the job demands-resources model. *Int. J. Environ. Res. Public Health* 19, 22 (Nov. 2022), 15263.
- [7] Bruno Bizzozero-Peroni, Valentina Díaz-Goñi, Estela Jiménez-López, Eva Rodríguez-Gutiérrez, Irene Sequí-Domínguez, Sergio Núñez de Arenas-Arroyo, José Francisco López-Gil, Vicente Martínez-Vizcaino, and Arthur Eumann Mesas. 2024. Daily Step Count and Depression in Adults: A Systematic Review and Meta-Analysis. *JAMA Network Open* 7, 12 (2024), e2451208–e2451208.
- [8] Rebecca Brauchli, Gregor J Jenny, Désirée Füllemann, and Georg F Bauer. 2015. Towards a Job Demands-Resources health model: empirical testing with generalizable indicators of job demands, job resources, and comprehensive health outcomes. *BioMed research international* 2015, 1 (2015), 959621.
- [9] Michelle L Byrne, Monika N Lind, Sarah R Horn, Kathryn L Mills, Benjamin W Nelson, Melissa L Barnes, George M Slavich, and Nicholas B Allen. 2021. Using mobile sensing data to assess stress: Associations with perceived and lifetime stress, mental health, sleep, and inflammation. *DIGITAL HEALTH* 7 (2021), 20552076211037227. doi:10.1177/20552076211037227 PMID: 34777852.
- [10] Chen Chen, Tong Liu, and Yiwen Chen. 2018. A study of career adaptability and work engagement of online teachers in the E-education industry: the mechanism of job insecurity and employability. In *Proceedings of the 6th International Conference on Information and Education Technology* (Osaka, Japan) (ICIET '18). Association for Computing Machinery, New York, NY, USA, 190–195. doi:10.1145/3178158.3178173
- [11] Victor P Cornet and Richard J Holden. 2018. Systematic review of smartphone-based passive sensing for health and wellbeing. *Journal of biomedical informatics* 77 (2018), 120–132.
- [12] Alex W DaSilva, Jeremy F Huckins, Rui Wang, Weichen Wang, Dylan D Wagner, and Andrew T Campbell. 2019. Correlates of stress in the college environment uncovered by the application of penalized generalized estimating equations to mobile sensing data. *JMIR mHealth and uHealth* 7, 3 (2019), e12084.
- [13] Evangelia Demerouti, Arnold B Bakker, Friedhelm Nachreiner, and Wilmar B Schaufeli. 2001. The job demands-resources model of burnout. *Journal of Applied psychology* 86, 3 (2001), 499.
- [14] Poorvish Dongre, Majid Behravan, Kunal Gupta, Mark Billingham, and Denis Gračanin. 2024. Integrating Physiological Data with Large Language Models for Empathic Human-AI Interaction. *arXiv e-prints*, Article arXiv:2404.15351 (April 2024), arXiv:2404.15351 pages. doi:10.48550/arXiv.2404.15351 arXiv:2404.15351 [eess.SP]
- [15] Zachary Enghardt, Chengqian Ma, Margaret E. Morris, Chun-Cheng Chang, Xuhai "Orson" Xu, Lianhui Qin, Daniel McDuff, Xin Liu, Shwetak Patel, and Vikram Iyer. 2024. From Classification to Clinical Insights: Towards Analyzing and Reasoning About Mobile and Behavioral Health Data With Large Language Models. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 2, Article 56 (May 2024), 25 pages. doi:10.1145/3659604
- [16] Alyssa Eversmeyer. 2024. *Social and Organizational Predictors of Burnout Among Health Service Psychology Doctoral Students: An Application of the Job Demands-Resources Model*. Ph.D. Dissertation. The University of Akron.
- [17] Joseph Firth, Marco Solmi, Robyn E Wootton, Davy Vancampfort, Felipe B Schuch, Erin Hoare, Simon Gilbody, John Torous, Scott B Teasdale, Sarah E Jackson, et al. 2020. A meta-review of "lifestyle psychiatry": the role of exercise, smoking, diet and sleep in the prevention and treatment of mental disorders. *World psychiatry* 19, 3 (2020), 360–380.
- [18] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR Ment Health* 4, 2 (06 Jun 2017), e19. doi:10.2196/mental.7785
- [19] Nan Gao, Zhuolei Yu, Yue Xu, Chun Yu, Yuntao Wang, Flora D. Salim, and Yuanchun Shi. 2024. Leveraging Large Language Models for Generating Mobile Sensing Strategies in Human Behavior Modeling. In *Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Melbourne VIC, Australia) (UbiComp '24). Association for Computing Machinery, New York, NY, USA, 729–735. doi:10.1145/3675094.3678423
- [20] Google. 2025. Places API - Nearby Search. <https://developers.google.com/maps/documentation/places/web-service/nearby-search>.
- [21] Google. 2025. Reverse Geocoding API. <https://developers.google.com/maps/documentation/geocoding/requests-reverse-geocoding>.
- [22] Google Cloud. 2021. Google Workspace for Education Data Protection Implementation Guide. https://services.google.com/fh/files/misc/google_workspace_edu_data_protection_implementation_guide.pdf

- [23] Zhijun Guo, Alvina Lai, Johan H Thygesen, Joseph Farrington, Thomas Keen, Kezhi Li, et al. 2024. Large language models for mental health applications: Systematic review. *JMIR mental health* 11, 1 (2024), e57400.
- [24] Jiying Han and Tongyan Wang and. 2025. Exploring graduate students' research characteristics, emotional exhaustion, mastery approach, and research career commitment: insights from the JD-R theory. *Studies in Higher Education* 50, 2 (2025), 306–320. doi:10.1080/03075079.2024.2336117
- [25] Sandra G Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. Sage publications Sage CA: Los Angeles, CA, 904–908.
- [26] Esther Howe, Jina Suh, Mehrab Bin Morshed, Daniel McDuff, Kael Rowan, Javier Hernandez, Marah Ihab Abdin, Gonzalo Ramos, Tracy Tran, and Mary P Czerwinski. 2022. Design of Digital Workplace Stress-Reduction Intervention Systems: Effects of Intervention Type and Timing. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 327, 16 pages. doi:10.1145/3491102.3502027
- [27] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024).
- [28] Becky Inkster, Shubhankar Sarda, and Vinod Subramanian. 2018. An Empathy-Driven, Conversational Artificial Intelligence Agent (Wysa) for Digital Mental Well-Being: Real-World Data Evaluation Mixed-Methods Study. *JMIR Mhealth Uhealth* 6, 11 (23 Nov 2018), e12106. doi:10.2196/12106
- [29] Sayash Kapoor, Matthew Sun, Mona Wang, Klaudia Jazwinska, and Elizabeth Anne Watkins. 2022. Weaving Privacy and Power: On the Privacy Practices of Labor Organizers in the U.S. Technology Industry. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 473 (Nov. 2022), 33 pages. doi:10.1145/3555574
- [30] Namwoo Kim, Jiseon Lee, Sung Hyeon Park, Yoonseo On, Jieun Lee, Musung Keum, Sanghoon Oh, Yoojin Song, Junhee Lee, Geun Hui Won, Joon Sung Shin, Silvia Kyungjin Lho, Yoon Jung Hwang, and Tae-Suk Kim. 2025. GPT-4 generated psychological reports in psychodynamic perspective: a pilot study on quality, risk of hallucination and client satisfaction. *Front. Psychiatry* 16 (March 2025), 1473614.
- [31] Yubin Kim, Xuhai Xu, Daniel McDuff, Cynthia Breazeal, and Hae Won Park. 2024. Health-llm: Large language models for health prediction via wearable sensor data. *arXiv preprint arXiv:2401.06866* (2024).
- [32] K. Kulikowski, A. Potoczek, E. Antipow, and S. Król. 2019. How to survive in academia: demands, resources and study satisfaction among polish phd students. *Educational Sciences Theory Practice* 19 (2019), 65–79. Issue 4. doi:10.12738/estp.2019.4.005
- [33] Bishal Lamichhane, Nidal Moukaddam, and Ashutosh Sabharwal. 2024. Mobile sensing-based depression severity assessment in participants with heterogeneous mental health conditions. *Scientific Reports* 14, 1 (2024), 18808.
- [34] Christine P. Lee, David Porfrio, Xinyu Jessica Wang, Kevin Chenkai Zhao, and Bilge Mutlu. 2025. VeriPlan: Integrating Formal Verification and LLMs into End-User Planning. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (CHI '25). Association for Computing Machinery, New York, NY, USA, Article 247, 19 pages. doi:10.1145/3706598.3714113
- [35] Tino Lesener, Leonard Santiago Pleiss, Burkhard Gusy, and Christine Wolter. 2020. The study demands-resources framework: An empirical introduction. *International Journal of Environmental Research and Public Health* 17, 14 (2020), 5183.
- [36] Katia Levecque, Frederik Anseel, Alain De Beuckelaer, Johan Van der Heyden, and Lydia Gisle. 2017. Work organization and mental health problems in PhD students. *Research policy* 46, 4 (2017), 868–879.
- [37] Xing Li and Yiwen Chen. 2021. The Impact of RD Employees' Income on Work Engagement in High-tech Industries: Based on A Moderated Mediation Model. In *Proceedings of the 2021 12th International Conference on E-Business, Management and Economics* (Beijing, China) (ICEME '21). Association for Computing Machinery, New York, NY, USA, 578–584. doi:10.1145/3481127.3481185
- [38] Robert LiKamWa, Yunxin Liu, Nicholas D. Lane, and Lin Zhong. 2013. MoodScope: building a mood sensor from smartphone usage patterns. In *Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services* (Taipei, Taiwan) (MobiSys '13). Association for Computing Machinery, New York, NY, USA, 389–402. doi:10.1145/2462456.2464449
- [39] Xin Liu, Daniel McDuff, Geza Kovacs, Isaac Galatzer-Levy, Jacob Sunshine, Jiening Zhan, Ming-Zher Poh, Shun Liao, Paolo Di Achille, and Shwetak Patel. 2023. Large language models are few-shot health learners. *arXiv preprint arXiv:2305.15525* (2023).
- [40] Xiao Ma, Swaroop Mishra, Ariel Liu, Sophie Ying Su, Jilin Chen, Chinmay Kulkarni, Heng-Tze Cheng, Quoc Le, and Ed Chi. 2024. Beyond ChatBots: ExploreLLM for Structured Thoughts and Personalized Model Responses. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI EA '24). Association for Computing Machinery, New York, NY, USA, Article 56, 12 pages. doi:10.1145/3613905.3651093
- [41] Gloria Mark, Shamsi T. Iqbal, Mary Czerwinski, Paul Johns, Akane Sano, and Yuliya Lutchyn. 2016. Email Duration, Batching and Self-interruption: Patterns of Email Use on Productivity and Stress. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) (CHI '16). Association for Computing Machinery, New York, NY, USA, 1717–1728. doi:10.1145/2858036.2858262
- [42] Greta Mazzetti, Enrique Robledo, Michela Vignoli, Gabriela Topa, Dina Guglielmi, and Wilmar B Schaufeli. 2023. Work engagement: A meta-analysis using the job demands-resources model. *Psychological reports* 126, 3 (2023), 1069–1107.
- [43] Shayan Mirjafari, Kizito Masaba, Ted Grover, Weichen Wang, Pino Audia, Andrew T. Campbell, Nitesh V. Chawla, Vedant Das Swain, Munmun De Choudhury, Anind K. Dey, Sidney K. D'Mello, Ge Gao, Julie M. Gregg, Krithika Jagannath, Kaifeng Jiang, Suwen Lin, Qiang Liu, Gloria Mark, Gonzalo J. Martinez, Stephen M. Mattingly, Edward Moskal, Raghu Mulukutla, Subigya Nepal, Kari Nies, Manikanta D. Reddy, Pablo Robles-Granda, Koustuv Saha, Anusha Sirigiri, and Aaron Striegel. 2019. Differentiating Higher and Lower Job Performers in the Workplace Using Mobile Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 2, Article 37 (June 2019), 24 pages. doi:10.1145/3328908

- [44] Payal Mohapatra, Vasudev Aravind, Marisa Bisram, Young-Joong Lee, Hyoyoung Jeong, Katherine Jenkins, Richard Gardner, Jill Streamer, Brent Bowers, Lora Cavuoto, Anthony Banks, Shuai Xu, John Rogers, Jian Cao, Qi Zhu, and Ping Guo. 2024. Wearable network for multilevel physical fatigue prediction in manufacturing workers. *PNAS Nexus* 3, 10 (10 2024), pgae421. doi:10.1093/pnasnexus/pgae421
- [45] Mehrab Bin Morshed, Javier Hernandez, Daniel McDuff, Jina Suh, Esther Howe, Kael Rowan, Marah Abidin, Gonzalo Ramos, Tracy Tran, and Mary Czerwinski. 2022. Advancing the understanding and measurement of workplace stress in remote information workers from passive sensors and behavioral data. In *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 1–8.
- [46] Mehrab Bin Morshed, Koustuv Saha, Richard Li, Sidney K. D’Mello, Munmun De Choudhury, Gregory D. Abowd, and Thomas Plötz. 2019. Prediction of Mood Instability with Passive Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 3, Article 75 (Sept. 2019), 21 pages. doi:10.1145/3351233
- [47] Subigya Nepal, Arvind Pillai, William Campbell, Talie Massachi, Michael V. Heinz, Ashmita Kunwar, Eunsol Soul Choi, Xuhai Xu, Joanna Kuc, Jeremy F. Huckins, Jason Holden, Sarah M. Preum, Colin Depp, Nicholas Jacobson, Mary P. Czerwinski, Eric Granholm, and Andrew T. Campbell. 2024. MindScape Study: Integrating LLM and Behavioral Sensing for Personalized AI-Driven Journaling Experiences. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 4, Article 186 (Nov. 2024), 44 pages. doi:10.1145/3699761
- [48] Subigya Kumar Nepal, Javier Hernandez, Judith Amores Fernandez, Mehrab Bin Morshed, Robert Lewis, Hemma Prafullchandra, and Mary P Czerwinski. 2023. Workplace Rhythm Variability and Emotional Distress in Information Workers. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (*CHI EA ’23*). Association for Computing Machinery, New York, NY, USA, Article 321, 8 pages. doi:10.1145/3544549.3585626
- [49] Sameer Neupane, Poorvash Dongre, Denis Gracanic, and Santosh Kumar. 2025. Wearable Meets LLM for Stress Management: A Duoethnographic Study Integrating Wearable-Triggered Stressors and LLM Chatbots for Personalized Interventions. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA ’25)*. Association for Computing Machinery, New York, NY, USA, Article 588, 8 pages. doi:10.1145/3706599.3720197
- [50] David B Olawade, Ojima Z Wada, Aderonke Odetayo, Aanuoluwapo Clement David-Olawade, Fiyinfoluwa Asaolu, and Judith Eberhardt. 2024. Enhancing mental health with Artificial Intelligence: Current trends and future prospects. *Journal of medicine, surgery, and public health* (2024), 100099.
- [51] OpenAI. 2025. ChatGPT Models. <https://platform.openai.com/docs/models>
- [52] OpenAI. 2025. Introducing GPT-5. <https://openai.com/index/introducing-gpt-5/>
- [53] Li-Chung Pien, Wan-Ju Cheng, Kuei-Ru Chou, and Li-Chiu Lin. 2021. Effect of work–family conflict, psychological job demand, and job control on the health status of nurses. *International journal of environmental research and public health* 18, 7 (2021), 3540.
- [54] Vinola Poliah and Saeeda Paruk. 2017. Depression, anxiety symptoms and substance use amongst sex workers attending a non-governmental organisation in KwaZulu-Natal, South Africa. *South African Family Practice* 59, 3 (2017), 116–122.
- [55] Shaoting Qi and Feng Li. 2022. Work Stress and Mental Health: A Moderated-Mediation Model. In *Proceedings of the 2022 13th International Conference on E-Business, Management and Economics* (Beijing, China) (*ICEME ’22*). Association for Computing Machinery, New York, NY, USA, 529–532. doi:10.1145/3556089.3556177
- [56] Sohrab Saeb, Mi Zhang, Christopher J Karr, Stephen M Schueller, Marya E Corden, Konrad P Kording, David C Mohr, et al. 2015. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *Journal of medical Internet research* 17, 7 (2015), e4273.
- [57] Akane Sano, Andrew J Phillips, Z Yu Amy, Andrew W McHill, Sara Taylor, Natasha Jaques, Charles A Czeisler, Elizabeth B Klerman, and Rosalind W Picard. 2015. Recognizing academic performance, sleep quality, stress level, and mental health using personality traits, wearable sensors and mobile phones. In *2015 IEEE 12th international conference on wearable and implantable body sensor networks (BSN)*. IEEE, 1–6.
- [58] Florian Schaule, Jan Ole Johanssen, Bernd Bruegge, and Vivian Loftness. 2018. Employing Consumer Wearables to Detect Office Workers’ Cognitive Load for Interruption Management. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 1, Article 32 (March 2018), 20 pages. doi:10.1145/3191764
- [59] Felipe B Schuch, Brendon Stubbs, Jacob Meyer, Andreas Heissel, Philipp Zech, Davy Vancampfort, Simon Rosenbaum, Jeroen Deenik, Joseph Firth, Philip B Ward, et al. 2019. Physical activity protects from incident anxiety: A meta-analysis of prospective cohort studies. *Depression and anxiety* 36, 9 (2019), 846–858.
- [60] Felipe B Schuch, Davy Vancampfort, Joseph Firth, Simon Rosenbaum, Philip B Ward, Edson S Silva, Mats Hallgren, Antonio Ponce De Leon, Andrea L Dunn, Andrea C Deslandes, et al. 2018. Physical activity and incident depression: a meta-analysis of prospective cohort studies. *American Journal of Psychiatry* 175, 7 (2018), 631–648.
- [61] Jaemin Shin, Hyungjun Yoon, Seungjoo Lee, Sungjoon Park, Yunxin Liu, Jinho Choi, and Sung-Ju Lee. 2023. FedTherapist: Mental Health Monitoring with User-Generated Linguistic Expressions on Smartphones via Federated Learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 11971–11988. doi:10.18653/v1/2023.emnlp-main.734
- [62] Stanford University. 2024. 6.1.1 Privacy and Access to Electronic Information. <https://adminguide.stanford.edu/chapters/computing/privacy-and-access-electronic-information/privacy-and-access-electronic>
- [63] Anna Sverdluk and Nathan C Hall. 2020. Not just a phase: Exploring the role of program stage on well-being and motivation in doctoral students. *Journal of Adult and Continuing Education* 26, 1 (2020), 97–124.
- [64] Vincent W. S. Tseng, Michael Merrill, Franziska Wittleder, Saeed Abdullah, Min Hane Aung, and Tanzeem Choudhury. 2016. Assessing mental health issues on college campuses: preliminary findings from a pilot study. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive*

- and *Ubiquitous Computing: Adjunct* (Heidelberg, Germany) (*UbiComp '16*). Association for Computing Machinery, New York, NY, USA, 1200–1208. doi:10.1145/2968219.2968308
- [65] Henrico van Roekel, Enno F. J. Wigger, Bernard P. Veldkamp, and Arnold B. Bakker. 2024. What is work engagement? A text mining approach using employees' self-narratives. *Applied Psychology* 73, 3 (2024), 1071–1102. doi:10.1111/apps.12501 arXiv:https://iaap-journals.onlinelibrary.wiley.com/doi/pdf/10.1111/apps.12501
 - [66] M.J.P.M. van Veldhoven, J. Prins, P.A. van der Laken, and L. Dijkstra. 2015. *QEEW2.0: 42 short scales for survey research on work, well-being and performance*. SKB.
 - [67] Maria Magdalena Van Zyl-Cillié, Jacoba H. Bührmann, Alwiena J. Blignaut, Derya Demirtas, and Siedine K. Coetzee. 2024. A machine learning model to predict the risk factors causing feelings of burnout and emotional exhaustion amongst nursing staff in South Africa. *BMC Health Services Research* 24, 1 (Dec 2024). doi:10.1186/s12913-024-12184-5
 - [68] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T. Campbell. 2014. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Seattle, Washington) (*UbiComp '14*). Association for Computing Machinery, New York, NY, USA, 3–14. doi:10.1145/2632048.2632054
 - [69] Weichen Wang, Subigya Nepal, Jeremy F. Huckins, Lessley Hernandez, Vlado Vojdanovski, Dante Mack, Jane Plomp, Arvind Pillai, Mikio Obuchi, Alex daSilva, Eilis Murphy, Elin Hedlund, Courtney Rogers, Meghan Meyer, and Andrew Campbell. 2022. First-Gen Lens: Assessing Mental Health of First-Generation Students across Their First Year at College Using Mobile Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2, Article 95 (July 2022), 32 pages. doi:10.1145/3543194
 - [70] Xin Wang, Ting Dang, Vassilis Kostakos, and Hong Jia. 2024. Efficient and personalized mobile health event prediction via small language models. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*. 2353–2358.
 - [71] Hongyi Wen, Michael Sobolev, Rachel Vitale, James Kizer, JP Pollak, Frederick Muench, Deborah Estrin, et al. 2021. mPulse mobile sensing model for passive detection of impulsive behavior: exploratory prediction study. *JMIR Mental Health* 8, 1 (2021), e25019.
 - [72] Xiaoxia Xie, Chienchung Huang, Shannon P Cheung, Yuqing Zhou, and Jingbo Fang. 2021. Job demands and resources, burnout, and psychological distress of social workers in China: moderation effects of gender and age. *Frontiers in Psychology* 12 (2021), 741563.
 - [73] Huatao Xu, Panrong Tong, Mo Li, and Mani Srivastava. 2024. Autolife: Automatic life journaling with smartphones and llms. *arXiv preprint arXiv:2412.15714* (2024).
 - [74] Xuhai Xu, Perna Chikersal, Afsaneh Doryab, Daniella K. Villalba, Janine M. Dutcher, Michael J. Tumminia, Tim Althoff, Sheldon Cohen, Kasey G. Creswell, J. David Creswell, Jennifer Mankoff, and Anind K. Dey. 2019. Leveraging routine behavior and contextually-filtered features for depression detection among college students. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 3, Article 116 (Sept. 2019), 33 pages. doi:10.1145/3351274
 - [75] Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K. Dey, and Dakuo Wang. 2024. Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 1, Article 31 (March 2024), 32 pages. doi:10.1145/3643540
 - [76] Wei Xuan, Meghna Roy Chowdhury, Yi Ding, and Yixue Zhao. 2025. Unlocking Mental Health: Exploring College Students' Well-being through Smartphone Behaviors. *arXiv preprint arXiv:2502.08766* (2025).
 - [77] Jan Fekke Ybema, Anja Koopman, and Maria Peeters. 2020. Working in sheltered employment: A weekly diary study. *International Journal of Stress Management* 27, 2 (2020), 160.
 - [78] Consuela Cheriece Yousef, Ali Farooq, Gigi Amateau, Laila Carolina Abu Esba, Keisha Burnett, and Omar Anwar Alyas. 2024. The effect of job and personal demands and resources on healthcare workers' wellbeing: A cross-sectional study. *Plos one* 19, 5 (2024), e0303769.
 - [79] Tianyi Zhang, Songyan Teng, Hong Jia, and Simon D'Alfonso. 2024. Leveraging LLMs to Predict Affective States via Smartphone Sensor Features. In *Companion of the 2024 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Melbourne VIC, Australia) (*UbiComp '24*). Association for Computing Machinery, New York, NY, USA, 709–716. doi:10.1145/3675094.3678420
 - [80] Yuezhou Zhang, Amos Folarin, and Richard Dobson. 2024. Exploring Digital Biomarkers for Depression Using Mobile Technology. *International Journal of Population Data Science* 9, 4 (2024). doi:10.23889/ijpds.v9i4.2422

A Weekly JD-R Questionnaire

Job Demands	
<i>Each item was measured on a 5-point Likert scale (1 = never, 5 = very often)</i>	
Perceived Stress	(1) In the last week, how often have you felt nervous and 'stressed'?
	(2) In the last week, how often have you found that you could not cope with all the things that you had to do?
	(3) In the last week, how often have you been angered because of things that happened that were outside of your control?
	(4) In the last week, how often have you found yourself thinking about things that you have to accomplish?
	(5) In the last week, how often have you been unable to control the way you spend your time?
	(6) In the last week, how often have you felt difficulties were piling up so high that you could not overcome them?
Job Resources	
<i>Each item was measured on a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree)</i>	
Peer Support	(1) I am actively in contact with other graduate students.
	(2) I have identified a suitable peer group of fellow graduate students.
	(3) I get sufficient support from other graduate students.
	(4) I actively take part in the activities of the academic community in my field.
	(5) I get sufficient support from the surrounding academic society regarding my studies.
Supervision	(1) Supervision is available when I need it.
	(2) My supervisor(s) provide(s) me with additional information relevant to my topic.
	(3) My supervisor(s) provide(s) helpful feedback on my progress.
	(4) I am given good guidance in topic selection and refinement.
	(5) Overall, I am satisfied with the quality of my supervision.
Infrastructure	(1) I have access to a suitable working space.
	(2) I have good access to the technical support I need.
	(3) I am able to organize good access to necessary equipment.
	(4) I have access to a common room or a similar type of meeting place.
	(5) Overall, I am satisfied with the quality of the services and facilities.
Intellectual Climate	(1) I am provided with opportunities for social contact with other research students.
	(2) I feel integrated into the faculty community.
	(3) I am provided with opportunities to become involved in the broader research culture.
	(4) The research ambience in the faculty stimulates my work.
	(5) Research students are provided with a supportive working environment.

B Information Extraction Prompts

Perceived Stress

1. In the last week, how often have you felt nervous and 'stressed'?

Context	Can you derive the specific situations in the user's textual data (emails, chats, notes) from the past week that
Framing	indicate feeling nervous, anxious, or stressed? Summarize the approximate time frames or triggers, if available.
Pattern	Check whether these stress inducing situations or mentions of stress occur repeatedly or appear linked to
Recognition	recurring events (such as weekly deadlines). Note any patterns in timing or triggers.
Cross-Modality	Identify any direct or derived indication to stress or anxiety in the user's text data and compare those time
Validation	windows (days around it) with sensor data—such as sleep disruptions or changes in step counts. Indicate if there is a consistent correlation (e.g., more stress references coinciding with poor sleep or lower physical activity).

2. In the last week, how often have you found that you could not cope with all the things that you had to do?

Context	From the textual data, find or infer any instances where the user appears overwhelmed, overworked, or unable
Framing	to handle tasks. Specify which deadlines, responsibilities, or time blocks are linked to these statements.
Perceived	Extract how the user describes the severity of not coping (e.g., 'barely managing', 'completely swamped') and
Quality Analysis	highlight any emotional tone (frustration, panic, resignation).
Pattern	Identify or infer each instance in the user's text where they indicate being unable to cope. Check whether these
Recognition	instances recur at similar intervals or revolve around common triggers (e.g., project deadlines). Summarize any patterns or recurring factors that emerge.
Cross-Modality	Compare the user's explicit or inferred references to feeling overwhelmed with sensor data (e.g., late-night
Validation	location logs, reduced step counts). Determine whether these reported or inferred episodes coincide with extended work hours or other observable signs of overload in the data.

3. In the last week, how often have you been angered because of things that happened that were outside of your control?

Context	Look through the user's text data for any explicit or inferred indication that the user was angered by external
Framing	factors (e.g., administrative delays, equipment malfunctions, unexpected changes). If the timing is mentioned or can be inferred, note when these instances may have occurred.
Pattern	For each identified or inferred instance of anger tied to external causes, determine whether similar complaints
Recognition	appear repeatedly (e.g., repeated equipment issues) or if each one is unique. Summarize any recurring triggers or patterns that emerge.
Cross-Modality	For each anger-related event found (explicitly stated or inferred) in the textual data, align the reported or
Validation	approximate time frames with sensor data—such as abrupt changes in location routines (e.g., leaving campus early) or unusual usage/activity patterns. Indicate whether there is any observable correlation between these anger events and the user's behavior.

4. In the last week, how often have you found yourself thinking about things that you have to accomplish?

Context	Describe the events or interactions related to deadlines, goals, or unresolved tasks. Indicate how frequently
Framing	they appear (daily, sporadically).
Perceived	Look for descriptions of the user's emotional perspective regarding those tasks (e.g., feeling motivated, anxious,
Quality Analysis	or overwhelmed). Summarize any language indicating how the user perceives these upcoming responsibilities.
Pattern	Assess if the user repeatedly mentions the same tasks or if each mention concerns different goals. Note if the
Recognition	user cycles back to the same uncompleted tasks on multiple occasions.
Cross-Modality	Identify or infer any tasks or goals mentioned in the user's text. Then check the sensor data (e.g., late-night lab
Validation	presence, altered sleep patterns) to see if behavioral changes coincide with periods where the user appears to experience heightened rumination or worry related to those tasks.

5. In the last week, how often have you been unable to control the way you spend your time?

<i>Context Framing</i>	Find or infer from the user's text data any references to forced or unplanned schedule changes (e.g., last-minute meeting requests, personal emergencies). Summarize how frequently these changes appear and in what contexts they arise.
<i>Pattern Recognition</i>	For each mention or inference of schedule disruptions in the text, determine whether the same cause (e.g., advisor demands, personal errands) recurs or if the disruptions vary. Summarize any repeated factors that emerge.
<i>Cross-Modality Validation</i>	Compare the reported or inferred schedule disruptions with the user's calendar events and location logs for overlapping or double-booked time slots. Indicate whether these data confirm the user's stated difficulty in controlling their schedule.

6. In the last week, how often have you felt difficulties were piling up so high that you could not overcome them?

<i>Context Framing</i>	Search the user's text data for any references indicating multiple concurrent problems (e.g., deadlines in different classes, personal responsibilities). Specify whether these 'convergence points' of challenges happen frequently or just once.
<i>Pattern Recognition</i>	Check whether the user references similar pile-ups in the past or if this is the first time. Summarize any repeated occurrence of feeling overwhelmed by multiple challenges.
<i>Cross-Modality Validation</i>	Extract any self-description of severity (like 'felt impossible to catch up') and see if the user's sensor data (reduced campus visits, fewer steps) suggests behavior changes consistent with overwhelm. Extract any self-description of severity (like 'felt impossible to catch up') and see if the user's sensor data (reduced campus visits, fewer steps) suggests behavior changes consistent with overwhelm.

Peer Support

1. I am actively in contact with other graduate students.

<i>Context Framing</i>	Gather textual references (emails, chat logs) indicating the user's interactions with graduate peers, whether academic or social. Summarize how often these are mentioned, if time frames are noted, and the topics (research vs. casual).
<i>Pattern Recognition</i>	Assess if the user repeatedly contacts the same peers or meets new people each time. Note any recurring patterns, such as weekly group chats or frequent lab partner interactions.
<i>Perceived Quality Analysis</i>	Extract the user's interaction with his peers and assess if the interaction is helpful. Highlight any language suggesting satisfaction/dissatisfaction.

2. I have identified a suitable peer group of fellow graduate students.

<i>Context Framing</i>	Identify discussions referring to a specific group or circle of fellow students (e.g., lab partners, study buddies). Summarize who they are and how the user describes them.
<i>Perceived Quality Analysis</i>	Extract the user's sentiments about this peer group (shared interests, supportive atmosphere). List any positive or negative adjectives used to describe them.

3. I get sufficient support from other graduate students.

<i>Context Framing</i>	Locate all mentions of receiving help, advice, or emotional backing from peers. Note the types of support (academic vs. personal) and whether it's occasional or routine.
<i>Pattern Recognition</i>	Determine if the user repeatedly references the same individuals or types of support (e.g., one study buddy who always helps) or if it varies. Summarize any recurring patterns of peer assistance.
<i>Cross-Modality Validation</i>	Check location logs to see if the user and their peers meet in person on campus or off-campus. Summarize whether sensor data confirms or contradicts frequent contact.

4. I actively take part in the activities of the academic community in my field.

Context	Collect mentions in text of user attending seminars, workshops, or departmental gatherings. Summarize how often these events appear on the user's calendar or in email invites.
Framing	
Cross-Modality	Identify the user's emotional tone (enthusiastic, indifferent) toward these events, then confirm attendance via
Validation	GPS or calendar data. Note any mismatch between 'planning to attend' and actual presence.
5. I get sufficient support from the surrounding academic society regarding my studies.	
Context	Gather references to broader institutional help (e.g., faculties, departmental resources, student associations, etc).
Framing	Note how often the user is reaching out or receiving assistance.
Cross-Modality	From the references to broader institutional help (e.g., faculties, departmental resources, student associations, etc.) assess the quality of the help (helpful, insufficient, etc.). Consider the timestamps of these interactions to
Validation	determine if the help is timely or not. Summarize any repeated concerns.
Supervision	
1. Supervision is available when I need it.	
Context	Locate all mentions where the user attempted to contact or meet with the supervisor(s). Summarize how often
Framing	these attempts occur.
Perceived	Analyze response times from supervisors in emails and chats.
Quality Analysis	
2. My supervisor(s) provide(s) me with additional information relevant to my topic.	
Context	Identify any references to resources (papers, data sets, reading lists) shared by the supervisor. Note whether the
Framing	user acknowledges them multiple times or rarely.
Perceived	Are the additional resources provided by the supervisor(s) timely and aligned with the user's research needs?
Quality Analysis	
3. My supervisor(s) provide(s) helpful feedback on my progress.	
Context	Extract textual mentions of feedbacks received from the supervisor in chat, emails, comments, or meeting notes.
Framing	Note how frequently the user receives feedback and whether it's detailed or brief.
Perceived	Assess the quality of textual feedbacks received from the supervisor (chat, emails, comments, or meeting notes).
Quality Analysis	
4. I am given good guidance in topic selection and refinement.	
Context	Locate references to topic-related discussions with the supervisor. Determine if these guidance sessions appear
Framing	at multiple stages (initial idea, revision, advanced stages) or just once.
Perceived	Determine if the topic-related discussion sessions appear at multiple stages (initial idea, revision, advanced
Quality Analysis	stages) or just once. What is the depth of the discussion?
Pattern	Check whether these guidance-related interactions recur at multiple stages (initial idea, revision, advanced
Recognition	stages) or just once. Summarize any repeated guidance patterns.
5. Overall, I am satisfied with the quality of my supervision.	
Context	Check whether the user's calendar/email data shows consistent or erratic supervision patterns (e.g., last-minute
Framing	meeting cancellations).
Perceived	From the overall interaction with the user's supervisor summarize the quality.
Quality Analysis	
Infrastructure	

1. I have access to a suitable working space.

Context	Identify mentions of the user's workspace setup (shared office, lab bench), focusing on any complaints (noise, crowding) or praises (convenient, quiet). Summarize frequency.
Framing	
Cross-Modality	Check whether GPS logs confirm regular or inconsistent usage of this working space. Note if usage aligns with the user's stated comfort or dissatisfaction, if any.
Validation	

2. I have good access to the technical support I need.

Context	Locate all references to contacting IT or technical staff. Summarize how often the user mentions system failures, bug fixes, or other technical hurdles.
Framing	
Perceived	Extract the user's impression of how quickly and effectively support is provided (e.g., 'immediate resolution,' 'took a week'). Compare with email threads or ticket resolution timestamps to confirm or refute these claims.
Quality Analysis	

3. I am able to organize good access to necessary equipment.

Context	Identify any text describing the equipment reservation or sign-out processes (shared lab tools, booking schedules).
Framing	Summarize how frequently these issues come up.
Perceived	Determine if the user repeatedly complains about scarcity/delays, or if it's a one-time mention. Then highlight the user's sentiment (frustrated, neutral) about obtaining equipment on time.
Quality Analysis	
Pattern	See if the user repeatedly mentions the same equipment shortages or scheduling conflicts, or if each reference is unique. Summarize recurring obstacles, if any.
Recognition	

4. I have access to a common room or a similar type of meeting place.

Context	Collect mentions of using or attempting to use a group meeting place/common room.
Framing	
Cross-Modality	Check GPS data for repeated visits to the common room location. Compare any expressed ease/difficulty in scheduling or using the space with actual usage frequency.
Validation	

5. Overall, I am satisfied with the quality of the services and facilities.

Context	Gather all broad statements (positive or negative) about institutional facilities (labs, IT services, library).
Framing	Summarize repeated praise or complaint themes.
Perceived	Check if the user's usage patterns (heavy vs. minimal facility usage) match their stated satisfaction level. Note any conflicting evidence, such as frequent logs of facility usage despite strong negative remarks.
Quality Analysis	

Intellectual Climate

1. I am provided with opportunities for social contact with other research students.

Context	Identify references to departmental gatherings, social events, or co-working sessions with other students.
Framing	Summarize how often such opportunities arise.
Pattern	Determine if the user repeatedly attends the same type of gatherings (weekly seminars, monthly socials) or if each mention is different. Summarize any recurring pattern.
Recognition	
Cross-Modality	Compare user statements about attending group events with location or calendar data to confirm actual attendance or scheduling. Indicate if the data supports the user's account.
Validation	

2. I feel integrated into the faculty community.

Context	Assess the user interaction and departmental emails to determine the level of integration the users feels towards the department and faculty community (strong sense of community vs. feeling sidelined).
Framing	

3. I am provided with opportunities to become involved in the broader research culture.

<i>Context</i>	Identify mentions of cross-department or external conferences, special interest groups, or reading clubs. Sum-
<i>Framing</i>	marize how frequently the user encounters or acknowledges such opportunities.
<i>Cross-Modality</i>	Identify mentions of cross-department or external conferences, special interest groups, or reading clubs and
<i>Validation</i>	confirm actual attendance or sign-ups via location logs or email invites.

4. The research ambience in the faculty stimulates my work.

<i>Context</i>	What opportunities are available to engage with the broader research culture? The general research atmosphere
<i>Framing</i>	(e.g., collaborative vibe, active seminars).
<i>Perceived</i>	Extract any commentary on how motivating or uninspiring the user finds this ambience. Indicate whether such
<i>Quality Analysis</i>	remarks appear consistent or vary depending on lab visits, departmental events, etc.

5. Research students are provided with a supportive working environment.

<i>Context</i>	Identify textual references to overall support structures for research students (faculty mentorship, open-door
<i>Framing</i>	policies, resource centers). Summarize how often these come up.

C Additional Examples of LLM Responses to Information Extraction Prompts

Table 6. Example of two JD-R question from the Peer Support scale, branched into information extraction prompts to summarize and contextualize relevant textual and behavioral data

JD-R Question	"I get sufficient support from other graduate students."
Context Framing	<p>Prompt: "Locate all mentions of receiving help, advice, or emotional backing from peers. Note the types of support (academic vs. personal) and whether it's occasional or routine."</p> <p>Response: "During the analyzed week, [STUDENT NAME] received support from peers in both academic and personal contexts across multiple modalities. In chats, [PEER A] offered notes on [EVENT NAME] seminar to help with a draft deadline (academic, routine), while [PEER B] reassured [STUDENT NAME] after a stressful lab meeting on [DATE] using casual, encouraging language and emojis (emotional, occasional). In emails, [PEER C] shared coding resources relevant to [STUDENT NAME]'s [COURSE NAME] project, providing constructive academic advice (occasional). Calendar records showed participation in a "Peer Study Group" event, with GPS confirming attendance, representing structured, routine academic backing. Overall, the week revealed a blend of routine academic support (study groups, lab work, shared resources) and occasional emotional support (peer reassurance in stressful moments)."</p>
Pattern Recognition	<p>Prompt: "Determine if the user repeatedly references the same individuals or types of support (e.g., one study buddy who always helps) or if it varies. Summarize any recurring patterns of peer assistance."</p> <p>Response: "[STUDENT NAME]'s peer assistance shows both consistency and variation. [PEER A] emerges as a recurring figure, appearing multiple times in study-related interactions such as sharing notes and collaborating around seminar and draft preparation, suggesting a stable source of routine academic support. In contrast, other peers like [PEER B] and [PEER C] are referenced less frequently, stepping in to provide emotional reassurance or situational academic resources when needed, which points to occasional, context-dependent assistance. Taken together, the pattern indicates that [STUDENT NAME] relies on one core peer for ongoing academic help, while additional peers contribute in more variable and situational ways, creating a support network anchored by repeated collaboration with a key individual."</p>
Cross Modality Validation	<p>Prompt: "Check location logs to see if the user and their peers meet in person on campus or off-campus. Summarize whether sensor data confirms or contradicts frequent contact."</p> <p>Response: "Based on the location data from [DATES], there are several instances suggesting the user and peers likely meet on campus. The user consistently visits the FSU campus or research labs daily, often from morning until early evening. Events on [DAY3], like the [EVENT NAME], which lists numerous attendees, further supports potential group interaction, although it's an online event. Physical activity levels and sedentary patterns often align with typical group working and meeting behaviors. Off-campus interactions are less frequent. Overall, the mobile sensing data aligns with moderate to regular in-person contact on campus while providing limited confirmation for off-campus meetups."</p>
JD-R Question	"I actively take part in the activities of the academic community in my field."
Context Framing	<p>Prompt: "Collect mentions in text of user attending seminars, workshops, or departmental gatherings. Summarize how often these events appear on the user's calendar or in email invites."</p>

Continued on next page

Table 6 (continued)

	<p>Response: “Based on the data provided, [STUDENT NAME] appears to be fairly active in attending seminars, workshops, and gatherings. The calendar data details that [STUDENT NAME] organized and participated in several meetings and events. Notably, there’s the [EVENT NAME] event on [DATES], and the [EVENT NAME] on [DATES], which indicate involvement in technical or academic work. Moreover, the [EVENT NAME] on [DAY3] included [STUDENT NAME] as an attendee, showing participation in an executive committee setting. The calendar shows frequent events with both professional and academic relevance, while no specific email invitations pertaining directly to seminars or workshops are evident. However, the presence of these events in the schedule suggests regular engagement with professional or academic gatherings. This indicates a proactive approach towards attending and possibly organizing key intellectual events, which occur at least a few times a month based on the schedule data. ”</p>
Cross Modality Validation	<p>Prompt: “Identify the user’s emotional tone (enthusiastic, indifferent) toward these events, then confirm attendance via GPS or calendar data. Note any mismatch between ‘planning to attend’ and actual presence.”</p> <hr/> <p>Response: “Based on the interactions and available data, [STUDENT NAME]’s emotional tone can be described as professional and slightly deferential, particularly in communications with their professor regarding approvals and abstract submissions. There is no strong indication of either enthusiasm or indifference toward the events in the reviewed exchanges, suggesting a neutral but respectful approach. Attendance records further confirm alignment between planning and presence. For the [EVENT NAME] and [EVENT NAME] events on [DATES], both calendar entries and GPS data verify [STUDENT NAME]’s participation at the designated times and locations. Similarly, for the [EVENT NAME] on [DAY3], the calendar shows scheduled attendance, and GPS data corroborates presence during the event. Overall, the available evidence indicates that [STUDENT NAME] attended as planned, with no mismatches detected between intended participation and actual presence. ”</p>

Table 7. Example of one JD-R question from the Supervision scale, branched into information extraction prompts to summarize and contextualize relevant textual and behavioral data

JD-R Question	“Overall, I am satisfied with the quality of my supervision.”
Context Framing	<p>Prompt: “From the overall interaction with the user’s supervisor summarize the quality.”</p> <hr/> <p>Response: “Overall quality of interaction with the supervisor: high.</p> <ul style="list-style-type: none"> - Responsiveness: Very prompt (e.g., <2 minutes to approve EGU registration), enabling quick student action. - Supportiveness and trust: Clear authorization for expenses and autonomy for tasks; student feels comfortable asking and reporting back. - Clarity and tone: Polite, concise, and affirmative (“please go ahead and register”), reducing ambiguity. - Structure and availability: Proactive scheduling (one-on-one after [PEER] meeting; regular group meeting), indicating consistent engagement. - Professionalism: Interactions are respectful and efficient; no signs of conflict or miscommunication. - Impact on well-being/productivity: Likely positive—reduced decision friction, steady progress, regular touch-points. - Not observed: Detailed technical feedback in chat logs (may occur in meetings), so depth of scholarly guidance can’t be fully assessed here. <p>Bottom line: Supportive, responsive, and professional supervisory relationship with good cadence of communication; strong administrative and motivational support evident. ”</p>
Perceived Quality Analysis	<p>Prompt: “Check whether the user’s calendar/email data shows consistent or erratic supervision patterns (e.g., last-minute meeting cancellations).”</p> <hr/> <p>Response: “Based on the [WEEK] data, the supervision patterns appear consistent rather than erratic. [STUDENT NAME], a PhD student, demonstrates regular communication with supervisors and colleagues through planned meetings, as observed in the chat logs and calendar data. Meetings, such as those scheduled for [DAY4] and [DAY5], were organized well in advance, with explicit start and end times, and no indication of cancellations or abrupt changes. The mobile sensing data also suggests regular routines without disruptions on meeting days. Overall, the consistency of scheduled interactions and responses by supervisor suggest stable and predictable supervisory practices, with no signs of last-minute cancellations or erratic behavior.”</p>

Table 8. Example of one JD-R question from the Infrastructure scale, branched into information extraction prompts to summarize and contextualize relevant textual and behavioral data

JD-R Question	“ I have access to a suitable working space.”
Context Framing	<p>Prompt: “Identify mentions of the user’s workspace setup (shared office, lab bench), focusing on any complaints (noise, crowding) or praises (convenient, quiet). Summarize frequency.”</p> <hr/>

Continued on next page

Table 8 (continued)

	<p>Response: “Based on the available records, there are only limited references to [STUDENT NAME]’s workspace setup, and none suggest strong dissatisfaction. Mentions of being in a shared lab or office appear in routine scheduling and communication, but these are largely neutral in tone, without complaints about noise, crowding, or other discomforts. Similarly, there are no explicit praises highlighting convenience, quietness, or other positive qualities of the workspace. The absence of either complaints or compliments indicates that workspace conditions are not a salient issue for [STUDENT NAME] during this week, and references to the setup appear infrequent and purely descriptive rather than evaluative. Overall, the data show neutral and infrequent mentions of workspace setup, with no evidence of notable concerns or appreciations.”</p>
Cross Modality Validation	<p>Prompt: “Check whether GPS logs confirm regular or inconsistent usage of this working space. Note if usage aligns with the user’s stated comfort or dissatisfaction, if any.”</p> <hr/> <p>Response: “The GPS logs indicate regular use of the working space, with consistent patterns observed at the FAMU-FSU College of Engineering campus and lab locations on weekdays. For example, between [DATES], [STUDENT NAME] visited the FSU campus daily from approximately [TIME] AM to [TIME] PM and the lab in the evenings on some days. This suggests a stable routine, possibly indicating a level of comfort with the work environment. There is no explicit mention of dissatisfaction in the chat or email data that could contrast her regular schedule. Physical activity and sleep data seem balanced, also supporting a healthy work-life routine, suggesting that the user is likely comfortable with her current work arrangement. Overall, the data portrays a consistent use of the workspaces that aligns well with maintaining productivity.”</p>

D Participant Demographics

Table 9. Demographic characteristics of participants (N = 14).

Category	Subgroup	N (%)
Age (years)	24–34	14 (100%)
Gender	Male	7 (50.0%)
	Female	7 (50.0%)
Majors	Biological Science	3 (21.4%)
	Civil & Environmental Eng.	3 (21.4%)
	Electrical & Computer Eng.	2 (14.3%)
	Industrial & Manufacturing Eng.	2 (14.3%)
	Earth, Ocean & Atmospheric Sci.	1 (7.1%)
	Psychology	1 (7.1%)
	Biochemistry	1 (7.1%)
	Chemical Engineering	1 (7.1%)
Seniority	First Year	3 (21.4%)
	Second Year	6 (42.9%)
	Third Year	4 (28.6%)
	Fourth Year	1 (7.1%)

E Processed Data Format

```

1 {
2   "timezone": "America/New_York",
3   "participant_id": "[P###]",
4   "week_start": "YYYY-MM-DD",
5   "week_end": "YYYY-MM-DD",
6   "daily_records": {
7     "YYYY-MM-DD": {
8       "location_patterns": [
9         { "time_window": "HH:MM-HH:MM", "place": "Home" },
10        { "time_window": "HH:MM-HH:MM", "place": "Research Building" },
11        { "time_window": "HH:MM-HH:MM", "place": "Classroom Building" },
12        { "time_window": "HH:MM-HH:MM", "place": "Home" }
13      ],
14      "sleep": {
15        "bedtime": "HH:MM",           // converted to local time
16        "wake_time": "HH:MM",       // converted to local time
17        "quality": "Good|Fair|Poor|Excellent"
18      },
19      "activity": {
20        "steps": <int>,
21        "activity_level": "Low|Moderate|High",
22        "sedentary_minutes": <int>
23      }
24    }
25    /* ... one object per day of the week ... */
26  }
27 }

```

Listing 1. Weekly mobile sensing data format

```

1 {
2   "timezone": "America/New_York",
3   "participant_id": "[P###]",
4   "week_start": "YYYY-MM-DD",
5   "week_end": "YYYY-MM-DD",
6   "daily_records": {
7     "YYYY-MM-DD": [
8       {
9         "timestamp": "YYYY-MM-DDTHH:MM:SS",
10        "from": "[Advisor]|[Peer]|[Department]",
11        "to": "[Student]",
12        "subject": "<subject>",
13        "content": "<mail body>",
14      },
15      {},
16      {},
17      {}
18    ],
19    "YYYY-MM-DD": [],
20    "YYYY-MM-DD": [],
21    "YYYY-MM-DD": [],
22    "YYYY-MM-DD": [],
23    "YYYY-MM-DD": [],
24    "YYYY-MM-DD": []

```

```

25 },
26 }

```

Listing 2. Weekly Email data format

```

1 {
2   "timezone": "America/New_York",
3   "participant_id": "[P###]",
4   "week_start": "YYYY-MM-DD",
5   "week_end": "YYYY-MM-DD",
6   "daily_records": {
7     "YYYY-MM-DD": [
8       {
9         "timestamp": "YYYY-MM-DDTHH:MM:SS",
10        "sender": "[Peer A]",
11        "recipient": "[Student]",
12        "message": "<text>",
13      },
14      {},
15      {},
16      {}
17    ],
18    "YYYY-MM-DD": [],
19    "YYYY-MM-DD": [],
20    "YYYY-MM-DD": [],
21    "YYYY-MM-DD": [],
22    "YYYY-MM-DD": [],
23    "YYYY-MM-DD": []
24  },
25 }

```

Listing 3. Weekly Chat data format

```

1 {
2   "timezone": "America/New_York",
3   "participant_id": "[P###]",
4   "week_start": "YYYY-MM-DD",
5   "week_end": "YYYY-MM-DD",
6   "daily_records": {
7     "YYYY-MM-DD": [
8       {
9         "title": "<Event Title>",
10        "start": "YYYY-MM-DDTHH:MM:SS",
11        "end": "YYYY-MM-DDTHH:MM:SS",
12        "location": "[Address|Online]",
13        "attendees": ["[Student]", "[Advisor]", "[Peer A]"],
14        "is_recurring": false,
15        "attendance_alignment": {
16          "calendar_entry": true,
17          "notes": "<brief justification>"
18        }
19      },
20      {},
21      {},
22      {}

```

```
23     ],  
24     "YYYY-MM-DD": [],  
25     "YYYY-MM-DD": [],  
26     "YYYY-MM-DD": [],  
27     "YYYY-MM-DD": [],  
28     "YYYY-MM-DD": [],  
29     "YYYY-MM-DD": []  
30   },  
31 }
```

Listing 4. Weekly Calendar data format

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009