

A Multi-Stage, RAG-Enhanced Pipeline for Generating Longitudinal, Clinically Actionable Mental Health Reports from Wearable Sensor Data

Ugonna Oleh¹[0009-0009-2169-3102], Roman Obermaisser¹[0009-0002-4483-1503], Alla Malchulska¹[0000-0001-5968-429X] and Tim Klucken¹[0000-0003-2433-6652]

¹University of Siegen, Siegen, Germany
*ugonna.oleh@uni-siegen.de

This work builds a full automated pipeline that turns **raw wearable sensor data** into **useful mental health insights**.

* First, a **deep learning HAR model** recognizes daily activities from sensor data.
* Then, a **temporal bout analysis** looks at how long and how often physiological/behavioural events happen over time.
* Finally, a **RAG-enhanced LLM** uses an external **clinical knowledge base** to generate **clinically grounded** text reports

Abstract. Consumer-grade wearable devices generate vast streams of high-frequency physiological and behavioural data, yet this raw information is not directly interpretable or actionable within clinical mental healthcare workflows. This creates a significant gap between the potential of remote monitoring technology and its practical application. To address this challenge, we present a novel, multi-stage automated pipeline that transforms raw sensor data into longitudinal, clinically relevant insights. The pipeline integrates a deep learning model for Human Activity Recognition (HAR), a temporal "bout" analysis to contextualise physiological and behavioural events, and a Retrieval-Augmented Generation (RAG) enhanced Large Language Model (LLM) to ensure outputs are grounded in an external clinical knowledge base. We demonstrate the pipeline's efficacy through a case study of a volunteer with panic disorder. The system successfully performed a week-over-week comparative analysis, identifying a nuanced shift from frequent, short bouts of physiological arousal to fewer, but more sustained and intense, episodes. From this analysis, the pipeline generated three distinct, audience-appropriate reports: a data-driven summary with treatment considerations for a psychologist, an empathetic summary for the volunteer, and a comprehensive analytical report. This work presents a viable end-to-end system for translating complex, continuous sensor data into actionable insights that can support and enhance mental healthcare.

In a panic disorder case study, the system detected a shift from **many short arousal episodes** to **fewer but longer and stronger episodes** across weeks. It then created three tailored reports:

1. A clinical, data-focused summary with treatment hints for a psychologist,
2. An empathetic, easy report for the patient,
3. A detailed analytical report.

Overall, it shows a practical end-to-end way to turn continuous sensor data into **actionable insights for mental healthcare**.

Keywords: Mental Health Monitoring, RAG, HAR, LLM, Wearable Sensors.

1 Introduction

For decades, the evaluation of mental health has relied heavily on subjective self-reports and clinician observations, methods that are inherently limited by recall bias, social desirability effects, and the snapshot nature of in-clinic visits. The advent of consumer-grade wearable devices, such as smartwatches, presents a paradigm shift, offering the potential to capture objective, high-frequency physiological (e.g., heart rate from photoplethysmography, PPG) and behavioural (e.g., motion from Inertial Measurement Units, IMU) data as individuals navigate their daily lives [1]. The practice of mental healthcare is undergoing a profound transformation, driven by the proliferation

Traditional mental health assessment depends on self-reports and clinic visits, which are limited by memory errors and short, rare snapshots of a person's life.

Now, wearable devices (like smartwatches with heart-rate and motion sensors) allow continuous, real-world monitoring of physiology and behaviour, and patients generally accept and like these devices.

In mental health, Human Activity Recognition (HAR) can automatically detect subtle anxiety-related behaviours (e.g., fidgeting, nail-biting, hand tremors) from wearable sensor data using labeled datasets like ADAM-sense. This idea fits into the larger concept of digital phenotyping, where continuous digital data is used to quantify a person's behaviour and state.

However, there is a big interpretation gap:

The raw data is huge, noisy, and multi-modal.

Clinicians cannot directly use streams of numbers from sensors in their normal workflows.

The problem is not hardware or patient willingness, but the lack of a robust analytic layer that turns raw data into clear, clinic

2

of wearable and mobile technologies that are shifting the locus of assessment from episodic, clinic-based encounters to continuous, real-world monitoring [2]. A critical factor accelerating the adoption of these technologies is their high degree of acceptance among patients [3].

In the domain of mental health, HAR is increasingly being applied to detect non-verbal, psychomotor behaviours that are known clinical indicators of anxiety, stress, and agitation [4]. Anxiety disorders often manifest through physical symptoms, including restlessness and repetitive, non-goal-oriented movements (e.g., fidgeting) [5]. HAR systems can be trained to identify these subtle yet significant behavioural markers automatically [6]. Research in this area has focused on recognising a specific vocabulary of anxiety-related actions, such as nail-biting, hair-pulling, knuckle-cracking, hand-tapping, and other forms of fidgeting [7]. Datasets like ADAM-sense have been specifically curated to facilitate the development and benchmarking of models for this purpose, providing labelled examples of such behaviours captured from wearable sensors [7].

The conceptual framework that underpins the use of wearable sensor data for mental health is known as "digital phenotyping." First formally articulated in 2015, digital phenotyping is defined as the "moment-by-moment quantification of the individual-level human phenotype in situ using data from personal digital devices" [8]. However, the resulting datasets, a torrent of high-frequency, multi-modal data, are characterised by high volume, noise, and complex multi-modality. However, this raw data is not inherently meaningful; it is a stream of numbers and signals that, in its unprocessed form, is not directly interpretable or actionable for clinicians or patients [9].

Despite high patient acceptance and mature sensor technology creating a strong "pull" for data-driven mental healthcare [3], a fundamental mismatch persists: the healthcare system and its existing workflows are not equipped to ingest, process, or interpret continuous streams of raw sensor data for large patient populations [10]. The core impediment, the systemic bottleneck, is not hardware or patient compliance, but the interpretive gap. This gap is defined by the absence of a robust analytic and interpretive layer that can sit between the patient's device and the clinician's workflow. To bridge this gap, an automated, intelligent pipeline is required, one that can transform the chaotic deluge of raw data into concise, contextualised, and clinically actionable insights.

To address this challenge, we present a novel, end-to-end automated pipeline designed to process, analyse, and interpret multi-modal data from a commercial wearable device for the purpose of mental health monitoring. Our system architecture makes the following primary contributions:

- Integration of Human Activity Recognition (HAR) for Behavioural Biomarkers: We employ a specialised deep learning model (1D-CNN-BiLSTM) to systematically translate raw IMU data into a continuous stream of clinically relevant, categorised psychomotor behaviours (e.g., specific anxiety-related fidgeting). This step transforms noisy, high-dimensional sensor readings into discrete, interpretable behavioural biomarkers, which form the foundation for downstream temporal analysis.
- A Multi-Stage Data Enrichment Process: We introduce a pipeline that moves beyond raw classifications. It employs a temporal "bout analysis" to contextualise both

To solve this, the paper proposes a novel end-to-end automated pipeline that:

Uses a 1D-CNN-BiLSTM HAR model to convert raw IMU data into meaningful behavioural biomarkers (e.g., specific anxiety-related fidgeting)

Applies multi-stage enrichment with temporal bout analysis (frequency + duration of behaviours and heart-rate arousal) to give contextualised insights

Performs longitudinal, baseline-based analysis, comparing each week to a personal baseline and highlighting important changes.

Uses a RAG-enhanced LLM and a multi-prompt setup to create three types of reports from the same data:

A clinical report for psychologists (short, with treatment hints),

A patient report (empathetic, easy language),

A research report (detailed analytical).

Overall, the introduction explains why such a pipeline is needed and what main components and contributions the sy

3

recognised activities and physiological arousal (heart rate), transforming raw event data into meaningful metrics of frequency and duration.

- Longitudinal, Baseline-Referenced Analysis: Our system establishes an individual's personal baseline from an initial monitoring period and uses this baseline to perform a comparative, week-over-week analysis, automatically highlighting trends and significant deviations in physiological and behavioural patterns.
- RAG-Enhanced, Multi-Audience Report Generation: The pipeline leverages a Large Language Model (LLM) enhanced with Retrieval-Augmented Generation (RAG) to ensure outputs are grounded in a clinical knowledge base. Crucially, it uses a multi-prompt architecture to generate three distinct, audience-appropriate reports from the same underlying data: clinical report (a concise summary with treatment considerations for a psychologist), patient report (an empathetic, non-clinical summary for the study volunteer) and a research report (a comprehensive analytical report).

Big Picture

2 Methodology

The paper builds a four-stage automated pipeline that takes raw wearable sensor data and turns it into clinically meaningful, personalised text reports for different audiences (psychologist, patient, researcher).

To address the challenge of translating raw sensor data into actionable insights, we designed and implemented a four-stage automated pipeline. The system is designed to be modular, allowing for independent development and refinement of each stage. Figure 1 provides a high-level overview of the system architecture, from data acquisition to final report generation.

Stage 1 – Data Acquisition &

Ethics & Consent:
Study approved by the Unive of Siegen ethics committee; t volunteer gave informed consent. Data is anonymised used only for research.

Data Sources:

Device: Garmin Venu 3 smartwatch

Signals:

IMU (accelerometer + gyrosc → motion

PPG → heart rate

Data is logged via a custom Monkey C app, stored as .FIT files, then transferred to a PC.

The volunteer also fills a daily digital journal: mood, nervousness, sleep quality, p attacks.

HAR Model:

Uses a 1D-CNN + BiLSTM model to detect anxiety-related behaviours (e.g., hand scratching, hair pulling).

Works on 4-second overlapping windows of IMU data.

Outputs: activity label + confidence score for each window.

Performance (5-fold CV on Garmin data mirroring ADAM-Sense):

Accuracy ≈ 0.87 , Precision ≈ 0.86 , Recall ≈ 0.87 , Macro F1 ≈ 0.86 .

This stream of activity labels is the input for the next temporal analysis stage

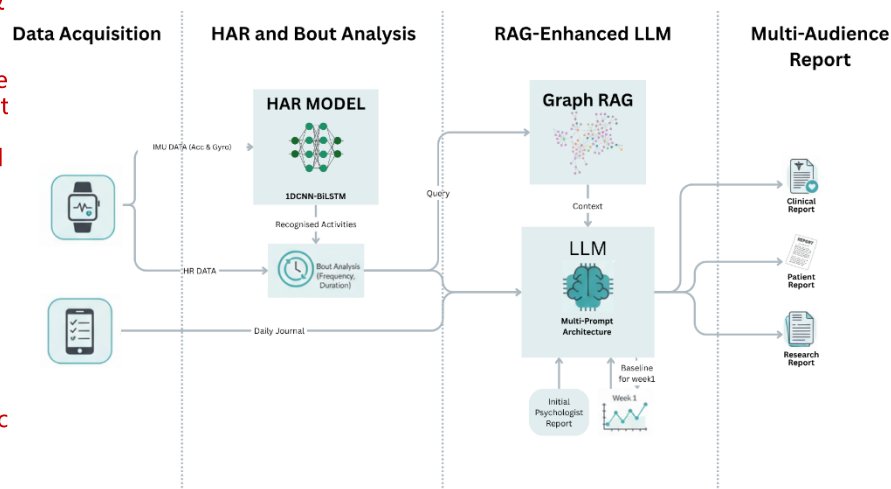


Fig. 1. Pipeline for Mental Health Report Generation using RAG-Enhanced LLM

2.1 Stage 1: Data Acquisition and Human Activity Recognition (HAR)

The initial stage of the pipeline is dedicated to collecting multi-modal data from the study volunteer and processing the raw motion data to identify clinically relevant

Raw classifications are turned into bouts and metrics.

Bout Analysis:
A bout = a continuous block of the same event/activity, with small gaps all

Heart-rate (HR) bouts: HR above the daily average, broken if the gap > 120 seconds.

Reflects autonomic nervous system recovery time → only sustained arousal counted.

Behavioural (HAR) bouts: same anxiety-related activity, broken if the gap > 2 seconds.

Longer gap to allow brief interruptions while the behaviour is ongoing.

Metrics Extracted per period (e.g., per week):

Total duration of each activity

Number of bouts

Average and max bout duration

Typical time of day when it happens

HR metrics are calculated relative to that day's personal mean HR
→ This turns a raw event log into a feature-rich summary of physiological and behavioural status.

4

behaviours. This stage involves two primary data streams: passive sensing from a commercial smartwatch and the behaviour sensing through subjective self-reports.

Ethical Compliance and Data Consent.: This study, including data acquisition from the volunteer, was reviewed and approved by the Ethics Committee of the University of Siegen. The volunteer provided informed consent after receiving an explanation of the study's purpose, the specific types of physiological and behavioural data collected, the methods of data storage and processing, and the intended use of the generated reports. All data was anonymized prior to analysis and utilized exclusively for research purposes.

Data Sources: We use a commercial Garmin Venu 3 smartwatch to collect continuous time-series data, specifically 3-axis accelerometer and gyroscope data for motion analysis, and photoplethysmography (PPG) for heart rate monitoring. Data acquisition was facilitated by a custom logging application programmed using Monkey C, which ensured access to the high-frequency raw IMU and PPG streams. The raw data was stored locally on the device in .FIT format and then manually extracted via a USB connection to a personal computer for subsequent processing. Complementing this passive sensor data, the volunteer provides daily self-reported subjective data through a digital journal, including ratings for mood, nervousness, and sleep quality, as well as details of any panic attacks.

Human Activity Recognition (HAR) Model: The raw IMU data is processed by a deep learning model to recognise a set of predefined, anxiety-related behaviours (e.g., hand scratching, hair pulling). As detailed in our previous work [11], the model utilises a 1D Convolutional Neural Network (CNN) combined with a Bidirectional Long Short-Term Memory (BiLSTM) network. The CNN layers extract salient features from the time-series segments, while the BiLSTM layers capture temporal dependencies. The HAR model achieved a stable performance across 5-fold cross-validation on custom data collected using the Garmin watch that mirrored the ADAMSense Dataset, yielding a mean Accuracy of 0.87(±0.02), Precision of 0.86(±0.02), Recall of 0.87(±0.01), and a macro F1 Score of 0.86(±0.01). The HAR model analyses the IMU data in short, overlapping windows. For each 4-second window, it outputs a classification label for the recognised activity (e.g., "hand scratching") and a corresponding confidence score. This continuous stream of classified activities serves as the input for the temporal processing and contextualisation stage of the pipeline.

2.2 Stage 2: Temporal Data Processing & Contextualisation

Recognising that raw classifications are insufficient, this stage transforms discrete events into meaningful behavioural and physiological metrics.

Temporal "Bout" Analysis: To understand the temporal patterns of behaviour, we apply a "bout" analysis algorithm. A bout is defined as a sequence of the same activity classification occurring in close temporal proximity. The algorithm groups consecutive data points into a single, contiguous bout if the time gap between them is less than a predefined threshold. Critically, this threshold is differentiated by data type: Physiological Arousal (HR) bouts (periods of heart rate above the daily average) are broken by a time gap exceeding 120 seconds (2 minutes). This duration is chosen to align with the

estimated time constant for autonomic nervous system recovery, ensuring that only sustained periods of arousal, rather than momentary transient noise, are grouped as a single event. Behavioural Activity (HAR) bouts (sequences of the same recognised anxiety-related activity) are broken by a gap exceeding 300 seconds (5 minutes). This longer threshold is used to ensure behavioural continuity, distinguishing a brief interruption (e.g., a hand adjustment) from a definitive shift in the psychomotor state. This differentiation reflects the expected temporal dynamics of physiological responses versus behavioural patterns.

Metric Extraction: From this bout analysis, we extract a set of contextualised metrics for each reporting period. For physiological metrics, "above-average heart rate" is calculated daily against the individual's mean heart rate for that specific 24-hour period. Extracted metrics include total duration of an activity, number of distinct bouts, average and maximum bout duration, and the primary time of day the activity occurs. This enrichment process turns a simple event log into a feature-rich summary of the volunteer's status.

2.3 Retrieval-Augmented Generation (RAG) Enhanced LLM

Once sensor data has been processed and contextualised into a set of meaningful metrics, the final challenge is to synthesise this information into a coherent, interpretable narrative for end-users. Large Language Models (LLMs) have recently emerged as a transformative technology with immense potential to automate this process of clinical interpretation and report generation [12]. These sophisticated neural networks, trained on vast corpora of text, have demonstrated a remarkable capacity for understanding and generating human-like language, making them well-suited for tasks involving the summarisation of complex information [13] or the generation of reports.

Despite their impressive capabilities, the direct application of general-purpose LLMs in high-stakes clinical settings has significant risks and challenges. A well-known concern is the phenomenon of factual inconsistency, or "hallucination." LLMs are generative models that can produce text that is fluent, plausible, and grammatically correct, but factually untethered from reality [14]. In a medical context, this can manifest in catastrophic ways, such as inventing fictitious drug recommendations, citing non-existent clinical trials as evidence, or misrepresenting patient data, all of which pose a direct and severe risk to patient safety [15].

In response to the critical limitations of standalone LLMs, the research community has developed Retrieval-Augmented Generation (RAG), a powerful architectural paradigm designed to enhance the factual accuracy and reliability of generative models [16]. RAG is a hybrid approach that synergistically combines the strengths of a retrieval system with the generative capabilities of an LLM, making it particularly well-suited for knowledge-intensive tasks where factual grounding is paramount [17].

RAG Implementation and Knowledge Base: To ensure the final reports are empirically grounded, we enhance our Large Language Model (LLM) with a Retrieval-Augmented Generation (RAG) system.

Knowledge Base and Graph Chunking: Our knowledge base is derived from a clinical Neo4j Knowledge Graph (KG), constructed from peer-reviewed literature on anxiety

and panic disorders. The graph was manually constructed using evidence extracted from clinical textbooks and systematic reviews. This systematic extraction was subsequently structured into linked CSV files (entities and relationships) and imported into the Neo4j database. This graph structure is visually represented in Figure 2, illustrating the network of relationships between diagnoses, symptoms, behaviours, and treatments. Unlike traditional RAG that relies on document splitting, our chunks are semantic triples, programmatically generated by traversing the KG relationships. Each chunk represents a single, verifiable statement structured as: (Source Node) [Relationship] (Target Node). This graph-based chunking yields fine-grained, highly factual context chunks.

Query Generation Logic: Before retrieval, a natural language query is programmatically generated by synthesising the top three salient findings from the Stage 2 summary metrics. This process uses a template-based approach to construct a clinically relevant query from quantitative data, combining the most frequent high-confidence activity, the highest physiological deviation (e.g., peak HR, HR bout counts change), and any self-reported significant events (e.g., panic attacks). For instance, an input query is generated as: 'Patient experienced a panic attack, elevated heart rate, and hand scratching.'

Retrieval Mechanism: The query is converted into a vector embedding using a SentenceTransformer model ('all-MiniLM-L6-v2'). We then calculate the cosine similarity between this query vector and the pre-computed embeddings of all graph chunks. The text from the top-k (k=5) most similar chunks is retrieved and provided as context to the LLM. This high value of k leverages the dense, high-precision nature of the graph chunks, ensuring multiple clinical connections are provided for synthesis.

Stage 3 – RAG-Enhanced LLM

Goal: turn metrics into text reports that are clinically grounded, not hallucinated.

Why RAG?

Plain LLMs can hallucinate (invent wrong clinical facts, fake studies, etc.), which is dangerous in health

RAG combines retrieval + generation so the LLM is guided by real, curated clinical knowledge.

Knowledge Base (Neo4j Graph):

Built from clinical textbooks and systematic reviews on anxiety and panic disorders.

Data is stored as entities and relationships in CSVs, then imported to Neo4j.

Nodes connect diagnoses, symptoms, behaviours, treatments (illustrated in Fig. 2).

Graph-based Chunking (instead of document chunks):

Each chunk is a semantic triple:
(Source Node) [Relationship] (Target Node)

Example: (Panic Disorder) –[ASSOCIATED_WITH]→ (Heart Palpitations)

This gives fine-grained, verifiable facts.

Query Generation:

From Stage 2 metrics, the system picks the top 3 salient findings, e.g.:

Most frequent high-confidence behaviour,

Largest physiological deviation (e.g., HR changes),

Important self-reported events (panic attacks).

It builds a natural language query like:

"Patient experienced a panic attack, elevated heart rate, and hand scratching."

Retrieval:

The query is embedded using SentenceTransformer (all-MiniLM-L6-v2).

Cosine similarity is computed against all graph triple embeddings.

Top k = 5 most similar chunks are retrieved and passed as context to the LLM.

Stage 4 – Longitudinal, Multi-Audience Report Generation

Uses Google's Gemma LLM + RAG context + temporal metrics to generate report

Initial Context Extraction:

The system first processes the Initial Psychologist Report (free text).

With a structured prompt + JSON output, it extracts:

Diagnosis (e.g., panic disorder with agoraphobia),

Most detailed; generates a comprehensive analytical report for documentation

7

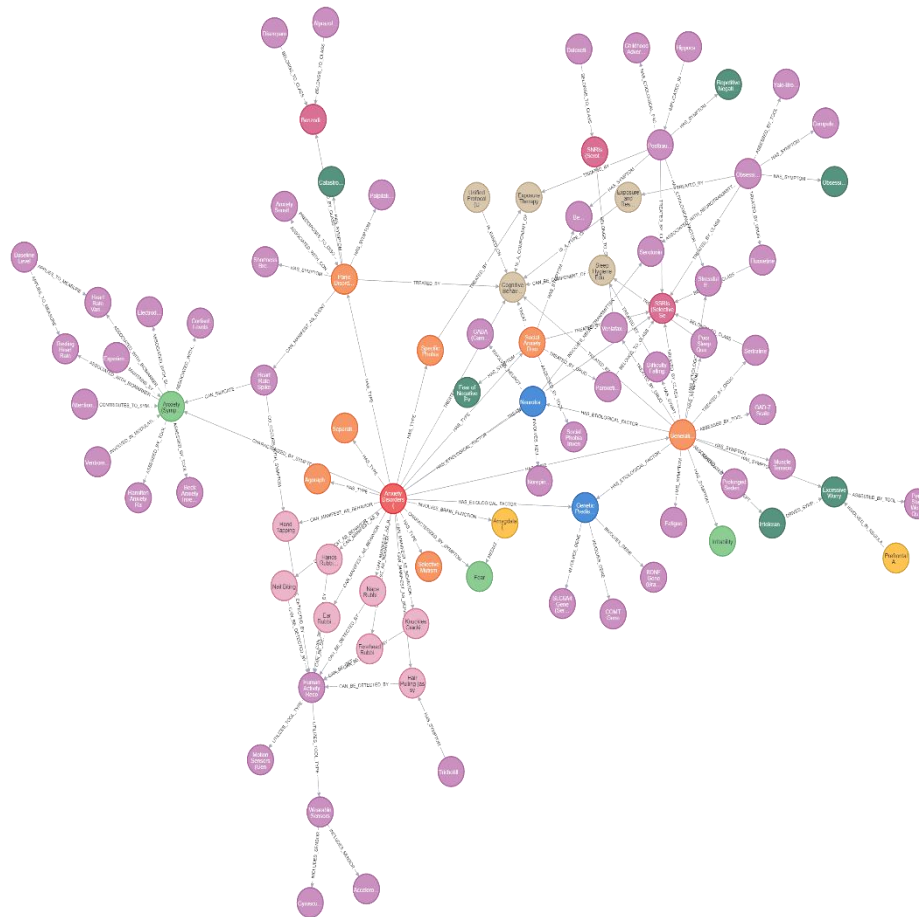


Fig. 2. Knowledge Graph for Anxiety Disorder

2.4 Stage 4: Longitudinal, Multi-Audience Report Generation

The final stage synthesises all processed data and retrieved context into coherent, audience-specific reports using Google's Gemma LLM.

Initial Context Extraction: Before generating the comparative report, the raw, free-text Initial Psychologist Report is first processed to distil high-level, static clinical context. This is achieved by using the LLM itself with a focused prompt and a JSON output constraint to extract two key pieces of information: the diagnosis (e.g., panic disorder with Agoraphobia) and the key symptoms and impairments mentioned in the initial file. This pre-processed context is then integrated into all subsequent prompts to personalise the analysis.

Baseline Comparison: For our case study, Week 1 is designated as the personal baseline, and Week 2 is compared directly against these initial metrics. More broadly, the system is architected to support longitudinal analysis where the baseline can be

Key symptoms and impairments.

This context is inserted into all future prompts to personalise the analysis.

Baseline Comparison (Longitudinal):

Case study: Week 1 = personal baseline, Week 2 compared against it.

In general: for longer monitoring, baseline can be the average of the first 4 stable

Baseline metrics are included in the prompt, and the LLM is instructed to:

Compare periods,

Highlight trends & deviations.

Why Multi-Audience?

Clinicians and patients have very different needs:

Clinician: time-poor, needs clear trends, deviations, risk factors, and clinical langu

Patient: needs gentle, motivating, non-technical feedback to avoid anxiety or con

Table 1 shows the differences in: purpose, tone, focus, language, and granularity.

8

dynamically defined. For cohorts with extended monitoring periods (e.g., eight or more weeks), the personal baseline can be calculated as the average of the initial four weeks of stable data. This baseline data is explicitly included in the prompt, with instructions for the LLM to perform a comparative analysis and highlight trends.

Multi-Prompt Architecture: While the technical challenges of data processing and interpretation are substantial, the ultimate success of any digital health intervention hinges on the effective communication of its outputs to the intended end-users [18]. In the context of a system designed to generate reports from mental health sensor data, two primary stakeholders are immediately apparent: the clinician (e.g., psychologist, psychiatrist, or therapist) who is responsible for guiding treatment, and the patient (or study volunteer) who is the source of the data and the ultimate beneficiary of the care. These two groups have fundamentally different relationships to the data, different levels of domain expertise, and distinct informational and emotional needs.

For healthcare professionals, the primary function of a report generated from patient data is to support clinical decision-making [19]. Clinicians operate in a time-constrained environment and are often subject to significant information overload [10]. Rather than receiving a continuous, unfiltered stream of raw data, clinicians strongly prefer to receive summarised information that highlights clinically significant events. The content should focus on identifying meaningful patterns, such as trends over time, significant deviations from a patient's personal baseline, and the emergence of potential risk factors (e.g., a sudden increase in nocturnal restlessness) [20]. The language should be clinical and precise, using established terminology to facilitate clear and unambiguous communication. The ultimate goal of a clinician-facing report is to distil the complexity of the raw data into a clear, digestible summary that can be quickly integrated into the therapeutic process to guide interventions, adjust treatment plans, and monitor progress.

The needs of the patient are starkly different. For the individual whose life is being monitored, the report is not just a collection of data points; it is a reflection of their personal experience and well-being. Therefore, reports designed for patients must be crafted with a high degree of psychological and emotional intelligence. The primary goal is to provide feedback that is empowering, educational, and supportive, without causing undue anxiety, confusion, or self-judgement.

The profound differences in the reporting requirements for these two key stakeholders are summarised in Table 1. This side-by-side comparison makes it clear that a single, generic, "one-size-fits-all" report is fundamentally inadequate to meet the needs of either group.

Table 1. Comparison of Reporting Requirements for Clinicians and Patients

Reporting Dimension	Clinician Requirement	Patient Requirement
Purpose	To inform diagnosis, treatment planning, and risk assessment.	To foster self-awareness, encourage positive behaviour change, and promote engagement in care.

Multi-Prompt Architecture:

Psychologist Prompt:

Objective, clinical, data-heavy.

Focus: trends, baseline deviations, risk factors, actionable points for review.

Volunteer (Patient) Prompt:

Empathetic, supportive, non-clinical language.

Focus: gentle observations, positive reinforcement, links between behaviour and fe

Research Prompt:

Most detailed; generates a comprehensive analytical report for documentation.

Tone	Objective, analytical, clinical.	Empathetic, encouraging, supportive, non-judgemental.
Content Focus	Trends, deviations from baseline, risk factors, quantitative data.	Gentle observations, positive reinforcement, and connections between behaviour and feelings.
Language	Precise, technical, clinical terminology	Simple, non-clinical, jargon-free, accessible language.
Data Granularity	Summarised metrics, critical value alerts, and on-demand access.	Qualitative descriptions, high-level summaries, gentle feedback.

- To generate reports tailored to the different stakeholders, we employ a multi-prompt architecture:
- The Psychologist Prompt is instructed to be objective, clinical, and data-driven, focusing on trend analysis of risk factors and suggesting actionable points for clinical review.
 - The Volunteer Prompt is instructed to be empathetic, encouraging, and use non-clinical language. It translates quantitative data into gentle, qualitative observations to provide feedback without causing alarm.
 - A third, more comprehensive prompt is used to generate the main analytical report for research documentation.

3 **Results: Case Study Demonstration**

To evaluate the efficacy of our pipeline, we present a case study analysing two separate weeks of data from a volunteer: 39-year-old female with a clinical diagnosis of panic disorder with agoraphobic avoidance. Week 1 (4 days of data) was used to establish the initial personal baseline, and Week 2 (5 days of data) was processed and compared against this baseline.

3.1 **Pipeline Input: Longitudinal Data Summary**

After passing through the first two stages of our pipeline (HAR and Temporal Processing), the raw sensor and journal data were transformed into a structured, feature-rich summary. This summary served as the primary input to the RAG and Report Generation stage of the pipeline. The initial report from the psychologist about the patient is also inputted. This raw clinical text is first processed by the LLM (Stage 4, Initial Context Extraction) to automatically generate a concise diagnosis and list of key symptoms, which are then used as high-level, persistent context for all generated reports. After the Week 1 reports were generated, the baseline file is also generated and saved and used as an input for Week 2 report generation. A key real-world challenge noted during this period was the absence of self-reported journal data in Week 2, which the pipeline was required to handle.

3. Results: What the Pipeline Actually Did in the Case S

The authors tested their pipeline on one real volunteer

39-year-old woman

Diagnosis: Panic disorder with agoraphobia

Week 1 (4 days) → used as baseline

Week 2 (5 days) → compared against Week 1

3.1 Pipeline Input – What Data Went In?

After HAR + temporal processing, the system had a feature-rich longitudinal summary of:

- Sensor data (heart rate + IMU / activities)
- Self-reports (only in Week 1, not in Week 2)

An initial psychologist report (free text).

The LLM:

- First processed the psychologist’s text → extracted
 - Diagnosis 10
 - Key symptoms

Generated Week 1 reports, then saved a baseline file. Used that baseline + Week 2 metrics (even without journal data in Week 2) for Week 2 report generation. So the pipeline had to work even when self-reported data was missing in Week 2.

3.2 Pipeline Output – What Did the System Say Clinically?

3.3 Analysis – Why This Is More Than a Simple Summary

The system produced three reports (psychologist, patient, and a summary report for Week 2). But the pipeline’s bout analysis story:

HR bout count (episodes of high HR):

- Week 1: 82 bouts
- Week 2: 54 bouts (fewer)

BUT:

- Total high-HR time in Week 2: 8 hours and 6 minutes (much longer than baseline)
- Max single bout in Week 2: 37.5 minutes (much longer than baseline)

So:

- Fewer episodes, but they last much longer and are more intense.

The LLM, using the longitudinal prompt + RAG, correctly interpreted this:

- Not just reading numbers, but summarising: "Intermittent, substantial physiological activation despite the lower average heart rate."
- From that, it: Marked likelihood of future panic attacks as increasing, and Suggested specific interventions targeting autonomic regulation.

3.2 Pipeline Output: Generated Clinical Summary

The pipeline's final stage generated three distinct reports. The most concise and clinically oriented of these is the Psychologist's Report. The system-generated assessment from this report for Week 2 identified the volunteer's physiological state as a primary risk factor, noting that "intermittent, substantial physiological activation" was occurring despite a lower average heart rate. Based on this trend analysis, the system's final synthesis produced a likelihood assessment and a corresponding treatment recommendation:

- Likelihood Assessment for Future Panic Attacks: Increasing. The report reasoned that "The intermittent, yet substantial, physiological arousal suggests a heightened potential for future panic episodes."
- Treatment Recommendation: The system suggested prioritising the exploration and implementation of "Autonomic Regulation Techniques," such as biofeedback or paced breathing exercises, to directly target the observed HR fluctuations.

3.3 Analysis of Results

The case study demonstrates the pipeline's ability to move beyond superficial data summarisation to uncover nuanced, empirically grounded insights. A simple analysis of the average daily heart rate might have misleadingly suggested a clinical improvement (a decrease from 81.67 to 76.45 bpm). However, by leveraging the temporal "bout" analysis and the longitudinal comparison, our system identified a more critical underlying pattern: a shift from frequent, shorter bouts of physiological arousal to fewer, but substantially longer and more intense, episodes. Quantitatively, the HR bout count dropped from 82 (Baseline Week 1) to 54 (Week 2), but the total cumulative duration increased to 8 hours and 6 minutes, with the maximum single bout duration reaching 37.5 minutes (a significant increase from the implied 14.5 minutes of baseline bouts). The LLM, guided by the longitudinal prompt, correctly synthesised this complex observation. It did not simply state the numbers but correctly interpreted the pattern as "intermittent, substantial physiological activation despite the lower average heart rate." This data-driven hypothesis directly informed the generated likelihood assessment ("Increasing") and led to a highly specific and relevant treatment recommendation, focusing on "Autonomic Regulation Techniques." This demonstrates the system's capacity to automatically process complex data, identify non-obvious trends, and translate them into actionable points for clinical consideration.

Key findings:

- It detected "intermittent, substantial physiological activation"
- Even though the average HR in Week 2 was lower, the pattern was still risky.
- The system gave:
 - Likelihood of future panic attacks: Increasing
 - Because of repeated strong arousal episodes.
 - Treatment recommendation: Focus on autonomic regulation techniques (e.g., biofeedback, paced breathing) to better manage heart rate fluctuations.

4 Discussions

Our work addresses the critical challenge of translating high-frequency, multi-modal data from wearable sensors into clinically actionable insights for mental health monitoring. The case study presented demonstrates the successful implementation of our end-to-end pipeline, moving beyond simple data aggregation to perform a nuanced, longitudinal analysis and generate audience-specific, interpretable reports.

4.1 Summary of Findings

The pipeline can detect subtle, clinically important patterns that simple statistics would miss. Example:

- Average HR dropped from 81.67 → 76.45 bpm → looks like "better".
- But bout analysis + longitudinal comparison showed:
 - Bouts: 82 → 54 (fewer)
 - But episodes became much longer, max bout 37.5 minutes.
- The RAG-enhanced LLM:
 - Didn't just repeat numbers.
 - Correctly summarised the pattern as "intermittent, substantial physiological activation despite lower average HR".
 - Concluded: likelihood of future panic attacks = Increasing.
 - Recommended Autonomic Regulation Techniques (e.g., biofeedback, paced breathing), grounded in clinical knowledge retrieval.

This shows the system can go from raw data → structured metrics → clinical-level inference automatically.

4.2 Implications for Clinical Practice & Patients

For clinicians:

They face time pressure and information overload.

The system acts as an analytical filter, turning raw streams into:

Short summaries of key trends, deviations, and risk factors.

A hypothesis, not just numbers → works like a decision-support tool.

For patients:

The separate empathetic report is crucial.

11

4.1 Summary of Findings

It uses simple, non-judgmental language, aims to:

Build self-awareness,

Support behaviour change,

Avoid triggering extra anxiety,

This can improve engagement and therapeutic alliance in digital health

The primary contribution of this work is the demonstration of a system capable of uncovering subtle but clinically significant patterns that a superficial analysis might miss. A conventional review of the volunteer's data might have concluded that a decrease in average daily heart rate (from 81.67 bpm to 76.45 bpm) signified an improvement. However, by employing a temporal "bout" analysis and a longitudinal comparison, our pipeline identified a more critical underlying trend: a shift from 82 frequent, shorter bouts in the baseline period to 54 bouts in Week 2, with the maximum bout duration increasing substantially to 37.5 minutes.

Crucially, the RAG-enhanced LLM did not merely report these metrics. Guided by a clinically focused prompt, it correctly synthesised this observation into a coherent hypothesis, identifying "intermittent, substantial physiological activation despite the lower average." This data-driven hypothesis directly informed its assessment that the likelihood of future panic attacks was "Increasing," leading to the specific, actionable recommendation to explore "Autonomic Regulation Techniques." This recommendation was empirically grounded by the RAG system, which retrieved clinical knowledge linking sustained physiological arousal to biofeedback-based regulation techniques. This outcome validates the pipeline's ability to automate the process of moving from raw data to a clinical-level inference.

4.2 Implications for Clinical Practice and Patient Engagement

The methodology presented has significant implications for two key stakeholders: clinicians and patients. For clinicians who operate in time-constrained environments and face information overload [10], this system acts as an analytical filter. It automates the laborious task of sifting through raw data streams and instead delivers a concise summary of clinically salient events and trends. By providing a hypothesis rather than just data points, it elevates its utility from a mere monitoring tool to a decision-support system, directly addressing the need for summarised, clinically significant information [20].

Simultaneously, our multi-prompt architecture acknowledges the distinct needs of the patient, as outlined in Table 1. The generation of a separate, empathetic, and non-judgmental report for the volunteer is a key feature. This approach aims to foster self-awareness and engagement in the care process without inducing the anxiety that overly clinical or quantitative feedback might cause. By empowering the individual with gentle, qualitative observations, we can enhance therapeutic alliance and sustained engagement with digital health interventions.

4.3 Limitations

While promising, this study has several limitations. First, the evaluation is based on a single case study. While this is sufficient to demonstrate the architectural functionality and proof-of-concept of the pipeline's end-to-end integration and unique analytical capability, the findings are not generalizable. The patterns observed are specific to this

4.3 Limitations

Single Case Study

Only one participant → results are not generalisable.

It shows the architecture works, but not population-level performance.

Incomplete Data (Real-World Challenge)

Week 2 lacked journal data and some high-confidence activity data.

The system still produced useful insights and explicitly mentioned the missing data in the report, which is itself clinically

LLM Reasoning Risk

RAG grounds facts, but the final reasoning (e.g., "Likelihood: Increasing") depends on the LLM.

LLM may misinterpret complex statistics or temporal patterns → risk of hallucinated interpretations.

The LLM is not a statistical engine, just an interpreter of metrics.

Suggested solution: human-in-the-loop + combine LLM + explicit statistical methods.

No Formal Usability Study Yet

Reports are designed to be "clinically actionable," but:

Not yet tested with real clinicians and patients in practice.

The system should be seen as an add-on to support therapy, not a replacement for full clinical assessment.

4.4 Future Work

Planned next steps:

Larger, Diverse Cohort

Test the pipeline on more participants and different conditions to get generalizable results.

Usability Study with Clinicians

Evaluate the clarity, relevance, and usefulness of the psychologist-facing reports.

Use clinician feedback to refine prompts and report format.

Expand the RAG Knowledge Base
Performance metrics require a statistically powered study cohort, which should be the focus of future work.

Add more clinical literature to improve the depth and accuracy of generated interpretations.

Efficiency & Deployment

Large models like Gemma-3 are heavy.

Explore:

Model compression (e.g., quantization),

Smaller efficient models,

To enable low-latency, practical clinical use.

Feedback Loop Integration

Allow clinicians and patients to annotate or comment on reports.

Use this feedback to:

Personalise monitoring,

Improve long-term performance and trust.

individual and this timeframe. Formal clinical validation and establishing generalizable performance metrics require a statistically powered study cohort, which should be the focus of future work.

Second, the case study highlighted the real-world challenge of incomplete data. The absence of self-reported journal and high-confidence activity data in Week 2 is a limitation of the dataset, but it also demonstrates the robustness of the pipeline. The system was still able to derive meaningful physiological insights and explicitly noted the data gap in its report, which is in itself a clinically relevant piece of information.

Third, while the RAG process grounds the LLM in the knowledge base, the final inferential step, synthesising the bout analysis, baseline comparison, and RAG context into a clinical conclusion (e.g., "Likelihood: Increasing"), relies on the inherent reasoning capabilities of the Large Language Model. The system is therefore susceptible to potential misinterpretations or "hallucinations" of complex statistical or temporal patterns that may not be explicitly covered in the retrieved clinical text. The LLM acts as an interpreter of pre-processed metrics, not a statistical engine, highlighting a limitation in relying on generative models for intricate quantitative analysis. A potential solution of this problem is the need for a human-in-the-loop and a combination between LLM and statistical approaches in the future.

Finally, while the generated reports are designed to be "clinically actionable," we have not yet formally evaluated their utility or usability in a clinical setting. The next phase of research must involve clinicians and patients interacting with these reports to validate their effectiveness in practice. Crucially, this technology is not intended to replace a clinician and thorough clinical exploration but can be regarded as a meaningful add-on to increase the efficiency and effectiveness of the psychotherapeutic process.

4.4 Future Work

Based on these findings and limitations, we have identified several directions for future work. The immediate next step is to expand the study to a larger and more diverse cohort of participants to validate the pipeline's performance across different individuals and conditions. Second, we plan to conduct a formal usability study with mental health professionals to evaluate the clarity, relevance, and clinical utility of the psychologist-facing reports. This will involve gathering qualitative and quantitative feedback to iteratively refine the report generation prompts. Third, we will continue to expand the RAG knowledge base with a broader range of clinical literature to enhance the depth and accuracy of the LLM's outputs. Fourth, we recognize that the computational requirements of large models like Gemma-3 present a challenge for real-time deployment; future work will explore optimization via model compression (e.g., quantization) or deployment on smaller, highly efficient models to ensure practical, low-latency clinical integration. Finally, exploring the integration of a feedback loop, where both clinician and patient can annotate or comment on the reports, could further personalise the monitoring process and improve the system's long-term efficacy.

Wearables could change mental healthcare, but the main problem is that raw sensor data doesn't fit easily into clinical practice. This proposes a full automated pipeline that fixes this gap by combining:

- HAR (to detect behaviours),
- Temporal bout analysis (to understand patterns over time), and
- A RAG-enhanced LLM (to turn data into grounded, clinical text).

In the case study, the system found a subtle but important change in physiological arousal that normal averages would have missed. It turned this into:

- A data-based hypothesis,
- A risk assessment, and
- A specific treatment suggestion.

Using different prompts, it also tailored the output for clinicians (concise, evidence-based) and patients (empathetic, supportive). Overall, the system shows a concrete way to make digital phenotyping actually usable in real mental healthcare, enabling more data-driven, personalised, and proactive treatment.

5 Conclusion

The proliferation of wearable technology presents a transformative opportunity for mental healthcare, yet the primary obstacle to its widespread adoption is the absence of a robust interpretive layer between raw sensor data and clinical workflows. This paper presented a novel, end-to-end automated pipeline designed to bridge this gap. By integrating human activity recognition, temporal bout analysis, and a RAG-enhanced large language model, our system successfully transforms high-frequency, multi-modal data into longitudinal, clinically actionable insights.

Our case study demonstrated the pipeline's ability to move beyond superficial metrics, uncovering a nuanced and clinically significant shift in a volunteer's physiological arousal that would likely be missed by manual or less sophisticated analysis. The system's capacity to synthesise this finding into a data-driven hypothesis, a risk assessment, and a specific treatment recommendation validates its potential as a clinical decision-support tool. Furthermore, the multi-prompt architecture ensures that these insights are communicated effectively to different stakeholders, empowering clinicians with concise, relevant data while engaging patients with empathetic, supportive feedback. This work represents a significant step toward translating the promise of digital phenotyping into a practical, scalable tool that can support a more data-driven, personalised, and proactive approach to mental healthcare.

References

1. A. Abd-Alrazaq *et al.*, "Wearable Artificial Intelligence for Detecting Anxiety: Systematic Review and Meta-Analysis," *J Med Internet Res*, vol. 25, p. e48754, Nov. 2023, doi: 10.2196/48754.
2. N. Gomes, M. Pato, A. R. Lourenço, and N. Datia, "A Survey on Wearable Sensors for Mental Health Monitoring," *Sensors*, vol. 23, no. 3, 2023, doi: 10.3390/s23031330.
3. A. Barrera *et al.*, "A Survey of Preferences for Sensing Technologies in People with Severe Mental Illness Admitted to an Acute Psychiatric Unit," *Ann Biomed Eng*, vol. 53, no. 9, pp. 2372–2383, Sep. 2025, doi: 10.1007/s10439-025-03769-0.
4. J. Chen, M. Abbod, and J.-S. Shieh, "Pain and Stress Detection Using Wearable Sensors and Devices—A Review," *Sensors*, vol. 21, no. 4, 2021, doi: 10.3390/s21041030.
5. N. K. Sahu, S. Gupta, and H. R. Lone, "AnxietyFaceTrack: A Smartphone-Based Non-Intrusive Approach for Detecting Social Anxiety Using Facial Features," 2025. [Online]. Available: <https://arxiv.org/abs/2502.16106>
6. U. Oleh, R. Obermaisser, and A. S. Ahammed, "A Review of Recent Techniques for Human Activity Recognition: Multimodality, Reinforcement Learning, and Language Models," *Algorithms*, vol. 17, no. 10, 2024, doi: 10.3390/a17100434.
7. N. Khan, M. Ghani, and G. Anjum, "ADAM-sense: Anxiety-displaying activities recognition by motion sensors," *Pervasive Mob Comput*, vol. 78, p. 101485, Oct. 2021, doi: 10.1016/j.pmcj.2021.101485.
8. D. Adam, "Digital phenotyping using smartphones could help steer mental health treatment," *Proceedings of the National Academy of Sciences*, vol. 122, no. 14, p. e2505700122, 2025, doi: 10.1073/pnas.2505700122.

Here's a clear, structured summary of the main points from the page you're viewing:

9. A. Ng, B. Kornfield, S. M. Schueller, A. K. Zalta, M. Brennan, and M. Reddy, "Provider Perspectives on Integrating Sensor-Captured Patient-Generated Data in Mental Health Care," *Proc ACM Hum Comput Interact*, vol. 3, no. CSCW, pp. 1–25, Nov. 2019, doi: 10.1145/3359217.
10. S. Sadhu, D. Solanki, J. A. Brick, N. R. Nugent, and K. Mankodiya, "Designing a Clinician-Centered Wearable Data Dashboard (CarePortal): Participatory Design Study," *JMIR Form Res*, vol. 7, p. e46866, Dec. 2023, doi: 10.2196/46866.
11. U. Oleh and R. Obermaisser, "Recognition of Anxiety-Related Activities using 1DCNNBiLSTM on Sensor Data from a Commercial Wearable Device," in *The 15th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH 2025)*, Istanbul, Türkiye, Oct. 2025.
12. R. Rubin, M. Li, S. Zhao, B. Chen, H. R. Chang, and L. Yao, "In-context learning for zero-shot medical report generation," in *Proceedings of the 32nd ACM international conference on multimedia*, 2024, pp. 8721–8730.
13. D. and G.-B. C. and E. A. K. and Z. J. and Z. Y. and B. A. and B. M. and L. C. Bodnarczyk Lydie and Reichenpfader, "Scientific Evidence for Clinical Text Summarization Using Large Language Models: Stoping Reviews," *J Med Internet Res*, vol. 27, p. e68998, May 2025, doi: 10.2196/68998.
14. L. Tang *et al.*, "Evaluating large language models on medical evidence summarization," *NPI Digit Med*, vol. 6, no. 1, p. 158, Aug. 2023, doi: 10.1038/s41746-023-00896-7.
15. Google Gemini enhanced with Retrieval-Augmented Generation (RAG). Knowledge base built from a Neo4j clinical graph of anxiety/panic disorder literature. Ensures outputs are grounded in verified language models.
16. N. D. Bhati, and D. K. Shukla, "Retrieval-Augmented Generation (RAG) in Healthcare: A Comprehensive Review," *AI*, vol. 6, no. 9, 2025, doi: 10.3390/ai6090226.
17. R. Uapadhyay and M. Viviani, "Enhancing Health Information Retrieval with RAG by Prioritizing Topical Relevance and Factual Accuracy," 2025. [Online]. Available: <https://arxiv.org/abs/2502.04666>
18. P. Godage, P. Sribaddana, and R. Hewapathirana, "Sustaining digital health systems through key stakeholder engagement: a critical systems heuristics perspective," *Ox-Open Digital Health*, vol. 1, p. eod0104, Oct. 2023, doi: 10.1093/oxford/odh0104.
19. H. E. Dzmitrowicz, L. J. Blakely, L. W. Jones, and T. W. LeBlanc, "Bridging New Technology Into Clinical Practice With Mobile Apps, Electronic Patient-Reported Outcomes, and Wearables," *American Society of Clinical Oncology Educational Book*, no. 42, pp. 94–99, 2022, doi: 10.1200/EDBK_350550.
20. E. Neeman *et al.*, "Mobile app and wearable sensor-based patient and caregiver physical function and reported outcomes: A survey on oncologists' current practices and preferences on data delivery," *Journal of Clinical Oncology*, vol. 39, no. 28_suppl, p. 174, 2021, doi: 10.1200/JCO.2020.39.28_suppl.174.
21. S. Sadhu, D. Solanki, J. A. Brick, N. R. Nugent, and K. Mankodiya, "Designing a Clinician-Centered Wearable Data Dashboard (CarePortal): Participatory Design Study," *JMIR Form Res*, vol. 7, p. e46866, Dec. 2023, doi: 10.2196/46866.
22. R. Rubin, M. Li, S. Zhao, B. Chen, H. R. Chang, and L. Yao, "In-context learning for zero-shot medical report generation," in *Proceedings of the 32nd ACM international conference on multimedia*, 2024, pp. 8721–8730.
23. D. and G.-B. C. and E. A. K. and Z. J. and Z. Y. and B. A. and B. M. and L. C. Bodnarczyk Lydie and Reichenpfader, "Scientific Evidence for Clinical Text Summarization Using Large Language Models: Stoping Reviews," *J Med Internet Res*, vol. 27, p. e68998, May 2025, doi: 10.2196/68998.
24. L. Tang *et al.*, "Evaluating large language models on medical evidence summarization," *NPI Digit Med*, vol. 6, no. 1, p. 158, Aug. 2023, doi: 10.1038/s41746-023-00896-7.
25. Google Gemini enhanced with Retrieval-Augmented Generation (RAG). Knowledge base built from a Neo4j clinical graph of anxiety/panic disorder literature. Ensures outputs are grounded in verified language models.
26. N. D. Bhati, and D. K. Shukla, "Retrieval-Augmented Generation (RAG) in Healthcare: A Comprehensive Review," *AI*, vol. 6, no. 9, 2025, doi: 10.3390/ai6090226.
27. R. Uapadhyay and M. Viviani, "Enhancing Health Information Retrieval with RAG by Prioritizing Topical Relevance and Factual Accuracy," 2025. [Online]. Available: <https://arxiv.org/abs/2502.04666>
28. P. Godage, P. Sribaddana, and R. Hewapathirana, "Sustaining digital health systems through key stakeholder engagement: a critical systems heuristics perspective," *Ox-Open Digital Health*, vol. 1, p. eod0104, Oct. 2023, doi: 10.1093/oxford/odh0104.
29. H. E. Dzmitrowicz, L. J. Blakely, L. W. Jones, and T. W. LeBlanc, "Bridging New Technology Into Clinical Practice With Mobile Apps, Electronic Patient-Reported Outcomes, and Wearables," *American Society of Clinical Oncology Educational Book*, no. 42, pp. 94–99, 2022, doi: 10.1200/EDBK_350550.
30. E. Neeman *et al.*, "Mobile app and wearable sensor-based patient and caregiver physical function and reported outcomes: A survey on oncologists' current practices and preferences on data delivery," *Journal of Clinical Oncology*, vol. 39, no. 28_suppl, p. 174, 2021, doi: 10.1200/JCO.2020.39.28_suppl.174.

In short: This paper presents a **RAG-enhanced, multi-stage pipeline** that converts wearable sensor data into **clinically actionable, audience-specific mental health reports**, demonstrating promise in bridging the gap between raw data and practical healthcare.

Would you like me to create a **diagram of the pipeline stages** so you can visualize the flow more clearly?