Wearable Data Aggregation Framework for Fine-Tuning of LLMs to Forecasting Future-Aware Insights*

Linas Petkevičius[1,*,†], Povilas Gudžius[2,†] and Ernestas Filatovas[3,†]

[1]*Quantum Wander, Lithuania*
[2]*Spike Technologies, Lithuania*
[3]*Vilnius University, Institute of Data Science and Digital Technologies*
Akademijos str. 4, LT-08412 Vilnius, Lithuania

Abstract

Wearable sensing and digital health analytics generate large volumes of heterogeneous physiological time series, yet converting these data into accurate and computationally efficient predictive insights remains challenging. While large language models (LLMs) are increasingly applied to health-related dialogue and recommendation tasks, their suitability for direct numerical forecasting from wearable data has not been systematically studied. This paper presents a unified statistical–semantic aggregation framework that transforms daily wearable-device time series into complementary statistical and structured natural-language representations suitable for text-to-regression forecasting with LLMs. We define a one-day-ahead benchmark in which next-day physiological targets are predicted exclusively from aggregated textual summaries, and investigate parameter-efficient fine-tuning using Low-Rank Adaptation (LoRA) for open-weight LLMs. Experiments on multiple public and internal wearable datasets demonstrate that lightweight adaptation consistently improves forecasting accuracy while preserving computational efficiency. In particular, mid-sized open-weight models achieve single-digit heart-rate prediction error and offer a favorable accuracy–cost trade-off compared to substantially larger alternatives. These results indicate that text-based aggregation combined with parameter-efficient adaptation enables practical, scalable, and cost-effective wearable-data forecasting using large language models.

Keywords

LLMs, AI-based Recommendations, Wearable Data, Text-Based Aggregation, LoRA Fine-Tuning, Digital Health

CEUR-WS.org/Vol-4213/paper37.pdf

1. Introduction

Wearable sensors and smartwatches increasingly support preventive and personalized healthcare by continuously capturing real-world physiological and behavioral data and transforming them into individualized guidance on physical activity, diet, sleep, and mental well-being [1, 2]. However, the analysis of high-frequency wearable sensor signals—combined with data preparation requires substantial time, domain expertise, and computational resources, thereby significantly increasing implementation costs and limiting scalability [3, 4].

Over the past decade, personalized wellness applications have evolved from deterministic rule-based systems [5, 6] to data-driven and machine-learning-based approaches [7, 8]. More recently, large language models (LLMs) have emerged as a promising paradigm for health-related applications, enabling flexible, context-aware interaction and natural-language reasoning over complex and heterogeneous data sources [9, 10]. A rapid breakthroughs in LLMs leverages them to summarize wearable data streams and generate personalized recommendations for physical activity, nutrition, sleep, and psychological resilience.

Despite this progress, the potential of LLMs for direct numerical forecasting from wearable data remains insufficiently explored. Existing studies primarily focus on dialogue, explanation, or recommendation tasks, while systematic evaluations of text-based forecasting accuracy, computational efficiency,

IVUS 2025: Information Society and University Studies, May 15, 2025, Kaunas, Lithuania
*Research was funded by UAB Corner Case Technologies" Grant (Nr. 02-020-K-0032).
*Corresponding author.
†These authors contributed equally.
info@quantumwander.com (L. Petkevičius); povilas@spikeapi.com (P. Gudžius); ernestas.filatovas@mif.vu.lt (E. Filatovas)
0000-0003-2416-0431 (L. Petkevičius); 0000-0002-2611-3383 (P. Gudžius); 0000-0002-9329-6431 (E. Filatovas)
© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-06-12

and fine-tuning benefits are largely absent. In particular, it is unclear to what extent aggregated natural-language representations can retain predictive information from wearable time series and whether lightweight adaptation of open-weight LLMs can provide an effective accuracy–cost trade-off.

In this study, we present a unified statistical–semantic aggregation framework that transforms daily wearable-device time series into structured natural-language representations for text-to-regression forecasting and define a one-day-ahead benchmark in which next-day physiological targets are predicted solely from aggregated textual summaries of prior-day activity. Using multiple public and internal wearable datasets, we systematically evaluate open-weight LLMs under parameter-efficient fine-tuning, analyzing forecasting accuracy and computational efficiency and demonstrating that lightweight adaptation enables accurate and cost-effective wearable-data forecasting. The main contributions of this work are:

- A unified statistical–semantic aggregation pipeline that converts heterogeneous wearable time series into structured textual inputs for large language models.
- A text-only, one-day-ahead forecasting benchmark for evaluating predictive performance from natural-language summaries.
- An empirical comparison of fine-tuned open-weight LLMs, analyzing accuracy–efficiency trade-offs and the effect of parameter-efficient adaptation in wearable forecasting.

The remainder of this paper is organized as follows. Section 2 reviews related work on the use of large language models for health-related signal interpretation. Section 3 describes the proposed methodology and data processing pipeline. Section 4 presents the experimental setup and empirical results. Finally, Section 5 concludes the paper and outlines directions for future work.

2. Background and Related Work

This section reviews prior work on the use of large language models for wearable and health-related data analysis and provides background on the open-weight LLM families considered in this study.

Recent research [11] has investigated the intersection of large language models and wearable sensor data for health monitoring, behavior modeling, and personalized inference. Survey-level analyses highlight the potential of LLMs to model human activity and health status from wearable signals, while noting challenges related to data quality, scalability, and integration with traditional machine-learning pipelines [11]. Researchers has explored zero-shot and few-shot inference of physiological and metabolic states from multimodal wearable sensor data [12], with further studies demonstrating that the incorporation of richer contextual information improves predictive performance for health-related outcomes [13].

Several studies have focused on summarization and explanatory tasks, leveraging LLMs to generate natural-language descriptions of human activity from wearable time-series data [14] as well as combining statistical aggregation of sensor signals with LLM-based reasoning to infer wellness indicators such as stress levels and sleep quality [15]. The majority of existing research in this domain has primarily addressed interpretability, textual summarization, and classification objectives. In contrast, the application of text-based representations of wearable sensor data for numerical time-series forecasting remains markedly underexplored. To address this gap, we present work which evaluates text-to-regression forecasting from aggregated wearable summaries and systematically analyzes the accuracy-efficiency trade-offs of parameter-efficient adaptation for open-weight LLMs.

To assess the feasibility of training and adapting LLMs for health-related and physiological signal interpretation, this work focuses on widely adopted open-weight transformer architectures that differ in scale, domain specialization, and computational characteristics. *Gemma3* [16] is a family of decoder-only transformer models ranging from 1B to 27B parameters, designed with an emphasis on computational efficiency and multimodality. The architecture integrates a *SigLIP* visual encoder [17], supports long-context processing up to 128k tokens, and employs hybrid local–global attention mechanisms to balance performance and resource usage. Trained on large multilingual corpora with distillation-based

optimization, *Gemma3* models are widely used as general-purpose foundation models, particularly in resource-constrained or privacy-sensitive deployment scenarios. On standard medical question-answering benchmarks, base *Gemma3* models demonstrate moderate performance, with reported accuracies of 50.7% on MedQA, 45.4% on MedMCQA, and 68.4% on PubMedQA [18]. Prior studies indicate that domain adaptation or task-specific fine-tuning can substantially improve performance.

GPT-OSS denotes a family of open-weight models released in 2025, including *gpt-oss-20b* and *gpt-oss-120b*, which are based on a Mixture-of-Experts (MoE) architecture [19, 20]. By activating only a subset of experts per token, these models support very long context lengths (up to 128k tokens) while maintaining inference efficiency at scale. Following large-scale pre-training, *GPT-OSS* models undergo instruction tuning and reinforcement-learning-based post-training to enhance reasoning and tool-use capabilities. In healthcare-oriented evaluations, *GPT-OSS* models demonstrate strong performance in medical dialogue and decision-support tasks, achieving a reported accuracy of 57.6% on the HealthBench benchmark [20], although at higher computational cost compared to smaller models.

MedGemma models are domain-adapted variants of *Gemma3*, trained on large-scale medical text and image corpora [18, 21]. Available in 4B and 27B parameter configurations, these models are optimized for multimodal medical tasks such as clinical question answering, medical report generation, and diagnostic image interpretation. Empirical evaluations show that *MedGemma* substantially outperforms general-purpose models on medical benchmarks, achieving accuracies of 64.4–87.7% on MedQA, 55.7–74.2% on MedMCQA, and 73.4–76.8% on PubMedQA [18]. While these results highlight the benefits of domain-specific pre-training for clinical applications, they also raise questions about transferability to non-clinical wearable and lifestyle data. Table 1 provides an overview of the open-weight LLM families considered in this study, summarizing their scale, domain orientation, supported modalities, context length, and licensing.

Table 1
Open-weight large language model families considered in this study

Model	Provider	Domain	Modality	Par. count (bn)	Context length	License
Gemma3 [16]	Google DeepMind	General-purpose	Text, image	1–27	up to 128k	Gemma License (open weights)
MedGemma3 [21]	Google DeepMind	Medical / clinical	Text, image	4, 27	up to 128k	Health AI DF (model), Apache 2.0 (code)
GPT-OSS [19]	OpenAI	General-purpose	Text, code	20, 120	up to ~131k	Apache 2.0

3. Methodology

This section presents the datasets, preprocessing, forecasting benchmark, and model adaptation procedures used in this study.

3.1. Dataset

This study utilizes three widely used public wearable-device datasets and one internal dataset. All datasets include core physical activity and physiological signals—step count, heart rate, and energy expenditure—making them suitable for evaluating language models in health and wellness applications.

LifeSnaps [22] is a four-month multimodal dataset collected from 71 participants wearing *Fitbit Sense* devices. In addition to objective physiological signals, the dataset includes ecological momentary assessments, combining sensor data with subjective self-reports. Recorded modalities include step count, heart rate, burned calories, sleep stages, and additional physiological signals such as EDA, temperature, SpO_2, and VO_2 max.

The HUPA-UCM dataset [23] constitutes a clinically oriented multimodal recordings from 25 individuals with type 1 diabetes. Participants wore Fitbit Ionic smartwatches alongside continuous glucose

monitoring (CGM) systems for a minimum duration of two weeks. The dataset integrates metabolic and physiological time-series data, encompassing glucose concentration trajectories, step counts, heart rate measurements, calories burned, and sleep features.

The FitBit Fitness Tracker Data dataset [24] is a widely used open-access collection of physiological and behavioral measurements from 30 Fitbit users recorded over 30 days. It provides multi-resolution time-series data (daily, hourly, minute-, and second-level) including step counts, caloric expenditure, heart rate, activity intensity, sleep stages, and body mass metrics.

The Spike dataset, collected by Spike Technologies, includes multimodal recordings from 13 participants throughout 2025, capturing daily physical activity, environmental interactions, and smart-device usage. It comprises 359 hourly records with notable heterogeneity in sensor availability, reflecting real-world conditions, and includes all variables accessible via the *Spike API*[1].

3.2. Preprocessing and Standardization

3.2.1. Data Harmonization and Aggregation

The variations in data structure were standardized across datasets. All physical activity and physiological signals-including step counts, estimated energy expenditure (calories), and heart rate-were uniformly aggregated to an hourly resolution (1-hour epochs). The resulting standardized hourly time series sizes are reported in Table 2. After standardization, the final dataset comprises a total of 9 860 hourly records collected from 142 users across four datasets. This unified hourly dataset serves as the common experimental basis for all subsequent stages of the study.

Table 2
Hourly aggregated data volumes.

Dataset	User Count	Record Count
FitBit Fitness Tracker	34	1 893
HUPA-UCM Diabetes Dataset	25	1 068
LifeSnaps	71	6 589
Spike	12	310
Total	**142**	**1 893**

3.2.2. Data Profiling and Temporal Window Definition

Here we desrcibe the profiling procedure used to transform standardized wearable-device signals into unified, model-ready time-series representations. The methodology targets fixed 24-hour activity windows aggregated at an hourly resolution.

Data definition For each user $u \in U$, two time series are considered: hourly energy expenditure and step counts,

$$C_u = \{(t_{u,i}, c_{u,i})\}_{i=1}^{N_u}, \qquad S_u = \{(t_{u,i}, s_{u,i})\}_{i=1}^{N_u},$$

where $t_{u,i}$ denotes the timestamp, $c_{u,i}$ the calories burned, and $s_{u,i}$ the number of steps during hour i.

Only the first 24 hours from the initial observation are retained:

$$\tilde{C}_u = \{(t_{u,i}, c_{u,i}) \,:\, t_{u,i} \le t_{u,1} + 24h\}, \qquad \tilde{S}_u = \{(t_{u,i}, s_{u,i}) \,:\, t_{u,i} \le t_{u,1} + 24h\}.$$

Each user is thus represented by $n_u = \min(24, |\tilde{C}_u|)$ hourly observations.

Statistical profiling of activity signals For each signal $x \in \{c, s\}$, descriptive statistics are computed, including mean, variance, skewness, excess kurtosis, minimum and maximum values, and the proportion of missing observations.

[1]Spike API could be access spikeapi.com

Temporal characteristics and regularity Measurement regularity is assessed via the observed time span $\Delta t_u = \max_i t_{u,i} - \min_i t_{u,i}$ and the most frequent inter-sample interval $\delta_u^{\text{mode}} = \text{mode}(t_{u,i} - t_{u,i-1})$.

Outlier detection After removing missing values, outliers are identified using the interquartile range (IQR),

$$\text{IQR}_u = Q_{3,u} - Q_{1,u},$$

where $Q_{1,u}$ and $Q_{3,u}$ denote the 25th and 75th percentiles, respectively. Observations outside $[Q_{1,u} - 1.5\,\text{IQR}_u,\ Q_{3,u} + 1.5\,\text{IQR}_u]$ are flagged as outliers.

3.3. Textual Representations of Wearable Data

To enable consistent comparison of user behavior representations across language model inputs, the standardized and profiled wearable time-series data are converted into two complementary textual representations: a statistical description and an aggregated (semantic) description. Representative examples of both representations are shown in Table 3.

The statistical description is deterministic and directly derived from the underlying time series. Such format encodes key distributional and temporal features of the time series (extrema, central tendency, dispersion, outlier frequency, temporal coverage, and observation count). Its reproducibility and comparability make it a consistent textual baseline across heterogeneous datasets. In contrast, the aggregated semantic description provides a concise natural-language daily summary, highlighting dominant activity patterns or iregularities. While closer to the project's end goal of human-interpretable summaries, this representation introduces higher semantic variability and imposes stronger requirements on output consistency and linguistic stability.

The joint use of both description types enables systematic analysis of information retention under different textual encodings and supports downstream tasks such as forecasting benchmarks and model adaptation.

Table 3
Statistical description of user data and aggregated semantic description.

Statistical description

Statistic	Value
Number of observations	24
Min / Max	48.0 / 270.0
Mean	87.50
Median	59.00
Std. deviation	62.07
Variance	3852.35
Skewness	1.86
Kurtosis	2.35
Missing values	0 (0.00%)
Outliers (IQR)	3
Outlier values	189, 237, 270

Time characteristics

Start time	2016-03-13 00:00:00
End time	2016-03-13 23:00:00
Time span	23 hours
Most common interval	1 hour

Aggregated (semantic) description

- Daily energy expenditure estimated at 2000–2200 kcal
- Moderate physical activity level throughout the day
- Peak caloric burn around 9:00 am and 6:00 pm (post-breakfast and pre-dinner)
- Approximate daily breakdown:
  - Breakfast/active morning: 60% (1200–1400 kcal)
  - Lunch: 20% (400–500 kcal)
  - Snacks + dinner/evening: 20% (600–800 kcal)
- Outliers (189, 237, 270 kcal/h) likely correspond to intense workouts or strenuous activity bouts
- Morning hours show consistently higher energy expenditure, suggesting a more active morning routine

3.4. Daily Forecasting Benchmark

To assess the practical suitability of large language models for forecasting health-related outcomes from wearable-device data, a daily forecasting benchmark is constructed. The task consists of predicting a

numeric target for day $t+1$ from an aggregated natural-language description of day t (as defined in Section 3.3). This defines a one-day-ahead regression problem with textual input encoding the statistical and semantic characteristics of the preceding day.

Daily target aggregation Since the raw data are available at minute-, hour-, or mixed-level granularities, all signals are first standardized to daily aggregates. Let $X_{u,t}$ denote the set of observations for user u on day t. Daily targets are defined as:

$$x_{u,t}^{(\text{steps})} = \sum_{h=1}^{24} \text{steps}_{u,t,h}, \qquad x_{u,t}^{(\text{hr})} = \frac{1}{24} \sum_{h=1}^{24} \text{HR}_{u,t,h}, \qquad x_{u,t}^{(\text{cal})} = \sum_{h=1}^{24} \text{calories}_{u,t,h}.$$

These values constitute the ground-truth targets for next-day prediction. For each day t, an aggregated textual description $d_{u,t}$ is generated to summarize the user's activity and physiological patterns over that day.

The target for the subsequent day is defined as:

$$y_{u,t+1} \in \mathbb{R},$$

where $y_{u,t+1}$ denotes one of the following daily aggregates: total step count, mean heart rate, or total energy expenditure, for user u at time $t+1$. A parameterized forecasting model g_ϕ produces a numeric prediction from the textual description of the current day:

$$\hat{y}_{u,t+1} = g_\phi(d_{u,t}).$$

Performance is tracked using standard regression metrics (MAE, MSE, MAPE).

3.5. Model Selection and Adaptation

3.5.1. Selected Models and Adaptation Protocol

The experimental evaluation employs three open-weight LLMs: *Gemma3* [16], *Med-Gemma3* [21], and *GPT-OSS* [19]. Large language models are typically trained in two stages—large-scale pre-training on generic corpora, followed by task-specific adaptation. Here, adaptation focuses on aligning the models with a text-to-regression forecasting task under strict output constraints, requiring generation of a single numeric prediction.

To ensure stable training and consistent inference behavior, each training example is encoded using the exact chat template required by the corresponding model family. Preliminary experiments showed that even minor deviations from the prescribed templates can substantially affect output stability and numerical parsing, leading to unstable optimization or loss of output control. The exact chat templates used for *Gemma3*, *MedGemma3*, and *GPT-OSS* are summarized in Table 4.

3.5.2. Parameter-Efficient Fine-Tuning via LoRA

To enable efficient adaptation of large language models under limited computational and memory budgets, this study employs *parameter-efficient fine-tuning* (PEFT) rather than full model fine-tuning. PEFT methods keep the base model parameters fixed and train only a small set of additional parameters that capture task-specific adaptation. Among available PEFT approaches, Low-Rank Adaptation (LoRA) [25] is used due to its strong empirical performance and favorable efficiency–accuracy trade-off.

LoRA assumes that the task-induced weight update ΔW is approximately low-rank and can be decomposed as

$$\Delta W = BA, \qquad B \in \mathbb{R}^{d \times r},\ A \in \mathbb{R}^{r \times k},$$

where $r \ll \min(d, k)$. The adapted weight matrix is therefore given by

$$W' = W + BA.$$

Table 4
Instructions for Gemma3, MedGemma3, GPT-OSS models

Model	Exact Chat Template Tokens
Gemma3 *(text & vision)*	<bos><start _of_turn>user ··· <end_of_turn> <start _of_turn>model ··· <end_of_turn><eos>
MedGemma3 *(multimodal)*	Same as Gemma3, plus image placeholder: <bos><start _of_turn>user <image> ... <end_of_turn> <start _of_turn>model ··· <end_of_turn><eos>
ChatGPT / GPT	<start|><|system|><|message|>| ··· <|end|><|start|><|user|><|message|>| ··· <|end|><|start|><|assistant|><|message|>| ··· <|end|><|start|><|user|><|message|>| ··· <|end|><|start|><|assistant|><|channel|>final<|message|>| ··· <|end|>

In practice, LoRA adapters are injected into key transformer components [26], most notably the attention projection matrices (W_Q, W_K, W_V, W_O) and, where appropriate, selected MLP layers were fine-tuned.

Compared to full fine-tuning, LoRA substantially reduces GPU memory requirements by eliminating the need to store gradients and optimizer states for the full parameter set.

To enhance computational efficiency, Quantized LoRA (QLoRA) is adopted: the base model weights are stored in 4-bit quantized format, while LoRA adapters are trained in higher precision (FP16 or BF16).

4. Experimental Results and Model Comparison

This section evaluates various LLM configurations on the proposed text-to-regression forecasting task.

4.1. Experimental setup

The evaluation includes the model families *Gemma 3*, *Med-Gemma*, and *GPT-OSS*, with both base (pre-trained) and fine-tuned (`_f`) variants assessed to quantify the impact of parameter-efficient fine-tuning across scales.

Given textual activity summary $d_{u,t}$ for user u on day t, predict

$$\mathbf{y}_{u,t+1} = (y^{(\mathrm{HR})}_{u,t+1}, y^{(\mathrm{dist})}_{u,t+1}, y^{(\mathrm{cal})}_{u,t+1})$$

where $y^{(\mathrm{HR})}_{u,t+1}$ represents mean daily heart rate, $y^{(\mathrm{dist})}_{u,t+1}$ total daily distance (or step-equivalent activity), and $y^{(\mathrm{cal})}_{u,t+1}$ total daily energy expenditure.

Models were constrained to produce predictions in strict JSON format via structured outputs:

```
{
  "heartrate_mean": <float>,
  "steps_total": <float>,
  "calories_total": <float>
}
```

This enforces deterministic structured output, enabling direct error calculation and fair comparison. Experiments ran on an NVIDIA Spark GPU cluster providing ~1PFLOPS (FP4) compute and 128GB GPU memory per executor, supporting stable fine-tuning and inference of models up to 27B parameters.

4.2. Experimental Results

Each model configuration was evaluated on the aggregated dataset using MAPE for heart rate, distance, and calories, with heart-rate MAPE as the primary accuracy metric. Deployment metrics include model size, peak memory, inference throughput (tokens/s), and training/fine-tuning duration. Results are shown in Table 5.

As shown in Table 5, heart-rate prediction accuracy varies substantially across model scales and adaptation strategies. Lightweight models such as Gemma3 (270M) yield heart-rate MAPE values above 10% , indicating limited predictive capacity from compact architectures. In contrast, mid-sized models achieve markedly lower error: Gemma3 4B and 12B reach MAPE values of 7.46% and 6.77% in their base configurations, representing the strongest accuracy–efficiency trade-off observed in this study. Larger models do not consistently outperform these mid-scale configurations. While MedGemma 27B and GPT-OSS 20B achieve heart-rate MAPE values of 7.24% and 7.64%, respectively, these gains are marginal relative to the substantial increase in memory footprint and training cost. Parameter-efficient fine-tuning with LoRA leads to modest but consistent improvements across most models, typically reducing heart-rate MAPE by approximately 0.2–0.3 percentage points. For example, Gemma3 4B improves from 7.46% to 7.44%, and GPT-OSS 20B from 7.64% to 7.60%.

Overall, the results demonstrate that aggregated textual representations of wearable data enable single-digit heart-rate prediction error, and that mid-sized open-weight LLMs (4B–12B parameters) provide the most favorable balance between predictive accuracy and computational efficiency.

Table 5
Comparison of LLM configurations for text-based wearable-data forecasting.

Model	Params (bn)	RAM (GB)	GPU cost (€/hr)	MAPE (HR)	Tokens/s	Training + fine-tuning (hrs)	Cost (€/100k)
Base models							
Gemma3:270m	0.27	0.6	0.18	10.33	350	2.6	14.38
Gemma3:4b	4	4.8	0.18	7.46	75	10.8	66.76
Gemma3:12b	12	9.8	0.18	**6.77**	22	16.8	227.36
MedGemma:4b	4	4.8	0.18	9.89	60	11.2	83.42
MedGemma:27b	27	17.1	0.35	7.24	4	63.0	2430.73
GPT-OSS:20b	20	12.6	0.18	7.64	8	43.0	625.09
Fine-tuned models (LoRA)							
Gemma3:270m_f	0.27	0.6	0.18	10.30	350	4.4	14.38
Gemma3:4b_f	4	4.8	0.18	7.44	75	16.7	66.76
Gemma3:12b_f	12	9.8	0.18	6.79	22	25.7	227.36
MedGemma:4b_f	4	4.8	0.18	9.86	60	17.3	83.42
MedGemma:27b_f	27	17.1	0.35	7.26	4	95.0	2430.73
GPT-OSS:20b_f	20	12.6	0.18	7.60	8	65.0	625.09

Note. Models with the suffix *_f* are additionally fine-tuned using LoRA. Cost values correspond to processing 100 000 requests at the specified GPU hourly price.

5. Conclusions

This study shows that wearable-device time series can be transformed into complementary statistical and aggregated textual representations that retain sufficient information for numerical forecasting. Within a text-to-regression benchmark, parameter-efficient fine-tuning consistently improves predictive performance, demonstrating that lightweight adaptation is sufficient to align open-weight large language models with wearable forecasting tasks.

Experimental results indicate that mid-sized models (e.g. Gemma3:12b) offer the best accuracy–efficiency trade-off, achieving stable single-digit heart-rate prediction error at substantially lower computational cost than very large architectures. Domain-specialized and large-scale models do not consistently outperform these configurations, highlighting diminishing returns beyond a moderate

model scale when operating on text-only summaries.

Overall, the proposed aggregation framework combined with LoRA-based adaptation enables accurate and cost-effective LLM-driven forecasting, making it suitable for practical deployment in digital health analytics. Future work will extend the approach to additional wearable modalities, longer forecasting horizons, and alternative model families, as well as larger and more diverse datasets.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] J. Longhini, M. Peruzzini, et al., Wearable devices to improve physical activity and reduce sedentary behaviour in adults: A systematic review and meta-analysis, Sports Medicine - Open 10 (2024) 19. PMCID: PMC10788327.

[2] C. Köhler, colleagues, The value of smartwatches in the health care sector for patient monitoring and disease prevention: A review, Journal of Personalized Medicine 14 (2024) 607. PMCID: PMC11549588.

[3] G. S. Ginsburg, R. W. Picard, S. H. Friend, Key issues as wearable digital health technologies enter clinical care, New England Journal of Medicine 390 (2024) 1118–1127. doi:10.1056/NEJMra2307160.

[4] C. Doherty, M. Baldwin, R. Lambe, M. Altini, B. Caulfield, Privacy in consumer wearable technologies: A living systematic analysis of data policies across leading manufacturers, npj Digital Medicine 8 (2025) 1–11. doi:10.1038/s41746-025-01757-1.

[5] E. Kuhn, B. J. Weiss, K. L. Taylor, J. E. Hoffman, K. M. Ramsey, R. Manber, et al., Cbt-i coach: A description and clinician perceptions of a mobile app for cognitive behavioral therapy for insomnia, Journal of Clinical Sleep Medicine 12 (2016) 597–606. doi:10.5664/jcsm.5700.

[6] K. Stefanidis, et al., Ontology-driven personalised nutrition with protein ai advisor, Journal of Biomedical Semantics (2022).

[7] M. Zhou, Y. Fukuoka, Y. Mintz, K. Goldberg, P. Kaminsky, E. Flowers, A. Aswani, Evaluating machine learning–based automated personalized daily step goals delivered through a mobile phone app: Randomized controlled trial, JMIR mHealth and uHealth 6 (2018) e28. doi:10.2196/mhealth.9117.

[8] C. L. Trevenen, L. A. Walmsley, T. Manser, J. P. Devos, W. J. Wilkinson, Using hidden markov models with raw triaxial wrist accelerometry to determine sleep stages, Physiological Measurement 40 (2019) 034001. URL: https://doi.org/10.1088/1361-6579/ab03d3. doi:10.1088/1361-6579/ab03d3.

[9] M. Jörke, S. Sapkota, L. Warkenthien, N. Vainio, P. Schmiedmayer, E. Brunskill, J. A. Landay, GPTCoach: Towards LLM-based physical activity coaching, in: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI '25, Association for Computing Machinery, 2025. doi:10.1145/3706598.3713819.

[10] J. A. Logan, S. Sadhu, C. Hazlewood, M. Denton, S. E. Burke, C. A. Simone-Soule, C. Black, C. Ciaverelli, J. Stulb, H. Nourzadeh, Y. Vinogradskiy, et al., Bridging gaps in cancer care: Utilizing large language models for accessible dietary recommendations, Nutrients 17 (2025) 1176. doi:10.3390/nu17071176.

[11] E. Ferrara, Large language models for human activity and health monitoring from wearable sensors: A survey, Sensors 24 (2024) 5045. doi:10.3390/s24155045.

[12] A. Böhi, B. Gashi, Large language models for wearable sensor data analysis, arXiv preprint arXiv:2403.01234 (2024).

[13] J. Kim, M. Park, S. Lee, Health-llm: Large language models for health prediction from wearable sensor data, arXiv preprint arXiv:2401.06866 (2024).

[14] Y. Zhang, R. Chen, H. Li, Motionteller: Integrating wearable time series with large language models for activity understanding, arXiv preprint arXiv:2512.21506 (2025).
[15] L. Yu, A. Kumar, S. Thompson, Hybridsense: A multimodal llm framework for wellness estimation from wearable sensor data, IEEE Journal of Biomedical and Health Informatics (2026). Early access.
[16] G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière, et al., Gemma 3 technical report, arXiv preprint arXiv:2503.19786 (2025).
[17] X. Zhai, B. Mustafa, A. Kolesnikov, L. Beyer, Sigmoid loss for language image pre-training, in: Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 11975–11986. doi:10.1109/ICCV51070.2023.01100.
[18] A. Sellergren, S. Kazemzadeh, T. Jaroensri, A. Kiraly, M. Traverse, T. Kohlberger, S. Xu, F. Jamil, C. Hughes, C. Lau, et al., Medgemma technical report, arXiv preprint arXiv:2507.05201 (2025).
[19] OpenAI, Introducing gpt-oss: Open-weight models, OpenAI News (2025).
[20] S. Agarwal, L. Ahmad, J. Ai, S. Altman, A. Applebaum, E. Arbus, R. K. Arora, Y. Bai, B. Baker, H. Bao, et al., gpt-oss-120b & 20b model card, arXiv preprint arXiv:2508.10925 (2025).
[21] D. Golden, et al., Medgemma: Our most capable open models for health ai, Google AI Blog (2025).
[22] S. Yfantidou, C. Karagianni, S. Efstathiou, A. Vakali, J. Palotti, D. P. Giakatos, T. Marchioro, A. Kazlouski, E. Ferrari, Š. Girdzijauskas, Lifesnaps, a 4-month multi-modal dataset capturing unobtrusive snapshots of our lives in the wild, Scientific Data 9 (2022) 663.
[23] J. I. Hidalgo, J. Alvarado, M. Botella, A. Aramendi, J. M. Velasco, O. Garnica, Hupa-ucm diabetes dataset, Data in Brief 55 (2024) 110559.
[24] R. Furberg, J. Brinton, M. Keating, A. Ortiz, Fitbit fitness tracker data, https://www.kaggle.com/datasets/arashnic/fitbit, 2016. doi:10.34740/KAGGLE/DSV/108, crowdsourced from 30 Fitbit users via Amazon Mechanical Turk, collected 2016-04-12 to 2016-05-12.
[25] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, L. Wang, W. Chen, A. Raj, Lora: Low-rank adaptation of large language models, arXiv preprint arXiv:2106.09685 (2021).
[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).