BENCHMARKING LLM SUMMARIES OF MULTIMODAL CLINICAL TIME SERIES FOR REMOTE MONITORING

Aditya Shukla[1]*, Yining Yuan [1]*, J. Ben Tamo[1] Yifei Wang[1] Micky C. Nnamdi[1] Benoit L. Marteau[1] Shaun Tan[1] Jieru Li[1] Brad Willingham[2] May D. Wang[1] †

[1]Georgia Institute of Technology **[2]Shepherd Center**

ABSTRACT

Large language models (LLMs) can generate fluent clinical summaries of remote therapeutic monitoring time series, yet it remains unclear whether these narratives faithfully capture clinically significant events such as sustained abnormalities. Existing evaluation metrics emphasize semantic similarity and linguistic quality, leaving event-level correctness largely unmeasured. We introduce an event-based evaluation framework for multimodal time-series summarization using the technology-integrated health management (TIHM)-1.5 dementia monitoring data. Clinically grounded daily events are derived via rule-based abnormal thresholds and temporal persistence, and model-generated summaries are aligned to these structured facts. Our protocol measures abnormality recall, duration recall, measurement coverage, and hallucinated event mentions. Benchmarking zero-shot, statistical prompting, and vision-based pipeline using rendered time-series visualizations reveals a striking decoupling: models with high conventional scores often exhibit near-zero abnormality recall, while the vision-based approach achieves the strongest event alignment (45.7% abnormality recall; 100% duration recall). These results highlight the need for event-aware evaluation to ensure reliable clinical time-series summarization.

Track: Research

1 INTRODUCTION

Home-based remote therapeutic monitoring (RTM) is increasingly used to support people living with dementia, combining physiological devices (e.g., blood-pressure cuffs, scales) with ambient sensors, such as bed, motion, and door sensors. Large deployments such as technology-integrated health management (TIHM)-1.5 collect continuous multimodal time series that can reveal early deterioration and support proactive care (Palermo et al., 2023). However, in current clinical workflows, practitioners must manually synthesize these data across disparate plots and alert logs, mentally integrating trends across multiple modalities and temporal scales, a process that is both cognitively demanding and prone to oversight.

Language models have introduced powerful new primitives for time-series analysis. Foundational models such as Lag-Llama, Chronos, and TimesFM have set new benchmarks for zero-shot numerical forecasting by capturing deep temporal dependencies (Rasul et al., 2023; Ansari et al., 2024; Das et al., 2024). Simultaneously, general-purpose Large Language Models (LLMs) like GPT-4o have demonstrated sophisticated reasoning and summarization capabilities (Achiam et al., 2023). Despite these gains, a critical validation gap remains: while these models excel at capturing statistical patterns or following linguistic instructions, their ability to generate clinically faithful prose, grounded in raw sensor evidence, remains unproven.

Current evaluation paradigms for LLM summarization typically rely on semantic overlap (e.g., ROUGE, BERTScore, Moverscore) or Natural Language Inference (NLI)-based consistency metrics such as SummaC or AlignScore (Lin, 2004; Zhang et al., 2019; Zhao et al., 2019; Laban et al.,

*Equal contribution.
†Correspondence to maywang@gatech.edu

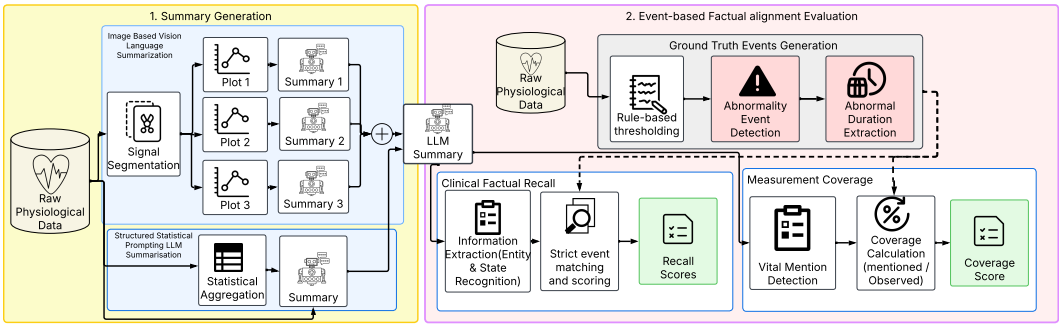

Figure 1: **Overview of the proposed event-based evaluation framework for clinical time-series summarization. Left:** Summary generation pipelines, including (1) image-based vision–language summarization via signal segmentation and plot rendering, and (2) structured statistical prompting with aggregated numerical features. Both pipelines generate daily LLM summaries from raw physiological data. **Right:** Event-based factual alignment evaluation. Ground-truth clinical events are derived through rule-based thresholding to detect abnormalities and sustained durations. Generated summaries are evaluated using strict event matching (clinical factual recall) and measurement coverage analysis, producing recall and coverage scores that quantify event-level correctness beyond reference-based semantic similarity metrics.

2022; Zha et al., 2023). However, these metrics prioritize linguistic fluency and topical similarity over the precise, threshold-based factual accuracy required in medical monitoring. In the context of dementia care, the omission of a subtle "sustained deviation" (e.g., a patient remaining in bed for an abnormal duration) is a catastrophic failure that standard NLP metrics are poorly equipped to detect.

To bridge this gap, we introduce an event-based evaluation framework designed to measure clinical faithfulness in multimodal time-series narratives. Leveraging the TIHM-1.5 dataset, we construct a rule-grounded evaluation system that extracts structured clinical events, such as abnormal vital sign means and sustained behavioral deviations, to serve as an immutable ground truth. We then assess LLM performance through the lens of event-level factual recall, duration accuracy, and modality coverage, allowing for a rigorous decomposition of omissions versus hallucinations.

Main contributions. This work (1) introduces an event-based evaluation framework for multimodal clinical time-series summarization, measuring abnormality recall, duration faithfulness, and modality coverage, (2) demonstrates a striking decoupling between conventional summarization metrics and event-level clinical correctness on TIHM-1.5 dementia monitoring data, and (3) benchmarks prompting- and visualization-grounded multimodal pipelines, showing that summaries based on rendered time-series plots achieve stronger alignment with underlying physiological events.

2 Methodology

2.1 Dataset and Problem Setup

TIHM-1.5 (Palermo et al., 2023) is a public remote monitoring dataset covering 56 people living with dementia across 2,803 patient-days, recording physiology (heart rate, blood pressure, temperature), room-level activity, and sleep staging with cardiorespiratory signals. We define each patient-day as the unit of analysis:

$$\mathcal{D}_{i,t} = \{x_{i,t}^{(m)}\}_{m=1}^{M}, \tag{1}$$

where $x_{i,t}^{(m)}$ denotes timestamped observations from modality m. A stratified test set of N=100 patient-days is held out for evaluation; remaining days are reserved for prompt development. Given $\mathcal{D}_{i,t}$, the goal is to generate a clinically grounded narrative $y_{i,t}$; pipelines differ only in how the time series is represented to the model.

2.2 Summarization Pipelines

Raw signal prompting. The LLM receives a direct textual serialization of $\mathcal{D}_{i,t}$ and must perform all numerical reasoning internally.

Statistical conditioning. Because LLMs struggle with long floating-point sequences, we also condition on a compact summary vector $S_{i,t} = f(\mathcal{D}_{i,t})$, where $f(\cdot)$ extracts per-modality descriptive statistics (mean, min, max, std) and discrete abnormality indicators.

Visualization grounding. Each patient-day is rendered into clinical-style plots $I_{i,t} = r(\mathcal{D}_{i,t})$, annotated with thresholds and sleep-stage context. A vision-language model generates $y_{i,t}$ from $I_{i,t}$, reporting only visually supported events.

Across all settings, the output is a structured daily narrative $y_{i,t}$, conditioned on $\mathcal{D}_{i,t}$, $S_{i,t}$, or $I_{i,t}$ respectively.

2.3 Ground-Truth Fact Extraction

We derive clinical event facts from established clinical standard vital-sign reference ranges without manual annotation. For each patient-day, we instantiate an *abnormality fact* when any vital falls outside predefined bounds (e.g., HR $<$50 or $>$90 bpm) and a *duration fact* when out-of-range values persist for $\geq\Delta$=30 minutes. Each patient-day yields a structured fact set

$$F_{i,t} = \{f_k\}_{k=1}^{K}, \tag{2}$$

encoding (*vital*, *type*, *direction*, *value/duration*), together with a record of modality availability.

2.4 Event-Based Evaluation

Generated summaries are aligned to $F_{i,t}$ using a conservative mention policy requiring co-occurrence of the correct vital name and an explicit abnormality indicator (e.g., "high blood pressure," "outside normal range"). We report two complementary metrics:

$$\text{Recall}(y_{i,t}) = \frac{|\{f \in F_{i,t} : f \text{ is mentioned in } y_{i,t}\}|}{|F_{i,t}|}, \tag{3}$$

$$\text{Coverage}(y_{i,t}) = \frac{|\{m : x_{i,t}^{(m)} \text{ observed and mentioned}\}|}{|\{m : x_{i,t}^{(m)} \text{ observed}\}|}. \tag{4}$$

Recall quantifies omission of clinically significant events; coverage measures whether all observed modalities are acknowledged. Claims matching no fact in $F_{i,t}$ are flagged as hallucinations.

3 Experiments and Results

3.1 Experimental Setup

We evaluate generated summaries along three complementary dimensions: event-level factual alignment, semantic faithfulness, and perceived clinical clarity.

Event-Level Factual Alignment: We decompose clinical faithfulness into three granular metrics: (i) *Abnormality Recall*, measuring the retrieval of point-wise threshold violations; (ii) *Duration Recall*, measuring the identification of sustained deviations (persisting $>$ 30 mins); and (iii) *Measurement Coverage*, quantifying the model's acknowledgment of available data streams versus hallucinated omissions.

Semantic faithfulness. We further evaluate reference-free consistency using AlignScore (Zha et al., 2023) and SummaC (Laban et al., 2022), treating the structured tabular cues as the source document and the generated summary as the hypothesis.

LLM-as-judge clarity. We utilize GPT-4o-mini as a proxy evaluator, scoring summaries on a 1–5 Likert scale for coherence, readability, and professional tone.

Table 1: Event-level clinical correctness diverges from conventional NLP metrics. Statistical conditioning improves grounding, and the visualization-grounded pipeline achieves the strongest event alignment despite lower NLP scores.

Method	**Backbone**	**Clinical Event Metrics - Ours** (%)			**Standard NLP Metrics** (↑)		
		Abnormality	Duration	Coverage	SummaC	Align	Clarity
Zero-shot	Llama-3.2-3B	0.00	0.00	46.44	0.59	0.32	3.27
	Llama-3-8B	0.00	0.00	41.40	0.57	0.31	**4.00**
	Gemma-3-12B	0.00	0.00	31.50	0.28	0.39	1.70
	Gemma-3-27B	2.10	0.00	15.60	0.55	**0.44**	1.44
Stat Based	Llama-3.2-3B	18.80	0.00	99.32	0.44	0.23	3.38
	Llama-3-8B	33.30	40.00	94.20	0.64	0.28	**4.00**
	Gemma-3-12B	27.10	20.00	56.61	**0.65**	0.39	1.87
	Gemma-3-27B	36.70	60.00	15.60	0.55	0.39	1.79
Image-Based	Gemini-2.5-Pro	**45.70**	**100.00**	**100.00**	0.38	0.16	2.71

3.2 Results

Table 1 reveals a striking decoupling between conventional summarization metrics and event-level clinical correctness. Zero-shot text-only models achieve moderate semantic faithfulness and high clarity scores, yet fail almost entirely to report clinically significant events, with near-zero abnormality recall (0.0–2.1%) and no detected duration abnormalities. In contrast, statistical conditioning substantially improves grounding: Llama-3-8B rises from 0.0% to 33.3% abnormality recall and reaches 94.2% measurement coverage, suggesting that quantitative signal extraction is a major bottleneck for text-based LLMs. The visualization-grounded Gemini-2.5-Pro pipeline achieves the strongest event alignment, with 45.7% abnormality recall and perfect 100% duration recall and coverage, despite lower SummaC and clarity scores. Overall, these results demonstrate that standard NLP metrics can substantially overestimate clinical reliability, whereas event-based evaluation directly exposes omission and grounding failures in RTM summarization.

4 Discussion and Conclusion

Our results reveal three central insights about multimodal clinical time-series summarization.

The "Fluency Illusion": standard metrics can mask clinically important omission. Zero-shot text-only models achieve moderate semantic scores and relatively strong clarity ratings, yet exhibit near-zero abnormality recall (0.0–2.1%) and no duration detection. This shows that summaries can appear fluent and topically aligned while still failing to surface the events most relevant for remote therapeutic monitoring. In safety-critical settings such as dementia care, omission of a sustained physiological abnormality may be more consequential than minor stylistic weakness, but conventional summarization metrics do not reliably expose this failure mode.

Statistical conditioning mitigates, but does not resolve, numerical grounding limitations. Providing threshold-aware summary statistics substantially improves factual alignment across text-only backbones. For example, Llama-3-8B improves from 0.0% to 33.3% abnormality recall and from 41.4% to 94.2% measurement coverage. This suggests that a major bottleneck in raw prompting is the difficulty of aggregating long, irregular numerical sequences into stable clinical statements. However, even the best text-based prompted systems still miss a substantial fraction of rule-defined abnormalities and sustained events, indicating that prompt scaffolding alone cannot guarantee dependable event completeness.

Visualization-grounded summarization yields the strongest event sensitivity, but the source of that gain should be interpreted carefully. The image-based Gemini-2.5-Pro pipeline achieves the highest abnormality recall (45.7%), perfect duration recall (100%), and complete measurement coverage (100%), despite lower conventional semantic scores. This suggests that visual grounding may improve sensitivity to sustained physiological deviations relative to purely textual conditioning. However, this should not be interpreted as a clean modality-only advantage, since the image pipeline uses a stronger frontier model than the text-only baselines. We therefore view the result as

evidence of a promising *pipeline-level* strategy rather than definitive proof that visualization alone is responsible for the observed gains.

Event-based evaluation complements, rather than replaces, semantic faithfulness metrics. We do not argue that metrics such as SummaC and AlignScore are uninformative; rather, our experiments show that they are insufficient on their own for high-stakes time-series summarization. Semantic consistency remains useful for judging whether a narrative broadly agrees with its source, but it does not adequately measure whether threshold-defined abnormalities, sustained deviations, and observed modalities are actually preserved. Our event-based protocol adds this missing axis by grounding evaluation in explicit clinical facts derived from the time series itself.

The framework is transparent, but presently limited to recall-oriented event extraction. A core strength of our approach is that the evaluation target is explicit and reproducible: abnormality and duration facts are derived from auditable threshold rules rather than subjective manual labels. At the same time, the current implementation is recall-centric. Because our present matcher detects whether gold events are mentioned, but does not yet comprehensively enumerate unsupported predicted event claims, we cannot reliably compute full false-positive-based metrics such as precision and F1 for the current experiments. This is an important limitation, since a system could in principle increase recall by over-predicting abnormalities. We therefore interpret our results as characterizing *event omission sensitivity* rather than a complete precision–recall tradeoff.

Generalizability depends on portable clinical rules, not on the specific TIHM variables alone. Our ground-truth library is built from threshold-based abnormality detection and persistence rules tailored to the vital streams available in TIHM-1.5, so the framework is not automatically plug-and-play for arbitrary time-series domains. However, the broader methodology is portable: for a new domain, one can define clinically meaningful event rules, derive structured fact sets, and evaluate summaries according to event alignment. In this sense, what is domain-specific is the *event definition layer*, not the overall evaluation paradigm.

Clinical validity remains bounded by the rule set. Our abnormality definitions are based on population-level reference ranges and a fixed persistence threshold, which provides clarity and reproducibility but does not capture all patient-specific context. In real deployments, clinically meaningful interpretation may depend on comorbidities, medications, baseline physiology, and sensor quality. Thus, the benchmark should be understood as evaluating whether summaries preserve a transparent operationalization of salient events, rather than whether they fully match clinician judgment in every case.

Taken together, these findings validate the need for event-aware evaluation in clinical time-series summarization. Fluency and general semantic consistency can substantially overestimate clinical reliability, while rule-grounded event alignment reveals omission and mischaracterization errors that are directly relevant to patient safety. Future work should prioritize richer event libraries, explicit false-positive accounting, model-matched ablations across modalities, and clinically validated rule sets as prerequisites for trustworthy deployment.

Visualization-grounded reasoning yields the strongest event-level alignment. The image-based Gemini-2.5-Pro pipeline achieves the highest abnormality recall (45.7%), perfect duration recall (100%), and complete modality coverage (100%). Interestingly, these gains occur alongside lower conventional semantic scores. This decoupling suggests that visual grounding may promote more faithful detection of sustained physiological deviations, even when linguistic polish is reduced. For multimodal monitoring, perceptual grounding appears to support event sensitivity more effectively than purely textual conditioning.

REFERENCES

Salar Abbaspourazad, Oussama Elachqar, Andrew C Miller, Saba Emrani, Udhyakumar Nallasamy, and Ian Shapiro. Large-scale training of foundation models for wearable biosignals. *arXiv preprint arXiv:2312.05409*, 2023.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Abdul Fatir Ansari, Lorenzo Stella, Ali Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Bernie Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=gerNCVqqtR. Expert Certification.

Brinnae Bent, Benjamin A Goldstein, Warren A Kibbe, and Jessilyn P Dunn. Investigating sources of inaccuracy in wearable optical heart rate sensors. *NPJ digital medicine*, 3(1):18, 2020.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Forty-first International Conference on Machine Learning*, 2024.

David Fraile Navarro, Enrico Coiera, Thomas W Hambly, Zoe Triplett, Nahyan Asif, Anindya Susanto, Anamika Chowdhury, Amaya Azcoaga Lorenzo, Mark Dras, and Shlomo Berkovsky. Expert evaluation of large language models for clinical dialogue summarization. *Scientific reports*, 15(1):1195, 2025.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *International conference on machine learning*, pp. 2151–2159. PMLR, 2019.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Benjamin Kompa, Jasper Snoek, and Andrew L Beam. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digital Medicine*, 4(1):4, 2021.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177, 2022. doi: 10.1162/tacl_a_00453. URL https://aclanthology.org/2022.tacl-1.10/.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Kaden McKeen, Sameer Masood, Augustin Toma, Barry Rubin, and Bo Wang. Ecg-fm: An open electrocardiogram foundation model. *Jamia Open*, 8(5):ooaf122, 2025.

Jungwoo Oh, Gyubok Lee, Seongsu Bae, Joon-myoung Kwon, and Edward Choi. Ecg-qa: A comprehensive question answering dataset combined with electrocardiogram. *Advances in Neural Information Processing Systems*, 36:66277–66288, 2023.

Christina Orphanidou, Timothy Bonnici, Peter Charlton, David Clifton, David Vallance, and Lionel Tarassenko. Signal-quality indices for the electrocardiogram and photoplethysmogram: Derivation and applications to wireless monitoring. *IEEE journal of biomedical and health informatics*, 19(3):832–838, 2014.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

Francesca Palermo, Yu Chen, Alexander Capstick, Nan Fletcher-Loyd, Chloe Walsh, Samaneh Kouchaki, Jessica True, Olga Balazikova, Eyal Soreq, Gregory Scott, et al. Tihm: An open dataset for remote healthcare monitoring in dementia. *Scientific data*, 10(1):606, 2023.

Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Arian Khorasani, George Adamopoulos, Rishika Bhagwatkar, Marin Biloš, Hena Ghonia, Nadhir Hassen, Anderson Schneider, et al. Lag-llama: Towards foundation models for time series forecasting. In *R0-FoMo: Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*, 2023.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

Anna Shcherbina, C Mikael Mattsson, Daryl Waggott, Heidi Salisbury, Jeffrey W Christle, Trevor Hastie, Matthew T Wheeler, and Euan A Ashley. Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort. *Journal of personalized medicine*, 7(2):3, 2017.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerova, et al. Clinical text summarization: adapting large language models can outperform human experts. *Research Square*, pp. rs–3, 2023.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine*, 30(4):1134–1142, 2024.

Yuriy Vasilev, Irina Raznitsyna, Anastasia Pamova, Tikhon Burtsev, Tatiana Bobrovskaya, Pavel Kosov, Anton Vladzymyrskyy, Olga Omelyanskaya, and Kirill Arzamasov. Evaluating medical text summaries using automatic evaluation metrics and llm-as-a-judge approach: A pilot study. *Diagnostics*, 16(1):3, 2025.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. AlignScore: Evaluating factual consistency with a unified alignment function. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11328–11348, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.634. URL https://aclanthology.org/2023.acl-long.634/.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 563–578, 2019.

A Appendix

A.1 LLM Usage

Large language models (LLMs) were used only as general-purpose writing aids to improve the clarity and presentation of the manuscript. In particular, LLMs assisted with grammar, style, and suggested alternative phrasings for prompt instructions. All scientific contributions, including the research ideas, experimental design, and core arguments, were developed by the authors, and all factual claims were independently verified.

A.2 Dataset Acknowledgement

This study uses the TIHM (Technology Integrated Health Management) dataset for remote healthcare monitoring in dementiaPalermo et al. (2023). We acknowledge the Surrey and Borders Partnership NHS Foundation Trust for providing access to the dataset and for supporting its use in research.

A.3 Related Works

Time-series foundation models and remote monitoring. Recent time-series foundation models such as Lag-Llama, Chronos, and TimesFM learn generic temporal representations that transfer across forecasting domains (Rasul et al., 2023; Ansari et al., 2024; Das et al., 2024). In parallel, the TIHM project has demonstrated that dense in-home monitoring for people living with dementia can support early detection of deterioration through multimodal physiological and behavioural signals (Palermo et al., 2023). Our work sits at the intersection: rather than proposing a new forecasting backbone, we treat TIHM-1.5 as a testbed for evaluating how well existing LLMs and vision–language models can turn RTM time series into clinically faithful daily narratives.

Clinical Summarization with Large Language Models. Large language models have recently been applied to clinical summarization tasks, including discharge summaries, radiology reports, and longitudinal EHR compression (Van Veen et al., 2023; 2024). While these systems demonstrate promising linguistic fluency and task-specific performance, evaluation typically relies on ROUGE-style overlap metrics or clinician preference studiesFraile Navarro et al. (2025); Vasilev et al. (2025). In contrast, our setting involves multimodal physiological time series rather than free-text EHR inputs, and we focus explicitly on whether generated summaries capture threshold-defined abnormalities and sustained deviations — properties that are not directly measured by conventional summarization benchmarks.

Faithfulness and physiological reasoning. A large body of work has highlighted gaps between surface-level summarization quality and factual consistency, leading to metrics such as SummaC and AlignScore that compare generated text against source documents using learned entailment or alignment models (Laban et al., 2022; Zha et al., 2023). However, these metrics operate over unstructured text and do not directly assess whether specific physiological events are correctly captured. Our event-based evaluation follows this spirit in the RTM setting: by grounding abnormalities and sustained deviations in transparent rules over time series, we expose omission and hallucination errors that are largely invisible to standard semantic faithfulness scores. A further challenge in remote monitoring is that physiological streams are affected by artifacts, missingness, and device-dependent noise (Shcherbina et al., 2017; Bent et al., 2020). Prior work has proposed signal-quality indices for ECG/PPG to quantify reliability in ambulatory monitoring (Orphanidou et al., 2014), and has empirically analyzed systematic error sources in wearable optical heart-rate sensing (Bent et al., 2020). Because such imperfections can cause both false alarms and missed abnormalities, our event-based protocol is intentionally conservative: it evaluates whether summaries explicitly capture threshold-defined abnormalities and sustained deviations, while separately measuring modality coverage to expose omissions under realistic sensing conditions (Kompa et al., 2021). More broadly, high-stakes clinical AI motivates evaluation protocols that are transparent and auditable rather than purely similarity-based. Interpretability arguments in such settings emphasize methods that can be scrutinized and verified (Rudin, 2019; Guo et al., 2017; Ovadia et al., 2019; Geifman & El-Yaniv, 2019). Our rule-grounded event library and strict matching criteria follow this spirit by making the evaluation target explicit and reproducible (Wachter et al., 2017).

Physiological waveform and wearable foundation models. Beyond generic time-series forecasting, recent work has trained large self-supervised foundation models directly on physiological biosignals such as ECG and PPG. For example, ECG-FM provides an open foundation model for electrocardiograms, enabling transfer across downstream clinical ECG tasks (McKeen et al., 2025). Complementarily, large-scale pretraining on consumer wearable biosignals has been demonstrated using hundreds of thousands of participants' PPG and ECG recordings, showing that representations learned without labels can encode clinically relevant attributes and conditions (Abbaspourazad et al., 2023). Datasets that pair ECG signals with question-answering supervision (rather than only diagnostic labels) have also been proposed as a step toward more structured reasoning interfaces (Oh et al., 2023). These efforts are closely related to our setting in that they target physiological sensing at scale; however, our focus is not representation learning, but *clinically faithful narrative summarization* and *event-level verifiability* from daily remote-monitoring time series.

A.4 Prompt Templates

Prompt: Zero-Shot

You are a clinical AI assistant specializing in remote monitoring for dementia patients. Your task is to provide a concise, clinically relevant summary of a patient's raw time-series data for a doctor.

Patient data block:

```
PATIENT DATA FOR {target_date.strftime('%Y-%m-%d')}:
---
{structured_patient_text}
---
```

Instructions: You MUST adhere strictly to the data provided in the "PATIENT DATA" section. Your primary goal is factuality. Based only on the raw time-series data provided, generate a summary that addresses the following:

1. **Grounding:** Every statement you make MUST be directly supported by a data point in the provided text. Do not infer trends from single data points. Do not add information that is not present.
2. **Handling Missing Data:** If a category like "Sleep Patterns" is missing from the input, you MUST state: "No data available for this category."
3. **Overall Status:** Provide a one-sentence overview of the patient's day.
4. **Physiological Analysis:** Analyze the time-series vitals. Note any trends, stability, or significant spikes/dips throughout the day.
5. **Behavioral Analysis:** Analyze the activity and sleep logs. Describe the patient's routine (e.g., when they were active, when they slept) and sleep quality.
6. **Clinically Significant Events:** If a labeled event is present, you MUST highlight it and try to correlate it with the sensor data.

Format the output as a clean, bulleted list. Be specific and refer to times if necessary.

Prompt: Statistical

You are a clinical summarization assistant trained to generate safe, factual, and structured remote monitoring reports for elderly patients with dementia. You operate under clinical supervision and your outputs will be evaluated by physicians for factual accuracy, actionability, and clarity.

Patient data block:

```
PATIENT DATA FOR {target_date}:
---
{structured_patient_text}
---
```

Instructions. You MUST only use and infer information explicitly present in the *PATIENT DATA* section. You are NOT allowed to fabricate, generalize, or assume any patterns without specific supporting data. You must flag any uncertainty or data absence explicitly.

Format your output using the following structure:

OVERALL STATUS

- One-sentence overview of the patient's day.

PHYSIOLOGICAL ANALYSIS
For each vital sign (Heart Rate, Systolic/Diastolic Blood Pressure, Body Temperature), perform:

1. **Abnormality Check:** Compare average and peak values to the ranges below. If outside range, state as "Abnormally High" or "Abnormally Low" and include specific values:
   - Heart Rate: 50–90 bpm
   - Systolic BP: 90–140 mmHg
   - Diastolic BP: 60–90 mmHg
   - Temperature: 35.0–37.5 °C
2. **Trend Analysis:** Identify any clear increasing or decreasing trends over several hours, supported by multiple timestamps. Flag uncertain or noisy data.
3. **Duration Analysis:** If abnormalities were sustained over consecutive readings (e.g. >30 minutes), report duration and time range.

BEHAVIORAL ANALYSIS

- Summarize daily activity patterns and sleep data.
- Include periods of peak movement or long inactivity.
- For sleep, report total duration and breakdown by sleep stage (if available).
- Flag any missing data explicitly (e.g. "No data available for sleep patterns").

CLINICALLY SIGNIFICANT EVENTS

- If labeled events are present (e.g., Agitation, Fall), describe timing and attempt correlation with physiology or behavior.

Ensure all bullet points are supported by timestamped data. Do not infer anything not backed by the provided input.

Prompt: Vision Based

Prompt: Vision-Based Clinical Summarization
You are a clinical summarization assistant trained to generate safe, factual, and structured remote monitoring reports for elderly patients with dementia. Your outputs are evaluated for factual accuracy, actionability, and clarity.
The provided image contains a single vital sign time series for {`signal`}. You must report only what is explicitly visible in the image.
You MUST use only information directly supported by the visual data. Do not fabricate, generalize, or assume patterns without clear supporting evidence. Explicitly state any uncertainty or missing information.
Your response MUST follow this structure exactly:
OVERALL STATUS Provide a one-sentence overview of the patient's day based only on this {`signal`} plot.
PHYSIOLOGICAL ANALYSIS Perform:

- Abnormality check relative to clinical thresholds.
- Trend analysis across multiple timestamps.
- Duration analysis for sustained abnormalities (>30 minutes).

For abnormality reporting, you MUST include the vital name exactly as written: "{`signal`}". You MUST use one of the following exact templates:

- "{`signal`} was Abnormally High (value: X.X)."
- "{`signal`} was Abnormally Low (value: X.X)."
- "{`signal`} was within normal range."

Clinical reference ranges:

- Heart Rate: 50–90 bpm
- Systolic BP: 90–140 mmHg
- Diastolic BP: 60–90 mmHg
- Temperature: 35.0–37.5 °C

BEHAVIORAL ANALYSIS If activity or sleep context is visible, summarize it. Otherwise state: "No data available for activity/sleep from this image."
CLINICALLY SIGNIFICANT EVENTS If labeled events are visible, describe timing and correlation. Otherwise state: "No labeled events visible."
Every statement must be grounded in timestamped data visible in the image. Do not infer anything not supported by the provided visual evidence.

A.5 Case Study: Omission/misclassification of systolic BP abnormality

We present a case that illustrates the type of safety-critical failure our evaluation framework is designed to detect: the omission and misclassification of a clinically significant abnormality despite a fluent and confident narrative in the LLM-generated summary. On this patient day, our event-based ground truth identifies an abnormal systolic blood pressure (SBP) of 177 mmHg, which exceeds the predefined threshold of 140 mmHg. However, the LLM summary states that "Blood pressure remained within normal limits," directly contradicting the physiological evidence.

This error reflects both omission, because the elevated SBP is not explicitly mentioned, and misclassification, because the summary provides a reassuring statement that contradicts the underlying data. In clinical practice, such an inaccurate summary could obscure a clinically relevant signal and potentially delay appropriate patient intervention.

Notably, the LLM-generated summary remains linguistically coherent and topically aligned with the concept of blood pressure, meaning that conventional similarity-based evaluation metrics could still rate the summary as acceptable. By explicitly aligning narrative claims with transparent, threshold-defined clinical events, our event-based evaluation framework makes this failure detectable. This example demonstrates that the omission of true abnormalities—rather than overt hallucination—may represent the primary safety risk in RTM time-series summarization.

Table 2: **Case study: omission/misclassification of systolic BP abnormality.** Our event-based evaluation flags this day as *Abnormally High* SBP, while the text-only LLM summary incorrectly reports normal blood pressure.

Patient: 8a835 **Date:** 2019-04-20

Ground-truth fact (rule-derived):

- **Systolic BP = 177 mmHg** (**Abnormally High**; threshold $>$ 140 mmHg)

LLM summary (excerpt):

"Blood pressure remained within normal limits."

Evaluation outcome (ours):

- **Abnormality recall for SBP:** Missed (no explicit mention of high/elevated SBP)
- **Failure type: Omission / misclassification** (reassuring statement contradicts rule-derived abnormality)

Why this matters:

- Conventional fluency/faithfulness metrics can remain high even when the summary *fails to surface a clinically significant event*.
- This illustrates the primary reliability gap our evaluation captures: **event-level omission not penalized by generic text metrics**.

A.6 Case Study: Mischaracterization of systolic BP abnormality (wrong magnitude/timing)

We present a case that illustrates a safety-relevant failure mode our evaluation framework is designed to detect: a clinically meaningful abnormality is mentioned, but its severity and temporal localization are incorrect, creating a misleading clinical picture despite a fluent narrative.

For patient d7a46 on 2019-06-11, the raw physiology contains multiple systolic blood pressure (SBP) readings far above the predefined threshold of 140 mmHg. In particular, SBP reaches 188 mmHg at 14:36 and remains in a markedly elevated range (approximately 173–183 mmHg) across several measurements between 14:36 and 17:55. However, the LLM-generated summary reports a lower peak SBP of 150 mmHg and assigns the abnormal episode to a different time window (13:00–13:30). This is not an omission of the concept of "high BP," but rather a mischaracterization of the event's magnitude and timing.

In clinical practice, underestimating the peak SBP and shifting the episode to an incorrect time interval can affect downstream interpretation, including severity assessment, correlation with activity or medication, and escalation decisions. This example motivates event-level evaluation that verifies not only whether an abnormality was mentioned, but also whether the reported values and time intervals accurately match the underlying time-series evidence.

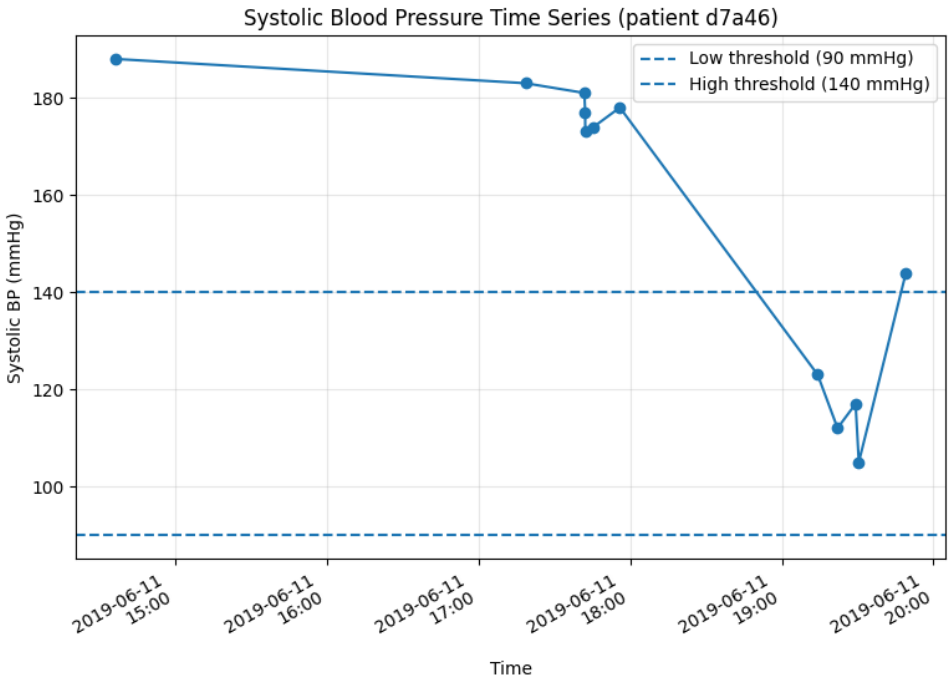

Figure 2: **Systolic blood pressure (SBP) over time for patient d7a46 on 2019-06-11.** The SBP peaks at 188 mmHg (14:36) and remains markedly elevated across multiple readings through 17:55, contradicting the summary's reported peak (150 mmHg) and time window (13:00–13:30).

Table 3: **Case study: mischaracterization of systolic BP abnormality (magnitude/timing mismatch).** Our event-based evaluation flags this day as *Abnormally High* SBP based on the raw time-series, while the LLM summary reports a lower peak and an incorrect time window.

Patient: d7a46 **Date:** 2019-06-11

Ground-truth evidence (rule-derived from time-series):

- **SBP threshold:** Abnormally High if > 140 mmHg
- **Observed SBP values (subset):** 14:36 → **188**, 17:18 → **183**, 17:41 → **181**, 17:42 → **177/173**, 17:45 → **174**, 17:55 → **178** (mmHg)
- **Peak SBP: 188 mmHg** at **14:36**
- **Temporal localization:** markedly elevated readings recur across **14:36–17:55**

LLM summary (excerpt):

> *"The systolic blood pressure was abnormally high at 150 mmHg (peak value) . . . sustained . . . from 13:00 to 13:30."*

Evaluation outcome (ours):

- **Abnormality detection (SBP):** Partially correct (mentions "abnormally high SBP")
- **Value accuracy: Incorrect** (reported peak 150 vs observed peak 188)
- **Temporal accuracy: Incorrect** (reported 13:00–13:30 vs observed elevated measurements 14:36–17:55)
- **Failure type: Mischaracterization** (wrong magnitude / wrong time window)

Why this matters:

- A summary can sound clinically plausible while *understating severity* and *misplacing timing*, which can distort clinician interpretation and downstream decision-making.
- This highlights a key reliability gap beyond binary "mentioned vs omitted": **event fidelity requires correct values and intervals, not just correct topic**.