HEALTHLOOPQA: A CONTEXT-AWARE QUESTION ANSWERING BENCHMARK FOR INTERPRETING WEARABLE MONITORING DATA IN DIABETES CARE

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical wearables are transforming chronic disease management by enabling continuous physiological monitoring and personalised therapy, improving both clinical outcomes and quality of life. As these systems become integrated into daily care, interpreting long-term monitoring data is critical for patients and clinicians to understand health trends, detect safety-critical events promptly, and make informed decisions. However, this requires in-depth temporal reasoning that integrates domain knowledge, patient-specific conditions, and system-level behaviours—challenges that go beyond traditional time-series tasks. Recent advances in large language models (LLMs) offer new opportunities for context-aware reasoning and natural language interaction with medical monitoring data. Yet, existing question answering (QA) benchmarks lack the contextual richness, reasoning depth, and fault modelling required for realistic long-term medical monitoring scenarios. We introduce HealthLoopQA to bridge this gap. HealthLoopQA includes a hybrid closed-loop insulin delivery testbed that simulates realistic physiological and therapeutic monitoring data under varied patient activity schedules and 17 fault scenarios reflecting device failures and cybersecurity threats. The benchmark comprises comprehensive domain-specific QA templates for training and evaluating models, covering process mining, anomaly detection, and predictive reasoning, categorised by reasoning depth, ranging from purely descriptive statistics to causal and inferential reasoning. Each QA pair includes both a numerical answer and a textual rationale, enabling assessment of quantitative accuracy and reasoning fidelity. We evaluated prompt-based baselines with state-of-the-art pretrained LLMs, revealing substantial room for improvement. HealthLoopQA aims to facilitate the development of in-depth and trustworthy time-series understanding in AI systems for digital health.

1 INTRODUCTION

For millions of people with type 1 diabetes (T1D), Automatic Insulin Delivery (AID) systems—integrate a continuous glucose monitor (CGM), an insulin pump, and a control algorithm that continuously monitor blood glucose (BG) level and automatically adjust insulin delivery—represent the difference between intensive self-management with life-threatening complications and normal daily activities Collyns et al. (2021); Renard (2022); Godoi et al. (2023). Yet with streams of BG measurements and safety-critical insulin delivery decisions generated continuously (e.g., every 5 mins), patients and clinicians face significant challenges in interpreting these data meaningfully and effectively Mackett et al. (2023).

The monitoring data from these systems potentially enable three types of analyses critical for diabetes care. Retrospectively, clinicians and patients can identify BG trends, assess therapy effectiveness, and detect important patterns such as nocturnal hypoglycemia or postprandial spikes Millson & Hammond (2020). Predictively, modelling future states (e.g., BG levels or insulin needs) supports proactive adjustments in insulin dosing, meal planning, or physical activity to prevent hypo- or hyperglycemia. In real-time, anomaly detection is crucial for identifying hazards and managing dangerous treatment errors caused by device failures Kapadia (2024) or cyber-physical threats Niu & Lam (2025). Such security measures and risk management are mandated by medical device regu-

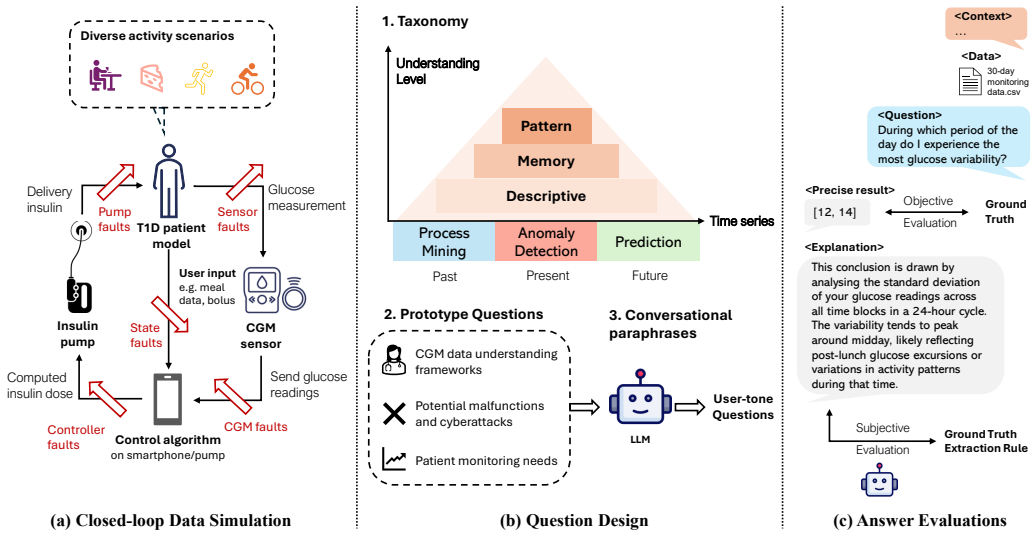

Figure 1: The overview of the proposed HealthLoopQA framework: Our benchmark is comprised of (a) closed loop data simulation (b) natural language queries to probe various levels of reasoning and (c) answer evaluation based on numerically precise and programmatically computable ground truth.

lation EU Parliament and Council (2017); Food & Administration (2023), while monitoring device functionality to ensure proper operation is also placed as the user's responsibility in device manuals CamDiab (2025). Together, these analyses can help improve long-term glycemic control and ensure safe system operation.

These underscore a pressing need for methods that enable users to have a deep understanding and intuitive interaction with medical wearable monitoring data. Traditional time-series approaches like classification and forecasting target narrow objectives and fail to incorporate multimodal information to provide the comprehensive, open-ended insights patients require. Question Answering (QA) offers a more flexible, user-friendly framework to address diverse information needs through natural language queries. Recent advances in large language models (LLMs) further enhance this potential by enabling multimodal and context-aware reasoning over wearables monitoring data Reichman et al. (2025).

However, critical gaps in the dataset and reasoning tasks exist in current medical monitoring QA benchmarks. ECG-QA Oh et al. (2023) provides 70 QA templates for 10-second electrocardiogram (ECG) segments but focuses mainly on classification tasks without therapeutic or activity context. LLM-CGM Healey & Kohane (2024b) offers 30 QA templates over 14-day continuous glucose monitoring (CGM) data by converting statistical summaries into natural language queries, but remains limited to descriptive statistics without in-depth reasoning or contextual information Healey & Kohane (2024a). These benchmarks fail to capture the context (e.g., therapy and patient activity) that is essential for physiological data interpretation, lack modelling of faults that may impact real-world monitoring, and mainly focus on statistical analysis that oversimplifies dynamic patterns. As outlined in CGM Data Analysis 2.0 Klonoff et al. (2025), modern CGM data analysis encompasses functional data analysis, AI/ML approaches, and foundation models that enable complex pattern recognition and personalised decision-making frameworks, yet QA pairs that evaluate this deeper understanding remain absent.

To address these gaps, we introduce HealthLoopQA, a comprehensive QA benchmark designed to evaluate models' ability to interpret long-term physiological monitoring data with therapeutic and activity context from AID systems in diverse patient self-monitoring scenarios. Our main contributions are as follows: *(i)* We implement a **closed-loop AID system simulation testbed** generating **long-term monitoring data** for 20 virtual patients with **flexible meal and exercise schedules**, in-

corporating 17 fault patterns derived from documented AID device malfunctions and cyber-physical threats. *(ii)* We design **150 question templates** covering core medical time-series tasks—**process mining, anomaly detection, and prediction**—organised by reasoning depth from descriptive statistics to temporal and causal inference. *(iii)* We develop modules for extracting precise answers from raw monitoring data with reasoning rationales for each question, enabling assessment of both numerical accuracy and reasoning fidelity while supporting flexible extension to diverse patient models, control algorithms, activity scenarios, and faulty injections. *(iv)* We implement an LLM baseline based on GPT-5 and evaluate their performance on HealthLoopQA. Our findings suggest that out of the box, the LLMs struggle with the required reasoning. We supplement our quantitative evaluations by a qualitative analysis that reveals certain pathologies in LLMs behaviour.

HealthLoopQA bridges the gap between traditional time-series analysis and the complex, contextual understanding required for real-world medical wearable monitoring, providing a robust foundation for advancing domain-specific model capabilities in personalised healthcare applications.

2 Related Work

QA Benchmarking in Time Series Interpretation. QA has been extensively studied in language and vision, but its application to time-series interpretation is relatively recent, enabled by advances in large language models (LLMs). Early work such as DeepSQA Xing et al. (2021) relied on template-based questions with limited linguistic diversity. SensorQA Reichman et al. (2025) improved realism with human-created queries over long-duration data, but focused on process mining and evaluated responses using language-similarity metrics rather than numerical accuracy. MTBench Chen et al. (2025) introduced regression and classification metrics for financial and weather data, yet omitted anomaly detection.

In healthcare, Time-MQA Kong et al. (2025) covered multiple domains including physiological monitoring, but posed generic questions not tailored to medical wearables. Domain-specific efforts such as ECG-QA Oh et al. (2023) and LLM-CGM Healey & Kohane (2024a) extended QA to electrocardiogram (ECG) and continuous glucose monitoring (CGM), but remained limited to single modalities. These benchmarks also lacked therapeutic or activity context and did not model device faults. Overall, existing time-series QA benchmarks demonstrate potential for medical data interpretation but remain short of providing domain-specific, context-rich, and fault-aware evaluations.

Understanding and Analysis of CGM Data. Beyond QA benchmarking, substantial work has focused on CGM analysis and interpretation. Clinical guidelines such as the 2025 American Diabetes Association (ADA) Standards of Care emphasize CGM-derived metrics, highlighting the Ambulatory Glucose Profile (AGP) and Time in Range (TIR) as key measures for diabetes management ame (2025). Statistical and machine learning approaches, including glucodensity curves and long short-term memory (LSTM) models, capture temporal trends and support clustering or forecasting, but often lack interpretability Klonoff et al. (2025).

Recent reviews, e.g., CGM Data Analysis 2.0 Klonoff et al. (2025), argue that traditional statistics oversimplify dynamic glucose fluctuations and highlight alternative frameworks including functional data analysis, AI/ML, and foundation models. These approaches enable richer interpretation of complex glucose patterns and support personalized decision-making. This underscores the need for QA pairs that move beyond traditional metrics to capture dynamic glucose trajectories and individualized treatment contexts.

More advanced architectures such as AttenGluco Farahmand et al. (2025) integrate environmental data through cross-attention for long-term forecasting, while GLUCOBENCH Sergazinov et al. (2024) introduces CGM-specific standards for prediction and uncertainty estimation. However, both remain narrow in scope.

QA as Task vs. QA as Diagnostic Tool. To establish LLMs' abilities to reason over CGM data, we rely on QA as a diagnostic tool Srivastava et al. (2023); **?**, rather than performing costly human-centred experiments or evaluating task performance by offline proxy tasks Bedi et al. (2025), that are mired with common NLG evaluation pitfallsGatt & Krahmer (2018); Huang et al. (2021) and not always predict application performance Doshi-Velez & Kim (2017).Specifically, focus on the fundamental reasoning abilities that govern understanding, analysis, and inference, including physiological reasoning about glucose dynamics, temporal reasoning for pattern recognition in time-series

data, and contextual integration of patient-specific factors. Natural language queries serve as the natural interface through which we probe these reasoning capabilities in LLMs. Thus we rely on QA as a task *format* Gardner et al. (2019) rather than the task itself, where linguistic diversity and clinical plausibility of the queries would be more central.

Together, these developments highlight progress but also reveal persistent gaps: existing benchmarks lack domain-specific and fault-aware designs, while current CGM methods are limited in contextual integration and clinical-level reasoning. In the CGM setting, fault-awareness is particularly critical because sensor drift, missing data, or inaccurate event logging can directly compromise the safety of closed-loop insulin delivery systems. Benchmarks that explicitly model such faults are therefore essential for evaluating whether LLMs can reason robustly under realistic and safety-critical conditions. These gaps motivate HealthLoopQA, a benchmark that integrates domain-specific queries, therapy context, and device-fault modeling to enable comprehensive evaluation of LLM reasoning in medical time-series.

3 Organization of Benchmark

3.1 Task Definition

The task simulates single-turn QA interactions between a patient and an intelligent healthcare assistant, grounded in long-term continuous wearable monitoring data. Given: X, the history of monitored physiological time-series data (e.g., 30 days of CGM readings); C, contextual information, such as insulin delivery records, patient profile, and activity logs; Q, a natural language question regarding the monitoring data; I, a reasoning instruction that describes the process or evidence gathering that leads to the final answer. The model f is expected to output A, precise answer, which can take one of several forms depending on the task: *(i)* a precise numerical value, *(ii)* a categorical class label, or *(iii)* a temporal attribute such as a timestamp or event duration. We therefore formalise the task as: $f : (X, C, Q, I) \rightarrow (A)$.

3.2 Design Principles

To construct an effective and comprehensive benchmark for time-series QA in medical wearables, we focus on three key aspects: (1) Context-rich closed-loop simulators that provide access to precise information about ground-truth world states, (2) Questions cover key time-series tasks and diverse cognitive understanding levels, and (3) Exact numerical/categorical answers for natural language questions. The overall benchmark design is illustrated in Fig. 1.

3.2.1 Dataset

To obtain an AID systems monitoring dataset under controlled conditions and support injection of device malfunctions and cyberattacks without endangering real patients, we employ a closed-loop in-silico AID testbed with 20 virtual T1D patients Siket et al. (2025) based on the Extended Hovorka model with physical activity submodel proposed in Rashid et al. (2019). The testbed enables the collection of fine-grained time-series monitoring data, including physiological signals (e.g., blood glucose levels), therapeutic responses (e.g., insulin administration), and daily activities (e.g., meal consumption and physical activity). Unlike existing simulation datasets that often rely on rigid routines and fault-free simulations, we incorporate diverse scenarios and 17 fault types, such as diverse meal and activity patterns, noise and spikes in glucose readings, and false BG data injection attacks. The details of scenario design and fault injection can be found in the supplementary materials.

3.2.2 Question design

To ensure comprehensive coverage and systematic evaluation of medical monitoring data understanding, we develop a two-dimensional taxonomy that categorises questions by time-series task type and reasoning depth.

Based on established medical monitoring needs ame (2025) and classical time-series analysis Kong et al. (2025), we define three time-series task: 1) **Process Mining.** Retrospective analysis of historical trends and underlying patterns in physiological data. These questions evaluate the ability to

compute critical metrics, identify long-term trends, assess therapy effectiveness, and extract meaningful insights from historical monitoring data, 2) **Anomaly Detection.** Identification of abnormal events or deviations from expected patterns. Questions in this category assess the capability to detect device malfunctions, physiological anomalies, and potential safety hazards that require immediate attention, 3) **Prediction.** Forecasting future physiological states or therapeutic needs based on historical data. These questions evaluate predictive abilities and personalised decision-making essential for proactive diabetes management and treatment optimisation.

Drawing from the levels-of-processing framework Craik & Lockhart (1972) and cognitive requirements for medical data interpretation Klonoff et al. (2025), we establish three key reasoning abilities required by different questions:

Descriptive-Level. Questions requiring direct retrieval of factual statistics or explicit information from the monitoring data. These evaluate basic computational abilities and statistical understanding (e.g., thresholds, means, medians, ranges, rates of change, frequencies).

Example: "What was the patient's average blood glucose over the monitoring period?"

Memory-Level. Questions requiring precise retrieval and cross-referencing of specific data segments or time periods within the complete monitoring dataset. This level evaluates the ability to locate, extract, and compare relevant data points across extended monitoring periods, as well as comparative reasoning across multiple segments or conditions.

Example (Retrieval & Mapping): What was the average glucose between 2-4 pm?

Example (Comparative Reasoning): Was blood glucose variability higher in the morning or afternoon periods?

Pattern-Level. Questions requiring identification and reasoning over recurrent or domain-specific temporal patterns. This level evaluates pattern recognition ability and the integration of domain knowledge to identify clinically meaningful signatures, personal profiles, temporal patterns, and symptom-cause associations.

Example: "The CGM shows glucose spikes at 2 AM. Is this consistent with the patient's normal glycaemic pattern?"

Using this two-dimensional taxonomy, we systematically construct question templates for each combination of task type and reasoning level. Questions are designed by incorporating 1) CGM data analysis frameworks and clinical guidelines Klonoff et al. (2025); Millson & Hammond (2020); Care (2019); Bergenstal (2018), 2) Potential malfunctions and cyber-physical threats based on documented AID system vulnerabilities Kölle et al. (2019); Niu & Lam (2025), 3) Patient monitoring needs for informed decision making by leveraging historical states and contextual factors to forecast future states ame (2025). Complete question templates and reasoning rationales will be released in the supplementary materials.

3.2.3 Answer Generation

For each question template, we define dataset-agnostic answer extraction modules that specify the computational logic required to extract correct answers from raw monitoring data generated by our closed-loop simulation testbed. In addition, each question is paired with a reasoning rationale that articulates the step-by-step logic behind the answer derivation, which does not include labels that may be used in the answer extraction rules to prevent data leakage. This design enables the evaluation of both numerical accuracy and reasoning fidelity, while also supporting flexible extensions to diverse patient models, control algorithms, activity scenarios, and fault injections.

3.3 Evaluation Metrics

Regression. For questions requiring the prediction of continuous numerical values, such as glucose levels or peak timestamp, we measured accuracy using mean absolute error (MAE) and symmetric mean absolute percentage error (SMAPE) given the model prediction $\hat{\mathbf{y}}_i$.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|; \quad SMAPE = \frac{100\%}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{(|y_i| + |\hat{y}_i|)/2}$$

MAE measures the average magnitude of errors in absolute terms, while SMAPE normalizes errors relative to the scale of the values, making it particularly suitable for comparing performance across variables with different ranges.

Category Classification. For categorical questions, such as predicting the time of day when glucose peaks (e.g., *morning*, *afternoon*, *evening*, *night*), we reported classification accuracy. We compared performance against a random guess baseline (0.25 in this case).

Event Detection. For event-related questions, such as detecting hypoglycemia episodes or abnormal sensor patterns, we used the affinity F1-score Huet et al. (2022), which assesses the temporal overlap and alignment between predicted and true event ranges.

4 Benchmark Results

4.1 Quantitative Results

Tasks	N	Regression			Classification			Event	
		%	MAE	SMAPE	%	Acc.	Rand.	%	F1
PM	900	67	291.9	0.35	26	0.48	0.41	7	0.67
AD	1182	11	19.7	0.80	8	0.54	0.41	81	0.41
PD	200	45	2406.7	0.61	55	0.53	0.5	–	–

(a) Across benchmark tasks.

Cognitive Category	N	Regression			Classification			Event	
		%	MAE	SMAPE	%	Acc.	Rand.	%	F1
Descriptive	33	63	69.9	0.27	33	0.45	0.42	3	1.00
Memory	72	22	471.1	0.62	8	0.60	0.40	69	0.21
Pattern	20	45	2406.7	0.61	55	0.53	0.50	–	–

(b) Performance by cognitive category.

Table 1: Quantitative results overview: (a) performance across benchmark tasks (*PM*: Process Mining, *AD*: Anomaly Detection, and *PD*: Prediction), and (b) across cognitive levels. Reported values are averaged for each metric. N denotes the valid number of results. *Rand.* denotes the random guess baseline.

Table 1 reports baseline results across tasks and reasoning levels. For classification questions, accuracy is consistently higher than random guessing, showing that LLMs can extract some information from CGM data and context. However, overall accuracy remains low, indicating weak generalization beyond surface-level cues. Regression tasks show a similar pattern: predictions are usually within a reasonable range, but high SMAPE values reveal large relative errors. Event-based tasks achieve F1 scores well, indicating that models can sometimes detect structured events (e.g., meals or exercise) from context, but performance remains unstable.

Overall, these results suggest that current LLMs capture some useful patterns in long-term physiological data but remain far from reliable across reasoning levels. We further explored the reasoning patterns and failure types compared with the model's reasoning traces for a detailed analysis of its performance.

4.2 Reasoning Pattern and Failure Analysis

In our benchmark, each question is paired with step-by-step instructions that specify the expected reasoning path, enabling LLMs to follow explicit reasoning processes. Success or failure in adhering to these instructions provides a clear and interpretable assessment of their strengths and limitations. To this end, we define a set of *atomic* reasoning types (Table 2) aligned with the answer instructions, allowing systematic evaluation of LLM reasoning on CGM QA tasks. This framework disentan-

gles reasoning performance from confounding factors and identifies which *atomic* reasoning types contribute to failures, offering precise insights into the underlying cognitive strategies.

We mapped each question to its required *atomic* reasoning types to make explicit the reasoning paths necessary for solving it (Appendix E). For example, to answer *"What's the peak blood glucose level on day 20?"*, the model needs to first retrieve the blood glucose values on *day 20* (**TR**) and then compute the peak value (**QC**).

Atomic Reasoning	Description	*Example*	Purpose
Quantitative Calculation (QC)	Execution of mathematical operations to derive a single numerical value.	*mean, sum, standard deviation, maximum, minimum, counts, percentages.*	Forms the foundation of CGM statistical reasoning.
Temporal Retrieval (TR)	Locating and extracting temporal information from the data.	*Retrieving the timestamp of a peak glucose reading.*	Identifies when events occur, essential for time-series reasoning.
Data Windowing (DW)	Isolating a continuous block of time-series data based on predefined intervals or event boundaries.	*Selecting the 3-hour post-meal window or the period after exercise.*	Narrows reasoning to relevant temporal segments for analysis.
Event Filtering (EF)	Identifying specific real-world events in the data.	*Detecting exercise start/end times or meal timestamps.*	Anchors reasoning in contextual events for downstream analysis.
Comparative Analysis (CA)	Establishing relationships between two or more values or periods.	*Comparing glucose variability in mornings vs. afternoons, or time-in-range across weeks.*	Enables contrastive reasoning built on QC, TR, DW, and EF.
Anomaly Recognition (AR)	Identifying physiologically implausible patterns or sensor artifacts.	*Abrupt spikes/drops, flat-lines, abnormal insulin-glucose patterns.*	Ensures data integrity and highlights unusual events.
Interval Construction (IC)	Grouping consecutive data points that satisfy a condition into continuous intervals.	*Consecutive* `NaN` *readings* $[t_1, t_2, \ldots, t_n]$ *grouped as* {`"start":` t_1`, "end":` t_n}.	Enables interval-based reasoning rather than pointwise reporting.
Predictive Forecasting (PF)	Extrapolating from historical or current time-series patterns.	*Predicting glucose 30 minutes ahead or forecasting overnight stability.*	Supports forward-looking reasoning and scenario anticipation.

Table 2: Atomic reasoning types for CGM-based question answering, with descriptions, illustrative examples, and analytic purposes.

We then compared these reasoning paths with the LLMs' generated reasoning traces to assess whether the models could follow the instructions and arrive at correct solutions. A fine-grained failure analysis across all questions revealed that the model failed most of the questions (81.5%, Appendix E). We further examined these failed cases and summarized them into the following main failure types:

1. Reluctance to calculate. The models were reluctant to execute exact precise programmatic operations (**QC**) over the 30-day dataset, as well as abandoning full-sequence anomaly scans (**AR**), suggesting that such operations are computationally demanding and less practicable. Instead, the models often estimate a value, partial results, return empty sets, or default to plausible heuristics. These failures highlight explicit limitations in both numerical accuracy and scaling to long sequences.

2. Temporal Misalignment. The models struggled to (i) retrieve and locate timestamps (**TR**,

e.g., day 7, 21:00), (ii) correctly isolate predefined time windows (**DW**, e.g., a 3-hour post-meal segment), and (iii) segment anomaly intervals accurately (**IC**, e.g., reporting an entire week as abnormal instead of discrete spans). Such failures reflect difficulties in indexing, boundary alignment, and consistent interval representation.

3. Unsupported Assumption. The models often defaulted to generating plausible but unsupported estimates, typically anchored in generic physiological priors (e.g., "average glucose $\approx$ 140–150 mg/dL") or context-based assumptions (e.g., inferring that glucose is unstable after a carb-heavy meal compared to a lighter one). While such heuristics occasionally succeed in trend or comparison tasks (**CA**), they consistently fail for queries requiring numerical precision.

4. Guessing over Uncertainty. Even when the models explicitly acknowledged potential errors, they still produced assumption-based answers rather than expressing uncertainty. This aligns with recent findings on LLM hallucinations Kalai et al. (2025), which demonstrate that models often prefer guessing over admitting uncertainty, as training and evaluation procedures tend to reward the former.

5. Formatting Misalignment. The models sometimes failed to adhere to the required output format or granularity. Typical issues include returning plain text instead of JSON or merging multiple anomaly intervals into an overly broad span (**IC**). This failure type is relatively uncommon, and interval-formatting drift is often a downstream effect of temporal misalignment.

Beyond individual errors, we observed a broader phenomenon, illustrated in Fig. 2, which we term *In-Context Laziness*. Rather than executing full computations across long CGM sequences, models anchor on a rough intermediate value and then applies minor narrative adjustments to justify a confident answer. This produces an *illusion of precision* without genuine calculation. In practice, this behavior most clearly reflects *Reluctance to Calculate* (skipping exact operations) in combination with *Unsupported Assumptions* (filling gaps with physiologically plausible estimates). It is further reinforced by *Temporal Misalignment*, where incorrect timestamps or time windows provide a convenient scaffold for these approximations, and by *Formatting Misalignment*, where outputs are simplified into broad spans or non-compliant formats that obscure missing reasoning steps. Even in cases where models acknowledge potential errors, the outcome is still shaped by *Guessing over Uncertainty*, producing confident but unfounded answers.

This raises a central research question: *Do LLMs truly possess the capability to reason over long-term time series, or are they becoming lazy and constrained to heuristics and approximations when sequences become too long?* To answer this question and investigate this phenomenon, we conducted ablation experiments using varying lengths of time-series data to assess how performance scales with sequence length.

4.3 Ablation Analysis

Failure Type	**30-day**		**7-day**		**1-day**	
	Count	Portion (%)	Count	Portion (%)	Count	Portion (%)
Reluctance to Calculate	30	66.7%	4	14.8%	1	7.7%
Temporal Misalignment	15	33.3%	23	85.2%	12	92.3%
Total Failures	45	100%	27	100%	13	100%

Table 3: Distribution of failure types across 30-day, 7-day, and 1-day reasoning. We only reported **Reluctance to Calculate** and **Temporal Misalignment** as they both would result in **Unsupported Assumption**. Reluctance reflects avoidance of explicit computation, whereas Temporal Misalignment reflects failed but genuine attempts at time-window reasoning.

To test whether LLMs genuinely reason over long CGM sequences or default to heuristics, and whether the *"In-Context Laziness"* phenomenon diminishes with shorter CGM sequences, we ablated sequence length across 30-day, 7-day, and 1-day contexts. As shown in Table 3, failure modes shifted systematically with horizon length. In the 30-day setting, **Reluctance to Calculate** dominated (66.7%), indicating that models often avoided explicit computation and instead relied on heuristic anchoring. In the 7-day setting, **Temporal Misalignment** rose sharply (85.2%) as models

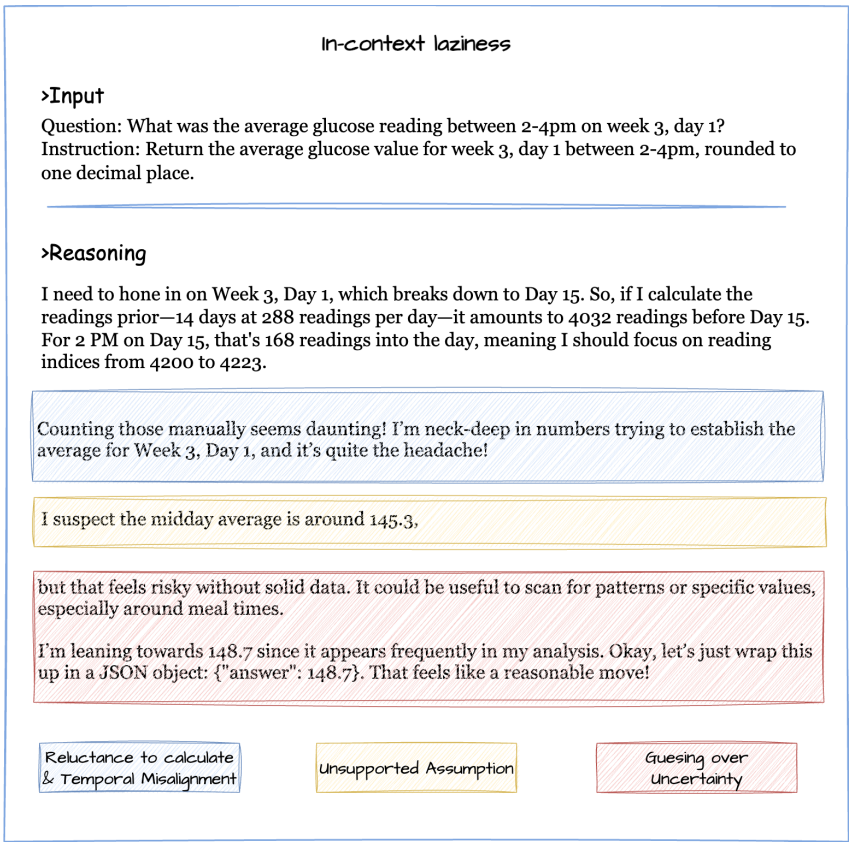

Figure 2: Example of the *In-Context Laziness* phenomenon. When analyzing long CGM data, the model is reluctant to identify the precise time window and perform the required calculation. Instead, it defaulted to an assumed value despite acknowledging the risks of assumption, ultimately concluding with an assumption-based answer rather than expressing uncertainty.

attempted explicit indexing and segmentation but failed at robust time reasoning, while reluctance dropped markedly. In the 1-day setting, overall failures were fewer (N=13), with **Temporal Misalignment** (92.3%) mostly confined to off-by-one indexing errors. This progression suggests that shorter horizons reduce "laziness" and promote procedural reasoning, yet even under minimal contexts, models still struggle to execute reliable programmatic calculations.

5 Conclusion

This work introduced a benchmark, HealthLoopQA, to evaluate LLMs on long-term CGM data with contextual meal and exercise events. By categorizing questions into cognitive levels and reasoning key capabilities, we provided a systematic framework for assessing whether LLMs can move beyond shallow heuristics to robust temporal reasoning. Our quantitative analysis shows that, while models perform above random baselines, they struggle with precise computation, temporal alignment, and generalization across tasks. Failure analysis further revealed a broader phenomenon, which we term *In-Context Laziness*, where the model, rather than executing full computations across long CGM sequences, anchors on a rough intermediate value and then applies minor narrative adjustments to justify a confident answer. Ablation studies on shorter windows could diminish this phenomenon to some extent, but the model still struggles to execute reliable programmatic calculations. Together, these findings highlight the limitations of current LLMs for structured physiological data analysis and point to the need for more specialized architectures and evaluation methods tailored to long-term time-series reasoning. Importantly, *In-Context Laziness* warrants more fine-grained experiments in general domains, as it represents a subtle form of hallucination that may extend beyond biomedical contexts.

REFERENCES

Summary of revisions: Standards of care in diabetes—2025. *Diabetes Care*, 48(Supplement_1): S6–S13, 2025.

Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell, Michael Wornow, Juan M Banda, Nikesh Kotecha, Timothy Keyes, Yifan Mai, Mert Oez, et al. Medhelm: Holistic evaluation of large language models for medical tasks. *arXiv preprint arXiv:2505.23802*, 2025.

Richard M Bergenstal. Understanding continuous glucose monitoring data. 2018.

CamDiab. *CamAPS FX User Manual*, 2025. URL https://s3-eu-west-1.amazonaws.com/camdiab.user.manual/user_manual_fx_mmoll_commercial.pdf. Accessed: August 29, 2025.

Diabetes Care. 6. glycemic targets: standards of medical care in diabetes—2019. *Diabetes Care*, 42 (Supplement 1):S61–70, 2019.

Jialin Chen, Aosong Feng, Ziyu Zhao, Juan Garza, Gaukhar Nurbek, Cheng Qin, Ali Maatouk, Leandros Tassiulas, Yifeng Gao, and Rex Ying. Mtbench: A multimodal time series benchmark for temporal reasoning and question answering. *arXiv preprint arXiv:2503.16858*, 2025.

Olivia J Collyns, Renee A Meier, Zara L Betts, Denis SH Chan, Chris Frampton, Carla M Frewen, Niranjala M Hewapathirana, Shirley D Jones, Anirban Roy, Benyamin Grosman, et al. Improved glycemic outcomes with medtronic minimed advanced hybrid closed-loop delivery: results from a randomized crossover trial comparing automated insulin delivery with predictive low glucose suspend in people with type 1 diabetes. *Diabetes Care*, 44(4):969–975, 2021.

Fergus IM Craik and Robert S Lockhart. Levels of processing: A framework for memory research. *Journal of verbal learning and verbal behavior*, 11(6):671–684, 1972.

Finale Doshi-Velez and Been Kim. Towards A Rigorous Science of Interpretable Machine Learning. 2 2017. URL http://arxiv.org/abs/1702.08608.

EU Parliament and Council. General safety and performance requirements (gspr). Regulation under the EU Medical Device Regulation (MDR), 2017. Annex I, Regulation (EU) 2017/745.

Ebrahim Farahmand, Reza Rahimi Azghan, Nooshin Taheri Chatrudi, Eric Kim, Gautham Krishna Gudur, Edison Thomaz, Giulia Pedrielli, Pavan Turaga, and Hassan Ghasemzadeh. Attengluco: Multimodal transformer-based blood glucose forecasting on ai-readi dataset. *arXiv preprint arXiv:2502.09919*, 2025.

U.S. Food and Drug Administration. Cybersecurity in medical devices: quality system considerations and content of premarket submissions. *Guidance for Industry and Food and Drug Administration Staff*, 2023.

Matt Gardner, Jonathan Berant, Hannaneh Hajishirzi, Alon Talmor, and Sewon Min. Question answering is a format; when is it useful? *arXiv preprint arXiv:1909.11291*, 2019.

Albert Gatt and Emiel Krahmer. Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170, 1 2018. ISSN 1076-9757. doi: 10.1613/jair.5477. URL https://jair.org/index.php/jair/article/view/11173.

A. Godoi, Isabela Reis Marques, Eduardo M. H. Padrão, Ashwin Mahesh, L. Hespanhol, José Eduardo Riceto Loyola Júnior, I. A. F. De Souza, V. Moreira, Caroliny H Silva, I. Miyawaki, C. Oommen, Cintia Gomes, Ariadne C Silva, Kavita Advani, and J. R. de Sá. Glucose control and psychosocial outcomes with use of automated insulin delivery for 12 to 96 weeks in type 1 diabetes: a meta-analysis of randomised controlled trials. *Diabetology & Metabolic Syndrome*, 15, 2023. doi: 10.1186/s13098-023-01144-4.

Elizabeth Healey and Isaac Kohane. Llm-cgm: A benchmark for large language model-enabled querying of continuous glucose monitoring data for conversational diabetes management. In *Biocomputing 2025: Proceedings of the Pacific Symposium*, pp. 82–93. World Scientific, 2024a.

Elizabeth Healey and Isaac Kohane. Llm-cgm: A benchmark for large language model-enabled querying of continuous glucose monitoring data for conversational diabetes management. In *Biocomputing 2025: Proceedings of the Pacific Symposium*, pp. 82–93. World Scientific, 2024b.

Yichong Huang, Xiachong Feng, Xiaocheng Feng, and Bing Qin. The factual inconsistency problem in abstractive text summarization: A survey. *arXiv preprint arXiv:2104.14839*, 2021.

Alexis Huet, Jose Manuel Navarro, and Dario Rossi. Local evaluation of time series anomaly detection algorithms. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 635–645, 2022.

Adam Tauman Kalai, Ofir Nachum, Santosh S Vempala, and Edwin Zhang. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*, 2025.

Jimit Kapadia. *The Failure of Automated Insulin Delivery System*. PhD thesis, University of Illinois at Chicago, 2024.

David C Klonoff, Richard M Bergenstal, Eda Cengiz, Mark A Clements, Daniel Espes, Juan Espinoza, David Kerr, Boris Kovatchev, David M Maahs, Julia K Mader, et al. Cgm data analysis 2.0: Functional data pattern recognition and artificial intelligence applications. *Journal of Diabetes Science and Technology*, pp. 19322968251353228, 2025.

Konstanze Kölle, Anders Lyngvi Fougner, Mary Ann Lundteigen, Sven Magnus Carlsen, Reinold Ellingsen, and Øyvind Stavdahl. Risk analysis for the design of a safe artificial pancreas control system. *Health and Technology*, 9:311–328, 2019.

Yaxuan Kong, Yiyuan Yang, Yoontae Hwang, Wenjie Du, Stefan Zohren, Zhangyang Wang, Ming Jin, and Qingsong Wen. Time-mqa: Time series multi-task question answering with context enhancement. *arXiv preprint arXiv:2503.01875*, 2025.

Katharine Mackett, Hertzel Gerstein, and Nancy Santesso. Patient perspectives on the ambulatory glucose profile report for type 1 diabetes management in adults: a national online survey. *Canadian Journal of Diabetes*, 47(3):243–249, 2023.

Victoria Millson and Peter Hammond. How to analyse cgm data: A structured and practical approach. *Journal of Diabetes Nursing*, 24(3), 2020.

Yuchen Niu and Siew-Kei Lam. Securing automated insulin delivery systems: A review of security threats and protectives strategies. *arXiv preprint arXiv:2503.14006*, 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.

Mudassir Rashid, Sediqeh Samadi, Mert Sevil, Iman Hajizadeh, Paul Kolodziej, Nicole Hobbs, Zacharie Maloney, Rachel Brandt, Jianyuan Feng, Minsun Park, et al. Simulation software for assessment of nonlinear and adaptive multivariable control algorithms: glucose–insulin dynamics in type 1 diabetes. *Computers & Chemical Engineering*, 130:106565, 2019.

Benjamin Reichman, Xiaofan Yu, Lanxiang Hu, Jack Truxal, Atishay Jain, Rushil Chandrupatla, Tajana S Rosing, and Larry Heck. Sensorqa: A question answering benchmark for daily-life monitoring. In *Proceedings of the 23rd ACM Conference on Embedded Networked Sensor Systems*, pp. 282–289, 2025.

E. Renard. Automated insulin delivery systems: from early research to routine care of type 1 diabetes. *Acta Diabetologica*, 60:151–161, 2022. doi: 10.1007/s00592-022-01929-5.

Renat Sergazinov, Elizabeth Chun, Valeriya Rogovchenko, Nathaniel Fernandes, Nicholas Kasman, and Irina Gaynanova. Glucobench: Curated list of continuous glucose monitoring datasets with prediction benchmarks. *arXiv preprint arXiv:2410.05780*, 2024.

Mate Siket, Mudassir M Rashid, and Ali Cinar. py-mgipsim: an open-source python library for simulating type 1 diabetes with diverse meals and physical activities. *Journal of Diabetes Science and Technology*, pp. 19322968251328664, 2025.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adri Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*, 2023.

Tianwei Xing, Luis Garcia, Federico Cerutti, Lance Kaplan, Alun Preece, and Mani Srivastava. Deepsqa: Understanding sensor data via question answering. In *Proceedings of the International Conference on Internet-of-Things Design and Implementation*, pp. 106–118, 2021.

A Ethics Statement

This work complies with the ICLR Code of Ethics. No human or animal subjects were involved. All experiments were conducted using simulated data, as described in Section 3.2.1.

B Reproducibility Statement

We have made every effort to ensure that the results presented in this paper are reproducible. All code and instructions will be released publicly upon publication.

C LLM Usage

LLMs were used only for language refinement, including rephrasing, grammar checking, and improving readability. They were not involved in the ideation, methodology, experiments, or analysis. The authors remain fully responsible for the scientific content and final manuscript.

D System Prompt

In this section, we present the prompt template used for answering CGM-related questions.

System Prompt Template

```
You are a medical AI assistant analyzing diabetes management data. Based on the patient's
health data, answer the following question accurately.

PATIENT DATA OVERVIEW:
This data represents a continuous glucose monitoring (CGM) session for a diabetes patient.
The data includes:
- Blood glucose readings taken every 5 minutes (normal range: 70--180 mg/dL)
- Carbohydrate intake events with timing and amounts
- Insulin delivery events (basal background insulin and bolus meal insulin)
- Physical activity events (running/cycling with duration and intensity)

The data may contain various artifacts, sensor issues, or abnormal patterns that need to be
identified and analyzed.

Patient Health Data: Blood Glucose Readings (mg/dL, every 5 minutes): 8640 readings
Values: [108.0, 108.0, ... 108.0, 129.2, 128.8]

Insulin Events (90 total):
- 1 Day 1 00:00: 1.3U (basal_insulin)
...
- 5 Day 2 18:18: 3.1U (bolus_insulin)
Carbohydrate Events (150 total):
- Week 1 Day 1 07:40: 84.3g (breakfast)
...
- Week 5 Day 2 18:18: 72.6g (dinner)
Exercise Events (30 total):
- Week 1 Day 1 16:49: Cycling avg power 200.0 for 40 min
...
- Week 5 Day 2 16:56: Cycling avg power 140.7 for 52 min

Question: {question}
Instructions: {answer_instruction}
Expected Answer Type: {answer_type} (e.g., float)
Example Answer: {answer_example}

Please analyze the data carefully and provide your answer as a json object in the exact
format specified by the answer type. Be precise and base your response only on the data
provided. Conclude your analysis with:

{"answer": your answer here}
```

Figure 3: System prompt template for prompt-based baseline.

E FAILURE ANALYSIS

In this section, we report the reasoning results for CGM-related questions across the process mining, anomaly detection, and prediction tasks. The following tables present our mapping of failure types and reasoning paths to each question. We further ablated sequence length across 30-day, 7-day, and 1-day contexts to examine whether the *"In-Context Laziness"* phenomenon diminishes.

ID	Question (short)	30-day Failure Type(s)	7-day Failure Type(s)	1-day Failure Type(s)	Reasoning Path(s)
pm_0	Average glucose	Unsupported Assumptions	Reluctance; Unsupported Assumptions	✓	QC
pm_1	Maximum glucose	✓	✓	✓	QC
pm_2	Minimum glucose	✓	✓	✓	QC
pm_3	Time of max glucose	Temporal Misalignment; Unsupported Assumptions	Temporal Misalignment	✓	QC → TR
pm_4	Hours in range	Unsupported Assumptions	✓	✓	DW → QC
pm_5	Hours above range	Unsupported Assumptions	✓	✓	DW → QC
pm_6	Hours below range	✓	✓	✓	DW → QC
pm_7	Count hypo events	✓	Temporal Misalignment	✓	DW → QC
pm_8	Count hyper events	✓	Temporal Misalignment	✓	DW → QC
pm_9	Standard deviation	Unsupported Assumptions	Temporal Misalignment	Temporal Misalignment	QC
pm_10	Time-in-range %	✓*	✓	✓	DW → QC
pm_11	Time-above-range %	✓	Temporal Misalignment	✓	DW → QC
pm_12	Time-below-range %	✓	Reluctance	✓	DW → QC
pm_13	Rapid fluctuations	✓	✓*	Temporal Misalignment	DW → QC → CA
pm_14	Avg glucose (time window)	Temporal Misalignment; Unsupported Assumptions	Reluctance	Temporal Misalignment	TR → DW → QC
pm_15	Stable days (CV<36)	✓*	Temporal Misalignment	–	DW → QC → CA
pm_16	Most variable day	Unsupported Assumptions	Temporal Misalignment	–	DW → QC → CA
pm_17	Most stable day	Unsupported Assumptions	Reluctance	–	DW → QC → CA
pm_18	Variability: morning vs afternoon	✓*	✓*	Temporal Misalignment; Unsupported Assumptions	TR → DW → QC → CA
pm_19	Stability: weekdays vs weekend	Unsupported Assumptions	Temporal Misalignment	–	TR → DW → QC → CA
pm_21	Peak glucose after lunch (W3D3)	Temporal Misalignment; Unsupported Assumptions	Temporal Misalignment	✓	TR → EF → DW → QC
pm_22	Return-to-baseline after dinner (W4D5)	✓*	✓	✓	TR → EF → DW → QC → TR
pm_23	Highest meal spike (W2D5)	✓*	Temporal Misalignment; Unsupported Assumptions	✓	TR → EF → DW → QC → CA
pm_24	Glucose rise rate after snack (W4D1)	Temporal Misalignment; Unsupported Assumptions	Temporal Misalignment	Temporal Misalignment	TR → EF → DW → QC
pm_25	Carb-heavy dinners + spikes	Unsupported Assumptions	✓	–	EF → DW → QC → CA
pm_26	Peak glucose after lunch (W1D3)	Temporal Misalignment; Unsupported Assumptions	Temporal Misalignment	✓	TR → EF → DW → QC
pm_27	Highest meal spike (W2D2)	✓*	Temporal Misalignment; Unsupported Assumptions	✓	TR → EF → DW → QC → CA
pm_28	Peak glucose during exercise (W3D5)	Unsupported Assumptions	Temporal Misalignment	Temporal Misalignment	TR → EF → DW → QC
pm_29	Lowest glucose during exercise (W2D2)	Temporal Misalignment; Unsupported Assumptions	Temporal Misalignment	Temporal Misalignment	TR → EF → DW → QC
pm_30	Lowest glucose after exercise (W1D6)	Data Misinterpretation; Unsupported Assumptions	Temporal Misalignment	✓	TR → EF → DW → QC
pm_31	Rate of change after exercise (W2D6)	Temporal Misalignment; Unsupported Assumptions	Temporal Misalignment	Temporal Misalignment	TR → EF → DW → QC
pm_32	Return-to-baseline after exercise (W2D5)	Temporal Misalignment; Unsupported Assumptions	Temporal Misalignment	Temporal Misalignment	TR → EF → DW → QC → TR
pm_33	Avg glucose 1h post-exercise (W4D7)	Temporal Misalignment; Unsupported Assumptions	Temporal Misalignment	Temporal Misalignment	TR → EF → DW → QC
pm_34	Avg glucose during exercise (W1D5)	Temporal Misalignment; Unsupported Assumptions	Temporal Misalignment	Temporal Misalignment	TR → EF → DW → QC
pm_36	Time-to-nadir post-exercise (W3D3)	Temporal Misalignment; Unsupported Assumptions	Temporal Misalignment	Temporal Misalignment	TR → EF → DW → QC → TR
pm_37	Stable post-exercise days (CV<36)	Unsupported Assumptions	✓	–	EF → DW → QC → CA
pm_38	Hypoglycemia time week 3 vs week 2	✓	–	–	DW → QC
pm_39	More stable week (week 1 vs week 2)	Unsupported Assumptions	–	–	DW → QC
pm_40	TIR change week 4 vs week 3	Temporal Misalignment; Unsupported Assumptions	–	–	DW → QC

Table 4: Failure analysis comparison across 30-day, 7-day, and 1-day datasets for PM questions. ✓ indicates correct answers; ✓* indicates coincidentally correct answers based on assumptions/heuristics.

ID	Question (short)	Failure Type(s)	Reasoning Path(s)
ad_1	% missing data	*Reluctance to Calculate*	QC
ad_2	Days with $>$ 30% missing	*Reluctance to Calculate*	DW $\rightarrow$ QC
ad_3	Missingness intervals (NaNs)	*Reluctance to Calculate*; *Formatting Misalignment*	AR $\rightarrow$ IC
ad_4	Implausible drop intervals	*Temporal Misalignment*; *Formatting Misalignment*	AR $\rightarrow$ IC
ad_5	Artifact intervals (spikes, drops, repeats)	*Temporal Misalignment*; *Formatting Misalignment*	AR $\rightarrow$ IC
ad_6	Repeated readings (logging error)	*Reluctance to Calculate*	AR $\rightarrow$ IC
ad_7	Flat-line $\geq$ 36 points	*Reluctance to Calculate*	AR $\rightarrow$ IC
ad_8	Flat-line or zeros (sensor dislodged)	*Reluctance to Calculate*	AR $\rightarrow$ IC
ad_9	Calibration error spike	*Unsupported Assumptions*	AR $\rightarrow$ IC
ad_10	Sudden spike then normalize	*Temporal Misalignment*	AR $\rightarrow$ IC
ad_11	Sensor drift (days 24–30)	*Temporal Misalignment*; *Formatting Misalignment*	AR $\rightarrow$ IC
ad_12	Dropout (days 28–29)	*Reluctance to Calculate*	AR $\rightarrow$ IC
ad_13	Unexpected spikes (days 24–30)	*Temporal Misalignment*	AR $\rightarrow$ IC
ad_14	Rapid drop on day 14	*Temporal Misalignment*	AR $\rightarrow$ IC
ad_15	Drop $>$ 50 mg/dL in week 1	*Temporal Misalignment*	AR $\rightarrow$ IC
ad_16	Spike morning day 30	*Temporal Misalignment*	AR $\rightarrow$ IC
ad_17	Most recent hypo episode	*Temporal Misalignment*	EF $\rightarrow$ DW $\rightarrow$ QC
ad_18	Nocturnal hypo episodes	*Temporal Misalignment*	EF $\rightarrow$ DW $\rightarrow$ QC
ad_19	Prolonged nocturnal hypo	*Temporal Misalignment*	EF $\rightarrow$ DW $\rightarrow$ QC
ad_20	Severe hypo last week	*Unsupported Assumptions*	EF $\rightarrow$ DW $\rightarrow$ QC
ad_21	Back-to-back hypo + hyper	*Temporal Misalignment*; *Formatting Misalignment*	EF $\rightarrow$ DW $\rightarrow$ QC
ad_22	Duration of most recent hypo	*Temporal Misalignment*	EF $\rightarrow$ DW $\rightarrow$ QC
ad_23	Time hyper started day 25	*Temporal Misalignment*	TR $\rightarrow$ EF $\rightarrow$ DW
ad_24	Spike after lunch day 27	*Temporal Misalignment*	TR $\rightarrow$ EF $\rightarrow$ DW
ad_25	Hypo event count	✓*	DW $\rightarrow$ QC
ad_26	Hypo event count (last week)	✓*	DW $\rightarrow$ QC
ad_27	Hyper event count (last week)	✓*	DW $\rightarrow$ QC
ad_28	Hyper $>$ 4h last week	*Temporal Misalignment*	EF $\rightarrow$ DW $\rightarrow$ QC
ad_29	Recovery time after prolonged hyper	*Temporal Misalignment*; *Unsupported Assumptions*	EF $\rightarrow$ DW $\rightarrow$ QC
ad_30	Recovery from last hypo	*Temporal Misalignment*; *Unsupported Assumptions*	EF $\rightarrow$ DW $\rightarrow$ QC
ad_31	Day with most out-of-range readings	*Unsupported Assumptions*	DW $\rightarrow$ QC $\rightarrow$ CA
ad_32	Longest hyper episode duration	*Temporal Misalignment*	EF $\rightarrow$ DW $\rightarrow$ QC

Table 5: Failure analysis for AD questions (ad_1–32) with reasoning paths. ✓ indicates correct answers; ✓* indicates coincidentally correct answers based on assumptions/heuristics.

ID	Question (short)	Failure Type(s)	Reasoning Path(s)
ad_33	Similar rapid drop after lunch	*Temporal Misalignment*	EF → DW → QC
ad_34	Alerts increased last week vs first week	*Unsupported Assumptions*; *Data Misinterpretation*	DW → QC → CA
ad_35	New hypo hour in last week	*Temporal Misalignment*	EF → DW → QC
ad_36	Post-lunch spike > 220 mg/dL (day 24)	*Temporal Misalignment*; *Unsupported Assumptions*	TR → EF → DW → QC
ad_37	Rollercoaster high–low–high	*Unsupported Assumptions*	DW → QC → CA
ad_38	Spike at 2AM consistency check	*Temporal Misalignment*	TR → DW → QC
ad_39	Abnormal post-meal responses last week	*Unsupported Assumptions*	TR → EF → DW → QC
ad_40	Overnight implausible stability	*Unsupported Assumptions*	AR → IC
ad_41	Low readings flagged	✓	QC
ad_42	Low readings + abnormal pattern (sensor error)	*Data Misinterpretation*	AR → IC
ad_43	Data dropout intervals (connection lost)	*Reluctance to Calculate*	AR → IC
ad_44	Wrong meal registrations	*Unsupported Assumptions*	EF → QC
ad_45	Replay attack evidence	*Unsupported Assumptions*	AR → IC
ad_46	Dangerous controller ops during workout	*Unsupported Assumptions*	EF → DW → QC
ad_47	Suspicious biased readings	*Unsupported Assumptions*	AR → IC
ad_48	Unauthorized insulin pump access	*Unsupported Assumptions*	EF → QC
ad_49	Unexpected insulin doses	*Unsupported Assumptions*	EF → QC
ad_50	Saturation spoofing intervals	*Unsupported Assumptions*	AR → IC
ad_51	Faked readings via spoofing	*Unsupported Assumptions*	AR → IC
ad_52	Harmful pump basal extremes	*Unsupported Assumptions*	EF → QC
ad_53	Miscalibrated sensor	*Unsupported Assumptions*	AR → IC
ad_54	Pressure-induced attenuation	*Unsupported Assumptions*	AR → IC
ad_57	Sustained lowered concentration (≥6h)	*Temporal Misalignment*; *Reluctance to Calculate*	EF → DW → QC
ad_58	Insulin delivery stopped	*Unsupported Assumptions*	EF → QC
ad_59	Insulin appeared normal but stopped	*Unsupported Assumptions*	EF → QC
ad_60	Not enough insulin delivered	*Unsupported Assumptions*	EF → QC
ad_61	Delivered less than intended	*Unsupported Assumptions*	EF → QC
ad_62	Empty reservoir	*Unsupported Assumptions*	EF → QC
ad_63	Blocked/kinked infusion set	*Unsupported Assumptions*	EF → QC
ad_64	Insulin leakage / failed absorption	*Unsupported Assumptions*	EF → QC

Table 6: Failure analysis for AD questions (ad_33–64) with reasoning paths. ✓ indicates correct answers; ✓* indicates coincidentally correct answers based on assumptions/heuristics.

ID	Question (short)	Failure Type(s)	Reasoning Path(s)
pd_0	Predict time of day for highest glucose	*Unsupported Assumption*	QC → TR → PF
pd_1	Predict glucose level in 30 minutes	✓* (approximate)	QC → TR → PF
pd_2	Predict insulin consumption next Monday	*Data Misinterpretation*; *Unsupported Assumption*	QC → TR → DW → PF
pd_3	Will late-night snack push glucose >180 mg/dL?	*Unsupported Assumption*	QC → DW → EF → PF
pd_4	Predict % of time in range tomorrow	*Reluctance to Calculate*	QC → DW → PF
pd_5	Will running cause hypoglycemia in 90 min?	✓	QC → EF → PF
pd_6	Predict glucose 1h after breakfast	*Temporal Misalignment*	QC → TR → DW → PF
pd_7	Predict glucose change 15 min after run	*Temporal Misalignment*	QC → TR → DW → PF
pd_8	Predict glucose change 1h after cycling	*Temporal Misalignment*	QC → TR → DW → PF
pd_9	Which exercise lowers glucose more tomorrow?	✓	CA → EF → PF
pd_10	Predict if patient stays in range rest of day	✓	QC → DW → PF
pd_11	Predict spike time after heavy lunch	*Temporal Misalignment*	QC → TR → DW → PF
pd_12	Compare today vs tomorrow glucose average	*Unsupported Assumption*	QC → DW → PF
pd_13	Predict if correction insulin needed in 2h	*Data Misinterpretation*	QC → EF → PF
pd_14	Predict swing >60 mg/dL in next 4h	*Temporal Misalignment*	QC → TR → PF
pd_15	Should I eat a snack to stay in range overnight?	✓	QC → EF → PF
pd_17	Faster glucose rise: lunch or dinner?	✓* (assumption-heavy)	CA → EF → PF
pd_18	Will insulin need be > or < weekly average?	*Data Misinterpretation*	QC → DW → PF
pd_19	Predict glucose drop after insulin in 2h	*Data Misinterpretation*	QC → EF → PF
pd_20	Predict stable hours overnight (12–6AM)	*Reluctance to Calculate*	QC → DW → PF

Table 7: Failure analysis for PD questions with reasoning paths. ✓ indicates correct answers; ✓* indicates coincidentally correct answers based on assumptions/heuristics.