

Article

An IoT-Ready Context-Aware Patient State Framework with LLM-Driven Recommendation for Shoulder Rehabilitation

Jonghyeok Mun ¹, Nackhwan Kim ² and Jongsun Choi ^{3,*}¹ ACE Team, BankwareGlobal, Seoul 04513, Republic of Korea; jonghyeok.mun@bankwareglobal.com² Institute of Human Behavior and Genetics, Korea University College of Medicine, Seoul 02841, Republic of Korea; nackhwan@gmail.com³ Department of Computer Software, Dong Seoul University, Seongnam-si 13117, Republic of Korea

* Correspondence: jongsun.choi@du.ac.kr

Abstract

In Internet-of-Things (IoT)-based rehabilitation, patient data from wearable sensors (IMU, EMG, heart rate), clinical assessments, and surveys differ in format and granularity, complicating unified patient-state construction. Existing AI-based systems often omit fatigue level or rehabilitation stage, or they place large language models (LLMs) in the clinical decision-making role, exposing patients to hallucination risk. We propose an IoT-ready framework comprising three components: an input-source-independent context abstraction pipeline, a digital twin that simulates clinical-score trajectories, and an LLM Agent that interprets the outputs of a deterministic algorithm and a digital twin as (subject, predicate, object) Triplets to produce a retrieval-augmented clinical report. The algorithm—not the LLM—selects the 13-exercise sequence from Shoulder Pain and Disability Index (SPADI) item-level responses. Explicit per-layer schemas let IoT-sensor branches be added without changing the downstream interface. Here we evaluate the framework on the clinical-score path (three patient-reported outcome measures and six range-of-motion measures); the multi-modal IoT branches maintain deployment-target functionality. On 48 IRB-approved shoulder rehabilitation patients (144 longitudinal records; augmented to 7200 only to train the trajectory generator), real-only leave-one-subject-out evaluation gave VAS RMSE 0.658 (0–10) and SPADI RMSE 6.499 (0–100). Ablation across four surface forms revealed a fidelity–accuracy trade-off—Narrative highest on template fidelity (BERTScore F1 0.250), raw JSON highest on judge-rated accuracy—and the framework adopts the balanced-midpoint Triplet form. A claim-level audit of the Triplet-form reports left 22–31% of atomic claims unsupported (Claude Sonnet and GPT-4o judges) regardless of guideline retrieval. A raw-LLM control never reproduced the algorithm-defined sequence exactly (0 of 48), supporting deterministic, auditable sequence selection and the need for clinician review before clinical use.



Academic Editors: Francisco Gómez-Rodríguez, Francisco Luna-Perejón, Lourdes Miró Amarante, Javier Civit Masot and Luis Muñoz

Received: 29 May 2026

Revised: 17 July 2026

Accepted: 22 July 2026

Published: 27 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: Internet of Things (IoT); digital twin; shoulder rehabilitation; large language model (LLM) agent; clinical report generation

1. Introduction

The demand for rehabilitation services continues to grow as populations age and chronic diseases become more prevalent. For patients with musculoskeletal or neurological disorders, systematic rehabilitation exercise directly contributes to motor function recovery and return to daily activities [1]. However, conventional rehabilitation services depend

on in-person assessments by clinicians, limiting the ability to continuously incorporate changes in a patient's condition into exercise prescription.

Advances in wearable sensors and Internet of Things (IoT) technologies can enable real-time collection of biosignals and movement data through heart rate monitors, Inertial Measurement Units (IMUs), and electromyography (EMG) sensors, while clinical assessments and patient surveys capture subjective symptoms, functional level, and pain severity. Jameil et al. [2] proposed a digital twin framework for real-time IoT-based healthcare monitoring, and subsequent studies [3,4] have reported sensor data-driven exercise recommendation methods. Nevertheless, three limitations persist when integrating these heterogeneous inputs for rehabilitation protocol recommendation.

First, individual data items—heart rate (bpm), acceleration (m/s^2), joint angle (degrees), and survey response scores—are single-dimensional numerical values from which one cannot determine whether a patient is fatigued or whether rehabilitation is progressing on track. Second, data collected from different sources—IOT sensors, clinical assessments, and patient surveys—vary in collection frequency, format, and semantic unit, making it difficult to integrate them into a single coherent representation of patient state. Third, many existing AI-based rehabilitation recommendation systems either feed raw data directly into models or rely on a single metric (e.g., heart rate zone) for exercise selection, failing to account for fatigue level, rehabilitation stage, or accumulated exercise history [5,6].

Digital twin technology models and synchronizes the state of a physical entity in a digital environment in real time. Originating in manufacturing [7], it has been extended to healthcare [8,9]. Applying digital twins to patients enables physiological status, activity level, and rehabilitation stage to be represented as a unified multi-dimensional state vector derived from heterogeneous data sources, including IoT sensors, clinical assessments, and patient surveys. Bidirectional synchronization—comprising simulation-based state evolution and feedback from recommendation outcomes—allows continuous tracking of the patient's current condition.

Recent generalist medical AI foundation models [10] have demonstrated broad clinical capabilities; however, their tendency toward hallucination and the difficulty of auditing their outputs remain barriers to safe clinical deployment. Several studies have explored the use of Large Language Models (LLMs) as agents across multiple domains [11]. An LLM Agent retrieves evidence from external knowledge bases via Retrieval-Augmented Generation (RAG), performs database queries and computations through tool invocation [12], and synthesizes these inputs to generate natural-language explanations. When applied to rehabilitation protocol prescription, however, having an LLM directly determine the exercise sequence renders the system vulnerable to hallucination [13], and recommendations drift with each model substitution, weakening auditability. A safer complementary architecture restricts the LLM to interpreting the output of a deterministic recommendation algorithm, separating clinical decision-making from natural-language generation. This separation aligns with the safety expectations articulated in the FDA's *Software as a Medical Device* (SaMD) guidance [14], yet no prior study has integrated it with IoT-driven digital twin patient-state modeling.

Building on the concept of context-aware computing [15], this study proposes a framework that integrates heterogeneous patient data through an input-source-independent context abstraction pipeline, models the patient state as a digital twin, and produces rehabilitation recommendations using a deterministic protocol algorithm and a digital twin simulation, while restricting the LLM Agent to interpreting the simulation output as a natural-language clinical report. The context abstraction pipeline transforms patient data into low-level context (activity intensity, movement patterns) and high-level context (fatigue level, rehabilitation stage). The digital twin patient state model integrates these into a

multi-dimensional state vector. A deterministic protocol algorithm then derives a sequence of 13 target exercises from the patient's SPADI item-level responses, and the digital twin simulates the trajectory of clinical measures under that sequence. The LLM Agent receives the simulation outcome as a set of (subject, predicate, object) Triplets and produces a natural-language clinical report grounded in evidence retrieved through RAG, without determining the type, order, or intensity of exercises itself. In this paper the framework is instantiated and evaluated on the clinical-score path of this pipeline (patient-reported outcome scales and range-of-motion measures); the IoT-sensor branches are defined by the same layer schemas, and their empirical validation is left as deployment-stage work.

The main contributions of this paper are as follows:

1. We propose a digital twin patient state model that integrates physiological status, activity status, and rehabilitation progress into a multi-dimensional state vector with simulation-based state evolution and a feedback loop, and we evaluate its clinical-score instantiation across nine clinical measures (three patient-reported outcome scales and six range-of-motion angles) in experiment E1.
2. We design an input-source-independent context abstraction pipeline whose explicit per-layer schemas transform patient data from low-level context (activity intensity, movement patterns) into high-level context (fatigue level, rehabilitation stage). The prototype realizes the clinical-score path of this pipeline; the IoT-stream branches are specified as deployment-target extensions admitted by the same schemas and are not empirically evaluated here.
3. We propose an LLM-agnostic framework in which a deterministic protocol algorithm and a digital twin simulation jointly produce the recommendation, and an LLM Agent interprets the simulation outcome into a natural-language clinical report grounded in RAG-retrieved evidence. We validate its feasibility through experiments E2 and E3 on 144 longitudinal records (augmented to 7200) from 48 patients undergoing shoulder rehabilitation under IRB approval.

The remainder of this paper is organized as follows. Section 2 reviews related work on IoT-based rehabilitation systems, healthcare digital twins, and LLM Agents. Section 3 describes the overall architecture and components of the proposed framework. Section 4 presents the experimental design, results, and analysis. Section 5 concludes the paper and discusses future research directions.

2. Related Work

2.1. IoT-Based Healthcare and Rehabilitation Monitoring

IoT-based remote patient monitoring (RPM) combines wearable sensors with cloud architectures to collect biosignals and support clinical decisions. Shaik et al. [5] reviewed the canonical “sensor → model → alert” RPM pipeline, and Baiense et al. [6] documented in a systematic review that the field remains fragmented across hardware and vendor ecosystems. In the rehabilitation context, Oh et al. [16] applied a 1D-CNN with axis-rotation augmentation to IMU data from 15 hemiparetic stroke patients, achieving approximately 11% average F1 improvement, and Huang et al. [4] combined a 5-IMU lower-limb rehabilitation system with an LSTM and dynamic-time-warping similarity scoring; integrated all-in-one wearable platforms further consolidate multiple physiological channels on commodity hardware, such as a smartphone-based photonic system for plantar and cardiorespiratory monitoring [17]. These prior works are typically bound to fixed sensor configurations and feed raw signals directly into classifiers, outputting a single label (activity class or normal/abnormal flag). This work introduces an input-source-independent context abstraction pipeline designed to integrate IoT sensors, clinical assessments, and patient surveys into a

unified state vector through explicit per-layer schemas; this input-source independence is a design property, and the present paper evaluates the clinical-score path of the pipeline (Section 3.2).

2.2. Digital Twins in Healthcare

The digital twin paradigm originated in manufacturing [7] and has been extended to healthcare [8], with patient-level applications becoming visible in the last three years. Mikołajewska et al. [9] presented a conceptual framework for rehabilitation patient digital twins but did not define an explicit state vector or transition function. Khan et al. [18] implemented bidirectional synchronization in a robot-assisted telerehabilitation system; however, the synchronization is tied to a specific robotic platform's telemetry, and the system supports therapist-driven teleoperation rather than automated recommendation generation. Makarov et al. [19] recently proposed DT-GPT, which uses an LLM to forecast patient health trajectories but relies on end-to-end generation without RAG or tool calls, offering passive prediction rather than actionable exercise recommendation. This work explicitly defines a multi-dimensional state vector $s(t)$ with a transition function $s(t+1) = f(s_t, u_t, w_t)$, and it produces exercise recommendations through a deterministic protocol algorithm coupled with the digital twin simulation while restricting the LLM Agent to interpreting the simulation outcome as a RAG-grounded natural-language clinical report.

2.3. LLM Agents and Evidence-Grounded Medical Recommendation

Yao et al. [11] established the ReAct paradigm, in which an LLM interleaves reasoning with tool calls so that subsequent reasoning is conditioned on tool observations; Schick et al. [20] showed an alternative path in Toolformer, which internalizes external API invocation into the model weights through self-supervised learning. In the medical domain, Zakka et al. [21] applied RAG to a clinical QA system that outperformed a general-purpose ChatGPT baseline on factuality and completeness across 130 clinical scenarios, and Kim et al. [22] proposed MDAgents, which adaptively routes queries between solo and multi-agent configurations based on complexity. Closer to the present context-framework methodology, other clinical pipelines retrieve heterogeneous structured and unstructured knowledge to ground LLM decision support [23] or use an LLM for context-aware augmentation of structured patient records [24]. In the recommendation domain, Huang et al. [25] introduced InteRecAgent, which wraps traditional recommenders as tools orchestrated by an LLM. However, these agents typically consume raw text (e.g., clinical questions, dialog history) and leave state abstraction to the LLM's implicit reasoning. They also place the recommendation or clinical decision itself in the LLM output, which exposes the system to hallucination [13] and weakens auditability because the recommendation drifts with each model substitution. This work differs from the above in three respects. The recommendation, produced by a deterministic protocol algorithm and a digital twin simulation, is decoupled from the LLM Agent's role, which is restricted to interpreting that simulation outcome in natural language. An input-source-independent context abstraction layer is placed upstream of the agent. And tool access is standardized through the Model Context Protocol [12], preventing the agent from accessing the framework's internal logic directly. As a result, clinical recommendations are deterministic and auditable at the algorithm level, and the recommended sequence is invariant to LLM model substitution, yielding a genuinely LLM-agnostic design.

2.4. Context-Aware Recommendation

Foundations and recent neural CARS. Context-aware recommendation (CARS) has been built on the three paradigms—contextual pre-filtering, post-filtering, and contextual modeling—formalized by Adomavicius and Tuzhilin [26]. Representative recent neural

CARS include the multi-context multi-criteria recommender of Afzal et al. [27], a deep learning model that performs multi-criteria utility prediction across a fixed context schema, and the temporal preference-transition extension of matrix factorization by Jeong and Kim [28], which models how user preferences shift over time. LLM-augmented CARS has emerged in two directions: Dai et al. [29] serialized user history as text and probed ChatGPT’s ability to rank candidate items directly (point-, pair-, and list-wise) but provided no retrieval grounding or tool invocation; Huang et al. [25], reviewed in Section 2.3, instead placed the LLM as an orchestrator over traditional recommenders called tools.

Why classical CARS does not transfer to clinical rehabilitation. Clinical rehabilitation recommendation differs from the consumer-recommendation systems above in ways that prevent direct transfer. The user state in rehabilitation is longitudinal and clinical (pain, ROM, disability scales over weeks) rather than a session-level interaction history, and it is reconstructed from heterogeneous IoT, clinical-assessment, and survey inputs that consumer CARS systems do not handle. Liu et al. [3], the closest health-adjacent example, addressed fitness recommendation from Fitbit streams but this was coupled to a specific sensor schema and did not abstract a multi-dimensional patient-state representation. The recommendation must also be auditable at the algorithm level for downstream clinical use, whereas the CARS systems above place the personalization decision inside an embedding model or an LLM and become difficult to audit when the underlying model changes. Finally, an LLM in this setting is more useful as an explainer of an externally produced clinical recommendation than as the recommender itself, a role that classical CARS does not consider. The framework proposed in this paper is positioned at the intersection of these three properties: it constructs a longitudinal multi-dimensional patient state from heterogeneous inputs, derives the 13-exercise sequence through a deterministic protocol algorithm so that the recommendation is auditable independent of the LLM, and uses the LLM only to interpret the digital twin simulation outcome of that sequence in natural language.

2.5. Comparison and Positioning

Table 1 compares the proposed framework with representative prior work along the four axes introduced above.

Table 1. Comparison of representative prior work with the proposed framework across the four axes of input-source independence, state integration, LLM/ Agent usage, and personalization mechanism. Rows are ordered by thematic grouping rather than publication year.

Work	Input Indep.	State Repr.	LLM/Agent Usage	Personalization
Shaik et al. [5]	Partial	Raw	None (DL)	Single-metric
Oh et al. [16]	None	Raw	None (CNN)	Single-metric
Mikołajewska et al. [9]	Partial	Raw	None	Single-metric
Khan et al. [18]	None	Raw	None	Fixed
Makarov et al. [19]	Partial	Raw tokens	LLM w/o grounding	Single-metric
Yao et al. [11]	Partial	Raw	ReAct (LLM + tools)	None
Zakka et al. [21]	Strong (text)	Raw	LLM + RAG (single-shot)	None
Huang et al. [25]	Partial	Raw	LLM + tools (ReAct-adj.)	Single-memory
Liu et al. [3]	None	Raw	None	Embedding
Dai et al. [29]	Partial	Raw	LLM w/o grounding	LLM ranking
This work	Strong (design)	Abstracted multi-dim	LLM reporter + RAG + MCP tools	Multi-dim state

As the table shows, none of the reviewed frameworks combines input-source independence, a multi-dimensional state vector, and an LLM Agent restricted to interpreting algorithmic and digital twin outputs through RAG-grounded tool invocation. Section 3 addresses this gap in the rehabilitation protocol recommendation domain by separating

the deterministic, auditable recommendation from the LLM, whose role is restricted to natural-language clinical reporting.

3. Proposed Framework

3.1. System Overview

The proposed framework transforms patient data collected from input sources (IoT sensor streams, clinical assessments, and patient surveys) into a unified state representation, produces recommendations through a deterministic protocol algorithm and a digital twin simulation, and employs an LLM Agent to interpret the simulation outcome as a natural-language clinical report. The framework comprises three components. First, the context abstraction pipeline converts input data into low-level context and high-level context (Section 3.2). Second, the digital twin patient state model integrates the transformed context into a multi-dimensional state vector and simulates trajectories of clinical measures under the exercise sequence produced by the deterministic protocol algorithm (Section 3.3). Third, the LLM Agent receives the simulation trajectory as a set of (subject, predicate, object) Triplets and produces a natural-language clinical report grounded in RAG-retrieved evidence; the agent accesses the algorithm and simulation only through standardized tool interfaces (Section 3.4). The overall data flow is shown in Figure 1.

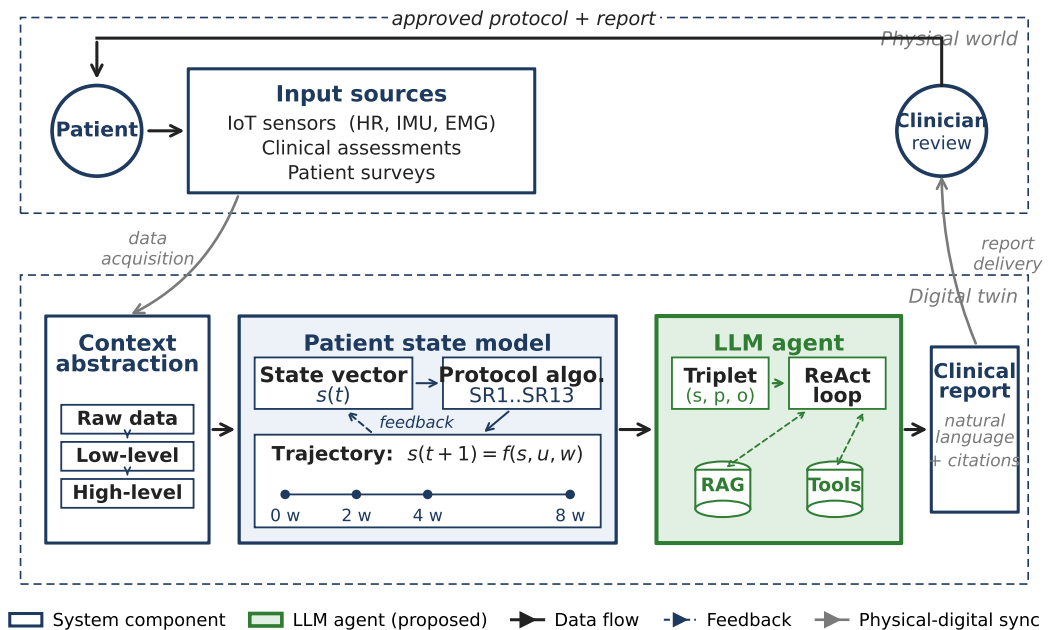


Figure 1. Overall architecture of the proposed framework. Heterogeneous input sources (IoT sensors, clinical assessments, and patient surveys) are unified by the context abstraction pipeline into low-level and high-level context. The digital twin patient state model integrates the high-level context into a multi-dimensional state vector $s(t)$ and updates it through a state transition function. A deterministic protocol algorithm derives a sequence of 13 exercises from the patient state, and the digital twin simulates the trajectory of clinical measures under that sequence. The simulation trajectory is abstracted into a set of (subject, predicate, object) Triplets, which the LLM Agent interprets into a natural-language clinical report through a ReAct loop with RAG-grounded evidence. The agent invokes the algorithm and simulation only through standardized tool interfaces and does not determine the exercise sequence itself. Bidirectional synchronization between the physical patient and the digital state is maintained by the outer cycle: the clinician-reviewed recommendation is delivered to the patient, whose exercise outcomes and biosignals are subsequently captured by the input sources and reflected back into the state vector.

3.2. Context Abstraction Pipeline

In IoT-based rehabilitation, input sources are heterogeneous. IoT devices such as heart rate monitors, Inertial Measurement Units (IMUs), and electromyography (EMG) sensors use different collection frequencies and value formats, while clinical assessments and patient surveys are recorded in natural language or scored form. The proposed framework transforms these heterogeneous inputs into a structured representation of the patient state through an input-source-independent context abstraction pipeline. The pipeline comprises three layers—the Raw Data Layer, the Low-level Context Layer, and the High-level Context Layer—where the input and output of each layer are defined by explicit schemas. The overall structure is shown in Figure 1.

The Raw Data Layer accepts raw data collected from input sources and classifies each input into one of three data types. A *simpleType* holds a single scalar value (e.g., heart rate, step count, speed); a *complexType* holds structured data (e.g., three-axis IMU acceleration, multi-channel ECG); and an *unstructuredType* holds non-tabular data (e.g., posture images, audio, patient survey responses). Each raw observation is stored with metadata describing its length, format, and unit; this enables subsequent layers to operate independently of the specific input source. The input is raw data from input sources, the processing consists of metadata tagging and type classification, and the output is typed raw observations.

The Low-level Context Layer transforms typed raw observations into quantitative context that is numerically interpretable. The transformation is performed per input type. Posture images are converted into Range of Motion (ROM) values after skeleton coordinate extraction; three-axis IMU acceleration is mapped to movement intensity; and step counts from a smartwatch are aggregated into time-windowed activity levels. The input is typed raw observations, the processing consists of per-type transformation functions (coordinate extraction, signal feature computation, aggregation, and the like), and the output is low-level context such as activity intensity, movement patterns, and ROM.

The High-level Context Layer integrates low-level context to derive clinically meaningful high-level state. Fatigue level is estimated from temporal changes in activity intensity and heart rate variability; rehabilitation stage is inferred by combining the ROM trajectory with rehabilitation history. The input is sequences of low-level context, the processing consists of temporal integration and clinical rule application, and the output is high-level context such as fatigue level and rehabilitation stage.

The final output of the pipeline serves as the input to the digital twin patient state model described in Section 3.3.

Scope of the prototype implementation. The architecture above describes the input-source-independent design across all three data types. The prototype evaluated in Section 4 implements the clinical-scoring path only: the three patient-reported outcome scales and six range-of-motion measures detailed in Section 4.1 enter the Raw Data Layer as *simpleType* observations and are passed through to the High-level Context Layer as patient-state features for the digital twin. The IMU signal-processing path (acceleration → movement intensity), the posture-image path (skeleton extraction → ROM), and the heart-rate-variability fatigue estimator are part of the deployment-target architecture and are not exercised in the experiments of this paper. The IRB-approved cohort dataset (Section 4.1) does not contain IoT sensor streams, and full multi-modal validation is left as deployment-stage work. The framework's input-source independence claim is therefore a design property (each layer is defined by explicit schemas that admit additional input types without changing the downstream interface), not an empirical claim of this paper.

3.3. Digital Twin Patient State Model

The digital twin patient state model takes the output of the context abstraction pipeline as input, integrates patient state data into a multi-dimensional state vector, and continuously updates it through bidirectional synchronization comprising simulation-based state evolution and exercise outcome feedback.

State vector definition. The patient state vector $s_t \in \mathbb{R}^d$ comprises three parts: physiological status, activity status, and rehabilitation progress. Physiological status includes fatigue-related indicators derived from heart rate variability and activity intensity; activity status includes movement patterns and ROM variability; and rehabilitation progress includes exercise history and temporal trajectories of clinical measures (e.g., SPADI, K-DASH, VAS, and six ROM metrics).

State transition function. The transition from time t to $t + 1$ is defined as follows:

$$s_{t+1} = f(s_t, u_t, w_t)$$

where u_t is the rehabilitation protocol applied during the interval—the 13-exercise sequence produced by the deterministic protocol algorithm from the patient state (Section 3.4 step 1)—and w_t is process noise capturing individual variability and measurement uncertainty. The function f is realized by per-feature random-forest regressors [30] trained independently for each of the nine clinical measures (three patient-reported outcome measures and six range-of-motion measures). In the leave-one-subject-out backend comparison of Section 4.3.1 (Table 3), TabPFN [31]—explored in prior work—ranked best on all nine measures (mean rank 1.00, mean %RMSE 4.02%), but it requires a proprietary license token and therefore cannot be part of a fully reproducible pipeline. We consequently adopt random forest, the best-performing license-free backend (lowest mean %RMSE and best mean rank among the six alternatives Ridge, ExtraTrees, GradientBoosting, HistGradientBoosting, XGBoost, and CatBoost), and we leave TabPFN as a drop-in option for licensed users. Each model takes the patient profile and the prior state vector as input and predicts the measure values at the next time step. The current implementation does not directly condition each per-feature regressor on the exercise sequence u_t itself; this is a known simulator limitation discussed in Section 4.4 (Limitation 5). As a consequence, the trajectory generated under a fixed algorithm output should be interpreted as state evolution over the rehabilitation interval rather than as a counterfactual simulation comparing alternative protocols. This restricts the use of the simulator in this paper to a single-protocol forward roll-out, which is sufficient for the feasibility experiments in Section 4 but not for protocol-vs-protocol comparison.

State estimation pipeline. State updates follow a standardized five-stage procedure: observation preprocessing (*preprocess*), control input augmentation (*augment*), core prediction (*core estimate*), output post-processing (*postprocess*), and validity checking (*validate*). The procedure is summarized in Algorithm 1, where each stage is defined as a replaceable component through an abstract interface. This design allows different prediction models to be combined across clinical measures while preserving a consistent state update flow.

Bidirectional synchronization. The state model is synchronized with the physical patient through two update channels.

- *Physical → Digital (sensor synchronization):* Data collected from input sources are converted to high-level context through the context abstraction pipeline and then reflected in the state vector. This update is performed asynchronously at the sensor sampling rate and the clinical assessment schedule.
- *Digital → Physical (feedback loop):* The rehabilitation protocol produced by the deterministic protocol algorithm and the natural-language clinical report generated by the LLM Agent (Section 3.4) are delivered to the patient after clinician review, and the

patient's exercise outcomes are collected as observations and reflected back into the state vector. This update is performed at the end of each exercise session.

Algorithm 1: State estimation pipeline

Input: previous state s_t , observation o_{t+1} , control input u_t

Output: updated state s_{t+1}

$o_{\text{norm}} \leftarrow \text{PREPROCESS}(o_{t+1});$

$o_{\text{aug}} \leftarrow \text{AUGMENT}(o_{\text{norm}}, u_t);$

$s_{\text{new}} \leftarrow \text{COREESTIMATE}(o_{\text{aug}}, s_t);$

$s_{\text{new}} \leftarrow \text{POSTPROCESS}(s_{\text{new}});$

$\text{VALIDATE}(s_{\text{new}});$

return s_{new}

Iterative simulation. The state transition function is applied iteratively at the clinical assessment cadence (baseline, week 2, week 4, and week 8) to simulate long-term rehabilitation trajectories. The framework uses an open-loop simulation over the fixed sequence produced by the deterministic protocol algorithm; the resulting trajectory is then handed to the LLM Agent for natural-language interpretation (the input representation, in particular the Triplet abstraction, is defined in Section 3.4). Closed-loop operation—in which clinician-validated patient outcomes feed back into the next cycle's algorithm input—is supported by the same architecture and is activated at clinical-deployment time.

3.4. LLM Agent-Based Clinical Reporting

The role of the LLM Agent is to interpret simulation outputs, not to make clinical decisions. In the proposed framework, the type, order, and intensity of exercises are produced by a deterministic protocol algorithm operating on the patient state, and the resulting trajectories of clinical measures under that protocol are simulated by the digital twin patient state model (Section 3.3). The LLM Agent's role is restricted to (1) synthesizing algorithm and simulation outputs into a natural-language clinical report, (2) grounding the report in evidence retrieved through RAG, and (3) orchestrating tool invocations to access the algorithm and simulation through standardized interfaces.

This role separation aligns with general safety constraints on medical AI—for example, the clinical-evaluation expectations articulated in the FDA Software as a Medical Device guidance [14] (an architectural-alignment observation, not a regulatory clearance)—and structurally prevents the LLM from altering the exercise sequence itself, the pathway by which an LLM hallucination [13] could otherwise reach the patient; residual hallucination in the report narrative is a separate concern quantified in Section 4.3.4 (Limitation 6).

Workflow. A single recommendation cycle for one patient comprises the following seven steps.

1. **Protocol generation (algorithm).** A deterministic protocol algorithm produces a sequence u_t of 13 exercises (SR1 through SR13) from the patient's SPADI item-level responses (N1–N13), using the deterministic SR pain-ranking formula from our previous work [32].
2. **Trajectory simulation (DT).** The digital twin state model (Section 3.3) takes u_t as a control input and generates trajectories for nine clinical measures (VAS, SPADI, K-DASH, and six ROM measures) over the assessment cadence (baseline $\rightarrow$ week 2 $\rightarrow$ week 4 $\rightarrow$ week 8).
3. **Result abstraction (Triplet).** The simulation trajectory is converted into a set of (subject, predicate, object) Triplets—for example, (AnalysisResult, hasAverageEffect,

“10%”), (AnalysisResult, isGoalAchievable, “true”), (AnalysisResult, hasMaximumEffectMetric, “SPADI”).

4. **LLM reporting.** Given the Triplet set and patient context (demographics, baseline state), the LLM Agent generates a natural-language clinical report covering (i) a summary of the simulation outcome, (ii) the clinical interpretation of the recovery trajectory, and (iii) the rationale for the prescribed exercises.
5. **Evidence grounding (RAG).** The LLM retrieves relevant chunks from a rehabilitation guideline corpus (AAOS, AMA criteria, clinical protocol) and adds inline citations (e.g., [ref_aaos_rom]) to the report.
6. **Tool invocation.** The LLM accesses the algorithm module and digital twin simulation through a standardized Model Context Protocol interface [12]; it does not access the framework’s internal logic directly.
7. **Clinician review (human-in-the-loop).** The final report is reviewed by a clinician before being delivered to the patient.

Structural safety and LLM-agnostic design.

- Because the LLM does not determine exercise order, clinical recommendations remain deterministic and auditable at the protocol-algorithm level.
- The recommended sequence is invariant under LLM model substitution, realizing a genuinely LLM-agnostic framework.
- Patient raw data passes through the algorithm and the digital twin, reaching the LLM only as abstracted Triplets, minimizing direct exposure of identifiable patient information when an external LLM API is used.

4. Experiments

4.1. Dataset

The dataset was collected at Korea University Anam Hospital as part of a shoulder rehabilitation project. Data collection was approved by the Korea University Anam Hospital Institutional Review Board (IRB No. 2025AN0595, approved 9 December 2025; valid through 30 November 2026), and written informed consent was obtained from all participants.

The cohort comprised 48 patients undergoing shoulder rehabilitation. Each patient underwent a baseline evaluation and three follow-up assessments (week 2, week 4, and week 8), yielding 144 longitudinal records. The rehabilitation protocol consisted of a personalized sequence drawn from 13 target exercises. Two categories of variables were recorded. Patient profile attributes comprised age, sex, height, weight, smoking status, and alcohol consumption. Clinical measures comprised three patient-reported outcome measures (PROMs)—a 0–10 shoulder-pain rating (recorded as “VAS (NRS)” in the source dataset and abbreviated VAS hereafter; a 0–10 range is consistent with a numeric rating scale), the Shoulder Pain and Disability Index (SPADI; 0–100) [33], and the Korean version of the Disabilities of the Arm, Shoulder, and Hand questionnaire (K-DASH; 0–100) [34,35]—together with six range-of-motion (ROM) indicators (flexion, abduction, extension, adduction, internal rotation, and external rotation).

Because the 144 longitudinal records alone were insufficient for training the multi-step trajectory generator used in E2/E3, the dataset was augmented to 7200 records through a three-stage pipeline: weekly preprocessing expanded the 48 baseline records to 144 longitudinal records, a CatBoost ensemble regression model [36] was trained on the preprocessed data, and the trained model generated additional synthetic records with injected noise. To prevent data leakage, the train/test split was performed before augmentation.

Because the augmentation pipeline is a prior-work product, we validated the quality of its output rather than the generation process. (i) *Marginals*: the augmented follow-up values were statistically indistinguishable from the real observed follow-ups for seven of the nine measures (two-sample Kolmogorov–Smirnov, $p > 0.05$), deviating significantly only for VAS and ROM_ADD. (ii) *Input–output structure*: the baseline-to-follow-up correlation in the augmented data closely tracked that of the real data (e.g., ROM_FLEX 0.986 vs. 0.986), preserving the autoregressive recovery relationship. (iii) *Range*: injected noise pushed a fraction of input values outside the valid clinical range (VAS 10.6%, ROM_ADD 17.3%). (iv) *Utility*: on an 80/20 holdout, a predictor trained on augmented rows did not improve accuracy on held-out real patients over real-only training (mean RMSE difference +0.76; real-only was better on seven of nine measures). Accordingly, every accuracy figure reported in this paper (Table 4) is computed under real-only LOSO with no augmentation; augmentation was used only to train the E2/E3 trajectory generator, and because E2/E3 compare surface representations of the same trajectory and make no trajectory-accuracy claim (Section 4.2.5), this choice does not affect their conclusions.

The corresponding statements required by MDPI (Institutional Review Board Statement and Informed Consent Statement) are provided at the end of the manuscript.

4.2. Experimental Setup

4.2.1. Evaluation Metrics

Because the LLM Agent does not determine exercise order, sequence-ranking metrics (PAS, Kendall's τ , Hit@K) do not apply to its interpreter role, and the primary evaluation target is the semantic quality of the LLM-generated clinical report. These sequence-ranking metrics are reported only for an auxiliary control baseline (Section 4.4) in which a raw LLM is asked to produce the exercise sequence directly, to quantify the failure mode that the role separation avoids; the full per-condition metrics are available in the released result files.

Primary metrics: **BERTScoreF1** [37] (DeBERTa-xl-large-mnli [38], baseline-rescaled): token-level embedding similarity between the LLM report and a reference report. We follow the semantic-similarity evaluation in our previous work [32] (Section 4.3.9 of that dissertation) and reproduce it on the IRB-approved cohort reported in this paper. **Sentence-Transformer Cosine Similarity** [39] (all-MiniLM-L6-v2): document-level embedding cosine similarity.

Both metrics directly compare candidate reports against the same reference; differences between input-form variants therefore reflect differences in the LLM's interpretation rather than in the underlying simulation.

Secondary metrics: **RAG citation rate**: the number of inline citations of retrieved chunks per patient under the RAG-on configuration. This confirms that the RAG mechanism operates as designed. **Word count**: the report-length distribution. All reports fall within the 250–350-word band, so BERTScore comparisons are not confounded by length differences. **DT trajectory plausibility rate**: out-of-clinical-range rate and reverse-direction step rate of simulated trajectories. Reported as part of E1.

All 95% confidence intervals are estimated using 1000-replicate bootstrap resampling [40].

4.2.2. Experimental Design (E1/E2/E3)

The experiments are organized into three sub-experiments, each mapping to one framework contribution.

- **E1—Digital Twin State Prediction and Simulation Reliability**: We report held-out RMSE of the random-forest-based state predictor under leave-one-subject-out (LOSO) 48-fold evaluation [41], disaggregated across nine clinical measures and three follow-up timepoints, along with the trajectory out-of-range and reverse-direction rates,

to quantify the feasibility of the state transition function f and simulation stability (Contribution #1).

- **E2—Effect of Input Representation on LLM Interpretation:** We measure BERTScore F1 and Cosine Similarity differences when the same simulation outcome is fed to the LLM in four surface forms: (i) the Triplet abstraction, (ii) a Quintuple form that extends the Triplet with explicit (timepoint, source) provenance metadata, (iii) a deterministic Narrative prose summary pre-rendered from the same fields, and (iv) the raw trajectory JSON (Table 2). This validates the contribution of the explicit context abstraction layer and probes the marginal effect of stronger surface abstraction on LLM interpretation quality (Contribution #2).
- **E3—RAG-Grounded Citation Effect and Clinical Scenario Validation:** We compare Triplet-input reports with and without RAG to measure the BERTScore and Cosine impact of evidence retrieval, and we report the RAG citation rate as a mechanism check. We additionally include an LLM-as-judge evaluation [42] in which Claude Sonnet rates the Haiku reports for accuracy, completeness, and evidence clarity, and we present input/output illustrations for three patient profiles (Contribution #3).

Table 2. Evaluated variants in the context-representation ablation matrix. The proposed configuration is Triplet + RAG. Quintuple extends the Triplet form with explicit (timepoint, source) provenance metadata; Narrative pre-renders the same content as a deterministic prose summary; and JSON is the raw trajectory form. All four surface representations carry the same underlying information (patient context, four-timepoint trajectory, protocol summary, AAOs-normal goal-achievement flags); only the surface form fed to the LLM differs.

Variant	Input Form	RAG	Verification Target
Triplet + RAG (proposed)	(subject, predicate, object)	on	Reference
Triplet	(subject, predicate, object)	off	RAG effect (Contribution #3)
Quintuple	(subject, predicate, object, timepoint, source)	off	Marginal effect of adding temporal/provenance metadata to Triplet
Narrative	deterministic prose summary pre-rendered from the same fields	off	Effect of stronger surface abstraction (Triplet → natural-language prose)
JSON + RAG	raw trajectory JSON	on	Auxiliary input-form ablation
JSON	raw trajectory JSON	off	Input-form ablation (Contribution #2)

Bold indicates the proposed Triplet + RAG configuration.

4.2.3. Evaluated Variants (Context-Representation Ablation Matrix)

The Random, Frequency, and Rule baselines used in the previous design were intended for sequence-alignment comparison. They do not occupy a baseline slot in the present evaluation: with the underlying simulation outcome held constant across variants, the comparison axis is the surface form in which that outcome is fed to the LLM. These sequence baselines therefore appear in Section 4.3 only as auxiliary checks on the output of the protocol algorithm.

4.2.4. Reference Reports and Their Limitations

The reference report is a clinical report synthesized deterministically from the simulation Triplets via a fixed template. All LLM output variants are compared against the same reference, so differences across variants reflect differences in LLM generation rather than in the underlying simulation. However, because the reference is template-generated rather than physician-authored, absolute scores have limited clinical interpretability (see Section 4.4 Limitations). To partially compensate, we include an LLM-as-judge evaluation in Section 4.3.3 as a complementary indicator.

4.2.5. Implementation Details

- **Evaluation cohort and the two roles of the state predictor:** The state predictor is used in two distinct roles that rely on different data splits, and the two must not be conflated. (i) *Accuracy evaluation (E1, Table 4)*. Held-out RMSE is measured under leave-one-subject-out (LOSO) 48-fold cross-validation on the original 48-patient cohort; every reported error is therefore a genuine held-out prediction, and these are the only predictive-accuracy figures claimed in this paper. (ii) *Trajectory-input generation (E2/E3)*. To supply the trajectory that the LLM interprets in E2/E3, a single predictor is trained once on the 38-patient train split of an 80/20 holdout and then generates trajectories for all 48 patients; the 38 training patients' trajectories are thus in-sample. This does not affect the E2/E3 conclusions, which compare surface representations of the *same* trajectory and make no trajectory-level accuracy claim. The protocol algorithm and the digital twin simulation are deterministic per patient.
- **LLM:** Claude Haiku (claude-haiku-4-5-20251001) for report generation; Claude Sonnet 4.6 for the LLM-as-judge auxiliary evaluation.
- **Invocation:** Headless `claude -p` mode with OAuth authentication (no external API key). This setup keeps the LLM evaluation reproducible without introducing an additional API-cost dependency, and is consistent with the framework's principle of treating the LLM as a swappable interpreter. The deployment-target architecture (Section 3.4, step 6) describes live MCP tool invocation; in the present experiments, the protocol algorithm and the digital twin simulation are executed offline and their outputs are injected as a pre-computed Triplet set in the LLM system prompt. Run-time MCP integration is therefore an architectural target rather than an empirically validated path in the current experiments.
- **RAG corpus:** Twenty chunks drawn from AAOS shoulder rehabilitation guidelines, AMA functional disability criteria, and the dissertation protocol document. Retrieval uses BM25Okapi [43].
- **Simulation:** Three-step autoregressive application from baseline through week 2, week 4, and week 8, feeding the predicted state at one step as the input at the next.
- **Statistics:** 95% CIs are estimated by 1000-replicate bootstrap resampling [40]. Direction of paired effects is confirmed by Wilcoxon signed-rank tests on per-patient differences (E2 BERTScore F1, E2 Cosine, and the three E3 judge metrics).

4.3. Results

4.3.1. E1—Digital Twin State Prediction and Simulation Reliability

Table 3 compares eight regression backends under the same LOSO protocol. TabPFN—explored in prior work—ranked first on all nine measures (mean rank 1.00, mean %RMSE 4.02%), but it requires a proprietary license token and cannot be part of a fully reproducible pipeline. Among the seven license-free backends, random forest achieved the lowest mean %RMSE (5.18%) and the best mean rank and was selected as the state-predictor backend. Ridge was competitive—best on several individual measures—but random forest was preferred for its consistent performance across all nine measures. Table 4 reports the per-feature per-timepoint RMSE of the selected predictor under LOSO 48-fold evaluation. The mean RMSE over three follow-up timepoints was 0.658 for VAS (scale 0–10), 6.499 for SPADI (scale 0–100), and 8.419 for K-DASH (scale 0–100). Among ROM measures, ROM_ADD showed the smallest absolute RMSE (mean 1.219°), while ROM_FLEX (mean 7.310°) and ROM_ABD (mean 6.743°) were the largest. RMSE at the 8-week timepoint was equal to or slightly higher than at 4 weeks, consistent with the doubled assessment interval (4 weeks vs. 2 weeks) amplifying prediction uncertainty. Across all nine measures, the relative RMSE—expressed as a percentage of the scale range (or, for ROM measures, the

AAOS normal value)—ranges from 3.0% (ROM_ADD) to 8.4% (K-DASH) (Table 4, %RMSE column). For shoulder ROM, the two largest errors remain ROM_FLEX and ROM_ABD (mean RMSE 7.310° and 6.743°); we therefore position the predictor at the feasibility level—adequate to generate plausible state trajectories for the LLM reporting stage—while it is not yet suitable for fine-grained clinical change detection, which requires prospective validation. A per-feature per-timepoint RMSE heatmap is shown in Figure 2.

Table 3. State-predictor backend selection: mean %RMSE and mean rank across the nine clinical measures under the same LOSO 48-fold protocol on the original (non-augmented) data. TabPFN ranks best on every measure but requires a proprietary license token; RandomForest, the best license-free backend, is adopted for a fully reproducible pipeline.

Backend	Mean %RMSE	Mean Rank
TabPFN (license-gated)	4.02%	1.00
RandomForest (adopted)	5.18%	3.44
Ridge	5.58%	3.56
ExtraTrees	5.49%	4.11
XGBoost	5.22%	4.44
GradientBoosting	5.59%	5.22
CatBoost	6.83%	6.89
HistGradientBoosting	7.36%	7.33

Table 4. LOSO 48-fold held-out RMSE per clinical measure and follow-up timepoint (random-forest backend). Patient-reported outcome measures (VAS, SPADI, K-DASH) and six ROM angles (degrees) are reported. %RMSE is computed as the mean RMSE divided by the AAOS normal value of each measure (ROM) or by the scale range (VAS, SPADI, K-DASH) as a coarse normalization for cross-measure comparison.

Feature	Scale	Week 2	Week 4	Week 8	Mean	%RMSE
VAS pain	0–10	0.623	0.686	0.664	0.658	6.58%
K-DASH disability	0–100	6.780	8.929	9.549	8.419	8.42%
SPADI	0–100	6.678	6.327	6.494	6.499	6.50%
ROM Forward Flexion	0–180°	6.914	7.413	7.602	7.310	4.06%
ROM Extension	0–60°	2.224	2.374	3.004	2.534	4.22%
ROM Abduction	0–180°	6.443	6.814	6.974	6.743	3.75%
ROM Adduction	0–40°	0.903	1.046	1.709	1.219	3.05%
ROM Internal Rotation	0–70°	3.293	3.413	3.864	3.523	5.03%
ROM External Rotation	0–90°	4.374	4.262	4.952	4.529	5.03%

Plausibility of simulated trajectories. Three-step autoregressive simulations from baseline through week 8 were generated for all 48 patients using the single trajectory-generation predictor of Section 4.2.5 (the same trajectories consumed by the LLM in E2/E3); this plausibility check therefore probes the stability of the deployment-configuration trajectories rather than the LOSO held-out accuracy reported in Table 4. The three patient-reported outcome measures (VAS, SPADI, K-DASH) remained inside their valid clinical ranges across every patient and every timepoint, with zero recovery-direction reversals. Five of the six ROM measures (FLEX, EXT, ABD, IR, ER) showed out-of-range rates of 0–0.5% and reversal rates of 0–0.7%. The exception was ROM_ADD with 8.9% out-of-range and 13.2% reversal. All 17 out-of-range values are baseline measurements—not predictions—that exceed the AAOS-normal validity cap for adduction ($40^\circ \times 1.05 = 42^\circ$), so they originate in the inherited baseline data rather than in the simulation (a rollout figure distinct from the augmented-input range check in Section 4.1). Because anatomic adduction is limited by the torso, these elevated baseline values more likely reflect a measurement- or coding-convention inconsistency in the source dataset than genuine anatomical variation; whether to re-scope the validity range or exclude the affected baselines is a question for

source-record review. The genuine simulator limitation is instead the 13.2% reversal rate (19 of 144 trajectory steps, median -7.1°), which arises because the per-feature regressors do not enforce monotonic recovery (Section 4.4). The simulations are sufficiently stable to serve as input to the LLM reporting stage for the framework’s feasibility check.

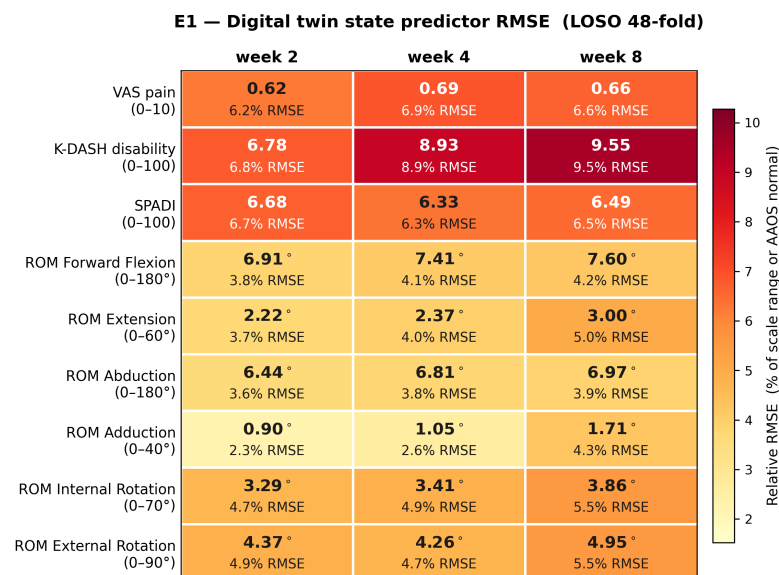


Figure 2. E1: LOSO held-out RMSE per clinical measure and follow-up timepoint. Rows: nine clinical measures; columns: three follow-up timepoints (week 2, week 4, week 8). Color intensity indicates RMSE magnitude within each measure’s own scale.

4.3.2. E2—Effect of Input Representation on LLM Interpretation

The same simulation outcome was fed to the LLM (Haiku 4.5) in four surface forms—(i) the Triplet abstraction, (ii) a Quintuple form that extends Triplet with explicit (timepoint, source) provenance, (iii) a deterministic Narrative prose summary pre-rendered from the same fields, and (iv) the raw trajectory JSON—and the resulting reports were compared against the same template reference using BERTScore F1 and Cosine Similarity (Table 5); the per-form score distributions are shown in Figure 3. All four input forms encode the same information units (patient context, four-timepoint trajectory, protocol summary, AAOS-normal goal-achievement flags); differences across variants therefore reflect the surface-representation effect rather than information content.

Table 5. E2—four context-representation forms, semantic similarity vs. reference report ($n = 48$, 95% bootstrap CI). BERTScore [37] uses DeBERTa-xlarge-mnli [38] with baseline rescaling; cosine similarity uses Sentence-Transformer [39] all-MiniLM-L6-v2. All four input forms encode the same information (patient context, four-timepoint trajectory, protocol summary, AAOS-normal goal-achievement flags); the gap therefore reflects representation effect rather than information content. Reference is the rule-based template report (Section 4.2).

Input Form	BERTScore F1	Cosine Similarity
Triplet (proposed)	0.223 [0.216, 0.230]	0.727 [0.711, 0.741]
Quintuple	0.226 [0.219, 0.233] ^{n.s.}	0.742 [0.729, 0.757] ^{n.s.}
Narrative	0.250 [0.243, 0.257] ***	0.728 [0.711, 0.745] ^{n.s.}
JSON (raw)	0.215 [0.208, 0.222] ^{n.s.}	0.713 [0.696, 0.731] ^{n.s.}

Bold indicates the proposed Triplet form. Paired Wilcoxon vs. Triplet: *** $p < 10^{-3}$, ^{n.s.} $p \geq 0.05$. $n = 48$ patient pairs.

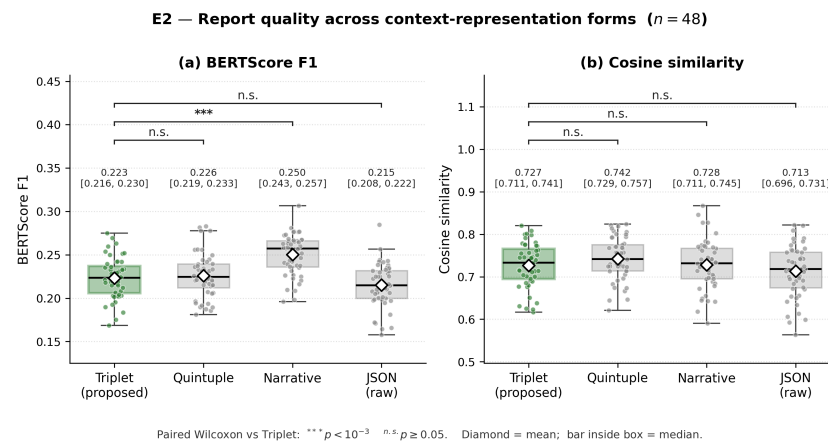


Figure 3. E2: BERTScore F1 and Cosine Similarity distributions across 48 patients for the four context-representation forms (Triplet, Quintuple, Narrative, JSON). Box plots show median, IQR, and whiskers within $1.5 \times \text{IQR}$. All three abstracted forms score above raw JSON; Narrative reaches the highest BERTScore F1 (the only statistically significant template-fidelity gain) by surface-level alignment to the prose reference, while Triplet and Quintuple sit at a balanced point on the abstraction–fidelity trade-off (Section 4.3.3).

The Triplet condition reached BERTScore F1 0.223 [95% CI 0.216, 0.230] and Cosine 0.727 [0.711, 0.741], compared with 0.215 [0.208, 0.222] and 0.713 [0.696, 0.731] for raw JSON. Triplet is numerically higher than JSON on both metrics, but the per-patient paired Wilcoxon signed-rank tests are not significant in this cohort (BERTScore F1: $W = 431$, $p = 0.109$; Cosine: $W = 510$, $p = 0.43$; $n = 48$). Quintuple matched Triplet within sampling noise (BERTScore F1 0.226, Cosine 0.742; Wilcoxon vs. Triplet $p = 0.64$ on F1, $p = 0.21$ on cosine), indicating that explicitly annotating temporal and provenance metadata on top of the Triplet structure yields no measurable marginal gain on template-fidelity metrics. The Narrative form, by contrast, significantly raised BERTScore F1 to 0.250 [0.243, 0.257] over the Triplet baseline (Wilcoxon $W = 177$, $p = 8.3 \times 10^{-6}$) while remaining tied on cosine (0.728, $p = 0.77$). This BERTScore gain is consistent with the Narrative prose exposing a prose-level surface structure that aligns directly with the prose-style template reference at the token level; the only statistically significant template-fidelity gain over Triplet is therefore the Narrative form’s.

Every abstracted surface form (Triplet, Quintuple, Narrative) scores numerically above raw JSON on template-fidelity metrics, though only the Narrative gain is statistically significant; this indicates that the explicit context-abstraction layer of Contribution #2 is not specific to a single surface choice, while the statistically confirmed improvement is concentrated in the prose form. The marginal effect of richer abstraction is non-uniform: adding (timepoint, source) tags to Triplet yields essentially no change, whereas pre-rendering as prose increases BERTScore F1 substantially. Combined with the LLM-as-judge results in Section 4.3.3, where stronger surface abstraction trades against raw factual fidelity, this locates Triplet at a balanced point on the abstraction–fidelity trade-off curve.

The direction is consistent with the Triplet vs. JSON results reported in our previous work [32] (BERTScore 0.71 vs. 0.77, Cosine 0.79 vs. 0.89). The absolute BERTScore F1 values (0.21–0.25) are lower than typical same-domain text comparisons because the reference is a fixed template-generated report rather than a clinician-authored narrative; this evaluation therefore targets relative comparison across input forms, not absolute clinical-quality estimation. Absolute scoring against a clinician-authored reference is left as the follow-up study discussed in Section 4.4.

4.3.3. E3—LLM-as-Judge Auxiliary Evaluation

Whereas BERTScore and Cosine Similarity measure semantic agreement against the reference template, the LLM-as-judge auxiliary evaluation [42] directly assesses factual fidelity to the simulation outcome. Claude Sonnet 4.6 rates each Haiku report on a 1–5 Likert scale across accuracy, completeness, and evidence clarity (Table 6). To make the comparison fair across input forms, the judge is given the full trajectory (baseline/week 2/week 4/week 8) together with the patient demographics as the ground truth—the same factual basis applies to all four input forms.

Table 6. E3—LLM-as-judge ratings of Haiku reports by Claude Sonnet 4.6 against full-trajectory + patient-demographics ground truth ($n = 48$, 1–5 Likert, 95% bootstrap CI). Ground truth (baseline/week 2/week 4/week 8 + demographics) is provided identically to all four conditions. Accuracy decreases monotonically with surface-abstraction strength (JSON $\rightarrow$ Quintuple $\rightarrow$ Triplet $\rightarrow$ Narrative): forms that expose raw numbers make verbatim factual citation easier, while pre-rendered prose paraphrases values and loses raw-fidelity points. Completeness and evidence clarity vary within a narrower band (~ 0.3 points).

Input Form	Accuracy	Completeness	Evidence Clarity
Triplet (proposed)	3.65 [3.42, 3.90]	3.23 [3.10, 3.35]	2.96 [2.81, 3.10]
Quintuple	3.96 [3.77, 4.19] ^{n.s.}	3.23 [3.12, 3.35] ^{n.s.}	3.04 [2.96, 3.15] ^{n.s.}
Narrative	3.25 [3.04, 3.46] *	3.38 [3.23, 3.52] ^{n.s.}	3.04 [2.92, 3.17] ^{n.s.}
JSON (raw)	4.00 [3.79, 4.23] *	3.48 [3.33, 3.62] *	3.27 [3.15, 3.42] **

Bold indicates the proposed Triplet form. Paired Wilcoxon vs. Triplet: ** $p < 10^{-2}$, * $p < 0.05$, ^{n.s.} $p \geq 0.05$. Sonnet 4.6 judge; $n = 48$ patient pairs.

Mean accuracy decreases monotonically with surface-abstraction strength: JSON 4.00 > Quintuple 3.96 > Triplet 3.65 > Narrative 3.25. Relative to Triplet, JSON is about 0.35 points higher (paired Wilcoxon $W = 161.5$, $p = 0.025$, $n = 48$; 33 non-tied pairs) and Narrative is about 0.40 points lower ($W = 104.5$, $p = 0.012$; 29 non-tied pairs). Quintuple is statistically equivalent to Triplet on accuracy ($p = 0.089$), indicating that explicitly tagging temporal and provenance metadata does not measurably help the LLM produce factually verifiable claims. Completeness and evidence clarity vary within a narrower band across the four conditions (completeness 3.23–3.48, evidence clarity 2.96–3.27); on these two dimensions JSON also scored significantly above Triplet (completeness $W = 139.5$, $p = 0.028$; evidence clarity $W = 40.0$, $p = 0.004$), whereas Quintuple and Narrative were statistically equivalent to Triplet. In short, raw JSON—the least abstracted form—was rated highest by the Sonnet judge on all three verifiable dimensions, at the cost of the lowest template fidelity (Section 4.3.2).

Combined with the E2 BERTScore and Cosine results, these ratings trace a clear trade-off curve. Stronger surface abstraction raises template-fidelity (BERTScore F1) but lowers raw-fidelity (judge accuracy): JSON is strongest on raw fidelity but weakest on template fidelity, Narrative is the opposite, and Triplet and Quintuple sit at the balanced midpoint. The framework adopts the Triplet form as default because (i) it satisfies the clinical demand for templated-structure conformance while losing less raw fidelity than Narrative, (ii) abstracting upstream of the LLM reduces direct exposure of raw patient data when external LLM APIs are used (privacy considerations in Section 4.4), and (iii) the Quintuple ablation confirms that adding provenance metadata provides no marginal benefit, leaving the simpler Triplet form preferable on interpretability grounds.

Cross-judge robustness check (GPT-4o-2024-11-20). To assess possible single-judge bias, the same 192 Haiku reports were independently scored by OpenAI’s GPT-4o using the same prompt and ground truth (Table 7). On this task GPT-4o was substantially more lenient than Sonnet, with mean ratings compressed into the upper 4.88–5.00 band—assigning a

constant 5.00 on completeness for three of the four forms (Triplet, Quintuple, JSON). Inter-rater Spearman ρ on per-patient scores fell in the 0.04–0.31 range (mostly non-significant; undefined on completeness, where GPT-4o had zero variance), with quadratic-weighted Cohen's κ near zero, reflecting the Sonnet 2.96–4.00 vs. GPT-4o 4.9–5.0 calibration gap. However, the mean accuracy ranking across the four input forms was preserved under both judges: JSON and Quintuple at the top, Narrative at the bottom. This pattern is consistent with the known cross-family calibration drift of LLM-as-judge methods [42] and indicates that the qualitative trade-off finding (monotonic inverse relationship between surface-abstraction strength and raw factual fidelity) is not specific to the Sonnet judge. Absolute score interpretation is discussed further in Section 4.4.

Table 7. Cross-judge mean ratings ($n = 48$ each). GPT-4o-2024-11-20 was run as a second judge with the same prompt and ground truth as Sonnet 4.6. Absolute means differ substantially (Sonnet ~ 3.0 – 4.0 , GPT-4o ~ 4.9 – 5.0), reflecting known cross-family LLM-judge calibration drift, but the qualitative ranking across the four input forms is preserved.

Input Form	Sonnet 4.6			GPT-4o		
	Acc.	Comp.	Clarity	Acc.	Comp.	Clarity
Triplet (proposed)	3.65	3.23	2.96	4.94	5.00	4.88
Quintuple	3.96	3.23	3.04	4.96	5.00	4.92
Narrative	3.25	3.38	3.04	4.88	4.98	4.88
JSON (raw)	4.00	3.48	3.27	4.98	5.00	4.96

Bold indicates the proposed Triplet form. $n = 48$ per judge per condition. Means only (CIs and Wilcoxon tests are reported per judge in Table 6).

To verify that the RAG mechanism operates as designed, the Triplet condition was also run with RAG retrieval enabled. All 48 patient reports under the RAG-on configuration contained inline citations (mean 1.85 per patient, median 2), while the RAG-off configuration produced zero citations across all 48 patients, confirming the mechanism check (detailed numbers in Section 4.3.4).

4.3.4. RAG Citation Effect and Scenario Validation

RAG citation rate. Under Triplet + RAG, every one of the 48 patient reports included inline citations of guideline chunks (mean 1.85 per patient, median 2; e.g., [ref_aaos_rom], [principle_least_pain_first]). The RAG-off configuration produced zero citations across all 48 patients, directly verifying the RAG mechanism.

Claim-level factual audit (auxiliary analysis). To test whether guideline retrieval also improves factual grounding, we decomposed each report into atomic claims and labeled every claim, using a Claude Sonnet 4.6 auditor, as supported by the input Triplets, supported by a retrieved guideline chunk, or unsupported ($n = 48$ reports per configuration; per-claim labels are in the released result files). Retrieval did not lower the unsupported-claim rate. Without RAG, 22.1% of claims were unsupported (398 of 1798); with Triplet + RAG the rate was 25.9% (515 of 1989), and only 1.1% of claims were traceable to a retrieved chunk. Retrieval mainly raised claim density (37.5 vs. 41.4 claims per report on average) without improving grounding. Most unsupported claims are exercise-selection rationale or prognostic statements that the retrieval step does not constrain, but a minority directly contradict a specific input value (for example, a range-of-motion percentage attributed to the wrong movement, or a protocol length stated in weeks that does not match the 13-step sequence), so the figure is best read as an upper-bound indicator rather than a calibrated error count. Repeating the audit with an independent-family judge (GPT-4o) raised the absolute rate (30.8% without RAG, 31.2% with it) but reproduced the key finding that

retrieval did not lower it, so the direction is robust across judge families while the absolute level is judge-dependent (22–31%; Section 4.4, Limitation 6).

Scenario-based qualitative validation. End-to-end algorithm output → simulation trajectory → LLM report illustrations are provided in Section 4.3.5 for three patients spanning Functional Disability Classification (FDC; based on AMA functional-disability criteria, a range-of-motion-sum scale with Class 0–4) classes 1, 2, and 4. In every case, the LLM report reproduced the simulation Triplet values without omission, and the recovery wording differentiated between mild (“substantial improvement”) and severe (“meaningful recovery across all domains”) cases. Minor surface variations such as “13-week” vs. “13-session,” together with the protocol-rationale paragraphs discussed in Section 4.4 (Limitation 6), indicate that clinician review remains a necessary step before delivery to the patient.

4.3.5. End-to-End Patient Examples

Three end-to-end cases were inspected, spanning FDC class 1 (mild), 2 (moderate), and 4 (severe). Scenario A (patient ID 2386041, 45-year-old female, FDC 1) improved from baseline VAS 3.0/SPADI 14.6/K-DASH 26.0 to week 8 VAS 1.94 (−35.3%), SPADI 7.92 (−45.8%), and K-DASH 12.07 (−53.6%). Scenario B (patient ID 466197, 50-year-old male, FDC 2) moved from VAS 4.0 to 2.08 (−48.0%), SPADI from 41.5 to 31.9 (−23.2%), and K-DASH from 47.0 to 24.04 (−48.8%). Scenario C (patient ID 926027, 56-year-old female, FDC 4) moved from VAS 4.0 to 1.76 (−56.1%), SPADI from 36.2 to 24.58 (−32.0%), and K-DASH from 36.0 to 20.56 (−42.9%).

In all three cases the LLM report reproduced the four-timepoint trajectory values and the algorithm-determined 13-exercise sequence (SR1–SR13) without addition or omission. The Sonnet-based LLM-as-judge rated accuracy 4/5 and completeness 4/5 for scenarios A and B, and accuracy 5/5 for scenario C, quantitatively supporting that the generated report consistently reflects the deterministic algorithm output and the simulation outcome. The full reports for the three scenarios are reproduced in Appendix A.

4.4. Discussion

Implications of Role Separation. The framework produces recommendations from a deterministic algorithm and digital twin simulation while restricting the LLM to result interpretation. This separation yields three properties. Recommendations are deterministic and traceable to the patient’s SPADI responses, supporting clinical audit. The recommended sequence is invariant to LLM model substitution, yielding a system-level LLM-agnostic design. Patient raw data passes through the algorithm and the simulation and reaches the LLM only as abstracted Triplets, which minimizes exposure of identifiable information when an external LLM API is used. An auxiliary control baseline quantifies the failure mode this separation avoids: when a raw LLM is asked to produce the 13-exercise sequence directly, it does not reproduce the algorithm-defined reference ordering exactly (0 of 48 exact matches for both ablations), though it retains moderate rank agreement (Kendall $\tau \approx 0.51$; Hit@1 6–10%, Hit@3 $\approx 54\%$, Hit@5 $\approx 64\%$). Because the deterministic algorithm reproduces that reference ordering by construction, this comparison speaks to reproducibility and auditability, not to the clinical correctness of the ordering; supplying the digital twin predicted state raised Hit@1 (10.4% vs. 6.25%) but did not yield exact reproduction. These properties align with the safety and auditability expectations for medical AI articulated in the FDA Software as a Medical Device guidance [14] and are designed to reduce the hallucination risk that arises when an LLM is placed directly in a clinical decision role. These three properties concern process integrity—determinism and auditability of how a recommendation is produced—and are separate from the clinical correctness of the exercise

ordering itself, which is inherited from the underlying protocol algorithm and is addressed as a limitation below.

Contribution of Context Abstraction. Feeding the same information to the LLM as four surface forms—Triplet, Quintuple, Narrative, and raw JSON—produced opposite-direction signals on the two evaluation axes (Sections 4.3.2 and 4.3.3). On template fidelity to the reference clinical report, BERTScore F1 rose monotonically with surface-abstraction strength (JSON 0.215 < Triplet 0.223 $\approx$ Quintuple 0.226 < Narrative 0.250). On raw factual fidelity scored by the LLM-as-judge, accuracy moved monotonically in the opposite direction (JSON 4.00 > Quintuple 3.96 > Triplet 3.65 > Narrative 3.25). Because all four inputs carry the same information, this trade-off is attributable to the surface representation itself. The statistically significant endpoints of the two axes are Narrative (highest template fidelity, BERTScore F1 vs. Triplet $p = 8.3 \times 10^{-6}$) and JSON (highest judge accuracy, vs. Triplet $p = 0.025$), with Triplet and Quintuple occupying the statistically indistinguishable middle of both axes. A supplementary composite rubric averaging five clinical dimensions (faithfulness, patient-specific grounding, clinical safety, completeness, and readability; scored by the same Claude Sonnet 4.6 judge on all four forms plus the deterministic reference report, $n = 48$ each) reproduced the same rank order as the E3 accuracy score (Table 8): JSON 3.94 > Quintuple 3.76 > Triplet 3.55 > Narrative 3.50 $\approx$ reference 3.50 (composite means on the 1–5 scale). This indicates the ranking is not an artifact of the specific E3 accuracy-rubric wording; an independent GPT-4o judge, however, saturated near the ceiling on this rubric (per-form means 4.96–5.00, near-zero variance), reproducing the leniency/ceiling effect reported for E3, and so could not discriminate among forms. The ranking therefore rests on the more discriminating Sonnet judge. Triplet is numerically above raw JSON on template fidelity, though this gap is not significant in the present cohort (the significant gain is Narrative’s); it loses less raw fidelity than Narrative (judge accuracy 3.65: -0.35 from JSON, $+0.40$ over Narrative), and it is statistically indistinguishable from Quintuple on every metric—so the simpler Triplet form is preferred on interpretability grounds. The framework adopts the Triplet form by default: it satisfies the clinical demand for templated-structure conformance, preserves the privacy benefit of upstream abstraction (the external LLM API never sees raw patient data), and sits at the balanced midpoint of the abstraction–fidelity trade-off traced by the four-way comparison. Relative to raw JSON and free narrative specifically, the (subject, predicate, object) form makes each reported fact an explicitly enumerable unit that maps onto both the templated report structure and the RAG knowledge graph—which is precisely what enables the claim-level audit of Section 4.3.4—while remaining compact, whereas raw JSON is more verbose and narrative entangles facts in prose. As the accuracy axis shows, however, raw JSON’s literal values still yield the highest judge-rated factual fidelity, so the Triplet form is adopted as a balanced, not a dominant, choice. These results provide the experimental grounding for the measurable, representation-dependent effect of the explicit context-abstraction layer (Contribution #2).

Table 8. Auxiliary clinical-rubric evaluation: mean per-dimension and composite scores (1–5 scale) from the Claude Sonnet 4.6 judge for each surface form and the deterministic reference report ($n = 48$ each). The rank order reproduces the E3 accuracy ranking, indicating it is not an artifact of the specific E3 rubric wording. An independent GPT-4o judge saturated near the ceiling on this rubric (per-form means 4.96–5.00, near-zero variance) and could not discriminate among forms.

Form	Faithful.	Clin. Safety	Grounding	Complete.	Readabil.	Mean
JSON (raw)	3.79	4.73	3.00	3.46	4.71	3.94
Quintuple	3.65	4.33	2.96	3.46	4.42	3.76
Triplet (adopted)	3.23	4.02	2.77	3.42	4.33	3.55
Narrative	3.00	4.02	2.79	3.40	4.31	3.50
Reference (template)	3.40	4.38	2.58	3.04	4.10	3.50

Reliability of the Digital Twin Simulation. The random-forest state predictor—selected over six alternative backends in the LOSO comparison of Section 4.3.1—achieved within-scale RMSE on all nine clinical measures (e.g., VAS 0.658, SPADI 6.499), and the baseline → week 8 simulations stayed within valid clinical ranges for the patient-reported outcome measures across every patient and timepoint. The 8.9% out-of-range rate observed for ROM_ADD is not a simulation error but stems from baseline measurements above the AAOS-normal validity cap (42°), likely a measurement- or coding-convention artifact in the source data (Section 4.3.1). The genuine simulator limitation is the 13.2% reversal rate, which reflects that the per-feature regressors do not enforce monotonic recovery. The simulation is stable enough to serve as the LLM's upstream input (experimental support for Contribution #1).

RAG Mechanism Check. Under Triplet + RAG, an average of 1.85 inline citations per patient (median 2) were observed, compared with zero in the RAG-off condition, confirming that the retrieval pipeline operates as designed and that the LLM integrates retrieved chunks into its output. However, the quantitative impact of RAG on BERTScore is limited here because the reference text itself contains no inline citations.

Evaluation Limitation—Reference Report. BERTScore and Cosine in this study measure semantic similarity to a deterministic template-generated reference. This is appropriate for relative comparison across variant pairs (Triplet vs. JSON), but it does not directly estimate clinical excellence in absolute terms. The LLM-as-judge evaluation partially compensates, but absolute interpretation becomes more robust once a clinician-authored reference is evaluated on a sufficient sample. A follow-up study with clinician adjudication is therefore warranted.

Privacy and Security Considerations. The framework handles sensitive patient biosignals and clinical data; deployment requires encryption of data in transit, de-identification of stored records, and access-control management. Because the LLM in this design receives only abstracted Triplets that have already passed through the algorithm and the simulation, raw patient data is not directly transmitted to an external LLM API, which is favorable for privacy-preserving operation. Stronger guarantees through federated learning [44] and edge-based lightweight inference are proposed as future work.

Generalization to Other Conditions and Populations. The framework's boundary—a schema-defined context-abstraction layer feeding a deterministic protocol algorithm and a digital twin, with the LLM confined to interpretation—is not specific to shoulder rehabilitation. Porting it to another domain requires substituting three components: the protocol algorithm that maps patient state to a recommendation, the clinical-score set the digital twin predicts, and the guideline corpus the RAG stage cites; the layer interfaces and the LLM reporter are unchanged. Because the abstraction layer is defined by explicit per-layer schemas, additional input modalities—IMU/EMG streams (Section 3.2) or medical-imaging studies such as DICOM data analysis [45]—can be admitted without altering the downstream interface. The same pattern applies to other musculoskeletal protocols (e.g., knee or hip rehabilitation) and, more broadly, to chronic conditions with longitudinal clinical trajectories and protocolized management (e.g., cardiac, metabolic, or post-stroke neurorehabilitation). We emphasize, however, that each new domain requires its own validated protocol algorithm and an independent cohort for evaluation; generalization here is a property of the architecture, not an empirical claim of this single-cohort study.

Limitations. First, the evaluation was conducted on a single cohort (48 shoulder rehabilitation patients); generalization to other rehabilitation domains and larger cohorts requires additional study. Second, because the reference report is a deterministic template rather than a clinician-authored narrative, the absolute values of BERTScore and Cosine have limited clinical interpretability (partially mitigated by the LLM-as-judge auxiliary

evaluation). Third, the LLM report generator was limited to Haiku; further validation with stronger models such as Sonnet or Opus is left as future work. Fourth, the LLM-as-judge evaluation was cross-validated using two judges (Sonnet 4.6 and GPT-4o-2024-11-20). The qualitative accuracy ranking (JSON and Quintuple highest, Narrative lowest) was preserved under both judges, but per-patient inter-rater agreement was weak (Spearman ρ 0.04–0.31, mostly non-significant, and undefined on completeness where GPT-4o assigned a constant top score of 5.0; quadratic-weighted Cohen's κ near zero) because GPT-4o compressed its ratings into the upper end of the 1–5 scale (4.88–5.00) while Sonnet used a broader range (2.96–4.00). We therefore treat the more discriminating Sonnet judge as primary and GPT-4o as a leniency/ceiling check rather than an independent calibration. This is consistent with documented cross-family LLM-judge calibration drift [42], and clinician-panel adjudication is still required to anchor the absolute scores. Fifth, the per-feature digital twin regressors are not directly conditioned on the exercise sequence, limiting the simulator's ability to differentiate trajectories across protocols (a limitation also noted in our previous work [32]); this can be improved in future work by integrating exercise-sequence inputs into the trajectory model. Sixth, the protocol-rationale paragraphs in the LLM-generated reports (Appendix A) include exercise-specific clinical reasoning (“SR5 establishes foundational pain modulation,” etc.) that goes beyond what is strictly derivable from the input Triplets; a claim-level audit (Section 4.3.4) put this residual unsupported-claim rate at 22.1%/25.9% under a Claude Sonnet auditor and 30.8%/31.2% under an independent GPT-4o auditor (without/with retrieval); both judge families agree that retrieval did not reduce it, though the absolute level is judge-dependent (22–31%), and although most such claims are unverifiable rationale, a minority directly contradict an input value. This is a residual hallucination of the type the architecture is designed to constrain, indicating that the Triplet vocabulary should be extended with per-SR muscle and ROM bindings before the system is used without clinician review. Seventh, the deterministic protocol algorithm guarantees that recommendations are reproducible and auditable, but these are properties of the *process*, not of the *clinical correctness* of the exercise ordering. The algorithm is inherited from the first author's prior dissertation [32], where its clinical validity was likewise not empirically established and was explicitly identified as future work; this study therefore treats the algorithm as a fixed, auditable component and does not claim that its orderings are clinically optimal. Prospective clinician evaluation of the recommended sequences remains necessary and is a direct subject of future work. Eighth, the per-feature digital twin regressors do not enforce monotonic recovery, so a subset of trajectory steps show recovery-direction reversals (ROM_ADD: 13.2%, 19 of 144 steps, median -7.1°); this can be addressed with monotonicity constraints or stronger per-feature models.

5. Conclusions

This study demonstrated the feasibility of an LLM-agnostic, IoT-ready framework for shoulder rehabilitation that enforces a clean architectural boundary: a deterministic protocol algorithm and a digital twin model produce all clinical recommendations, while an LLM Agent is confined to interpreting those outputs as RAG-grounded natural-language reports. The framework comprises an input-source-independent context abstraction pipeline, a random-forest-based digital twin patient-state model coupled with a multi-step trajectory simulator, and an LLM Agent that consumes Triplet-form analysis results to produce a clinical report. The experiments validate the digital twin patient-state model and the LLM-as-reporter component on a single IRB-approved cohort using clinical-score inputs; full evaluation of the IoT-stream branches of the context abstraction pipeline is deferred to deployment-stage validation.

Leave-one-subject-out evaluation on 48 shoulder rehabilitation patients showed that the random-forest predictor produced within-scale trajectory simulation at the feasibility level (VAS RMSE 0.658, SPADI RMSE 6.499), and input-representation ablation traced a fidelity–accuracy trade-off across four surface forms: all three abstracted forms scored numerically above raw JSON on template fidelity, with a Narrative form attaining the highest template fidelity (BERTScore F1 0.250, the only statistically significant gain over Triplet) at the cost of the lowest raw factual accuracy (LLM-judge 3.25 vs. 4.00 for JSON), while the Triplet form sits at the balanced midpoint adopted by the framework. The RAG-on configuration produced an average of 1.85 inline guideline citations per patient versus zero under RAG-off, verifying the retrieval mechanism. A control baseline further showed that a raw LLM asked to select the exercise sequence directly never reproduced the algorithm-defined ordering exactly (0 of 48), supporting the decision to keep sequence selection in the deterministic layer.

Key limitations include the single-cohort scope (48 shoulder rehabilitation patients), reliance on a template-generated reference for the BERTScore evaluation, LLM evaluation restricted to the Haiku model, a residual unsupported-claim rate in the generated reports that guideline retrieval did not reduce (22–31% across Claude Sonnet and GPT-4o judges), and the fact that the protocol algorithm’s clinical correctness is inherited from prior work rather than established here. Future directions include follow-up evaluation with a clinician-authored reference, extension to additional rehabilitation domains, report-quality validation with stronger LLMs (Sonnet- or Opus-class), enhancement of the trajectory simulator with explicit exercise-sequence conditioning, and privacy-preserving deployment through federated learning [44] and edge inference.

Author Contributions: Conceptualization, J.M. and J.C.; methodology, J.M.; software, J.M.; validation, J.M. and N.K.; formal analysis, J.M.; investigation, J.M.; clinical resources and domain validation, N.K.; data curation, J.M.; writing—original draft preparation, J.M.; writing—review and editing, J.M., N.K., and J.C.; visualization, J.M.; supervision, J.C.; project administration, J.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (Grant number: RS-2026-25490832).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of Korea University Anam Hospital (IRB No. 2025AN0595, approval date 9 December 2025).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The IRB-approved patient dataset is not publicly available due to privacy and research-ethics restrictions. Aggregated results, augmented data, evaluation scripts, and trained model artifacts that support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments: During the preparation of this manuscript, the authors used Claude Opus (Anthropic) for English language proofreading and copy editing. The authors reviewed and edited all content and take full responsibility for the published article.

Conflicts of Interest: Author Jonghyeok Mun was employed by the company ACE Team, Bankware-Global. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Appendix A. End-to-End Patient Examples (Full Reports)

Three patient profiles spanning FDC class 1 (mild), 2 (moderate), and 4 (severe). Each scenario shows the deterministic algorithm output, the digital twin–

simulated trajectory, and the LLM-generated clinical report (Triplet input, no RAG, $n = 48$ v3 run with information-matched Triplets and full-trajectory ground truth for the LLM-as-judge).

Appendix A.1. Scenario A—FDC 1 (Mild Articular, Female, Age 45)

Patient ID: 2386041 **Protocol algorithm output:** SR5 → SR2 → SR4 → SR8 → SR9 → SR3 → SR12 → SR1 → SR6 → SR13 → SR11 → SR10 → SR7 **Baseline:** VAS 3.0, SPADI 14.6, K-DASH 26.0 **Week-8 simulation:** VAS 1.94 (−35.3%), SPADI 7.92 (−45.8%), K-DASH 12.07 (−53.6%) **LLM-as-judge ratings (Sonnet, full-trajectory ground truth):** accuracy 4/completeness 4/evidence_clarity 3

LLM clinical report (Haiku 4.5, Triplet input, full report):

Clinical Analysis Report: Digital Twin Simulation Outcome

This 45-year-old female with mild articular shoulder pathology presented with baseline pain (VAS 3.0) and moderate functional limitation (K-DASH 26.0, SPADI 14.62). The deterministic protocol algorithm generated a 13-exercise sequence initiating with SR5 and concluding with SR7. The digital twin simulation projects substantial recovery across pain and disability metrics by week 8, though with important caveats regarding full goal achievement.

The simulated trajectory demonstrates early and sustained gains in patient-reported outcomes. By week 2, VAS declined to 2.2 and K-DASH improved to 21.36, representing the steepest phase of recovery. ... Week-8 projections of VAS 1.94 (35.3% improvement), K-DASH 12.07 (53.6% improvement), and SPADI 7.92 (45.8% improvement). The K-DASH score shows the largest treatment effect ...

The SR5 initiation and SR7 conclusion were selected by the deterministic protocol to address the patient's mild articular condition and mixed pain-ROM profile. The prominent K-DASH improvement alongside the controlled ROM trajectory suggests SR5 establishes foundational pain modulation and stabilization, while SR7 likely maintains these functional gains while accepting strategic range limitations ...

Appendix A.2. Scenario B—FDC 2 (Moderate Bursal, Male, Age 50)

Patient ID: 466197 **Protocol algorithm output:** SR2 → SR4 → SR8 → SR9 → SR1 → SR5 → SR12 → SR3 → SR6 → SR13 → SR10 → SR11 → SR7 **Baseline:** VAS 4.0, SPADI 41.5, K-DASH 47.0 **Week-8 simulation:** VAS 2.08 (−48.0%), SPADI 31.90 (−23.2%), K-DASH 24.04 (−48.8%) **LLM-as-judge ratings:** accuracy 4/completeness 4/evidence_clarity 3

LLM clinical report:

Clinical Report: Rehabilitation Digital Twin Analysis

Patient Summary and Week-8 Outcome

This 50-year-old male presents with mild bursal-origin shoulder pain and functional limitation (FDC category 2). ... By week 8, pain is projected to decrease to 2.08 (48% improvement) and functional disability to improve substantially, with K-DASH declining to 24.04 (48.8% reduction) and SPADI to 31.9 (23.2% improvement).

Recovery Trajectory and Outcome Metrics

The simulated trajectory demonstrates rapid early improvement with gradual plateau. Pain reduction accelerates in the first two weeks (VAS 4.0 → 2.85), with continued but slower decay through weeks 4 and 8 (2.32 → 2.08). ... The

maximum predicted benefit appears in the K-DASH domain (48.8%), reflecting substantial restoration of upper-extremity function ...

Protocol Rationale

The SR2-through-SR7 sequence addresses the characteristic pathomechanics of bursal impingement. Initial exercises (SR2) likely establish pain-free scapular mobility ... Later exercises (SR7) reinforce external rotation and internal rotation capacity ... While the protocol achieves the pain-reduction goal, the persistence of adduction limitation and incomplete ROM normalization (2 of 6 measures) suggests functional recovery, though substantial, remains incomplete at eight weeks.

Appendix A.3. Scenario C—FDC 4 (Severe Articular, Female, Age 56)

Patient ID: 926027 **Protocol algorithm output:** SR8 → SR9 → SR2 → SR4 → SR5 → SR3 → SR12 → SR1 → SR13 → SR10 → SR6 → SR11 → SR7 **Baseline:** VAS 4.0, SPADI 36.2, K-DASH 36.0 **Week-8 simulation:** VAS 1.76 (−56.1%), SPADI 24.58 (−32.0%), K-DASH 20.56 (−42.9%) **LLM-as-judge ratings:** accuracy 5/completeness 3/evidence_clarity 3

LLM clinical report:

This 56-year-old female with severe articular shoulder pathology presents a favorable pain trajectory under the prescribed 13-exercise protocol. ... By week 8, pain decreased to 1.76 (56.1% improvement), with functional disability metrics declining to K-DASH 20.56 and SPADI 24.58, representing improvements of 42.9% and 32.0%, respectively.

The recovery trajectory demonstrates characteristic post-rehabilitation patterns: most substantial pain reduction occurred in the first two weeks (VAS 4.0 → 2.61), with continued but decelerated improvement through weeks 4 and 8. ... Extension mobility showed the most robust response, improving 62.6% from baseline (21° → 34.14°), while adduction demonstrated the most modest gains at 11.8% (25° → 27.94°). ... yielding an average improvement across all metrics of 34.9%.

The protocol's opening exercise (SR8) and closing exercise (SR7) were selected to address this patient's severity profile and motion restrictions. ... SR8, positioned early in the sequence, likely establishes foundational pain modulation and mobility tolerance to permit progressive loading. SR7's terminal placement reflects the rehabilitative principle of consolidating gains under increasing functional demand ...

While the simulated trajectory predicts achievement of the pain management goal, projections indicate that no individual range-of-motion measure reaches normative thresholds by week 8.

Appendix A.4. Observations

1. **Algorithm output diverges across patients**—Scenarios A and B start with SR5/SR2 (low joint-load) while C starts with SR8 (minimal joint-load) due to severe ROM restriction; the deterministic ordering responds to per-patient SPADI patterns. All three end with SR7 (broadest activation), consistent with the least-painful-first principle.
2. **Reports correctly mirror the underlying analysis**—VAS / SPADI/K-DASH numbers and intermediate-week values match the simulation Triplets without invention, supported by the LLM-as-judge accuracy scores (4/4/5 across the three scenarios).
3. **Severity-dependent narrative**—Reports differentiate “substantial recovery” (A, FDC 1) from “favorable but incomplete” (B, FDC 2) and “meaningful symptomatic relief ...

some residual motion restrictions” (C, FDC 4), reflecting the magnitudes captured in `hasAverageEffect`, `isPainGoalAchievable`, and `reachesNormalROMRatio` Triplets.

4. **Goal-achievability Triplets are correctly used**—In all three scenarios, the report explicitly states whether the pain goal is achieved and whether full ROM normalization is reached, matching the corresponding Triplet predicates.

Appendix A.5. Caveats

- Surface terminology in the reports occasionally drifts (e.g., the protocol is variously referred to as “13 weeks,” “13 sessions,” or “13 exercises”); this does not affect numeric content but reinforces the requirement for clinician review before delivery.
- The protocol-rationale paragraphs use general clinical language that goes beyond what is strictly derivable from the Triplets (e.g., “scapular mobility and posture correction”); a future iteration should extend the Triplet vocabulary with per-SR muscle and ROM bindings to ground these rationales more tightly.

References

1. Cieza, A.; Causey, K.; Kamenov, K.; Hanson, S.W.; Chatterji, S.; Vos, T. Global estimates of the need for rehabilitation based on the Global Burden of Disease study 2019: A systematic analysis for the Global Burden of Disease Study 2019. *Lancet* **2020**, *396*, 2006–2017. [CrossRef]
2. Jameil, A.K.; Al-Rawashidy, H. A digital twin framework for real-time healthcare monitoring: Leveraging AI and secure systems for enhanced patient outcomes. *Discov. Internet Things* **2025**, *5*, 37. [CrossRef]
3. Liu, X.; Gao, B.; Suleiman, B.; You, H.; Ma, Z.; Liu, Y.; Anaissi, A. Privacy-Preserving Personalized Fitness Recommender System P³FitRec: A Multi-Level Deep Learning Approach. *ACM Trans. Knowl. Discov. Data* **2023**, *17*, 1–24. [CrossRef]
4. Huang, Y.J.; Chang, C.S.; Wu, Y.C.; Han, C.C.; Cheng, Y.Y.; Chen, H.M. Development of Wearable Devices for Collecting Digital Rehabilitation/Fitness Data from Lower Limbs. *Sensors* **2024**, *24*, 1935. [CrossRef] [PubMed]
5. Shaik, T.; Tao, X.; Higgins, N.; Li, L.; Gururajan, R.; Zhou, X.; Acharya, U.R. Remote patient monitoring using artificial intelligence: Current state, applications, and challenges. *WIREs Data Min. Knowl. Discov.* **2023**, *13*, e1485. [CrossRef]
6. Baiense, J.P.; Zdravevski, E.; Coelho, P.J.; Pires, I.M.; Velez, F.J. Driving Healthcare Monitoring with IoT and Wearable Devices: A Systematic Review. *ACM Comput. Surv.* **2025**, *57*, 1–38. [CrossRef]
7. Tao, F.; Cheng, J.; Qi, Q.; Zhang, M.; Zhang, H.; Sui, F. Digital twin-driven product design, manufacturing and service with big data. *Int. J. Adv. Manuf. Technol.* **2018**, *94*, 3563–3576. [CrossRef]
8. Abd Elaziz, M.; Al-qaness, M.A.A.; Dahou, A.; Al-Betar, M.A.; Mohamed, M.M.; El-Shinawi, M.; Ali, A.; Ewees, A.A. Digital twins in healthcare: Applications, technologies, simulations, and future trends. *WIREs Data Min. Knowl. Discov.* **2024**, *14*, e1559. [CrossRef]
9. Mikołajewska, E.; Masiak, J.; Mikołajewski, D. Applications of Artificial Intelligence-Based Patient Digital Twins in Decision Support in Rehabilitation and Physical Therapy. *Electronics* **2024**, *13*, 4994. [CrossRef]
10. Moor, M.; Banerjee, O.; Abad, Z.S.H.; Krumholz, H.M.; Leskovec, J.; Topol, E.J.; Rajpurkar, P. Foundation models for generalist medical artificial intelligence. *Nature* **2023**, *616*, 259–265. [CrossRef] [PubMed]
11. Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; Cao, Y. ReAct: Synergizing reasoning and acting in language models. In Proceedings of the International Conference on Learning Representations (ICLR), Kigali, Rwanda, 1–5 May 2023.
12. Anthropic. Model Context Protocol (MCP) Specification. 2024. Available online: <https://modelcontextprotocol.io/> (accessed on 20 May 2026).
13. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* **2023**, *55*, 1–38. [CrossRef]
14. U.S. Food and Drug Administration. *Software as a Medical Device (SaMD): Clinical Evaluation*; FDA Guidance Document; FDA: Silver Spring, MD, USA, 2017.
15. Dey, A.K. Understanding and Using Context. *Pers. Ubiquitous Comput.* **2001**, *5*, 4–7. [CrossRef]
16. Oh, Y.; Choi, S.A.; Shin, Y.; Jeong, Y.; Lim, J.; Kim, S. Investigating Activity Recognition for Hemiparetic Stroke Patients Using Wearable Sensors: A Deep Learning Approach with Data Augmentation. *Sensors* **2024**, *24*, 210. [CrossRef] [PubMed]
17. Bao, Y.; Wang, Z.; Zhang, Q.; Xiao, K.; Tosi, D.; Massaroni, C.; Li, X.; Min, R. All-in-One: Smartphone-Based Intelligent Photonic mHealth System for Plantar and Cardiorespiratory Monitoring. *IEEE Internet Things J.* **2026**, *13*, 5393–5406. [CrossRef]

18. Khan, M.M.R.; Sunny, M.S.H.; Ahmed, T.; Shahria, M.T.; Modi, P.P.; Zarif, M.I.I.; Sanjuan, J.; Ahamed, S.I.; Ahmed, H.U.; Mcgonigle Ketchum, E.M.; et al. Development of a Robot-Assisted Telerehabilitation System With Integrated IIoT and Digital Twin. *IEEE Access* **2023**, *11*, 70174–70189. [[CrossRef](#)]
19. Makarov, N.; Bordukova, M.; Quengdaeng, P.; Garger, D.; Rodriguez-Esteban, R.; Schmich, F.; Menden, M.P. Large language models forecast patient health trajectories enabling digital twins. *npj Digit. Med.* **2025**, *8*, 588. [[CrossRef](#)] [[PubMed](#)]
20. Schick, T.; Dwivedi-Yu, J.; Dessì, R.; Raileanu, R.; Lomeli, M.; Hambro, E.; Zettlemoyer, L.; Cancedda, N.; Scialom, T. Toolformer: Language models can teach themselves to use tools. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, USA, 10–16 December 2023; Curran Associates, Inc.: New York, NY, USA, 2023; Volume 36, pp. 68539–68551.
21. Zakka, C.; Shad, R.; Chaurasia, A.; Dalal, A.R.; Kim, J.L.; Moor, M.; Fong, R.; Phillips, C.; Alexander, K.; Ashley, E.A.; et al. Almanac—Retrieval-Augmented Language Models for Clinical Medicine. *NEJM AI* **2024**, *1*, AIoa2300068. [[CrossRef](#)] [[PubMed](#)]
22. Kim, Y.; Park, C.; Jeong, H.; Chan, Y.S.; Xu, X.; McDuff, D.; Lee, H.; Ghassemi, M.; Breazeal, C.; Park, H.W. MDAgents: An adaptive collaboration of LLMs for medical decision-making. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, Vancouver, BC, Canada, 10–15 December 2024; Curran Associates, Inc.: New York, NY, USA, 2024; Volume 37, pp. 79410–79452. [[CrossRef](#)]
23. Lu, Y.; Zhao, X.; Wang, J. ClinicalRAG: Enhancing Clinical Decision Support through Heterogeneous Knowledge Retrieval. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, Bangkok, Thailand, 16 August 2024; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 64–68.
24. Hsieh, C.; Moreira, C.; Nobre, I.B.; Costa Sousa, S.; Ouyang, C.; Brereton, M.; Jorge, J.; Nascimento, J.C. DALL-M: Context-Aware Clinical Data Augmentation with Large Language Models. *Comput. Biol. Med.* **2025**, *190*, 110022. [[CrossRef](#)] [[PubMed](#)]
25. Huang, X.; Lian, J.; Lei, Y.; Yao, J.; Lian, D.; Xie, X. Recommender AI Agent: Integrating Large Language Models for Interactive Recommendations. *ACM Trans. Inf. Syst.* **2025**, *43*, 1–33. [[CrossRef](#)]
26. Adomavicius, G.; Tuzhilin, A. Context-aware recommender systems. *AI Mag.* **2011**, *32*, 67–80. [[CrossRef](#)]
27. Afzal, I.; Yilmazel, Ö.; Kaleli, C. An Approach for Multi-Context-Aware Multi-Criteria Recommender Systems Based on Deep Learning. *IEEE Access* **2024**, *12*, 99936–99948. [[CrossRef](#)]
28. Jeong, S.; Kim, Y.K. Deep Learning-Based Context-Aware Recommender System Considering Change in Preference. *Electronics* **2023**, *12*, 2337. [[CrossRef](#)]
29. Dai, S.; Shao, N.; Zhao, H.; Yu, W.; Si, Z.; Xu, C.; Sun, Z.; Zhang, X.; Xu, J. Uncovering ChatGPT's capabilities in recommender systems. In *Proceedings of the 17th ACM Conference on Recommender Systems (RecSys '23)*, Association for Computing Machinery, Singapore, 18–22 September 2023; pp. 1126–1132. [[CrossRef](#)]
30. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
31. Hollmann, N.; Müller, S.; Eggensperger, K.; Hutter, F. TabPFN: A transformer that solves small tabular classification problems in a second. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Kigali, Rwanda, 1–5 May 2023.
32. Mun, J. A Digital-Twin-Based Rehabilitation Exercise Recommendation Framework Using Enterprise Integration Patterns. Ph.D. Dissertation, Soongsil University, Seoul, Republic of Korea, 2025.
33. Roach, K.E.; Budiman-Mak, E.; Songsirdej, N.; Lertratanakul, Y. Development of a Shoulder Pain and Disability Index. *Arthritis Care Res.* **1991**, *4*, 143–149. [[CrossRef](#)]
34. Hudak, P.L.; Amadio, P.C.; Bombardier, C. Development of an upper extremity outcome measure: The DASH (Disabilities of the Arm, Shoulder, and Hand). *Am. J. Ind. Med.* **1996**, *29*, 602–608. [[CrossRef](#)]
35. Lee, J.Y.; Lim, J.Y.; Oh, J.H.; Ko, Y.M. Cross-cultural adaptation and clinical evaluation of a Korean version of the disabilities of arm, shoulder, and hand outcome questionnaire (K-DASH). *J. Shoulder Elb. Surg.* **2008**, *17*, 570–574. [[CrossRef](#)] [[PubMed](#)]
36. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased Boosting with Categorical Features. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, Montréal, QC, Canada, 3–8 December 2018.
37. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. BERTScore: Evaluating Text Generation with BERT. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Addis Ababa, Ethiopia, 26 April–1 May 2020.
38. He, P.; Liu, X.; Gao, J.; Chen, W. DeBERTa: Decoding-Enhanced BERT with Disentangled Attention. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Virtual, 3–7 May 2021.
39. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 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)*, Hong Kong, China, 3–7 November 2019; pp. 3982–3992. [[CrossRef](#)]
40. Efron, B. Bootstrap Methods: Another Look at the Jackknife. *Ann. Stat.* **1979**, *7*, 1–26. [[CrossRef](#)]
41. Wong, T.T. Performance evaluation of classification algorithms by *k*-fold and leave-one-out cross validation. *Pattern Recognit.* **2015**, *48*, 2839–2846. [[CrossRef](#)]

42. Zheng, L.; Chiang, W.L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.P.; et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), New Orleans, LA, USA, 10–16 December 2023*; Datasets and Benchmarks Track; Curran Associates, Inc.: New York, NY, USA, 2023; Volume 36, pp. 46595–46623.
43. Robertson, S.; Zaragoza, H. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.* **2009**, *4*, 1–174. [[CrossRef](#)]
44. McMahan, H.B.; Moore, E.; Ramage, D.; Hampson, S.; Agüera y Arcas, B. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*. PMLR, Fort Lauderdale, FL, USA, 20–22 April 2017; pp. 1273–1282.
45. Shakya, A.K.; Vidyarthi, A. Comprehensive Study of Compression and Texture Integration for DICOM Data Analysis. *Technologies* **2024**, *12*, 17. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.