

RuleEdit: Failure-Guided Human-AI Model Editing with Prospective Impact Preview

MIN HUN LEE and JUSTIN YU FENG TEO, Singapore Management University, Singapore

Despite the promise of AI to assist complex decisions, practitioners still lack ways to detect likely failures and inspect the consequences of model edits before committing them. We present RuleEdit, an interactive, rule-guided human-AI model editing system that (i) surfaces likely failures through interpretable mismatch signals from rule tables and (ii) supports user-authored rule feedback with prospective previews of projected performance changes and embedding shifts. We instantiate RuleEdit in stroke rehabilitation assessment and evaluate it with health professionals and students. Rule-guided failure detection significantly increased Human + AI performance by 14.16% ($p < 0.001$) while improving rejection of incorrect AI and reducing both over- and under- reliance as well as ChangedToWrong decisions. In addition, presenting prospective embedding previews improved participants' feedback for model adaptation, increasing post-update local performance gains from 11.50% to 36.38% after incorporating users' rule-based feedback ($p < 0.001$). Our findings show that mismatch-based failure cues and prospective impact previews can support failure-aware human-AI model editing, while also revealing a local-global tradeoff: edits that help a specific case can degrade performance when transferred globally. We discuss implications of designing failure-aware and controllable human-AI systems.

1 Introduction

Human-AI collaborative decision-making [10, 13, 14, 27, 37, 48, 66] examines how humans and AI systems can combine complementary strengths to improve decisions in high-stakes domains such as healthcare [13, 31, 48], law [51], and public services [18, 30, 36, 68]. To support end users, prior work has investigated data-driven scaffolds including explanations that highlight important features [48, 74], retrieve similar cases [13, 15], counterfactuals [43, 78], and uncertainty cues [49, 50, 57, 79]. These mechanisms aim to clarify why a model produced a given output [1, 4, 39, 58, 73] and to quantify confidence [49, 50, 57, 79] so that people can judge when to rely on AI.

While explanations and confidence cues can improve users' mental models of AI [7, 13, 43, 48, 57, 72], some studies show that explanations can inflate unwarranted trust and induce over-reliance on erroneous outputs [5, 10, 11, 43]. Achieving calibrated reliance and human-AI complementarity [5, 10, 27, 37] remains challenging. A core gap remains: most systems emphasize informational transparency showing explanations and scores, but provide limited support for two practical tasks. First, users often lack actionable cues for identifying when the AI is likely wrong on a specific case. Second, even when users want to intervene, they rarely receive previews of how a proposed edit may affect model behavior before it is committed.

To address this gap, we present RuleEdit, an interactive, failure-guided human-AI model editing system that surfaces likely failures through interpretable mismatch signals derived from rule tables and supports user-authored rule and label feedback for model editing. Beyond helping users inspect failures after an AI prediction is shown, RuleEdit supports anticipatory interaction by previewing projected performance changes and embedding shifts (on class centroids and data) before users decide whether to commit an edit.

We instantiate RuleEdit in physical stroke rehabilitation assessment and conduct a study with 21 participants including health professionals and students to investigate two research questions: (RQ1) How do rule-guided mismatch cues affect users' accept/override decisions and Human + AI team performance? and (RQ2) How does prospective impact preview affect the quality of user-authored feedback for model editing?

Authors' Contact Information: Min Hun Lee, mhlee@smu.edu.sg; Justin Yu Feng Teo, justin.teo.2022@scis.smu.edu.sg, Singapore Management University, Singapore, Singapore.

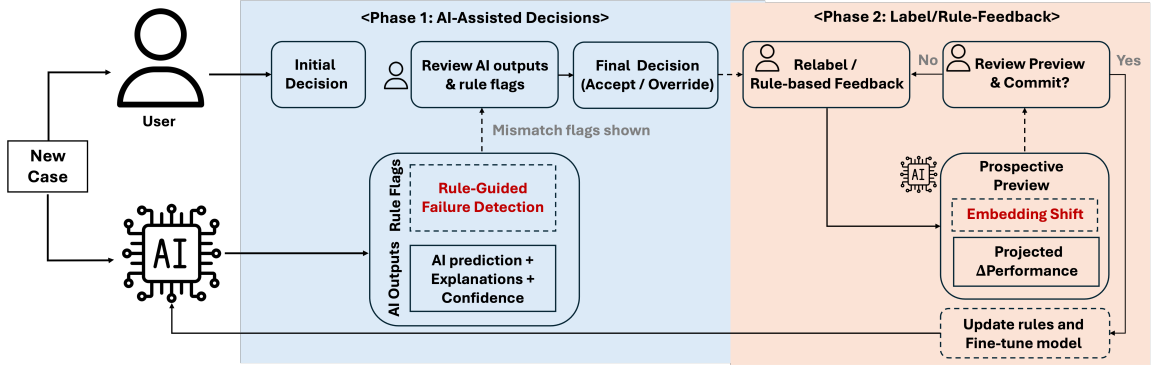


Fig. 1. RuleEdit supports failure inspection, user-authored editing, and prospective preview in AI-assisted decision-making. Given a new case, users first make an initial judgment and then review the system’s AI output together with rule-guided mismatch cues that highlight likely inconsistencies between the model prediction and rule-based evidence. Users then finalize their decision (accept/override). For refinement, users can provide label feedback (relabel) and/or rule feedback (select rules and edit thresholds). The system computes a prospective preview including projected Δ performance and embedding-shift summaries before users decide whether to commit the update.

To explore these research questions, we use a dataset of 15 post-stroke survivors and 10 healthy participants [28, 46] and developed an AI-based decision-support system with a feed-forward neural network to classify post-stroke patients’ quality of motion. Grounded in prior findings that health professionals value interpretable, important feature explanations [48], our system also presents feature-importance explanations to assist users’ rehabilitation assessment tasks (Figure 2a). To support AI failure detection, we implemented a rule-based table that flags case-level contradictions between rule verdicts and model predictions (Figure 2a). For this rule-based table, we empirically selected the top three motion-quality rules by validation performance. To preview prospective embedding updates, showing how data points and class centroids would shift after applying rule feedback, we utilize supervised Uniform Manifold Approximation and Projection (UMAP) [6] with supervision given by user-provided relabels on disagreed cases and updated AI predictions from an ensemble approach that integrates rule-based feedback while keeping the underlying ML model frozen.

To evaluate the effectiveness of our rule-guided failure detection and model adaptation with prospective performance estimates and prospective embedding previews, we conducted a within-subjects experiment with domain experts (i.e. therapists with experience in stroke rehabilitation) and novices (i.e. other health professionals with limited experience in stroke rehabilitation and students in medicine, nursing, and therapy). The study separates two questions: whether rule-guided mismatch cues improve AI-assisted decision-making at decision time, and whether prospective impact preview improves the quality of user-authored edits at refinement time.

Our study shows that rule-guided failure detection increased Human+AI team performance by 14.16% ($p < 0.001$) and improved failure identification while reducing both over- and under-reliance and ChangedToWrong decisions by 20.53% ($p < 0.01$), 7.15% ($p < 0.01$), and 8.00% ($p < 0.01$) respectively. In addition, presenting prospective impact previews improved the quality of user-authored model edits, raising post-update local performance gains from 11.50% to 36.38% ($p < 0.001$) after incorporating users’ rule-based feedback. At the same time, our results reveal a local–global tradeoff: edits that improve a specific case locally may not transfer safely when pooled into broader global updates.

Overall, our work contributes to research on human-AI collaborative decision-making by (i) presenting RuleEdit, an interactive, failure-guided human-AI model editing system that combines mismatch-based failure inspection with

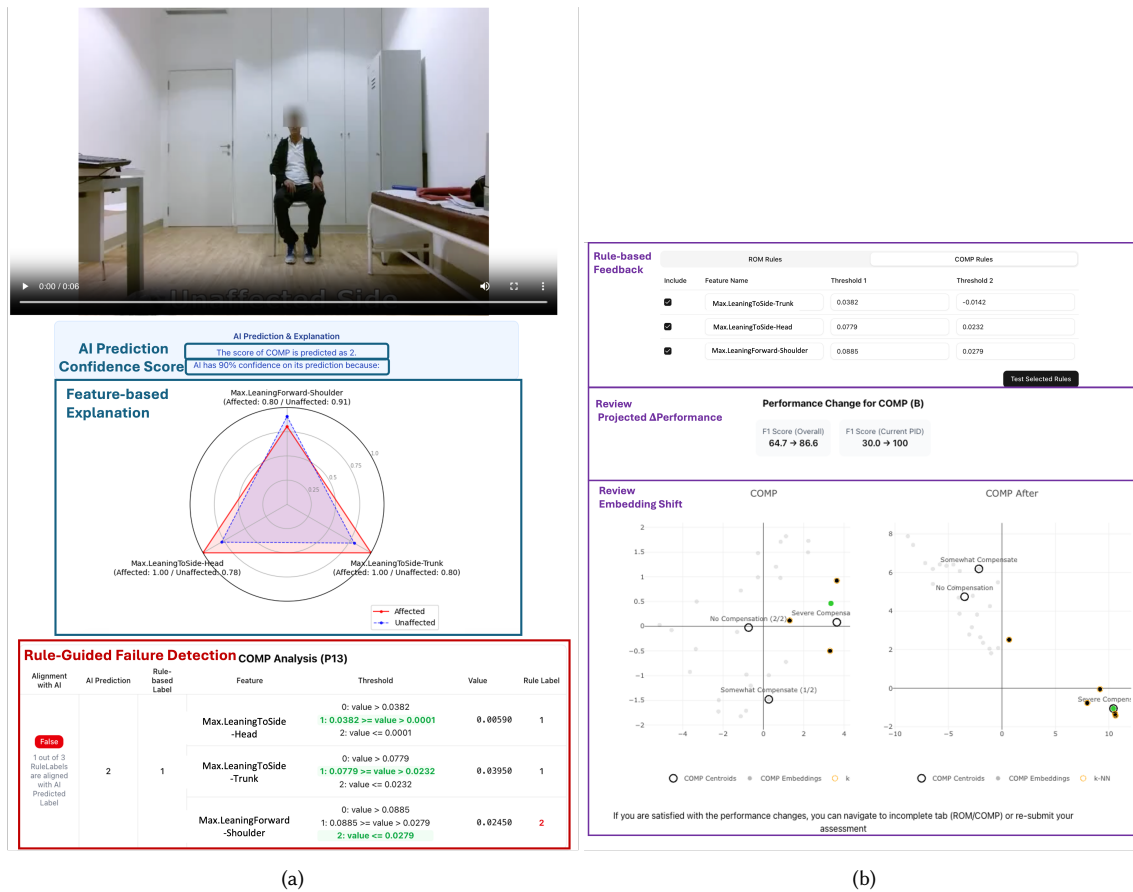


Fig. 2. The system interface includes the video of a post-stroke survivor and the corresponding AI output (prediction, confidence score, and feature-based explanation). (a) The failure-inspection view, where mismatch cues surface likely disagreements between the AI prediction and rule-based evidence to support accept/override decisions, and (b) the editing view, where users select or edit rules and inspect projected performance change together with before/after embedding-shift previews (updated neighborhoods and class centroids) before committing an update.

user-authored edits and prospective impact preview; (ii) showing that rule-guided mismatch cues can improve calibrated reliance and Human+AI team performance during AI-assisted decision-making; and (iii) demonstrating that prospective preview supports higher-quality local refinement while exposing an important boundary: naively pooled global updates can regress despite strong local gains. These findings motivate safeguards for failure-aware and controllable human-AI systems.

2 Related Work

2.1 Human-AI Collaborative Decision-Making

Human-Artificial Intelligence (AI) collaborative decision making research [10, 13, 14, 27, 37, 48, 66] investigates how humans and AI can leverage complementary strengths to improve decision-making in various fields (e.g. healthcare

[13, 31, 48], law [51], and other public services [18, 30, 36, 68]). To support human decision-making, prior work provides decision-support aids, such as feature-based explanations [48, 74], similar-case retrieval [13, 15, 45], counterfactual explanations [43, 78] alongside uncertainty cues [32, 49, 50, 57, 79] and model documentation (e.g., model cards) [44, 54]. These approaches aim to help users interpret why an AI model produced a prediction [1, 4, 39, 58, 73], estimate when that prediction may be uncertain [49, 50, 57, 79], or communicate system-level performance characteristics [54] to support users to determine when to trust AI.

These techniques can improve humans' accuracy and efficiency in practice [7, 13, 43, 48, 57, 72]. However, findings on their effectiveness are mixed. While some studies find limited effects of explanations on trust [16] or improve decision quality [38, 48, 57, 79], other studies show that AI explanations could inflate trust in AI and induce overreliance on incorrect AI outputs [5, 10, 11, 43]. Fostering calibrated reliance remains challenging as over-trust propagates errors and undermines human-AI complementarity [5, 10, 27, 37]. Much of this work helps users interpret AI outputs, but offers limited support for two practical tasks: identifying likely failures on a specific case [12, 69, 76] and intervening in a way that makes the consequences of an edit visible before it is applied.

Our work builds on this literature by treating human-AI decision-making not only as an interpretive task, but also an editing task: users must inspect likely failures, decide whether to accept or override the AI, and, when needed, propose refinements to system behavior. RuleEdit contributes an approach that supports these actions through mismatch-based failure cues and prospective preview before commit for trustworthy human-AI collaboration.

2.2 Interactive Failure Detection and Debugging

Moving beyond presenting static explanations, researchers have developed interactive approaches that let users probe, test, and refine model behavior. Example-based views, counterfactual exploration, and what-if analysis help people interrogate decision boundaries and inspect failure modes [34, 40, 43]. Interactive debugging paradigms further enable users to flag errors, adjust system behavior, and provide feedback that influences subsequent predictions [2, 34, 70, 77]. Within computational interaction, carefully designed feedback channels, such as labeling and co-auditing, have been shown to support oversight and iterative improvement [2, 16, 34].

A central challenge in these systems is making the path from user feedback to model change understandable and trustworthy. Users need not only mechanisms to provide corrections, but also ways to anticipate what a proposed intervention will do before it is committed. Prior work investigates various approaches for interactive debugging from re-labeling [41] and allocating human effort where uncertainty is high [49, 59, 75], to manipulating confusion matrix to steer a model [29], and adjusting feature weights [35] or feature relevance [25, 34, 48]. However, most prior systems emphasize retrospective correction: users identify an issue, provide feedback, and only then observe updated behavior. Support for anticipatory intervention, where users preview the likely downstream effects of an edit before commit, remains limited.

RuleEdit extends this line of work by making candidate edits legible before commitment. The system previews projected performance changes together with representation shifts, such as embedding neighborhoods [9] and class centroids, under a proposed update. These previews support anticipatory auditing by helping users inspect whether an intervention is likely to improve behavior locally or introduce broader regressions.

2.3 Rule-Based Representations for Interpretability and Control

Rules provide a structured, interpretable interface for users to interact with AI systems. Unlike open-ended or example-based feedback [19], rules establish explicit, auditable links between human understanding and model behavior, enabling

practitioners to see when and how their inputs will take effect [62, 63]. Prior work on interpretable decision sets and human-in-the-loop approaches shows that such structured representations can both explain predictions and constrain them at inference or training time [20, 25, 34, 40, 48, 62].

Rule-based approaches have long served as a way to externalize and revise decision logic [67]. In contemporary AI systems, they remain attractive as interaction primitives because they can capture domain constraints, encode safety conditions, and provide users with an interpretable handle for influencing model behavior [23, 64]. However, we still know relatively little about how to design rule-guided workflows that are usable by non-ML experts during real decision tasks, especially when those workflows must support both failure inspection and safe refinement.

We address this gap with a rule-guided editing system in which users leverage interpretable rules to (i) surface likely failure conditions and (ii) propose controllable adaptations. In contrast to static explanations and uncertainty displays [13, 43, 48–50, 57, 74, 79], our approach makes candidate interventions explicit and previews their likely effects before commitment. We empirically examine how these interactions shape human-AI collaborative decision-making for rehabilitation assessment with health professionals and health-related students.

3 Study Design

This work investigates whether interpretable rules can assist users in detecting AI failures and enacting controllable adaptations for human-AI collaborative decision-making (i.e. physical stroke rehabilitation assessment). Building on prior work that explores the utility of informational aids [10, 43, 57], we examine two interaction questions: whether rules can (i) surface likely failure cases in situ to improve how users rely on or override the model during decision-making and (ii) support user-authored model refinements through pre-commit preview. To make these refinements more understandable and auditable, we examine the effect of showing projected performance change together with updated representation views including embedding points and class centroids, before an edit is committed. This framing emphasizes anticipatory rather than purely reactive interaction.

To study these questions, we proceed in two stages. First, we evaluate how well rule-derived signals identify likely AI failures relative to the model’s predictions. Second, we conduct a user study with health professionals and medical/health students to examine two interface effects: whether rule-guided mismatch cues improve accept/override decisions during AI-assisted assessment, and whether pre-commit preview improves the quality of user-authored edits during model refinement.

3.1 Study Context: Physical Stroke Rehabilitation Assessment

We situate our study in a clinical decision-making task: assessing the movement quality of post-stroke survivors. Following prior work on human–AI collaboration for this domain [48], we focus on a functional task in which participants raise the wrist to the mouth, approximating a drinking motion. For rehabilitation assessment, we consider two commonly used dimensions: Range of Motion (ROM) and compensation [47]. ROM captures how completely the target posture is achieved (i.e. whether the hand reaches the intended position). Compensation captures the use of unintended strategies to accomplish the task, such as shoulder elevation or trunk leaning, indicating reliance on non-target joints.

3.2 System Implementations

Our system for AI-assisted stroke rehabilitation assessment combines a feed-forward neural network classifier with two interaction components: rule-guided failure inspection at decision time and user-authored editing with prospective

preview at refinement time (Figure 2). The system presents each rehabilitation video alongside the AI output (prediction, confidence, and feature-based explanation) and mismatch-based failure cues that surface likely inconsistencies between the model prediction and rule-based evidence. Users make an initial decision, review the AI output and mismatch cues, and then finalize the decision by accepting or overriding the AI prediction (Figure 1). For interactive refinement (Figure 2b), the system supports label feedback (i.e. re-label disagreed cases) and rule-based feedback (i.e. select/edit rules and adjust thresholds). Before committing an update, an on-demand, interactive preview module estimates the expected impact on prospective Δ performance change and prospective embedding-shift (updated neighborhoods and class centroids) while keeping the underlying NN frozen. The interface is implemented using a client-server architecture: a React frontend renders the study flow and interactive views, while a Python/Flask [24] backend with PostgreSQL (SQLAlchemy) supports session-based APIs and stores user study data.

3.2.1 Dataset. We draw on an upper-limb rehabilitation dataset on the “bring a cup to the mouth” task. The dataset comprises (i) 300 videos from 15 post-stroke survivors with varied motor impairments, including the affected and unaffected side, [46] and 60 videos from 10 healthy participants who performed one correct motion and five acted impairments designed to mimic common post-stroke movement errors [28], (ii) 3D skeletal joint trajectories extracted using a Microsoft Kinect v2, and (iii) ground truth labels by expert therapists conducted a clinically validated assessment (e.g. Fugl-Meyer Assessment [22]) at recruitment.

3.2.2 AI Model. Following prior work on rehabilitation assessment [46], we extracted a set of kinematic features of each trial (Appendix A.1). As feed-forward neural network (NN) models perform competitively for assessing quality of rehabilitation exercises [46], we implemented feed-forward NN models in PyTorch [56] for Range of Motion (ROM) and Compensation. For the ROM, the model utilizes three hidden layers of 64 units each with a learning rate of 1e-2. For Compensation, the model uses three hidden layers (32, 32, 64 units) with a learning rate of 5e-3. Using leave-one-subject-out (LOSO) cross-validation, the models achieved average F1 scores of 82.83% for ROM and 80.74% for Compensation.

3.2.3 Rule-based Model and Hybrid Inference. We implemented a family of lightweight rule-based models [71] to map each kinematic feature to a clinical label ($y \in \{0, 1, 2\}$) for each performance element (i.e. ROM and COMP). For a given element and feature f , we learn a pair of thresholds (T_1, T_2) from the training split within a LOSO cross-validation, yielding an interpretable decision rule that partitions the feature value into three regions. We considered four thresholding strategies: **RuleMedian** and **RuleAvg**, which compute class representatives (median or mean of x_f within each label) and set $T_1 = (r_0 + r_1)/2, T_2 = (r_1 + r_2)/2$; **RulePercentile**, which sets thresholds using fixed percentiles (33rd/67th) with the direction reversed for COMP to reflect that lower compensation indicates better performance; and **RuleTree**, which fits a shallow decision tree ($\text{max_depth} = 2$) from x_f to y and extracts split points to form (T_1, T_2). At inference time, each rule outputs both a discrete label and a soft probability vector over $\{0, 1, 2\}$ using element-specific monotonic mappings (i.e. piecewise-linear for ROM and smooth sigmoid transitions for COMP), enabling aggregation across multiple rules. To construct concise per-subject explanations, we ranked candidate rules by their held-out performance and selected the Top- K rules ($K \in \{3, \dots, 10\}$) for rule-guided failure detection and hybrid inference/model adaptation. Overall, a compact Top-3 rule set learned via **RuleTree** provided a strong interpretable rule baseline (Appendix. Table 5) for failure detection and a practical choice for hybrid ML-rule aggregation

For the hybrid of ML and rule-based models, we combined the base ML model prediction with the selected rule outputs using either (i) majority voting over predicted labels or (ii) probability aggregation, which averages the

probability vectors from the ML model and rules (uniform weights by default) and predicts argmax of the aggregated distribution. This produces a final label while preserving interpretable rule-level rationales and allowing us to quantify agreement/misalignment between the ML prediction and rule-based evidence. Based on validation results, we use probability aggregation in the user study and fix the rule set to **RuleTree** with Top-3 rules.

3.2.4 Comparison to Alternative Failure Signals. To examine whether rule-guided mismatch cues provide value beyond generic AI unreliability cues, we compared our rule-guided disagreement signal against several alternative failure signals computed under the same leave-one-subject-out (LOSO) protocol used throughout the study. Specifically, we evaluated three uncertainty-derived signals from the base neural network output—prediction confidence, prediction margin, and predictive entropy [21], as well as an ensemble-disagreement baseline [60] computed from three independently trained model seeds. For each test case, these methods produced a scalar score intended to indicate the likelihood that the AI prediction was incorrect. We then evaluated how well each signal identified AI-error cases by measuring area under the precision–recall curve (AUPRC), treating incorrect AI predictions as the positive class.

3.2.5 AI Explanations. As the domain experts, therapists prefer feature-oriented explanations for rehabilitation assessment [47], the interface includes feature-based explanations to assist case review. For each instance, we computed local feature attributions over our feed-forward NN models using SHAP [52] and surfaced only the top three features to manage cognitive load in line with human-AI guidelines [3].

3.2.6 Interactive Preview Module. To support low-latency what-if feedback during refinement, we implement an interactive preview module (Figure 2b) that estimates how a candidate update would change (i) projected performance and (ii) the embedding space. The preview is computed on demand and **does not retrain** the underlying neural network model. A preview is triggered whenever a participant proposes an update: label feedback on disagreed cases or rule-based feedback (e.g. selecting rules and adjusting thresholds).

For projected performance, we use two approaches depending on feedback type. For label feedback, we evaluate a lightweight k-nearest-neighbors classifier on the updated embedding and report local (edited patient) and global (overall) before/after summaries as a fast proxy of change. For rule-based feedback, we apply hybrid inference that combines the frozen ML model with rule outputs (e.g. probability aggregation) to obtain predictions on a held-out evaluation split, and we report the estimated performance change as a pre-commit signal.

To preview embedding shifts, we re-project NN penultimate-layer embeddings into 2D using supervised Uniform Manifold Approximation and Projection (UMAP) [65], conditioning on the updated labels (from label or rule-based feedback). We then aggregate trial-wise points into patient-side points (median across trials for each patient) and refresh kNN neighborhoods for similar cases inspection. We also summarize class-level movement using per-class centroids in the 2D space. For each class $c \in 0, 1, 2$, we compute the centroid as the component-wise median of $(PC1, PC2)$ over participant-side points assigned to that class c to reduce sensitivity to outliers. If a class becomes empty after an edit, we fall back to the last available centroid to avoid unstable behavior.

As the preview is grounded in the neural model’s learned representation, neighborhood and centroid shifts provide an additional check on whether a proposed edit makes cases more locally consistent with nearby examples in embedding space. Together with projected performance change, these views help users inspect the likely consequences of an edit before commitment.

3.3 Experimental Designs

In this section, we describe our experimental setup, including study conditions, participants, protocol, and evaluation metrics. Our study materials and procedures were reviewed and approved by the Institutional Review Board (IRB). The study is designed to evaluate two questions: whether rule-guided mismatch cues improve AI-assisted decisions at decision time, and whether prospective preview improves the quality of user-authored edits at refinement time.

3.3.1 Experimental Conditions. We investigate two interface manipulations across two phases: Phase 1 examines rule-based failure detection during AI-assisted decision making and Phase 2 examines feedback-based model adaptation with/without an embedding-shift preview. For Phase 1 (Figure 2a), participants completed AI-assisted assessment tasks under two conditions that differed only in whether rule-guided mismatch cues were shown.

- Condition A (baseline AI assistance): participants review the AI output (prediction, confidence scores, and explanations) while completing each case. No rule-guided mismatch cues were shown.
- Condition B (rule-guided failure inspection): the system includes all baseline elements and additional mismatch cues from a rule table, highlighting potential inconsistencies between the AI prediction and rule-based evidence. These cues were intended to assist participants to identify likely AI failures and calibrate accept/override decisions.

For Phase 2, participants provide label and rule-based feedback (e.g. selecting a rule, adjusting rule thresholds) to refine the system (Figure 2b). Both conditions showed projected performance change before commitment, but differed in whether the system also visualized embedding shifts.

- Condition A (Performance-only preview): after participants propose feedback, the system showed prospective performance changes only to support the commit decision.
- Condition B (Performance + embedding-shift preview): the system showed both projected performance changes and embedding shifts (i.e. how the representation of cases and class centroids would move under the proposed updates) before users decided whether to commit.

3.3.2 Participants. Our final analyzed sample consisted of 21 participants (domain experts and novices), recruited through an advertisement sent to hospital staff, mailing lists, and the research team’s professional contacts. Domain experts were rehabilitation therapists with stroke rehabilitation experience (e.g. occupational or physiotherapy practice). Novices were health professionals with limited stroke rehabilitation experience and students majoring in medicine, health or therapy. Our participants includes 9 domain experts (5 occupational therapists and 4 physiotherapists) and 12 novices (1 healthcare worker, 1 therapist, 3 nurses, 3 medical students, 1 nursing student, and 2 physiotherapy students). Detailed demographics are provided in Appendix. Table 6.

3.3.3 Protocol. We conducted a two-phase within-subject design study. Phase 1 examined how rule-guided mismatch cues influenced reliance during AI-assisted decision-making. Phase 2 examined how prospective preview affected the quality of user-authored refinements under performance-only versus performance and embedding-shift previews. After providing informed consent, participants completed a tutorial on the system and procedures and then proceeded through the study using Condition A and Condition B (Figure 3).

During the tutorial, participants were introduced to the rehabilitation assessment task and labeling criteria, the system implementations, the interface elements (e.g. AI predictions, confidence scores, feature-based explanations,

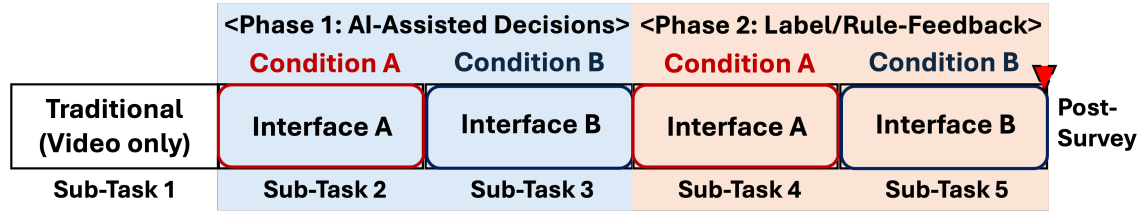


Fig. 3. Two-phase within-subject study. In Phase 1 (AI-assisted decisions), participants complete tasks under Condition A (baseline) and Condition B with rule-based failure detection. In Phase 2 (Label/rule-based feedback), participants review prospective updates under Condition A (projected performance-only preview) and Condition B (projected performance + embedding shifts preview).

and rule-based interactions), and the overall workflow. After the tutorial, participants first reviewed 14 post-stroke survivors’ videos and provided an initial assessment without seeing AI outputs to establish a human-only baseline.

Phase 1 (AI-assisted decisions): Participants reviewed the AI outputs (predictions, confidence scores, and feature-based explanations). In Condition B, the system additionally displayed rule-guided mismatch cues for failure inspection. After reviewing the AI outputs (and rule-based failure detection), participants submitted a final assessment by either accepting or overriding the AI suggestion (Appendix. Figure 4). Each participant completed 14 decisions per condition in Phase 1. To ensure that participants encountered both correct and incorrect AI recommendations, we curated cases such that each condition included 10 AI-correct and 4 AI-incorrect outputs (approximately matching the AI model’s performance). We counterbalanced the order of Condition A and B across participants and randomized case presentation order within each condition. The AI prediction, confidence, and explanation remained fixed across conditions; the only additional interface element in Condition B was the mismatch cue.

Phase 2 (Label/rule feedback for model adaptation): Participants provided label and rule-based feedback (e.g. selecting a rule and adjusting thresholds) to improve and adapt the system. For Condition A (performance-only preview), after participants proposed feedback, the system displayed projected performance changes only. For Condition B, (performance + embedding-shift preview), the system additionally visualized embedding shifts. The order of Conditions A and B in Phase 2 was counterbalanced across participants. Phase 2 always followed Phase 1 to preserve a natural workflow from AI-assisted decision-making, decision-time failure detection to refinement. For the cases, in which participants disagreed with the AI recommendation, participants provided label feedback. For rule-based feedback, participants provided six feedback actions across three cases per condition (selected from cases they disagreed with and/or cases with incorrect AI outputs). Finally, participants completed post-study questionnaires and received fixed compensation based on the rate recommended by the domain experts from a national professional organization.

3.3.4 Metrics for Phase 1. We build on prior work on human-AI decision-making and reliance [5, 13, 26, 37, 49] and utilize performance, reliance, and safety-oriented metrics [42]. Full formal definitions are provided in Appendix (A.4). We summarize the metrics used in the main analysis:

For performance, we report the following metrics: (1) **Human** (h_0) indicates the percentage of cases where the participant’s initial decision matches the ground truth. (2) **Human–AI Team** (Acc_{team}): indicates the percentage of cases where the participant’s final decision after seeing AI matches the ground truth. (3) **Team gain vs. Human** (TeamGain): the improvement from human-only to AI-assisted decision making. To quantify avoidable errors relative to what was achievable on each trial, we also compute (4) **Regret vs. best-of** (Regret_best): the percentage of cases where the final team decision is incorrect despite either the initial human decision or the AI prediction being correct [26].

To measure reliance quality, we compute accept/reject rates conditioned on AI correctness that measure appropriate reliance (accepting correct AI and reject incorrect AI) and failures, such as over-reliance: (5) **Accept-on-correct**: how often participants accepted the AI when the AI was correct, (6) **Reject-on-wrong**: how often participants overrode the AI when the AI was wrong, (7) **Reject-on-correct**: how often participants overrode the AI when the AI was correct, and (8) **Accept-on-wrong**: how often participants accepted the AI when the AI was wrong. As avoiding harm from incorrect reliance is central to safety and responsible deployment, we additionally report (9) **AI-harm rate**: the percentage of cases where the participant was initially correct but became incorrect after seeing AI.

To characterize how participants changed their decisions after seeing AI, we also report: (10) **Changed_rate**: how often the participant changed their decision after seeing AI, (11) **ChangedToRight_rate**: the percentage of cases in which a participant changed an initially wrong decision to a correct one, and (12) **ChangedToWrong_rate**: the percentage of cases in which a participant changed an initially correct decision to a wrong one. Together, these metrics capture overall outcome improvement, calibrated reliance, and safety-relevant decision changes.

3.3.5 Phase 2: Interactive Refinement and Adaptation. We evaluate how participant feedback improves the system under two preview conditions: performance-only preview (Condition A) and performance + embedding-shift preview (Condition B). In Phase 2, participants provide interactive refinement feedback (label feedback on disagreed cases and rule-based feedback by selecting/editing rules and thresholds).

For re-label feedback, we treat each correction as an additional supervised instance and perform offline adaptation by augmenting the baseline training data with these feedback instances, followed by fine-tuning from the same baseline checkpoint. We apply lightweight quality control (dropping incomplete feedback, deduplicating identical instances, and consolidating conflicting labels with a majority vote).

Rule-based feedback allows participants to select or edit interpretable feature–threshold rules that deterministically adjust the model output at inference time. We implement these edits as a *patient-specific rule layer* composed with the baseline ML predictor without changing the learned model parameters. Rules are authored for a given patient by default; to assess whether they capture broader patterns or are overly specific, we evaluate both within-patient effects and cross-patient transfer by applying the same edited rules to held-out data from other patients.

Across both feedback types, we construct *local* updates using only feedback from a specific patient i (within a condition $c \in \{A, B\}$) and *global* updates using pooled feedback across patients. For re-label feedback, the global update corresponds to a single condition-level fine-tuned model. For rule-based feedback, the global analysis evaluates cross-patient transfer of patient-authored rules. We conduct offline adaptation evaluations to quantify how these feedback mechanisms translate into measurable performance change relative to the same baseline model on a held-out evaluation set.

4 User Study Results

We report the results for the two phases of the study. In Phase 1, we evaluate whether rule-guided failure inspection improves AI-assisted decision making. In Phase 2, we examine whether previewing prospective performance changes and embedding shifts helps participants provide higher-quality label and rule-based feedback when refining the model. For each metric in Phase 1 and performance improvement measure in Phase 2, we report descriptive statistics and conducted significance tests of within-participant differences. After assessing data normality using the Kolmogorov-Smirnov test [53], we conducted paired t-tests for normally distributed metrics and the Wilcoxon signed-rank tests [17] otherwise.

4.1 Phase 1: AI-Assisted Decision-Making & Failure Detection

Comparison to Alternative Failure Signals. Before examining participants’ decisions, we first compared whether rule-guided mismatch cues identified likely AI failures more effectively than alternative failure signals. Under the same LOSO protocol, uncertainty-based signals from the base model achieved AUPRC values in the range of 0.46–0.49, while ensemble disagreement across three model seeds achieved an AUPRC of 0.52. In contrast, our rule-guided disagreement signal achieved an AUPRC of 0.64. This suggests that rule-derived mismatch captures complementary information not fully reflected in confidence, margin, entropy, or seed-level prediction instability, and provides a plausible mechanism for the improved decision outcomes observed in Phase 1.

We first examine whether exposing mismatch cues improved participants’ final decisions after viewing AI. Under Condition A (baseline AI assistance without rule-guided failure detection), Human–AI Team accuracy increased only marginally from 67.69% to 69.71% ($\Delta=+2.02\%$, $p = 0.47$ - Appendix. Table 7). In contrast, under Condition B (rule-guided failure detection), Human–AI Team accuracy increased significantly from 67.86% to 82.02% ($\Delta=+14.16\%$, $p < 0.001$ - Appendix. Table 7).

Comparing conditions directly (Table 1), Condition B (rule-guided failure detection) yielded significantly higher Human–AI Team accuracy than Condition A ($\Delta=+12.31\%$, $p < 0.001$) and significantly larger team gain vs. human-only decisions ($\Delta=+12.14\%$, $p < 0.01$). Rule-guided failure detection (Condition B) also reduced avoidable errors as measured by Regret_best (16.46% \rightarrow 4.80%; $\Delta=-11.66\%$, $p < 0.01$), indicating fewer cases where the final team decision was incorrect even though either the initial human decision or the AI was correct.

Table 1. Performance and team-gain metrics. Compared with baseline AI assistance (Condition A), rule-guided mismatch cues (Condition B) significantly improved Human–AI Team accuracy and team gain, while reducing avoidable errors, Regret_best.

	Human	Human–AI Team	Team gain vs. Human	Regret_best
Condition A	67.69	69.71	2.02	16.46
Condition B	67.86	82.02	14.16	4.80
Δ B-A (p-value)	0.17 (p=0.86)	\uparrow 12.31 (p<0.001)	\uparrow 12.14 (p<0.01)	\downarrow -11.66 (p<0.01)

Table 2. Reliance metrics conditioned on AI correctness. Rule-guided mismatch cues improved reliance calibration by increasing acceptance of correct AI and rejection of incorrect AI, while reducing both over/under-reliance and AI-induced harm.

	Accept on_correct	Reject on_wrong	Reject on_correct	Accept on_wrong	AI_harm rate
ConditionA	88.09	41.61	11.91	58.39	8.29
ConditionB	95.24	62.14	4.76	37.86	3.70
Δ B-A (p-value)	\uparrow 7.15 (p<0.01)	\uparrow 20.53 (p<0.05)	\downarrow -7.15 (p<0.01)	\downarrow -20.53 (p<0.01)	\downarrow -4.59 (p<0.05)

Subgroup trends (Appendix. Tables 7 and 8). The same pattern held for both experts and novices. Experts showed no improvement of Human–AI Team under Condition A (71.35% \rightarrow 70.19%, $p = 0.56$), but improved significantly in Condition B (72.22% \rightarrow 83.43%, $p < 0.05$). Novices improved modestly the Human–AI Team metric under Condition A (64.94% \rightarrow 69.35%, $p = 0.18$; n.s.) and improved significantly under Condition B (64.58% \rightarrow 80.95%, $p < 0.01$). Across subgroups, Regret_best decreased in Condition B (Appendix Table 8), suggesting that mismatch cues assisted participants to better select between their initial decision and the AI recommendation, thereby reducing avoidable errors even though oracle-best accuracy (i.e. the availability of a correct answer from either human or AI) remained similar across conditions.

4.1.1 Reliance Behavior (Accept/Reject Patterns). To understand how mismatch cues changed reliance behavior, we analyze accept/reject behaviors conditioned on AI correctness (Table 2). Compared to the baseline (Condition A), rule-guided failure detection (Condition B) significantly increased *Accept-on-correct* (88.09% \rightarrow 95.24%, $\Delta=+7.15\%$, $p < 0.01$) and *Reject-on-wrong* (41.61% \rightarrow 62.14%, $\Delta=+20.53\%$, $p < 0.05$). In parallel, it significantly decreased *Reject-on-correct* (11.91% \rightarrow 4.76%, $\Delta=-7.15\%$, $p < 0.01$) and *Accept-on-wrong* (58.39% \rightarrow 37.86%, $\Delta=-20.53\%$, $p < 0.01$). Taken together, these results indicate better reliance calibration under Condition B (rule-guided failure detection): participants were more likely to follow correct AI, **more likely to reject incorrect AI**, and less likely to show either over-reliance or under-reliance.

As a safety-oriented reliance measure, rule-guided failure detection (Condition B) also reduced the *AI-harm rate* (8.29% \rightarrow 3.70%, $\Delta=-4.59\%$, $p < 0.05$), indicating fewer cases where participants moved from an initially correct decision to an incorrect final decision after viewing AI.

Subgroup trends. (Appendix Table 9) Experts showed strong improvements in *Accept-on-correct* (86.40% \rightarrow 98.89%, $p < 0.01$) and reductions in *Reject-on-correct* (13.60% \rightarrow 1.11%, $p < 0.01$). For novices, the accept/reject metrics changed in the same direction, such as higher *Reject-on-wrong* and lower *Accept-on-wrong*, but these differences were not statistically significant (often $p \approx 0.1$), consistent with lower power in subgroup analyses. Help-oriented metrics (e.g. *AI-help*, *missed-help*, *correct-ignore*) were directionally positive under rule-guided failure detection (Condition B) (i.e. slightly higher AI-help and lower missed-help) but were not statistically significant.

4.1.2 Decision-Change Behavior After Viewing AI. We then examine how participants changed their decisions after seeing AI, in order to better understand how the improvements above were achieved (Table 3). Overall, rule-guided failure detection (Condition B) reduced the *Changed* rate (34.98% \rightarrow 26.66%, $\Delta=-8.32\%$, $p < 0.1$), while showing a directional increase in *ChangedToRight* (14.74% \rightarrow 18.88%, $\Delta=+4.14\%$, $p = 0.13$). In addition, rule-guided failure detection (Condition B) significantly reduced *ChangedToWrong* decisions (12.72% \rightarrow 4.72%, $\Delta=-8.00\%$, $p < 0.01$), indicating that rule-guided failure detection reduced harmful switches from initially correct judgments to incorrect final decisions.

Table 3. Decision-change behaviors after viewing AI outputs. Rule-guided mismatch cues (Condition B) reduced overall switching and significantly reduced harmful changes from correct initial judgments to incorrect final decisions.

	Changed	ChangedToRight	ChangedToWrong
ConditionA	34.98	14.74	12.72
ConditionB	26.66	18.88	4.72
Δ B-A (p-value)	\downarrow 8.32 (p<0.1)	4.14 (p=0.13)	\downarrow 8.00 (p<0.01)

Subgroup trends. (Appendix Table 10) The reduction in *ChangedToWrong* holds for both experts (13.73% \rightarrow 6.25%, $p < 0.05$) and novices (11.96% \rightarrow 3.57%, $p < 0.05$). While the *Changed* and *ChangedToRight* showed similar directional patterns (increases) within subgroups, these effects were not statistically significant, again consistent with reduced power.

4.2 Phase 2: Model Adaptation

We analyzed how participants' feedback affected model refinement under the two preview conditions (Table 4). Overall, Phase 2 refinement results show a clear asymmetry between *local* versus *global* updating (Table 4). Local (participant-specific) updates improved performance relative to the baseline in both conditions, with significantly larger gains in

Table 4. Phase 2 refinement performance for all participants. We report improvement over the baseline model on a held-out evaluation set for local (participant-specific) and global (across-patients) updates. p -values are from paired Wilcoxon signed-rank tests across participants.

Metric	A	B	$\Delta B-A$	p
Local update Δ Perf (overall)	11.50	36.38	24.89	< 0.001
Global update Δ Perf (overall)	2.38	-5.05	-7.43	< 0.001

Performance+Embedding-Shift Preview (Condition B) (A: **+11.50** vs B: **+36.38**; $\Delta B-A=+24.89$, $p < 0.001$). In contrast, global (across patients) updates did not provide reliable improvement: Performance-only Preview (Condition A) produced only a small gain (**+2.38**), while Performance+Embedding-Shift Preview (Condition B) performed below baseline (**-5.05**), yielding a significant negative difference ($\Delta B-A=-7.43$, $p < 0.001$). Together, these findings suggest that richer preview is particularly useful for local personalization while revealing that edits beneficial at the local level may not remain beneficial when aggregated into a single global update.

Subgroup trends. (Appendix Table 11) The local-update advantage of Performance+Embedding-Shift Preview (Condition B) held for both experts (A: **+11.20** \rightarrow B: **+33.89**, $\Delta = +22.69$, $p < 0.001$) and novices (A: **+11.92** \rightarrow B: **+38.88**, $\Delta = +26.96$, $p < 0.05$). For global updates, both subgroups show the same directional pattern of degradation under Condition B relative to Condition A (experts: **+5.58** \rightarrow **-2.95**, $p < 0.001$; novices: **0.00** \rightarrow **-6.46**, $p < 0.01$), consistent with pooled updates being less stable than personalized refinement.

Additional analysis of global direct retraining showed that ROM improved only slightly (best $\Delta \approx +1.99$), while COMP degraded substantially (best $\Delta \approx -14.92$), leading to an overall decrease in performance (best $\Delta \approx -6.46$). This pattern indicates that naively aggregating feedback into a single global update can introduce task-specific trade-offs and may require additional safeguards (e.g. per-task updates, data curation, or constraints) to avoid regressions. In contrast, rule-based refinement yields strong *local* gains but limited cross-patient transfer.

5 Discussion

Our findings suggest that failure-aware human-AI model editing benefits from two complementary interface capabilities: decision-time mismatch cues and refinement-time prospective preview. In Phase 1, rule-guided mismatch cues improved Human-AI team performance and reliance calibration by helping participants decide when to accept or override the AI. In Phase 2, prospective preview improved the quality of local refinements, while also revealing that edits that appear beneficial locally may not remain beneficial when pooled into broader global updates. Together, these results shift the design focus from merely showing AI outputs to supporting failure inspection, user-authored intervention, and pre-commit reasoning about model change, enabling more actionable and anticipatory human-AI collaboration.

5.1 Rule-Guided Mismatch Cues Support Calibrated Reliance

Our Phase 1 findings show that rule-guided mismatch cues can calibrate reliance by helping users decide *when to follow* AI and *when to override* it. Adding interpretable mismatch cues from rule-based evidence (Condition B) yielded a large improvement in Human-AI Team accuracy compared to baseline AI assistance (Condition A): 82.02% vs. 69.71% ($\Delta=+12.31\%$, $p < 0.001$), while human-only performance did not differ across conditions (67.86% vs. 67.69%, $p = 0.86$). This extends prior work showing that AI outputs (e.g. explanations and confidence scores) can induce overreliance:

beyond helping users interpret the AI, a lightweight mismatch cue can help them act on potential failures at decision time.

Beyond team performance, rule-guided mismatch cues also delivered measurable safety benefits. They significantly reduced harmful reliance events, lowering AI-induced harm (*AI-harm*: 8.29% \rightarrow 3.70%, $p < 0.05$) and reduced harmful switches from correct initial judgments to incorrect final decisions (*ChangedToWrong*: 12.72% \rightarrow 4.72%, $p < 0.01$). In addition, rule-guided mismatch cues reduced avoidable errors in a decision-theoretic sense: *Regret_best* dropped from 16.46% to 4.80% ($p < 0.01$), meaning teams were less likely to end wrong even when either the human or AI had the correct answer available [26]. Together, these results suggest that *interpretable consistency checks* can act as a practical guardrail for both reliance calibration and harm reduction, complementing prior approaches that emphasize richer explanations or uncertainty displays [5, 33, 49, 50].

Rules functioned as simple, interpretable heuristics that prompted participants to scrutinize potentially inconsistent AI outputs, supporting reflective rather than automatic acceptance. This resonates with recent findings on confidence alignment [15, 50] and uncertainty-aware delegation [49], yet extends them by offering a structured and interpretable mechanism for trust calibration. A key next step is to understand how such cues should be learned or specified, validated, and personalized across different error profiles, tasks, and domains. Another is to examine how mismatch-based signals can be combined with uncertainty-aware or decision theoretic reliance objectives to jointly optimize accuracy and safety [26].

5.2 Prospective Previews for Anticipatory Model Refinement

Our Phase 2 results reveal that presenting prospective embedding updates during model refinement improved participants' ability to provide targeted, high-quality feedback. When participants could inspect how their selected rules would reshape the representation space (e.g. shifts in class centroids and data points) alongside projected performance changes, the resulting *local* (per-patient) improvement increased substantially (Condition A: 11.50% \rightarrow Condition B: 36.38%, $p < 0.001$). This suggests that preview helps users reason about the likely consequences of an edit before commitment, enabling more targeted refinements that are better aligned with the model's learned representation. At the same time, these benefits were primarily *local*: when feedback was aggregated for a single *global* update, performance did not reliably improve and could regress (global Δ Perf: Condition A +2.38 vs Condition B -5.05, $p < 0.001$; Table 4). Consistent with our global direct-retraining analysis, pooled global updates can introduce task-specific trade-offs (e.g. COMP requires patient-specific assessment [47]), yielding a net decrease in performance. This local-global asymmetry indicates that systems supporting user-authored editing should not assume that locally beneficial edits are globally safe. Instead, broader deployment may require safeguards [8], such as per-task updates, stricter data curation, conflict-aware aggregation, or constrained optimization (e.g. partial freezing, conservative learning rates).

We interpret this effect as a proactive form of anticipatory calibration and alignment: users learn not only from retrospective outcomes (i.e. observing an error after it occurs), but also from prospective consequences (i.e. previewing how an intervention will change the model) before an intervention is committed. Prior interactive and explanatory debugging systems largely emphasize retrospective error correction (i.e. users identify failures, provide corrective feedback, and then observe the updated behavior) [25, 34, 48]. In contrast, our approach makes the likely direction and magnitude of change legible before commitment, supporting planning-oriented interaction and helping users reason about potential failure propagation in advance. By improving users' mental models of how feedback translates into model change, prospective preview can support more transparent and controllable refinement for trustworthy human-AI collaboration, while reducing reliance on trial-and-error editing [10, 27].

5.3 From Explanations to Failure-Aware Model Editing

Our study highlights a shift from explanation-centered systems towards failure-aware model editing. Traditional explainable AI (XAI) methods, such as feature importance [48, 74], counterfactual [43, 78] or uncertainty scores [49, 50, 57, 79], offer informational transparency [4, 73], but they often fail to foster calibrated reliance or provide users with mechanisms to intervene when the system fails. RuleEdit connects explanation-like evidence to controllable mechanisms, such as rejecting a likely-wrong AI output or proposing targeted feedback for refinement. Through these mechanisms, users move from passively interpreting AI outputs to actively inspecting likely failures and editing model behavior.

This approach aligns with interactive auditing [69] and human-centered model governance [23] by enabling users to trace, inspect, and justify their interventions. Importantly, our controllable interaction moves beyond documenting failures [69] to mitigating them: instead of post-hoc interpretation, participants can intervene on failures in situ and inspect prospective consequences before commitment.

5.4 Limitations and Future Work

While our findings provide promising evidence for rule-guided failure detection and model refinement, this study also has several limitations. First, our study was conducted in a single domain, stroke rehabilitation assessment with a modest sample size. Although this scale is comparable to prior work on human-AI collaborative decision-making in health [11, 47], future work should examine whether observed improvements in reliance calibration and model refinement generalize to other domains, such as diagnostic imaging [61] or child welfare services [18, 68].

In addition, the model adaptation process was evaluated offline rather than within real or longitudinal decision-making workflows. Field deployment studies would provide deeper insights into how trust, learning, and editing behavior evolve over time. As rehabilitation practices are typically led by individual therapists, our study focused on individual feedback. Future work should extend this setting to multi-user workflows in which teams co-create, review, and validate rules and edits for broader organizational use and accountability [55].

6 Conclusion

In this work, we examined how failure-aware human-AI model editing can be supported through two complementary capabilities: rule-guided mismatch cues and prospective impact preview. Our findings show that rule-guided mismatch cues improved AI-assisted decision-making by helping users identify likely AI failures, decide when to accept or override the AI, and reduce both overreliance and underreliance. Furthermore, prospective preview substantially improved the quality of local user-authored refinements, demonstrating that users' structured, rule-based feedback can meaningfully enhance AI systems in a controlled and auditable workflow.

At the same time, our results reveal an important boundary: edits that appear beneficial locally may not remain beneficial when pooled into broader global updates. This local-global tradeoff underscores the need for supporting pre-commit reasoning about model change rather than trial-and-error editing alone.

Overall, RuleEdit points toward a shift from explanation-centered AI support to failure inspection, user-authored intervention, and prospective reasoning about model behavior. By linking interpretable rules to controllable editing workflows, our work contributes to the design of more transparent, controllable, and trustworthy human-AI systems.

References

- [1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–18.
- [2] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI magazine* 35, 4 (2014), 105–120.
- [3] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.
- [4] Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. 2019. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012* (2019).
- [5] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–16.
- [6] Etienne Becht, Leland McInnes, John Healy, Charles-Antoine Dutertre, Immanuel WH Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W Newell. 2019. Dimensionality reduction for visualizing single-cell data using UMAP. *Nature biotechnology* 37, 1 (2019), 38–44.
- [7] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M Vardoulakis. 2020. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–12.
- [8] Yoshua Bengio, Stephen Clare, Carina Prunkl, Maksym Andriushchenko, Ben Bucknall, Philip Fox, Nestor Maslej, Conor McGlynn, Malcolm Murray, Shalaleh Rismani, et al. 2025. International ai safety report 2025: Second key update: Technical safeguards and risk management. *arXiv preprint arXiv:2511.19863* (2025).
- [9] Angie Boggust, Brandon Carter, and Arvind Satyanarayan. 2022. Embedding comparator: Visualizing differences in global structure and local neighborhoods via small multiples. In *27th international conference on intelligent user interfaces*. 746–766.
- [10] Zana Bućinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction* 5, CSCW1 (2021), 1–21.
- [11] Adrian Bussone, Simone Stumpf, and Dympna O’Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *2015 international conference on healthcare informatics*. IEEE, 160–169.
- [12] Ángel Alexander Cabrera, Abraham J Druck, Jason I Hong, and Adam Perer. 2021. Discovering and validating ai errors with crowdsourced failure reports. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–22.
- [13] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. 2019. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–14.
- [14] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. " Hello AI": uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. *Proceedings of the ACM on Human-computer Interaction* 3, CSCW (2019), 1–24.
- [15] Valerie Chen, Q Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. 2023. Understanding the role of human intuition on reliance in human-AI decision-making with explanations. *Proceedings of the ACM on Human-computer Interaction* 7, CSCW2 (2023), 1–32.
- [16] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O’connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. 2019. Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.
- [17] Jack Cuzick. 1985. A Wilcoxon-type test for trend. *Statistics in medicine* 4, 1 (1985), 87–90.
- [18] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. 2020. A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.
- [19] John J Dudley and Per Ola Kristensson. 2018. A review of user interface design for interactive machine learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8, 2 (2018), 1–37.
- [20] Alan M Frisch, Christopher Jefferson, Bernadette Martínez Hernández, and Ian Miguel. 2005. The rules of constraint modelling. In *IJCAI* 109–116.
- [21] Jakob Gawlikowski, Cedrique Rovile Njéutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. 2023. A survey of uncertainty in deep neural networks. *Artificial intelligence review* 56, Suppl 1 (2023), 1513–1589.
- [22] David J Gladstone, Cynthia J Danells, and Sandra E Black. 2002. The Fugl-Meyer assessment of motor recovery after stroke: a critical review of its measurement properties. *Neurorehabilitation and neural repair* 16, 3 (2002), 232–240.
- [23] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–24.
- [24] Miguel Grinberg. 2018. *Flask web development: developing web applications with python*. " O’Reilly Media, Inc."
- [25] Lijie Guo, Elizabeth M Daly, Ozgur Alkan, Massimiliano Mattetti, Owen Corneic, and Bart Knijnenburg. 2022. Building trust in interactive machine learning via user contributed interpretable rules. In *Proceedings of the 27th international conference on intelligent user interfaces*. 537–548.

- [26] Ziyang Guo, Yifan Wu, Jason D Hartline, and Jessica Hullman. 2024. A decision theoretic framework for measuring AI reliance. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 221–236.
- [27] Kenneth Holstein, Maria De-Arteaga, Lakshmi Tumati, and Yanghui Cheng. 2023. Toward supporting perceptual complementarity in human-AI collaboration via reflection on unobservables. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–20.
- [28] Min Hun Lee, Daniel P Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermudez i Badia. 2023. Design, development, and evaluation of an interactive personalized social robot to monitor and coach post-stroke rehabilitation exercises. *User Modeling and User-Adapted Interaction* 33, 2 (2023), 545–569.
- [29] Ashish Kapoor, Bongshin Lee, Desney Tan, and Eric Horvitz. 2010. Interactive optimization for steering machine classification. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1343–1352.
- [30] Anna Kawakami, Venkatesh Sivaraman, Hao-Fei Cheng, Logan Stapleton, Yanghui Cheng, Diana Qing, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. 2022. Improving human-AI partnerships in child welfare: understanding worker practices, challenges, and desires for algorithmic decision support. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–18.
- [31] Dajung Kim, Niko Vegt, Valentijn Visch, and Marina Bos-De Vos. 2024. How Much Decision Power Should (A) I Have?: Investigating Patients’ Preferences Towards AI Autonomy in Healthcare Decision Making. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17.
- [32] Sunnie SY Kim, Q Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. "I'm Not Sure, But...": Examining the Impact of Large Language Models’ Uncertainty Expression on User Reliance and Trust. In *Proceedings of the 2024 ACM conference on fairness, accountability, and transparency*. 822–835.
- [33] Sunnie SY Kim, Jennifer Wortman Vaughan, Q Vera Liao, Tania Lombrozo, and Olga Russakovsky. 2025. Fostering appropriate reliance on large language models: The role of explanations, sources, and inconsistencies. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–19.
- [34] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*. 126–137.
- [35] Todd Kulesza, Simone Stumpf, Weng-Keen Wong, Margaret M Burnett, Stephen Perona, Amy J Ko, and Ian Oberst. 2011. Why-oriented end-user debugging of naive Bayes text classification. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 1, 1 (2011), 1–31.
- [36] Tzu-Sheng Kuo, Hong Shen, Jisoo Geum, Nev Jones, Jason I Hong, Haiyi Zhu, and Kenneth Holstein. 2023. Understanding Frontline Workers’ and Unhoused Individuals’ Perspectives on AI Used in Homeless Services. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.
- [37] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a science of human-ai decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471* (2021).
- [38] Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency*. 29–38.
- [39] Himabindu Lakkaraju, Julius Adebayo, and Sameer Singh. 2020. Explaining machine learning predictions: State-of-the-art, challenges, and opportunities. *NeurIPS Tutorial* (2020).
- [40] Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. 2016. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1675–1684.
- [41] Kimin Lee, Laura Smith, and Pieter Abbeel. 2021. Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. *arXiv preprint arXiv:2106.05091* (2021).
- [42] Min Hun Lee. 2026. From Accuracy to Readiness: Metrics and Benchmarks for Human-AI Decision-Making. *arXiv preprint arXiv:2603.18895* (2026).
- [43] Min Hun Lee and Chong Jun Chew. 2023. Understanding the Effect of Counterfactual Explanations on Trust and Reliance on AI for Human-AI Collaborative Clinical Decision Making. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–22.
- [44] Min Hun Lee, Silvana Xin Yi Choo, Shamala D Thilarajah, et al. 2024. Improving Health Professionals’ Onboarding with AI and XAI for Trustworthy Human-AI Collaborative Decision Making. *arXiv preprint arXiv:2405.16424* (2024).
- [45] Min Hun Lee, Renee Bao Xuan Ng, Silvana Xinyi Choo, and Shamala Thilarajah. 2024. Interactive Example-based Explanations to Improve Health Professionals’ Onboarding with AI for Human-AI Collaborative Decision Making. *arXiv preprint arXiv:2409.15814* (2024).
- [46] Min Hun Lee, Daniel P Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2019. Learning to assess the quality of stroke rehabilitation exercises. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 218–228.
- [47] Min Hun Lee, Daniel P Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2020. Co-design and evaluation of an intelligent decision support system for stroke rehabilitation assessment. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–27.
- [48] Min Hun Lee, Daniel P Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2021. A human-ai collaborative approach for clinical decision making on rehabilitation assessment. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–14.
- [49] Min Hun Lee and Martyn Zhe Yu Tok. 2025. Towards Uncertainty Aware Task Delegation and Human-AI Collaborative Decision-Making. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. 2274–2289.
- [50] Jingshu Li, Yitian Yang, Q Vera Liao, Junti Zhang, and Yi-Chieh Lee. 2025. As Confidence Aligns: Understanding the Effect of AI Confidence on Human Self-confidence in Human-AI Decision Making. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–16.

- [51] Gabriel Lima, Nina Grgić-Hlača, and Meeyoung Cha. 2021. Human perceptions on moral responsibility of AI: A case study in AI-assisted bail decision-making. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–17.
- [52] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 4765–4774. <http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf>
- [53] Frank J Massey Jr. 1951. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association* 46, 253 (1951), 68–78.
- [54] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.
- [55] Claudio Novelli, Mariarosaria Taddeo, and Luciano Floridi. 2024. Accountability in artificial intelligence: What it is and how it works. *Ai & Society* 39, 4 (2024), 1871–1882.
- [56] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).
- [57] Snehal Prabhudesai, Leyao Yang, Sumit Asthana, Xun Huan, Q Vera Liao, and Nikola Banovic. 2023. Understanding uncertainty: how lay decision-makers perceive and interpret uncertainty in human-AI decision making. In *Proceedings of the 28th international conference on intelligent user interfaces*. 379–396.
- [58] Alun Preece. 2018. Asking ‘Why’ in AI: Explainability of intelligent systems—perspectives and challenges. *Intelligent Systems in Accounting, Finance and Management* 25, 2 (2018), 63–72.
- [59] Maitra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. 2019. Transfusion: Understanding transfer learning for medical imaging. *Advances in neural information processing systems* 32 (2019).
- [60] Rahul Rahaman et al. 2021. Uncertainty quantification and deep ensembles. *Advances in neural information processing systems* 34 (2021), 20063–20075.
- [61] Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J Topol. 2022. AI in health and medicine. *Nature medicine* 28, 1 (2022), 31–38.
- [62] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.
- [63] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* 1, 5 (2019), 206–215.
- [64] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. 2022. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys* 16 (2022), 1–85.
- [65] Tim Sainburg, Leland McInnes, and Timothy Q Gentner. 2021. Parametric UMAP embeddings for representation and semisupervised learning. *Neural Computation* 33, 11 (2021), 2881–2907.
- [66] Sara Salimzadeh, Gaole He, and Ujwal Gadiraju. 2024. Dealing with Uncertainty: Understanding the Impact of Prognostic Versus Diagnostic Tasks on Trust and Reliance in Human-AI Decision Making. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17.
- [67] Edward H Shortliffe. 1986. Medical expert systems—knowledge tools for physicians. *Western Journal of Medicine* 145, 6 (1986), 830.
- [68] Logan Stapleton, Min Hun Lee, Diana Qing, Marya Wright, Alexandra Chouldechova, Ken Holstein, Zhiwei Steven Wu, and Haiyi Zhu. 2022. Imagining new futures beyond predictive systems in child welfare: A qualitative study with impacted stakeholders. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1162–1177.
- [69] Ningjing Tang, Jiayin Zhi, Tzu-Sheng Kuo, Calla Kainaroi, Jeremy J Northup, Kenneth Holstein, Haiyi Zhu, Hoda Heidari, and Hong Shen. 2024. AI failure cards: Understanding and supporting grassroots efforts to mitigate AI failures in homeless services. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 713–732.
- [70] Stefano Teso and Kristian Kersting. 2019. Explanatory interactive machine learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 239–245.
- [71] Ayush K Varshney and Vicenç Torra. 2023. Literature review of the recent trends and applications in various fuzzy rule-based systems. *International Journal of Fuzzy Systems* 25, 6 (2023), 2163–2186.
- [72] Dakuo Wang, Liuping Wang, Zhan Zhang, Ding Wang, Haiyi Zhu, Yvonne Gao, Xiangmin Fan, and Feng Tian. 2021. “Brilliant AI doctor” in rural clinics: Challenges in AI-powered clinical decision support system deployment. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–18.
- [73] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.
- [74] Xinru Wang and Ming Yin. 2021. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th international conference on intelligent user interfaces*. 318–328.
- [75] Bryan Wilder, Eric Horvitz, and Ece Kamar. 2020. Learning to complement humans. *arXiv preprint arXiv:2005.00582* (2020).
- [76] Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. 2019. Errudite: Scalable, reproducible, and testable error analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 747–763.
- [77] Qian Yang, Jina Suh, Nan-Chen Chen, and Gonzalo Ramos. 2018. Grounding interactive machine learning tool design in how non-experts actually build models. In *Proceedings of the 2018 designing interactive systems conference*. 573–584.

Table 5. Top- K rule performance across rule construction methods for ROM, COMP, and Overall.

Top- K	ROM				COMP				Overall			
	RuleMedian	RuleAvg	RulePercentile	RuleTree	RuleMedian	RuleAvg	RulePercentile	RuleTree	RuleMedian	RuleAvg	RulePercentile	RuleTree
Top10	0.8544	0.8548	0.6505	0.8094	0.8455	0.8503	0.8311	0.8226	0.8500	0.8525	0.7408	0.8160
Top9	0.8701	0.8673	0.6719	0.8311	0.8496	0.8533	0.8369	0.8312	0.8598	0.8603	0.7544	0.8312
Top8	0.8833	0.8774	0.6915	0.8442	0.8522	0.8559	0.8410	0.8407	0.8678	0.8667	0.7663	0.8424
Top7	0.8970	0.8883	0.7154	0.8599	0.8547	0.8587	0.8445	0.8505	0.8759	0.8735	0.7800	0.8552
Top6	0.9077	0.8983	0.7439	0.8769	0.8574	0.8612	0.8471	0.8593	0.8826	0.8797	0.7955	0.8681
Top5	0.9216	0.9113	0.7690	0.8918	0.8594	0.8647	0.8507	0.8686	0.8905	0.8880	0.8098	0.8802
Top4	0.9378	0.9288	0.8020	0.9074	0.8612	0.8661	0.8539	0.8809	0.8995	0.8975	0.8279	0.8942
Top3	0.9574	0.9476	0.8400	0.9297	0.8630	0.8668	0.8574	0.8961	0.9102	0.9072	0.8487	0.9129*

[78] Wencan Zhang and Brian Y Lim. 2022. Towards relatable explainable AI with the perceptual process. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–24.

[79] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 295–305.

A System Implementation Details

A.1 Feature Extractions

We extend prior rehabilitation assessment work [46] by transforming tracked 3D joint trajectories of post-stroke survivors’ exercises into kinematic features.

For Range of Motion (ROM), we derive joint-angle features (e.g., elbow flexion/extension, shoulder flexion) and distance related features. Specifically, we compute pairwise Euclidean distances between key joints, such as head–wrist and head–elbow, and wrist distances to the head and shoulder, along with their per axis components (x, y, z) [46]. For Compensation, we quantify non-target movement by measuring normalized displacements of the head, spine, and shoulder from the trial’s initial frame over time. These per-axis (x, y, z) trajectories summarize behaviors, such as shoulder elevation or trunk lean that indicate compensatory strategies [46].

A.2 Results of Machine Learning Models

We trained the feed-forward neural network (NN) model through a grid search over depth (1-4 hidden layers with 32, 64, 128, 256, 512 hidden units) along with learning rates (i.e. 1e-5, 5e-4, 1e-4, 5e-3, 1e-3, 1e-2). Models were trained by minimizing cross-entropy loss and evaluated with leave-one-subject-out cross-validation: in each fold, all data from a single participant were held out for testing, and the model was trained on the remaining data.

A.3 Results of Rule-based Models

Across all three experiments on ROM, COMP, and Overall, performance of rule-based models improves monotonically as we reduce K (Top 10 \rightarrow Top 3), indicating that a small set of high-quality rules is more effective than including weaker rules. For the ROM performance component, RuleMedian achieves the highest performance of 0.9574 at Top3 (vs. RuleTree 0.9297), while for the COMP performance component, RuleTree achieves the best Top3 score (0.8961). Overall, the best Top3 performance is 0.9129 (RuleTree), suggesting that selecting a compact Top3 rule set can yield a strong, interpretable rule baseline and a practical candidate for hybrid of ML and rules aggregation.

A.4 Detail of Data Analysis Metrics for Phase 1

Let y denote the ground-truth label, h_0 the participant’s initial (human-only) decision, a the AI prediction, and h_1 the participant’s final decision after viewing the AI output (and optionally overriding). We use the indicator function $\mathbb{I}[\cdot]$.

- **Human** (h_0): $\mathbb{I}[h_0 = y]$ indicates the percentage of cases where the participant’s initial decision matches the ground truth.
- **AI accuracy** (Acc_{ai}): $\mathbb{I}[a = y]$ indicates the percentage of cases where the AI prediction matches the ground truth.
- **Human–AI Team** (Acc_{team}): $\mathbb{I}[h_1 = y]$ indicates the percentage of cases where the participant’s final decision after seeing AI matches the ground truth.
- **Team gain vs. Human** (TeamGain_{human}): $\text{Acc}_{team} - h_0$, the improvement from human-only to AI-assisted decision making.
- **Team gain vs. AI** (TeamGain_{ai}): $\text{Acc}_{team} - \text{Acc}_{ai}$, the improvement from AI accuracy to the team’s final accuracy, computed as $\text{Acc}_{team} - \text{Acc}_{ai}$.
- **Regret_best** ($\text{Oracle} - \mathbb{I}[h_1 = y]$): the percentage of cases where the final team decision is incorrect despite either the initial human decision or the AI prediction being correct [26]. This equals 1 when the team is wrong despite at least one agent (human-only or AI) being correct, and 0 otherwise [26].
- **Oracle best accuracy** (Acc_{oracle}): $\mathbb{I}[(h_0 = y) \vee (a = y)]$, Upper bound accuracy if one could always choose the correct answer between the initial human decision and the AI prediction.

In addition, we compute accept/reject rates conditioned on AI correctness that measure appropriate reliance (accepting correct AI and reject incorrect AI) and failures, such as over-reliance:

- **Accept-on-correct** ($\Pr(h_1 = a \mid a = y)$) indicates when the participant’s final decision agrees with AI among cases with correct AI outputs.
- **Reject-on-wrong**: $\Pr(h_1 \neq a \mid a \neq y)$ indicates the percentage where the participant’s final decision disagree with AI among cases with wrong AI outputs.
- **Reject-on-correct**: $\Pr(h_1 \neq a \mid a = y)$ indicates the percentage where the participant’s final decision disagrees with AI among cases with correct AI outputs.
- **Accept-on-wrong**: $\Pr(h_1 = a \mid a \neq y)$ indicates the percentage where the participant’s final decision agrees with AI among cases with wrong AI outputs

To further characterize safety-relevant reliance outcomes, we report a help–harm decomposition. N the number of cases for a participant in a given condition.

- **AI-help rate** ($\frac{1}{N} \sum_{j=1}^N \mathbb{I}[h_{0j} \neq y_j \wedge h_{1j} = y_j]$): quantifies cases where the participant is initially incorrect but becomes correct after reviewing AI
- **AI-harm rate** ($\frac{1}{N} \sum_{j=1}^N \mathbb{I}[h_{0j} = y_j \wedge h_{1j} \neq y_j]$): quantifies cases where the participant is initially correct but becomes incorrect after reviewing AI.
- **Missed-help rate** ($\frac{1}{N} \sum_{j=1}^N \mathbb{I}[h_{0j} \neq y_j \wedge a_j = y_j \wedge h_{1j} \neq a_j]$): captures cases where the participant is initially incorrect and the AI prediction is correct, but the participant does not follow the AI and remains incorrect.
- **Correct-ignore rate** ($\frac{1}{N} \sum_{j=1}^N \mathbb{I}[h_{0j} = y_j \wedge a_j \neq y_j \wedge h_{1j} \neq a_j]$): captures cases where the participant is initially correct and the AI prediction is incorrect, and the participant rejects the AI and remains correct.

To characterize when improvements arise and whether decision changes introduce harm, we report decision-change behaviors (Changed/ChangedToRight/ChangedToWrong):

- **Changed_rate** ($\mathbb{I}[h_1 \neq h_0]$) indicates the percentage of cases where the participant changes their initial decision after viewing AI.
- **ChangedToRight_rate** ($\mathbb{I}[h_1 \neq h_0] \cdot \mathbb{I}[h_0 \neq y] \cdot \mathbb{I}[h_1 = y]$) indicates the percentage of cases where the participant changes their decision and ends up correct.
- **ChangedToWrong_rate** ($\mathbb{I}[h_1 \neq h_0] \cdot \mathbb{I}[h_0 = y] \cdot \mathbb{I}[h_1 \neq y]$) indicates the percentage of cases where the participant changes their decision and ends up incorrect.

Table 6. Detailed Demographics of Participants: Domain experts, therapists, who have experience in stroke rehabilitation more than one year (TP1 - TP9) and novices (NV1 - NV12) (i.e. other health professionals and students majoring in medicine, therapy)

ID	Age	Gender	Role	Setting / Others	Experience
TP1	35 - 44 years	Female	Occupational Therapist (OT)	Inpatient Rehabilitation	13
TP2	25 - 34 years	Female	Occupational Therapist (OT)	Outpatient Clinic	8
TP3	25 - 34 years	Female	Occupational Therapist (OT)	Outpatient Clinic	9
TP4	25 - 34 years	Female	Occupational Therapist (OT)	Skilled Nursing Facility	6
TP5	25 - 34 years	Male	Occupational Therapist (OT)	Other (please specify)	7
TP6	55 - 64 years	Female	PhysioTherapist (PT)	Skilled Nursing Facility	30
TP7	25 - 34 years	Male	PhysioTherapist (PT)	Outpatient Clinic	2
TP8	35 - 44 years	Female	PhysioTherapist (PT)	Outpatient Clinic	4
TP9	35 - 44 years	Male	PhysioTherapist (PT)	Skilled Nursing Facility	12
NV1	25 - 34 years	Male	Student-Physiotherapy		
NV2	18 - 24 years	Female	Student-Physiotherapy		
NV3	25 - 34 years	Female	Speech Therapist	Private Therapy Centre	2
NV4	35 - 44 years	Female	Nurse	Home Care Centre Acute Care Centre	13
NV5	25 - 34 years	Male	Nurse	(experience with post-stroke survivors) Acute Care Centre	7
NV6	25 - 34 years	Male	Nurse	(experience with post-stroke survivors)	8
NV7	25 - 34 years	Female	Nurse	Nursing home	10
NV8	25 - 34 years	Male	Healthcare Worker	Community Hospital, Inpatient	6
NV9	18 - 24 years	Female	Student-Medicine		
NV10	18 - 24 years	Male	Student-Medicine		
NV11	18 - 24 years	Female	Student-Medicine		
NV12	18 - 24 years	Female	Student-Nursing		

Table 7. Performance of decision-making without/with AI-assistance by all participants (All), domain experts (EXPs), and health professionals and students (NVs)

	All			Experts			Novices		
	Human	Human-AI	Changes	Human	Human-AI	Changes	Human	Human-AI	Changes
ConditionA	67.69	69.71	↑ 2.02 (p=0.47)	71.35	70.19	↓ -1.16 (p=0.56)	64.94	69.35	↑ 4.40 (p=0.18)
ConditionB	67.86	82.02	↑ 14.16 (p<0.001)	72.22	83.43	↑ 11.21 (p<0.05)	64.58	80.95	↑ 16.37 (p<0.01)

A.5 Phase 2: Interactive Refinement and Offline Adaptation

A.5.1 *Feedback-to-update mechanisms. Re-label feedback (dataset curation for offline adaptation).* For each trial where a participant disagrees with the AI output and provides a corrected label, we create a feedback instance (x, \tilde{y}) ,

User Interface
Phase 1

COMP Assessment
Complete your evaluation for COMP predictions

Agreement AI Confidence Final Assessment

Step 1: Decision on AI prediction ✓

Do you agree with the AI prediction?

Agree Disagree

Step 2: Rate your confidence in AI's COMP prediction ✓

How confident are you that the AI's COMP prediction is correct?

Not Sure Very Sure

0% 20% 40% 60% 80% 100%

Selected: 33 %

Submit new COMP Assessment

Your COMP assessment:

1

Complete COMP

Fig. 4. User Interface of Phase 1.

Table 8. Detailed performance and team-gain metrics for all participants (All), domain experts (EXPs), and health professionals and students (NVs)

Metrics	All			Experts			Novices		
	CondA	CondB	Δ B-A (P-Value)	CondA	CondB	Δ B-A (P-Value)	CondA	CondB	Δ B-A (P-Value)
Human	67.69	67.86	0.17 (p=0.86)	71.35	72.22	0.87 (p=0.61)	64.94	64.58	-0.36 (p=0.90)
AI	70.76	73.00	2.24 (p=0.14)	70.58	73.31	2.73 (p=0.5)	70.89	72.77	1.88 (p=1.00)
Human-AI Team	69.71	82.02	\uparrow 12.31 (p<0.001)	70.19	83.43	\uparrow 13.25 (p<0.05)	69.35	80.95	\uparrow 11.61 (p<0.05)
Human-AI Team_gain vs Human	2.02	14.16	\uparrow 12.14 (p<0.01)	-1.16	11.21	\uparrow 12.37 (p<0.05)	4.4	16.37	\uparrow 11.96 (p<0.05)
Human-AI Team_gain vs_AI	-1.05	9.01	\uparrow 10.07 (p<0.05)	-0.4	10.12	10.52 (p=0.20)	-1.55	8.18	\uparrow 9.73 (p<0.1)
Regret_best	16.46	4.80	\downarrow 11.66 (p<0.01)	17.46	3.87	\downarrow 13.59 (p<0.05)	15.71	5.51	\downarrow 10.21 (p<0.05)
Oracle_best	86.17	86.82	0.65 (p=0.63)	87.65	87.30	0.34 (p=1)	85.06	86.46	1.4 (p=0.68)

where x denotes the original model input (e.g., patient trial/window features) and \tilde{y} is the participant-provided label. We operationalize these instances as additional supervised training data by augmenting the original training set:

$$\mathcal{D}_c^{\text{global}} = \mathcal{D}_{\text{base}} \cup \mathcal{F}_c^{\text{global}},$$

Table 9. Detailed acceptance/rejection metrics capturing reliance on AI for all participants (All), domain experts (EXPs), and health professionals and students (NVs)

Metrics	All			Experts			Novices		
	CondA	CondB	Δ B-A (P-Value)	CondA	CondB	Δ B-A (P-Value)	CondA	CondB	Δ B-A (P-Value)
Accept_on_correct	88.09	95.24	\uparrow 7.15 (p<0.01)	86.4	98.89	\uparrow 12.49 (p<0.01)	89.36	92.5	3.14 (p=0.34)
Reject_on_wrong	41.61	62.14	\uparrow 20.53 (p<0.05)	42.33	56.11	13.78 (p=0.40)	41.07	66.67	25.6 (p=0.1)
Reject_on_correct	11.91	4.76	\downarrow -7.15 (p<0.01)	13.6	1.11	\downarrow -12.49 (p<0.01)	10.64	7.5	-3.14 (p=0.34)
Accept_on_wrong	58.39	37.86	\downarrow -20.53 (p<0.01)	57.67	43.89	-13.78 (p=0.40)	58.93	33.33	-25.6 (p=0.1)
AI_help_rate	13.76	15.9	2.14 (p=0.58)	10.85	15.08	4.23 (p=0.62)	15.95	16.52	0.57 (p=0.88)
AI_harm_rate	8.29	3.7	\downarrow -4.59 (p<0.05)	7.91	3.87	-4.04 (p=0.21)	8.57	3.57	-5.00 (p=0.12)
Missed_help_rate	4.72	3.06	-1.66 (p=0.29)	5.45	0	\downarrow -5.45 (p<0.1)	4.17	5.36	1.19 (p=0.76)
Correct_ignore_rate	7.12	10.12	3.00 (p=0.22)	9.15	10.12	0.97 (p=0.64)	5.6	10.12	4.52 (p=0.27)

Table 10. Ratio of Changed, ChangedToRight, and ChangedToWrong rates after reviewing AI outputs (Human + AI) for all participants (All), domain experts (EXPs), and health professionals and students (NVs)

Metrics	All			Experts			Novices		
	CondA	CondB	Δ B-A (P-Value)	CondA	CondB	Δ B-A (P-Value)	CondA	CondB	Δ B-A (P-Value)
Changed_Rate	34.98	26.66	\downarrow -8.32 (p<0.1)	32.41	26.09	-6.32 (p=0.41)	36.9	27.08	-9.82 (p=0.10)
ChangedToRight_Rate	14.74	18.88	4.14 (p=0.13)	12.57	17.46	4.89 (p=0.34)	16.37	19.94	3.57 (p=0.27)
ChangedToWrong_rate	12.72	4.72	\downarrow -8.00 (p<0.01)	13.73	6.25	\downarrow -7.48 (p<0.05)	11.96	3.57	\downarrow -8.39 (p<0.05)

where $\mathcal{F}_c^{\text{global}}$ pools re-label feedback across patients under condition c . We then perform offline adaptation by fine-tuning the same baseline checkpoint using supervised learning on $\mathcal{D}_c^{\text{global}}$ (global), and evaluate on a disjoint held-out set to measure generalization beyond the corrected cases.

Quality control. Before augmentation, we (i) exclude incomplete feedback (missing \tilde{y}), (ii) deduplicate identical instances (same x and \tilde{y}), and (iii) if multiple feedback labels exist for the same case, we consolidate them via a predefined conflict-resolution policy (e.g., majority vote), yielding one consolidated label per case.

Rule-based feedback (explicit inference-time update). Participants can create or edit rules that map detectable conditions (feature/threshold patterns) to corrective actions on the model output.

We treat rules as *patient-specific* by default (authored for a given patient), and evaluate their transfer by applying the same rule layer to other patients’ held-out data.

A.5.2 Local vs. global updates and generalization checks. For each patient i and condition $c \in \{A, B\}$, we construct and evaluate local and global updates.

Table 11. Phase 2 refinement performance. We report improvement over the baseline model on a held-out evaluation set for local (participant-specific) and global (across-patients) updates. p from paired Wilcoxon signed-rank tests across participants.

Metric	All (Experts+Novices)				Experts				Novices			
	A	B	$\Delta B-A$	p	A	B	$\Delta B-A$	p	A	B	$\Delta B-A$	p
Local update ΔPerf (overall)	11.50	36.38	24.89	<0.001	11.20	33.89	22.69	<0.001	11.92	38.88	26.96	<0.05
Global update ΔPerf (overall)	2.38	-5.05	-7.43	<0.001	5.58	-2.95	-8.52	<0.001	0.00	-6.46	-6.46	<0.01

Local updates. use only feedback associated with patient i under condition c : *local rule update* by applying the edited patient-specific rule for patient i .

Global updates / transfer checks. We use two complementary notions of “global” depending on feedback type: (i) *global re-label model* by fine-tuning a single model on $\mathcal{D}_c^{\text{global}}$, and (ii) *cross-patient rule transfer* by applying a patient’s edited rule layer to held-out data from other patients $j \neq i$, measuring spillover effects beyond the authoring patient.

Refinement-performance metrics. Let $\text{Perf}(\cdot)$ denote model performance (percentage; higher is better) on the held-out set. We report improvement scores relative to the baseline:

$$\Delta\text{Perf}_{i,c}^{\text{local}} = \text{Perf}(\text{local update}_{i,c}) - \text{Perf}(\text{baseline})$$

$$\Delta\text{Perf}_c^{\text{global}} = \text{Perf}(\text{global update}_c) - \text{Perf}(\text{baseline}).$$