AGENTIC ADAPTIVE AI FOR CLINICAL CARE: AUTONOMOUS CONSTRUCTION OF PATIENT-SPECIFIC MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

We present an adaptive framework for personalized medicine designed to discover and deploy predictive models tailored to specific patients and their unique physiological dynamics. While static population models fail to capture individual physiological heterogeneity, standard patient-specific fine-tuning is not uniformly beneficial; sparse calibration budgets, high-frequency sensor noise, and severe domain shifts frequently degrade model accuracy below the original baseline, yet conventional systems lack mechanisms to assess whether adaptation should occur or when an unreliable model must be rejected. To address this, our architecture pairs an autonomous AutoResearch agent that iteratively searches across adaptation spaces (architectures, adapter ranks, and loss schedules) with a deterministic verification gate and a domain-specialized diagnostic agent. When a candidate breaches stability or physical invariants, the system triggers a diagnostic recovery loop, providing clinical oversight to pinpoint failure modes and safely arbitrate between matched cohort retrieval, training a scratch patient-only baseline, or refusing inference via data recollection. We evaluate our framework across two distinct clinical modalities: continuous ICU arterial blood pressure tracking from ECG and PPG waveforms (144 unseen PulseDB subjects), and simulation-to-real bioimpedance monitoring for dynamic fluid-shift tracking. Most notably, across both tasks, the framework cuts prediction error by approximately 50% relative to baseline model.

1 INTRODUCTION

Every human body is built differently, possessing a unique genome, size, and physiological makeup. Since the early days of modern medicine, this variety has made it difficult to create a single set of rules or workflows that work for everyone. For a long time, the lack of advanced monitoring and data processing technology meant that general guidelines were a necessity—clinicians had to rely on "average" benchmarks because there was no practical way to account for individual differences at scale. As a result, medical protocols were designed for the majority, often overlooking the specific needs of the individual.

Recent years have brought a paradigm shift toward personalized medicine, moving away from generalized care to strategies tailored to the individual Abrahams & Silver (2009); Johnson et al. (2021). This progress is largely expedited by the integration of big data and advanced manufacturing, which allow for a level of customization previously impossible with traditional technology. Central to these developments is the emergence of large-scale data models capable of processing vast datasets to identify patterns. By leveraging these broad pools of clinical examples, these systems can accurately infer outcomes for new, unseen patients, bridging the gap between historical medical data and individual diagnostic needs.

Within the medical AI domain, it has become increasingly recognized that traditional locked or static models struggle to accommodate the evolving dynamics of clinical practice Lee & Lee (2020). Effective clinical tools must instead be capable of evolving in response to new data Parekh et al. (2023). Currently, model adaptation generally follows three trajectories: data and input adaptation, where the model adjusts to varying data types; weight fine-tuning, which utilizes new datasets to update

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

model parameters Parekh et al. (2023); and structural plasticity, involving real-time architectural shifts as demonstrated in patient-specific learning frameworks for adaptive monitoring in critical care Zhang & Szolovits (2008).

However, patient-specific adaptation is not uniformly beneficial. A population model may already generalize well to some individuals, while adaptation may improve predictions for others or cause negative transfer when calibration data are sparse, noisy, or poorly matched to the population prior. Moreover, different patients may require different adaptation strategies. Personalization therefore presents not only the problem of how to adapt a model, but also whether adaptation should occur, which adaptation strategy should be selected, and when an unreliable adaptation should be rejected.

Parallel to these domain-adaptation efforts, the rise of Large Language Models (LLMs) and autonomous agent frameworks has unlocked new modalities for automated scientific discovery Weng et al. (2025). Moving beyond passive zero-shot prompting, agentic architectures now formulate hypotheses, execute closed-loop code synthesis, and iteratively refine models against execution feedback. Recent systems like AgenticSciML and AutoML-Agent Jiang et al. (2026); Trirat et al. (2025) demonstrate that collaborative multi-agent loops can automate discovery in scientific machine learning, while Karpathy's autoresearch paradigm Karpathy (2026) formalizes continuous iterative program search under constrained compute budgets. Concurrently, specialized foundation agents such as PHIA Merrill et al. (2026) have demonstrated that language models coupled with execution tools can reliably interpret continuous sensor time-series and wearable telemetry, outperforming handcrafted pipelines in personalized health reasoning. In that setting, the agent acts as an interpretability interface—writing code to analyze sensor trends, answer user queries, and extract behavioral insights from existing records, rather than training or optimizing downstream models for real-time predictive tasks. Building predictive pipelines from complex, high-frequency physiological streams requires moving beyond question-answering toward agents that can autonomously adapt and refine the core predictive architectures themselves.

In this work, we present a failure-aware agentic framework that dynamically personalizes physiological predictive pipelines around an individual patient's unique dynamics. Rather than relying on rigid, one-size-fits-all fine-tuning, our framework deploys an AutoResearch agent atop established pretrained backbones to iteratively search for and discover a dedicated, patient-specific model. Crucially, candidate models are audited against deterministic performance and stability thresholds: when a proposed model succeeds, it is retained directly; if it fails, the system triggers a domain-specialized, failure-aware agentic pipeline. This diagnostic loop pinpoints the root failure mode and incorporates human-in-the-loop supervision to select the safest clinical intervention—whether routing to matched cohort retrieval, falling back to a patient-only baseline, or abstaining via data recollection. By uniting autonomous iterative search with supervised failure awareness, our system ensures robust individual customization while explicitly safeguarding against harmful adaptation.

We demonstrate the versatility of this framework across two distinct clinical challenges:

1. **Continuous Blood Pressure Estimation from Patient-Calibrated ECG and PPG:** Using continuous dual-channel ECG and PPG telemetry across 144 unseen patients from the PulseDB benchmark Wang et al. (2023), our framework predicts continuous arterial blood pressure (SBP and DBP). Rather than relying on a static global model, the agentic pipeline rapidly tailors a dedicated regressor to each individual patient's exact blood-pressure baseline using same-day calibration signals, dynamically verifying adaptation viability to safeguard against negative transfer.
2. **Simulation-to-Real Bioimpedance for Fluid Shift Estimation:** Trained on a synthetic archive of over one million spectra simulating physiological variability, fluid-dependent impedance changes, and electrode positioning uncertainty, our agent personalizes a frozen multi-frequency neural backbone to sparse, *in-vivo* measurements collected from real human participants across multiple sensor placements and sessions. By navigating low-rank adaptation parameters, physics-informed branches, and synthetic cohort rehearsal, the framework successfully tracks individual fluid kinetics under severe simulation-to-real domain shift.

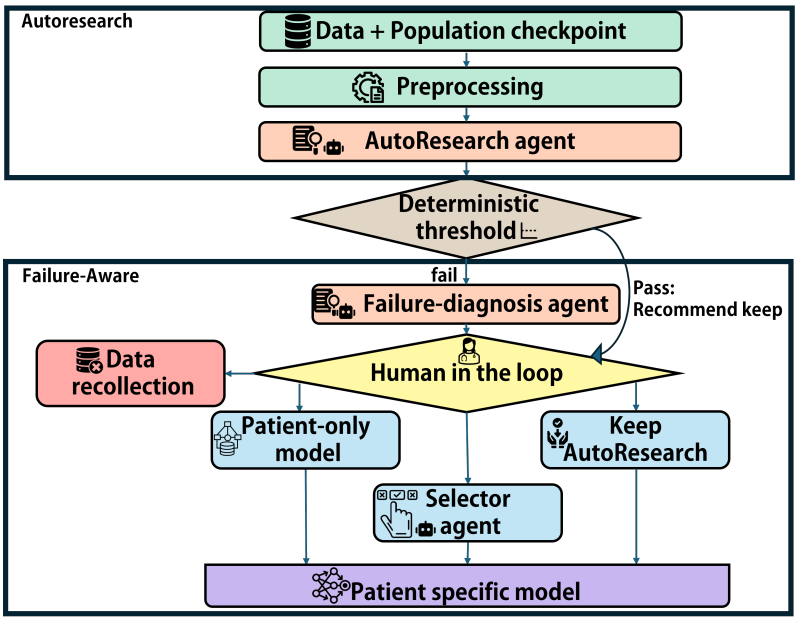

Figure 1: Overview of the multi-agent failure-aware personalization pipeline.

2 METHODS

2.1 MULTI-AGENT FAILURE-AWARE PERSONALIZATION ARCHITECTURE

The end-to-end personalization workflow follows a structured, staged decision pipeline designed to handle physiological heterogeneity while guarding against catastrophic adaptation failure (Fig. 1). Across both domains, the system executes the following operational stages:

1. **Data Ingestion and Population Checkpoints:** Inflow begins with a frozen population backbone trained on broad historical cohorts.
2. **Preprocessing and Signal Conditioning:** Signals undergo quality verification, filtering, and temporal or spectral standardization to align with the population checkpoint's input.
3. **AutoResearch Agent Exploration:** A local LLM agent (`Qwen3-30B`) serves as an AutoResearch engine, exploring bounded adaptation spaces over fine-tuning schedules, adapter ranks (LoRA), loss objectives, and feature representations on development data.
4. **Deterministic Threshold Gate:** Candidate models are evaluated on a held-out validation segment against preregistered stability and non-inferiority margins:
   - If the top candidate clears the criteria, it bypasses diagnosis and proceeds directly to deployment (*Keep AutoResearch*).
   - If the candidate breaches the threshold, execution enters into the failure-diagnosis loop.
5. **Failure-Diagnosis Agent (Domain Specialist):** Upon a threshold breach, a domain-specialized diagnostic agent is invoked to diagnose the failure mode and provide actionable recovery recommendations. Rather than applying generic prompt heuristics, the agent is skilled with task-specific physiological priors and diagnostic capabilities tailored to the target modality:
   - *Data Recollection Recommendation:* Triggered when the calibration data appears corrupted, unphysical, or unrelated to the population distribution due to operational errors (e.g., severe sensor displacement, detached leads, or artifact corruption), advising an outright refusal to predict.

- *Patient-Only Model Recommendation:* Triggered when the data is valid but the patient's physiology diverges severely from the population prior, recommending training a lightweight model exclusively on the patient's calibration data.
- *Selector-Agent Retrieval Recommendation:* Triggered when population priors remain viable, recommending that the *Selector Agent* retrieve nearest-neighbor patients or matched synthetic cohorts to transfer proven adaptation recipes.
- *Keep AutoResearch Recommendation:* Triggered for marginal threshold failures where qualitative stability and error bounds remain clinically acceptable.

6. **Human-in-the-Loop Supervision:**While all candidate routes were evaluated offline to enable comprehensive benchmark comparisons, the final pipeline is designed for real-time human-in-the-loop (HITL) deployment. Here, a clinician or domain specialist leverages the LLM's diagnostic advisory alongside raw telemetry to inspect the failure mode and arbitrate the optimal recovery route dynamically.
7. **Patient-Specific Deployment:** The validated model pipeline selected from either the direct path or the supervised rescue route is deployed as the final patient-specific model.

2.2 CASE 1: CONTINUOUS ICU BLOOD PRESSURE TRACKING (PULSEDB)

Data Cohorts and Preprocessing. We evaluate continuous personalization on the VitalDB supplementary subset of PulseDB v2.0 Wang et al. (2023), whose waveform sources include the MIMIC-III Waveform Database Matched Subset Moody et al. (2020). The dataset comprises quality-screened, bandpass-filtered, and min-max normalized 10-second windows sampled at 125 Hz. Inputs are structured as tensors $X \in \mathbb{R}^{2\times 1250}$ corresponding to synchronized lead-II ECG and fingertip PPG channels. Ground-truth systolic (SBP) and diastolic (DBP) blood pressures are extracted from aligned invasive arterial blood pressure lines. The upstream population training corpus consists of 465,480 windows across 1,293 subjects (360 windows per subject; SBP: 115.5 ± 18.9 mmHg; DBP: 62.9 ± 12.1 mmHg). The out-of-distribution test cohort named (`CalFree`) comprises 57,600 windows from 144 distinct subjects with zero patient overlap with the population set.

Pretrained Population Regressors. Separate 1D ResNet-18 architectures serve as frozen population backbones for SBP and DBP. Following the benchmark by Wang et al. (2023), 1D ResNets consistently demonstrate superior performance over alternative neural architectures for raw physiological time-series regression. Each network features four residual stages followed by global average pooling and a linear projection head. Checkpoints were trained on the 465,480 population windows for 100 epochs using Adam ($\eta = 10^{-6}$, batch size 32) under a mean squared error (MSE) loss.

Calibration Partitioning and Waveform Selection. For each `CalFree` patient, the 400 chronological windows are structured into distinct functional partitions to support staged adaptation without data leakage:

- **Acquisition Pool ($N = 160$):** Serves as the candidate bank from which patient calibration windows are sampled to fit candidate adaptation recipes.
- **Search-Validation Split ($N = 40$):** Used exclusively to evaluate and rank agent-proposed candidates during the exploratory search phase.
- **Routing-Validation Split ($N = 20$):** Held out from search to evaluate candidate stability and inform the deterministic routing decision.
- **Locked Evaluation Segment ($N = 180$):** Completely sealed during search and routing; unlocked exactly once to compute final reported performance.

To mimic bedside data acquisition where reference blood pressure labels are unavailable at capture time, calibration windows are chosen via a label-blind, diversity-maximizing farthest-first traversal over an 18-dimensional feature space (characterizing cardiac cycle duration, pulse arrival latency, signal variance, and spectral flatness).

Blood Pressure Instantiation Across Pipeline Stages. The operational execution and candidate models map directly through the seven-stage architecture:

1. **Data Ingestion & Population Checkpoint (Stage 1):** Calibration windows are ingested alongside the frozen upstream 1D ResNet-18 population regressor, which provides unmodified zero-shot population inferences without patient calibration He et al. (2016).
2. **Preprocessing & Signal Conditioning (Stage 2):** Calibration windows from the patient's acquisition pool undergo signal quality checks, artifact rejection, and temporal standardization before candidate optimization begins.
3. **AutoResearch Agent Exploration & Baseline Candidates (Stage 3):** Candidates are fitted on the patient's calibration pool and ranked according to performance on the 40-window search-validation split. The candidate and baseline comparisons comprise:
   - *Fixed Calibration Baseline:* Standard fine-tuning restricted to the linear projection head atop the frozen backbone (Adam, $\eta = 10^{-4}$, 20 epochs), serving as the reference anchor.
   - *Random Hyperparameter Optimization (Random HPO):* Uniform random sampling across the exact same hyperparameter search space explored by AutoResearch (tuning depth, learning rates, loss functions, optimizer schedules, and backbone architectures $\leq$ 100k parameters) evaluated under an identical computational trial budget, serving as an unguided exploration baseline.
   - *Diagnosis Route without AutoResearch:* An ablation baseline that foregoes agent search entirely and directly executes the domain-specialist diagnostic and retrieval rescue pipeline (Stage 5) starting from the fixed calibration model.
   - *AutoResearch Agent Only:* Closed-loop, feedback-driven optimization and structural configurations proposed by `Qwen3-30B`.
4. **Deterministic Threshold Gate (Stage 4):** The top AutoResearch candidate is evaluated on the independent 20-window routing-validation partition. If it satisfies stability criteria and passes the non-inferiority margin against the fixed baseline, execution branches directly to *Keep AutoResearch*, bypassing downstream diagnostic stages.
5. **Failure-Diagnosis Specialist (Stage 5):** If the candidate breaches the threshold, the domain-specialist diagnostic agent reviews the error distribution and ECG–PPG waveform morphology, issuing a structured recommendation:
   - *Flag for Data Recollection:* Recommended if telemetry reveals detached leads, flatlining, or severe motion artifacts unsuited for inference.
   - *Recommend Patient-Specific Reset:* Recommended if the patient's arterial dynamics severely diverge from population priors, indicating that population weights risk harmful negative transfer.
   - *Recommend Neighbor Model Retrieval:* Recommended if population priors remain viable, advising the *Selector Agent* to transfer recipes from matched patients in the training-set database.
   - *Retain AutoResearch:* Recommended if the threshold violation is borderline and qualitative error bounds remain clinically acceptable.
6. **Human-in-the-Loop Supervision & Rescue Candidates (Stage 6):** The human supervisor reviews the diagnostic advisory and telemetry to lock the recovery path. If a fallback route is selected, the *Selector Agent* instantiates and evaluates rescue candidates on the routing split:
   - *Pretrained Training-Set Model DB:* An offline database containing historical adaptation trajectories from 192 training-set subjects (768 adaptation experiences), from which the agent retrieves the $k = 6$ nearest neighbor subjects via morphological waveform distance to warm-start or transfer verified adaptation recipes.
   - *Patient-Specific Resets:* Four compact models (CNN or feature-MLP, $\leq$ 100k parameters) trained strictly from scratch on the patient's calibration data without utilizing population weights.
   - *Data Recollection / Abstention:* If confirmed by the supervisor, the system halts inference and issues a recollection request.
   - *Keep AutoResearch Override:* The supervisor confirms retention of the agent-adapted model despite the gate breach.
7. **Patient-Specific Deployment (Stage 7):** The finalized model—selected via direct AutoResearch pass, supervised retention, or rescue routing—is permanently frozen and evaluated on the locked 180-window test segment.

2.3 CASE 2: SIMULATION-TO-REAL TRANSFER FOR FLUID SHIFTS (BIOIMPEDANCE)

Fluid-Shift Bioimpedance Simulation Engine. To resolve the scarcity of continuous ground truth, we synthesized 1,064,376 multi-frequency impedance spectra using a combined finite-element and equivalent-circuit simulator. The engine models dynamic fluid depletion ($\Delta V \in [-100, 0]$ mL) while varying tissue geometry, layer dielectric properties, contact impedances, and electrode misplacements. Each spectrum covers 30 logarithmically spaced frequencies (0.1 Hz to 100 kHz) across six spectral channels ($R, X, |Z|, \theta, \partial R/\partial f, \partial X/\partial f$) concatenated with 25 physics-derived features.

Real Self-Collected Acquisition and Preprocessing. Real-world validation data consisted of a self-collected dataset acquired from healthy volunteers ($N = 2$) team using the same tetrapolar bioimpedance hardware across repeated acquisition sessions and electrode placements. The acquisition protocol comprised 11 nominal volume-shift targets, $\Delta V \in \{-0.1, -1, -5, -10, -20, -40, -50, -60, -70, -80, -90\}$ mL, with the initial state ($\Delta V = -0.1$ mL) serving as the reference baseline. Importantly, these 11 values represent target levels rather than individual measurements: each target was measured repeatedly across multiple sessions and electrode placements, yielding multiple spectra per target. During preprocessing, spectra were checked for contact continuity and aligned via log-frequency cubic splines onto the checkpoint's exact 30-frequency grid, followed by feature standardization using frozen synthetic statistics.

Bioimpedance Instantiation Across Pipeline Stages. Adaptation to individual fluid dynamics maps directly onto the seven-stage architecture:

1. **Data Ingestion and Population Checkpoints (Stage 1):** Inflow begins with sparse real-patient calibration points and an upstream frozen backbone: an 11.6-million parameter 1D convolutional residual network with squeeze-and-excitation blocks, stochastic depth, and physics-conditioned FiLM modules, selected as the top performer across synthetic sweeps (detailed architecture in Fig. 3, Appendix). The model was pretrained on 100,000 synthetic spectra using AdamW under a Huber–$\mathcal{L}_1$ loss, preserving dataset normalization statistics directly within the checkpoint.
2. **Preprocessing and Signal Conditioning (Stage 2):** Real-world spectra undergo contact impedance continuity verification, logarithmic spline alignment onto the exact 30-frequency simulation grid, and feature standardization using the population checkpoint's summary statistics.
3. **AutoResearch Agent Exploration & Baseline Adaptation (Stage 3):** Due to the acute sample sparsity of in-vivo nursing sessions, adaptation spaces are explored over the development split rather than full fine-tuning:
   - *Fixed LoRA Adaptation:* Standard low-rank adaptation applied across linear projection layers using default hyperparameters ($r = 8$, $\alpha = 16$, Huber loss) without iterative tuning.
   - *Random HPO:* Evaluates random combinations sampled across the identical search space as the agent (LoRA ranks $r \in \{2, 4, 8, 16, 32\}$, scaling factors α, loss objectives, and FiLM layer conditioning) matched to the exact iteration budget of AutoResearch.
   - *AutoResearch Agent Exploration:* The local LLM agent (`Qwen3-30B`) navigates this bounded adaptation space via feedback-driven hypothesis generation, dynamically optimizing rank parameters, loss formulations (Huber, $\mathcal{L}_1$, or MSE), and selective adaptation of the FiLM physics-conditioning modules based on validation trajectory errors.
4. **Deterministic Threshold Gate (Stage 4):** Candidate models are assessed on the routing partition against strict physical invariants. For the dose–response sweep, candidates must achieve MAE ≤ 10 mL, Spearman correlation $\rho \geq 0.90$, and zero monotonicity violations (ΔV strictly non-increasing across consumption). Passing candidates bypass downstream diagnosis and proceed directly to deployment via *Keep AutoResearch*.
5. **Failure-Diagnosis Agent (Domain Specialist) (Stage 5):** Upon a threshold breach, a domain-specialist diagnostic agent skilled in bioimpedance signal analysis evaluates error trajectories, frequency-dependent phase profiles, and Cole-Cole arc parameters. By tracking the characteristic impedance circle radius to distinguish genuine physiological volume

loss from measurement artifacts or electrode decoupling Fu & Freeborn (2020), the agent issues a structured recovery advisory selecting among four distinct pathways:

- *Data Recollection:* Advised when anomalous phase shifts or saturated contact impedances indicate severe electrode detachment or motion corruption, recommending an outright refusal to estimate ΔV.
- *Patient-Only Rescue Model:* Advised when high-dimensional representation transfer proves unstable, proposing a small model based on power-law fit trained exclusively on the sparse in-vivo calibration points.
- *Selector-Agent Retrieval:* Advised when transfer remains viable but calibration data under-constrains adaptation, directing the *Selector Agent* to query an 81-trajectory synthetic reference bank (810 spectra) to retrieve $m = 8$ matched trajectories mirroring the participant's baseline curvature for replay fine-tuning.
- *Keep AutoResearch:* Advised for marginal metric breaches where physical monotonicity is strictly preserved.

6. **Human-in-the-Loop Supervision (Stage 6):** The clinician or supervisor reviews the diagnostic advisory alongside raw spectral telemetry and validation metrics. The supervisor either ratifies the agent's recommended pathway or overrides it with an alternative route from the diagnostic action space.
7. **Patient-Specific Deployment (Stage 7):** The selected, validated model is frozen and deployed as the patient-specific estimator.

2.4 Methodological Robustness and Claim Scope

The integrity of our results relies on deterministic, reproducible constraints: label-blind calibration sampling, immutable cryptographic ledgering, strict physical constraint verification, and single-opening test evaluation. These safeguards mitigate common failure modes in adaptive learning, including validation leakage, hyperparameter cherry-picking, and silent failure under domain shift.

Scope and Boundaries. We distinguish algorithmic validation from broad clinical trial validation. PulseDB evaluations reflect retrospective performance on ICU-derived waveforms within a single institutional database. The bioimpedance study provides proof-of-concept verification of simulation-to-real transfer on a small dataset acquired across multiple placements and sessions; it does not claim multi-patient demographic generalization. Broader deployment across diverse clinical cohorts and trials remains essential prior to operational clinical translation.

3 Results

3.1 Case 1: Continuous ICU Blood Pressure Tracking (PulseDB)

Across the locked evaluation segment of the `CalFree` cohort, static population models fail to generalize reliably, yielding negative explained variance across both targets (Table 1). While fixed personalization reduces this error, it remains considerably less effective than patient-specific search. Autonomous iterative search via AutoResearch cuts prediction error by over half relative to the population baseline, outperforming unguided random exploration (Random HPO) and isolated diagnosis routing. The complete failure-aware pipeline, scored strictly according to the diagnostic agent's routing decisions, achieves the lowest overall error across both SBP and DBP targets (Table 1).

The calibration-budget analysis demonstrates that while fixed personalization improves gradually with additional data, both AutoResearch and the failure-aware pipeline consistently maintain lower estimation errors across all data regimes (Fig. 2A–B). The close aggregate performance between AutoResearch and the failure-aware pipeline is by design: AutoResearch drives optimization under standard conditions, whereas the failure-aware module acts strictly as a safety-critical fallback to catastrophic adaptation failures. When the deterministic gate flags an unstable AutoResearch candidate, the diagnostic router effectively identifies the optimal recovery route in the majority of cases (76.9% for SBP and 70.2% for DBP; Fig. 2D). While our benchmark evaluation executes the route selected by the diagnostic agent to evaluate empirical limits—eliminating negative transfer across all 144 patients at 160 calibration windows (Fig. 2C)—within clinical deployment the router operates

Method	SBP MAE ↓	DBP MAE ↓	SBP R^2 ↑	DBP R^2 ↑
Population	12.65	8.53	-0.779	-1.645
Fixed personalization	8.85	5.35	0.224	0.012
Random HPO	6.92	5.27	0.577	0.523
Diagnosis Route without AutoResearch	6.03	3.51	0.61	0.541
AutoResearch	5.77	3.32	0.656	0.599
Failure-aware	**5.72**	**3.28**	**0.661**	**0.604**

Table 1: PulseDB results on the locked evaluation segment (`CalFree` cohort). Lower MAE is better; higher R^2 is better.

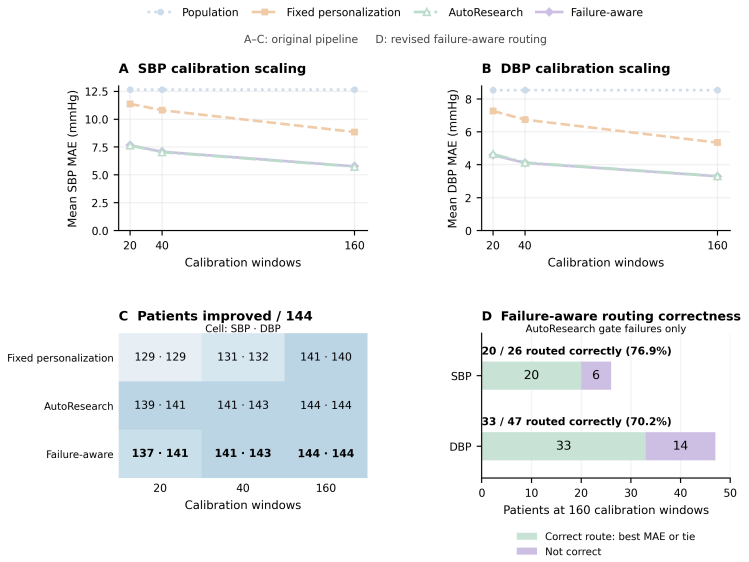

Figure 2: PulseDB results overview at 160 calibration windows. The figure summarizes patient-level error distributions and the fraction of patients improved relative to the frozen population model.

only as a suggestion agent, providing grounded clinical rationale to empower human-in-the-loop oversight and safeguard against individual failure modes.

3.2 Case 2: Simulation-to-Real Transfer for Fluid Shifts (Bioimpedance)

On the self-collected bioimpedance data under severe simulation-to-real shift, the unadapted synthetic backbone exhibits substantial systematic bias, yielding near-zero explained variance (Table 2). Under extreme data scarcity with only one real calibration target, standard parameter-efficient adaptation methods(including Fixed LoRA, Random HPO, and AutoResearch) provide limited corrective capacity. Because the underlying transfer representation is compromised by unmodeled contact dynamics and morphological disparities, the deterministic gate flags the candidate, triggering the domain-specialist diagnostic agent. Rather than enforcing an ill-conditioned neural adaptation, the failure-aware pipeline routes execution to the patient-only calibration model, cutting estimation error by nearly half and restoring physical consistency (Table 2).

When two calibration targets are introduced, neural fine-tuning strategies begin to narrow the simulation-to-real gap, yet the failure-aware route consistently maintains superior accuracy and trajectory fidelity (Table 2). Across both calibration regimes, the router rejects the transferred neural representation in favor of the lightweight patient-specific model. Crucially, this arbitration is guided by the diagnostic agent's grounded reasoning over underlying biophysical invariants rather than heuristic thresholding.

As summarized in Table 2, the diagnostic agent explicitly evaluates Cole–Cole dispersion dynamics to decouple true tissue volume changes from interface artifacts. In the single-target setting, the agent detects that while spectral shape mirrors synthetic expectations, absolute magnitude calibration fails,

Real calibration	Personalization strategy	MAE ↓ (mL)	R^2 ↑	Selected by router	Routing message
1 target	Population model (no adaptation)	23.53	0.036	—	"The decision is driven by the divergence
	Fixed LoRA adaptation	23.18	0.182	—	between spectral shape agreement and
	Random HPO	21.13	0.227	—	magnitude prediction accuracy." Here,
	AutoResearch adaptation	21.26	0.180	—	one calibration label sets the scale; the
	Failure-aware	**12.18**	**0.701**	✓	exponent is fixed at 1.
2 targets	Population model (no adaptation)	23.53	0.036	—	"The calibration_only route avoids this
	Fixed LoRA adaptation	16.27	0.524	—	specific transfer error by relying on local
	Random HPO	12.42	0.596	—	nonlinear fit supported by available labels,
	AutoResearch adaptation	14.38	0.512	—	though its extrapolation remains an
	Failure-aware	**8.62**	**0.931**	✓	assumption given the lack of independent contact measurements."

Table 2: Real calibration routing results on the bioimpedance task. The router selects the calibration-only model under both budgets; routing messages are shown for each calibration budget.

motivating an intervention that fixes the scaling behavior to the solitary reference. When two points are available, the diagnostic loop attributes the residual transfer gap to unobserved electrode contact variations rather than true physiological shifts, opting for a local non-linear fit that circumvents sim-to-real distortion entirely. These diagnostic explanations demonstrate that under severe domain mismatch, the primary strength of the failure-aware framework lies beyond numerical parameter search: it provides an interpretable, physically grounded safety layer. **Patient-only rescue model.** For a calibration budget of k target levels, the power law model ($\leq$ 100k parameters) is trained from scratch using the $6k$ real spectra corresponding to those k calibration targets.

4 Discussion and Conclusion

Static population models struggle to capture the complex, evolving dynamics of individual physiology, frequently inducing negative transfer when applied uniformly across heterogeneous clinical cohorts Lee & Lee (2020); Parekh et al. (2023). While autonomous agent frameworks have demonstrated significant promise in scientific exploration Weng et al. (2025); Jiang et al. (2026); Karpathy (2026), deploying agentic discovery directly in clinical monitoring requires verifiable safety guarantees. In this work, we introduced a failure-aware personalization framework that unifies autonomous architecture search with domain-specialized diagnostic agents with human-in-the-loop oversight to ensure robust patient customization without risking catastrophic adaptation.

Our empirical findings across two distinct physiological tasks demonstrate necessity of this approach. In continuous ICU arterial blood pressure tracking on PulseDB Wang et al. (2023), AutoResearch candidate search halved estimation errors relative to static population backbones, while failure-aware routing eliminated negative transfer across all 144 unseen test patients (Table 1, Fig. 2). Crucially, the diagnostic loop routed unstable candidates to matched cohort retrieval or patient-specific resets, accommodating the vast heterogeneity observed across individual patient trajectories and between SBP and DBP targets (Appendix Fig. 4). In simulation-to-real bioimpedance monitoring for fluid shift tracking, the system successfully enforced strict physical monotonicity invariants, rejecting corrupted transfer representations in favor of lightweight patient-only models that restored physically plausible volume estimation under severe data scarcity (Table 2, Appendix Fig. 5).

Limitations and Future Work. While these results validate the algorithmic viability of deterministic safety routing, several boundaries remain. The blood pressure evaluation relies on retrospective ICU telemetry from a single institutional cohort, and the bioimpedance analysis serves only as a proof-of-concept for simulation-to-real transfer under repeated sessions and placements. Operational translation will require validating these systems prospectively in diverse demographic populations. Additionally, our exploratory search and diagnostic loop were instantiated using a local open-weight coding model (`Qwen3-30B`); while it yielded compelling personalization recipes, evaluating more advanced reasoning backbones or frontier foundation models could further expand search efficiency, improve failure diagnosis, and elevate overall downstream accuracy. Expanding the diagnostic agent's intervention space to handle broader real-time bedside anomalies remains an active avenue for future research. Ultimately, our findings suggest that safe personalized medicine requires AI systems capable not only of discovering patient-tailored models, but also of auditing, diagnosing, and refusing harmful adaptations.

Reproducibility Statement

To facilitate reproducibility, detailed implementation and experimental configuration information is provided in Appendix A.2, including the agent configuration, verbatim prompt structure, search and routing procedures, deterministic thresholds, randomization protocol, data partitions, and training settings used for the PulseDB experiments. Pretrained population models and dataset splits follow the PulseDB experimental protocol described in Section 2.2. For the bioimpedance evaluation, raw participant measurements cannot be shared publicly due to privacy restrictions governing human-subject data. Detailed simulation parameters, equivalent-circuit formulations, preprocessing procedures, and model architecture are documented in Section 2.3 and Appendix A.1.1 to support independent algorithmic replication.

Ethics Statement

The bioimpedance data collection involving human participants was conducted under approval from the appropriate Institutional Review Board (IRB), with informed consent obtained in accordance with the approved study protocol. The specific protocol identifier and institutional information are withheld during double-blind review to preserve author anonymity and will be provided in the final version. The PulseDB experiments use previously released de-identified physiological waveform data. Our experiments are intended as algorithmic and proof-of-concept evaluations and do not constitute clinical validation or support autonomous clinical decision-making.

Use of AI Assistance

Large language models were utilized to assist with experimental code templates, exploratory hyperparameter scripting, and drafting initial text portions. All core conceptual frameworks, scientific hypotheses, pipeline architectures, empirical analyses, and final interpretations are original to the authors. All AI-assisted code, model outputs, and manuscript text were thoroughly audited, verified, and edited under direct human supervision.

References

Edward Abrahams and Mike Silver. The case for personalized medicine, 2009.

Bo Fu and Todd J. Freeborn. Cole-impedance parameters representing biceps tissue bioimpedance in healthy adults and their alterations following eccentric exercise. *Journal of Advanced Research*, 25:285–293, 2020. doi: 10.1016/j.jare.2020.05.016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Qile Jiang et al. Agenticsciml: collaborative multi-agent systems for emergent discovery in scientific machine learning. *Nature Machine Intelligence*, 2026. doi: 10.1038/s44387-026-00102-5.

Kevin B Johnson, Wei-Qi Wei, Dilhan Weeraratne, Mark E Frisse, Karl Misulis, Kyu Rhee, Juan Zhao, and Jane L Snowdon. Precision medicine, ai, and the future of personalized health care. *Clinical and Translational Science*, 14(1):86–93, 2021.

Andrej Karpathy. autoresearch. `https://github.com/karpathy/autoresearch`, 2026.

Cecilia S Lee and Aaron Y Lee. Applications of continual learning machine learning in clinical practice. *The Lancet Digital Health*, 2(6):e279–e281, 2020.

Mike A. Merrill, Akshay Paruchuri, et al. Transforming wearable data into personal health insights using large language model agents. *Nature Communications*, 17(1):1143, 2026. doi: 10.1038/s41467-025-67922-y.

Benjamin Moody, George Moody, Mauricio Villarroel, Gari D. Clifford, and Ikaro Silva. MIMIC-III Waveform Database Matched Subset. *PhysioNet*, April 2020. doi: 10.13026/c2294b. URL `https://doi.org/10.13026/c2294b`. Version 1.0.

National Institute of Standards and Technology. Secure hash standard (shs). Technical Report FIPS PUB 180-4, National Institute of Standards and Technology, 2015.

Ollama. Ollama: Run large language models locally. https://github.com/ollama/ollama, 2026. Software.

Ad-Duhaa E Parekh, Omer A Shaikh, Simran, Sadia Manan, and Md Al Hasibuzzaman. Artificial intelligence (ai) in personalized medicine: Ai-generated personalized therapy regimens based on genetic and medical history: short communication. *Annals of Medicine and Surgery*, 85(11): 5831–5833, 2023.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research: A report from the neurips 2019 reproducibility program. *Journal of Machine Learning Research*, 22(164):1–20, 2021.

Qwen Team. Qwen3-coder-30b-a3b-instruct. https://huggingface.co/Qwen/Qwen3-Coder-30B-A3B-Instruct, 2025. Open-weight model card.

Patara Trirat, Wonyong Jeong, and Sung Ju Hwang. AutoML-agent: A multi-agent LLM framework for full-pipeline AutoML. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 60099–60146. PMLR, 2025.

Weinan Wang, Pedram Mohseni, Kevin L. Kilgore, and Laleh Najafizadeh. PulseDB: A large, cleaned dataset based on MIMIC-III and VitalDB for benchmarking cuff-less blood pressure estimation methods. *Frontiers in Digital Health*, 4:1090854, 2023. doi: 10.3389/fdgth.2022.1090854.

Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. Cycleresearcher: Improving automated research via automated review. In *International Conference on Learning Representations*, pp. 3669–3709, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Ying Zhang and Peter Szolovits. Patient-specific learning in real time for adaptive monitoring in critical care. *Journal of Biomedical Informatics*, 41(3):452–460, 2008.

A APPENDIX

A.1 ADDITIONAL FIGURES

This appendix contains additional figures and analyses that are referenced but not shown in the main text.

A.1.1 BACKBONE ARCHITECTURE TRAINED ON SYNTHETIC DATASET

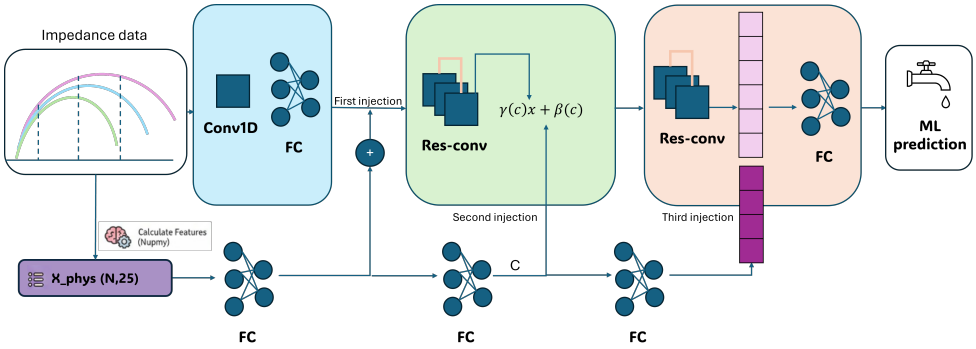

Figure 3: The model integrates multi-channel spectral sequences with 25 derived physiological features ($\mathbf{X}_{phys}$). These features are processed through an MLP and injected at three critical hierarchies built on residual feature extraction: (1) **Early Injection**, which applies a spatial bias to the initial feature maps; (2) **Mid-Stage FiLM**, utilizing Feature-wise Linear Modulation to dynamically scale and shift residual features Perez et al. (2018); and (3) **Late Injection**, concatenation before the final regression head to preserve physical priors for the volume prediction.

A.1.2 PULSEDB

PulseDB: representative patient heterogeneity. PulseDB provides large-scale synchronized physiological waveform data for cuff-less blood-pressure estimation and contains substantial inter-subject variability in cardiovascular signals **?**. Such variability motivates patient-specific learning approaches in physiological monitoring, where adaptation to an individual's observed dynamics can complement population-level models Zhang & Szolovits (2008).

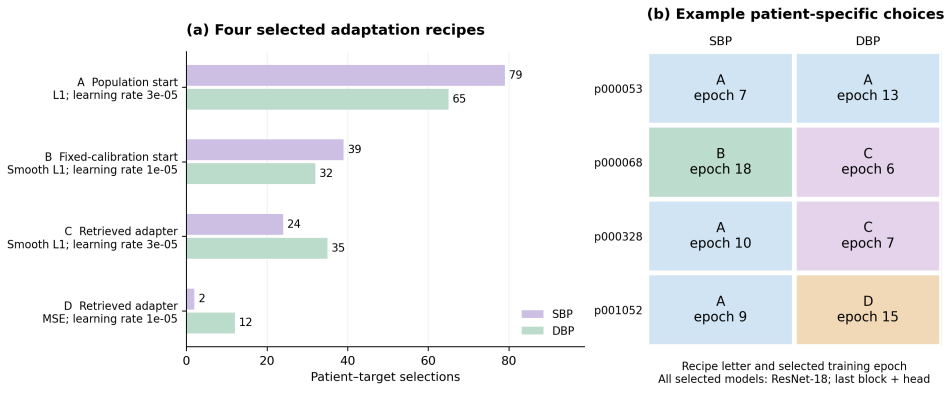

Figure 4: PulseDB patient heterogeneity and model-behavior variation. Representative examples illustrating that waveform morphology and baseline shifts can yield divergent personalization outcomes, motivating failure-aware routing rather than unconditional adaptation. (a) Patient-level optimization trajectories differ across warm-starts and retrieved neighbors, indicating that no single fixed recipe is uniformly reliable. (b) Even within one individual, SBP and DBP can prefer different adaptation horizons/recipes, supporting our design choice to route and personalize targets separately.

A.1.3 BIOIMPEDANCE

Bioimpedance measurements reflect frequency-dependent electrical properties of biological tissue and can also be affected by measurement and electrode-interface conditions. Cole-domain impedance characteristics provide a useful representation for examining physiologically related changes in tissue impedance across measurement conditions Fu & Freeborn (2020).

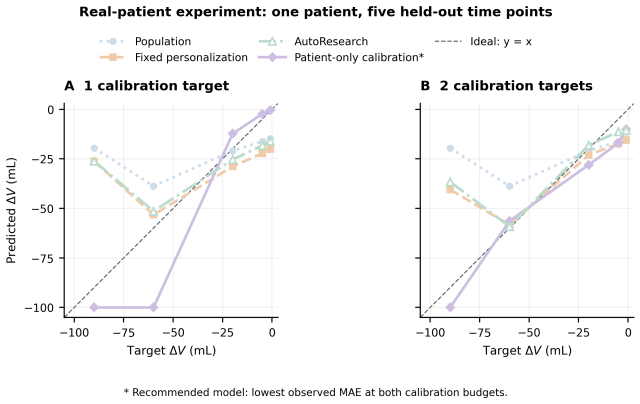

Figure 5: Predicted vs. target volume shift (ΔV) on the self-collected sweep for a representative participant (Subject 1 of $N = 2$). Each target corresponds to multiple samples; points show the mean predicted and mean target ΔV for that target. Points near the diagonal indicate accurate volume-shift estimation; systematic deviations indicate bias under sim-to-real mismatch. This figure supports our claim that the router should reject unstable transfer (e.g., synthetic-to-real adaptation that violates physical expectations) and instead select a patient-only rescue model that more closely tracks the target trajectory under sparse calibration.

A.2 PULSEDB AGENT REPRODUCIBILITY DETAILS

All PulseDB AutoResearch experiments were executed on a single NVIDIA L40 GPU. The agent was instantiated using Qwen3-Coder-30B (`qwen3-coder:30b`), a coding-oriented model from the Qwen3 family Yang et al. (2025); Qwen Team (2025), with `Q4_K_M` quantization and served locally using Ollama version 0.30.10 Ollama (2026). AutoResearch inference used a temperature of 0.2, a maximum of 1,800 generated tokens, a 35 s request timeout, and one retry.

The agent prompt instructed the model to modify a supplied JSON template while preserving exactly four candidate configurations, initializing the first two from the population model and the remaining two from fixed personalization, and relying exclusively on evidence supplied in the prompt. Agent observations contained the subject and prediction target, calibration budget, data-integrity summaries, and the population-checkpoint hash. The returned `candidates` array specified `id`, `start`, `scope` (encoding initialization and trainable architecture/scope), `loss`, `learning_rate`, `weight_decay`, `epochs`, `batch_size`, `patience`, and `hypothesis`.

Candidate selection minimized MAE on the 40-window search-validation split, with candidate ID used as a deterministic tie breaker. An independent 20-window routing-validation split governed subsequent routing decisions, while the locked 180-window evaluation segment was excluded from both search and routing and was accessed only for final scoring. Separating development and model-selection evidence from the final evaluation set is consistent with established recommendations for transparent and reproducible machine-learning experimentation Pineau et al. (2021). ResNet candidate trials terminated at the proposed epoch limit (5–40 epochs), after the proposed number of consecutive non-improving search-validation epochs (2–8), or after the configured 25 s fitting timeout, retaining the best checkpoint observed during training.

Protocol-level sampling used seed 1337. For each subject–target–budget case, the training seed was defined as the integer represented by the first eight hexadecimal digits of the SHA-256 hash National Institute of Standards and Technology (2015) of the compact JSON representation `[subject,target,budget]`. For zero-based AutoResearch candidate index i, the candidate-

specific training seed was defined as

$$\text{seed}_i = \text{seed}_{\text{case}} + 100 + i.$$

The original LLM proposal calls did not specify a generation seed; therefore, exact reproduction of the proposal sequence requires the stored agent responses, whereas all downstream candidate training and evaluation stages are deterministically seeded.

For deterministic safety routing, let S denote the AutoResearch search-validation MAE, V its independent routing-validation MAE, and F the routing-validation MAE of fixed personalization. For SBP, the frozen gate retained AutoResearch only when

$$V \leq 8.4979919049265261, \qquad V \leq 1.10F, \qquad \frac{V}{\max(S, 10^{-5})} \leq 1.50.$$

For DBP, AutoResearch was retained only when

$$V \leq 5.934352619117226, \qquad V \leq F, \qquad \frac{V}{\max(S, 10^{-5})} \leq 1.25.$$

The deterministic gate was used to decide whether the AutoResearch candidate could proceed directly to deployment or whether execution entered the failure-diagnosis stage. The gate thresholds were determined during training from the average validation error observed on the development subjects and were subsequently frozen before evaluation on the held-out test cohort. When the gate criteria were satisfied, the AutoResearch candidate was retained. When the gate was breached, the domain-specialized diagnostic agent was invoked to recommend a recovery route from the action space described in Section 2.2, including retention of AutoResearch, retrieved-memory adaptation, patient-specific reset, or data recollection/abstention.

For offline benchmarking, all available candidate routes were instantiated and evaluated to enable comprehensive method comparisons. However, the reported *Failure-aware* results in Table 1 and Fig. 2 follow strictly the route recommended by the diagnostic agent; they do not correspond to an oracle selection of the lowest test error across routes. The diagnostic router was provided with calibration and cohort descriptors together with the available search- and routing-validation evidence and returned a structured JSON routing decision. Router inference used temperature 0, seed 1337, a maximum of 400 generated tokens, and at most three attempts to obtain a valid JSON response. All routing decisions were frozen before the locked evaluation segment was accessed. In the intended deployment setting, the diagnostic recommendation is presented to a human supervisor for review, as described in Section 2.1.

Representative execution traces include cases in which the router retained the AutoResearch candidate and cases in which it selected the retrieved-memory route. Evaluation outcomes from the locked test segment were stored separately from the observations provided to the agent and therefore could not influence search or routing decisions. These traces are included solely to illustrate execution of the pipeline and are not treated as an independent evaluation.