IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 3 June 2026, accepted 17 June 2026, date of publication 23 June 2026, date of current version 30 June 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3706484

RESEARCH ARTICLE

Policy-First Architecture for Offline Validation of Emotion-Aware Conversational AI Under Identifiability Constraints

KESARA HARISARA PUNCHIHEWA[1], MANINDA EDIRISOORIYA[2], RAJKUMAR BUYYA[3], (Fellow, IEEE), LAKSHIKA CHANDRADEVA[4], AND ACHALA APONSO[5], (Member, IEEE)

[1]School of Computing, IIT, Colombo, Sri Lanka
[2]Department of Decision Sciences, University of Moratuwa, Moratuwa, Sri Lanka
[3]Quantum Cloud Computing and Distributed Systems (qCLOUDS) Laboratory, School of Computing and Information Systems, The University of Melbourne, Melbourne, VIC 3053, Australia
[4]School of Computing, Dublin City University, Dublin, Ireland
[5]School of Electrical Engineering, Computing and Mathematical Sciences (EECMS), Faculty of Science and Engineering, Curtin University, Bentley, WA 6102, Australia

Corresponding authors: Maninda Edirisooriya (manindaw@uom.lk) and Lakshika Chandradeva (lakshika.chandradeva2@mail.dcu.ie)

ABSTRACT Adaptive conversational AI systems in high-stakes domains require rigorous architectural validation prior to deployment. However, existing evaluation approaches often fail to distinguish which system-level properties can be reliably assessed under fully offline conditions, particularly when ethical constraints preclude real-time experimentation and outcome-based evaluations. This paper proposes a policy-first architecture that decouples intervention decision making from response generation by introducing an explicit intervention policy, while the generation module operates strictly as a downstream policy-gated component. Building on this design, we introduce a deterministic offline replay framework for reproducible validation of system-level properties under fixed logged data. We formalize the offline identifiability boundary, which distinguishes properties evaluable under offline conditions from those that remain non-identifiable without counterfactual action support or outcome-linked feedback. The evaluation focused on architectural properties including intervention controllability, execution stability, auditability of decision logic, explicit non-intervention behavior, and invocation-level structural constraints. Experiments on the Multimodal EmotionLines Dataset (MELD) across 11,351 decision points show that both behavioral cloning and reward-driven offline reinforcement learning converge toward degenerate intervention behavior under restricted action support and absent outcome-linked feedback. These findings demonstrate the limitations of emotion-labelled conversational datasets for intervention-policy learning while showing that deterministic replay can validate execution-level properties such as policy-gated invocation, intervention controllability, auditability, and structural compliance. Content-level safety lies outside the offline identifiability boundary and cannot be established through deterministic offline replay alone, thus requiring human-in-the-loop validation.

INDEX TERMS Affective computing, emotion-aware conversational AI, offline evaluation, identifiability, policy-first architecture, decision-gated systems, artificial intelligence safety, just-in-time adaptive interventions (JITAI).

I. INTRODUCTION

Emotion-aware conversational systems have expanded rapidly in recent years, driven by advances in multimodal affect recognition and large language models (LLMs) [1],

The associate editor coordinating the review of this manuscript and approving it for publication was Maria Chiara Caschera.

[2]. By integrating signals from language, speech, and facial expressions, these systems aim to provide adaptive conversational support that responds to users' moment-to-moment emotional states [3], [4]. Simultaneously, the fluency and contextual capabilities of LLMs have enabled scalable conversational agents for mental health-related applications [5], [6]. However, deploying such systems in safety-critical domains introduces significant challenges because inappropriate, excessive, or poorly timed interventions may negatively affect user trust and engagement [7], [8], [9].

A central challenge in adaptive conversational systems concerns not only what to say, but also when intervening. In many existing implementations, intervention timing and language generation are tightly coupled within end-to-end generative architectures [10], [11]. As a result, the decision to intervene emerges implicitly from model behavior, rather than being governed by an explicit and auditable mechanism. This coupling limits controllability, obscures system-level decision logic, and complicates the enforcement and validation of safety-related constraints.

Existing approaches address aspects of this problem but leave critical gaps. End-to-end generative systems treat intervention behavior as an emergent property of language model outputs, making it difficult to independently regulate or evaluate intervention timing [8], [9], [10]. Just-in-Time Adaptive Intervention (JITAI) frameworks provide a conceptual structure for adaptive decision-making but do not specify how system-level properties can be validated under offline and ethically constrained conditions [11], [12].

Offline reinforcement learning enables policy learning from fixed logged data, but meaningful intervention-policy evaluation generally requires logged actions, reward or outcome signals, and sufficient action-support coverage [13], [14]. Public emotion-labelled conversational datasets such as MELD are designed for emotion recognition rather than intervention learning, and therefore do not provide genuine intervention actions, outcome-linked feedback, or counterfactual action support [15]. Under such constraints, reward-driven offline policy learning is vulnerable to distributional shifts, extrapolation errors, and degenerate policy behavior [16], [17].

This limitation leads to a fundamental methodological question: Which properties of adaptive conversational systems can be reliably evaluated under fully offline conditions, and which are inherently non-identifiable without counterfactual action support or outcome-linked feedback? In the absence of such information, properties that depend on unobserved alternative actions or outcomes cannot be inferred from logged data alone. To address this gap, we formalize the concept of an offline identifiability boundary that distinguishes system properties that can be uniquely determined from fixed-logged data from those that remain fundamentally non-identifiable. To clarify this distinction, Fig. 1 provides a conceptual illustration of the offline identifiability boundary under a deterministic replay.

Figure 1 shows a conceptual map of the offline identifiability boundary formalized in this study. Specifically, it distinguishes execution-identifiable properties that can be validated under deterministic offline replay from outcome-dependent properties that remain non-identifiable without intervention outcomes, counterfactual support, or longitudinal feedback.

Building on this perspective, we propose a policy-first architecture in which the intervention timing is governed by an explicit decision layer, whereas language generation is invoked strictly downstream as a policy-gated process. This design externalizes intervention timing as an independent system variable, enabling explicit non-intervention behavior, deterministic control of intervention frequency, and architectural enforcement of language model invocation constraints.

To support the evaluation under offline constraints, we introduce a deterministic offline replay framework that enables the reproducible validation of system-level properties using fixed logged data.

This study adopts a validation-oriented perspective on adaptive conversational. It does not aim to evaluate the intervention effectiveness, policy optimality, or clinical outcomes. Instead, it focuses on validating system-level execution properties that are directly observable under deterministic offline replay, including intervention controllability, execution stability, decision logic auditability, and the enforcement of invocation-level architectural constraints.

The proposed framework was evaluated using deterministic offline replay on a multimodal emotional line dataset (MELD) [15] across 11,351 conversational decision points. The MELD dataset was used exclusively as a state-distribution dataset for controlled offline validation, rather than as a proxy for real intervention data. Accordingly, the evaluation was intended to be a system validation study, rather than a conversational performance benchmark.

The results demonstrated that both behavioral cloning and reward-driven offline reinforcement learning converged toward degenerate intervention behavior under restricted action support and in the absence of outcome-linked feedback. These observations reflect the structural limitations of emotion-labelled conversational datasets for intervention-policy learning rather than the algorithmic deficiencies of the underlying learning methods.

This study introduces an offline identifiability-aware framework to validate the execution-level properties of adaptive conversational systems under deterministic replay. The framework consisted of four tightly connected components.

- **Formalization of the offline identifiability boundary**, defining which system-level properties can be evaluated from fixed logged data and remain fundamentally non-identifiable. Unlike existing offline RL frameworks [13] and causal identifiability approaches [18], [19], this boundary operates at the level of system execution properties rather than outcome estimation or policy optimality.

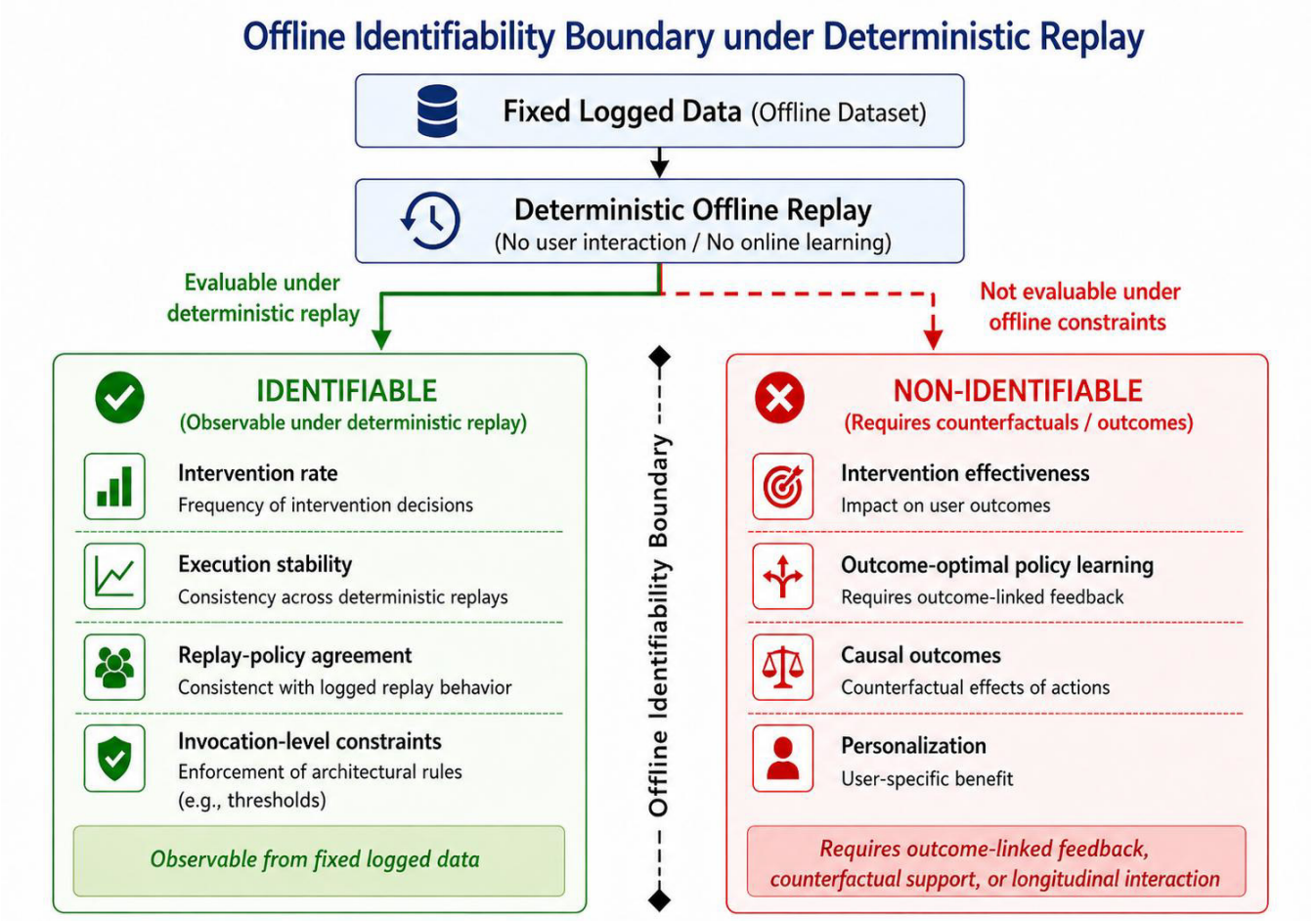

FIGURE 1. Conceptual illustration of the offline identifiability boundary under deterministic offline replay. Execution-level properties that depend only on fixed logged data and replay behavior are identifiable, including intervention rate, execution stability, replay-policy agreement, and invocation-level structural compliance. In contrast, outcome-dependent properties such as intervention effectiveness, optimal policy learning, causal outcomes, and personalization remain non-identifiable under fully offline conditions because they require intervention outcomes, counterfactual support, or longitudinal feedback.

- **Policy-first architectural design** separates intervention decision making from language generation to enable explicit control, auditability, and an explicit non-intervention pathway. Many end-to-end generative conversational architectures invoke response generation without an explicit intervention-decision layer or explicit non-intervention pathway [10].
- **Deterministic offline replay framework** enables the reproducible validation of execution behavior under fully offline conditions. Existing JITAI frameworks [11], [12] define the intervention structure but do not specify how system-level properties can be validated under ethically constrained offline settings.
- **Diagnostic characterization of reward-driven offline reinforcement learning** demonstrates its structural limitations under constrained action support and reframes its role as a diagnostic probe rather than a performance benchmark. Prior offline RL research has extensively studied instability, uncertainty, and extrapolation challenges under limited action support and distributional shifts [13], [16]. In the present study, the observed collapse behavior was interpreted within the context of emotion-labelled conversational datasets lacking intervention outcomes and counterfactual support.

Among these contributions, the primary contribution is the offline identifiability boundary as a property-level criterion for validating execution behavior under deterministic offline replay. The policy-first architecture, deterministic replay framework, and diagnostic reinforcement-learning analysis operationalize this central contribution by separating execution-identifiable properties from outcome-dependent properties that cannot be validated under fully offline conditions.

The remainder of this paper is organized as follows. Section II reviews related work. Section III presents the system architecture and its formal definitions. The experimental setup is described in Section IV. Section V reports the results. Section VI discusses the findings of this study. Section VII outlines limitations and threats to validity. Section VIII presents directions for future research. Finally, Section IX concludes the paper.

II. RELATED WORK

Understanding the motivation for the proposed architecture requires situating it within prior work on emotion-aware AI, adaptive intervention frameworks, offline policy learning, and LLM safety.

Emotion-aware AI systems integrate multimodal signals such as language, speech, and facial expressions to infer user affective states [1], [2]. Multimodal models have demonstrated improved robustness over unimodal approaches, particularly under noisy or partially missing inputs [3], [4]. However, affect recognition alone does not specify when or how a system should intervene [11]. Just-in-Time Adaptive Interventions (JITAIs) address this issue by formalizing adaptive support as structured decision processes with explicit tailoring variables and decision rules [11], [12]. Computational implementations commonly adopt reinforcement learning or contextual bandit formulations to learn intervention policies from data [13], [14]. However, such methods depend on genuine intervention annotations, sufficient action coverage, and outcome-linked feedback, which are systematically absent in publicly available emotion-labelled conversational datasets [13], [14]. In parallel, large language models have enabled scalable, fluent conversational agents with strong contextual capabilities [5], [6], motivating their application in mental health-related systems [8], [20], [21]. However, prior studies have documented risks, including hallucinations, boundary violations, and inappropriate guidance in vulnerable settings, motivating calls for structured governance of language model invocations [8], [22], [23].

Table 1 summarizes the key architectural and evaluation features of representative prior studies in comparison with those of the proposed framework. This feature-level comparison highlights the gaps motivating this study, without implying that the cited works lack value in their target domains.

As shown in Table 1, no prior work has simultaneously addressed multimodal affect recognition, explicit intervention policy separation, offline identifiability, deterministic replay validation, or invocation-level structural constraints within a unified framework. This gap motivated the development of the proposed policy-first architecture.

Despite these advances, however, a critical gap remains. End-to-end generative architectures treat intervention timing as an emergent property of model outputs rather than as an explicit, auditable decision variable that limits controllability and complicates safety validation [8], [10], [23]. The JITAI framework defines the structure of adaptive interventions, but does not specify how system-level properties can be verified under fully offline and ethically constrained evaluation conditions [11], [12]. Offline reinforcement learning offers policy learning from fixed logged data, but requires action diversity and outcome signals that emotion-labelled datasets do not provide, leading to known failure modes, including distributional shift, policy collapse, and extrapolation error under narrow action support [13], [16], [17]. Causal inference frameworks address the identifiability of treatment effects under structural assumptions [18], [19] but operate at the level of outcome estimation rather than system execution and do not characterize which architectural properties of a deployed system are verifiable from logged data alone. Consequently, no prior work has explicitly defined which system-level properties of adaptive conversational systems are identifiable from fixed-logged data, which are fundamentally non-identifiable without counterfactual or outcome-linked information. This gap motivates the offline identifiability boundary introduced in this study, and the accompanying validation-oriented framework for decision-gated conversational systems.

III. SYSTEM ARCHITECTURE AND METHODS

Definition 1 (Offline Identifiability Boundary): Let $D = \{(s_t, a_t)\}_{t=1}^{T}$ denote a fixed logged dataset of state–action pairs collected under a logging policy, where s_t denotes the observed system state and a_t denotes the logged intervention action. Let R denote the deterministic offline replay procedure applied to D under a fixed configuration, policy parameters, and random seeds.

A system property P is offline-identifiable with respect to D and R if P can be uniquely determined from the replay trace produced by $R(D)$ without requiring counterfactual action support, unobserved user responses, or outcome-linked feedback. Equivalently, P is offline-identifiable if there exists a deterministic mapping g for which $P = g(R(D))$.

Properties that depend on unobserved alternative actions, intervention outcomes, user responses, or longitudinal feedback are non-identifiable under fully offline conditions.

Observation 1: Under Definition 1, system properties that depend only on observed execution under fixed-logged data are offline-identifiable. These include the intervention rate, execution stability under deterministic replay, policy agreement with a reference logging policy, explicit non-intervention behavior, and invocation-level structural compliance. By contrast, properties that require counterfactual action support or outcome-linked feedback, including causal intervention effectiveness, optimal policy learning, content-level safety, and personalization quality, are non-identifiable under fully offline conditions.

Corollary 1: A system property may be causally non-identifiable while remaining execution-identifiable under deterministic replay. A property may be fully observable as part of a deterministic system execution, whereas its effect on user outcomes remains non-identifiable from the same logged data.

Proof sketch: Consider execution-level properties such as the intervention rate. Given D and deterministic replay procedure R, the action sequence produced during replay is fixed. Therefore, the intervention rate is uniquely determined from $R(D)$, which makes it identifiable offline under Definition 1. By contrast, causal intervention effectiveness depends on comparing outcomes under actions taken and actions not taken, such as the difference between the user response under $a_t = 1$ and the unobserved response under $a_t = 0$. Because D contains no counterfactual outcomes or outcome-linked

TABLE 1. Comparison of representative prior works and the proposed framework across key architectural and evaluation features. (✓ = feature addressed; ✗ = not addressed).

Feature	[3]	[4]	[11]	[12]	[13]	[14]	[8]	[10]	Proposed
Multimodal Affect Recognition	✓	✓	✗	✗	✗	✗	✗	✗	✓
Explicit Intervention Policy Layer	✗	✗	✗	✗	✗	✗	✗	✗	✓
Policy-Gated Language Generation	✗	✗	✗	✗	✗	✗	✗	✗	✓
Offline Evaluation Support	✗	✗	✗	✗	✓	✓	✗	✗	✓
Offline Identifiability Boundary	✗	✗	✗	✗	✗	✗	✗	✗	✓
Deterministic Replay Framework	✗	✗	✗	✗	✗	✗	✗	✗	✓
Explicit Non-Intervention Pathway	✗	✗	✗	✗	✗	✗	✗	✗	✓
Invocation-Level Structural Constraints	✗	✗	✗	✗	✗	✗	✗	✗	✓
LLM Safety Governance	✗	✗	✗	✗	✗	✗	✓	✗	✓
Ethical Offline Constraint Handling	✗	✗	✗	✗	✗	✗	✗	✗	✓

feedback, two distinct causal mechanisms can produce the same logged dataset D and the same replay trace $R(D)$ while yielding different intervention effects.

The notion of identifiability used here is related to but distinct from identifiability in causal inference. In causal inference, identifiability concerns whether causal effects can be uniquely estimated from observational data under assumptions, such as those formalized in Pearl's do-calculus [18] and Rubin's potential outcomes framework [19]. In this study, the offline identifiability boundary operates at the level of system execution rather than treatment-effect estimation. It defines the architectural and behavioral properties of a decision-gated system that can be validated under deterministic offline replay, without claiming that intervention effectiveness or clinical benefit can be inferred from the same data.

A. SYSTEMS-VALIDATION PERSPECTIVE ON OFFLINE EVALUATION

Most previous studies have evaluated adaptive mental health systems through the predictive accuracy of affect recognition models [2], [3], [4], policy optimization within adaptive intervention frameworks [11], [12], reinforcement learning formulations [13], [14], and simulated improvements in behavioral or clinical outcomes. By contrast, this study adopts a system validation perspective that explicitly distinguishes execution correctness from intervention effectiveness.

Under fully offline and ethically constrained conditions, public emotion-labelled conversational datasets originally constructed for affect recognition rather than intervention decision-making [15] do not permit the causal evaluation of intervention benefits or optimal policy learning [13], [14]. However, such datasets are suitable for examining the system-level properties relevant to safe deployment.

Specifically, offline evaluation enables the assessment of deterministic execution, explicit non-intervention, post-training controllability of intervention frequency, auditability of decision logic independent of language generation, and structural constraints on language model invocation. These properties are treated as architectural guarantees rather than learning outcomes and are validated through deterministic offline replay.

B. SYSTEM OVERVIEW

The architecture explicitly separates the data and control flows. The data flow consisted of multimodal inputs that were transformed into a fixed-dimensional affective state representation. Control flow is governed by the intervention policy, which determines whether the generation model is invoked. This separation ensures that the intervention timing is an explicit and auditable decision independent of language generation.

Fig. 2 provides an overview of the proposed policy-first architecture.

The system operates in a four-stage pipeline: (1) multimodal affective state construction, (2) policy-based intervention decisions, (3) policy-gated language generation, and (4) deterministic offline replay for validation. Deterministic offline replay refers to the sequential execution of logged conversational data without environmental interaction, exploration, or stochastic transition dynamics. At each conversational turn t, multimodal conversational inputs are

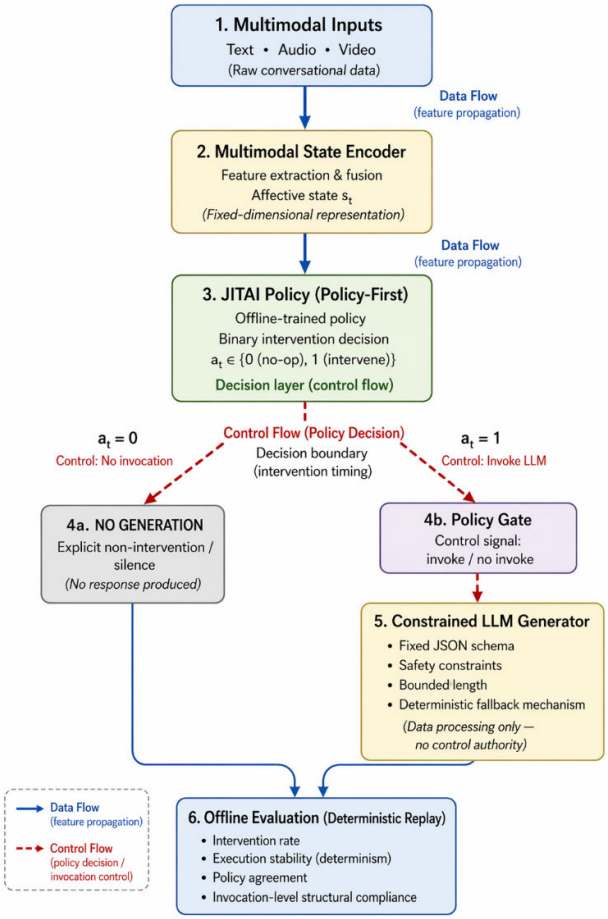

FIGURE 2. **Policy-first architecture for offline evaluation of adaptive conversational AI. Multimodal inputs are encoded into a fixed-dimensional affective state representation, processed by an offline JITAI policy to determine intervention timing, and routed to a policy-gated generation module for constrained response generation under deterministic offline replay, enabling invocation-level structural compliance.**

encoded into a fixed-dimensional affective representation s_t. This representation serves as the sole input to the intervention policy and is not directly exposed to the language model, thereby enforcing a strict separation between perception and generation. The JITAI policy maps each affective state to a binary action $a_t \in \{0, 1\}$, where $a_t = 1$ denotes the intervention and $a_t = 0$ denotes explicit non-intervention.

In the implemented architecture, the intervention policy determines the action before the response is generated. At the architectural level, the generation module is treated as an LLM-compatible conditional generator operating downstream of the policy layer. The generation module is conditioned by the intervention action emitted by the policy layer and the intervention timing is not determined by the generator itself. Thus, intervention timing is mediated by an explicit and auditable decision layer prior to response-generation.

In the offline replay implementation reported in this study, the response-generation component is instantiated as a deterministic structured output stub conditioned on affective valence. This implementation choice reflects the scope of the evaluation: validating policy-gated invocation, schema compliance, bounded response length, deterministic fallback substitution, and explicit non-intervention behavior. However, it did not evaluate the quality of the open-ended LLM.

The generated outputs are constrained by a structured output format, bounded response length, and a deterministic fallback mechanism. Fallback is triggered when the generated output violates predefined structural or rule-based constraints, including: (i) invalid or non-parseable JSON structure, (ii) missing required fields such as ''sentences'' or ''safety,'' (iii) exceeding predefined response length limits, or (iv) inclusion of disallowed content categories. In such cases, the generated output is discarded and replaced with a predefined neutral-fallback response. This fallback mechanism ensures deterministic structural compliance under evaluated offline replay conditions; however, it has not been presented as evidence of content-level clinical safety.

Example fallback response:

```
{
 "sentences": [
  ''I'm here with you.'',
  ''Let's take a moment to pause and breathe.''
 ],
 ''safety": "fallback''
}
```

Example outputs returned by the implemented deterministic generation stub when the policy selects intervention ($a_t = 1$) are shown below. These examples are included only to demonstrate structured output formatting, fallback handling, and policy-gated invocation behavior. They were not presented as clinically validated responses.

```
{
 "sentences": [
  ''I hear you. That sounds difficult.'',
  ''Try taking a few slow breaths to steady yourself.''
 ],
 ''safety": "ok''
}
{
 "sentences": [
  ''That sounds positive.'',
  ''Consider one small step to build on that.''
 ],
 ''safety": "ok''
}
```

C. MULTIMODAL AFFECTIVE STATE REPRESENTATION

The system operates exclusively on the pre-extracted multimodal features provided by the MELD dataset [15]. No end-to-end representation learning was performed using raw text, audio, or video signals. For each utterance, modality-specific embeddings were loaded from the disk and fused to construct a fixed-dimensional affective state representation, which was used as the sole input for the downstream intervention policies.

1) TEXT MODALITY

Textual features consisted of sentence-level contextual embeddings released from the MELD dataset. These embeddings were derived from pretrained transformer-based language models during dataset construction. Each utterance was represented using fixed-dimensional embedding of 768. The text encoder was not fine-tuned or modified in this study, and all textual representations remained fixed throughout training and evaluation.

2) AUDIO MODALITY

Audio features were obtained from the MELD multimodal release and corresponded to fixed-dimensional acoustic embeddings extracted using the dataset's standard preprocessing pipeline. These representations capture the prosodic and paralinguistic characteristics associated with emotional expressions, including pitch, intensity, and spectral attributes. Each audio embedding had a dimensionality of 1582. Neither raw waveform processing nor additional acoustic feature learning were performed.

3) VIDEO MODALITY

Visual features correspond to facial expression embeddings computed using the ArcFace framework [24], which applies additive angular margin loss to produce discriminative facial representations. Each utterance level visual embedding had a dimensionality of 512 and was aggregated from the detected face frames. Utterances without detectable faces were treated as missing-modality instances.

4) FUSION AND PROJECTION

For each utterance at time step t, modality-specific embeddings are concatenated to form a joint multimodal representation

$$x_t = [x_t^{text}; x_t^{audio}; x_t^{video}] \tag{1}$$

with total dimensionality:

$$768 + 1582 + 512 = 2862$$

This concatenated vector is projected into a shared affective state space using linear transformation.

$$s_t = Wx_t + b \tag{2}$$

where

$$W \in \mathbb{R}^{512\times 2862} \text{ and } b \in \mathbb{R}^{512}.$$

The resulting vector $s_t \in R^{512}$ constitutes the fixed-dimensional multimodal affective state used in all the experiments.

Notably, under deterministic offline replay using a synthetic logging policy, the contribution of individual modalities to intervention behavior may be suppressed by dataset-imposed constraints. Multimodal representation is used to verify the correct ingestion, fusion, and projection of heterogeneous inputs into a unified affective state space rather than to demonstrate predictive performance improvements.

D. POLICY-FIRST JITAI DECISION LAYER

Intervention timing is governed by an explicit offline JITAI policy that consumes affective state vectors and outputs a binary action indicating whether or not to intervene at each conversational turn.

The objective of policy learning in this study was system validation rather than intervention optimization. Policy learning was used to verify that multimodal state representations could be reliably consumed, decisions could be executed deterministically, and downstream system components could respond correctly under strict offline constraints.

E. SYNTHETIC LOGGING POLICY

Because public conversational datasets such as MELD do not contain explicit intervention annotations, a synthetic logging policy was introduced solely to construct reproducible offline state–action trajectories for architectural validation rather than to represent real intervention behavior. Let $s_t \in R^{512}$ denote the multimodal affective state at conversational turn t, and let $y_t \in Y$ denote the categorical emotion label associated with the utterance.

The intervention action is defined as:

$$a_t \in \{0, 1\}$$

where $a_t = 1$ indicates intervention and $a_t = 0$ denotes explicit non-intervention.

1) VALENCE GROUPING

Emotion labels are mapped into coarse affective valence categories:

1) Negative valence: anger, sadness, fear, disgust
2) Neutral valence: neutral
3) Positive valence: joy, surprise

This grouping is used solely to define a deterministic logging mechanism and does not represent psychological theory or clinical guidance.

2) DETERMINISTIC LOGGING RULE

The synthetic logging policy π_log maps emotion labels to intervention actions deterministically:

$$\begin{aligned} &\pi\_log(a\_t{=}\,1|y\_t) = p\_neg, if y\_t \in Y\_neg \\ &\pi\_log(a\_t{=}\,1|y\_t) = 0, otherwise \end{aligned} \tag{3}$$

where $Y\_neg = \{\text{anger, sadness, fear, }\}$. In the deterministic logging configuration, $p\_neg = 1.0$. In the logging-policy variation experiment, $p\_neg \in \{1.0, 0.6, 0.5\}$.

And

$$\pi log(a\_t{=}\,0|y\_t) = 1 - \pi log(a\_t{=}\,1|y\_t). \tag{4}$$

The logging policy operates independently at each conversational turn and does not incorporate dialogue history, temporal smoothing, or future context. Under this policy, interventions occur exclusively in utterances exhibiting negative emotional valence, whereas neutral and positive affective states result in explicit non-intervention.

3) PURPOSE AND LIMITATIONS

The synthetic logging policy was solely used to construct reproducible state–action trajectories for validation.

F. BEHAVIORAL CLONING

Behavioral cloning (BC) [25] was included as a supervised imitation baseline and diagnostic probe to evaluate intervention policy learnability under a restricted offline support. BC formulates policy learning as a supervised imitation of a logged behavior policy by estimating a conditional action distribution, as follows:

$$\pi_{\mathrm{BC}}(a \mid s)$$

from the observed state-action pairs (s_t, a_t) generated by the synthetic logging policy.

1) TRAINING OBJECTIVE

The policy network consumes a 512-dimensional affective state representation s_t and outputs logits over the binary intervention action space. The behavioral cloning policy was implemented as a feed-forward multilayer perceptron operating on a 512-dimensional affective state vector s_t. The network consisted of an input layer of size 512, two hidden layers of 256 units with rectified linear unit (ReLU) activations, and a final linear output layer producing logits over the two intervention actions $a_t \in \{0,1\}$. The model was trained using behavioral cloning as a supervised imitation-learning classifier with weighted cross-entropy loss and class weights [1.0, 10.0] for non-intervention and intervention actions, respectively. Training used the Adam optimizer with a learning rate of 1×10^{-4}, batch size 128, and 8 epochs. A 90/10 train-validation split was used, with the validation subset capped at 1000 transitions. Gradients were clipped to a maximum value of 1.0. During deterministic offline replay, the selected intervention action is obtained using the greedy argmax of the output logits, without online exploration, user interaction, or environmental dynamics. All training and evaluations were conducted offline using only the logged data.

Behavioral cloning was used as a diagnostic probe for offline intervention-policy learnability rather than as an intervention-optimization method. Because the logged actions were synthetically derived from emotion labels and the dataset contained no genuine intervention annotations or outcome-linked feedback, behavioral cloning was not interpreted as evidence of clinical policy learning. Instead, the learned policy behavior was analysed to determine whether supervised imitation could produce meaningful intervention decisions under the restricted support conditions imposed by the evaluated dataset.

G. DIAGNOSTIC FAILURE CASE: REWARD-DRIVEN OFFLINE REINFORCEMENT LEARNING AND ITS STRUCTURAL LIMITATIONS

The reinforcement learning component is included solely as a diagnostic probe rather than a performance benchmark. The observed policy behavior reflects the structural limitations of the dataset, including restricted action support and absence of outcome-linked feedback, rather than deficiencies in the learning paradigm. This observation is specific to datasets lacking action diversity and outcome annotations and does not reflect the general capability of offline RL methods.

A minimal policy-gradient model was implemented using a one-step contextual bandit formulation with proxy rewards derived from emotion-valence transitions. Under a deterministic offline replay, the learned policy converges toward a near-unity intervention strategy, selecting the intervention action at nearly every decision point.

This behavior is consistent with known failure modes of reward-driven offline reinforcement learning under restricted action support, logging bias, and absence of outcome-linked rewards [13], [16], [17]. In this dataset, the intervention actions were synthetically derived, the action diversity was limited, and no causal outcome feedback was available. Under these conditions, reward optimization cannot reliably distinguish between alternative policies, leading to degenerate solutions.

These findings reflect a structural mismatch between reward-driven optimization and datasets lacking intervention annotations and outcome signals, rather than a general limitation of reinforcement learning methods.

H. POLICY-CONSTRAINED LANGUAGE GENERATION

The generation module operates strictly as a policy-gated conditional generator and does not participate in intervention decision making.

The generation is constrained by structured JSON output schemas, bounded response length, explicit prohibition on clinical advice, and deterministic fallback responses for malformed outputs.

When the policy selects non-intervention ($a_t = 0$), the generation module is not invoked and no response is produced. This explicit silence pathway distinguishes the policy-first architecture from the unconstrained end-to-end generative baselines that produce a response at every turn.

When the policy selects an intervention ($a_t = 1$), the offline replay implementation invokes the deterministic structured output stub, as described above. Therefore, language-generation evaluation is limited to the invocation level and structural properties: whether generation occurs only when authorized by the policy, whether outputs satisfy the required structured format and bounded response length, and whether malformed outputs are replaced by deterministic fallback responses. The psychological appropriateness, clinical suitability, and content-level safety of the generated responses lie outside the offline identifiability boundary and require human-in-the-loop validation before any real-world deployment.

I. DETERMINISTIC OFFLINE REPLAY PIPELINE

All the system components were evaluated using deterministic offline replay over logged conversational data. Utterances

are processed sequentially, policy decisions are recorded, and the generated responses are stored without delivery to real users.

The framework supports policy ablation, enabling the comparison of learned, heuristic, random, and gated intervention strategies under identical offline conditions. Given fixed inputs and random seeds, the pipeline operates deterministically, ensuring full reproducibility and auditability.

Therefore, the deterministic replay framework is intended to validate execution-level architectural behavior under fixed logged data, rather than conversational effectiveness or real-world intervention outcomes.

J. EVALUATION QUESTIONS

The evaluation addresses the following questions under fully offline conditions:

1) EQ1: Does reward-driven offline reinforcement learning exhibit degenerate behavior under narrow action support?
2) EQ2: Does behavioral cloning produce stable but dataset-imposed intervention behavior under deterministic offline replay?
3) EQ3: Can intervention timing be deterministically regulated post-training through policy gating?
4) EQ4: Can invocation-level structural constraints be enforced through architectural design rather than relying solely on model behavior?

These evaluation questions collectively examine whether fully offline evaluation can be used to assess architectural properties, such as controllability, execution stability, auditability, and safety-related constraints, while remaining insufficient for the causal assessment of intervention effectiveness.

IV. EXPERIMENTAL SETUP

This section describes the experimental configuration used to evaluate the proposed policy-first architecture under the fully offline and ethically constrained conditions.

All the experiments were conducted using deterministic offline replays over prerecorded conversational data. No online interaction, exploration, user exposure, or stochastic environment simulations have been conducted. Accordingly, all results are interpreted strictly as evidence of architectural behavior and methodological limitations rather than as indicators of real-world benefits. The objective of this evaluation is not performance estimation, but the validation of execution properties under controlled offline conditions.

A. DATASET AND OFFLINE REPLAY CONSTRUCTION

All experiments were conducted using the Multimodal EmotionLines Dataset (MELD) [15], a publicly available benchmark consisting of multiparty conversational dialogues annotated with categorical emotion labels.

Although the MELD dataset was derived from scripted television dialogues rather than from clinical interactions, the evaluation in this study focused on architectural execution properties rather than ecological representativeness. The dataset was used exclusively as a source of multimodal sequential data with categorical annotations, and its clinical plausibility lies outside the offline identifiability boundary defined in Section III-A. Accordingly, the dataset supports the validation of system-level execution properties but is not used to infer real-world intervention effectiveness or clinical validity.

Each utterance constituted a discrete decision point. For every conversational turn, a fixed-dimensional multimodal affective state representation of dimension 512 was constructed, as described in Section III-B, and provided as the sole input to all evaluated intervention policies.

The dataset does not contain ground-truth intervention actions, timing annotations, outcome variables, or counterfactual feedback. Accordingly, the MELD dataset scale was not used to evaluate the intervention effectiveness, therapeutic benefits, or optimal decision-making performance. Instead, MELD was treated exclusively as a state distribution dataset for system-level validation under offline constraints.

Because no intervention actions were available, a deterministic synthetic logging policy was used to construct the offline state action trajectories required for the behavioral cloning and diagnostic reinforcement learning experiments.

Consecutive utterances within each dialogue are treated as independent one-step transitions, yielding a contextual bandit-style offline formulation that is commonly used in dynamic treatment regime modeling [26]. No temporal credit assignment, long-horizon return estimation, or modeling of environmental dynamics was incorporated. Therefore, the characteristics of the dataset delimit the scope of the subsequent analyses.

B. EVALUATED POLICIES

All the policies were evaluated under identical offline conditions using the same logged dataset and deterministic replay order.

1) BEHAVIORAL CLONING (BC)

The behavioral cloning policy was implemented as a supervised imitation model trained on state-action pairs generated by a synthetic logging policy [25]. In this study, BC was evaluated as a diagnostic probe for intervention policy learnability under restricted offline support, rather than as an optimized or clinically meaningful policy. Its behavior is interpreted in relation to the synthetic proxy policy and the absence of genuine intervention annotations.

2) PROXY RULE-BASED POLICY

A deterministic heuristic policy that exclusively intervened in negative emotional valence was included as a reference baseline. This policy corresponds precisely to the synthetic logging rule used to construct an offline dataset and enables direct measurement of policy agreements.

3) RANDOM POLICY

A uniform random intervention policy was used as a lower-bound baseline, consistent with standard offline policy evaluation practices [13]. This policy provides a contextual reference for the observed intervention rates, entropy values, and label-conditioned action distributions.

4) DEPLOYMENT TIME GATED POLICY

A deployment-time-gated variant of the behavioral cloning policy was evaluated, in which intervention timing was modified after training using explicit emotion-aware gating rules. No learned parameters were updated. This configuration was used solely to assess the post-training controllability.

5) END-TO-END GENERATIVE BASELINE

Prior LLM-based conversational systems motivate the need to distinguish response generation from explicit intervention control, particularly in safety-sensitive conversational settings [5], [6], [8]. To isolate this architectural issue, an unconstrained end-to-end generative baseline was included as architectural contrast. In this configuration, the response generator is invoked at every conversational turn without an external intervention policy, an explicit non-intervention pathway, or a separate policy-gated invocation mechanism. Therefore, the timing of the intervention was not mediated by the auditable decision layer prior to the response generation. This baseline is not intended to represent all modern LLM orchestration systems, many of which may incorporate guardrails, structured schemas, moderation layers, retrieval modules, and other control mechanisms. Rather, it represents the specific unconstrained configuration in which response generation occurs at every turn and intervention timing is treated as an implicit property of generation.

The purpose of this baseline is not to evaluate generative performance, response quality, or content safety. It was only included to isolate the architectural effect of adding an explicit intervention policy between affective state estimation and response generation.

C. EVALUATION METRICS

The evaluation metrics are deliberately restricted to the observable system behavior under deterministic offline replay. Because the dataset contains no ground truth intervention actions, outcome variables, or counterfactual feedback, reward-based performance measures and off-policy value estimation methods are not applicable and are, therefore, excluded. The reported metrics include the following.

1) Intervention rate: The proportion of decision points at which an intervention action is selected.
2) Action entropy: the entropy of the binary action distribution, reflecting behavioral dispersion.
3) Policy agreement: Proportion of matching actions between learned policies and proxy rule-based baselines.
4) Label conditioned intervention rates: intervention frequency stratified by categorical emotion labels.
5) Invocation-level structural compliance: Proportion of outputs adhering to the required invocation rules and structured response schemas.
6) Deterministic reproducibility: Consistency of action sequences, intervention rates, and invocation outcomes across repeated executions with fixed data and random seeds.

Collectively, these metrics characterize the execution behavior, controllability, and architectural safety properties under offline constraints.

D. ROBUSTNESS AND REPRODUCIBILITY ASSESSMENT

Because the proposed evaluation was conducted under deterministic offline replay rather than stochastic online interaction, robustness was assessed at the level of execution reproducibility. Specifically, the assessment examined whether repeated executions under fixed logged data D, fixed replay configurations, fixed policy parameters, and fixed random seeds produced identical action sequences, intervention rates, and invocation control behaviors.

Three execution-level robustness indicators were used. First, the intervention rate variance was computed across repeated runs to determine whether aggregate intervention frequency remained stable. Second, the unauthorized invocation variance was computed across repeated runs to determine whether the generation module was invoked when the selected action was non-intervention. Third, the action-sequence mismatch rate was computed relative to a reference replay trace. The action-sequence mismatch rate is defined as the proportion of decision points at which the action selected in a repeated run differs from the corresponding action in the reference trace. A value of zero indicates the exact replay stability.

In addition to the fixed-seed replay, robustness was examined under controlled synthetic logging policy variations. The negative intervention probability in the synthetic logging policy varied across $p\_neg \in \{1.0, 0.6, 0.5\}$. These variations were introduced to test whether the replay framework preserved the deterministic execution and invocation control consistency under different logged policy configurations. They were not intended to simulate real clinical intervention behaviors or establish real-world generalizations.

Accordingly, the robustness assessment was restricted to the offline execution stability, reproducibility, and invocation control consistency. It does not measure clinical robustness, psychological appropriateness, content-level safety, or real-world intervention effectiveness, all of which remain outside the offline identifiability boundary and require a human-in-the-loop evaluation with outcome-linked data.

V. RESULTS

The results are presented from a system validation perspective, focusing on properties that are identifiable under

TABLE 2. Policy behavior under synthetic logging policy variation (negative-intervention probability).

Negative Intervention Probability	Final Greedy Action Distribution
1.0	100% intervention
0.6	100% intervention
0.5	100% intervention

deterministic offline replay, rather than on predictive performance or policy optimality. The objective was to verify execution correctness, controllability, and architectural constraints within the offline identifiability boundary.

All results were obtained under identical deterministic replay conditions across 11,351 conversational decision points in the MELD dataset. The reported metrics represent the exact execution characteristics of the system under fixed data support and are interpreted as evidence of framework validity within the defined offline constraints.

A. REPRESENTATION PIPELINE SANITY CHECK

All utterances were successfully mapped to valid multimodal affective state vectors, yielding a 100% state coverage across the dataset. The resulting representations were consistent in dimensionality (512), and exhibited no numerical instability during feature loading, modality fusion, or linear projection.

These results confirm that the representation pipeline operates reliably under deterministic offline replay and that subsequent findings reflect decision-layer behavior rather than preprocessing artifacts.

B. EVALUATION QUESTION 1: BEHAVIOR OF REWARD-DRIVEN OFFLINE REINFORCEMENT LEARNING

Across all configurations, the learned policy converges to a near-unity intervention strategy, selecting an intervention action $a_t = 1$ at nearly every decision point s_t, as summarized in Table 2.

This behavior is consistent with the known failure modes of reward-driven offline reinforcement learning under restricted action support and the absence of outcome-linked feedback. Under these conditions, policy optimization over D cannot reliably distinguish between alternative actions and converges to a constant policy.

Accordingly, this result is interpreted as a diagnostic confirmation of a structural limitation induced by the dataset, rather than an evaluation of the reinforcement learning performance. The observed behavior reflects dataset-imposed constraints under a deterministic offline replay.

C. EVALUATION QUESTION 2: BEHAVIORAL CLONING STABILITY AND DATASET-IMPOSED BIAS

The learned behavioral cloning policy produced an intervention rate of 1.0 under deterministic replay, indicating convergence toward a near-unity intervention strategy. This behavior differs substantially from the synthetic logging policy intervention distribution and reflects the absence of genuine intervention supervision within the underlying emotion-labelled dataset.

TABLE 3. Policy-first system execution metrics under offline replay.

Metric	Observed Value
Decision points	11,351
Interventions	2,609
Silence (non-intervention)	8,742
Intervention rate	0.23
Unauthorized generation invocations	0

TABLE 4. aggregate intervention behavior under deterministic offline replay.

Policy	Intervention Rate	Action Entropy	Agreement With Synthetic Proxy	Unauthorized Invocation Violations
Behavioral Cloning (BC)	1.00	0.00	0.23	0
Proxy Rule-Based Policy	0.23	0.78	-	0
Random Policy	0.50	1.00	0.50	0

The behavioral cloning experiment functions as a diagnostic probe for offline intervention-policy learnability, rather than as a predictive benchmark. Under deterministic replay, the learned policy converges to near-unity intervention behavior, producing intervention actions at nearly all conversational decision points. This behavior indicates that the emotion-labelled conversational dataset does not provide sufficient intervention supervision to support meaningful intervention-policy imitation under the evaluated conditions.

Accordingly, the behavioral cloning results should not be interpreted as evidence of policy optimality or clinically meaningful intervention behavior. Rather, it illustrates the tendency of supervised policy imitation to collapse toward degenerate intervention strategies when genuine intervention annotations and outcome-linked feedback are absent.

The observed intervention rate (0.23) reflected the intervention distribution induced by the synthetic logging policy. No unauthorized generation module invocations were observed when the policy selected no intervention, thereby confirming that the decision layer fully mediated language generation. Table 3 presents the execution-level properties under the deterministic proxy replay configuration used for architectural validation. Table 4 compares the learned and baseline policies for the same offline replay setting.

1) REPRESENTATION ABLATION ANALYSIS

Table 5 presents the results of the representation ablation analysis under deterministic offline replay.

TABLE 5. Representation ablation under deterministic offline replay.

Input Representation	Intervention Rate	Agreement with Synthetic Proxy	Action Entropy (bits)
Text Only	1.00	0.23	0.00
Text + Audio	1.00	0.23	0.00
Full Multimodal	1.00	0.23	0.00

TABLE 6. Emotion-conditioned intervention rates under deterministic offline replay.

Emotion	Proxy Rule	BC	Random
angry	1.00	1.00	0.495
sad	1.00	1.00	0.488
neutral	0.00	1.00	0.498
happy	0.00	1.00	0.506

The invariance across representation variants reflects dataset-imposed degeneracy under deterministic replay rather than meaningful multimodal policy learning. Across all the evaluated input configurations, the learned behavioral cloning policy converged toward near-unity intervention behavior with identical intervention rates and near-zero action entropy.

The observed invariance was likely attributable to the MELD-derived replay construction and label-driven intervention supervision used in this study. Because the logged intervention actions were deterministically derived from emotion labels, the textual representation captured much of the semantic and affective information required to reproduce the logged decision pattern. Consequently, the addition of audio and visual modalities did not materially alter the learned intervention behavior under the evaluated offline replay conditions.

This result indicates that the underlying emotion-labelled dataset does not provide sufficient intervention supervision for representation-sensitive intervention policy imitation. Accordingly, representation ablation analysis should be interpreted as evidence of representation-invariant policy collapse under limited intervention support rather than evidence of multimodal intervention-learning effectiveness.

D. EMOTION-CONDITIONED INTERVENTION BEHAVIOR

To further characterize decision-layer behavior beyond aggregate intervention rates, intervention frequency was stratified by categorical emotion labels. Table 6 reports the label-conditioned intervention rates under deterministic offline replay.

Behavioral cloning exhibited near-unity intervention behavior across all evaluated emotional conditions, intervening not only under negative emotional states, but also under neutral and positive conversational states. This behavior differs substantially from the synthetic logging policy and reflects the absence of genuine intervention supervision within the underlying emotion-labelled dataset. By contrast, the random baseline exhibited approximately uniform intervention probabilities across emotional categories, confirming the absence of state-dependent control behavior.

E. EVALUATION QUESTION 3: DEPLOYMENT-TIME CONTROL VIA POLICY GATING

Applying explicit emotion-aware gating rules at deployment enables deterministic control of intervention timing, independent of π_log, without retraining or modifying the learned policy parameters. The intervention frequency and contextual sensitivity change predictably according to the specified gating logic, whereas the underlying policy distribution remains unchanged.

This demonstrates that intervention timing can be independently regulated through architectural separation of decision control and policy learning.

F. EVALUATION QUESTION 4: ARCHITECTURAL IMPLEMENTATION OF LANGUAGE MODEL INVOCATION-LEVEL STRUCTURAL CONSTRAINTS

Across all evaluated configurations, the generated outputs conformed to the required structured JSON schema under policy gating. Language generation is triggered exclusively when the intervention policy selects an action $a_t = 1$ and no output is produced under explicit non-intervention conditions. Malformed generations were replaced using deterministic fallback mechanisms, and no unauthorized generation module invocations were observed.

These results confirm that invocation-level structural constraints are correctly enforced at the architectural level under a deterministic offline replay.

However, structural compliance does not constitute content-level safety validation. The appropriateness, tone, and potential psychological risks of the generated responses were not evaluated in this study and require future human-in-the-loop assessments prior to deployment.

G. DETERMINISTIC EXECUTION CONSISTENCY CHECK

A consistency check was conducted to assess the reproducibility of system execution under identical data and policy configurations across all 11,351 conversational decision points. The observed intervention rate remained stable at 0.23, and no language model invocations occurred during the non-intervention decisions. The action distributions and safety-related outcomes were identical across repeated executions.

These results confirm that the observed controllability and safety properties arise from architectural design rather than stochastic variation. Robustness is evaluated through reproducibility under deterministic replay and invariance across controlled variations in the synthetic logging policy, rather than through stochastic performance metrics.

TABLE 7. Execution robustness under deterministic offline replay.

CONDITION	Repeated runs	Mean intervention rate	Intervention-rate variance	Unauthorized invocation variance	Action-sequence mismatch rate
Fixed replay, seed 1	3	0.2298	0.0000	0.0000	0.0000
Fixed replay, seed 2	3	0.2298	0.0000	0.0000	0.0000
Fixed replay, seed 3	3	0.2298	0.0000	0.0000	0.0000
Synthetic logging probability = 1.0	3	0.2298	0.0000	0.0000	0.0000
Synthetic logging probability = 0.6	3	0.1361	0.0000	0.0000	0.0000
Synthetic logging probability = 0.5	3	0.1144	0.0000	0.0000	0.0000

H. EXECUTION ROBUSTNESS UNDER DETERMINISTIC REPLAY

The robustness analysis reported in Table 7 was conducted under the deterministic proxy replay configuration used for architectural validation. Under fixed logged data, fixed replay configurations, and fixed random seeds, repeated deterministic replay produced identical action sequences, intervention rate values, and invocation control outcomes across all evaluated runs. The intervention rate variance, unauthorized invocation variance, and action-sequence mismatch rate were all 0.0000, indicating consistent behavior of the offline replay pipeline under identical experimental conditions.

The synthetic logging-policy variation results further show that intervention rates change predictably as the negative-intervention probability varies, while replay stability and invocation-control consistency are preserved. These robustness metrics are defined at the level of deterministic system execution. They do not measure clinical robustness, psychological appropriateness, or real-world generalization, which remains outside the offline identifiability boundary.

I. COMPARISON WITH AN UNCONSTRAINED END-TO-END GENERATIVE BASELINE

This comparison highlights differences in architectural execution properties rather than differences in response quality or task performance. Under a deterministic offline replay, an unconstrained end-to-end generative baseline produces responses at every conversational decision point under offline replay.

In this configuration, no explicit non-intervention pathway is defined, and the intervention timing emerges from the model behavior rather than from an external decision layer. For offline replay, the intervention rate was operationalized at the invocation level. Because the unconstrained baseline invokes the response generator at every replayed conversational decision point and provides no explicit silent action, each decision point is counted as an intervention, yielding an intervention rate of 1.00.

TABLE 8. Architectural comparison between the proposed policy-first framework and an unconstrained end-to-end baseline.

Property	Policy-First Architecture	Unconstrained End-to-End Baseline
Intervention timing	Explicit policy	Implicit
Intervention rate	~0.23	1.00
Silence possible	✓	✗
Post training control	✓	✗
Output boundedness enforced	✓	✗
Invocation-level structural constraints	Architectural	Behavioral
Offline auditability	✓	✗

The comparison in Table 8 should, therefore, be interpreted as an architectural contrast, not as a performance benchmark, or as a claim about all modern LLM orchestration systems.

Therefore, this comparison focuses on execution-level architectural properties that remain observable under deterministic offline replays, including explicit intervention gating, silence-pathway support, invocation-level constraint enforcement, and offline auditability. It does not evaluate conversational quality, therapeutic effectiveness, or human preferences.

J. SUMMARY OF RESULTS

Overall evaluation outcomes across all configurations are summarized in Table 9.

K. FINAL INTERPRETATION

Collectively, the results show that the evaluated offline setting is suitable for validating execution-level architectural

TABLE 9. Summary of evaluation outcomes under deterministic offline replay.

EQ	Focus	Observed Outcome	Interpretation
EQ1	Reward-driven offline RL behavior	Convergence to a constant near-unity intervention policy across all runs	Structural limitation under absent outcome-linked feedback
EQ2	Behavioral cloning behavior	Convergence to a near-unity intervention policy with intervention rate 1.00 and agreement with proxy of 0.23	Indicates limited intervention-policy learnability under emotion-labelled data lacking genuine intervention annotations
EQ3	Deployment-time controllability	Deterministic regulation of intervention timing achieved without retraining	Intervention timing controllable independently of learned policy parameters
EQ4	Invocation-level structural constraints	Full schema compliance; zero unauthorized invocations	Invocation-level structural constraints enforced architecturally; content-level safety not evaluated
Consistency Check	Deterministic execution consistency	Stable proxy replay intervention rate of 0.23; zero unauthorized invocations under repeated deterministic replay	Core architectural execution guarantees preserved under fixed replay conditions

properties but not for estimating intervention effectiveness or learning outcome-optimal intervention policies. Reward-driven optimization converged to near-unity intervention behavior under limited action support and absent outcome-linked feedback. Behavioral cloning similarly converged toward a near-unity intervention strategy, differing substantially from synthetic logging policy distribution. These outcomes should therefore be interpreted as diagnostic evidence of limited intervention-policy learnability under emotion-labelled offline data, rather than as evidence of policy optimality or real-world effectiveness.

The results operationalize the offline identifiability boundary defined in this study. Properties such as execution correctness, intervention controllability, auditability, and invocation-level structural compliance are directly observable from fixed-logged dataset D under deterministic replay. By contrast, properties that require outcome-linked feedback, including causal intervention effectiveness, personalization quality, and optimal policy learning, remain non-identifiable under the evaluated conditions.

L. STATISTICAL INTERPRETATION AND REPRODUCIBILITY

All the reported results were obtained under deterministic offline replay using fixed data, replay order, and random seeds. Therefore, the reported metrics represent the exact execution characteristics under the evaluated replay conditions rather than stochastic performance estimates. Reproducibility was assessed using the intervention rate variance, unauthorized invocation variance, and action-sequence mismatch rate across repeated replay runs, as shown in Table 7.

In the absence of stochastic environmental interactions, repeated executions produce identical policy behaviors. Robustness is operationalized through deterministic reproducibility, defined as identical outputs across repeated executions under fixed data and random seeds, together with invariant intervention rates and action distributions across controlled variations in the synthetic logging policy π_log, as reported in Table 7.

VI. DISCUSSION

This study demonstrates that offline evaluation can be useful when its purpose is restricted to system validation. Under the deterministic replay of fixed-logged data D, the policy-first architecture enables the verification of execution-level properties, including intervention timing control, explicit non-intervention, policy-gated invocation, and structural compliance of generated outputs. These properties are observable because they depend on the executed system behavior rather than on unobserved user outcomes.

The offline identifiability boundary provides a method to separate the two types of claims that are often blurred in adaptive conversational AI. The first type concerns execution-identifiable properties such as whether the system invokes the generation module only when authorized by the intervention policy. The second type concerns outcome-dependent properties such as whether an intervention improves the user's state, engagement, or clinical well-being. The present framework supports the first claim but not the second.

This distinction extends beyond emotion-aware conversational artificial intelligence (AI). Any decision-gated AI system evaluated under offline constraints faces a similar separation between properties that can be verified from logged execution and those that require interaction, counterfactual support, or outcome-linked feedback. In this sense, the offline identifiability boundary is not only a limitation statement but also a design principle; it helps determine which claims can be responsibly made before human-facing deployment.

These findings highlight the role of the architecture in safety-critical systems. By placing the intervention timing within an explicit policy layer $\pi(a_t \mid s_t)$, the proposed design prevents the response generator from implicitly determining when to intervene. This separation improves auditability, supports explicit silence when $a_t = 0$, and enables deterministic validation of invocation control under offline replay conditions. However, these architectural guarantees do not establish content-level safety, clinical appropriateness, or real-world effectiveness, which requires human-in-the-loop evaluations and outcome-linked data.

VII. LIMITATIONS AND THREATS TO VALIDITY

The primary limitation of this study arises from the use of a public emotion-labelled conversational dataset that is not designed for adaptive intervention research. The dataset did not contain ground-truth intervention actions, timing annotations, outcome variables, or longitudinal feedback. Consequently, all intervention actions and reward signals were synthetically derived, precluding causal inference, outcome evaluation, and assessment of real-world effectiveness. Beyond generalizability, the use of scripted entertainment dialogue introduces a threat to construct validity, and the emotional expressions and interaction dynamics present in MELD may not correspond to the constructs of genuine psychological distress or the therapeutic need that the proposed system is intended to address, implying that the validated system properties may not reflect behavior in the target construct domain. This threat to construct validity is acknowledged as a structural constraint of the evaluation and does not affect the validity of architectural and execution-level properties demonstrated under deterministic offline replay.

The use of a deterministic synthetic logging policy introduces strong action imbalance and limited action diversity. Under these conditions, behavioral cloning converges toward a dataset-imposed degenerate intervention strategy, whereas reward-driven optimization collapses toward near-unity intervention behavior. These behaviors reflect the structural properties of the synthetic data-generating process and are consistent with the known limitations of offline reinforcement learning under restricted action support and in the absence of outcome feedback [13], [16], [17].

The intervention space is intentionally simplified to a binary decision (intervention vs. non-intervention), isolating intervention timing as a controllable system variable. While this enables focused validation of architectural properties, it reduces realism and does not capture the multidimensional nature of real-world mental health interventions. Similarly, reward signals are defined using coarse emotion-valence transitions and do not correspond to clinically meaningful outcomes.

Although the architecture supports multimodal inputs, the effective representations are dominated by textual features owing to dataset characteristics, including limited visual consistency and weak alignment between modalities. Consequently, the observed modality effects reflect dataset constraints rather than the intrinsic importance of different modalities in real-world settings.

All evaluations were conducted under a deterministic offline replay using structural and behavioral metrics. No clinician review, user study, or longitudinal evaluation was performed. Therefore, the findings should be interpreted strictly as evidence of architectural and execution-level validation under offline constraints, rather than as indicators of clinical utility, content safety, or deployable system performance.

VIII. FUTURE WORK

Future research should focus on developing datasets with genuine intervention annotations, outcome-linked feedback, and longitudinal structures to enable causal evaluation of adaptive conversational systems. Expanding the intervention space to include multilevel, personalized, and temporally aware strategies is essential.

In addition, a human-in-the-loop evaluation, including a clinician's review of intervention appropriateness, tone, and safety, is required for real-world validations. Alignment with established evaluation frameworks such as micro-randomized trials (MRTs) [27] and sequential multiple assignment randomized trials (SMART) [12] is necessary to assess effectiveness in practical deployment settings.

IX. CONCLUSION

This study proposed a policy-first architecture for the offline validation of emotion-aware conversational AI under identifiability constraints. The central idea is to separate intervention decision making from language generation so that intervention timing becomes an explicit, auditable, and replay-testable system property rather than the emergent behavior of a generative model.

The main contribution is the offline identifiability boundary, which distinguishes execution-level properties that can be validated using fixed-logged data from outcome-dependent properties that cannot be identified without counterfactual action support or outcome-linked feedback. Within this boundary, the proposed deterministic replay framework validates controllability, execution stability, auditability, explicit non-intervention, and invocation-level structural compliance.

The results support the use of offline evaluation as a pre-deployment system validation step and not as evidence of clinical effectiveness or optimal intervention learning. Future studies should extend this framework using datasets with genuine intervention actions, richer action support, longitudinal outcomes, and ethically approved human evaluations.

APPENDIX

IMPLEMENTATION DETAILS

Implementation details, including the model architecture, training configuration, and deterministic offline replay pipeline, are available in the corresponding repository. The full experimental pipeline, including policy training, evaluation, and result generation, can be reproduced using scripts and instructions provided in the repository. Data and Code Availability: The codebase and experimental resources supporting this study are publicly available at https://doi.org/10.5281/zenodo.20158330 [28]. The MELD dataset is publicly available at https://affective-meld.github.io/ and is not redistributed due to usage restrictions.

ACKNOWLEDGMENT

This article is being submitted under Dublin City University's open access publishing agreement, subject to publisher eligibility and approval.

REFERENCES

[1] R. W. Picard, *Affective Computing*. Cambridge, MA, USA: MIT Press, 1997, doi: 10.7551/mitpress/1140.001.0001.

[2] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009, doi: 10.1109/TPAMI.2008.52.

[3] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, and R. Zimmermann, "Conversational memory network for emotion recognition in dyadic dialogue videos," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol.*, 2018, pp. 2122–2132, doi: 10.18653/v1/n18-1193.

[4] S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh, and A. Hussain, "Multimodal sentiment analysis: Addressing key issues and setting up the baselines," *IEEE Intell. Syst.*, vol. 33, no. 6, pp. 17–25, Nov. 2018, doi: 10.1109/MIS.2018.2882362.

[5] J. Wei et al., "Emergent abilities of large language models," *Trans. Mach. Learn. Res.*, 2022. Accessed: Jan. 25, 2026. [Online]. Available: https://openreview.net/forum?id=yzkSU5zdwD

[6] T. B. Brown et al., "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates, 2020, pp. 1871–1877. [Online]. Available: https://proceedings.neurips.cc/pa

[7] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, J. Teevan, R. Kikin-Gil, and E. Horvitz, "Guidelines for human-AI interaction," in *Proc. CHI Conf. Human Factors Comput. Syst.* Glasgow, U.K.: ACM, May 2019, pp. 1–13, doi: 10.1145/3290605.3300233.

[8] L. Weidinger et al., "Taxonomy of risks posed by language models," in *Proc. ACM Conf. Fairness Accountability Transparency*, Jun. 2022, pp. 214–229, doi: 10.1145/3531146.3533088.

[9] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in AI safety," 2016, *ArXiv:1606.06565*.

[10] S. Roller, E. Dinan, N. Goyal, D. Ju, M. Williamson, Y. Liu, J. Xu, M. Ott, E. M. Smith, Y.-L. Boureau, and J. Weston, "Recipes for building an open-domain chatbot," in *Proc. 16th Conf. Eur. Chapter Assoc. Comput. Linguistics: Main*, 2021, pp. 300–325, doi: 10.18653/v1/2021.eacl-main.24.

[11] I. Nahum-Shani, S. N. Smith, B. J. Spring, L. M. Collins, K. Witkiewitz, A. Tewari, and S. A. Murphy, "Just-in-time adaptive interventions (JITAIs) in mobile health: Key components and design principles for ongoing health behavior support," *Ann. Behav. Med.*, vol. 52, no. 6, pp. 446–462, May 2018, doi: 10.1007/s12160-016-9830-8.

[12] S. A. Murphy, "An experimental design for the development of adaptive treatment strategies," *Statist. Med.*, vol. 24, no. 10, pp. 1455–1481, May 2005, doi: 10.1002/sim.2022.

[13] S. Levine, A. Kumar, G. Tucker, and J. Fu, "Offline reinforcement learning: Tutorial, review, and perspectives on open problems," *J. Mach. Learn. Res.*, vol. 22, no. 279, pp. 1–68, 2021.

[14] O. Gottesman, F. Johansson, M. Komorowski, A. Faisal, D. Sontag, F. Doshi-Velez, and L. A. Celi, "Guidelines for reinforcement learning in healthcare," *Nature Med.*, vol. 25, no. 1, pp. 16–18, Jan. 2019, doi: 10.1038/s41591-018-0310-5.

[15] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*. Italy: Association for Computational Linguistics, 2019, pp. 527–536, doi: 10.18653/v1/p19-1050.

[16] K. Ghasemipour, S.S. Gu, and O. Nachum, "Why so pessimistic? Estimating uncertainties for offline RL through ensembles, and why their independence matters," in *Proc. Adv. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, Nov. 2022, pp. 18267–18281.

[17] A. Kumar, J. Fu, G. Tucker, and S. Levine, "Stabilizing off-policy Q-learning via bootstrapping error reduction," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 11784–11794. Accessed: Jan. 25, 2026. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/c2073ffa77b5357a498057413bb09d3a-Abstract.html

[18] J. Pearl, *Causality: Models, Reasoning, and Inference*. Cambridge, U.K.: Cambridge Univ. Press, 2009, doi: 10.1017/CBO9780511803161.

[19] D. B. Rubin, "Causal inference using potential outcomes: Design, modeling, decisions," *J. Amer. Stat. Assoc.*, vol. 100, no. 469, pp. 322–331, Mar. 2005, doi: 10.1198/016214504000001880.

[20] J. Bi, Z. Wang, H. Yuan, X. Shi, Z. Wang, J. Zhang, M. Zhou, and R. Buyya, "Large AI models and their applications: Classification, limitations, and potential solutions," *Software, Pract. Exper.*, vol. 55, no. 6, pp. 1003–1017, Jun. 2025, doi: 10.1002/spe.3408.

[21] T. Rousmaniere, S. B. Goldberg, and J. Torous, "Large language models as mental health providers," *Lancet Psychiatry*, vol. 13, no. 1, pp. 7–9, Jan. 2026, doi: 10.1016/s2215-0366(25)00269-x.

[22] H. Lalwani and H. Salam, "The supportiveness–safety tradeoff in LLM well-being agents," in *Companion Proc. 21st ACM/IEEE Int. Conf. Human-Robot Interact.*, Mar. 2026, pp. 1089–1093, doi: 10.1145/3776734.3794563.

[23] I. Steenstra, P. Pedrelli, W. Shi, S. Marsella, and T. W. Bickmore, "Assessing risks of large language models in mental health support: A framework for automated clinical AI red teaming," 2026, *arXiv:2602.19948*.

[24] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 5962–5979, Oct. 2022, doi: 10.1109/TPAMI.2021.3087709.

[25] S. Ross, G. J. Gordon, and J. A. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *Proc. 14th Int. Conf. Artif. Intell. Statist. (AISTATS)*, Fort Lauderdale, FL, USA, 2011, pp. 627–635. Accessed: Dec. 31, 2025. [Online]. Available: https://proceedings.mlr.press/v15/ross11a.html

[26] S. A. Murphy, M. J. van der Laan, and J. M. Robins, "Marginal mean models for dynamic regimes," *J. Amer. Stat. Assoc.*, vol. 96, no. 456, pp. 1410–1423, Dec. 2001, doi: 10.1198/016214501753382327.

[27] P. Klasnja, E. B. Hekler, S. Shiffman, A. Boruvka, D. Almirall, A. Tewari, and S. A. Murphy, "Microrandomized trials: An experimental design for developing just-in-time adaptive interventions.," *Health Psychol.*, vol. 34, no. Suppl, pp. 1220–1228, Dec. 2015, doi: 10.1037/hea0000305.

[28] Kharisara, "Kharisara/MER-JIT-LLM," *IEEE Access*, 2026, doi: 10.5281/ZENODO.20158330.

KESARA HARISARA PUNCHIHEWA is currently pursuing the B.Eng. degree (Hons.) in software engineering with the School of Computing, IIT, Colombo, Sri Lanka. His academic and research interests include conversational artificial intelligence, emotion-aware systems, and the evaluation and assurance of data-driven decision-making models. His current work focuses on trustworthy AI, offline evaluation methodologies, and safety-aware conversational systems.

MANINDA EDIRISOORIYA received the B.Sc. (Eng.) degree (Hons.) in computer science and engineering and the M.B.A. degree in management of technology from the University of Moratuwa, Sri Lanka, in 2011 and 2019, respectively.

He is currently a Senior Lecturer with the School of Computing, Informatics Institute of Technology (IIT), Sri Lanka. He is also a Visiting Lecturer with the University of Moratuwa. Prior to joining academia full-time, he spent more than a decade in the software industry, including engineering and technical leadership roles at WSO2, where he contributed to enterprise integration platforms, large-scale distributed systems, artificial intelligence solutions, and developer tooling. His research interests include artificial intelligence, machine learning, deep learning, generative AI, agentic AI systems, natural language processing, conversational AI, and enterprise AI applications. His work spans both academic research and industrial deployment of AI-driven solutions, including intelligent automation, sentiment analysis, and large-scale data analytics.

Mr. Edirisooriya is an Associate Member of the Institute of Engineers Sri Lanka (IESL) and a Corporate Life Member of the Radio Society of Sri Lanka (RSSL).

RAJKUMAR BUYYA (Fellow, IEEE) is currently a Redmond Barry Distinguished Professor and the Director of the Cloud Computing and Distributed Systems (CLOUDS) Laboratory, The University of Melbourne, Australia. He is the author or co-author of more than 800 publications, including seven textbooks. He is among the most cited authors in the fields of computer science and software engineering worldwide, with an H-index of 181, G-index of 394, and more than 170 000 citations as of 2026. His research interests include cloud computing, distributed systems, and service-oriented computing.

Prof. Buyya is a fellow of ACM. He was a recipient of the 2008 Distinguished Service Award from the IEEE Computer Society. He served as the Founding Editor-in-Chief for IEEE TRANSACTIONS ON CLOUD COMPUTING.

LAKSHIKA CHANDRADEVA received the B.Eng. degree in software engineering and the M.Sc. degree in cyber security and forensics from the University of Westminster.

She is a Ph.D. Researcher in computer applications with Dublin City University (DCU) with more than nine years of combined academic and industry experience in information security, cybersecurity governance, and health informatics. Her doctoral research focuses on standards-based frameworks for healthcare delivery organizations integrating medical devices with health information technology systems, addressing security, risk management, and regulatory compliance within the EU digital health ecosystem. Academically, she has a strong foundation in cybersecurity and applied machine learning. She actively contributes to national and international standardization efforts as a National Committee Member of the National Standards Authority of Ireland (NSAI). Professionally, she has held senior roles in information security, governance, and infrastructure consulting, most recently as a Senior Consultant with Virtusa and as a Senior Analyst with LOLC Technology Services. Her experience spans ISMS implementation, risk management, cloud security, audits, regulatory compliance, and enterprise security architecture, with hands-on leadership in ISO 27001, ISO 20000, PCI-DSS, and large-scale security operations. She is an established researcher with peer-reviewed publications and conference papers in areas, such as machine-learning-based fraud detection, phishing awareness, and cybersecurity analytics, presented at national and international conferences, including IEEE-indexed and ICICT venues. Her work reflects a strong commitment to translate advanced security research and standards into practical and secure digital healthcare solutions. Her research interests lie at the intersection of healthcare cybersecurity, medical device integration, health informatics standards (ISO 80001, ISO 81001, ISO/TC 215, CEN/TC 251), and EU regulatory frameworks, including GDPR, MDR, IVDR, and NIS2.

ACHALA APONSO (Member, IEEE) received the B.Sc. degree (Hons.) in software engineering and the M.Sc. degree in artificial intelligence and brings over a decade of combined academic and industrial experience from across Australia and overseas. He is currently pursuing the Ph.D. degree with Edith Cowan University, Australia.

He is a Sessional Academic with Curtin University. His doctoral research integrates machine and deep learning with psychological science, with an emphasis on behavior modeling, adaptive decision systems, system-level controllability, and responsible AI deployment. He served as the Senior Author and a Principal Supervisor of this work, leading to the conceptualization of the policy-first architecture, experimental design, and system validation framework. He has supervised undergraduate and postgraduate research projects in machine learning and data analytics, contributing to peer-reviewed publications, and applying AI system development. His work has been published in international peer-reviewed venues. His research focuses on multimodal affective computing, decision-centric artificial intelligence architectures, and safe integration of large language models in high-stakes domains, including digital mental health.

Mr. Aponso has received research recognition, including the President's Award for Scientific Research Publication.