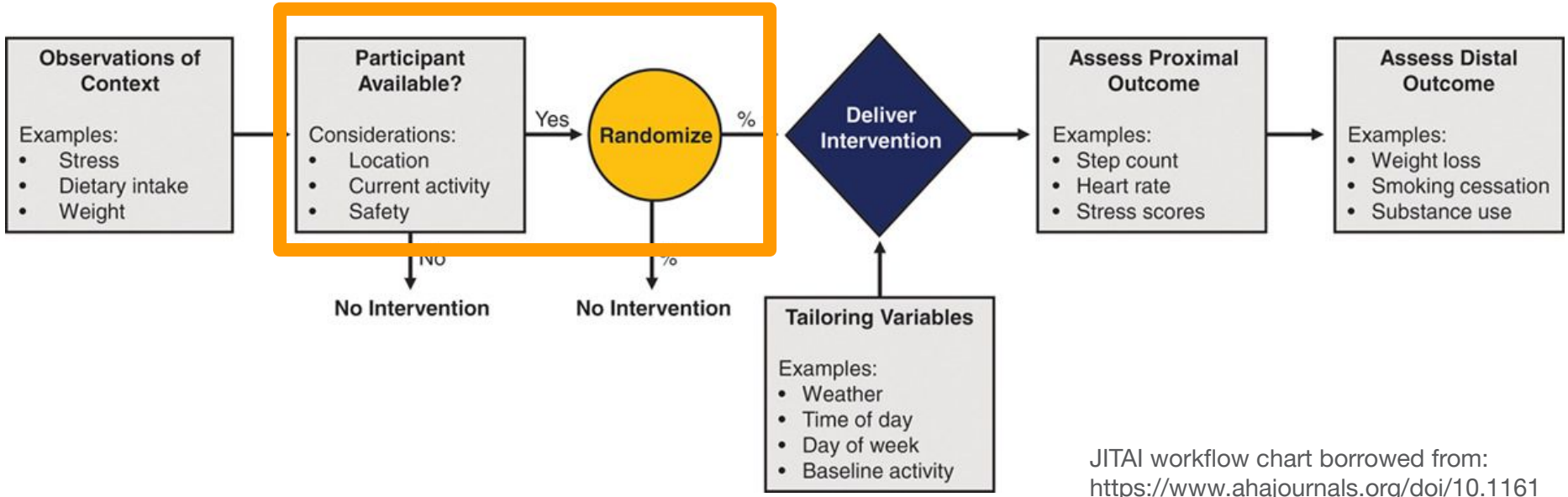
UCLA Samueli School of Engineering

Exploring LLM Reasoning in Just-in-time Adaptive Intervention

Tianyi Li

Motivation and Objectives

Just-in-time adaptive interventions use contextual information from mobile devices (location, calendar, etc) to determine when to provide behavioral interventions to individuals.



JITAI workflow chart borrowed from: https://www.ahajournals.org/doi/10.1161/CIRCOUTCOMES.120.006760

Technical Approach and Novelty

Current Landscape: How to make intervention more “Ontime”

1. **Smarter timing and decision-point scheduling**
2. **ML-driven JITAIs with refined/specialized algorithms for different application**
   a. Mental health & well-being
   b. Physical activity & diet / cardiometabolic risk
   c. General behavior change

Pattern:
Conventional JITAIs rely on predefined rules/functions/algorithms/models to map context to intervention.

Limitation:
Fixed utility; Hard to model complex multi-factor context. -> But LLM is good at reasoning

Technical Approach and Novelty

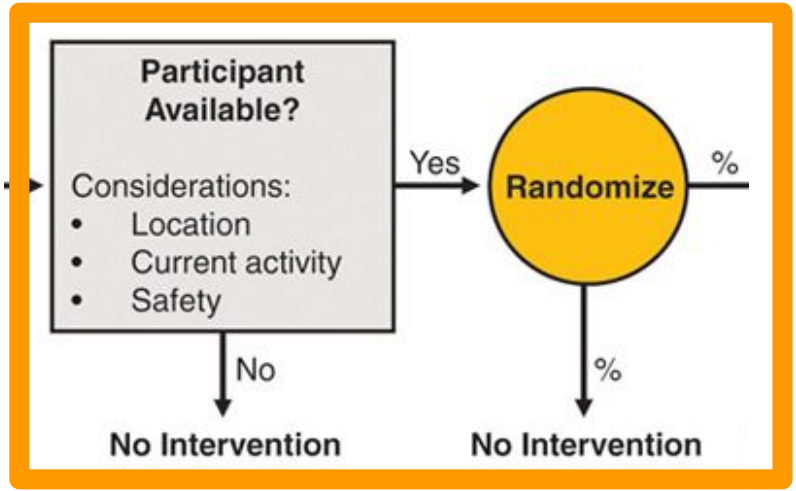
- **Haag et al. (2025) = proof-of-concept that LLMs can act as the "JITAI brain."**

- **No perception - Given as text input:**
- The LLM **reads:**
  a. a rich patient persona (cardiac rehab patient: medical status, lifestyle, preferences).
  b. a detailed momentary context (time of day, symptoms, mood, recent activity, barriers right now).
- **Also with text output:**
- LLM **generates:** a tailored, clinician-style message. -> **Rate by human to be much better than human generated message**

There is limited literature on working, end-to-end systems that truly integrate LLMs into JITAIs.

My Goal

Motivation and Objectives - cont.



Bridging the gap and Integrate LLM reasoning directly into the JITAI decision layer

Goal+Deliverables: A clear and working architecture/pipeline adaptable to different intervention domain.

Methods

Arduino Nicla Voice:
Mic with Syntiant NDP120

Model adapted from **Edge Impulse** Open Source Project NDP.onClassification→ {airport, anomaly, bathroom, construction, home, road} -> ~90%

BLE GATT Notification: {Detected Event}

Intervention Generated by LLM Pushed

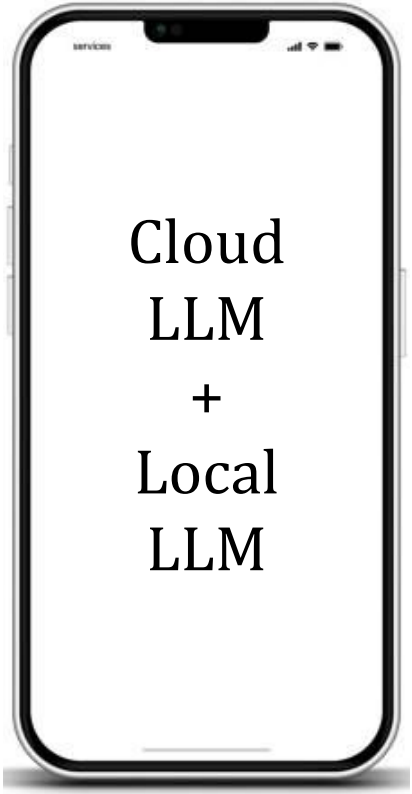
Cloud LLM (OpenAI GPT-4o): accessed via WebSocket API connection from the iPhone
Local LLM (TinyLlama 1.1B): executed on-device using the llama.cpp inference engine

Evaluation and Metrics

Conditions compared:

1. Cloud LLM only
2. Local LLM only
3. Hybrid: cloud + local in parallel, routed by context & latency needs

Stage-level ML classification accuracy;
System-level end-to-end latency;
(Target: < 1 minute)
Appropriateness of LLM prompts (qualitative);



Current Status and Next Steps

Current Status:
✅ **Edge ML model:** CNN trained and adapted for the Nicla Voice; Ready for on-device deployment

✅ **BLE transmission:** Edge → central device communication established; Messages received with timestamps on the central device

✅ **Central LLM layer:** Cloud LLM and local LLM both integrated; Responses generated and time stamped on the phone

Next Steps:
ML On-device deployment & benchmarking &
Full-pipeline integration & latency/accuracy measurement