DOJO: Distributed Open Justice Oversight

Executive Summary

A community-operated adversarial testing platform for AI models, particularly focused on healthcare applications. The “gym” will use LLeoMe as the baseline evaluator, combined with M3/M4 MCPs and community-contributed agents to identify biases, shortcuts, and blind spots before deployment. Models that pass rigorous testing receive certification valued above existing regulatory frameworks.

Core Concept

Problem Statement:

- General-purpose AI models are being used for healthcare without adequate regulation
- Existing certification (CHAI) lacks community input and real-world adversarial testing
- AI companies exploit regulatory gaps, especially in low-regulation environments
- Fiduciary alignment is missing from general-purpose models making clinical recommendations

Solution: Create a "gym" where models undergo adversarial training and testing through:

- Community-operated, community-owned evaluation infrastructure
- Multi-agent adversarial testing using LLeoMe + M3/M4 + contributed MCPs
- Certification that signals true robustness over checkbox compliance

Technical Architecture

Foundation Layer: LLeoMe

- **Training Data:** FathomAI transcripts, Claude conversations, published papers, WhatsApp shared documents
- **Base Model:** DeepSeek or LLaMa (open source)
- **Knowledge Base:** Conversations with philosophers, ethicists, indigenous leaders, religious community leaders, artists, students
- **Launch Date:** February 14, 2026 ("Choose Love") and February 17, 2026 ("Ride with Wisdom" - Lunar New Year)
- **Purpose:** Baseline evaluator trained on moral and ethical frameworks

Testing Layer: M3/M4 MCPs

- Direct connection to MIMIC and other PhysioNet datasets
- Open source, publicly accessible
- Guardrails implemented for data access

- Proven infrastructure from MIT Critical Data research and events

Community Layer: Contributed Agents

- Open call for MCPs/agents that test different aspects:
  - Shortcut learning detection
  - Bias identification across populations
  - Out-of-distribution robustness
  - Hidden assumptions in training data
  - Ethical reasoning consistency
  - Clinical decision quality under adversarial conditions

Inspiration & Evolution

Original Trigger: Molt (social media for agents where agents interrogate each other)

DOJO Innovation: Deliberate adversarial design - not just conversation, but systematic red-teaming to identify:

- Shortcuts in ML models
- Blind spots before deployment to diverse populations
- Misalignment with fiduciary duties
- Performance degradation in real-world conditions

Renato's Key Insight: Continuous adversarial testing provides value even if initial implementation is chaotic.

Certification Process

1. **Submission:** Developer submits model + write-up
2. **LLeoMe Evaluation:** Identifies blind spots based on ethical/philosophical frameworks
3. **M3/M4 Testing:** Real-world clinical data evaluation
4. **Community Agent Testing:** Multiple MCPs probe different vulnerability dimensions
5. **Iterative Training:** Model improves through DOJO feedback
6. **Certification:** Passing all tests grants certificate valued above CHAI

Strategic Value Propositions

For Developers

- Identify problems before deployment
- Address hidden biases early
- Improve chances of high-impact journal publication
- Gain credibility through rigorous community vetting

For Health Systems

- Higher confidence in model robustness
- Community-validated vs. expert-panel-only validation
- Reduced liability risk
- Ethical alignment verification

For Global Health Equity

- Catches exploitation before deployment to low-regulation regions
- Tests on out-of-distribution populations
- Community ownership prevents regulatory capture
- Multi-cultural perspective (through LLeoMe's training)

Operational Model

Governance: Community-operated, community-owned (contrasts with CHAI's expert-only approach)

Infrastructure:

- GitHub repository for LLeoMe
- Open recipe book for creating custom versions
- PhysioNet integration through M3/M4
- Continuous feedback from MIT Critical Data events and courses

Validation Loop: Models tested in actual courses and datathons, results fed back to developers

Multi-Pronged Defense Strategy

No single guardrail suffices - the gym addresses multiple exploitation vectors:

1. **Regulatory arbitrage:** Certification valuable even without legal mandate
2. **Direct-to-consumer AI:** Community validation provides user guidance
3. **Deployment in low-regulation regions:** Reputational risk of failing gym certification
4. **Lack of fiduciary alignment:** Explicit testing of alignment through adversarial scenarios
5. **Post-deployment harm:** Proactive testing reduces downstream litigation

Philosophical Underpinnings

LTARC Principles: Local, Task-specific, Agile, Reflective, Community-partnered

Epistemic Virtues (BODHI): Bridging, Open, Discerning, Humble, Inquiring

Federated Co-evolution: Human gyms (MIT Critical Data events) inform AI gym design; AI gym results inform human evaluation methods

Open Questions for Development

1. **Consensus mechanism:** How to handle disagreement among different agent evaluations?
2. **Selection bias:** Which tests are truly essential vs. nice-to-have?
3. **Scope:** Start US-focused or global from inception?
4. **Tracking:** How to verify LLeoMe/gym suggestions improve model quality?
5. **Scalability:** Cost management as submission volume grows

Competitive Positioning

vs. CHAI: Community-operated vs. expert panel; adversarial testing vs. framework compliance

vs. General-purpose LLMs: Moral underpinnings, ethical training corpus, healthcare-specific evaluation

Cultural Infiltration Strategy

LLeoMe serves dual purpose:

1. **Research tool:** Helps write papers for high-impact journals
2. **Cultural vector:** Trains next generation on ethical frameworks through daily use
3. **Gym foundation:** Provides moral baseline for evaluating other models

You're betting that students and investigators will adopt LLeoMe for its practical utility (better publications), thereby absorbing its ethical orientation, which then propagates through their own work and model development.

Key Insight: We are creating infrastructure for what regulation cannot achieve - community-driven adversarial testing that crosses cultural boundaries and prevents exploitation before deployment. The gym becomes simultaneously a technical testing platform and a cultural intervention in how AI models are developed and validated for healthcare.