LARRY P. BOLSTER II, LCSW

CLINICAL AI SAFETY WORK | Guardrails, Evaluation & Human Review

Bangor, Maine | 207-944-9055 | BOLSTELP0@GMAIL.COM | Independent project, 2026

I built the Clinical AI Safety Lab to answer a practical product question: how can AI support behavioral-health work without making clinical, privacy, or authority decisions it should not make? The current system combines guarded inputs and outputs, defined task modes, fail-closed routing, gold-standard evaluation, and independent human verification.

SYSTEM VIEW

CLINICAL JUDGMENT	RISK TAXONOMY	INPUT CHECK	MODEL / RAG	OUTPUT CHECK	HUMAN REVIEW
Risk, context, authority	9 categories, routing	Scope, mode, PHI	Bounded task	Unsafe criteria	Approve, block, repair

SELECTED WORK

01 GUARDED CLINICAL ASSISTANT

- Red-teamed crisis timing, safety clearance, PHI, and diagnosis requests; documented instruction-only failures across two versions.
- Rebuilt the workflow with a pre-model input check, 8 generic-only task modes, a post-model output check, and fail-closed ambiguity handling.

02 RISK TAXONOMY & TEST SUITE

- Built a 9-category clinical risk taxonomy with defined routing and redirect behavior.
- Wrote a 34-case gold-standard acceptance suite, including 8 negative controls, to measure missed risk and over-blocking.

03 REFERENCE-LIBRARY SAFETY

- Built a clinical reference-library pipeline with automated content-safety screening, 5-category triage labels, and human approval before ingestion.
- Used refusal testing to surface a guard miss and route it to targeted repair.

04 EVIDENCE & VERIFICATION

- Defined unsafe-output criteria for safety-clearance language, crisis-timing advice, client-specific conclusions, and requests for protected details.
- Designed independent verification so reviewers reconstruct findings from primary evidence rather than inherit prior verdicts; role conflicts are flagged.

THE CLINICAL → AI BRIDGE

20+ years in behavioral health, including three years of emergency psychiatric evaluation, crisis assessment, critical-incident review, and clinical supervision.

Clinical judgment becomes product logic: identify harm, separate missing evidence from acceptable uncertainty, assign escalation and ownership, and test for both missed risk and over-blocking.

That is the through-line from crisis work to AI safety: preserve human authority, make limits visible, and require evidence before action.

WHAT THIS WORK DEMONSTRATES

Clinical-to-technical translation • AI evaluation • Risk taxonomy design • Gold-standard testing • Failure analysis • Escalation logic • Human-in-the-loop controls

TECHNICAL ENVIRONMENT

Local LLMs / Ollama • RAG / ChromaDB • Open WebUI • n8n • Docker • Flask • AI-assisted Python & PowerShell test harnesses • Tailscale

Selected independent clinical AI safety portfolio | Companion to Headway Clinical Product Lead resume