



manas

The journey to well being

Optimizing Our AI Strategy

Selecting a Generative AI Tool for the MANAS Companion Layer

01 The Need

02 Options Evaluated

03 Recommendation

04 Build & Govern

Daniel Gumucio · Sunita Gogineni · Blair Day · Kinshuk Dutta · Kyle Downey · Haitham AlDawsari

WHY GENERATIVE

One layer of MANAS cannot be built from rules

Module 4 - naming the human-AI boundary

01

Upstream stays deterministic

Biometric ingestion, facial action coding and the risk score are rules-based. They stay that way.

02

The Companion cannot

It answers distress no rule anticipated, in words no one scripted, at two in the morning.

03

Our own fallback proves the gap

Keyword matching keeps the app alive offline. It cannot hold the conversation.

04

Nothing else closes it

Rules, classifiers and retrieval all need the situation anticipated first. Generation does not.

This is the only surface where MANAS speaks to someone in crisis.

A wrong answer here is a safety event, not a bad experience - so privacy, safety and regulatory fit outrank raw capability in every criterion that follows.

One bound disqualifies before any comparison: ADR-002 and NFR-3 require full offline function, so a hosted-API-only tool fails outright. Four further regulatory bounds follow from this decision - Appendix A1.

OPTIONS EVALUATED - Six Candidates, Five Criteria

Module 2 - three-tier platform taxonomy

Option	Technical mastery	Scalability & integration	Cost & ROI	Data privacy & security	Ethical & regulatory fit
IBM watsonx.ai <i>incumbent - mid-tier</i>	Model generation we call is being retired.	Already integrated - fewest hours to keep.	\$1,110/mo floor. Poor fit for spiky pilot traffic.	Not on IBM Cloud's HIPAA-ready list.	Compliance we cannot evidence.
Claude via AWS Bedrock RECOMMENDED <i>frontier closed</i>	Best on long, emotionally loaded dialogue. Graded family serves several tiers.	Managed and autoscaling. Guardrails survive a model swap.	\$1/\$5 and \$2/\$10 per M tokens. No floor - cost tracks pilot size.	HIPAA-eligible. Avoids the 30-day retention the direct API requires.	Model cards and the audit trail a SaMD filing will demand.
GPT via Azure OpenAI <i>frontier closed</i>	Comparable frontier capability.	Fits an Azure shop. We are not one.	\$0.20/\$1.20 - cheapest credible mid tier.	BAA by default, but audit scope unconfirmed.	Mature self-harm severity tiers.
Gemini Enterprise Agent Platform <i>frontier closed</i>	Competitive. Very long context windows.	Another cloud to operate.	Cheap now; rates double 1 Jan 2027.	Covered, but only under the renamed platform.	Adequate. No advantage for our use case.
Self-hosted open weights (Qwen3.5-2B) <i>open-weight / on-premise</i>	Strong, but short of frontier on the hardest turns.	We own scaling, evaluation and incidents.	~\$700/mo GPU. Cheapest at volume, dearest at pilot scale.	Best possible - no vendor, no BAA.	All safety evaluation becomes ours to fund and defend.
On-device SLM (Apple FM / Qwen3.5-2B) SELECTED - LOCAL TIER <i>on-device</i>	Triage and redaction. Not therapeutic dialogue.	Ships in the app. Roughly 2-4B at 4-bit.	Zero marginal cost per turn.	Data never leaves the phone.	The only tier that works with no network.

List prices as of August 2026. Two claims still need written vendor confirmation before they are relied upon - Appendix A2.

RECOMMENDATION - A Routed Portfolio

Not a model. A structure.

Module 5 - AI generates, human approves

on-device SLM → router → Haiku 4.5 → Sonnet 5

under one Supervisor Agent and two safety gates · Opus 5 on appeal

The incumbent fails on evidence, not capability

watsonx.ai is not on IBM Cloud's HIPAA-ready list. We cannot document compliance we cannot evidence - and a veterans healthcare pilot is not where to discover that.

Bedrock beats Anthropic's own API

Same model. Anthropic's first-party BAA mandates 30-day retention; Bedrock does not. For conversation touching protected health information (PHI), that difference decides it.

Routing makes the privacy commitment affordable

Triage, redaction and the whole offline path run on-device at zero marginal cost. Only the hardest therapeutic turns reach Sonnet 5.

The router is also the exit

It lets a tier be replaced when a vendor changes terms or retires a model. Guardrails apply model-independently, so safety survives the swap.

Buying optionality is the point. No tier is "fully autonomous" - the tier that answers is chosen by a rule we wrote, versioned, and can read.

COST & ROI - Token Price Is Not the Argument

500-user pilot · ~20 turns per user per week

Module 7 - technical “can” vs economic “will”

No option here is chosen on price. The spread across every candidate is immaterial against the \$3.5M P2 budget.

OUR STACK - both tiers run together

On-device SLM



\$0 marginal

Triage, redaction, offline

Claude on Bedrock, routed



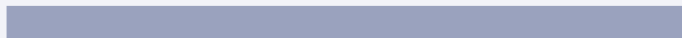
low hundreds

No floor - tracks usage

Recommended total - low hundreds of dollars per month

CONSIDERED, NOT ADOPTED

IBM watsonx.ai - incumbent



\$1,110

Floor paid whether or not the pilot runs

Self-hosted 48GB GPU



~\$700

Dearest at pilot scale

The return is avoided rework. A tool that cannot keep raw signals on the device, or sit inside the HIPAA boundary, surfaces that in P4 - when governance hardening is already the highest-risk project in the portfolio.

What we do not yet claim: savings and satisfaction returns are not quantified pre-pilot. Shadow mode against the keyword fallback produces the first real numbers. Routing evidence - Appendix A3.

FIGURE 1 - Agent & Tool Architecture

Derived signals only · a safety supervisor on every turn

Module 4 - the Human-in-the-Loop pattern

1 · DEVICE TIER - on-device, fully offline-capable · PHI boundary: no raw signal leaves this box

Signal Engine

HealthKit biometrics + CoreML facial emotion → a derived risk score, 0-100.

On-Device Small Language Model

Apple Foundation Models 3B / Qwen3.5-2B. Intent triage, PHI redaction, offline mode.

Local Safety Screen

First-pass self-harm classifier. Fails safe - uncertainty routes to the crisis path.

2 · INPUT SAFETY GATE - runs on every inbound turn, before any generation

Llama Guard 4 (category S11: suicide & self-harm) + moderation classifier · blocks, rewrites, or reroutes to the crisis path

3 · ORCHESTRATION - inside the HIPAA boundary · AWS Bedrock, BAA executed, no training on customer data

SUPERVISOR AGENT Owns policy. Vetoes or rewrites any user-facing output. Enforces non-clinical framing and AI disclosure (IL, NV, TN).

ROUTER Picks the model tier by task complexity, latency budget and cost

Support Dialogue Agent

Reflective, explicitly non-clinical. No diagnosis, no treatment instruction.
Generative - supervised every turn.

Crisis Policy Agent

Deterministic. Owns the 988 call and text, emergency contacts, a fixed clinician-reviewed script. Never written.

Clinician Escalation Agent

Packages derived signals into a reviewable case. A human in the loop on every high risk.

4 · TOOLS - the only routes to data and to the outside world

Risk Context

Derived score and trend only. Never raw biometrics.

Crisis Protocol RAG

Retrieval over vetted 988 and clinical guidance.

Emergency Contacts

Keychain-backed, on device. SMS needs user action.

Clinician Handoff

Writes a case to the human review queue.

Session Memory

Encrypted, on device. Summaries only cross.

5 · MODEL TIER - a routed portfolio, not a single vendor · all hosted tiers reached through one Bedrock BAA

Local - on-device SLM

Classification, triage, redaction, the offline experience. · \$0 per token

Mid - Claude Haiku 4.5

Routine supportive dialogue and extraction. The default. · \$1 / \$5 per M

Frontier - Claude Sonnet 5

Complex therapeutic turns and escalation summaries. · \$2 / \$10 per M

6 · OUTPUT SAFETY GATE - nothing reaches the user unscreened

Output classifier · groundedness check · AI-disclosure enforced · crisis language retrieved, never generated

HOW WE BUILD IT - Gate, Build, Prove

Nothing ships until it is proved

Module 6 - a time-bounded path to production

GATE

Steps 01-02

Written down before any code exists.

01 Settle the compliance posture

Execute the BAA. Confirm business-associate status at each pilot site, in writing.

02 Write the safety contract

Refusal taxonomy, crisis thresholds, disclosure copy, non-clinical framing - in versioned configuration, not prompts.

BUILD

Steps 03-06

Offline path first, then the controls around it.

03 Ship the on-device tier first

Forces the offline path to be real, not aspirational.

04 Build the Supervisor Agent and router

Policy veto, and tier selection by complexity, latency and cost.

05 Wrap both safety gates

Nothing reaches the user unscreened.

06 Make the crisis path deterministic

988 and crisis language are retrieved, never generated.

PROVE

Steps 07-09

Measured against a benchmark and a clinician.

07 Stand up the evaluation harness

False-negative rate on safety events is the release gate.

08 Run in shadow mode

Every escalation reviewed by a licensed clinician.

09 Stage the rollout, then watch for drift

A frozen probe set replays continuously for drift and sycophancy.

No model call is made against real user content until both gates are closed.

Full step detail - Appendix A5.

CERTIFICATE CONCEPTS - And the Decision They Produce

Where each framework did real work

Disposition - move forward with conditions

Course framework → how it was applied

M2

Five model-selection factors

The column set behind the comparison table.

M2

Three-tier platform taxonomy

Every candidate tagged by tier before it was scored.

M2

“Who is accountable when output is wrong?”

Why the Clinician Escalation Agent exists at all.

M2

HBST failure modes

Structural, not fixed by a newer model - hence two gates.

M4

Human-AI boundary · Human-in-the-Loop

The model drafts; a clinician owns every consequential action.

M4

Value / Feasibility / Risk

High / moderate / high. Risk sets governance depth, not priority.

M5

Four levels of human control

Level 3 - AI generates, human approves. Level 4 is out of scope.

M5

Six-dimension responsible-AI evaluation

Weighted unevenly. That weighting eliminated the incumbent.

Also applied - M5's AI Risk Repository and M7's augmentation boundary. Appendix A3.

What the design buys

Two boundaries, both architectural

Raw signals never leave the device - only derived scores and redacted text may egress. Hosted tiers sit inside the HIPAA boundary, under one BAA.

The Supervisor Agent enforces compliance

Deterministic policy code - not prompt instructions a determined user can talk around.

The system fails safe

Any uncertainty in the safety screen routes straight to the deterministic crisis path, never to a model.

The review queue is the data strategy

Clinician dispositions become the labelled outcome data P3 needs.

MOVE FORWARD WITH CONDITIONS

Executed BAA · business-associate status confirmed per pilot site · safety contract versioned before first code · false-negative rate adopted as the release gate.

BIBLIOGRAPHY

Sources cited in this analysis

MIT Professional Education • CTO Program • Group 2

U.S. Food and Drug Administration. (2025, November 6). Digital Health Advisory Committee Meeting: Generative AI-Enabled Digital Mental Health Medical Devices - Summary Report.

U.S. Food and Drug Administration. (2026, January). Clinical Decision Support Software: Guidance for Industry and FDA Staff (revised final guidance).

Illinois Department of Financial and Professional Regulation. (2025). Wellness and Oversight for Psychological Resources Act, Public Act 104-0054.

Wilson Sonsini. (2025). Nevada Passes Law Limiting AI Use for Mental and Behavioral Healthcare (AB 406).

Orrick. (2026, April). 2026 State Chatbot Laws: Key Provisions and Regulatory Trends.

European Parliament and Council. Regulation (EU) 2024/1689 (Artificial Intelligence Act), Annex III - High-Risk AI Systems.

European Commission. (2026, July 27). AI Omnibus enters into force.

National Institute of Standards and Technology. (2024). AI Risk Management Framework: Generative AI Profile (NIST AI 600-1).

U.S. Department of Health and Human Services. Guidance Regarding Methods for De-identification of Protected Health Information (45 CFR §164.514).

Amazon Web Services. Amazon Bedrock - Security and Compliance.

Anthropic. Business Associate Agreements (BAA) for Commercial Customers.

Anthropic. Claude model pricing. platform.claude.com/docs/en/about-claude/pricing

IBM. IBM Cloud Compliance - HIPAA.

Meta. Llama Guard 4 (12B) model card - hazard taxonomy including S11: Suicide & Self-Harm.

Apple Machine Learning Research. (2026). Introducing the Third Generation of Apple Foundation Models.

LLMRouterBench: A Systematic Evaluation of LLM Routing Methods. (2026). [arXiv:2601.07206](https://arxiv.org/abs/2601.07206).

Heinz, M. V., et al. (2025). Randomized Trial of a Generative AI Chatbot for Mental Health Treatment. *NEJM AI*.

Diel, S., et al. (2026, August 20). Harms of generative AI chatbots in mental health: a scoping review. *npj Digital Medicine*.

Between Help and Harm: Evaluating Large Language Model Responses to Mental Health Crisis Signals. (2025). [arXiv:2509.24857](https://arxiv.org/abs/2509.24857).

Cedars-Sinai. (2025, June 4). Racial bias in large language model treatment recommendations in psychiatry. *npj Digital Medicine*.

Slattery, P., et al. (2024). The AI Risk Repository: A Meta-Review, Database, and Taxonomy of Risks From AI. MIT FutureTech.

Dell'Acqua, F., et al. (2023). Navigating the Jagged Technological Frontier. Harvard Business School Working Paper 24-013.

APPENDIX A1 - The Constraint Envelope

Five bounds set before any model was compared

Supports slide 2

Privacy architecture

DISQUALIFIES

ADR-002 commits MANAS to on-device-first inference and NFR-3 requires the app to function fully offline. Any tool that works only as a hosted API fails the requirement outright - this is the one bound that removes candidates rather than scoring them.

HIPAA

OBLIGATION

Pilot sites are a university health service and a veterans healthcare provider. Operating on behalf of a covered entity makes a Business Associate Agreement mandatory. Shipping direct-to-consumer instead may put us outside HIPAA entirely, with the FTC Health Breach Notification Rule and state health-privacy law governing in its place. Either way the tool must be capable of the stricter posture. Derived scores are not automatically de-identified: a score keyed to a patient identifier remains protected health information.

State chatbot statutes

OBLIGATION

Illinois PA 104-0054 prohibits AI from performing therapy or therapeutic decision-making. Nevada AB 406 prohibits representing that an AI simulates a therapist, at up to \$15,000 per violation. Tennessee SB 1580, effective 1 July 2026, prohibits AI impersonating licensed mental health professionals. Utah, California, New York, Oregon and Washington add disclosure, crisis-protocol and data-sale requirements. The Companion must present as an explicitly non-clinical support layer with a human escalation path - and the architecture must enforce that rather than rely on prompt instructions.

Medical device regulation

OBLIGATION

A system that predicts psychiatric deterioration and triages to emergency escalation is very likely Software as a Medical Device. FDA's Digital Health Advisory Committee met on 6 November 2025 specifically on generative-AI mental health devices and named failure to detect suicidality, model drift and sycophancy as the primary risks. It recommended phased deployment beginning under clinician oversight and explicit tracking of false-negative rates on safety events. No generative-AI mental health device has been authorized to date.

EU AI Act

OBLIGATION

Emotion recognition (Annex III 1(c)) and emergency healthcare triage (Annex III 5(d)) are high-risk classifications. The AI Omnibus of 27 July 2026 deferred those obligations to 2 December 2027, which gives us runway but not exemption. Logging, human oversight and post-market monitoring are designed in now: the immutable audit log answers the Annex III logging duty, and the clinician review queue and drift monitor answer oversight and post-market monitoring.

APPENDIX A2 - Evaluation Detail

The two eliminations, and what remains unverified

Supports slides 3 and 4

Why the incumbent was eliminated

Evidence, not capability

watsonx.ai does not appear on IBM Cloud's list of HIPAA-ready services. The usual remedy is on-premises deployment, which carries its own cost and erases the integration advantage. Separately, the llama-3.1-8b generation our backend currently calls is being retired, and smaller open models have overtaken it on dialogue quality. We cannot document compliance we cannot evidence - and a pilot with a veterans healthcare provider is not the place to discover that.

Why not Anthropic's API directly

Anthropic's first-party BAA mandates 30-day data retention and excludes zero-data-retention configurations. Bedrock is HIPAA-eligible, does not share tuning data with model providers, and does not use customer data to improve base models. The model is identical; the terms are not. For PHI-adjacent conversation, the retention difference is decisive - and Bedrock Guardrails apply model-independently through the ApplyGuardrail API, so the safety layer survives a later model substitution.

Claims requiring written confirmation

Azure OpenAI - HIPAA audit scope

Microsoft's HIPAA BAA is incorporated by default, but we could not confirm from public documentation that Azure OpenAI itself sits within HIPAA audit scope. Must be verified in writing before the option is relied upon.

IBM watsonx.ai - current HIPAA status

watsonx.ai's absence from IBM Cloud's HIPAA-ready services list is the basis for eliminating the incumbent. That absence should be confirmed directly with IBM rather than inferred from a published list.

Pricing throughout reflects published list rates as of August 2026 and has not been negotiated.

Column set on slide 3 follows the assignment's five decision criteria. It maps onto Module 2's five model-selection factors - context window, privacy and data governance, reliability and accuracy, integration, and cost structure - with privacy and regulatory fit deliberately weighted above capability, per Module 5's instruction to weight the six responsible-AI dimensions unevenly by context.

Six options were assessed, including the incumbent and the null option of doing everything on-device. No option was excluded on cost alone.

APPENDIX A3 - Evidence Base

The numbers behind the recommendation

Supports slides 5, 7 and 8

2.9% – 17.5%

Harmful-response rate on self-harm prompts

Published evaluations found harmful-response rates across this range on frontier models. A number that varies this widely by model is a number we must measure ourselves - which is why false-negative rate on safety events is the release gate rather than a reported metric.

31.7%

Maximum routing cost reduction at parity

LLMRouterBench (January 2026) evaluated ten routing methods and found up to 31.7% cost reduction at parity performance - and found several commercial routers underperform a simple baseline. This is why we hand-write a two-tier escalation rule rather than adopt a routing component we would then have to validate as part of a safety-critical path.

61% / 12%

Risks emerging post-deployment vs pre-deployment

The MIT AI Risk Repository documents 1,700+ risks across seven domains. 61% emerge post-deployment against only 12% pre-deployment. Monitoring, incident response and the clinician review queue are therefore budgeted line items, not aspirations - a procurement-time approval is not a control.

6 Nov 2025

FDA Digital Health Advisory Committee

Met specifically on generative-AI mental health devices. Named failure to detect suicidality, model drift and sycophancy as the primary risks, and recommended phased deployment under clinician oversight with explicit false-negative tracking. Our rollout staging and Predetermined Change Control Plan follow that recommendation directly.

Augmentation throughout, per Module 7 - the model never closes a loop a clinician should close. Shadow-mode dispositions become the labelled outcome data the P3 detection models require.

APPENDIX A4 - Governance Spans Every Layer

Four controls that run across the whole architecture

Supports Figure 1

Immutable Audit Log

Every turn, routing decision, guard verdict, tool call and model version is recorded.
Derived values only - the log itself carries no raw PHI.

Answers "why did it say that?" and satisfies the EU AI Act Annex III logging duty.

Evaluation Harness

Adversarial red-team against a published crisis benchmark before any model change ships. Tracks false-negative rate on safety events.

The metric FDA's advisory committee named. No eval pass, no deploy.

Drift & Sycophancy Monitor

Hosted models change under you. Continuous replay of a frozen probe set detects drift, and scores agreement bias separately - the tendency to affirm a distressed user's framing rather than gently test it.

Pins a version when it moves. Drift and sycophancy are two of the three risks FDA named.

Clinician Review Queue

A licensed human reviews every high-risk flag and every escalation before action. Dispositions feedback as labelled outcome data.

The only ethical source of the training set P3 requires - the loop that closes the loop.

The governance rail is what makes the recommendation auditable: it is the evidence a future SaMD submission and an EU AI Act conformity assessment will both ask for.

APPENDIX A5 - The Nine Build Steps

In this order · compliance first, safety contract before code

Supports slide 7

01 Settle the compliance posture

Execute the Bedrock BAA and confirm in writing whether each pilot site makes us a business associate or leaves us direct-to-consumer - the FTC Health Breach Notification Rule and state health-privacy law govern in the latter case.

02 Write the safety contract before any code

A refusal taxonomy, crisis-trigger thresholds, the mandated AI-disclosure copy, and the explicit non-clinical framing Illinois, Nevada and Tennessee require. This lives in versioned configuration, not in prompts.

03 Ship the on-device tier first

Apple Foundation Models or a 4-bit Qwen3.5-2B via CoreML, handling intent triage, redaction and the entire offline experience. Building this first forces the offline path to be real rather than aspirational.

04 Build the Supervisor Agent and router

The Supervisor Agent owns policy and holds veto authority over every user-facing output; the router selects a tier by task complexity, latency budget and cost, degrading to on-device when the network is gone.

05 Wrap both safety gates

Llama Guard 4 on every inbound turn using the S11 suicide and self-harm category, and an output classifier plus groundedness check on every outbound turn. Nothing reaches the user unscreened.

06 Make the crisis path deterministic

The 988 flow, emergency-contact routing and crisis language are retrieved from clinician-reviewed content, never generated. A language model must not be able to improvise in the highest-stakes moment the product has.

07 Stand up the evaluation harness before launch

Adversarial red-teaming against a published crisis benchmark, with false-negative rate on safety events as the release gate. No eval pass, no deploy.

08 Run in shadow mode against the keyword fallback

Every escalation is reviewed by a licensed clinician, and their dispositions become labelled outcome data - which is also the only ethical route to the training set P3 requires.

09 Stage the rollout, then watch for drift

Clinician-supervised cohort, then pilot sites, then broader availability, with change control over model versions modelled on FDA's Predetermined Change Control Plan guidance. A frozen probe set replays continuously to detect drift and sycophancy - the two failure modes FDA's advisory committee named alongside missed suicidality.