

Part 02

STACKAI WHITEPAPER

AI Agents for Digital Health and Virtual Care

How virtual-first providers, digital health platforms, and care-management companies are deploying AI agents to grow their patient panel without growing headcount.

A practical guide for CIOs, COOs, CFOs, and Heads of Clinical Operations evaluating AI agent platforms in virtual care.



PREPARED BY

Shani Fargun

VP Healthcare at StackAI

sfargun@stack-ai.com



HIPAA · SOC 2 TYPE II · ISO 27001 · GDPR

BAAs & DPAs with service providers

Executive summary

Digital health and virtual care companies were built on the promise that technology could make care more accessible, more convenient, and more affordable. AI agents are the next layer of that promise: operational infrastructure that lets a 200-person company serve the patient volume that used to require 350.

This whitepaper consolidates StackAI's deployment learnings across virtual-first providers, behavioral health platforms, telehealth, and digital MSK into five production use cases: patient message classification and draft-reply, voice and live-chat patient assistance, intelligent intake and eligibility, message and call quality monitoring, and cancellations and retention.

At scale these run in production. On a single national telehealth deployment, the messaging stack alone runs well over 350,000 draft classifications and 380,000 intent-routing decisions per quarter, the patient-summary service backing every agent runs roughly 667,000 lookups per quarter, and the QA layer grades more than 440,000 interactions per quarter. Companies running these workflows report 25 to 35 percent reductions in cost per active patient per month, 15 to 20 point improvements in intake completion, 40 to 60 percent deflection of inbound support volume, near-total audit coverage of patient communications, and notes-compliance rates above 96 percent.

The companies that operationalize this first will serve more patients, in more states, across more conditions, with the same team. In a market where acquisition costs are rising and reimbursement is tightening, that operational leverage is what separates the companies that scale from the ones that stay small.

380K+intent-routing
decisions / quarter**667K**patient-summary
lookups / quarter**440K+**interactions QA-
graded / quarter**25–35%**lower cost per active
patient

The scale-up paradox in virtual care

Digital health companies face a contradiction that brick-and-mortar practices never had to confront so starkly: the entire value proposition is that technology makes care delivery more efficient, but behind the sleek patient app, operations often look remarkably manual.

A behavioral health platform might have 50,000 active patients. A telehealth provider might handle 30,000 visits a month. A digital MSK program might be embedded in a dozen employer contracts. In every case the business model rests on one ratio: cost per active patient per month. And that ratio is under constant pressure.

Clinical staff are an inherently constrained resource. But you can remove the operational friction that eats into their patient-facing time, and you can automate the administrative work that currently requires human judgment only because no one has built the system to handle it autonomously.

That is where AI agents come in, as operational infrastructure that lets a smaller team serve a larger panel.

The KPIs that define success

The metrics that matter here are precise: cost per active patient per month, time-to-resolution, ticket deflection rate, provider minutes saved, intake completion rate, and churn. Every use case in this whitepaper maps to one or more of them.

◆ INDUSTRY SIGNAL

Gartner projects that by 2029, agentic AI will autonomously resolve 80 percent of common customer-service issues without human intervention, driving a 30 percent reduction in operational costs. Virtual care is one of the verticals where this shift is happening fastest, because patient communications are already digital-native and the ratio of operational tasks per clinical hour is unusually high.

Why virtual care needs different AI infrastructure

A platform built for a hospital system is not automatically the right fit for a virtual-first provider operating in 40 states.

CHARACTERISTIC	WHY IT MATTERS FOR PLATFORM SELECTION
Multi-state complexity	A company operating in 40 states faces 40 sets of telehealth regulations, consent requirements, and scope-of-practice rules. Agents need to be configurable per jurisdiction, not just per organization.
Flexible architecture	Virtual care runs on a stack of SaaS tools: telehealth platform, separate EHR, CRM, engagement platform, billing. Agents need API and integration depth to orchestrate across that stack rather than assuming an EHR-centric workflow.
Speed of iteration	Digital health ships product weekly, not annually. The platform must support new intake questions, updated compliance rules, and revised escalation criteria without an engineering cycle for each change.
Patient experience as product	In virtual care the patient experience is the product. Every AI-mediated interaction is a product touchpoint, so the bar for tone, quality, and accuracy is materially higher than in back-office automation.
Data as a strategic asset	Virtual care generates rich longitudinal data. Agents that leverage it (spotting engagement patterns, predicting churn, optimizing provider matching) compound in value over time.

The five use cases

Across StackAI's digital health customers, five workflows recur with the highest measurable impact on unit economics and patient outcomes. They are presented in the order most virtual care leaders deploy them.

1

Patient Messaging: Classification, Draft-Reply, and QA

Classifies every inbound message, drafts a tone-matched clinical reply for the rep or provider to approve, and grades the result.

2

Voice and Live-Chat Patient Assistant

Resolves the majority of inbound patient conversations autonomously across chat and voice, with intelligent escalation to the care team.

3

Intelligent Intake, Eligibility, and Insurance Capture

Orchestrates identity, eligibility, clinical screening, consent, and provider matching as one adaptive conversation.

4

Message and Call Quality and Compliance Monitoring

Near-total audit coverage of every patient interaction against HIPAA, state telehealth rules, scope-of-practice, and internal quality standards.

5

Cancellations, Retention, and Patient-Flow Orchestration

Intercepts churn signals, offers evidence-based retention paths, and orchestrates the between-visit lifecycle.

1 Patient Messaging: Classification, Draft-Reply, and QA

THE PROBLEM

For a virtual-first provider, the message is the front door to care. Patients ask about scheduling, symptoms, medications, side effects, refills, insurance, and program questions, all day, at a volume that overwhelms human teams. Response times slip from minutes to hours to days, and every slow reply is a churn risk. Doing this well means getting the intent right, pulling accurate patient context, and replying in a clinically appropriate, on-tone voice, thousands of times a day.

HOW AN AI AGENT HANDLES IT

The stack has three layers that share one patient-summary object:

- A classifier reads every inbound portal message and routes it (provider queue, MSR queue, refill, cancellation, billing, escalation, auto-response). It is trained on the customer's own message history and continuously re-tuned against QA-graded outcomes.
- A draft-reply agent pulls the patient's full context, retrieves the SME-approved answer or protocol from the knowledge base, and writes a reply for the rep or provider to review and send. Separate flows exist for weight-management, women's health, renewals, and providers, each with its own tone rules and clinical guardrails.
- A QA agent grades the interaction so the loop improves over time.

WHY IT MATTERS

In subscription virtual care, churn correlates strongly with responsiveness. The classifier is the single most-invoked workflow in a mature deployment (roughly 380,000 routing decisions per quarter on one national provider), and the draft-reply layer runs even higher. Getting the intent right and the first draft written is what lets the human team clear the queue instead of drowning in it.

BUILT FOR	Director of Patient Support, VP Clinical Operations, Head of Member Services
KEY KPIs	Median response time, messages handled per rep, first-contact resolution, CSAT, churn
TYPICAL IMPACT	Hours to seconds on the routed tier, human queue shrinks so remaining messages resolve faster, tone and clinical accuracy held constant at volume

2 Voice and Live-Chat Patient Assistant

THE PROBLEM

Patient engagement matters, and patients do not keep business hours. A growing share of inbound volume (scheduling, program information, refills, general questions) does not need a clinician at all, but it still lands in a human queue.

HOW AN AI AGENT HANDLES IT

A production voice- and chat-enabled assistant handles inbound conversations with tool-calling access to the patient record, provider availability, and program-eligibility rules. It goes beyond a scripted FAQ bot: it understands multi-issue messages, queries live systems for specific answers, triages clinical concerns to the care team with a structured summary, and operates around the clock. The architecture uses subflow tools, so the same chat agent can call scheduling, eligibility, and knowledge-base subflows without one monolithic prompt. Voice calls are post-processed into structured intent and summary records for QA and analytics. A flagship live-chat surface runs roughly 20,000 production sessions per quarter.

WHY IT MATTERS

A patient who waits 18 hours for a simple answer is far more likely to disengage than one who gets an accurate answer in 30 seconds. The agent protects revenue as well as cost, and it shrinks the human queue so the remaining tickets resolve faster too.

BUILT FOR	Director of Patient Support, VP Operations, Head of Member Services
KEY KPIs	Ticket deflection rate, median response time, CSAT, churn
TYPICAL IMPACT	0 to 15 percent deflection becomes 40 to 60 percent, 24/7 coverage, faster resolution on the human-handled tier

3 Intelligent Intake, Eligibility, and Insurance Capture

THE PROBLEM

Virtual onboarding is more involved than it looks: identity verification across a digital channel, eligibility checking across many payers and plan types, clinical screening that varies by specialty (PHQ-9, GAD-7, medications, allergies), telehealth and HIPAA consents, provider matching on clinical need, availability, licensure, and language, and first-appointment scheduling that accounts for all of it. In most companies this touches three to five systems, takes 15 to 30 minutes of staff time per patient, and completes well below 100 percent. Every patient who starts intake but does not finish is a lost acquisition.

HOW AN AI AGENT HANDLES IT

The agent runs intake as one adaptive conversation: it collects demographics and insurance conversationally, verifies eligibility in real time, administers screening instruments and branches on the responses, presents consents in plain language, matches the patient to a provider, and offers scheduling. A companion clinical-eligibility validator checks each consult against program rules (contraindications, interactions, age or BMI thresholds, state prescribing restrictions) and returns a pass, flag, or hold with the specific rule cited. And when a patient uploads an insurance card or a paper intake form, an OCR-plus-extraction step returns a structured coverage record (payer, member ID, group, plan, copay) written straight to the chart, cutting the "wrong insurance on file" errors that drive denials weeks later.

WHY IT MATTERS

Intake completion is one of the most leveraged metrics in digital health. A 10-point improvement on 5,000 monthly starts is 500 additional patients entering care with zero additional marketing spend.

BUILT FOR	VP Patient Experience, Director of Operations, Director of Patient Acquisition
KEY KPIs	Intake completion rate, time-to-first-appointment, cost per completed intake, clean-insurance rate
TYPICAL IMPACT	65 to 75 percent completion becomes 80 to 90 percent, 5 to 10 day time-to-first-appointment becomes 1 to 3 days, staff time per intake cut by more than half

4 Message and Call Quality and Compliance Monitoring

THE PROBLEM

Support teams (care navigators, patient success, coaches) handle inbound and outbound communications subject to HIPAA, state telehealth rules, and internal quality standards. Manual QA typically covers 2 to 5 percent of interactions, which means 95 percent-plus go unreviewed. For a company handling 30,000 interactions a month across multiple regulatory regimes, that unreviewed volume is existential risk.

HOW AN AI AGENT HANDLES IT

An agent reviews every interaction (call transcript, chat log, message thread) against a structured rubric: HIPAA identity verification and minimum-necessary standards, scope-of-practice and state consent requirements, protocol and script adherence, resolution, tone, and escalation appropriateness, plus sentiment and safety-signal detection. Results are scored, rolled up by rep, team, program, and channel, and flagged for human review when issues surface. On a mature deployment this grades more than 440,000 interactions per quarter, up from a sampled few percent.

WHY IT MATTERS

Automated monitoring turns QA from a periodic spot-check into a continuous, audit-ready trail, at a cost that scales sub-linearly with volume. For a multi-state operator, that coverage directly reduces regulatory tail risk.

BUILT FOR	Compliance Director, VP Customer Operations, QA leads, Head of Clinical Operations
KEY KPIS	Percent of interactions audited, issue-detection rate, escalation accuracy, multi-state exposure
TYPICAL IMPACT	2 to 5 percent coverage becomes near-total, earlier detection of systemic issues, audit-ready trail per interaction

5 Cancellations, Retention, and Patient-Flow Orchestration

THE PROBLEM

Patient drop-off quietly erodes virtual care unit economics. A patient who enrolls, has two sessions, and disappears consumed onboarding cost without generating enough lifetime value to cover acquisition. The between-visit lifecycle (assessments, refills, adherence nudges, transitions, re-engagement) works at 5,000 patients and breaks at 50,000.

HOW AN AI AGENT HANDLES IT

A cancellation agent intercepts inbound cancellation requests, classifies the reason (cost, side effects, no longer needed, switching, clinical concern), pulls the patient's full history, and either offers an evidence-based retention path (pause, plan change, provider swap, coupon) or routes to a specialist with the case pre-summarized. For clinically sensitive cases (adverse events, protocol failures) it escalates straight to a clinical reviewer rather than trying to retain. A companion flow orchestrates the rest of the lifecycle: post-visit summaries reinforcing instructions, between-session monitoring of patient-reported data with threshold alerts, transition summaries, and multi-step re-engagement when disengagement signals appear.

WHY IT MATTERS

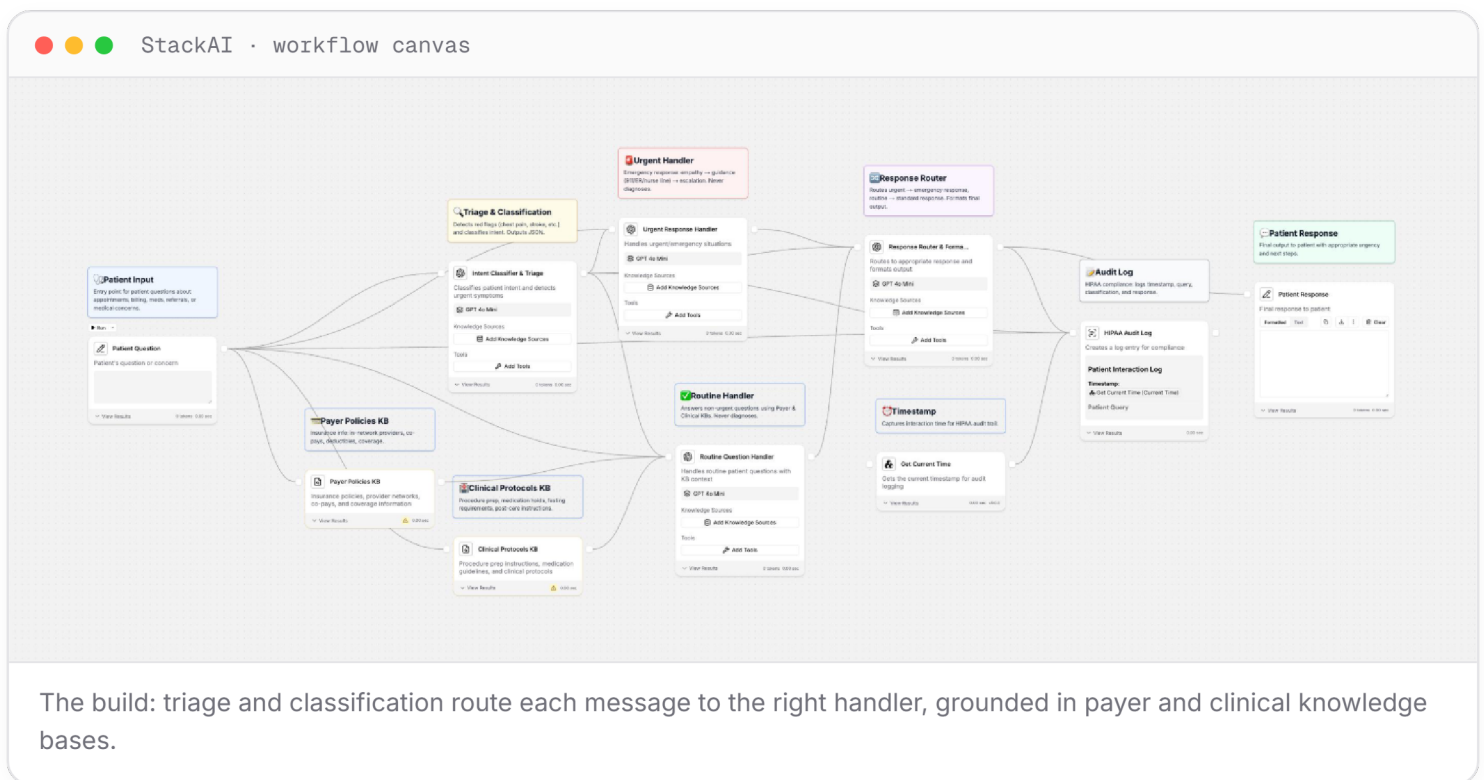
Automated flow orchestration protects the business model directly and frees coordinator capacity for the high-touch cases where human attention changes outcomes.

BUILT FOR	Retention leads, care managers, telehealth operations teams
KEY KPIs	Retention lift, adherence, follow-up completion, churn, lifetime value
TYPICAL IMPACT	Measurable retention lift with clinical guardrails intact, earlier detection of disengagement, coordinator capacity recovered

SEE IT IN PRODUCTION

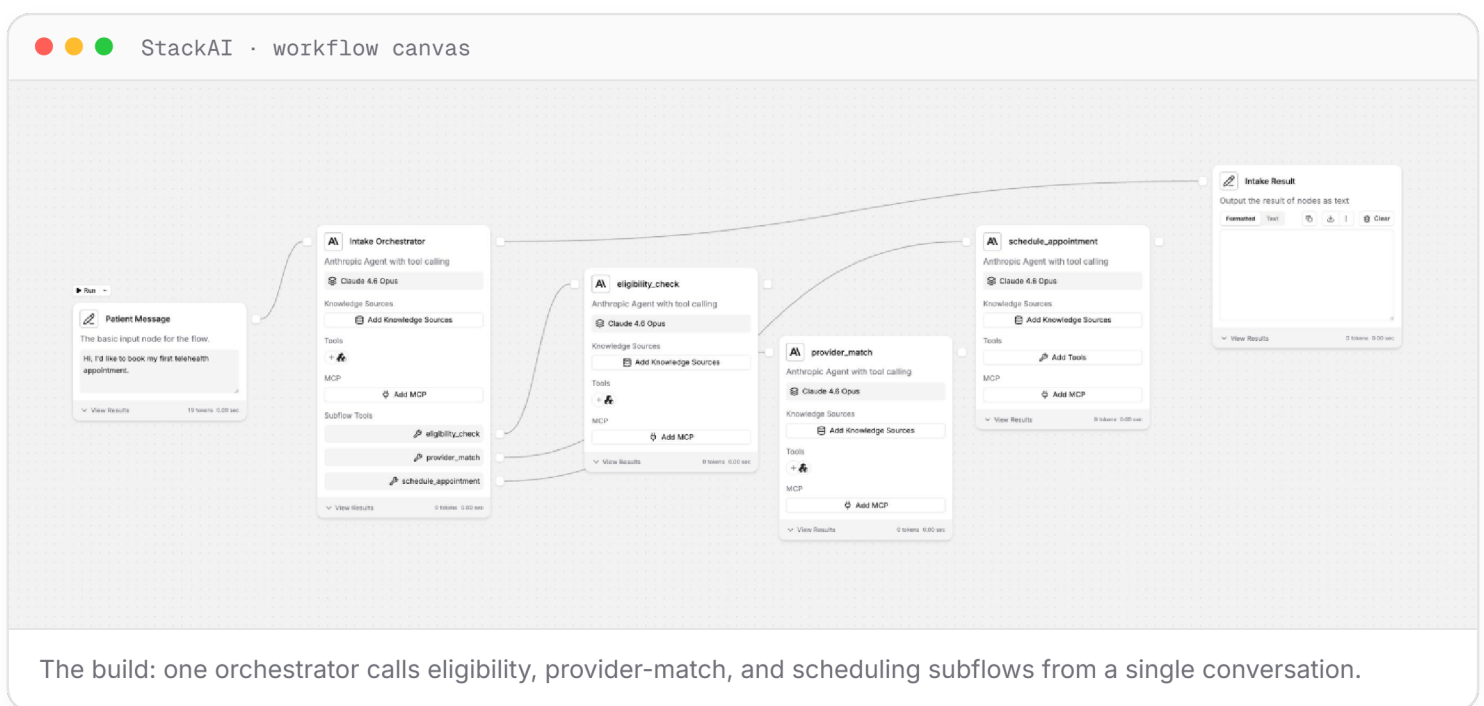
Four of these workflows, built and running on StackAI today.

An intent classifier reads each inbound patient message, detects urgent red flags, and routes to an urgent or routine handler, with every interaction written to a HIPAA audit log before the patient response goes out.



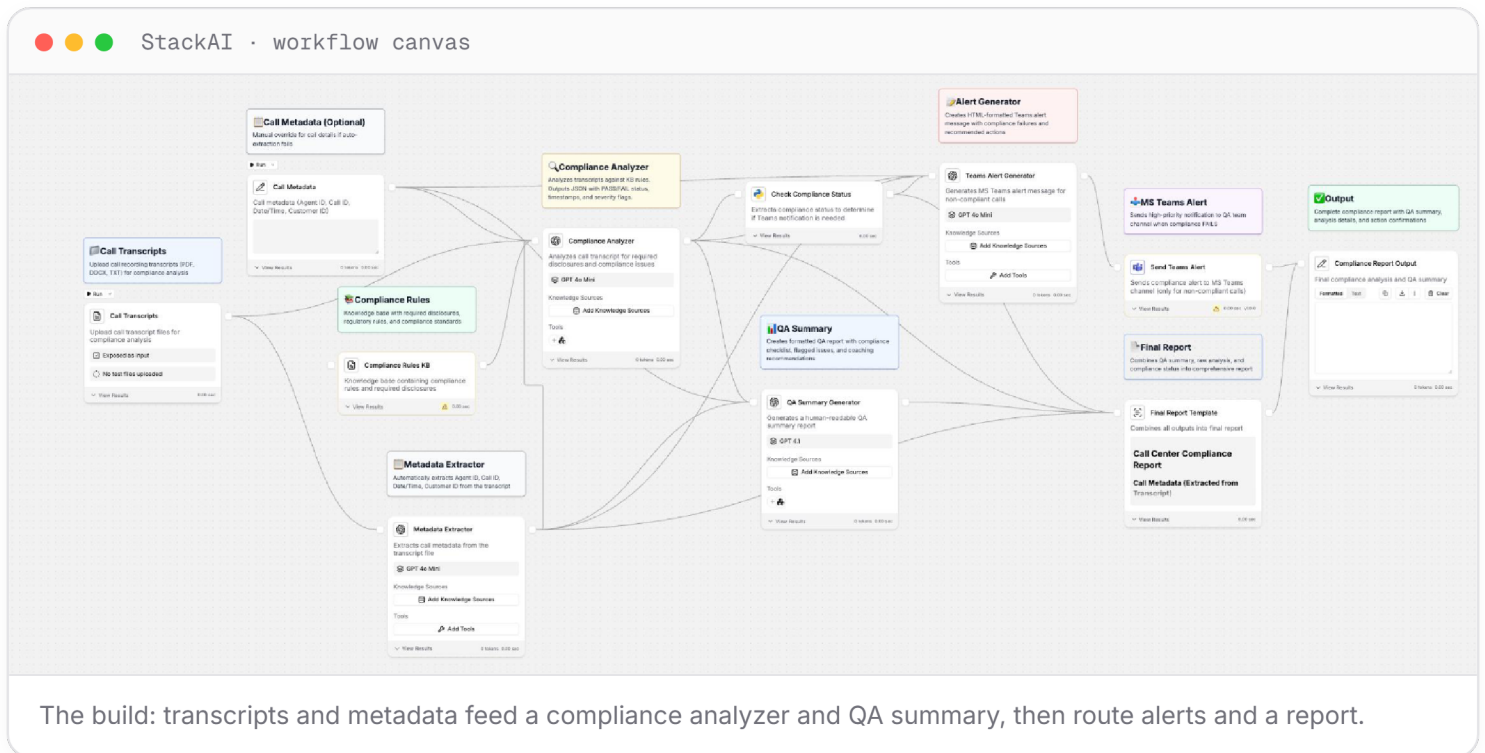
Intelligent intake and eligibility

An intake orchestrator runs eligibility, provider matching, and scheduling as subflow tools, turning a patient's first message into a structured, routed intake.



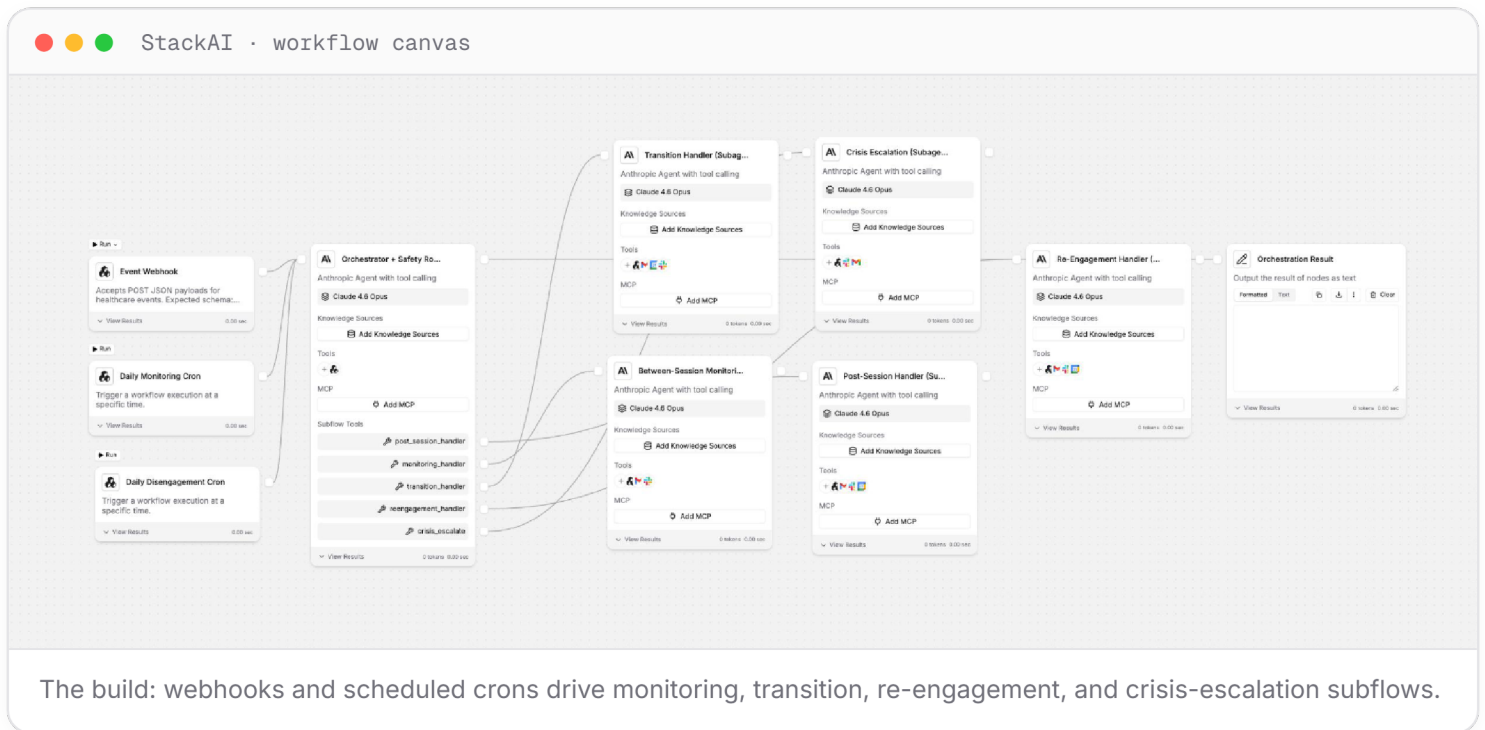
Message and call quality monitoring

Call transcripts are analyzed against a compliance rulebook, producing a QA summary, a pass-or-fail status, a Teams alert on non-compliant calls, and a full compliance report.



Cancellations, retention, and flow orchestration

An orchestrator with a safety router coordinates post-session summaries, between-session monitoring, transitions, re-engagement, and crisis escalation across the patient lifecycle.



Common questions, answered

The questions virtual care leaders raise in evaluation, with honest answers.

Q1 "Are you HIPAA compliant, and is our PHI safe?"

Yes. StackAI is HIPAA compliant and SOC 2 Type II certified, with BAAs signed. For any workflow touching PHI, the platform runs only on the LLM providers we hold BAAs with (Azure OpenAI, Anthropic, and others), backed by DPAs with every model provider. Everything you upload is encrypted and can be de-identified, redacted, or masked before it reaches a model, with a patient-ID key held outside the platform for the rare cases that need re-identification. Nothing you upload appears in logs or trains any model, and under the BAA the compliance liability is ours.

Q2 "We run on a telehealth platform plus a separate EHR. Can you integrate?"

This is exactly the environment the platform is built for. StackAI is an orchestration layer that connects across your SaaS stack (telehealth, EHR, CRM, engagement, billing) rather than assuming one EHR is the center of the world. Most of the five use cases above do not require deep read/write access to the system of record at all: they run against your communication channels and knowledge bases. Where EHR data is needed, we use targeted FHIR calls or an existing integration rather than a full custom build.

Q3 "How do we stop it from getting a clinical answer wrong?"

Three layers. Agents answer only from approved sources and cite them, and can be instructed not to respond unless confident. A second evaluator model grades outputs against your rubric. And every clinical reply ships with a human review-and-approve step, so a rep or provider signs off before anything sends. Every decision is logged with its reasoning, so you always know why the agent said what it said.

Q4 "Why not just use Copilot or build it ourselves?"

Copilot works inside the Microsoft world but struggles the moment you need patient-messaging logic, non-Microsoft data, custom interfaces, or per-state configuration, and it usually needs a dedicated team to stand up. Building from scratch means reinventing model fallback, SSO, audit logs, and governance before you write a single workflow. StackAI ships all of that, plus a weekly build-and-support session with an AI strategist and a forward-deployed engineer, for as long as you are a customer.

Q5 "How is it priced, and how do we know it pays off?"

A predictable monthly subscription, not usage-based. Engagements start with a three-month proof of value: up to three production workflows, three builder seats, all model tokens included, and the strategist plus engineer. During the POV we measure ROI on your own numbers so the business case is proven before any annual commitment.

How these use cases compound

Each use case is valuable alone, but the leverage is in the compounding.

- 01 **Quality and compliance monitoring** is the foundation that lets you deploy other agents confidently in a regulated, multi-state environment.
- 02 **The messaging stack and the voice or chat assistant** absorb inbound volume without growing the team, keeping cost per patient in check as the panel scales and feeding a richer signal to the monitoring layer.
- 03 **Better intake** means more patients enter care with clean data, correct insurance, and appropriate provider matches, which reduces downstream billing issues and improves show rates.
- 04 **Cancellations and flow orchestration** keep patients engaged between visits, which improves retention and lifetime value, and feeds the next cycle of clean intake, clean documentation, and clean compliance.

Together these address the core challenge of a scale-up: growing the panel from 10,000 to 100,000 without growing the team from 100 to 1,000.

The unit economics case

METRIC	BEFORE → AFTER AGENTS		IMPACT
QA audit coverage	before 2 to 5 percent	after Near-total	Continuous, audit-ready trail
Cost per active patient / month	before \$45 to \$80	after \$30 to \$55	25 to 35 percent reduction
Ticket deflection rate	before 0 to 15 percent	after 40 to 60 percent	Support capacity freed
Intake completion rate	before 65 to 75 percent	after 80 to 90 percent	15 to 20 points more entering care
Time-to-first-appointment	before 5 to 10 days	after 1 to 3 days	Reduced early-stage churn
Provider admin time / session	before 12 to 18 min	after 5 to 10 min	More sessions per provider per day

◆ REFERENCE DEPLOYMENT

A national virtual-first primary care provider deployed StackAI agents across patient communications, intake, eligibility, and documentation review. Across the agent fleet the deployment has logged over 475,000 hours of clinical and operational time saved, roughly 230 FTE-years, without proportional headcount growth.

475,000+

hours saved · ~230
FTE-years



[Read the LifeMD case study ↗](#)

Why StackAI for virtual care

StackAI is the AI transformation platform for healthcare: a HIPAA- and GDPR-compliant operating system for enterprise agentic AI, built by MIT PhDs and deployed across more than 200 organizations in regulated industries.



Security and compliance

HIPAA compliant, SOC 2 Type II, BAAs in place.
Secure deployment on your cloud or StackAI's.
Multi-state configurability for jurisdiction-specific rules.



Universal connectivity

500+ integrations including leading EHRs, telehealth platforms, CRMs, billing systems, and engagement tools, the SaaS stack virtual care runs on.



Unstructured data

Handles PDFs, scans, images, and audio natively, critical for transferred records, intake forms, and insurance cards.



No-code builder

Clinical ops, RCM, and compliance experts build and modify agents in days. New intake questions and rules do not need an engineering ticket.



Human-in-the-loop everywhere

Review, approval, and escalation on every clinical and revenue workflow, with full audit logging for regulatory readiness.



Model-agnostic and governed

No LLM lock-in, automatic fallback, plus access control, audit logs, and role-based permissions built in.

The bottom line

Digital health was built on the promise that technology could make care more accessible, convenient, and affordable. AI agents are the next layer of that promise: they remove the operational friction that keeps the care team from operating at the top of their capability.

The companies that figure this out first will serve more patients, in more states, across more conditions, with the same team.

Want to see these use cases in action?

Book a working session with the StackAI healthcare team to scope a three-month proof of value tailored to your panel size, payer mix, and tech stack. We will bring deployment patterns from comparable virtual care customers and an opinionated view on where your unit economics will move fastest.