

HIPAA-compliant AI IVR: 94-minute hold time down to 22 minutes.

Senior AI Solutions Architect (contract) · 8 months kickoff to production · Regional healthcare provider

WAIT TIME BEFORE	WAIT TIME AFTER	REDUCTION	SHIP TIME
94 min average hold	22 min post-deployment	77% no added clinical headcount	8 mo kickoff to production

THE SETUP

A regional healthcare provider’s urgent-care call line had drifted into operational crisis. Average hold time had climbed to 94 minutes. Abandonment was steep, clinical ops was taking internal heat, and the incumbent IVR vendor couldn’t move the number. Any replacement had to work inside the provider’s EMR, handle PHI correctly from day one, and pass compliance review on the first pass. The team didn’t have budget for a second attempt.

THE ARCHITECTURE

LLM IVR **Amazon Bedrock, intent + triage routing.** Replaced the rules-based system. Intent classification and triage handled per call. PHI-aware prompts kept identifying data out of the model boundary where possible.

EMR I/O **Bidirectional integration, idempotent + audited.** Read for patient context, write for triage notes and call disposition. Sync was idempotent and audited end-to-end.

Compliance **PHI handling, audit trails, encryption, IAM.** Designed in from the start. Not bolted on. Passed internal compliance review on the first pass.

Observability **Latency, errors, containment, eval loops.** Wired to Slack and PagerDuty so failures surfaced fast. Clinical-accuracy eval loop kept the model honest in production.

Delivery **8 months end-to-end, weekly cycle with clinical ops.** Shipped on AWS. Clinical ops and compliance embedded in the weekly cycle — not as a final-stage gate.

OUTCOME

HOLD TIME	COMPLIANCE REVIEW	CLINICAL HEADCOUNT
22 min down from 94	First pass no rework	Zero added same team operating

What this proves. Healthcare AI ships when compliance, clinical ops, and engineering work the same weekly cycle — not when compliance is a final gate. Same pattern applies to any regulated workload.