

AI Automation & RAG **Baseline**

The minimum architecture, safeguards, and readiness checks for production AI in healthcare.

INSIDE

- ✓ The 8-stage architecture behind production-grade RAG
- ✓ The safety & compliance checklist for systems that touch PHI
- ✓ A pilot roadmap, success metrics, and a readiness self-assessment

01 · THE OPPORTUNITY

Your documentation is an untapped asset

The problem

Healthcare teams are drowning in paperwork. A time-and-motion study published in the *Annals of Internal Medicine* found that physicians spent nearly **two hours on EHR and desk work for every hour of direct patient care**. The same pattern shows up for therapists, nurses, and front-office staff: notes, plans of care, referral summaries, prior authorization packets, and discharge paperwork.

Meanwhile, most organizations already hold years of high-quality documentation: assessments, treatment plans, progress notes, discharge summaries, policies, and payer guidelines. It sits in PDFs and EHR archives, and nobody can search it.

Retrieval-augmented generation (RAG) connects AI to that archive. The AI doesn't write from generic internet knowledge. It retrieves your organization's own relevant documentation and uses it to draft new work. A qualified professional then reviews and approves the draft.

Three tools, three different jobs

	Workflow automation	RAG systems	AI agents
What it does	Moves data between systems using fixed rules	Finds relevant information in your documents and drafts from it	Carries out multi-step tasks using several tools
Best for	Predictable, rule-based handoffs	Questions and drafts that depend on your own knowledge	Tasks that need gathering, deciding, and assembling
Healthcare example	Intake form → EHR record → appointment reminder	Drafting a progress note from the patient record and similar past documentation	Assembling a prior authorization packet for staff review

Rule of thumb: Automate what's predictable. Use RAG where the answer lives in your documents. Use agents only when a task genuinely needs several steps, and always with a human checkpoint.

Where to start: the 4–point filter

- High volume.** It happens dozens or hundreds of times a week.
- Repeatable structure.** The output follows a known format or template.
- Existing examples.** You have years of past documents showing what “good” looks like.
- A reviewer is already in place.** Someone qualified already checks or signs this work today.

Common candidates

- Clinical documentation drafts
 - Referral and intake summaries
 - Discharge summaries
- Prior authorization packet preparation
 - Denial appeal letter drafts
 - Staff Q&A on internal policies and procedures

02 · THE ARCHITECTURE

The 8-stage RAG blueprint

From raw archive to approved document. Use these stages as a checklist for any AI build or vendor.



🔄 **Feedback loop:** Stage 8 returns only human-approved documents to Stage 2 for re-indexing.

- 01

Collect: Build the knowledge source. Gather the historical documents the system will learn from: notes, assessments, treatment plans, discharge summaries, protocols, and payer guidelines.

Why it matters: *Draft quality can never exceed the quality of this archive.*
- 02

Index: Chunk, embed, and store. Split documents into meaningful sections, convert them to searchable vectors, and store them in a vector database with metadata such as document type, date, specialty, and access level.

Why it matters: *Good chunking decides whether the system finds the right paragraph or an irrelevant one.*
- 03

Contextualize: Pull the current case. Securely retrieve the current patient's record from the EHR (demographics, diagnoses, history, prior outcomes) and turn it into a precise search query.

Why it matters: *Without case context, retrieval returns generic results.*
- 04

Retrieve: Find relevant precedent, with fallbacks. Search the knowledge base for the most relevant prior documentation. Production systems use layered retrieval: meaning-based search first, then keyword search, then direct section lookup. If every layer fails, the system returns a clear "not found" message and never a guess.

Why it matters: *A system that fails silently is a system nobody trusts.*
- 05

Draft: Generate grounded output. The language model receives the patient context, the retrieved precedent, and your guidelines, then drafts the document in your required format. Every key statement links back to its source.

Why it matters: *Citations let reviewers verify the draft in seconds.*
- 06

Guard: Automated review before any human sees it. A two-layer check runs on every draft. Rule-based checks cover required sections, measurable goals, and documentation standards. An AI reviewer checks against your protocols. Drafts that pass go to the reviewer's dashboard. Drafts that fail go to an incident queue for investigation.

Why it matters: *It catches problems early and creates an audit trail.*
- 07

Approve: A professional reviews, edits, and signs. A qualified clinician or staff member reviews the draft, edits it, and approves or rejects it. Only approved documents are saved to the EHR.

Why it matters: *AI drafts, people decide. Accountability stays where it belongs.*
- 08

Learn: Close the feedback loop. Approvals, edits, and rejections are captured. Only final, human-approved documents return to the knowledge base, which is then re-indexed.

Why it matters: *The system improves with use, and it never learns from its own unreviewed output.*

DEMO VS. PRODUCTION

5 things that separate a demo from a production system

1 Fallbacks

It never fails silently.

2 Guardrails

Drafts are checked before a human sees them.

3 Human approval

Nothing reaches the record unsigned.

4 Citations

Every suggestion is traceable.

5 Gated learning

It learns only from approved work.

03 · SAFETY & COMPLIANCE

The safety & compliance baseline

Any AI system that touches protected health information (PHI) should meet this baseline **before** the pilot starts, not after.

<p>DATA & VENDORS</p> <ul style="list-style-type: none">☐ A signed Business Associate Agreement (BAA) with every vendor that stores, processes, or transmits PHI, including the AI model provider, vector database, and hosting.☐ Contractual confirmation that your data is not used to train vendor models.☐ Encryption in transit and at rest across every stage of the pipeline.☐ A defined data retention and deletion policy for documents, embeddings, prompts, and logs.	<p>PRIVACY & ACCESS</p> <ul style="list-style-type: none">☐ Cross-patient protection. RAG retrieves <i>other patients'</i> past records as precedent. De-identify the historical archive where possible, and apply the minimum-necessary principle to everything else.☐ Permission-aware retrieval. Users only retrieve documents their role already allows them to see. Access controls apply inside the AI, not just around it.☐ Role-based access to the review dashboard, incident queue, and admin tools.
<p>OUTPUT SAFETY</p> <ul style="list-style-type: none">☐ Two-layer guardrails. Use rule-based checks plus an AI reviewer. An AI checking an AI is not enough on its own.☐ Citations on every draft, so reviewers can independently verify the basis for each suggestion.☐ An incident queue with a named owner and a response process for flagged drafts.☐ A test set. A fixed set of real cases with known-good answers, re-run after every change to the system.	<p>ACCOUNTABILITY</p> <ul style="list-style-type: none">☐ Human approval is required before anything is saved to the record.☐ Full audit logs recording who generated, reviewed, edited, approved, or rejected each draft, and when.☐ Monitoring for accuracy, drift, and adoption, reviewed on a set schedule.

What AI should not do

- ✗ Make diagnoses or treatment decisions on its own
- ✗ Save anything to the patient record without human approval
- ✗ Present suggestions without showing their source
- ✗ Learn from drafts that were never reviewed by a person

Regulatory note: Systems that suggest clinical next steps may fall under FDA guidance on clinical decision support software. Designing for independent clinician review, with sources shown and the human making the final decision, is central to that guidance. Review your specific use case with qualified legal counsel.

04 · IMPLEMENTATION

From pilot to production

The 5-phase roadmap

Phase	Typical duration	What happens	Exit criteria
1. Discovery	1–2 weeks	Pick one use case using the 4-point filter. Map the current workflow and time spent.	One use case, one owner, baseline metrics captured
2. Data audit	1–2 weeks	Inventory historical documents: volume, format, quality, years covered, PHI handling.	Clean, approved knowledge source; BAAs signed
3. Pilot	4–6 weeks	Build stages 1–7 for a single document type with 2–3 reviewers.	Reviewers using it in real work
4. Evaluate	2 weeks	Measure against baseline and the test set. Review every incident.	Targets met; no unresolved safety issues
5. Scale	Ongoing	Add reviewers, document types, and the stage 8 feedback loop. Monitor continuously.	Regular monitoring and review in place

Durations are typical ranges and depend on data readiness and integration complexity.

What to measure

Metric	What it tells you
Documentation time per case	The headline efficiency gain, compared with baseline
Draft acceptance rate	% of drafts approved with no or minor edits
Edit depth	How much reviewers change each draft (falling over time = learning)
Rejection rate	% of drafts reviewers discard entirely
Guardrail incident rate	% of drafts flagged before human review
Retrieval hit rate	How often the right precedent is found on the first layer
Time to sign	Elapsed time from encounter to approved documentation
Adoption	% of eligible staff actively using the system

05 · READINESS

Readiness self-assessment

Answer honestly. Score 1 point for each “yes,” then find your band below.

01	We can name one high-volume documentation task that costs significant staff time.	☐
02	We have at least 2–3 years of digital historical documentation for that task.	☐
03	Those documents are reasonably consistent in format and quality.	☐

04	Our EHR or core system allows API or export-based integration.	☐
05	We know which vendors would touch PHI and can sign BAAs with them.	☐
06	A qualified professional already reviews or signs this work today.	☐
07	We have a clinical or operational champion who will own the pilot.	☐
08	We can measure how long this task takes today.	☐
09	Leadership has agreed on what “success” means for a pilot.	☐
10	We have a process for reporting and reviewing errors or incidents.	☐

8–10 Pilot-ready
You can likely start a scoped pilot within weeks.

5–7 Foundation needed
Close the gaps in data, integration, or ownership first.

0–4 Start with strategy
Begin with discovery and a data audit before building anything.

NEXT STEP

Book an AI Readiness Assessment

In a readiness assessment, our team will:

- ✓ Review your self-assessment score and identify the gaps
- ✓ Help you pick the highest-value first use case
- ✓ Outline what a scoped, compliant pilot would look like for your organization

Book your assessment →

Contact: support@cloudsstream.com

CloudsStream · AI Automation & RAG Systems · Business Growth & Marketing. This baseline is for general informational purposes only and does not constitute legal, regulatory, or clinical advice. Consult qualified counsel regarding HIPAA, FDA, and state requirements applicable to your organization.