UW HEALTH

An estimated 42,000+ clinician hours returned by Epic-integrated AI that waits for clinician sign-off

I audited 50+ AI use cases at UW Health, won ownership of the system's AI/ML roadmap and, with a cross-functional team of 30+, launched GenAI agents inside Epic that draft replies to patient messages, nurse shift summaries and routine documentation but never send them. In the first year of a 500-clinician pilot they returned an estimated 42,000+ clinician hours, at 93% clinician-judged task success and a 2.5% clinician-reviewed hallucination rate, and the use cases then scaled to the whole organization.

ROLE
Senior Technical Product Manager, AI/ML

WHEN
2023 to 2025

TEAM
Cross-functional team of 30+, including an internal development team of 8+ I led in an agile format

SCOPE
A 500-clinician pilot at Wisconsin's largest health system, then organization-wide

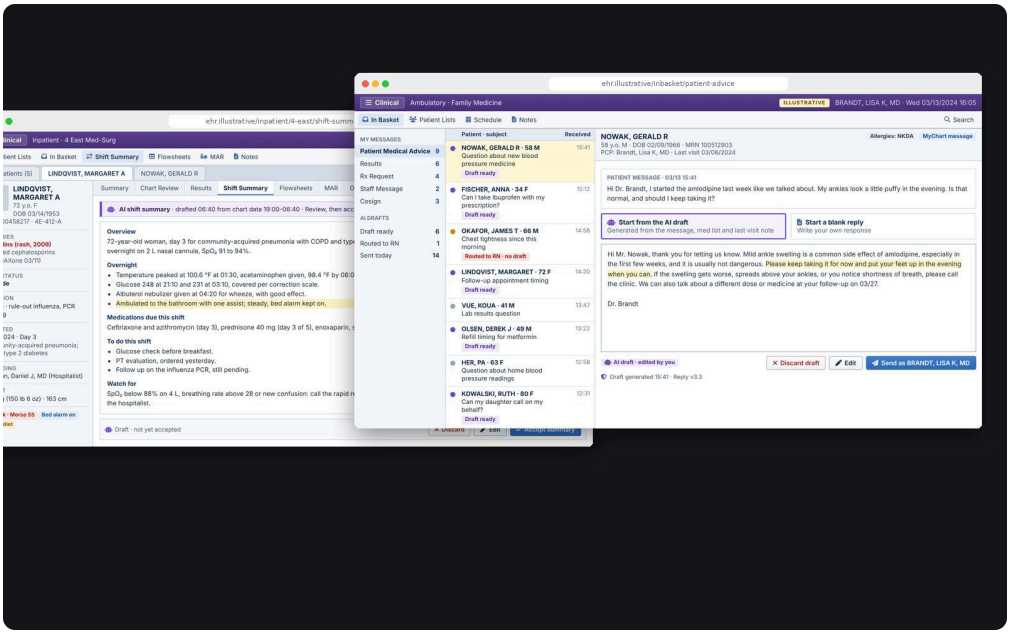
I OWNED
The AI/ML roadmap, the agent products, their evals, guardrails and governance analytics

42,000+
estimated clinician hours returned in the first year, the capacity of 20+ full-time clinicians

93%
task success, judged by clinicians

2.5%
clinician-reviewed hallucination rate

>90%
engagement among the 500 pilot clinicians, the highest of any piloted UW Health AI tool



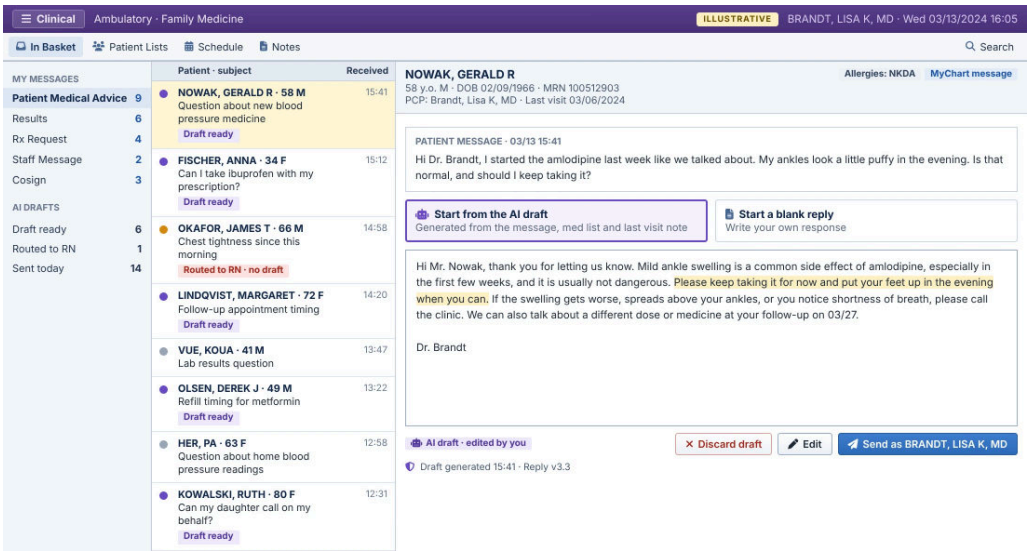
Overview

I was asked to audit AI across UW Health, Wisconsin's largest health system. I collected more than 50 use cases, scored them on one five-criteria rubric, took the ranking to senior leadership and executive stakeholders, and came out owning the AI/ML roadmap. The audit showed clinicians spending up to 40% of a shift on messaging, handoffs, documentation and queues: work around care rather than care itself.

From there I led a cross-functional team of 30+, including an internal development team of 8+ that I ran in an agile format, to build AI clinician workflow assistants on Azure OpenAI GPT-4 integrated with Epic: draft replies to patient messages in the In Basket, nurse shift summaries, and operational agents that predicted patients' length of stay and generated routine documentation.

Also delivered: Best Buy Health remote patient monitoring, an Epic Willow Ambulatory transition for two new pharmacy locations, and UiPath robotic process automation for finance and administrative teams.

The rule throughout: nothing reaches a patient's chart, or a patient, until a clinician sends or signs it.



Illustrative.

Problem and research

The audit started with a list of requests and a question I asked of every one of them: if this works, who gets time back, and if it's wrong, what breaks? I scored each use case on the same five criteria so that a request from a nursing unit and a request from finance could be compared side by side, and so that anyone who disagreed with the ranking could see exactly why it came out that way.

CRITERION	WHAT IT TOLD US
Clinician time today	How much of a shift the task takes. Time returned to care was the outcome clinicians would notice first.
Risk if it's wrong	A wrong draft a clinician reads first is a nuisance. A wrong message to a patient or a wrong entry in the chart is a safety event. This set how much human review each use case needed.
Fits the workflow	Whether it could live inside Epic and the tools around it. Anything outside the workflow would depend on people opening one more app.
Data ready	Whether the data the task needs was available and allowed for that use.
Rules that apply	Privacy, clinical decision support and nondiscrimination rules differ by use case and set the work needed before launch.

Who we built it for

1

Clinicians answering patient messages

Portal messages arrive all day and most need a short, careful reply. Writing each one from a blank page at the end of a clinic day is where a lot of time went.

2

Nurses at shift change

They hand over every patient twice a day, pulling the shift's vitals, medications and notes together by hand. A good summary is safety-critical, and it is also one of the most repetitive jobs on a unit.

3

Finance and administrative teams

Billing and back-office queues that follow the same steps every time. They needed a simple way to hand those steps to automation without writing code.

The first roadmap took the use cases that combined a lot of clinician time with a risk we could control through review: replies to patient messages in the In Basket, nurse shift summaries and routine documentation, plus length-of-stay prediction for operations and rules-based automation for finance and administration. Ideas with a high cost of being wrong and no clear place in the workflow waited, however popular they were.

What we built

Clinicians send every message

Agents draft replies to patient messages, shift summaries and routine documentation. A clinician reviews, edits and sends or signs each one, and nothing is written to the chart or sent to a patient before that.

Not chosen: Letting agents send routine replies on their own. It would have saved a little more time and turned every error into a message a patient reads.

Start from the draft, or start blank

Every In Basket reply offered both: start from the generated draft, or open a blank reply. Clinicians stayed in control of their own voice, and the choice itself told us when a draft wasn't worth using.

Not chosen: Pre-filling every reply with the draft, which makes it too easy to send without reading.

Work inside Epic, not next to it

Clinicians already live in Epic, so the drafts appear in the In Basket and the shift tools where they work.

Not chosen: A separate AI app. Faster to build, but adoption would depend on clinicians opening one more tool.

The right tool for each job

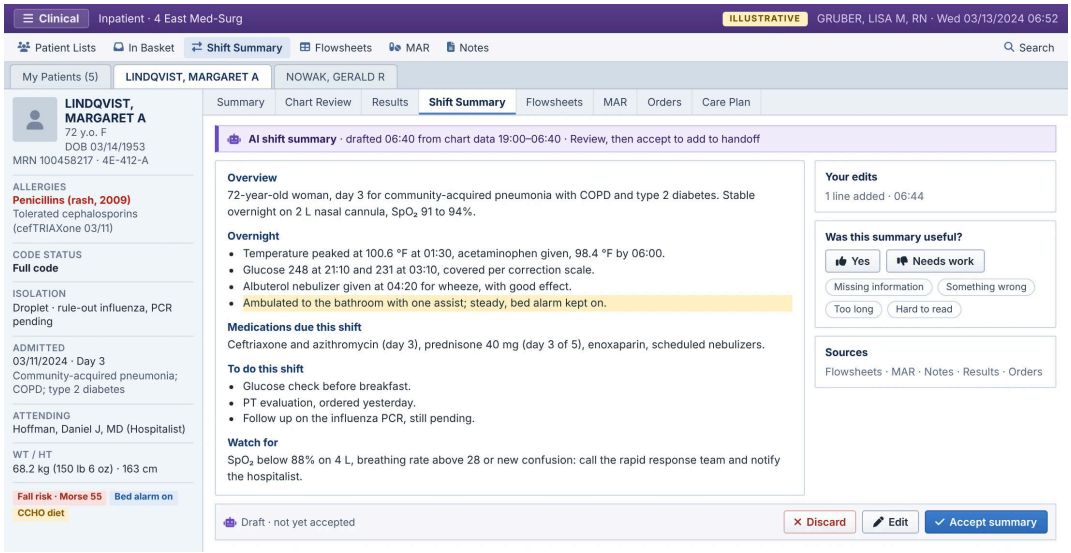
GenAI agents on Azure OpenAI GPT-4 for work that needs language. UiPath robotic process automation, with a simple interface for finance and administrative staff, for tasks that follow fixed steps.

Not chosen: A language model on everything. Fixed business rules don't need one, and rules are easier to audit.

Least privilege and an audit trail

Each agent runs under its own identity, reads only the data its task needs, and logs every tool call, so any action can be traced afterward.

Not chosen: Broad read access for convenience, which would have made the HIPAA review harder and the audit trail meaningless.

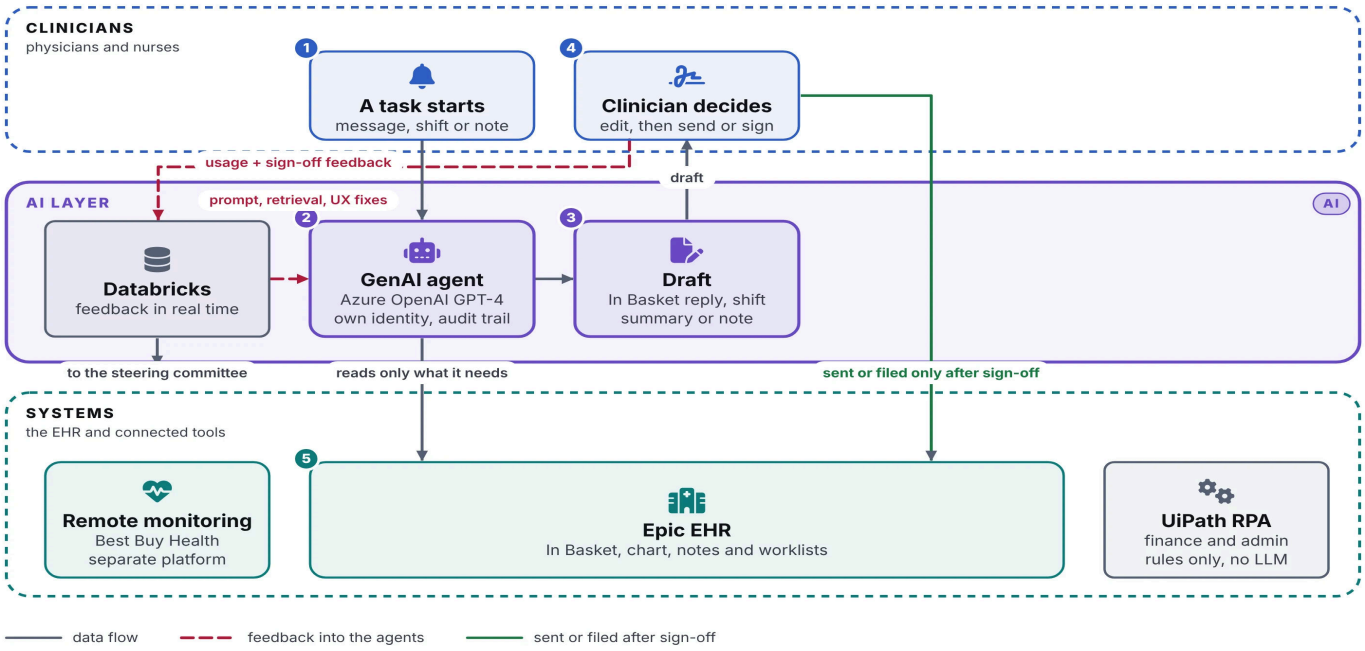


Illustrative.

Replies, shift summaries and documentation shared one pattern: the agent gathers the chart data a clinician would otherwise collect by hand, drafts the text, and waits for a person. The operational agents built with the development team followed the same rule, predicting each patient's length of stay and drafting routine documentation for the care team to review.

Two programs ran alongside the agents. Remote patient monitoring used Best Buy Health as a separate platform: patients received home kits so they didn't need to stay in the hospital, and their readings came to the care team through that platform rather than through Epic. And I led an Epic Willow Ambulatory transition for two new pharmacy locations: a full platform build in Epic and all the data migration onto it.

How it works



Illustrative.

Work starts in one of three ways: a patient message arrives, a shift is ending, or a routine note is due. The agent, built on Azure OpenAI GPT-4, reads only the data that task needs from Epic, under its own identity, and logs each step.

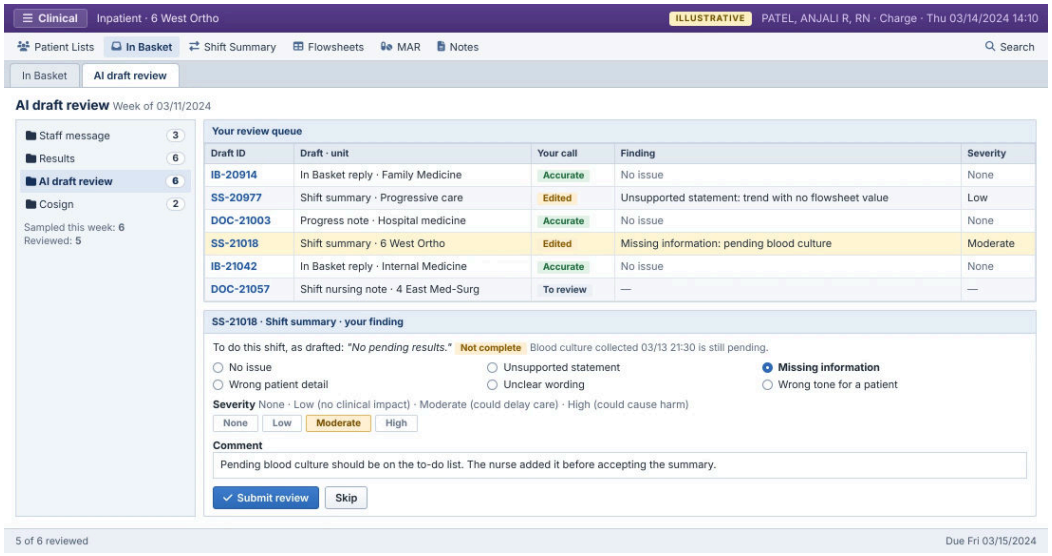
It writes a draft and hands it to a clinician. For a patient message, the clinician starts from the draft or from a blank reply, edits it, and sends it under their own name. For a shift summary, nothing is saved to the handoff until the nurse reviews and accepts it; for a note, the clinician reviews, edits and signs. Messages that describe urgent symptoms are routed to a nurse instead of being drafted. Finance and administrative tasks run separately on UiPath, with fixed rules and no language model.

What clinicians did with each draft, and what they told us about it, streamed into a real-time Databricks pipeline. For the product work, it moved our iteration cycle from months to days and showed how each prompt change affected usage. For the executive AI steering committee, it was the governance layer: legal and compliance measures such as how often a reply went out straight from the generated draft without edits, next to adoption, quality and time returned.

Evaluation and guardrails

The checks asked three questions: did the agent do the task it was given, was every claim supported by the chart, and did it stop and hand off when it should? Clinicians did the judging, not the model grading itself, and the checks kept running after launch.

WHAT WE CHECKED	HOW	RESULT
Task success	Clinicians judged whether each sampled draft did its task without a critical edit	**93%**
Hallucination	Clinicians reviewed sampled drafts for any statement the chart didn't support	**2.5%**
Escalation	Test cases where the agent should not draft and should hand off instead, such as urgent symptoms	**Launch gate**
Security	Prompt-injection and patient-data leakage tests before launch	**Launch gate**
Fairness	Results compared across patient groups: age, sex, race and ethnicity, preferred language and payer	**Launch gate**
Governance	How often drafts were sent without edits, edit rates and adoption, reported to the steering committee	**Tracked throughout**



Illustrative.

The rules we designed toward

RESPONSIBLE AI

HIPAA
Worked through with the legal and compliance team: patient data stays within agreements that cover it, each agent sees the minimum it needs, and every access is logged.

FDA clinical decision support guidance
Clinicians see and edit every draft before it goes anywhere, and the agent never acts on its own.

ACA Section 1557
Tools that support patient care have to be checked for discrimination risk, so performance was reviewed across patient groups, not only overall.

NIST AI Risk Management Framework
Its generative AI profile names risks like made-up answers and data leakage. Our tests map onto them.

Prompts got the same treatment. I co-authored a study of GPT-4 drafting replies to patient messages inside the EHR across provider clinics. It used a pre-post design to compare crowd-sourced manual prompts with a semi-automated prompt design approach, and was presented as a poster at SAIL 2024.

Results, and how to read them

NUMBER	WHAT IT IS	HOW TO READ IT
An estimated 42,000+ clinician hours returned in the first year	Time from the tasks the agents took over in the 500-clinician pilot, which had taken up to 40% of a shift across messaging, handoffs, documentation and queues	About the capacity of 20+ full-time clinicians. It is time returned on automated tasks, not a stopwatch study of every shift.
93% task success	The share of sampled drafts clinicians judged as doing their task without a critical edit	Judged by clinicians, not by the model grading itself.
2.5% hallucination rate	Counted per reviewed draft: the share of reviewed drafts with a claim the chart didn't support	Each one went into the feedback pipeline, and review before sending meant none reached a patient or the chart unchecked.
More than 90% engagement among the 500 pilot clinicians	Use of the pilot tool among the clinicians it was rolled out to	The highest of any piloted UW Health AI tool. It shows adoption, and the hours and quality numbers show whether it helped.
Organization-wide	Where the use cases went after the pilot year	The pilot results were what justified scaling them across the system.
Months to days	Time to ship a prompt, retrieval or screen fix	What the Databricks pipeline changed. It is why the agents kept improving after launch.

Rollout and governance

1
Audit
Collect and score 50+ AI use cases across the system on one rubric

2
Roadmap
Present the ranking to senior leadership and executives, and take ownership of the AI/ML roadmap

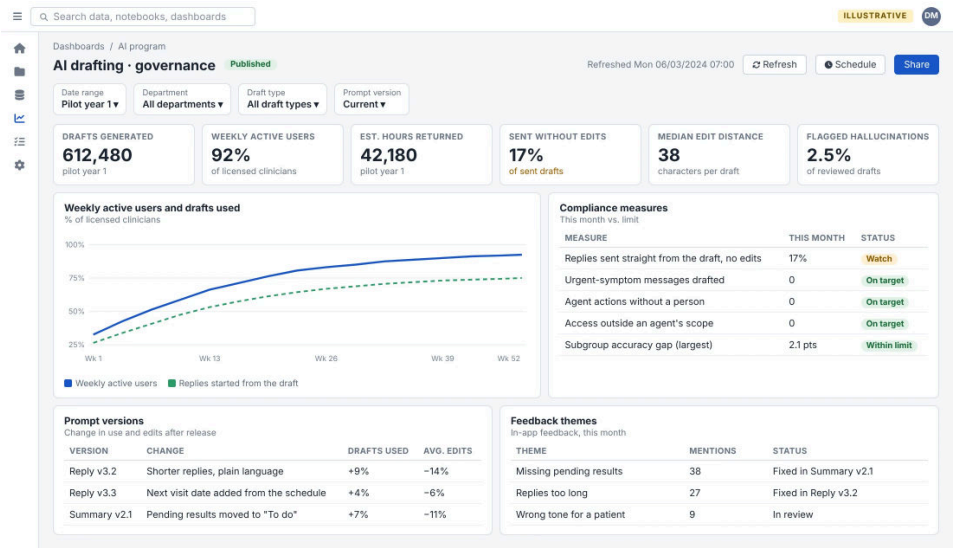
3
Build
Build the agents with an internal development team of 8+ in agile sprints, with the guardrails on the same ticket as the feature

4
Pilot
Roll out to 500 clinicians and judge it on time returned, task success and hallucination rate

5
Govern
Report adoption, quality and compliance measures from Databricks to the AI steering committee

6
Scale
Take the use cases organization-wide after the pilot year

Agreement across a matrixed organization came less from the ranking itself than from showing each group what it meant for their own work: which tasks would come off their plate, what they would review, and what would never happen without them. I co-led the AI steering committee, where we presented use cases to senior leadership, shared progress, earned their buy-in and proposed new use cases. And I created an AI education program for UW Health that covered what AI is and how it works, so clinicians, operations and leadership started from the same foundations.



Illustrative.

Looking back

- Define every number before launch. Engagement and hours returned are easy to overstate, and defining them up front is what lets you defend them later.
- Let people choose between the draft and a blank page. The choice kept clinicians in control and told us which drafts weren't earning their place.
- Shorten the time between fixes. Going from months to days between fixes is what let the agents keep improving after launch.
- Pick the tool before the model. Finance and admin work ran on rules with UiPath, and not every problem on the roadmap needed a language model.
- Plan the rollout with the pilot. The steering committee and the education program are what let a pilot become a system-wide rollout.
- What carries over. The draft-or-blank choice, governance metrics such as how often drafts are sent unedited, and a feedback pipeline that ships fixes in days apply to any regulated workflow where AI drafts and people decide.