When execution becomes cheap.

Speed, judgment and reliability in critical domains

Akarsh Verma
Director of Product Management, XCaliber Health



Akarsh Verma

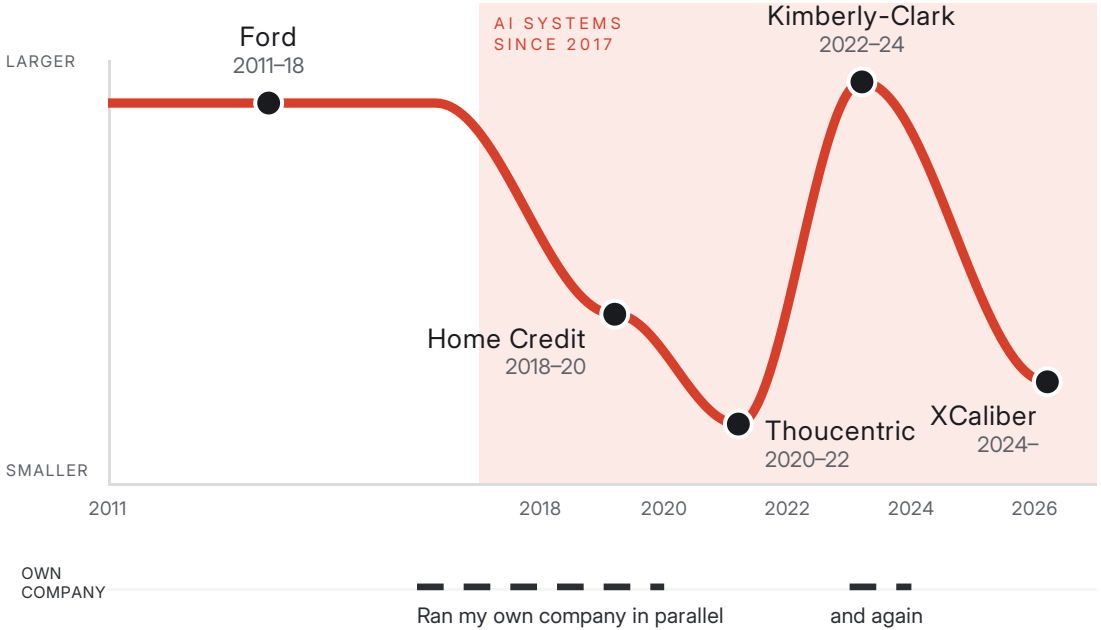
Engineer & Architect
Procrastinating YouTuber
Loves Technology, Books, Comics & Coffee
Evolving Father & Hopeful Husband

Director of Product Management
XCaliber Health

linkedin.com/in/
akarshverma

01 / MY CAREER, BY COMPANY SIZE

Big. Small. Smaller.
Very big. Small again.



One emergency visit.

01

ED discharge note arrives

A child is seen in the emergency department for an asthma exacerbation.

One emergency visit. Sixteen steps.



The model is 4 of 16 steps. The rest has to be right, every time.

03 / CROSSING DOMAINS

New domain.
Same two questions.

01 Which principles transpose?

02 What role should AI play here? Sometimes none.

BORROWED EXPERTISE

Doctors on the team.

They guide us. I never try to be the expert.

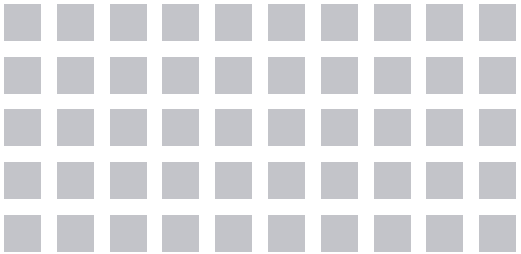
THE HARD PART

Turning clinical expertise into engineering is very hard.

That translation is human. Not AI.

Prototyping is good.
Real feedback is better.

50



inside the building

vs

5



with real patients

Elsewhere they are edge cases. In healthcare, every patient is one.

LLMs are great.
People still need deterministic software.

HEALTHCARE

Discharge summary

SPEED	Reads a 60-page chart in seconds.
SLOP	Carries forward a stopped medication.
RULES DECIDE	Med list reconciled against the EHR.

FINANCE

Credit risk engine

SPEED	Reads payslips and statements in seconds.
SLOP	A decline nobody can explain.
RULES DECIDE	Scorecard decides. Every decline has a reason code.

COST STILL MATTERS

LLM on a 60-page chart: seconds and real money. Purpose-built inference: milliseconds, at scale.

Most "agents" are not agents.

NOT AN AGENT

A prompt on an LLM.

NOT AN AGENT

Fifty conditions in LangChain.

AN AGENT

Owns one job, end to end.

A goal it can check

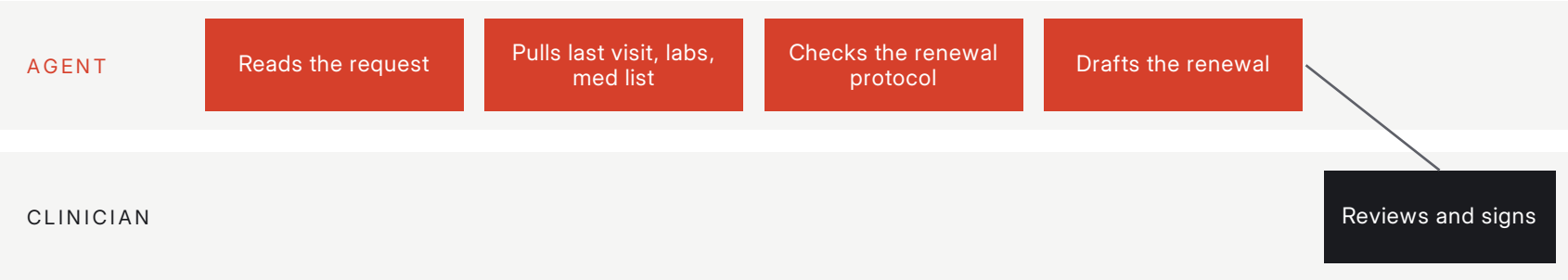
Tools it may use

Limits it cannot cross

A hand-over to a person

One workflow. Clear roles.
Hard boundaries.

EXAMPLE · MEDICATION RENEWAL REQUEST



The agent never: sends a prescription, changes a dose, or gives clinical advice.

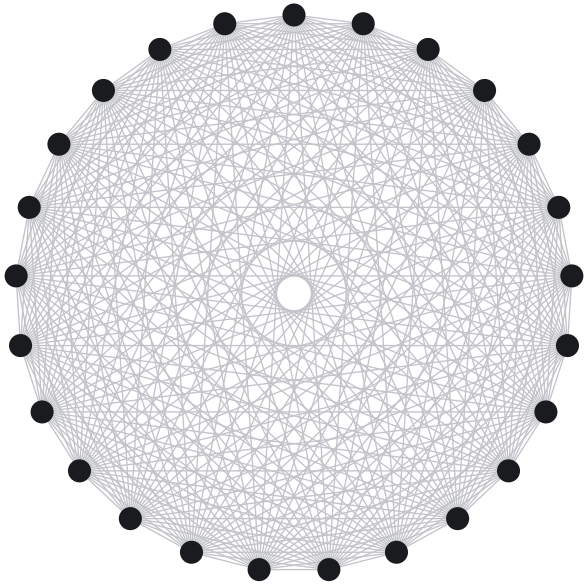
Agent prepares, clinician does

Agent completes, clinician signs

THIS WORKFLOW

Agent completes, clinician audits

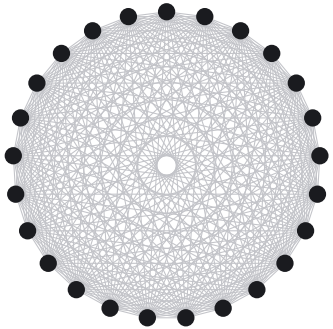
Twenty-five people.



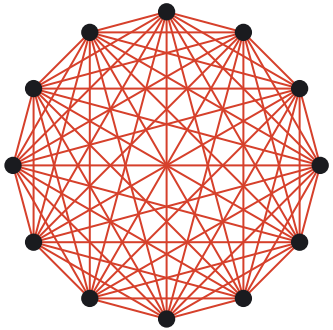
300

lines of communication

Remove the handoffs, and you need fewer people.



25 people
300 lines



~12 people + ~25 agents
66 lines

Go deep

Focused problems, not a wide org chart.

Pay better, hire fewer

Move people to customers

Customer success and FDE. Not support.

We used to have Stack Overflow and midnight oil. Now we buy a bigger Claude plan.

THE QUESTION

Are people still thinking?

Intelligence degradation? My honest answer: yes.

WHAT WE ARE TRYING

One-year internships

Time to actually learn the system.

Roles by problem, not title

Fewer layers. Less lost in translation.

A duty to question

Push back on the AI.

Z is the last letter.
Maybe they have something to teach us.

IF WE GIVE THEM THE RIGHT TOOLS

Accountability

Ownership

Depth

AT HOME

My daughter is six.

Her questions surprise me every day.

What do we teach them, and how?

Execution is cheap.
Judgment is not.
And we may be outsourcing it.

Deterministic where it matters.

The model reads and drafts. Software decides.

Small, flat teams that own problems.

Fewer handoffs. People closer to customers.

Grow judgment on purpose.

Long internships, ownership, the duty to question back.

Is there a good answer? I'm not sure. I would like to hear yours.

Thank you

