

From Hallucinations to Hard Truths

Real Lessons in Production AI



Rashi Agrawal

Head of AI Engineering

Rashi Agrawal

16+ years in Engineering & AI Leadership



Systems Engineer

Accenture

Engineering Leader

Yahoo

Head of AI Engineering

GoodLeap

**Masters in Cloud
Computing**

San Jose State University

Executive Education

Stanford Graduate School of
Business



Hard Truth # 1

You're Solving the Wrong Problem

Hard Truth #1

You're Solving the Wrong Problem

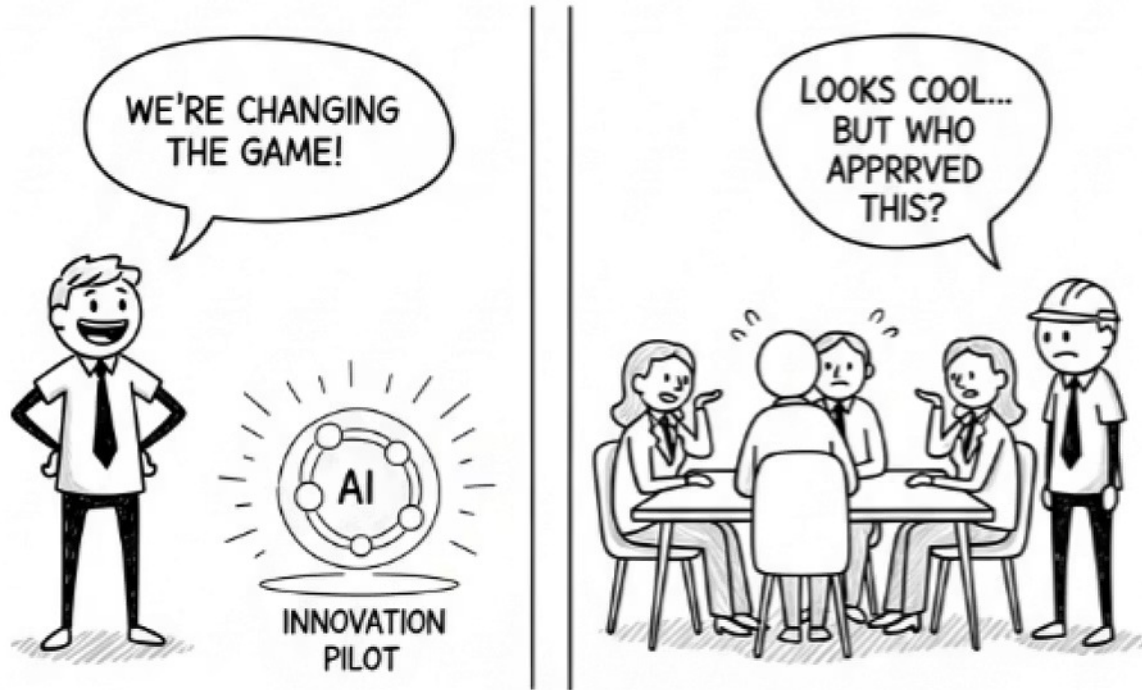
What We Did Wrong

- Didn't start with "What problem are we solving?"
- Chased impressive demos with no ROI
- Built AI features nobody asked for
- Ignored "Build vs Buy"

What You Should Do

- Start with measurable business problem
- Use an "Impact vs. Effort" framework to prioritize quick wins
- Build momentum with tangible results

📌 If you can't **measure the problem**, you can't measure the solution.



Hard Truth # 2

You Don't Have Executive Sponsors

Hard Truth #2

You Don't Have Executive Sponsors

⚠️ 85% of AI initiatives fail to scale due to absent executive sponsorship." — Gartner

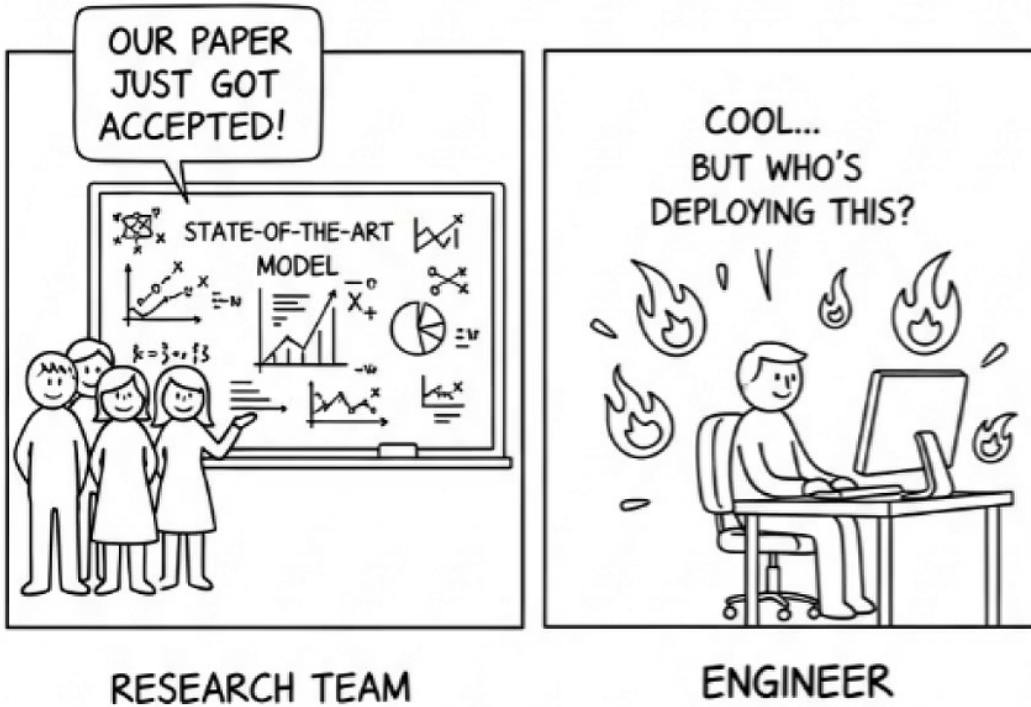
What We Did Wrong

- Led with "AI features" not "\$2M in savings"
- Never defined what "Success" looks like
- Lack of prioritization at company level

What You Should Do

- Translate to business metrics & existing strategy
- Start with "No-Brainer" wins
- Crisp one-pagers and sign-offs
- Lock Quarterly Check-ins

📌 Frame AI initiatives as a **core strategic priority**, not an R&D experiment.



Hard Truth # 3

Your Team Isn't Built To Ship

Hard Truth #3

Your Team Isn't Built to Ship

 "Most AI success is linked to fusion teams and manager-led adoption" — Gartner

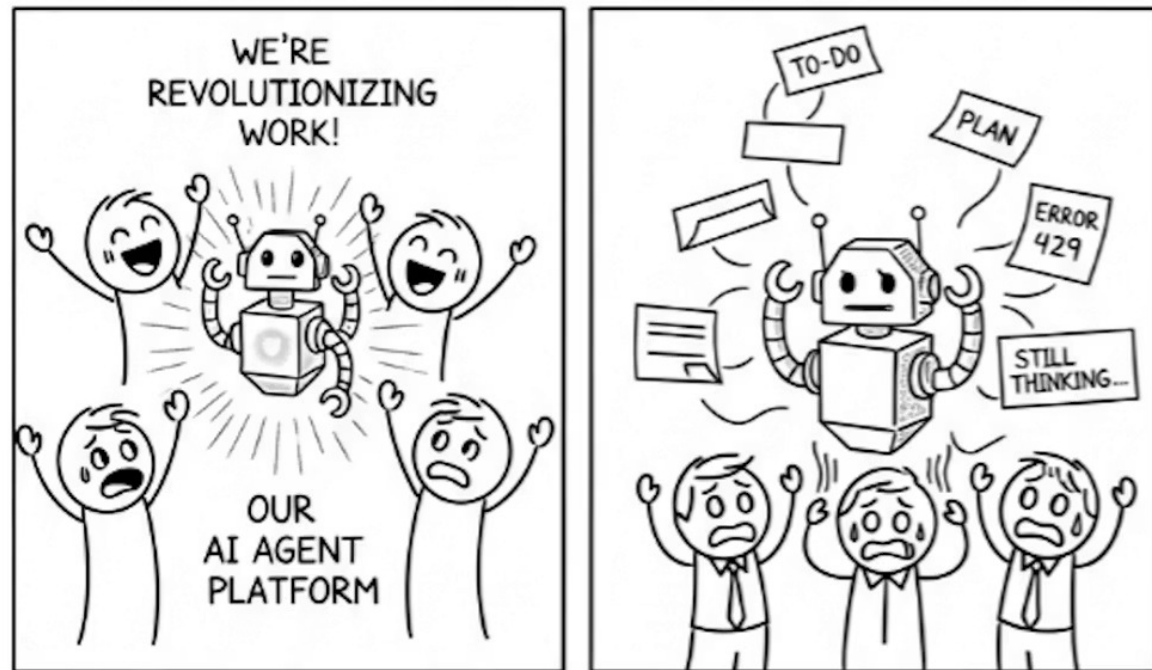
What We Did Wrong

- Spent endless cycles chasing "AI talent"
- Hired ML researchers not system builders
- First production feature took 6 months to ship
- Isolated central AI team

What You Should Do

- Interview for production instinct and LLMOps
- Hire system builders and operational excellence
- Ship infrastructure first, models second
- Embed or grow AI talent within product teams

 Stop hiring for what looks impressive on paper. **Hire for what ships.**



Hard Truth # 4

Your AI Agents Won't Fix Broken Processes

Hard Truth #4

Your AI Agents Won't Fix Broken Processes

⚠️ 40%+ Agentic AI projects to be canceled by 2027 — Gartner

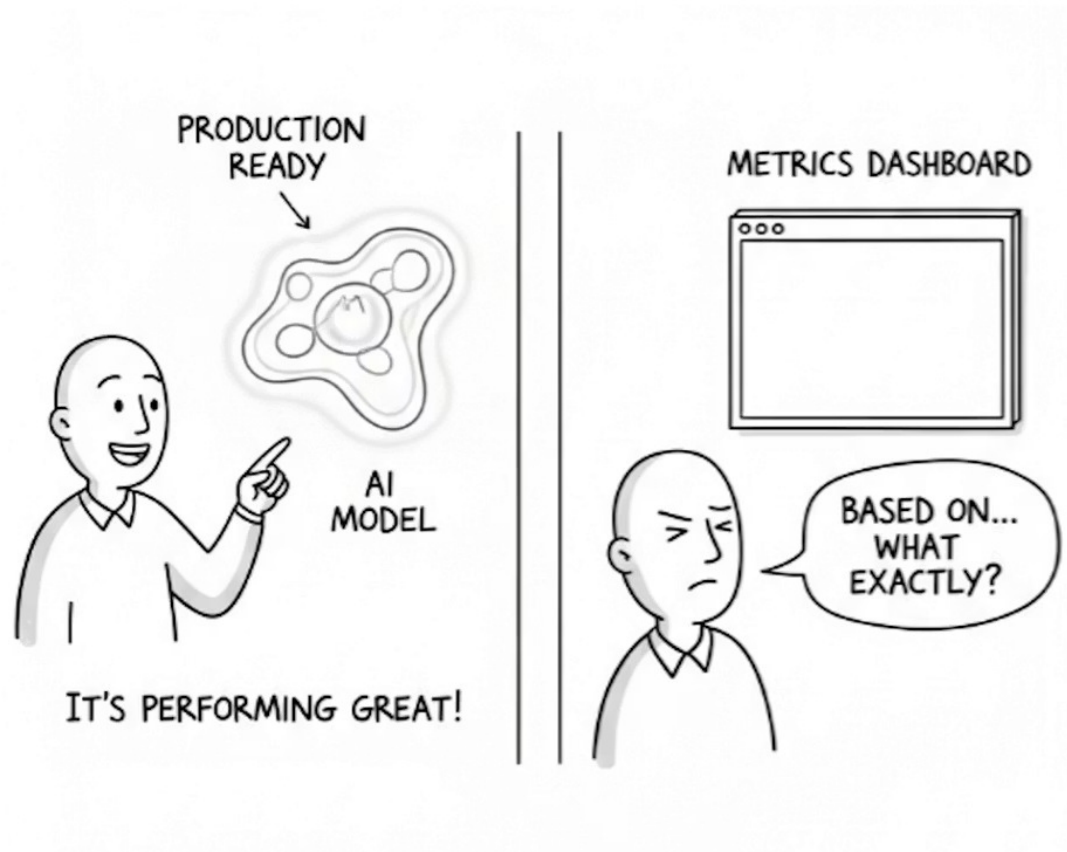
What We Did Wrong

- Believed every "agentic AI" hype post
- Confused autonomy with intelligence
- Built agents for problems humans couldn't even solve

What You Should Do

- Audit and fix the human process first
- Agents are tools, not magic
- Define clear boundaries: What can it decide? What can't it?
- Agents need guardrails and human-in-the-loop

📌 Autonomous agents on broken processes = **autonomous disasters**



Hard Truth # 5

You are Ignoring AI Evals

Hard Truth #5

You're ignoring AI Evals

"Evals" define and measure success. Without them, you're flying blind.

Business Metrics

Did we move the needle?

Revenue

Customer Satisfaction

Handle Time

Model Metrics

Is it reliable?

Hallucination

Accuracy

Precision

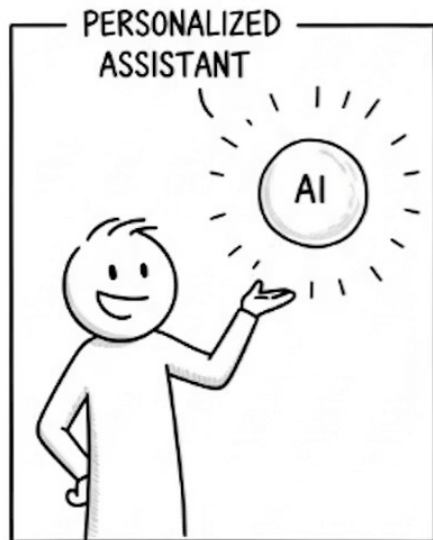
System Metrics

Can we afford to scale?

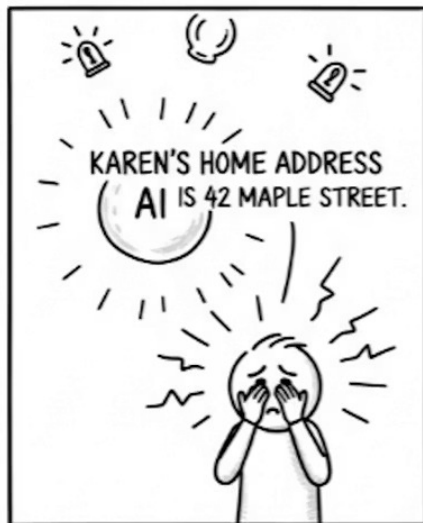
Cost/transaction

Latency

Uptime



IT LEARNS EVERYTHING
ABOUT USERS!



Hard Truth # 6

Your Biggest Risk Isn't the Model

Hard Truth #6

Your Biggest Risk Isn't the Model

⚠️ "97% of AI breach victims lacked access controls" - IBM

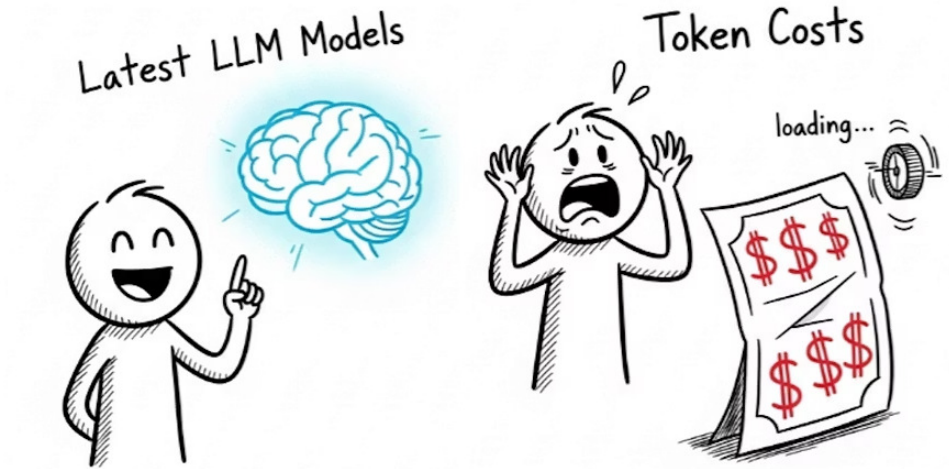
What We Did Wrong

- Treated security as "bolt-on later"
- Let PII flow through prompts and logs
- No governance until after our first scare

What You Should Do

- Design for security from day one
- PII never touches the model, redact it first
- Build AI governance council early
- Assume breach - design accordingly

📌 Your biggest risk isn't hallucinations. It's a **data breach**.



Hard Truth # 7

Your Model is
Slow and
Expensive

Hard Truth #7

Your Model is Slow and Expensive

⚠️ "AI project costs can balloon 500–1,000%" — Gartner

What We Did Wrong

- Used GPT-4 for everything—\$0.37 per inference
- No cost monitoring until 6-figure surprise
- Never considered caching or cheaper alternatives

What You Should Do

- Right-size models: GPT-4 vs 3.5 vs local
- Track cost per transaction from day one
- Optimize: Shorter prompts, caching, batching
- Stay model-agnostic, avoid vendor lock-in

📌 The best model today won't be the best tomorrow. **Build for flexibility.**

Our Playbook for Production AI

1

Start with the Problem

2

Secure Sponsorship

3

Hire to Ship

4

Fix Processes First

5

Build AI Evals

6

Security from Day 1

7

Manage Cost & Latency

Screenshot this. Share it. Skip our mistakes.

Thank you!



Connect with me on LinkedIn

Rashi Agrawal
Head of AI Engineering