

AI Observability Maturity

\$ whoami

- Saurabh Hirani
- Principal SRE at **One2N**
- Infrastructure and reliability engineering
- Joke pronouns - sed / awk
- Sed-istc and Awk-ward

What this talk is not about?

- A hello-world exploration of tools
- “I evaluated 20 tools in 2 hours so you don’t have to”
- “XYZ tool is the best”

What's *init* for you?

- A way to think about AI observability from first principles
- How to evaluate what's good enough at your stage
- Customer learnings compressed in demos
- An open source repo you can tinker with

Everybody wants to know...

- What tool should I use for AI observability?
- I don't want to waste time exploring what's already out there

Everybody should ask...

- Do I need instrumentation?
- What am I instrumenting for?
- Correctness - Is my AI application giving the right answers?
- Performance - Is it doing so fast enough and cheap enough?
- Is the tooling aligned with my *current* organizational priorities?

Am I adopting what is
trending?

Today - What if...

- You had that tool / that instrumentation library / that AI gateway
- It exposes all the metrics you needed to monitor AI applications

Will your reliability problems
disappear?

10 years ago - What if...

- You had a tool that exposes all the metrics you needed
- To monitor VMs, Databases, Caches, etc.

You ~~could've~~ had it all

- Rolling in the deep
- 100s of metrics per exporter
- Yet your systems find newer ways to fail
- Basics - disk / cpu / memory
- Beyond basics - Disk IO / File desc exhaustion / TIME WAIT issues
- Beyond beyond basics - Organizational service quirks
- Beyond beyond beyond.... Knowless

New tools, old mistakes

- AI instrumentation libraries expose metrics and spans
- Is your retrieval quality degrading?
- Is your cheaper model really cheap?
- Is that shiny dashboard pleasing to the eye or critical to ops?

Tools don't think about your
failure modes. You do.

Our approach

- Failure modes > tools
- What can go wrong?
- What signals can catch that scenario?
- What tool emits that signal?
- If the tool falls short - can I augment it manually?

What we built

- Worked with customers and saw decent AI observability gaps
- Addressed them with minimal intrusion
- 2 customers - both using RAG applications
- Compressed those learnings in a demo app with experiments

Demo-1

● Pending

Demo-2

● Pending

Demo-3

● Pending

Questions?

Twitter - @sphirani

LinkedIn -

[linkedin.com/in/shirani/](https://www.linkedin.com/in/shirani/)

