INFOBIP SHIFT

Infobip Shift Conference 2026

13–15 September 2026; Zadar, Croatia

PUBLIC

INFOBIP SHIFT

Infobip Shift 2026
Zadar, Croatia

Building a (mostly) reliable research assistant

Infobip Shift 2026



Dan Maher

Senior Technical Advocate & Researcher
Datadog

The thing that needed me to make a thing

- ***SSCSS***: 50+ questions, 8 sections, yes / no / unknown
- The evidence lives across source code, docs, blogs, RFC repos, policy pages, whitepapers, newsgroups, social media…
- Docs and reality don't always agree and simple web searches only get me the easy answers

PUBLIC

```
{
  "name": "Dan Maher",
  "handle": "phrawzty",
  "title": "Senior Technical
    Advocate & Researcher",
  "employer": "Datadog",
  "portfolio":
    "https://speaking.dark.ca/",
  "🤫": "secretly French"
}
```

INFOBIP SHIFT

Infobip Shift 2026
Zadar, Croatia



DATADOG

Warning: marketing statement follows!

Datadog is the leading observability and security platform for the AI era, providing businesses with unified visibility across the technology stack to manage complexity at scale. It brings applications, infrastructure, data, models, and security into one place, using AI to detect and resolve issues before they impact customers. Trusted globally by Fortune 500 companies and high-growth AI leaders, Datadog enables businesses to move faster with clarity and confidence.

Four attempts
(the third one will shock you!)

Attempt one: head-first into a wall

- One prompt in Claude: "here are 52 questions, here's a registry, go"
- Context limits, rate limits, truncated output
- LOL this did **not** work
- Lots of tokens though! 🙄

PUBLIC

Attempt two: the version you'd draw on a whiteboard

- Orchestrator + eight parallel researcher agents, one per section, shared state file
- Hanging agents, silent blank results reported as success, state file corruption (it’s just vibes y’all)
- *Bonne nouvelle !* I hardened it for weeks and it got more robust!

PUBLIC

Attempt three: engineering mode *activate*

- Full rewrite: JSON schema enforcement, real test suite, sandboxed egress, CI on every commit, parallelism…
- The kind of codebase you'd be proud to ship
- *"Oh yeah that’s hot." – me, at the time*

PUBLIC

NOTHING
(worse than)

Production-grade reliability theatre

- Confident YES answers backed by evidence that said NO if you actually read the friggen source. URLs that didn't exist. Notes attached to the wrong question.

- "Evidence" that turned out to be the *model narrating its own search process like what even*

PUBLIC

What even is truth?

- Farquhar et al., Nature, 2024: **confabulation**
- Kalai et al. (OpenAI), 2025: **models are trained to guess**, not to say "I don't know"
- Hicks, Humphries & Slater, 2024: it's not lying;
  it's indifferent to truth

PUBLIC

You cannot code your way out of an LLM quality problem

On the shoulders of giants

- Huyen, 2022: ML systems fail silently; quality problems are harder to notice than operational ones
- Husain & Shankar, 2025: start with error analysis, not infrastructure. *Read 20 outputs before you build anything*
- Yan, 2025: "**An LLM-as-Judge Won't Save The Product**"

PUBLIC

Attempt four: Stop—constraint time!

- Lead with **constraints, not requirements**
- The bar isn't "fast", it's "faster than what exists now"
- Reduce or remove the failure surfaces. Minimum viable product first, then evolve from there

PUBLIC

Models and harnesses matter

Models and harnesses matter

- Sonnet: good for process, bad for (this) research
- Opus: bad for process, good for playing around with words
- HumanLayer, 2026: "what looks like a model's personality is very often a harness's personality"

PUBLIC

Simple, but not simplistic

- Schluntz & Zhang (Anthropic), 2024: the best agent systems use simple, composable patterns, not frameworks

- Willison, 2025: "an LLM agent runs tools in a loop to achieve a goal" (spoiler: that's the whole abstraction)

PUBLIC

INFOBIP SHIFT

Infobip Shift 2026
Zadar, Croatia

Seven things that work

1. Guardrails; or "the art of scoping your sources"

- Decide where the agent is allowed to look, before it's allowed to think
- Prioritise good sources and deny bad ones up front
- Exhaust all good sources before hitting the information superhighway (it's not actually that safe)

PUBLIC

2. Discipline; or "inference is not evidence"

- Evidence includes a verbatim quote and a URL
- Every question asked **three different ways**, even after a confirming hit
- Gao et al., (HyDE), 2022: one phrasing is a sample, not a signal—**sample more**

PUBLIC

Absence of evidence is not evidence of absence

Altman & Bland (BMJ), 1995

3. Logs; or "what the heck did the agent even do?"

- Log the **reasoning, not just the answer**
- This is how you build, debug, and improve
- Can be **operational**, not just informational

PUBLIC

4. Rigour; or "you can’t shortcut the methodology"

- Once is not enough; use tokens **intelligently**
- He et al. (Thinking Machines Lab), 2025: rerun variance is a property of the inference system
- **Variance is signal**, be careful if there isn’t any

PUBLIC

5. Prompting; or "operational != behavioural"

- Operational: what to do, in what order, with what tool
- Behavioural: how to reason, communicate, set boundaries
- Shankar et al. (UIST), 2024: criteria drift; "you can't write the rules before you've seen them fail"

PUBLIC

6. Human-in-the-loop; or "can’t do my job yet"

- Involve a human when the variance is stable; domain expertise and intuition matter
- Where the runtime file and LLM experimentation shine
- Faster is only better if the results are good

PUBLIC

"Human oversight is still needed even with automated evaluators."

Eugene Yan, 2025

7. Sandboxes; or "basic agentic security hygiene"

- Willison's lethal trifecta: private data, untrusted content, external communication
- Greshake et al., 2023: hostile pages can hide instructions aimed at the agent, not the human reading it
- I use a kernel-level sandbox on a virtual machine, fwiw

PUBLIC

Everything, on one slide 

- Architecture can make you more robust without making you more correct
- Different tasks need different model and harnesses
- Guardrails, Discipline, Logs, Rigour, Prompting, HITL, and don’t forget Sandboxes

PUBLIC

Moral of the story

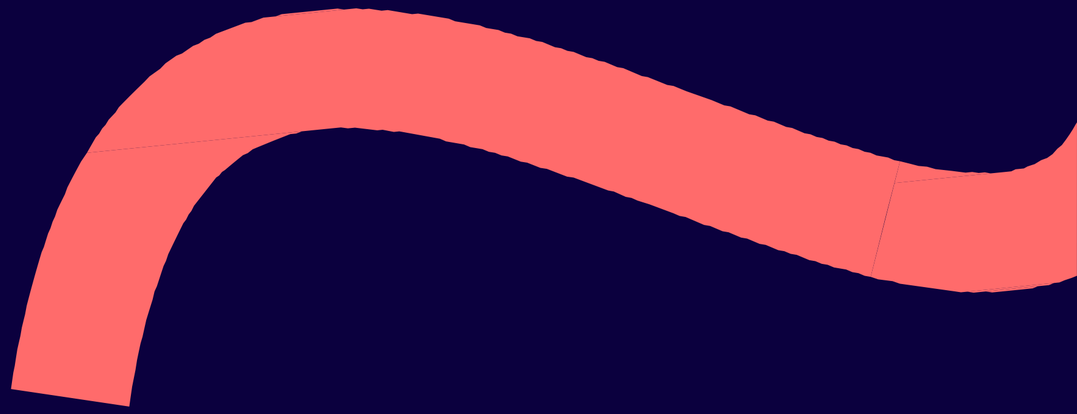
- Research **assistant, *not leader***
- Don't let the LLM boss you around
- Seriously though: you can build this too!

PUBLIC

INFOBIP SHIFT

Infobip Shift 2026
Zadar, Croatia

Thank you



Dan Maher

@phrawzty