Lessons Learned Writing Unit Tests for Chatbots

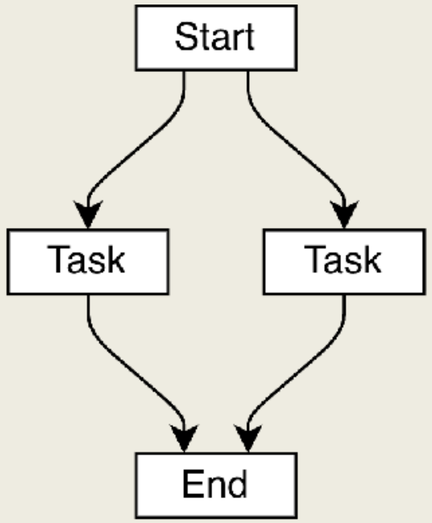
Dean Wampler
The AI Alliance and IBM
dwampler@thealliance.ai
Applied ML Conference 2026

aialliance.org

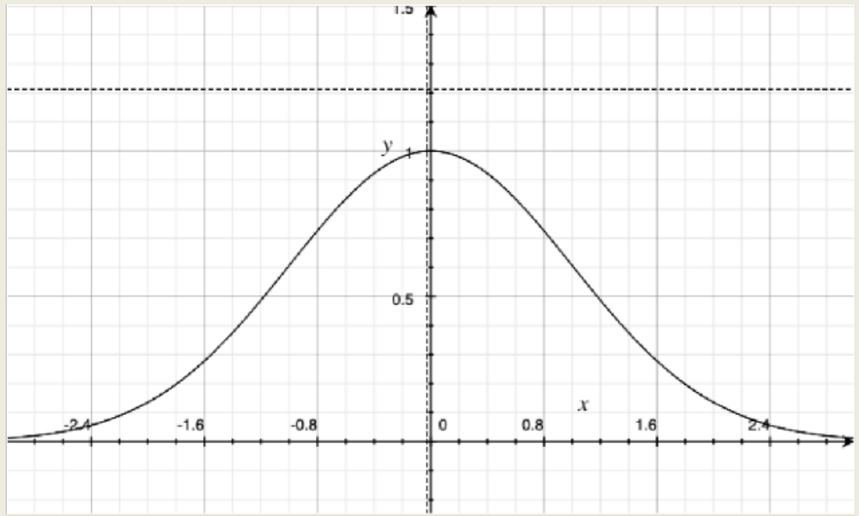
deanwampler.com/talks

© 2023-2026, Dean Wampler, except where noted.

Software Developer



Deterministic Behavior



Stochastic Behavior

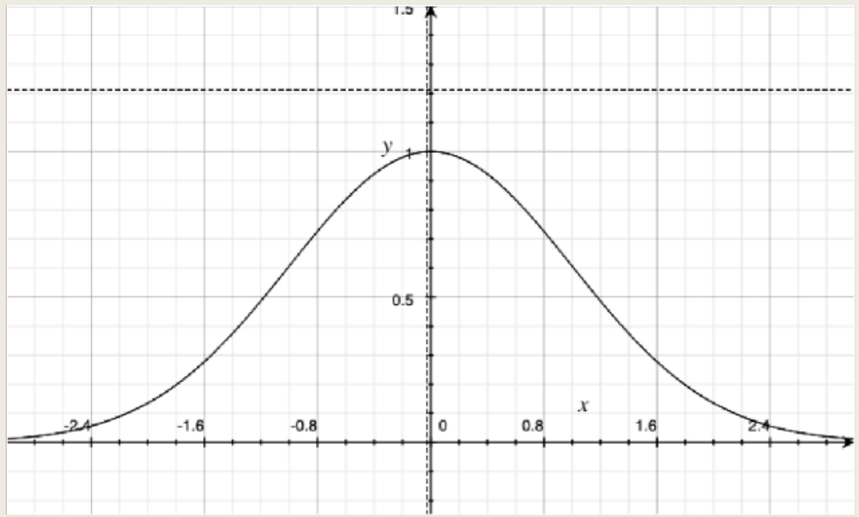
AI Expert

The Spectrum of Experience

The World Is Stochastic

Probabilities and Statistics:

The tools scientists (including data scientists), model builders, etc. use to understand their systems.



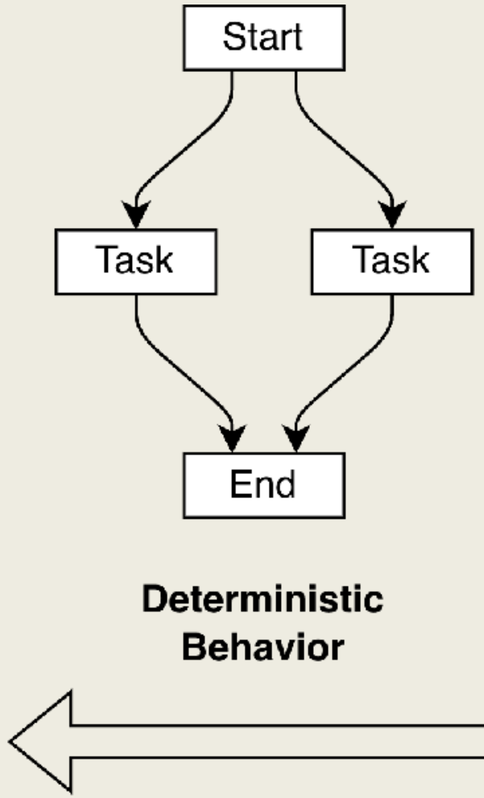
Stochastic Behavior

AI Expert

Spectrum of Experience

Developers Assume Determinism‡

Software Developer



Deterministic Behavior

Test-Driven Development

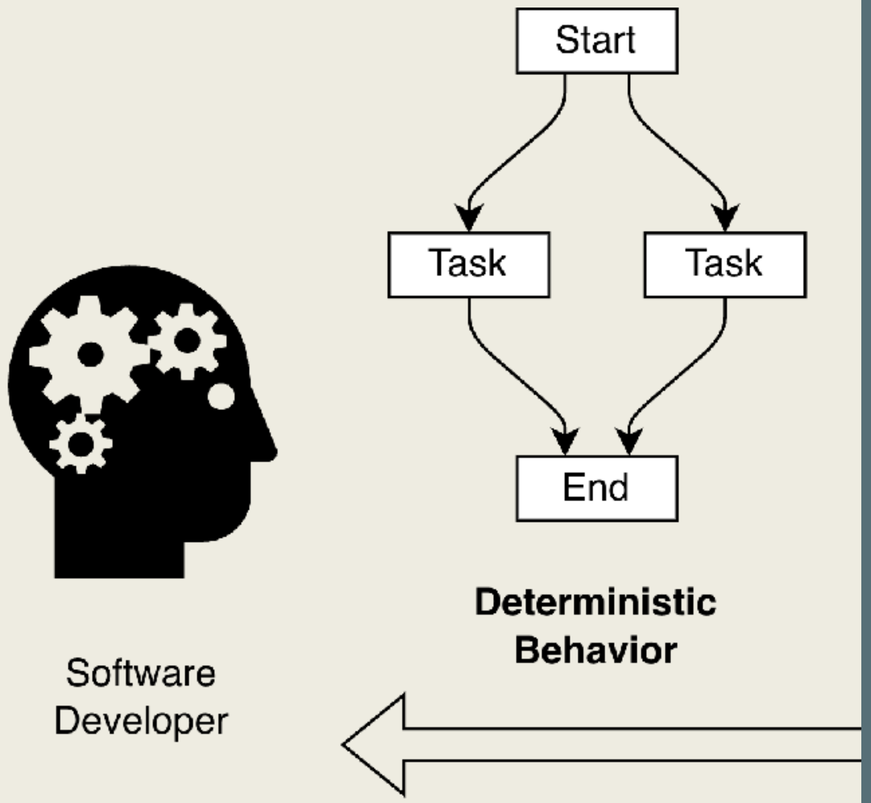
Refactor:
Prepare for new feature, using existing tests to catch regressions

Write New Test:
For the feature to be implemented

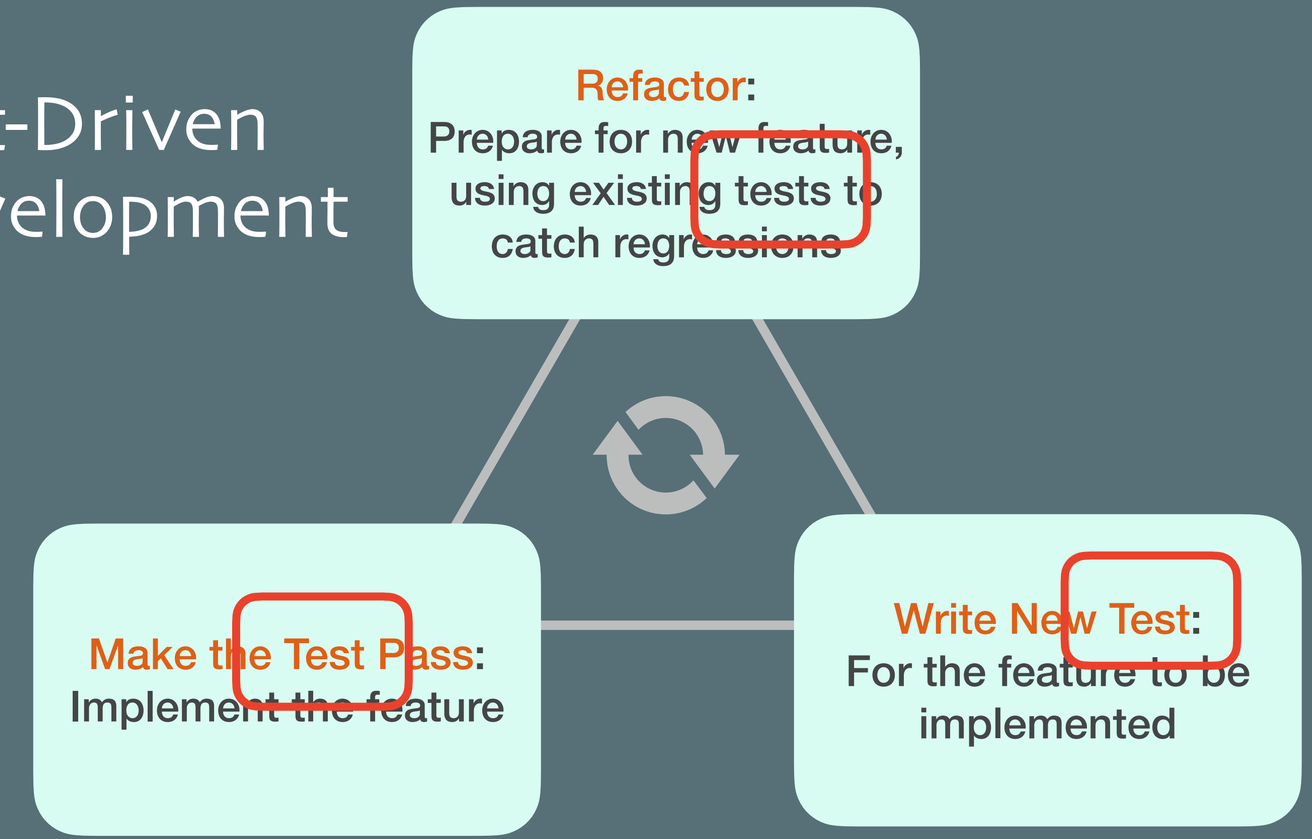
Make the Test Pass:
Implement the feature

‡ Well, distributed systems aren't…

Developers Assume Determinism‡



Test-Driven Development



‡ Well, distributed systems aren't…

AI ALLIANCE

Testing Generative AI Agent Applications

Search Testing Generative AI Agent Applications

Let's bridge the divide!

Join This Project | GitHub Repo

- Home
- Testing Problems
- Architecture and Design for

Testing Generative AI Agent Applications

(Previous Title: *Achieving Confidence in Enterprise AI Agent Applications*)

I am an enterprise developer; how do I test my AI agent applications??

I know how to test my traditional software, which is ***deterministic*** *(more or less...), but I don't know how to t*
applications, which are uniquely ***stochastic****, and therefore* ***nondeterministic****.*

Welcome to the **The AI Alliance** project to advance the state of the art for **Enterprise Testing of Generative**
Applications. We are building the knowledge and tools you need to achieve the same testing *confidence* for y
applications that you have for your traditional applications.

Note:

https://the-ai-alliance.github.io/ai-application-testing/

e resources ab
contains gene

Two Lessons from This Project

- Restore determinism where you can
- Adopt and adapt benchmarks → "Unit Benchmarks"

Restore determinism where you can

```
Welcome to the patient ChatBot. Type help or ? to list commands. Use "bye" to

input>
```

Healthcare Chatbot Example App

- A use case-specific ChatBot has “frequently asked questions”:
  - I need a prescription refill.
  - Where is your office?
  - I need to change my next appointment.
  - I am gasping for breath since I started taking Killsfatzpic.

“Please hang up and dial 911…”

Restore determinism where you can

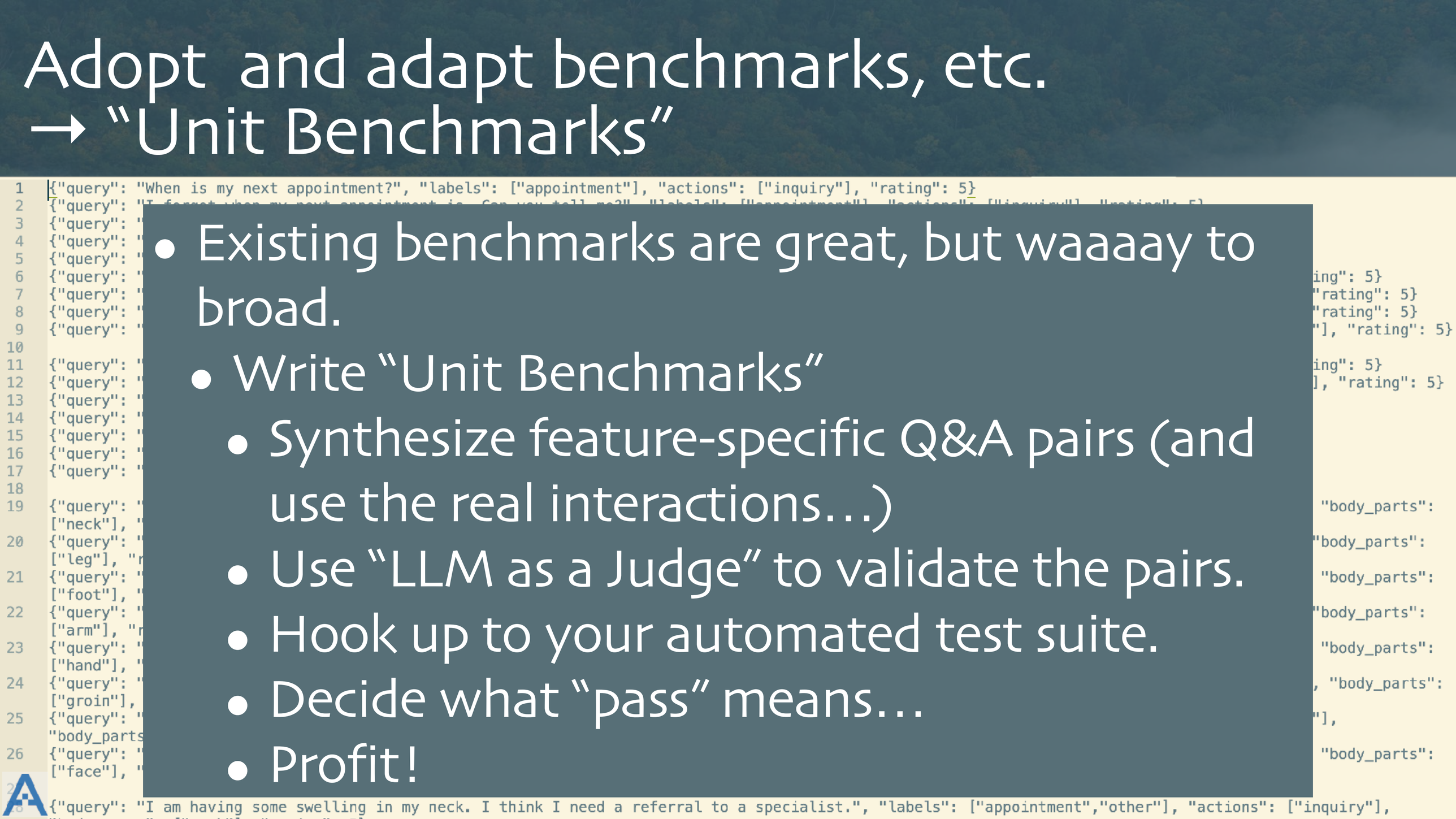
```
Welcome to the patient ChatBot. Type help or ? to list commands. Use "bye" to

input>
```

Healthcare Chatbot Example App

- LLMs are great at classifying the variety of human queries into categories.
  - v0.1 of your ChatBot can just do that; classify the query and route it to the right human or other system.
- In general, keep using “old-school” planners, logic engines, …

Adopt and adapt benchmarks, etc.
→ “Unit Benchmarks”

- Existing benchmarks are great, but waaaay to broad.
  - Write “Unit Benchmarks”
    - Synthesize feature-specific Q&A pairs (and use the real interactions…)
    - Use “LLM as a Judge” to validate the pairs.
    - Hook up to your automated test suite.
    - Decide what “pass” means…
    - Profit!



Thanks!

Dean Wampler
The AI Alliance and IBM
dwampler@thealliance.ai
Applied ML Conference 2026

https://the-ai-alliance.github.io/ai-application-testing/

aialliance.org

deanwampler.com/talks

© 2023-2026, Dean Wampler, except where noted.