

Lessons Learned Writing Unit Tests for Chatbots

Dean Wampler
The AI Alliance and IBM
dwampler@thealliance.ai
[Applied ML Conference 2026](#)

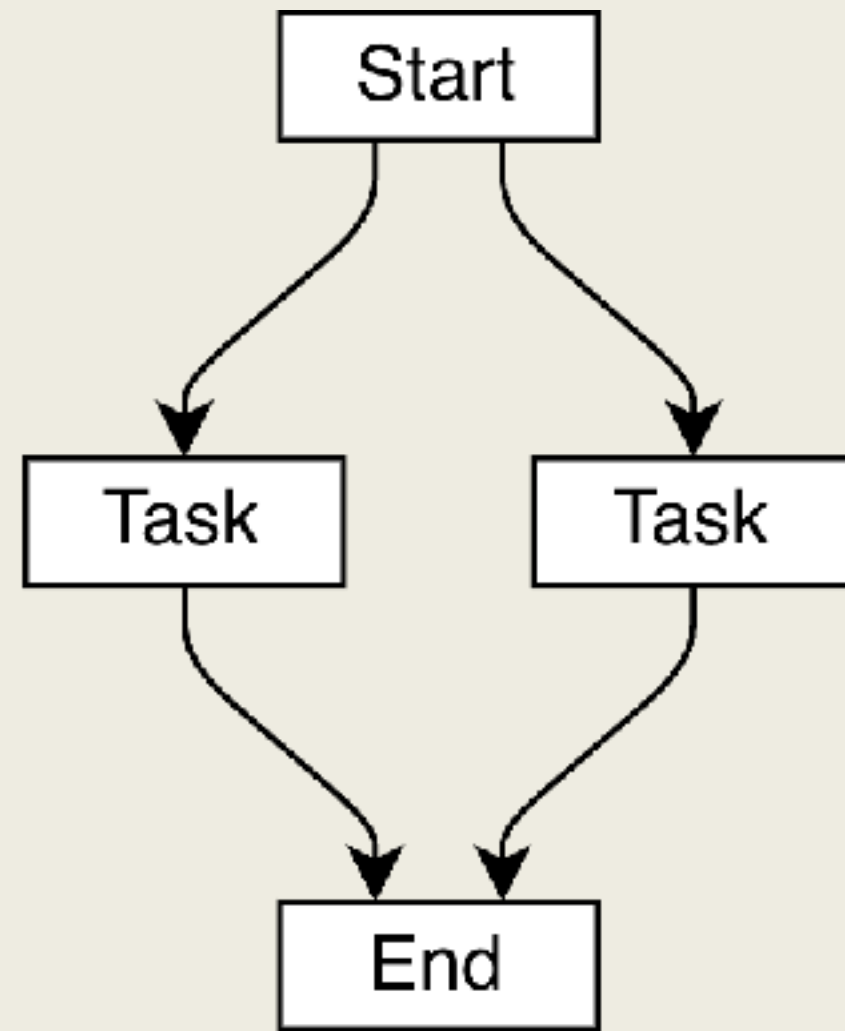
aialliance.org

deanwampler.com/talks

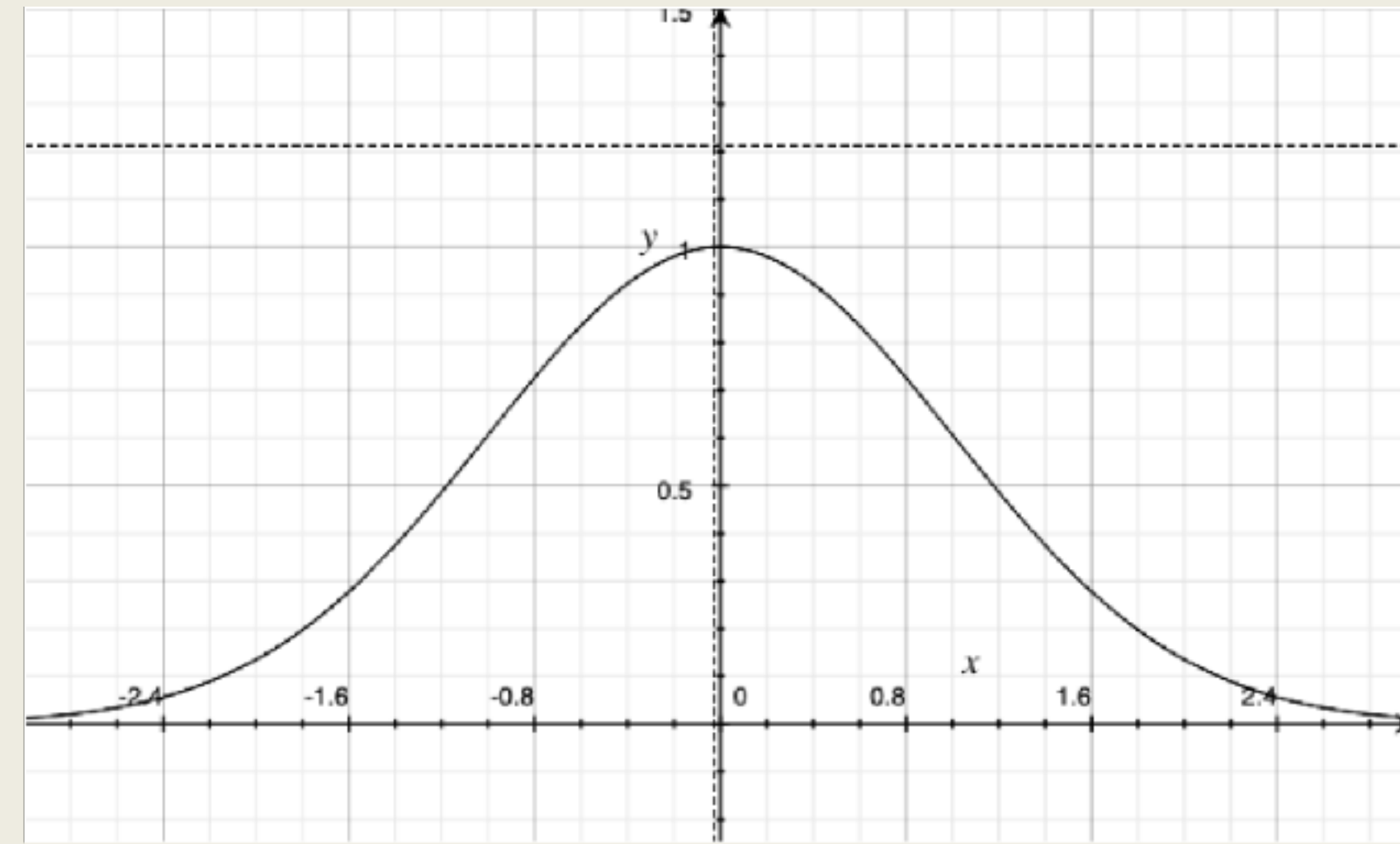




Software Developer



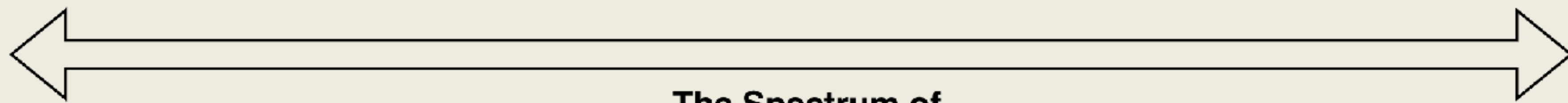
Deterministic Behavior



Stochastic Behavior



AI Expert

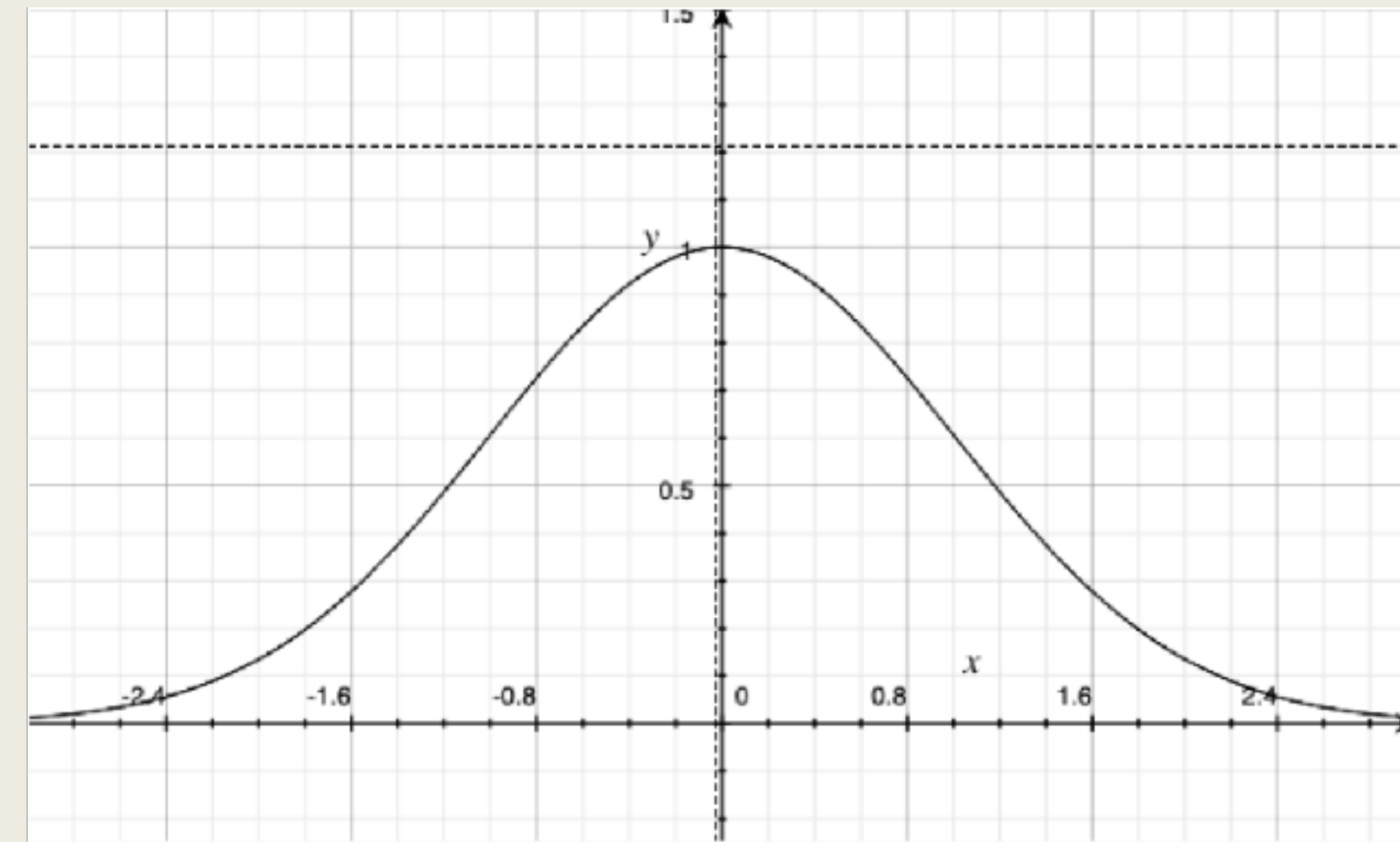


The Spectrum of Experience

The World Is Stochastic

Probabilities and Statistics:

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



**Stochastic
Behavior**



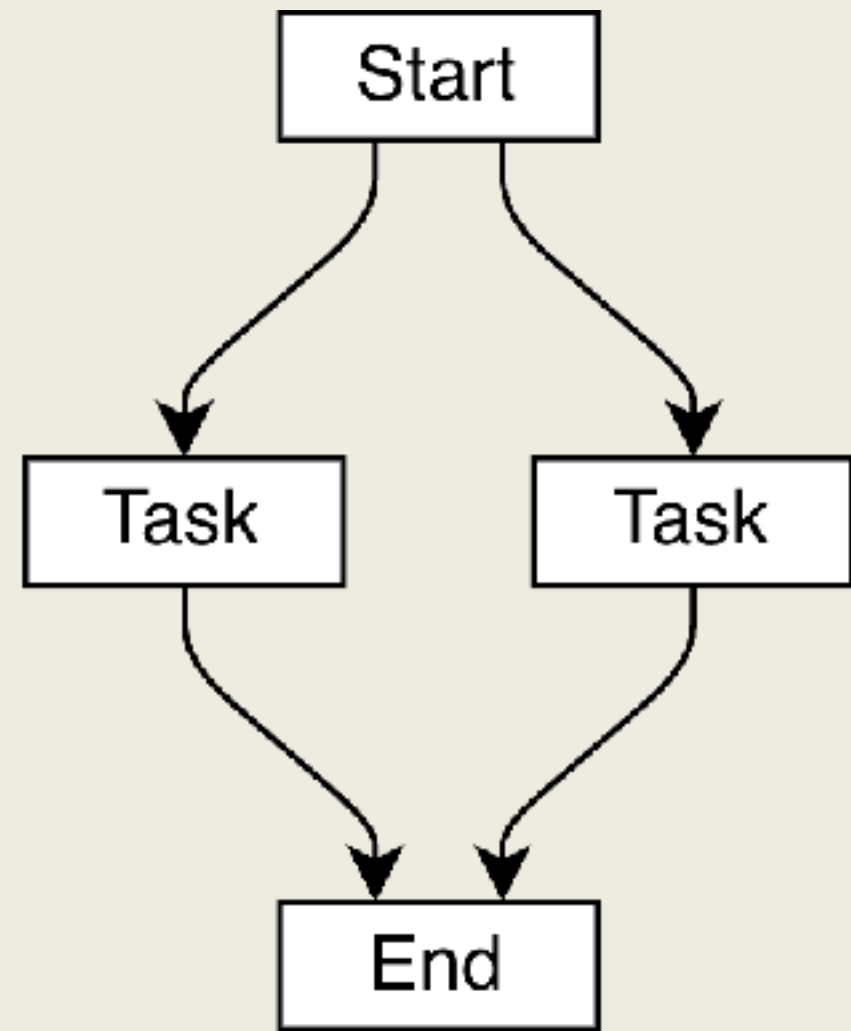
AI Expert

Spectrum of
experience

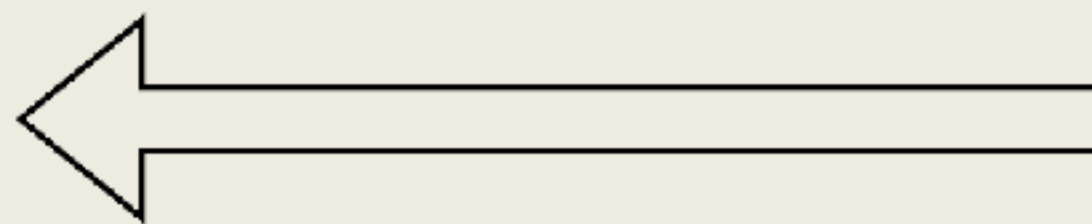
Developers Assume Determinism‡



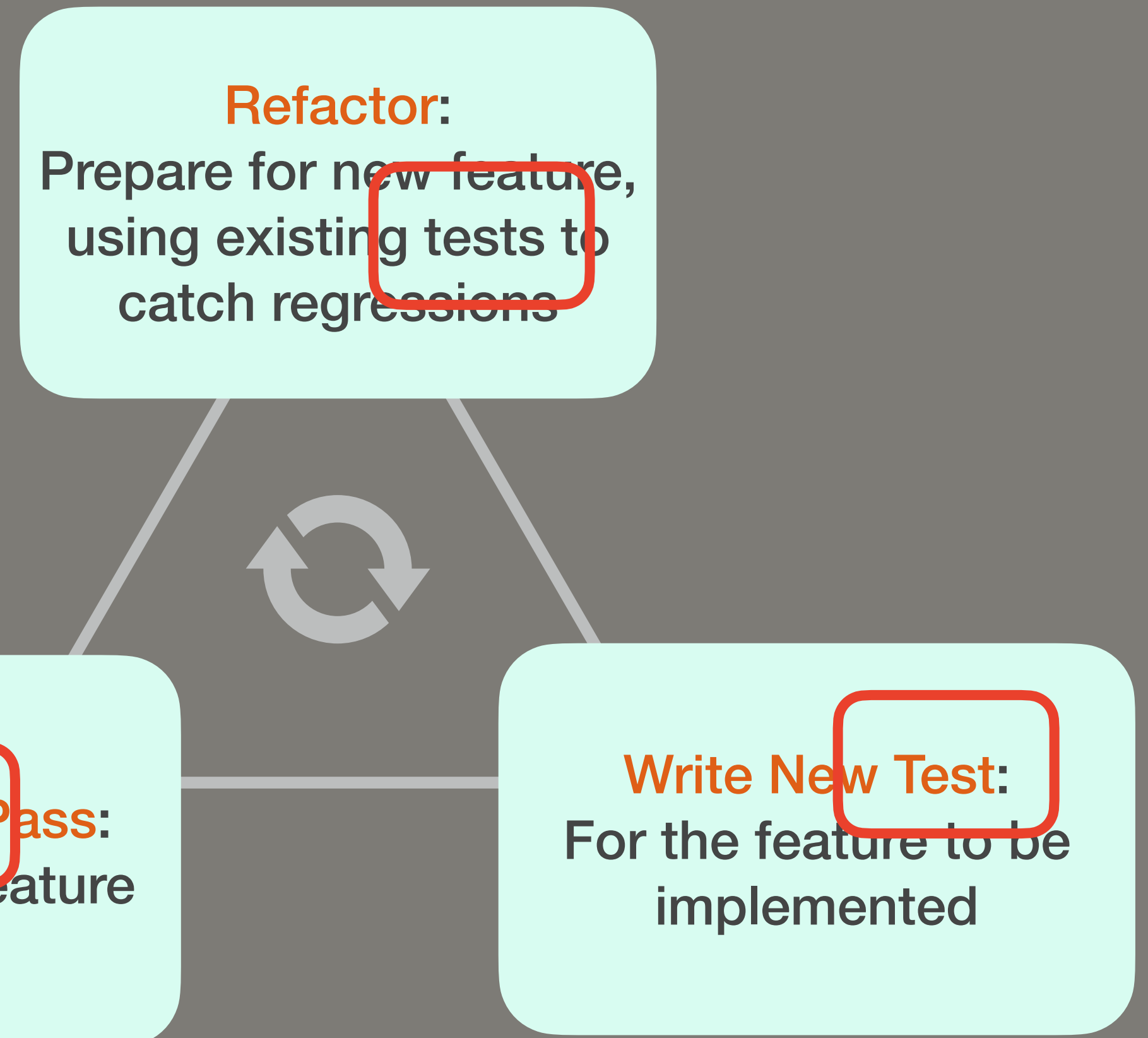
Software Developer



Deterministic Behavior



Test-Driven Development



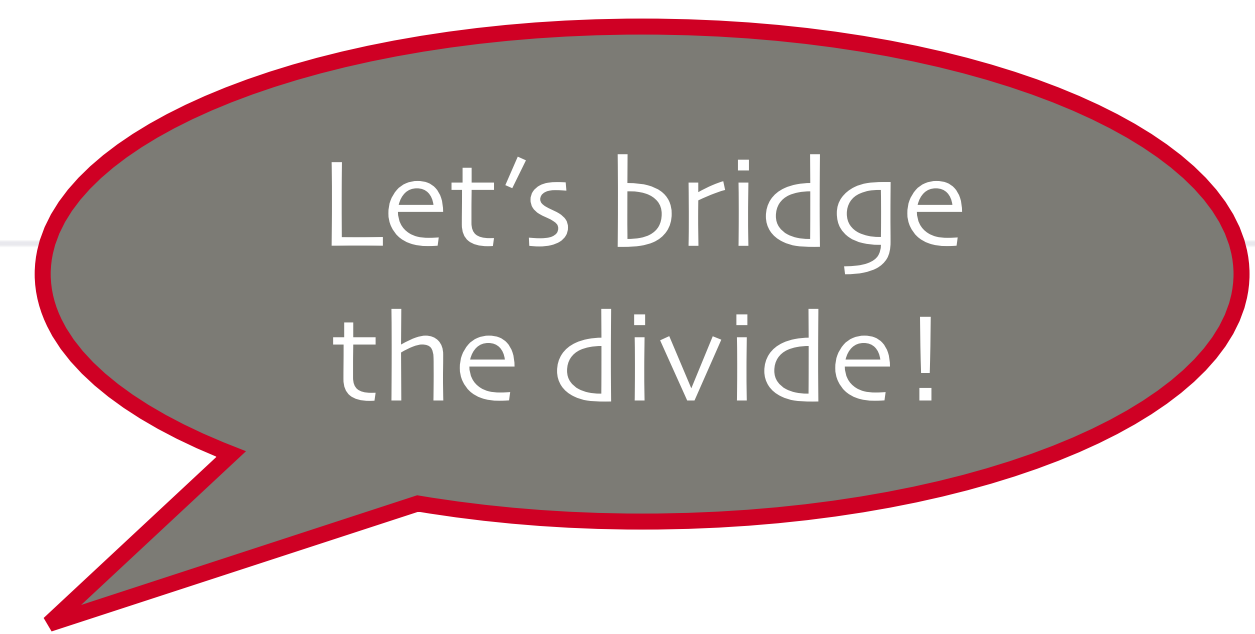
Make the Test Pass:
Implement the feature

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

Write New Test:
For the feature to be implemented

‡ Well, distributed systems aren't...





[Join This Project](#) [GitHub Repo](#)

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 test my AI agent 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 AI Applications**. We are building the knowledge and tools you need to achieve the same testing *confidence* for your AI agent applications that you have for your traditional applications.

Note:



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

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 waaaaay too 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!

```
1 {"query": "When is my next appointment?", "labels": ["appointment"], "actions": ["inquiry"], "rating": 5}
2 {"query": "I forgot when my next appointment is. Can you tell me?", "labels": ["appointment"], "actions": ["inquiry"], "rating": 5}
3 {"query": "..."
4 {"query": "..."
5 {"query": "..."
6 {"query": "..."
7 {"query": "..."
8 {"query": "..."
9 {"query": "..."
10
11 {"query": "..."
12 {"query": "..."
13 {"query": "..."
14 {"query": "..."
15 {"query": "..."
16 {"query": "..."
17 {"query": "..."
18
19 {"query": "I have a swelling in my neck. I think I need a referral to a specialist.", "labels": ["appointment", "other"], "actions": ["inquiry"],
20 {"query": "..."
21 {"query": "..."
22 {"query": "..."
23 {"query": "..."
24 {"query": "..."
25 {"query": "..."
26 {"query": "..."
27 {"query": "..."
28 {"query": "I am having some swelling in my neck. I think I need a referral to a specialist.", "labels": ["appointment", "other"], "actions": ["inquiry"],
```



Thanks!

Dean Wampler
The AI Alliance and IBM
dwampler@thealliance.ai
[Applied ML Conference 2026](#)



aialliance.org

deanwampler.com/talks



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

