

Model Evaluations Are Not Smoke Tests

**Jack Sullivan
ML Engineering
GetReal Security**

Defining the Parameters of Our Pipeline

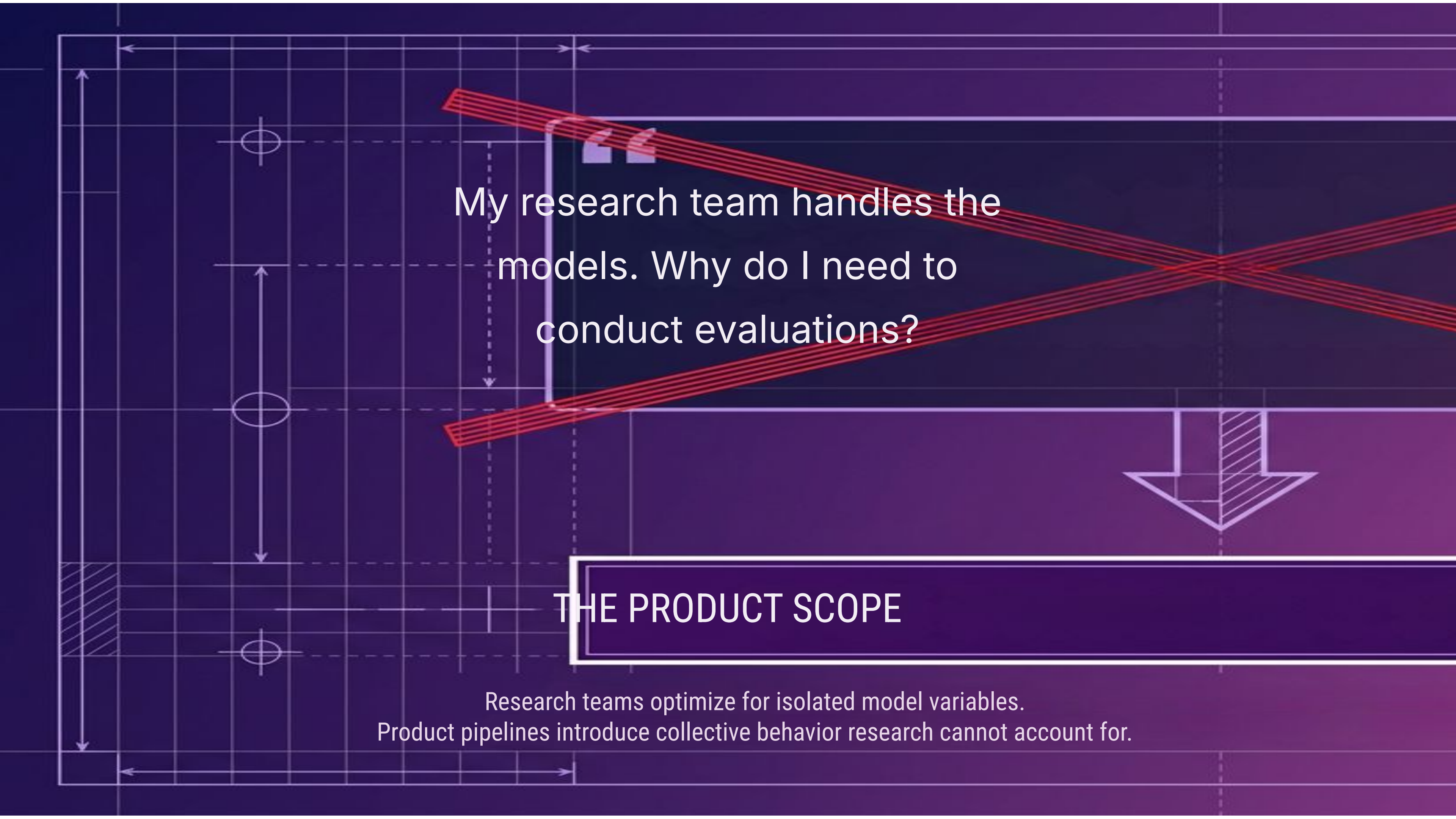
IN SCOPE

- General processes for managing ML business logic.
- Model testing misconceptions vs. deterministic testing.
- Integrating in-house or API wrapper models into a larger product pipeline

OUT OF SCOPE

- Specific model ingredients, recipes, or training mechanics.
- Uptime/downtime incident response scenarios.

Context: We are managing a model pipeline as a sequential gear within a larger product machine.



My research team handles the models. Why do I need to conduct evaluations?

THE PRODUCT SCOPE

Research teams optimize for isolated model variables.
Product pipelines introduce collective behavior research cannot account for.

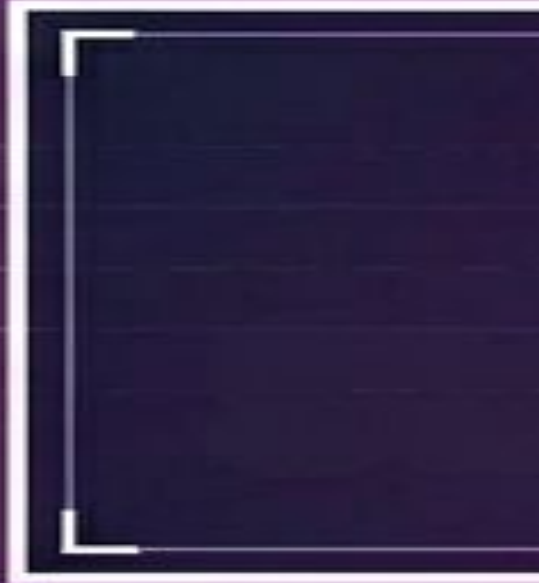


Test Cases != Ready for Launch

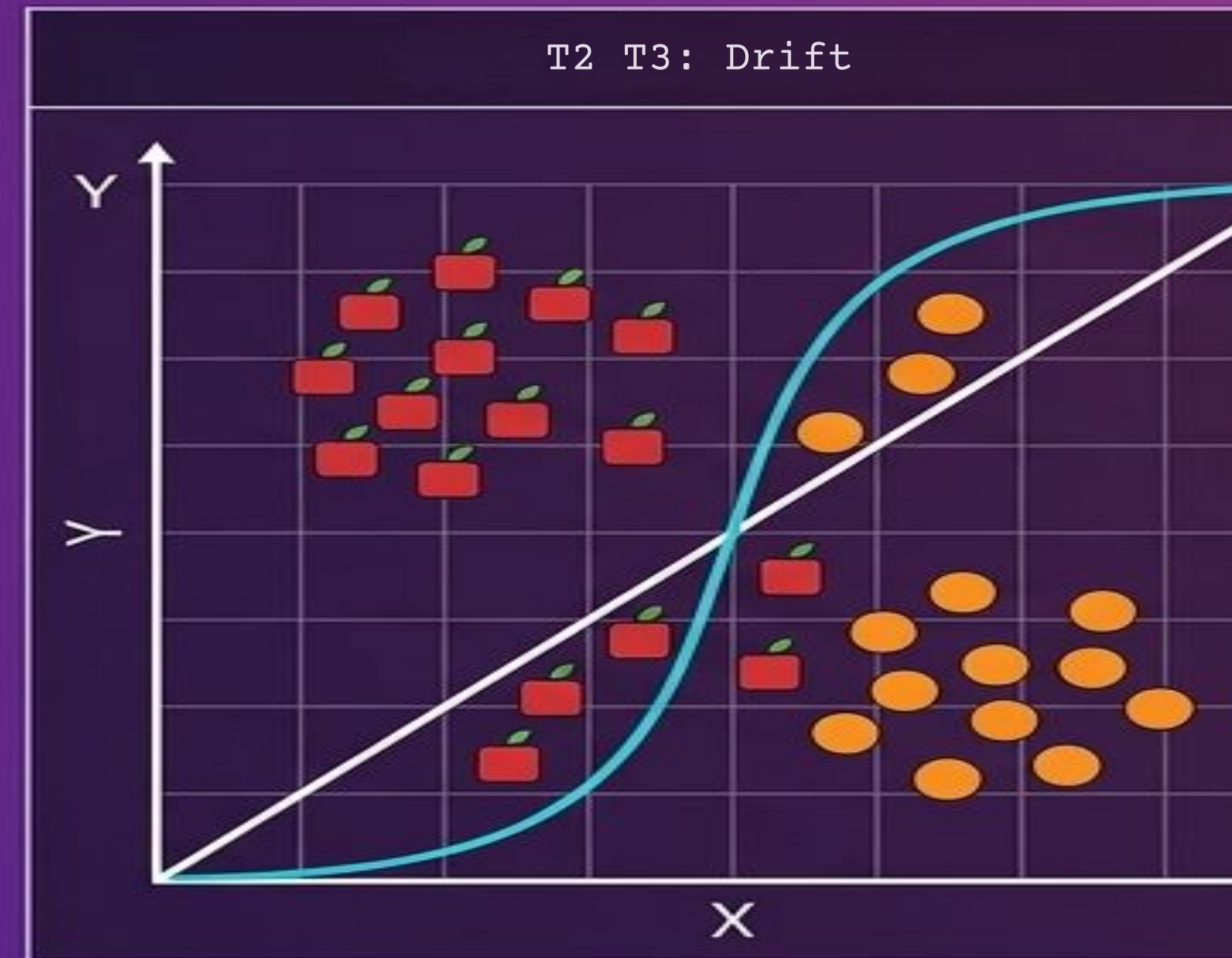
Perfect performance reports from research is not
the release criteria



We require
Statistical Power.



The Reality of Data Drift



The model remains stationary, but real-world data distributions constantly shift.

Defining 'Model Version' in the Product Pipeline

A Single Model Version

Weights

Model Code

Orchestration Code

('The Knowledge') The data representing the trained knowledge of the model

('The Algorithm') Model architecture, layers, activation functions.

('The Flow') Endpoint routing, data handling, and business logic.

Changing ANY of these alters the model and requires version change

Anticipating Model Drift Scenarios

Scenario 1: In-House Drift

- Real-world data shifts or upstream product changes alter the distribution.
- Action: Internal research team must recollect data and retrain the model to adapt to new baselines.

Scenario 2: Third-Party Deprecation

- An external API provider (e.g., OpenAI, Anthropic) discontinues the specific version optimized for your use-case.
- Action: Forced migration to a new, potentially un-optimized version.

Product-Level Consequences



If Model A is retrained or updated, the transformation fundamentally alters the downstream individual data results feeding into B and C.

Can't we simply revert to the previous model version?

The problem is larger than the model.

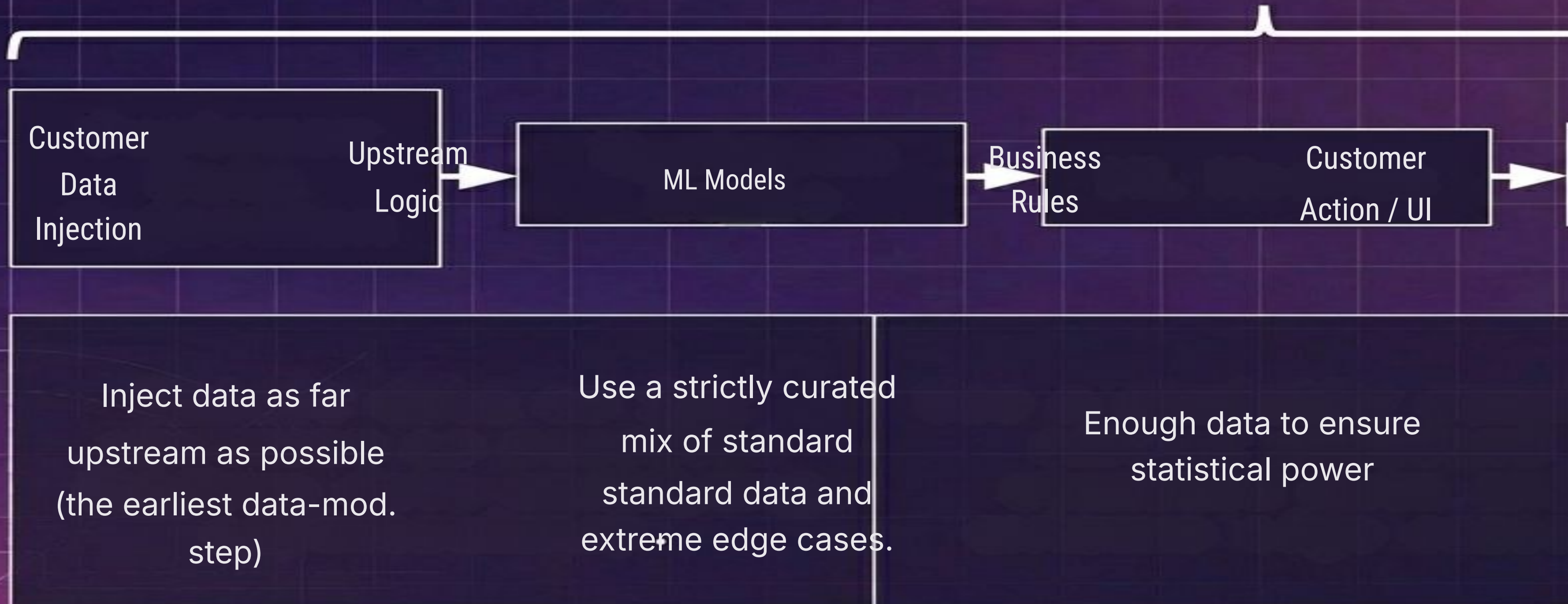
The Upstream Pipeline Break



Cross-team communication is never perfect.
Upstream changes silently break downstream ML mechanics.

Evaluate the Product, Not the Model

Imitate the customer's experience



Product-Level Evaluation: Modalities

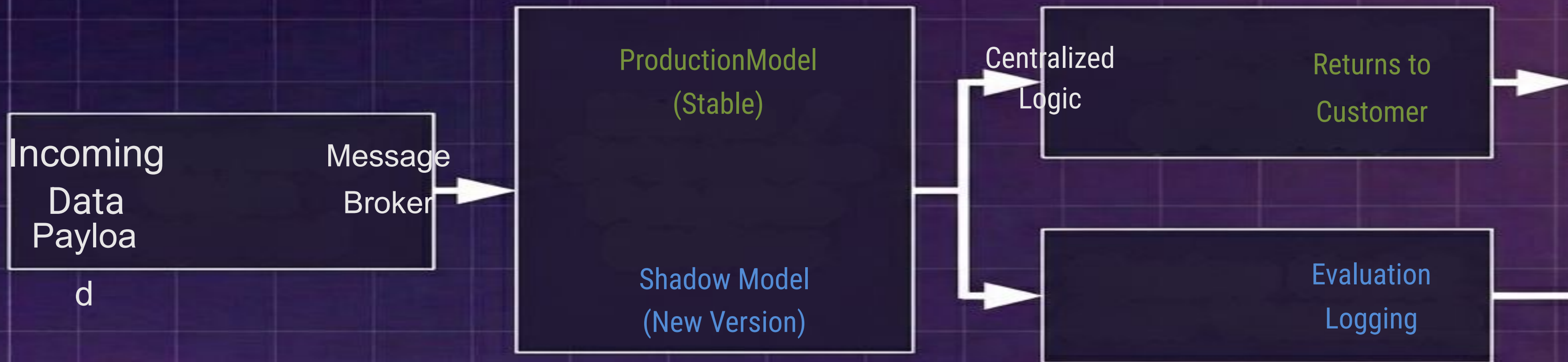
Staging Evaluations

- Curated data evaluation (from the field or from a vendor)
- Expected performance indicator

Production Evaluations

- Route subset of traffic to shadow deployments
- Actual performance indicator
- Shortcomings uncovered feed back into data curation for stage evals.

Shadow Deployment Operations



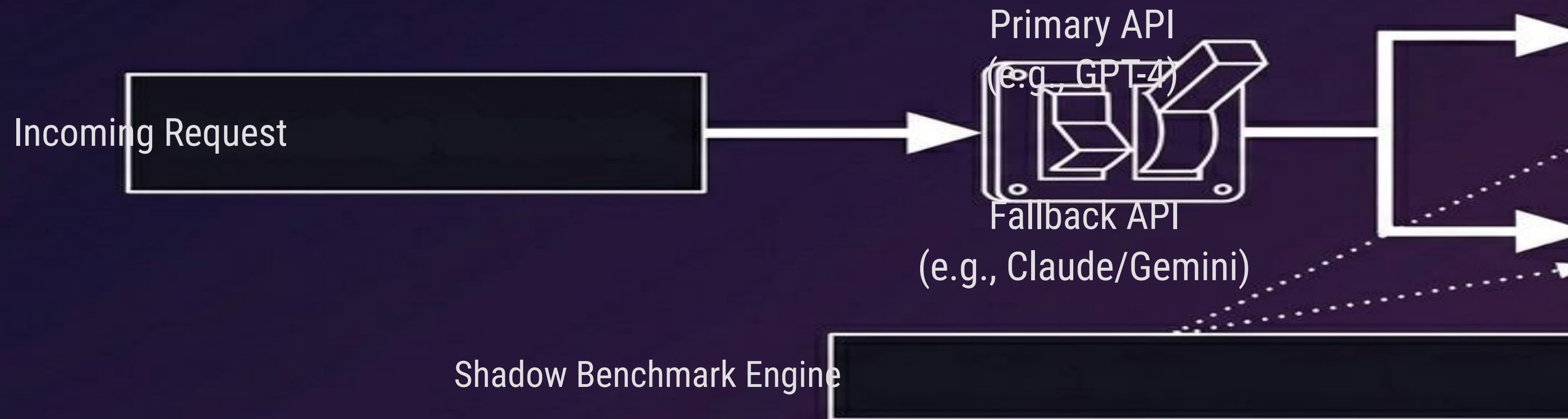
Duplicate payloads to test new models in real-world conditions without exposing results to the customer. Centralized logic prevents scattered business rules and securely intercepts shadow outputs.

Scenario 2: Third-Party Deprecation

- > **API_RESPONSE: 404 NOT FOUND**
- > ERROR: Model gpt-version-Y has been deprecated.

Big company didn't ask us for permission to discontinue support for model X.
What happens API reliance chain is abruptly broken by the provider?

API Contingencies & Benchmarking Fallbacks



Build AI pipelines utilizing a multitude of APIs.
Constantly shadow-deploy and evaluate endpoints in the pipeline on cadence: alternative is ready before deprecation occurs.

Inter-Team Bridge: Standard for Deployment

Unit Tests (Deterministic)

- Relies on stationary past benchmarks.
- Operates on pass/fail binary logic.
- Insufficient for pipelines

Release Criteria (Probabilistic)

- Establishes standardized deployment elevation.
- Based entirely on evaluated **statistical performance metrics**
- Requires agreed-upon thresholds from both business and technical stakeholders.

Summary



Find MLOps stack ingredients that fit your product needs.

Use Shadow Deployments to nip product drift in the bud and enable smooth rollback.



Partially Accept the unpredictability of AI systems in order to be prepared

Develop standardized release criteria based on evaluation metrics.

GetReal.

GetReal Security

Digital Trust & Authenticity Platform

GetReal is the world's leading authority on digital trust, authenticity, and continuous audio/visual identity assurance. Purpose-built by experts in cybersecurity, digital forensics, AI, and computer vision, GetReal helps enterprises, governments, and media organizations stop AI-powered identity attacks before the damage is done.

Detect

Deepfakes &
synthetic media

Verify

Real-time identity
assurance

Protect

Adaptive security
policies

Powered by cybersecurity & digital forensics expertise · Incubated by Ballistic Ventures

THE THREAT IS REAL

62%

of organizations experienced a deepfake attack in the past 12 months

Source: Gartner

41%

of enterprises unknowingly hired a physical imposter

Source: GetReal Study

25%

of all job applicants by 2028 will be fake

Source: Gartner

getrealsecurity.com

GetReal.

GetReal Security

Digital Trust & Authenticity Platform

LEARN MORE



Visit Our Website

getrealsecurity.com



Follow Us on LinkedIn

GetReal Security

The only multimodal identity defense that verifies who's really on the other side of the screen.