Testing Strategies for Deployed ML Models

After deploying an ML model into production, various tests ensure its stability, performance, and accuracy. These validate the model's real-world performance and monitor its behavior over time.

by Rahul Shetty

Integration Testing

Purpose

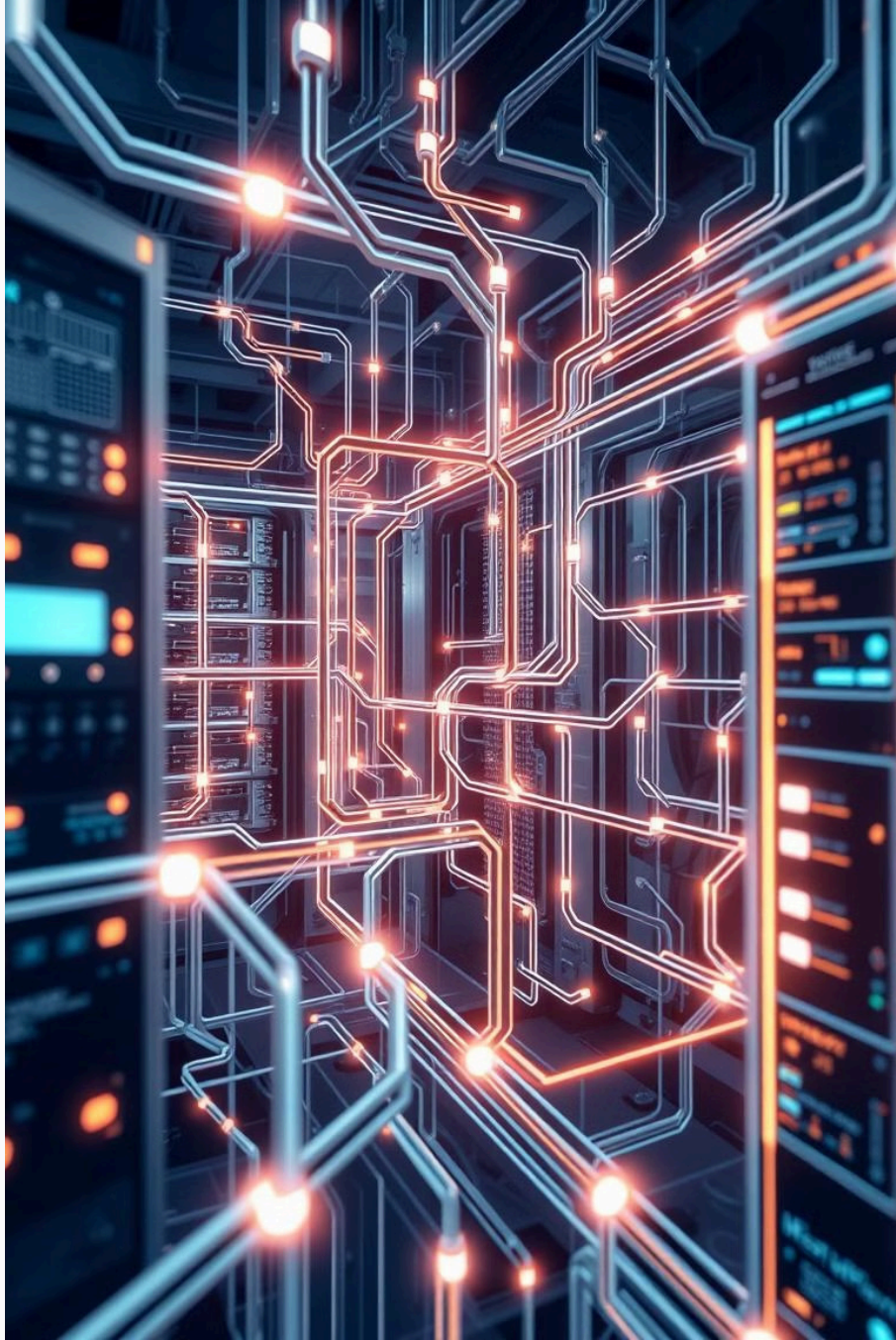
Ensures smooth integration with the larger production system.

Process

QA teams verify effective communication between model and system components.

Key Aspects

Correct data flow, API functionality, and input/output data format consistency.



Latency Testing

1

Purpose

Ensures predictions are made within an acceptable time frame.

2

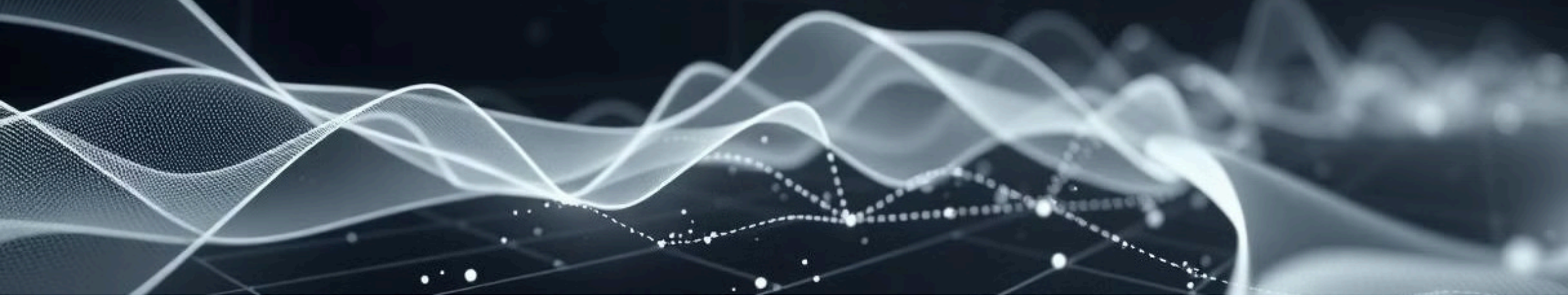
Process

Measure prediction time and ensure it meets defined SLAs.

3

Key Aspects

Response time under normal and peak load, scalability testing.



Drift Testing

Data Drift

Changes in distribution of input data.

Concept Drift

Changes in relationship between input and output.

Monitoring

Set up alerts for changes in data distributions or accuracy.

Shadow Testing

1

Purpose

Run new model in shadow mode without affecting real-world applications.

2

Process

Test with live data, log and compare predictions with production model.

3

Key Aspects

Validate performance on live data, identify discrepancies, ensure expected behavior.

A/B Testing

Purpose

Compares new model performance against existing model with real users.

Process

Split user traffic between current (A) and new (B) models.

Key Aspects

Compare metrics, evaluate user interactions, ensure statistical significance.

Continuous Monitoring

Real-time dashboards	Track accuracy, precision, recall, latency
Data drift monitoring	Trigger alerts for performance drops
Automated retraining	Update model with new data patterns