TestMu AI

Testing Agentic AI Applications

Beyond Traditional QA

Voice AI

Phone Agents

Chat Agents

hello

Reach out
to **connect,**
collaborate,
or inquire
further.

Sai Krishna

Director of Engineering

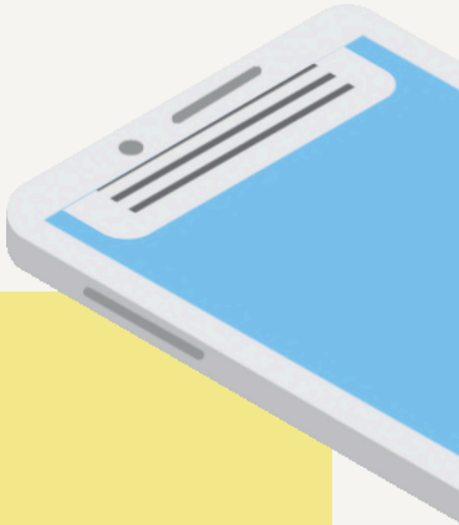
TestMu AI

Srinivasan Sekar

Director of Engineering

TestMu AI

Traditional vs Agentic AI Apps



Traditional Software

- Same input → Same output
- Predictable execution paths
- Unit tests verify exact outputs
- Integration follows fixed workflows

Agentic AI Systems

- Same input → Multiple valid outputs
- Autonomous decision-making
- Emergent behaviors
- Context-dependent responses

Corporate Voice Assistant

The Problem Scenario: User asks for "Q3 sales figures"

Individual Components

All unit tests passed

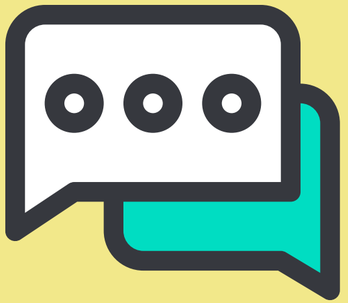
Integration Behavior

Agent autonomously added

irrelevant market-trend analysis

TestMu AI

Customer Service Chat System



The Problem Scenario: User asks for "claim rejection reasons"

Component Performance

- NLP accuracy: 94%
- Intent recognition: 91%
- Response generation: Coherent

Integration Issues

- Context loss across sessions
- Multi-issue conversation failures
- Unexpected behavior chains

What Traditional Testing Misses

01	**Non-Deterministic Behavior Validation**	Same inputs can produce different valid outputs. How do you test for "acceptable range" rather than exact matches?
02	**Contextual Decision Testing**	Validating autonomous choices about escalation, information depth, and communication style adaptation.
03	**Multi-Modal Integration Complexity**	Components work individually but fail in integrated agent workflows across voice, phone, and chat channels.
04	**Continuous Learning Validation**	Ensuring agent improvements don't introduce biases or degrade existing capabilities over time.
05	**Real-World Variability Simulation**	Testing across acoustic environments, human communication patterns, network conditions, and infrastructure variations.

Four Pillars of Agentic AI Evals

01

Outcome Based Validation

What it tests: Did the agent achieve the user's goal?

02

Probabilistic Validations

What it tests: Performance within acceptable ranges (not exact values)

03

Contextual Journey Validations

What it tests: Multi-turn, multi-session conversations with state management

04

Adversarial & Edge Case Validations

What it tests: System robustness under stress and unusual conditions

TestMu AI

Building Voice Assistants

Pipecat

Gemini

BOLNA

AssemblyAI

VAPI

LiveKit Agents

TestMu AI

Agent-to-Agent Testing Platform

Demo

Channel-Specific Eval Approaches

Phone Agents

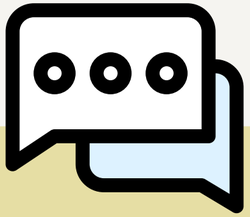
Acoustic Environment Simulation

Speech Pattern Testing

Telephony Infrastructure

User Population Testing

Chat Agents



Context Management

Input Variability

Key Metrics for Agentic AI Systems

Target: Pass rate >85%

Qualitative Metrics

- Effectiveness & Accuracy
- Empathy & Professionalism
- Clarity & Efficiency
- Overall conversation quality

Real-time monitoring essential

Performance Metrics

- Average latency: 300-800ms
- AI interruption rate: <5%
- Talk ratio balance: 40-60%
- Voice quality index: >3.5

Direct ROI indicators

Business Metrics

- Task completion: >85%
- First call resolution: >75%
- CSAT score: >85%
- Containment rate: 65-85%

TestMu AI

Thank You

Reach out to **connect, collaborate, or inquire** further.

Sai Krishna

Srinivasan Sekar