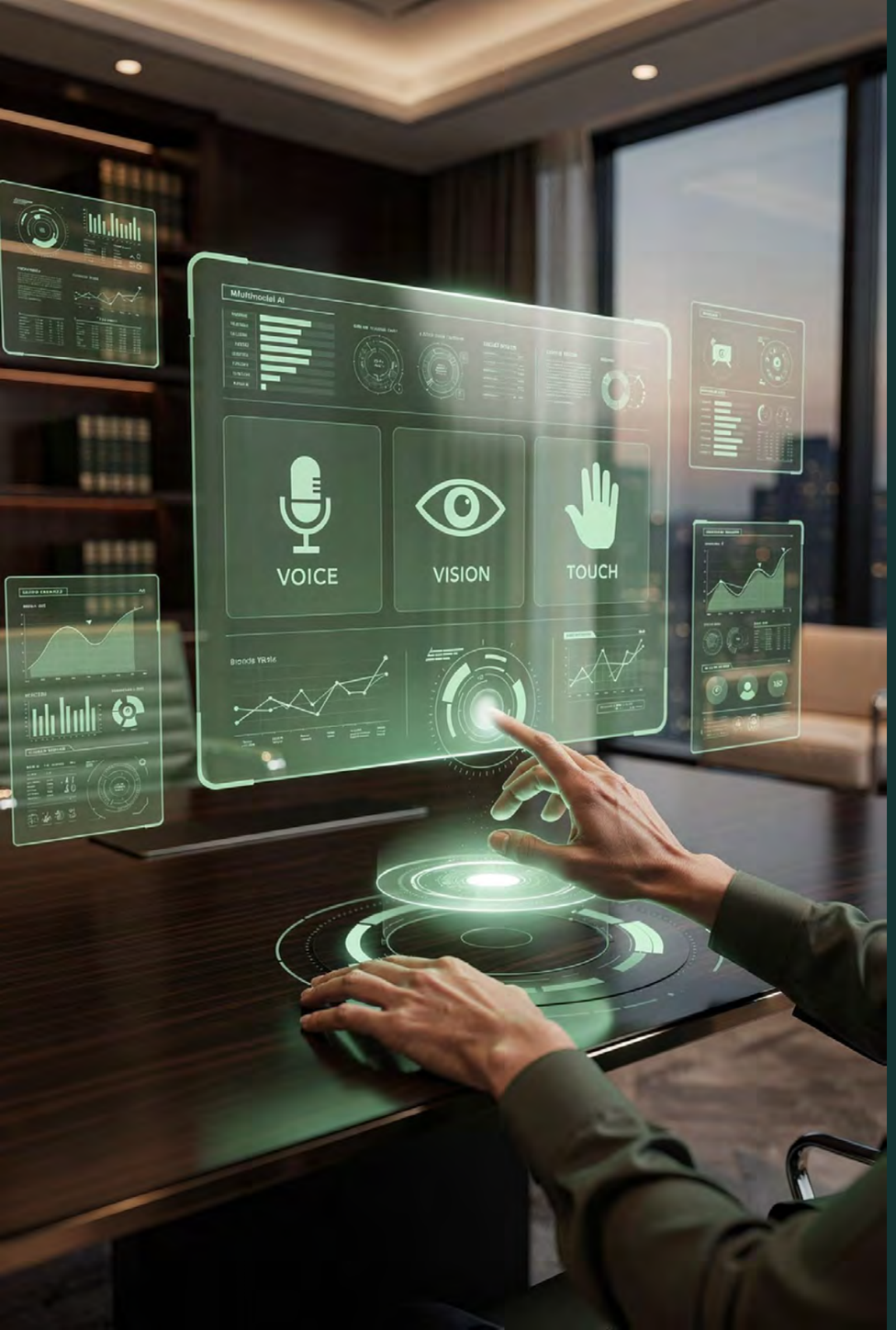


From Demo to Deployment: Building, Evaluating, and Observing Multimodal Conversational Agents

By : Gobu Natarajan

Amazon

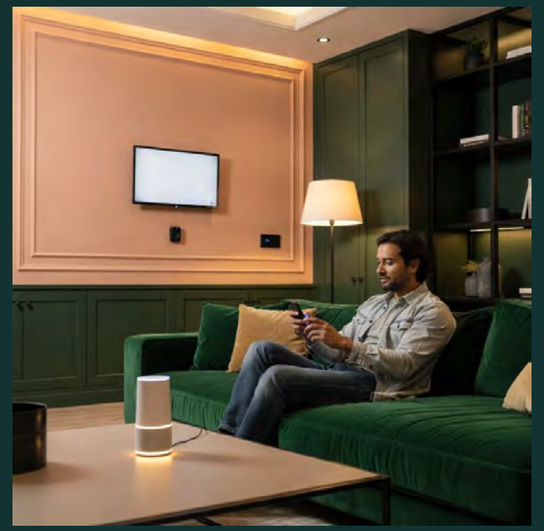
Conf42 DevOps 2026

INTRODUCTION

The Multimodal Revolution in Conversational AI

Multimodal conversational agents that seamlessly combine voice, vision, and touch are rapidly transitioning from experimental novelty to production-grade workloads.

Field research across 250 households demonstrated a **37% improvement in task completion** for complex information-seeking scenarios when users engaged through multiple modalities versus single-mode interfaces.



Within approximately three weeks, longitudinal behavioural data revealed users naturally shifting towards multimodal commands as they discovered richer, more intuitive interaction capabilities.

AGENDA

What We'll Cover Today

01

Reference Architectures

Core components for production multimodal pipelines

02

Fusion Strategies

Early, late, and hybrid approaches with operational trade-offs

03

MLOps & LLMOps Patterns

Evaluation harnesses and observability frameworks

04

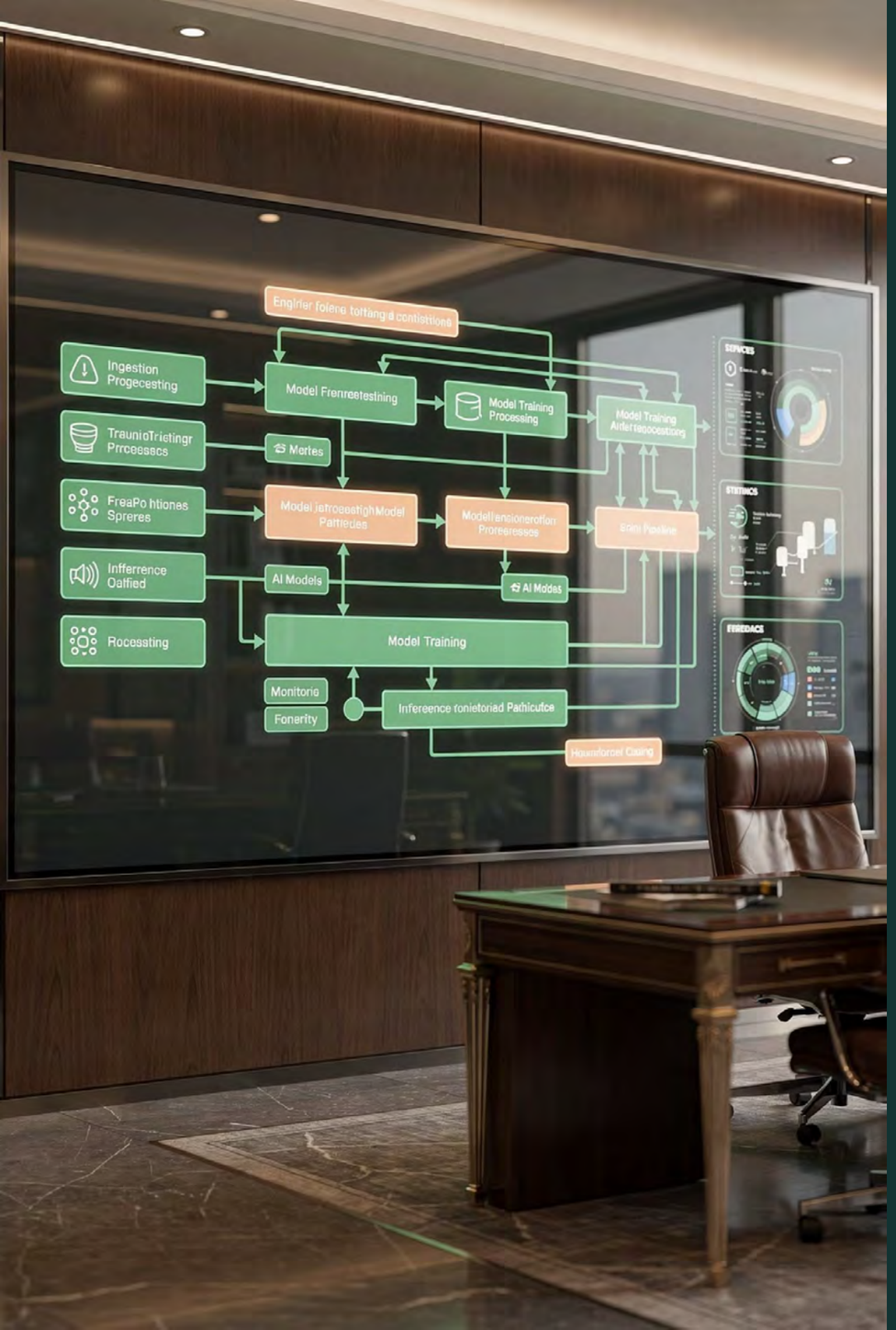
Production Safeguards

Incident-ready fallbacks and privacy boundaries

05

Real-World Applications

Deployment insights across smart home, automotive, wearables, AR/VR



ARCHITECTURE

Multimodal Pipeline Reference Architecture

Ingestion

Audio, video, touch, sensor data capture with timestamp synchronisation

Fusion

Multimodal signal combination using early, late, or hybrid strategies

Reasoning

LLM-powered intent understanding and context building

Action

Response generation and system integration for task execution

Each component must be independently scalable and fault-tolerant, with clear observability hooks for production monitoring and debugging.

FUSION STRATEGIES

Comparing Multimodal Fusion Approaches

Early Fusion

Combine signals immediately at the input layer before processing

- Lower latency for synchronous inputs
- Requires aligned timestamps
- High compute at ingestion

Late Fusion

Process independently and merge at decision layer

- Flexible modality handling
- Easier failure isolation
- Higher latency overhead

Hybrid Fusion

Selective combination based on context and complexity

- Optimises for cost and latency
- Increased system complexity
- Best for production workloads

Operational Trade-Offs in Fusion Strategy Selection

Key Decision Factors

1. **Latency Budgets**

   Real-time requirements for voice and video streams

2. **GPU & Edge Placement**

   Compute distribution between cloud and device

3. **Streaming vs Batch**

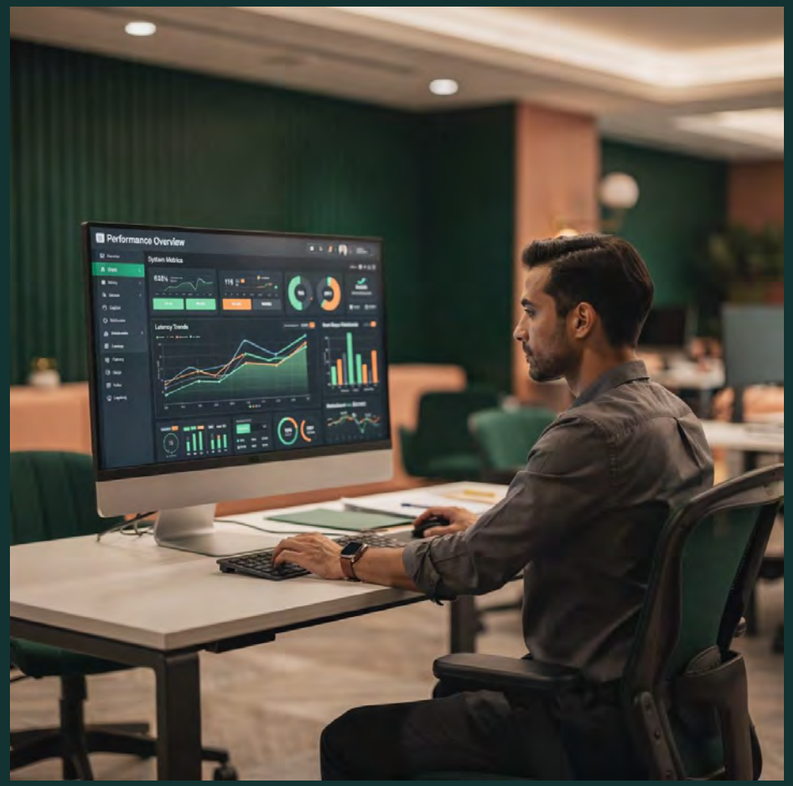
   Processing model impacts fusion timing

4. **Failure Modes**

   Graceful degradation when modalities fail

5. **Cost Controls**

   GPU utilisation and data transfer costs



EVALUATION

Beyond Text: Evaluation Harnesses for Multimodal Systems

Traditional text-only metrics fall short for multimodal agents. Production evaluation requires comprehensive harnesses that measure cross-modal performance and real-world task success.

- **Task Success Rate**

  End-to-end completion of user intents across modalities

- **Grounding Accuracy**

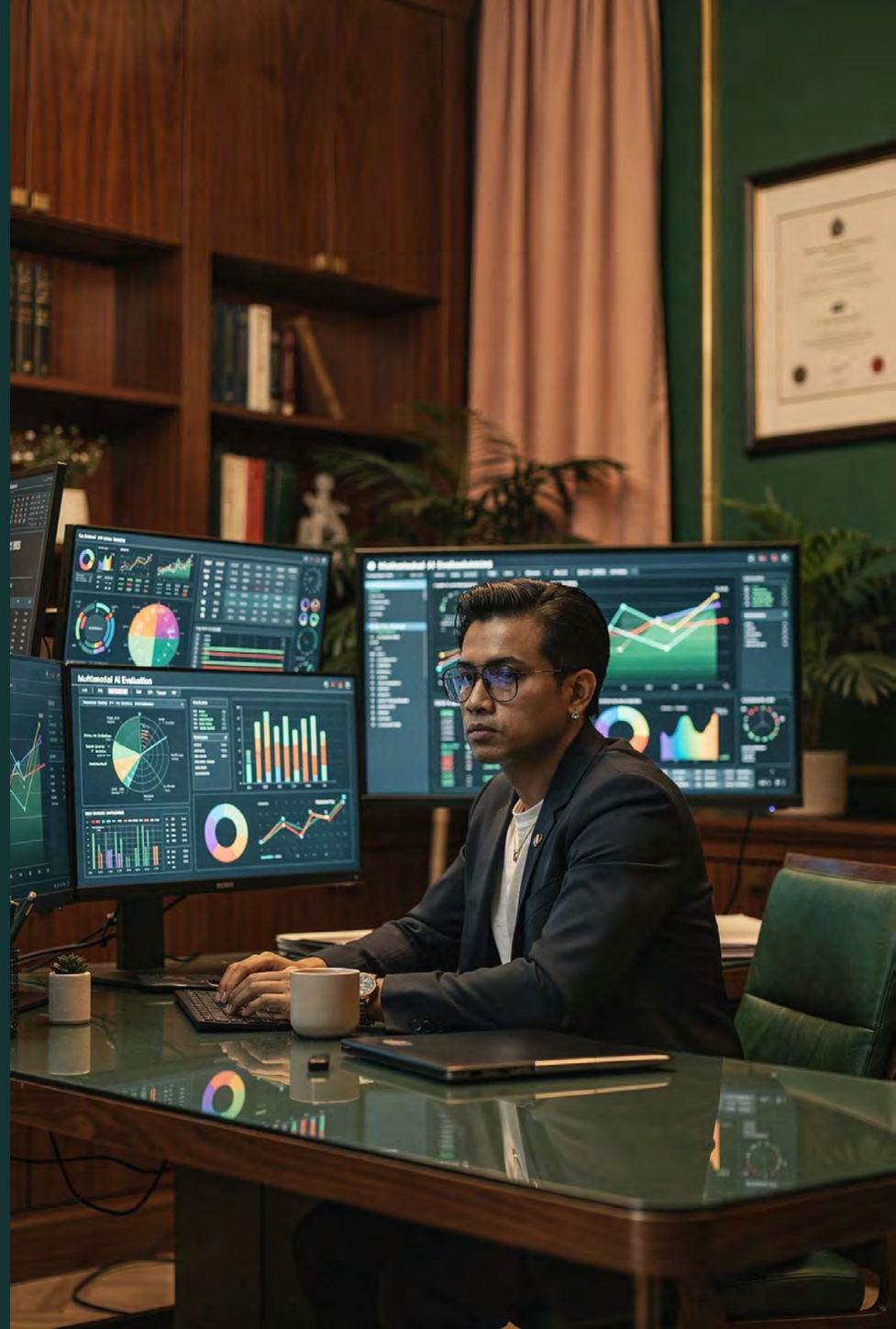
  Correctness of object recognition and spatial reasoning

- **Modality Agreement**

  Consistency between voice, vision, and touch signals

- **Temporal Alignment**

  Synchronisation quality across input streams



OBSERVABILITY

Production Observability for Multimodal Pipelines

Cross-Modal Tracing

Distributed traces capturing audio, video, and touch events with correlation IDs for debugging latency and dropped signals

Drift Detection

Statistical monitoring for input distribution shifts, model performance degradation, and behavioural anomalies

Compute Attribution

Per-modality resource consumption tracking for cost optimisation and capacity planning

Effective observability provides real-time visibility into system health, enables rapid incident response, and supports data-driven optimisation of multimodal agent performance at scale.

SAFEGUARDS

Incident-Ready Safeguards for Production

Single Modality Fallback

Automatic degradation to voice-only or touch-only when fusion fails, maintaining core functionality during partial system outages.

Graceful Degradation

Reduced capability modes that preserve essential features whilst limiting resource consumption during high load or component failures.

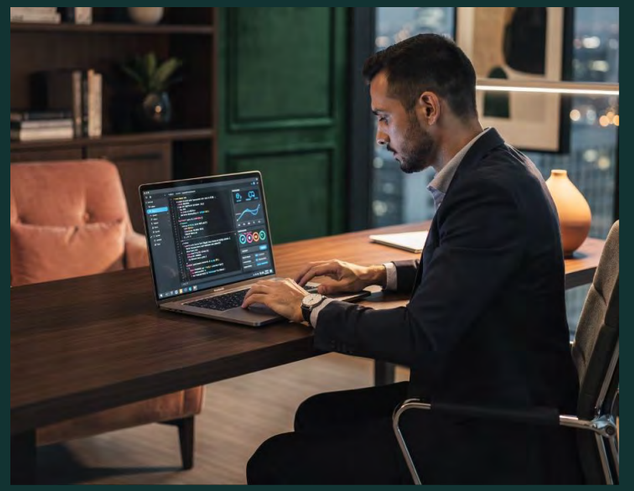
Privacy Boundaries

Strict access controls and encryption for audio and video data, with configurable retention policies and user consent management.

MLOps & LLMOps Patterns for Multimodal Systems

Model Management

- **Version control** for audio, vision, and language models with synchronised deployment
- **A/B testing frameworks** supporting multimodal experiments and phased rollouts
- **Model registry** with metadata for modality-specific performance characteristics



Data Operations

- **Multimodal dataset versioning** with timestamp alignment validation
- **Synthetic data generation** for edge cases and rare modality combinations
- **Privacy-preserving pipelines** with automated PII detection and redaction

APPLICATIONS

Production Deployments Across Domains

Smart Home

Voice, gesture, and screen interactions for lighting, climate, and entertainment control with sub-second response times.

Automotive

Driver attention monitoring combined with voice commands and touch interfaces for safer in-vehicle experiences.

Wearables

Compact form factors leveraging voice and haptic feedback for health monitoring and contextual notifications.

AR & VR

Spatial computing with gaze tracking, hand gestures, and voice creating immersive multimodal environments.

From Prototype to Production: Key Success Factors

- Establish Baseline Metrics

  Define success criteria for each modality and cross-modal performance before scaling.

- Implement Progressive Rollout

  Deploy to limited user segments with feature flags and comprehensive monitoring.

- Build Feedback Loops

  Capture user corrections and system failures to continuously improve model performance.

- Optimise Cost Structure

  Balance GPU compute, edge processing, and cloud resources based on usage patterns.

- Maintain Operational Readiness

  Establish runbooks, incident response procedures, and on-call rotations for multimodal systems.

KEY TAKEAWAYS

Essential Principles for Multimodal Agent Operations

- **Architecture Matters**

  Choose fusion strategies based on latency, cost, and reliability requirements, not theoretical performance.

- **Measure What Matters**

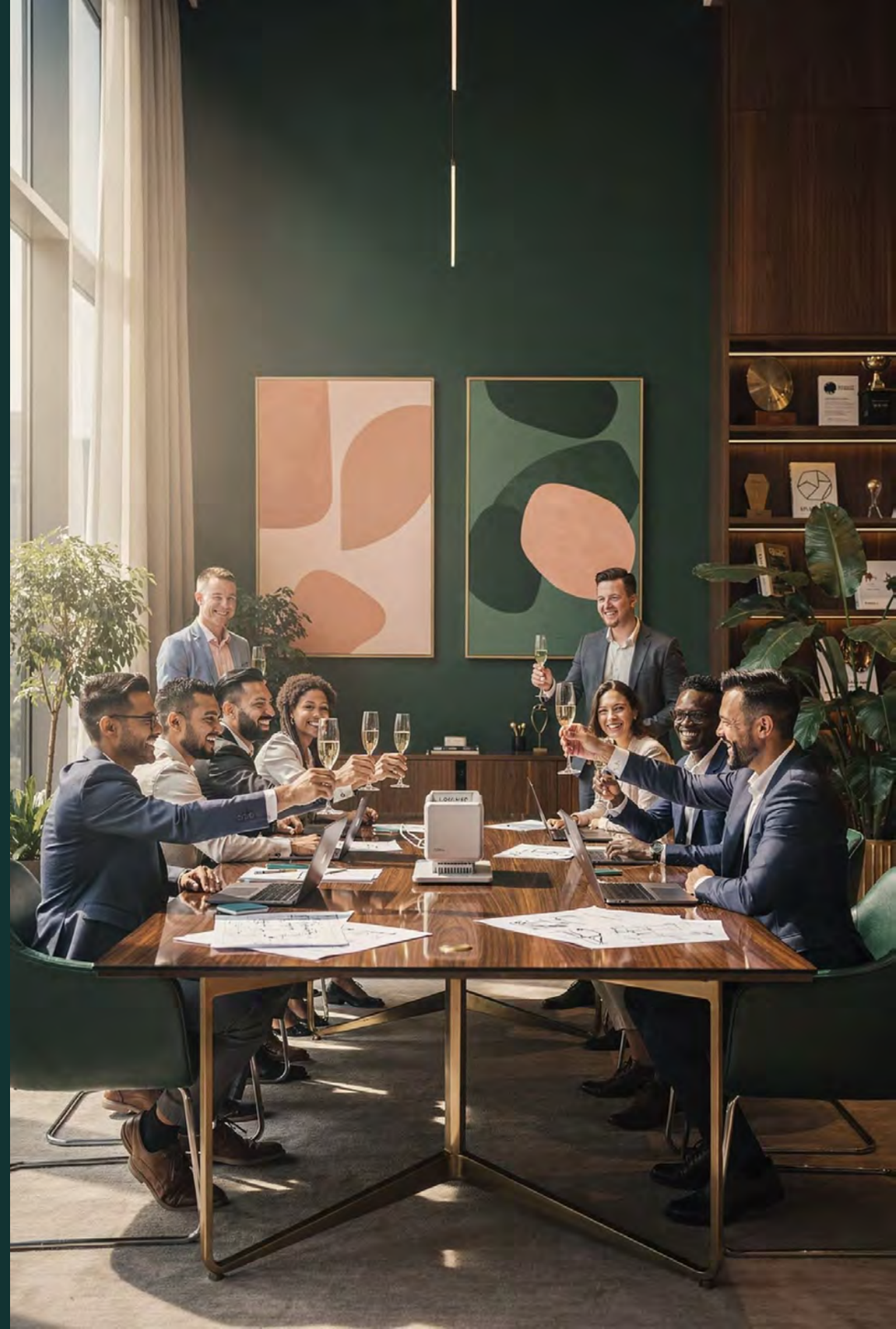
  Implement evaluation harnesses that capture cross-modal task success, not just per-modality accuracy.

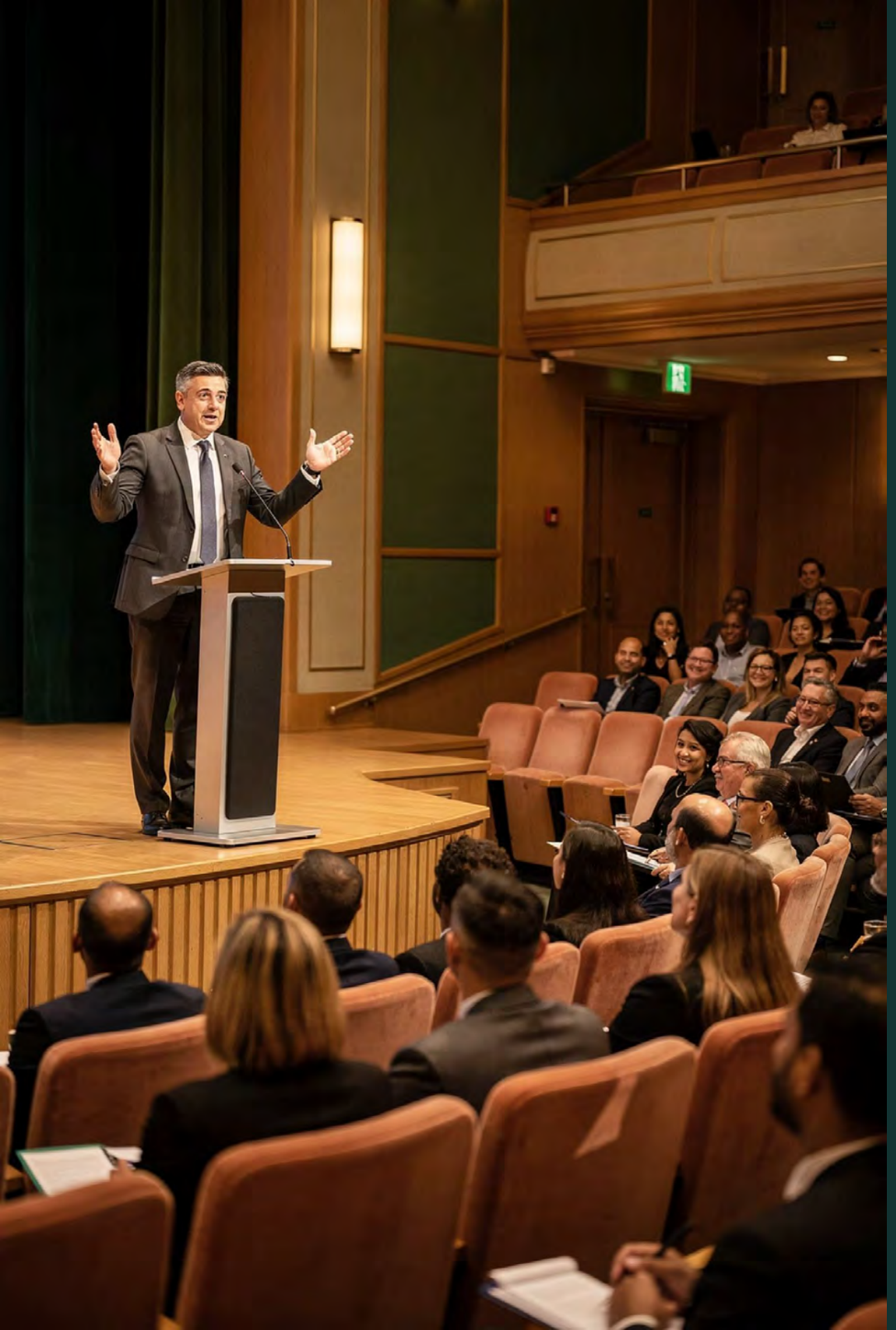
- **Observe Everything**

  Build comprehensive observability with cross-modal tracing, compute attribution, and drift detection from day one.

- **Plan for Failure**

  Design fallback mechanisms and graceful degradation before incidents occur, not during them.





Thank You!

Questions?

Gobu Natarajan

Amazon

Conf42 DevOps 2026

Let's discuss building production-ready multimodal conversational agents that your teams can deploy with confidence.