LLM Agents as Co-Pilots for Scalable Decision Systems in Production Environments

Moving beyond experimentation - a practical architecture for embedding LLM-driven agents into real-world decision workflows at scale.

*Opinions are mine and not representative of Walmart

About the Speaker

Mazdul Hasan Choudhury
Staff Product Manager, Walmart Inc.

Mazdul Hasan Choudhury is a Staff Product Manager at Walmart with over 15 years of experience building enterprise products across retail operations, supply chain, and fulfillment.

He has shipped GenAI-powered tools, redesigned onboarding systems that cut ramp time by 80%, and led mobile commerce platforms used by thousands of associates.

A Boston University alumnus with an MBA and MS in Information Systems Management, Mazdul brings a practitioner's lens to the intersection of AI, product strategy, and operational scale.

The Problem

Most LLM Deployments Stop Short

Organizations have broadly adopted LLMs for chat interfaces and isolated automation tasks. The harder - and more valuable - challenge is different:

Reliable at Scale

Consistent outputs across thousands of decisions per day

Decision-Oriented

Integrated into operational workflows, not just interfaces

Production-Grade

Latency, cost, and stability constraints fully respected

The Co-Pilot Paradigm

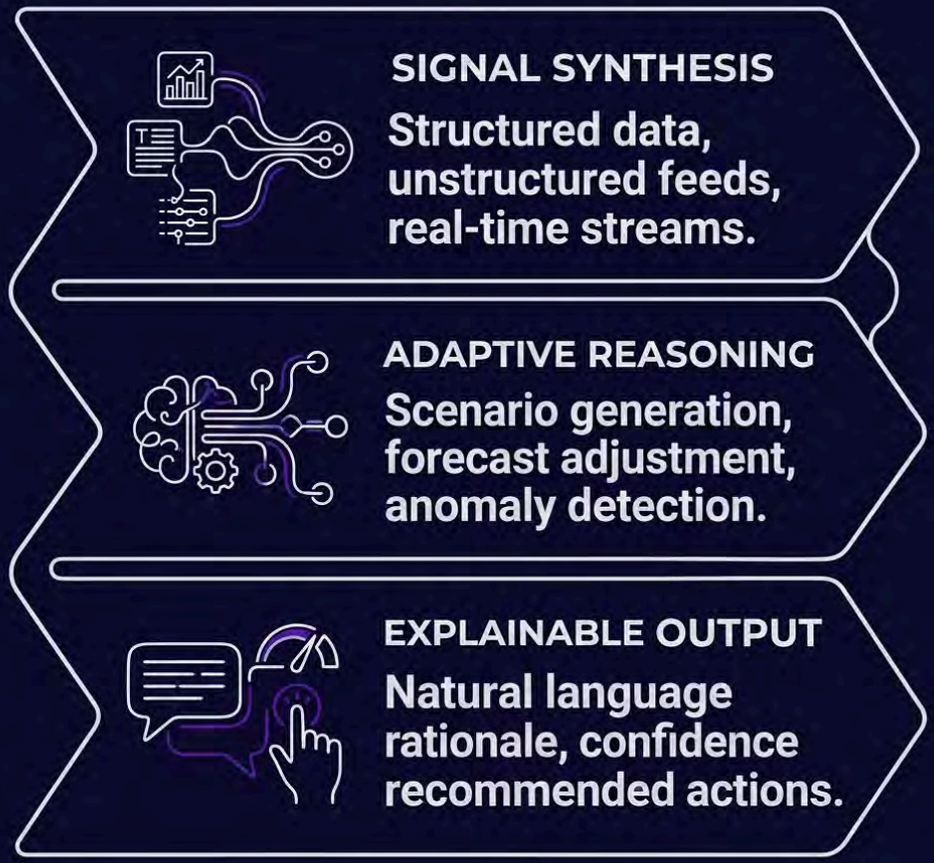
LLM agents don't replace human expertise - they **augment** it. By reducing cognitive load and accelerating scenario evaluation, agents act as intelligent co-pilots within complex decision systems: always on, always reasoning, never overriding human judgment.

Representative Use Case

Forecasting & Operational Planning

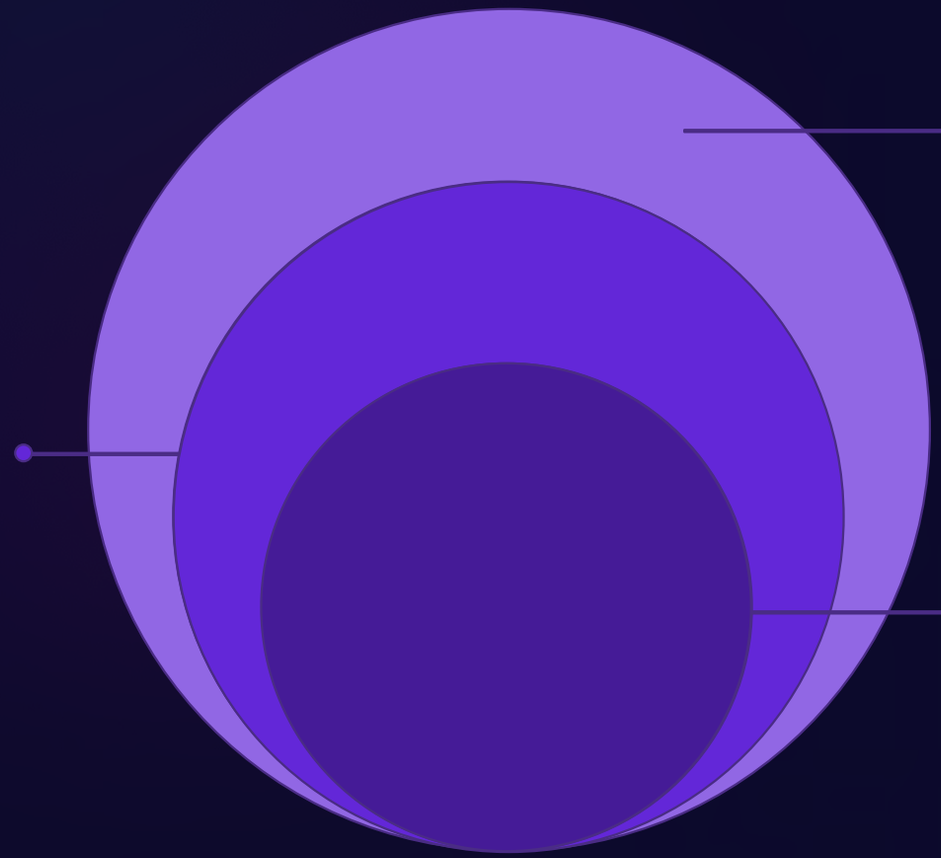
Forecasting and operational planning serve as a concrete testbed for LLM-agent integration. These domains require continuous synthesis of structured and unstructured signals, adaptive reasoning, and explainable recommendations - exactly where agents add the most value.

> The architectural patterns introduced here generalize across domains: supply chain, risk management, capacity planning, and more.



ARCHITECTURE

A Three-Layer Production Architecture



Human Interaction

Review, validate, and intervene as needed

LLM Agent Reasoning

Contextualize signals and generate actions

Signal Aggregation

Collect and normalize data streams

Each layer is independently addressable and designed for resilience - enabling teams to evolve one layer without destabilizing the others.

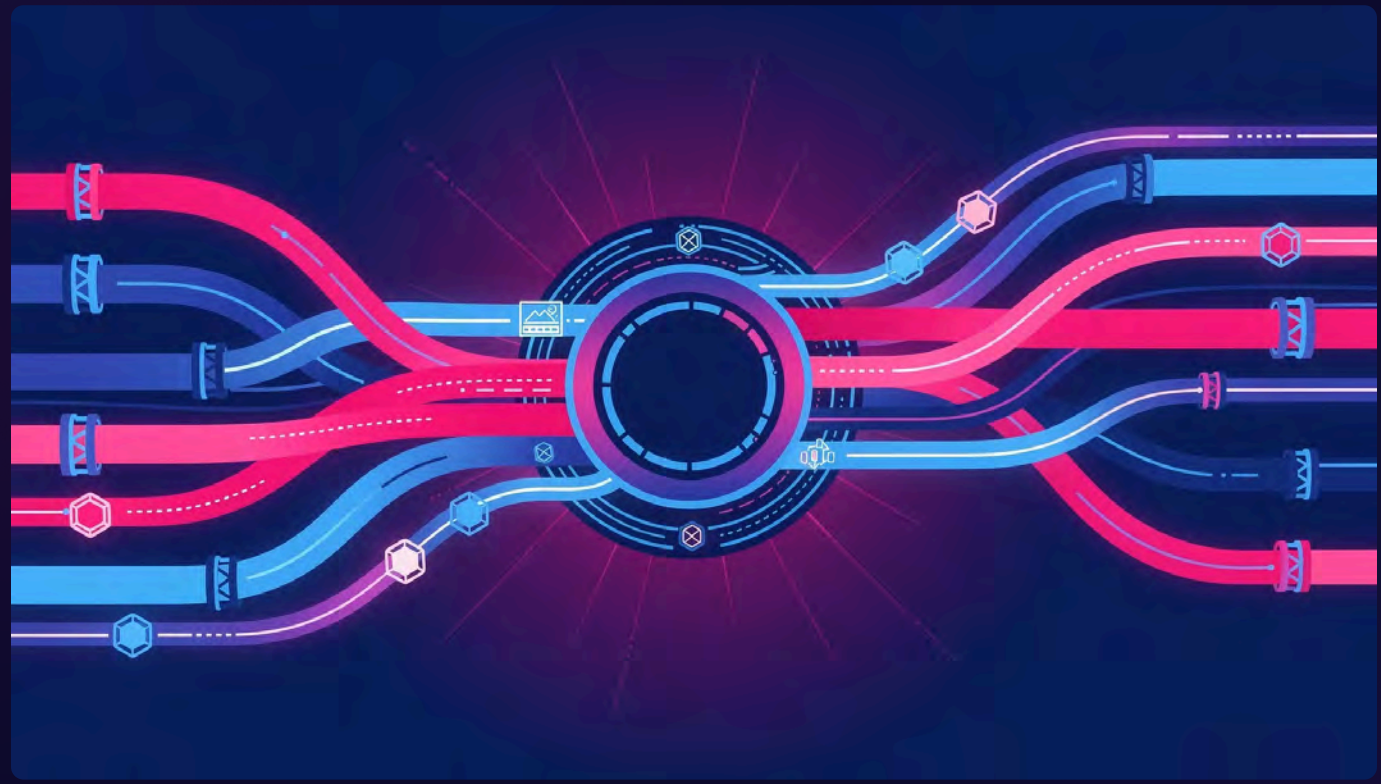
LAYER 1

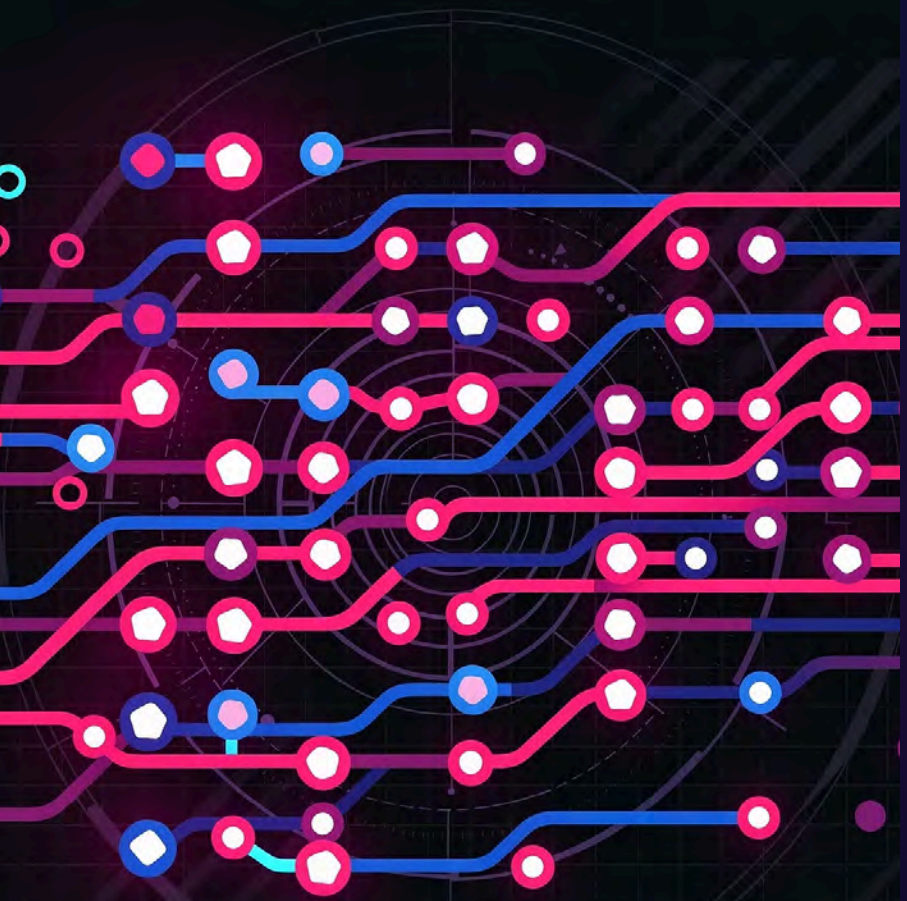
Signal Aggregation

Unified Context Creation

Structured and unstructured data sources are continuously processed into a coherent, queryable context layer. This is the foundation everything else depends on.

- Structured feeds: time-series, transactional, operational databases
- Unstructured feeds: news, reports, internal documents, logs
- Continuous ingestion with normalization and versioning





LAYER 2

LLM Agent Reasoning

Adaptive Forecasts

Agents reason over aggregated context to generate and revise probabilistic forecasts dynamically

Action Recommendations

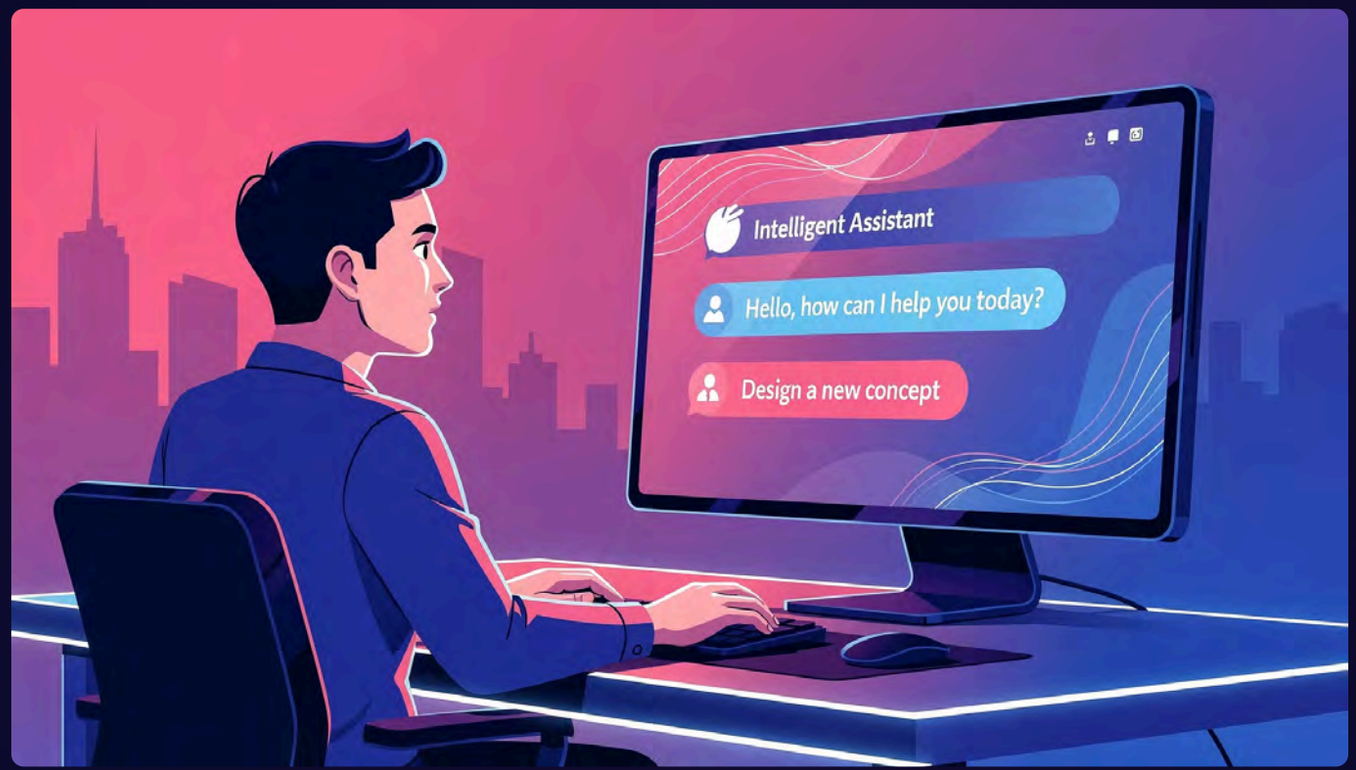
Ranked, explainable action suggestions grounded in enterprise data and current operational state

Natural Language Explanations

Every output is accompanied by a rationale - bridging the gap between model inference and human trust

LAYER 3

Human Interaction Layer



The interaction layer exposes the system's reasoning through conversational interfaces - enabling operators, analysts, and planners to:

Query the System

Ask free-form questions about forecasts, anomalies, or recommendations

Evaluate Scenarios

Run what-if analyses and compare outcomes across planning horizons

Guide Decisions

Provide feedback that loops back into agent context and future reasoning

CHALLENGES

Production Realities: The Hard Problems

Hallucination Management

LLMs can generate plausible-sounding but incorrect outputs. Grounding every response in retrieved enterprise data is non-negotiable.

Decision Consistency

Non-determinism across runs creates auditability and trust issues. Structured output schemas and temperature controls help, but require deliberate design.

Latency Constraints

Multi-step agent reasoning adds wall-clock time. Async execution, caching, and model-tier selection are essential levers.

Cost Efficiency at Scale

Token costs compound quickly in high-volume systems. Prompt compression, routing to smaller models, and output caching are practical mitigations.

Grounding LLM Outputs in Enterprise Data

Hallucination risk drops dramatically when agents operate within a **retrieval-augmented generation (RAG)** framework anchored to authoritative internal data sources.

- Vector stores for semantic search over documents and logs
- Structured query tools for exact lookups against operational databases
- Confidence scoring to flag low-certainty outputs for human review

Grounding is not a one-time configuration - it requires continuous data freshness and retrieval pipeline monitoring.

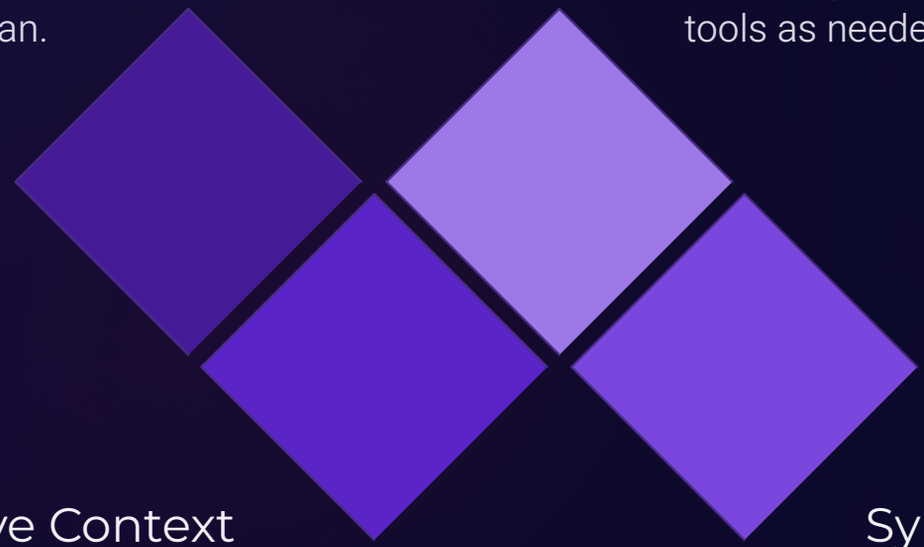
Orchestrating Multi-Step Reasoning

Plan Steps

Define a clear, ordered action plan.

Execute Tools

Invoke agents and external tools as needed.



Retrieve Context

Gather relevant data and past states.

Synthesize Output

Combine results into a final coherent response.

Complex decisions require chained reasoning - not a single prompt. Agent orchestration frameworks manage state, tool calls, and intermediate outputs across multiple inference steps, enabling workflows that parallel human analytical processes.

Integrating Without Disruption

LLM agents must be embedded as **additive layers** - not system replacements. Key integration patterns that preserve operational stability:

Adapter Interfaces

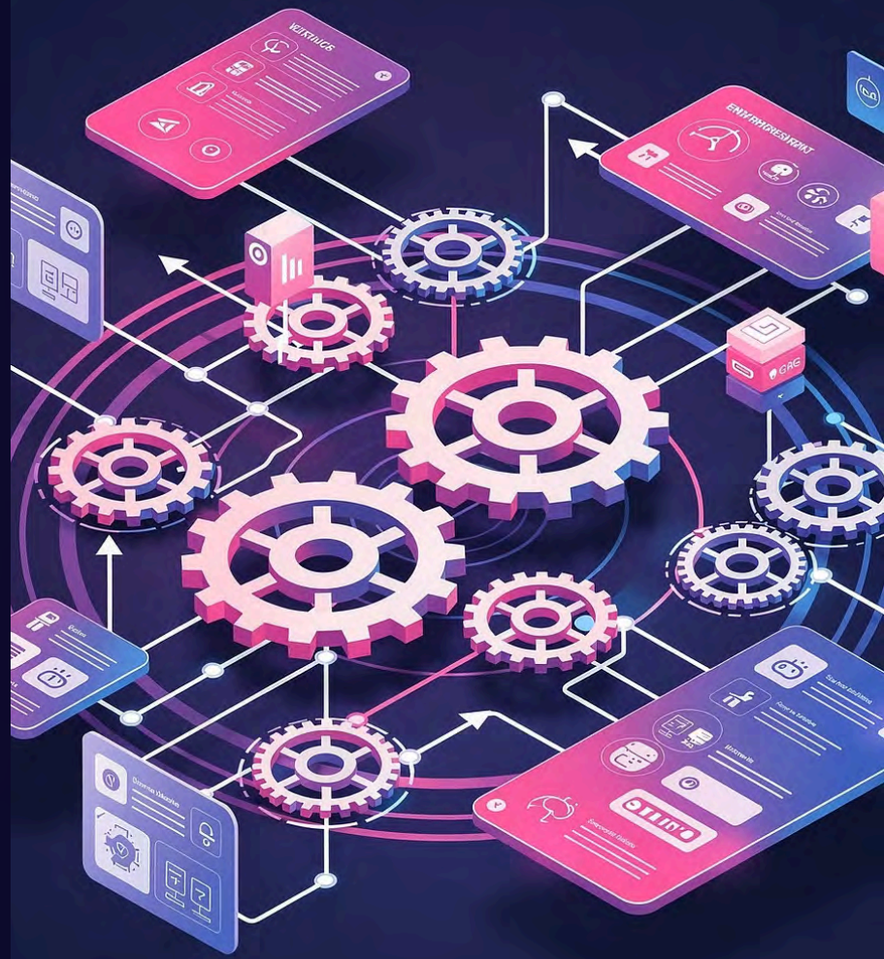
Thin wrappers around existing APIs and data contracts

Shadow Mode Deployment

Run agents in parallel without affecting live decisions during validation

Fallback Logic

Deterministic rule-based fallbacks when agent confidence is below threshold



IMPACT

Augmentation, Not Replacement

Human + Agent

The design goal is a system where human judgment remains authoritative, while agents handle the cognitive heavy lifting of synthesis, pattern detection, and scenario generation.

What Agents Handle

- Continuous monitoring of hundreds of signals simultaneously
- Rapid generation and scoring of alternative scenarios
- Consistent application of decision criteria across all cases
- Natural language summarization for stakeholder communication

What Humans Retain

- Final decision authority and accountability
- Strategic context and organizational judgment

From Experimentation to Measurable Business Impact

1

Prototype

Isolated agent on a single use case; validate reasoning quality and grounding

2

Shadow Deployment

Run alongside production; measure accuracy, latency, and cost against baselines

3

Controlled Rollout

Human-in-the-loop on all agent recommendations; build trust and refine prompts

4

Production Scale

Full integration with monitoring, fallback logic, and continuous evaluation loops

Key Takeaways for Practitioners

1 **Architecture first**

A three-layer design - signal aggregation, agent reasoning, human interaction - provides the structural foundation for production-grade systems.

2 **Ground everything**

Retrieval-augmented grounding in enterprise data is the single highest-leverage technique for production reliability.

3 **Integrate additively**

Shadow mode, fallback logic, and adapter interfaces let you move fast without risking operational stability.

4 **Measure continuously**

Define evaluation metrics before deployment. Accuracy, latency, cost, and human override rates all matter at scale.

Thank You

Mazdul Hasan Choudhury

mazdul@gmail.com

Questions or further conversation?