Microsoft Dynamics 365

Session 1: Agent evaluation framework foundation

Presenter(s):
Akshat Singh
Saurabh Bharati
Vishal Singh

Moderator:
Ashley Desiongco

TechTalk Series

- **Session 1 – Agent evaluation framework foundation**
- Session 2 – Designing Evaluation Sets, Metrics, and the Evaluation Blueprint
- Session 3 – Governance, Lifecycle Gates, and Operating Agents in Production

Agenda

- Agent evaluation framework overview
- Evaluations: Lifecycle & Design Considerations
- Evaluations: Dimensions, Methods and Goals

Microsoft Dynamics 365

Agent evaluation framework

Akshat Singh

Traditional Testing Isn’t Enough for Agents

Success by Design was built for deterministic software. AI agents are fundamentally different.

✓ Traditional D365 Testing

Deterministic	Same input → same output. Always.
Predictable	Forms, workflows & record-change paths
Pass / Fail	Binary, verifiable test criteria
What was built	Configuration tested, not judgment made

✓ **Confirms the system works as configured.**

≠

⚠ AI Agent Behaviour

Non-deterministic	Same input → different outputs. Both valid.
Autonomous	Acts on live data, often without human loop
Drifts over time	Model updates & data changes erode quality
No binary result	Correctness, safety & alignment need rubrics

❌ **Test scripts confirm a response - not that it is correct, safe, or useful.**

What are Agent Evaluations

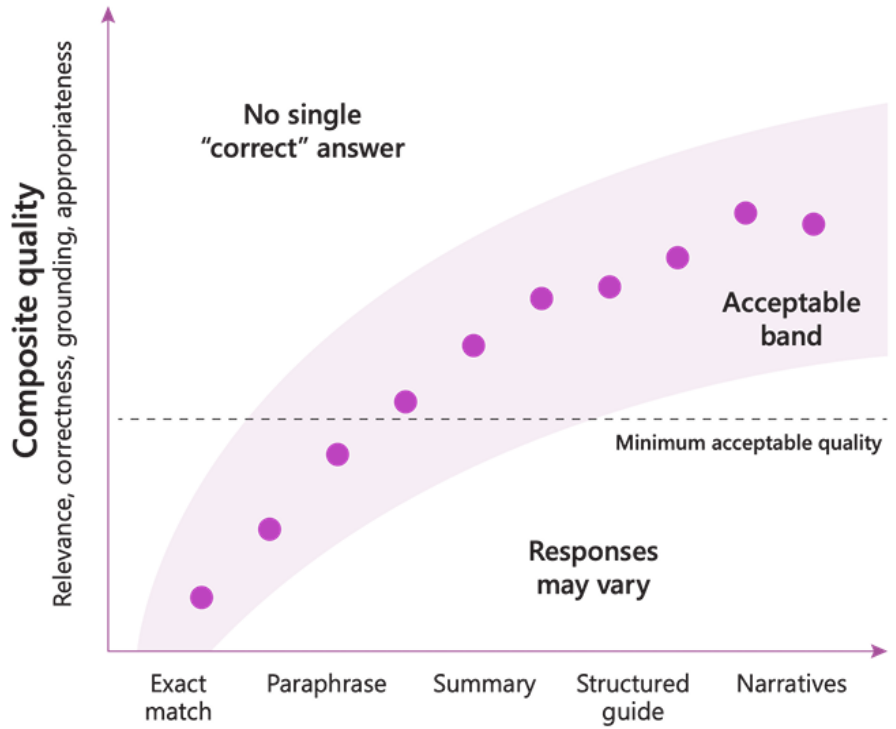
Agent evaluations are the structured process teams use to test how an agent responds to defined inputs — assessing whether those responses meet expectations for quality, reliability, and real-world behavior.

KEY DIFFERENCES FROM TRADITIONAL TESTING

1. ### No single correct answer
   Unlike unit tests, there is rarely one expected output. An agent can produce different valid responses to the same question.

2. ### Responses may vary — and that's OK
   Multiple valid outputs can satisfy the same intent. Evaluation must account for acceptable variation, not just exact matches.

3. ### Quality over exact output matching
   Judged on relevance, correctness, grounding, and appropriateness — a holistic quality score, not a binary pass/fail.



More than one right answer

Agent responses may vary while still meeting expectations. Quality considers relevance, correctness, grounding, and appropriateness — not just exact match.

Cost of Not Evaluating Agents

Business Risk

Wrong records in production

Unhandled order edge cases cause incorrect pricing or shipping addresses in live Dynamics 365 data.

Compliance Risk

PII exposed in outputs & logs

Financial summaries or customer-facing responses leak personal data, creating regulatory liability.

Trust Risk

One failure erodes AI adoption org-wide

A single visible failure in a business-critical workflow sets back stakeholder confidence in AI broadly.

✓ **Every one of these risks is preventable with structured evaluation — catch them before production, detect quickly if behavior changes after Go-Live.**

The Reason to Evaluate

Ensuring Quality Standards

Regular evaluation helps maintain AI agents at high-quality performance and reliability standards.

Early Regression Detection

Continuous assessments detect performance regressions early to prevent failures in production.

Alignment with Goals

Evaluation ensures AI agents stay aligned with changing objectives and business needs.

Risk Mitigation

Proactive evaluation enables risk reduction and improves safety in live environments.

Agent Evaluation Framework

A companion to Success by Design — not a replacement.

Structured. Governance-driven. Lifecycle-wide.

Evaluates AI agents across every stage of the Dynamics 365 delivery lifecycle. Defines what to evaluate, when, how to measure quality across five dimensions, and who is accountable at each delivery gate.

✓ **Not a replacement for Success by Design — a companion.**

WHAT to evaluate

Correctness · Safety · Reliability · Alignment · Efficiency

WHEN to evaluate

At every Success by Design gate — from Initiate through to Operate

HOW to measure quality

Structured rubrics, representative test datasets & production monitoring

WHO decides

Accountable roles and a signed decision documented at each delivery gate

Structured

Defined activities at every lifecycle stage

Governance-led

Ownership, criteria & decisions documented

Risk-based

Rigour scales with deployment stakes

Tool-agnostic

Copilot Studio · Azure · Custom code

Core Principles of the Agent Evaluation Framework

Five principles that govern every evaluation decision across the delivery lifecycle.

Business Value First

Every criterion must connect to a D365 business outcome. Metrics without business context are data, not insight.

Governance from Day One

Evaluation intent, ownership & success criteria defined before the first line of code is written.

Risk-Based Prioritisation

High-risk scenarios — finance, PII, compliance — demand more rigorous thresholds than low-risk automations.

Incremental Deployment

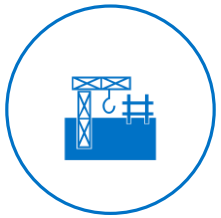
Start limited scope, validate thoroughly, then scale — guided by evaluation results, not timelines.

Continuous Improvement

Agent quality is not point-in-time. Maintain through monitoring, feedback loops & re-evaluation.

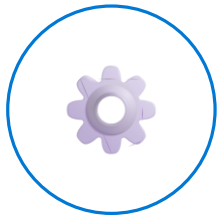
Who This Framework Is For

One framework, one shared language — for every team delivering and sustaining D365 solutions with AI agents.



Solution Architects

Design agent architecture & evaluation strategy for D365 engagements.



Delivery Managers

Oversee implementation quality across the full delivery lifecycle.

Implementation Partners

Build & deploy agents as part of D365 project delivery.

Business Owners & Sponsors

Define what success looks like & own the business outcomes for agent deployments.

Compliance & Governance

Ensure AI deployments meet compliance, privacy & accountability requirements.

📋 **One framework — shared by architects, engineers, partners, business owners & compliance. Same lifecycle, same language.**

Microsoft Dynamics 365

Evaluations: Lifecycle & Design Considerations

Saurabh Bharati

Lifecycle & Design Considerations: What we'll cover

01

Evaluation Lifecycle

How evaluation maps to the five Success by Design delivery phases

02

Evaluation Design Document (EDD)

The central governance artifact — single source of truth per agent

03

Responsible AI Checkpoints

Core RAI criteria and safety gates across every lifecycle stage

04

Quality vs Quantity in Evaluation Sets

Readiness tiers — start with 25–50 high-quality cases

Evaluation Lifecycle Overview

Mapped to the five Success by Design delivery phases

Each phase has a primary evaluation focus, a governance question, and activities that build on the previous stage.

The Evaluation Design Document (EDD)

The central governance artifact of the Agent Evaluation Framework

WHAT IS THE EDD?

Single source of truth

Created at the Design & Ideation stage and updated at every subsequent gate — for evaluation intent, decisions, and evidence across the delivery lifecycle.

WHAT THE EDD CAPTURES

01
Business context & objectives
D365 process, problem, measurable outcomes

02
Intent & success criteria
What "good" looks like with quantitative thresholds

03
Quality dimensions & risk
Critical dimensions and consequences of failure

04
Methods & test data strategy
How evals run, tools used, dataset management

05
Roles & responsibilities
Ownership, go/no-go approvers, production monitors

06
Gate decisions with rationale
Evidence reviewed, decisions, residual risks

EDD: Key Roles and Governance

Not a bureaucratic artifact — a communication tool that aligns all stakeholders

KEY ROLES

1	**Business Owner**	Defines success criteria, approves evaluation intent, owns go-live decisions
2	**Evaluation Lead**	Designs strategy, selects methods/tools, maintains EDD throughout lifecycle
3	**RAI Reviewer**	Validates safety, fairness, and compliance at each gate. Required sign-off for production
4	**Release Authority**	Enforces deployment gates. Accepts residual risk on behalf of the organization
5	**Service Owner**	Monitors production quality, responds to incidents, drives continuous improvement

WHY GOVERNANCE IS DIFFERENT

Agent quality is not static

Traditional D365 deployments reach a stable state after go-live. Agent quality degrades silently as models update, grounding data shifts, or usage patterns evolve.

Someone must own agent quality, monitor it continuously, and have authority to act when it degrades.

Responsible AI Checkpoints

At each stage gate, the following RAI criteria must be addressed explicitly — not assumed

A Core RAI Criteria

Must be addressed in every EDD

Fairness

Consistent performance across customer segments and input formats. Aggregate scores can mask failures in specific subgroups.

Privacy & Security

No PII in logs or outputs. Agent respects D365 security roles and row-level security.

Transparency & Accountability

Reasoning logged and traceable. Users aware they interact with AI. Escalation paths defined and tested.

B RAI Safety Gates

At each lifecycle stage

1. **Design (Initiate)**
   RAI requirements review — assess risk, define acceptable behavior boundaries
2. **Build (Implement)**
   Red-team testing, PII protection validation, fairness assessment across segments
3. **Pre-deployment (Prepare)**
   Production readiness gate — documented evidence of RAI compliance
4. **Production (Deploy/Operate)**
   Continuous monitoring for safety drift and compliance degradation
5. **Incident response**
   Post-incident review gate — root cause analysis feeding back into evaluation sets

Quality vs Quantity in Evaluation Sets

Start with 25–50 high-quality test cases, not "1,000 generic ones"

WHAT MAKES AN EVAL SET "HIGH QUALITY"?

- **Representative + edge/adversarial coverage (risk-based)**
- **Scenario-based — tests real end-to-end tasks**
- **Clear inputs + expected outcomes + criteria**
- **Versioned and evolved with production learnings**

RECOMMENDED BALANCE · *30–50% representative and 30–50% edge/adversarial tasks (depending on risk tolerance).*

READINESS TIERS — MATURITY ROADMAP

TIER 1
Foundational

25–50 cases covering core scenarios. Manual review. Basic correctness validation. Sandbox/pilots.

TIER 2
Developing

100+ cases with edge and adversarial inputs. Automated benchmarking. RAI review documented.

TIER 3
Mature

All five dimensions measured with thresholds. Full regression suite. Production monitoring with alerting. Shadow mode tested.

TIER 4
Optimized

CI/CD-integrated evaluation. Production telemetry feeding test datasets. Full audit trail. Periodic human review.

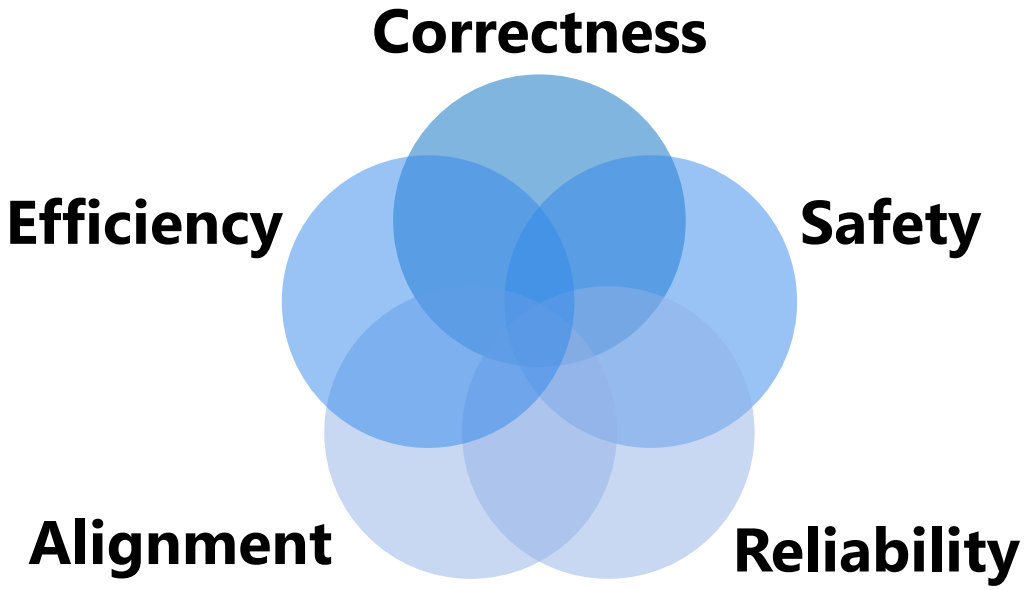
Microsoft Dynamics 365

Evaluations: Dimensions

Vishal Singh

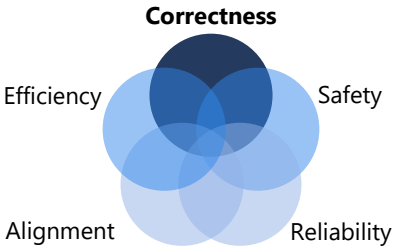
The Five Dimensions of Evaluation

Effective evaluations assess your Agentic AI solution across five interconnected dimensions:



Key Insight: These dimensions are not mutually exclusive. A comprehensive evaluation strategy addresses all five.

Dimension 1: Correctness



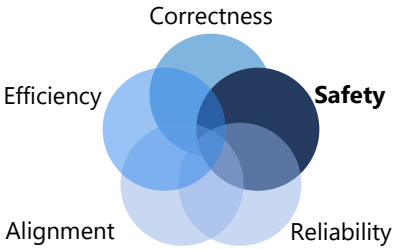
Definition: Does the agent provide accurate, factually correct, and contextually appropriate responses?

Measurement Approaches:

1. Ground truth comparison (comparing outputs to known correct answers)
2. Expert human review with scoring rubrics
3. Automated fact-checking against knowledge bases
4. Consistency checks across related queries

Why It Matters: Foundational for reliability and trust in agent outputs

Dimension 2: Safety



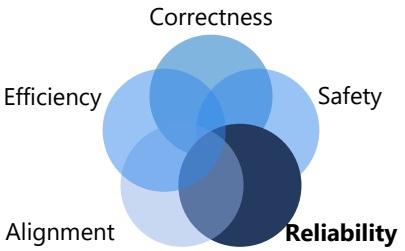
Definition: Does the agent avoid generating harmful, offensive, or inappropriate content and respect security boundaries?

Measurement Approaches:

1. Adversarial testing with known harmful prompts
2. Security penetration testing
3. Content classification and toxicity scoring
4. Privacy leak detection

Why It Matters: Protects users and ensures organizational compliance

Dimension 3: Reliability



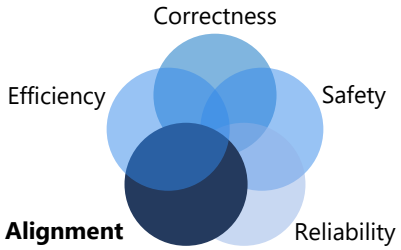
Definition: Does the agent perform consistently, handle errors gracefully, and maintain availability?

Measurement Approaches:

1. Load and stress testing
2. Monitoring of latency, throughput, and error rates
3. Variation analysis (measuring consistency across similar inputs)
4. Chaos engineering (introducing failures to test resilience)

Why It Matters: Ensures consistent performance under various conditions

Dimension 4: Alignment



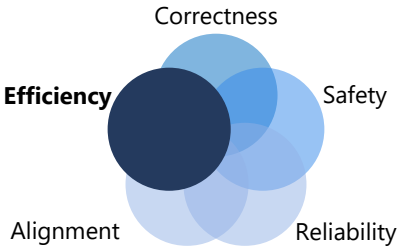
Definition: Does the agent's behavior align with business objectives, brand voice, and organizational values?

Measurement Approaches:

1	Business stakeholder review and scoring
2	Sentiment and tone analysis
3	Policy compliance checking
4	A/B testing against business KPIs (conversion rate, satisfaction scores)

Why It Matters: Ensures user goals are met, and business objectives are achieved

Dimension 5: Efficiency



Definition: Does the agent optimize resource usage, cost, and time to value?

Measurement Approaches:

1. Cost tracking (API calls, compute time, storage)
2. Performance benchmarking
3. Conversation length and resolution time analysis
4. Resource utilization monitoring

Why It Matters: Business usability, user experience, and operational cost control

Microsoft Dynamics 365

Evaluation Methods & Patterns

Vishal Singh

Evaluation Patterns & Methods

Evaluation Patterns

- *Strategy/Approach* and lifecycle placement of evaluation.
- *Define when, where,* and under *what* conditions an agent is evaluated, but not how scoring is done.

Evaluation Methods

- *Technique* used to judge or score an agent's behaviour.
- Define *how* quality is measured, what criteria are used, and who or what performs the judgment.

Common Examples:

Offline Evaluation	Golden Dataset Comparison
Online/In-Production Evaluation	LLM-as-a-Judge Scoring
Continuous Evaluation	Human-in-the-Loop Review
Regression Evaluation	Automated Benchmark Evaluations
A/B Testing	Task/Goal Success Scoring

Offline vs. Online Evaluations: Comparison

	OFFLINE (Pre-Production)	ONLINE (Production)
Test Data	Pre-defined test cases with known ground truth	Real production data with unknown ground truth
Repeatability	Can execute repeatedly without user impact	Must balance coverage with cost and latency
Iteration	Allows thorough debugging and iteration	Uses proxy signals and heuristics
Security	Works with test data	Need customer consent to use if PII info found.
Purpose	Provides evidence for evaluation gates	Catches issues not in test datasets

Both offline and online evaluations are essential. Offline evaluation establishes the baseline and provides controlled validation. Online evaluation ensures continued performance in real-world conditions and catches issues not represented in test datasets.

Evaluation Method: Human-in-the-Loop (HITL)

When to use: For subjective criteria like brand alignment, tone, or complex correctness judgments

How it works

1. Generate a sample of agent outputs (e.g., 100 conversations)
2. Present them to human evaluators (domain experts, stakeholders, crowdworkers)
3. Evaluators score outputs using standardized rubrics
4. Aggregate scores to measure performance

Best Practices

- ✓ Use multiple evaluators per item to ensure reliability
- ✓ Provide clear, specific scoring criteria
- ✓ Calibrate evaluators with training examples
- ✓ Track inter-rater agreement to validate consistency

Evaluation Method: Automated Benchmarks

When to use: For objective criteria with ground truth available (factual accuracy, calculation correctness)

How it works

1. Curate or generate a test set of input-output pairs with known correct answers
2. Run your agent on the test inputs
3. Automatically compare agent outputs to ground truth using appropriate metrics (exact match, semantic similarity, etc.)
4. Report aggregate performance statistics

Best Practices

- ✓ Ensure test sets are representative of production workloads
- ✓ Regularly update test sets to avoid overfitting
- ✓ Use diverse test cases covering edge cases and typical scenarios
- ✓ Version control test sets alongside code

Evaluation Method: LLM-as-Judge

When to use: For scalable evaluation of subjective criteria when human review is too expensive or slow

How it works

1. Use a powerful LLM (e.g., GPT-5) to evaluate your agent's outputs
2. Provide the judge LLM with evaluation criteria and scoring rubrics
3. The judge LLM scores outputs along specified dimensions
4. Aggregate scores for overall assessment

Best Practices

- ✓ Validate LLM-as-judge against human evaluations initially
- ✓ Use specific, detailed prompts for the judge LLM
- ✓ Consider potential biases of the judge LLM
- ✓ Combine with human review for high-stakes decisions

Evaluation Methods comparison

Method	Context	Strengths	Limitations	Best For
Human Evaluation	Offline & Online (sampled)	• Nuanced judgment • Captures subjective criteria • High validity	• Expensive • Slow • Inter-rater variability	Brand alignment, tone, complex correctness
Automated Benchmarks	Offline & Online	• Fast and cheap • Reproducible • Scales easily	• Requires ground truth • May not capture all aspects • Risk of overfitting	Factual accuracy, calculations, retrieval quality
LLM-as-Judge	Offline & Online	• Scalable • Faster than humans • Can handle subjectivity	• Judge LLM biases • Potential errors • Cost of API calls	Alignment, tone, summarization quality

Microsoft Dynamics 365

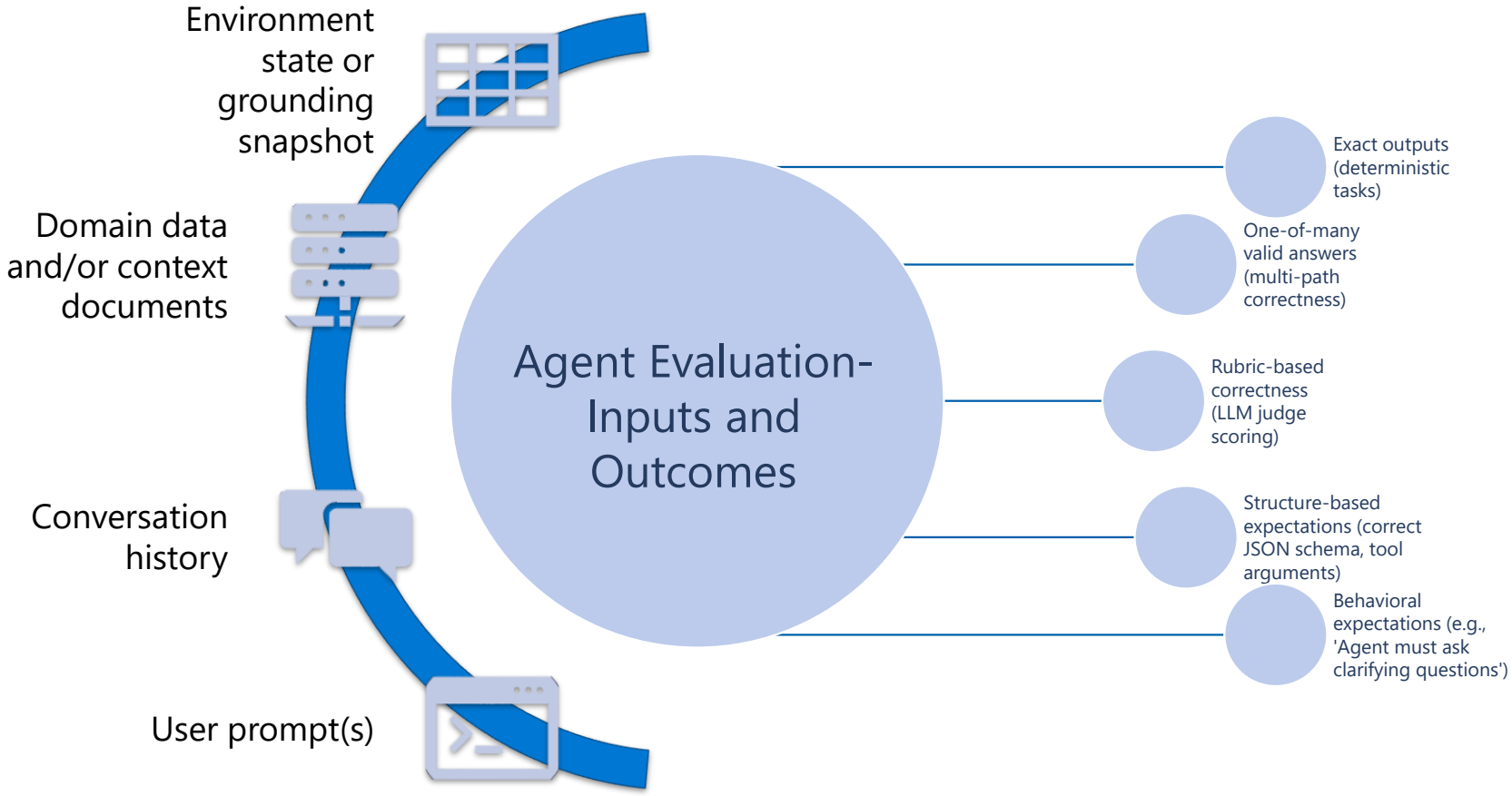
Evaluation: Goals and Criteria

Vishal Singh

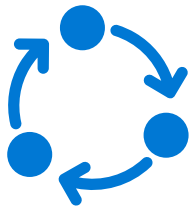
Evaluation Goals & Criteria

Every evaluation scenario must define clear structure and measurable expectations.

Evaluation Criteria:
Measure qualities such as correctness, safety, tone, compliance, and efficiency

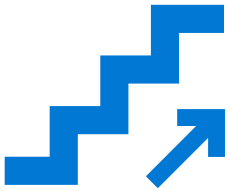


Key Takeaways



Evaluation is a **lifecycle**, not a checkpoint.

RAI is operationalized via **principles + safety gates.**

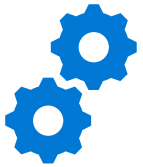


Start with **high-quality eval sets**, then scale + evolve

Key Takeaways

Five Dimensions: Evaluate across Accuracy, Safety, Reliability, Alignment, and Efficiency

Multiple Methods: Combine Human-in-the-Loop, Automated Benchmarks, and LLM-as-Judge for comprehensive coverage



Offline & Online: Both pre-production and production evaluations are essential

Clear Criteria: Define inputs, expected outcomes, and evaluation criteria for every scenario


QUESTIONS

Please share your feedback!

Dankie Faleminderit **Shukran** Chnorakaloutioun Hvala Blagodaria

Děkuji **Tak** Dank u Tänan Kiitos **Merci** Danke Ευχαριστώ A dank

Mahalo ההודה. **Dhanyavād** Köszönöm Takk Terima kasih **Grazie** Grazzi

Thank you!

감사합니다 Paldies Choukrane Ačiū **Благодарам** ありがとうございました

谢谢 Баярлалаа **Dziękuję** Obrigado Mulţumesc **Спасибо** Ngiyabonga

Ďakujem Tack Nandri Kop khun **Teşekkür ederim** Дякую Хвала Diolch


Microsoft Dynamics 365