Agent Evaluation – Solution Document

1. Solution Overview

Agent Evaluation is an enterprise-ready, AI-native solution designed to **evaluate and benchmark end-to-end (E2E) AI solutions**. It validates the performance, reliability, and compliance of **LLMs, AI Agents, and complete AI workflows** — covering everything from data retrieval to orchestration, tool integration, reasoning, and user-facing outputs.

The solution leverages a **multi-agent evaluation framework** orchestrated through an **Evaluation Orchestrator Agent** built on Python A2A SDK. It integrates **LangGraph, Ragas, LLM-as-a-Judge techniques, and sandboxed MCP servers** to provide holistic testing across accuracy, robustness, efficiency, safety, and fairness.

Evaluation results are delivered as **traceable metrics, dashboards, and narrative summaries**, enabling enterprises to **benchmark, monitor, and trust AI systems across the full pipeline**.

2. Problem Statement

Enterprises deploying AI solutions face key challenges:

- No consistent way to compare **E2E AI pipelines**, not just individual models.
- Difficulty validating **multi-step reasoning, retrieval, and tool orchestration**.
- Limited ability to detect **hallucinations, unsafe behavior, and bias** across workflows.
- Lack of systematic compliance checks for **Responsible AI in production**.
- Absence of full **observability and traceability** for model, agent, and workflow-level evaluations.

Agent Evaluation addresses these gaps by offering a **centralized platform for comprehensive E2E AI evaluation**.

3. Solution Detail

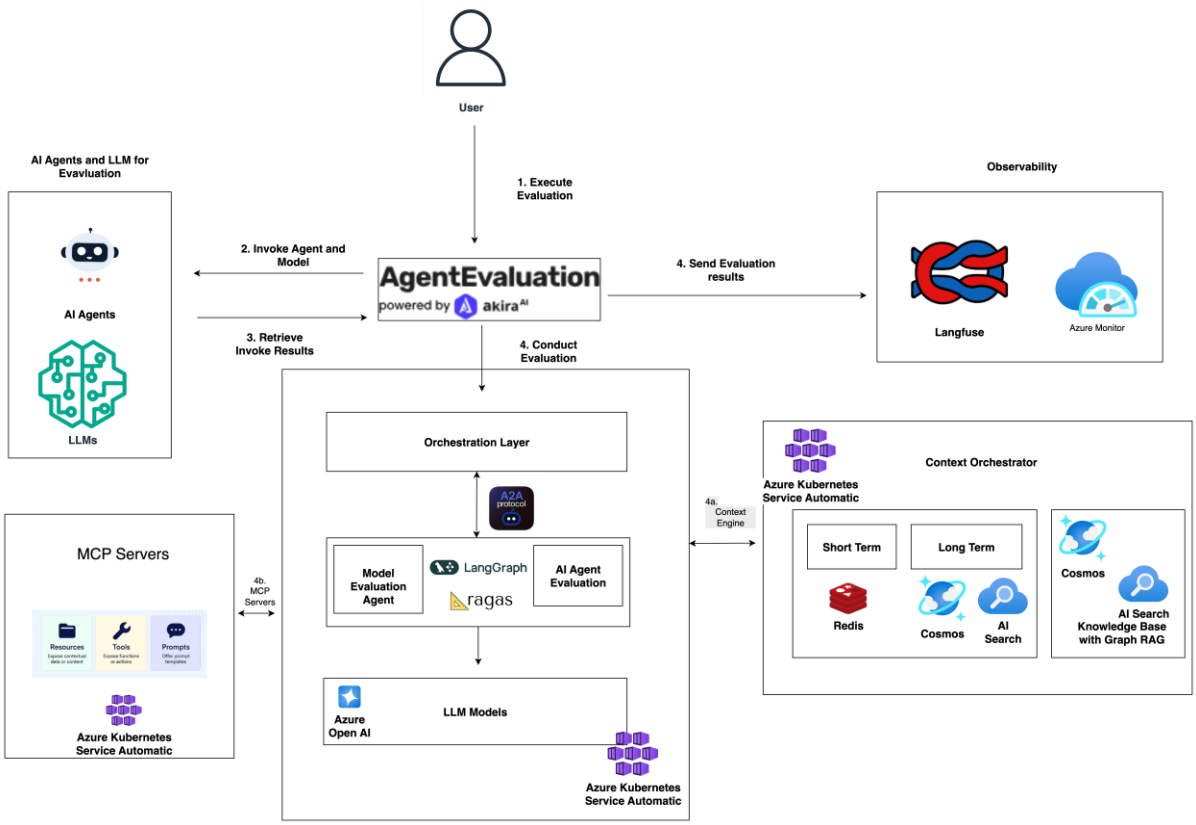
Agent Evaluation functions as a **multi-agent evaluation layer for E2E AI solutions**:

- **Evaluation Orchestrator Agent**: Controls workflows and routes tasks across evaluator agents via A2A protocol.
- **Model Evaluation Agent**: Benchmarks LLMs and other models for factuality, latency, fluency, efficiency, and fairness.
- **AI Agent Evaluation Agent**: Validates reasoning chains, tool usage, and task completion in agent workflows.
- **Solution Workflow Evaluation**: Assesses the full AI pipeline — data retrieval, orchestration, knowledge grounding, and output quality.
- **MCP Servers (Sandbox)**: Provide controlled environments for tool and API testing without touching live systems.

- **Context Orchestrator**: Supplies short- and long-term memory with Redis, Cosmos DB, AI Search, and Graph RAG.
- **Langfuse Observability**: Captures full evaluation traces, metrics, costs, and reports for transparency.
- **Azure-Native Deployment**: Runs on AKS with Azure OpenAI for scalable enterprise execution.

4. Technical Architecture

The architecture includes the following layers (*see diagram*):



- **User Interface Layer**: UI and APIs for initiating evaluations and accessing reports.
- **Responsible AI Layer**: Bias, fairness, and safety evaluation services.
- **Orchestration Layer**: Evaluation Orchestrator Agent using A2A SDK.
- **Evaluator Agents**: Model, AI Agent, and Workflow Evaluation Agents.
- **MCP Servers**: Controlled connectors for resources, tools, and prompts.
- **Context Orchestrator**: Redis, Cosmos DB, AI Search, Graph RAG for grounding and memory.
- **Observability Stack**: Langfuse, Azure Monitor, Data Explorer.

- **Deployment Layer**: Azure Kubernetes Service (AKS).

5. Key Components

1. **Evaluation Orchestrator Agent** – Manages execution, error handling, and sequencing.
2. **Model Evaluation Agent** – Uses Ragas, LangGraph, and LLM-as-a-Judge.
3. **AI Agent Evaluation Agent** – Validates reasoning, tool invocation, and multi-step execution.
4. **Workflow Evaluation Agent** – Measures full AI pipeline accuracy, efficiency, and compliance.
5. **MCP Servers** – Sandbox connectors for GitHub, SQL, Salesforce, Slack, etc.
6. **Context Orchestrator** – Provides grounding and memory.
7. **Langfuse** – Observability layer for traces and reports.
8. **Responsible AI Services** – Bias, safety, and fairness checks.

6. Integration Points

- **Models**: Azure OpenAI, HuggingFace, or custom models.
- **Agents**: LangGraph, LangChain, or custom orchestrators.
- **E2E Workflows**: Retrieval-augmented generation (RAG), text-to-SQL, recommendation pipelines.
- **MCP Servers**: GitHub, SQL, Salesforce, Slack, Jira, etc.
- **Knowledge Bases**: Cosmos DB, Azure AI Search, Graph RAG.
- **Observability**: Langfuse, Azure Monitor, Data Explorer.
- **Authentication**: Azure AD + MCP Auth Server.

7. Use Cases

- Benchmark GPT-4 vs LLaMA 3 across summarization and Q&A.
- Evaluate a text-to-SQL pipeline from query → SQL generation → result retrieval.
- Test a RAG workflow (retrieval → grounding → answer quality) using Ragas.
- Validate multi-agent collaboration (planner → executor → verifier).
- Audit E2E enterprise AI solutions for bias, fairness, and compliance.
- Regression testing for new model or pipeline versions.

8. Customer Pain Points Addressed

- Inconsistent E2E evaluation across AI systems.
- Manual and siloed benchmarking approaches.
- High cost and complexity of regression testing.
- Risk of unsafe, biased, or non-compliant AI outputs.
- Lack of observability and auditability for enterprise AI pipelines.

9. Industry-Specific Applications

- **Banking/Finance**: Validate fraud detection pipelines, compliance workflows.
- **Healthcare**: Evaluate claims analysis, clinical AI assistants, and medical Q&A pipelines.
- **Retail**: Benchmark recommendation systems and customer support agents.
- **Telecom**: Test service desk optimization agents and chatbots.
- **Manufacturing**: Validate predictive maintenance and supply chain workflows.

10. Sample Customer Journey

1. User uploads a new LLM, agent configuration, or E2E solution.
2. Orchestrator assigns tasks to evaluation agents.
3. MCP servers supply mock prompts, tools, and datasets.
4. Evaluator Agents run tests at model, agent, and workflow levels.
5. Context Orchestrator ensures factual grounding.
6. Langfuse logs evaluation metrics and traces.
7. Dashboard shows:

"Solution A had 15% lower hallucination rate and 20% faster response time than Solution B across the full workflow."

11. Technical Requirements

- Azure subscription with OpenAI and Prompt Flow.
- AKS cluster for evaluation deployment.

- Redis for short-term memory.
- Cosmos DB + AI Search for grounding.
- MCP connectors for sandbox evaluation.
- Langfuse for observability.

12. Security Architecture

- **MCP Auth Server**: Role-based access control.
- **Sandbox Mode**: Ensures tool evaluation without affecting production.
- **Private Networking**: VNets + Private Link.
- **Data Governance**: Azure Purview for dataset lineage.
- **Audit Logs**: Traces via Langfuse, Azure Monitor, App Insights.

13. Performance Considerations

- Parallel evaluations across agents and workflows.
- Memory-first design reduces duplication.
- AKS scaling ensures load-aware orchestration.
- Caching accelerates repeated tasks.
- Real-time traces identify regressions quickly.

14. Tools and Azure Services Used

- Azure OpenAI
- LangGraph, Ragas, LLM-as-a-Judge
- Cosmos DB, Redis, AI Search, Graph RAG
- Langfuse, Azure Monitor, Data Explorer
- Azure Purview, Confidential Computing, Private Link
- AKS for deployment
- MCP connectors (GitHub, SQL, Salesforce, Slack, etc.)

15. Users of Agent Evaluation

- **ML Engineers & Researchers** – Benchmark models.

- **Enterprise Architects** – Validate agents and workflows.
- **Compliance Teams** – Audit for Responsible AI.
- **Product Managers** – Compare solution readiness.
- **MLOps/DevOps Teams** – Automate regression pipelines.

16. Dependencies

- MCP servers for sandbox testing.
- Evaluation libraries: Ragas, LangGraph, LLM-as-a-Judge.
- Langfuse observability stack.
- Azure services: OpenAI, AKS, Cosmos DB, Redis, AI Search.

17. Key Benefits and Differentiators

- **E2E Evaluation** – Models, agents, and full AI workflows.
- **Sandbox MCP** – Safe validation of tool/API usage.
- **Responsible AI** – Fairness, safety, bias checks included.
- **Traceability-first** – Langfuse observability.
- **Azure-native deployment** – Scalable, secure, enterprise-ready.

18. Value Proposition

Agent Evaluation enables enterprises to **trust entire AI solutions** by providing rigorous, explainable, and automated E2E evaluation.

It ensures that models, agents, and workflows deliver safe, accurate, and compliant outcomes, closing the gap between **experimental AI and production-ready AI systems**.

19. Conclusion

Agent Evaluation is a **transformative platform for evaluating end-to-end AI solutions**. By combining evaluator agents, sandbox MCPs, contextual grounding, and Azure-native

observability, it ensures organizations can deploy AI solutions with full confidence in their accuracy, fairness, and compliance.