Sudhir Kumar Kakumanu , 2026, 14:4
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Well-Architected Framework for Agentic AI

Sudhir Kumar Kakumanu, Mukesh Shukla, Srushti Anil Patil
LTIMindtree Blueverse, Applied Research and Innovation

Abstract- AI agents are changing how enterprises solve problems. While traditional AI models addressed specific tasks like classification or prediction, AI agents offer goal-driven, autonomous capabilities—sensing, reasoning, planning, and acting. These data-driven autonomous systems depend on high-quality, well-governed data to reason, ground decisions, and deliver reliable outcomes. However, deploying AI agents in complex enterprise environments is difficult. This paper presents the Well-Architected Framework (WAF) for AI Agents and an executable WAF Auditor Agent. The framework defines nine pillars—spanning data foundation, security, governance, and beyond—from Agent Complexity to Sustainability, while the Auditor Agent turns these guidelines into an automated assessment workflow. This allows organizations to evaluate architectural designs and receive recommendations before and during development, providing a clear path to production readiness.

Keywords: AI Agents, Large Language Models, Enterprise Architecture.

I. INTRODUCTION

Background and Motivation

The transition from static Large Language Models (LLMs) to autonomous AI agents represents a major change in enterprise computing. Unlike passive models that wait for prompts, agents are goal-driven systems capable of sensing their environment, reasoning through complex workflows, and executing actions to achieve business outcomes. This shift offers significant value by automating end-to-end processes rather than just generating text. AI agents use planning and reasoning capabilities from generative AI foundation models to perform tasks, while maintaining their own memory to further learn and improve. They can be fully autonomous or semi-automated, with a human in the loop. The core architecture enables the agent to listen to its environment, sense changes, reason about them, plan actions, and execute tasks accordingly.

As shown in Figure 1, the foundational features of AI agents are reasoning and acting, but multiple components must be integrated to create a modern, enterprise-ready agent. Every agent begins with a Trigger, such as an API call, a WhatsApp message, or an incoming email. Once invoked, the agent relies on Perception to understand the environmental context required for informed decision-making. This is supported by Context and Memory, which preserves insights from past interactions and retrieves relevant enterprise data. The core of the agent lies in Reasoning and Planning, where advanced generative AI models formulate strategic plans. These plans are realized through Action Execution, allowing the agent to interact with third-party applications. To ensure reliability, a Feedback and Evaluation layer continuously monitors performance, enabling the final component, Learning and Adaptation, which allows the agent to self-reflect and improve its accuracy over time.

The journey from traditional Machine Learning (ML) models to autonomous AI agents represents a significant paradigm shift, moving from systems that primarily predict or classify to entities that can understand, plan, act, and adapt within complex environments. Traditional ML models are typically Reactive and Task-Specific, designed to solve well-defined problems like classification or regression based on historical data. Their performance is heavily Data-Dependent, relying on the quality and quantity of training sets, with capabilities that remain fixed post-training. These models operate with Limited Autonomy, functioning within predefined constraints that require human intervention for new scenarios. Furthermore, they are Stateless, making predictions independently without retaining memory of past interactions.

Generative AI marks a shift towards Content Generation, moving beyond analysis to create novel text, images, and code. It demonstrates a remarkable capacity for Understanding and Nuance, grasping context, style, and subjective subtleties. By

leveraging Pre-trained Power from vast datasets, GenAI provides the "brain" for agents, enabling the reasoning and planning capabilities required for autonomy. Key evolutionary leaps include autonomy, multi-step reasoning and execution, tool usage, memory and statefulness, adaptive learning, and robust architectures.

Problem Statement

While the potential of AI agents is clear, the path to production is difficult. Recent studies of agents in production [16] highlight a gap between research capabilities and enterprise reality. Contrary to the vision of fully autonomous "set-and-forget" agents, successful real-world deployments today are defined by constraint and oversight. Production agents typically exhibit Bounded Complexity, with 68% executing fewer than 10 steps per task. There is also a significant Human Dependence, as 74% of deployments rely primarily on human evaluation to ensure correctness. Furthermore, organizations prioritize Simplicity over Training, with 70% choosing to prompt off-the-shelf models rather than investing in complex fine-tuning.

However, the rush to deploy agentic systems has outpaced the development of the engineering rigor required to sustain them. As organizations move from proof-of-concept to production, they are hitting reliability problems. Seminal work by Pan et al. (2025) [16] on "Measuring Agents in Production" has exposed critical gaps in how agents are built and evaluated, a finding reinforced by broader industry data: Gartner predicts that over 40% of agentic AI projects will be canceled by 2027 due to escalating costs and unclear value [8]; 61% of companies report significant accuracy issues with their AI tools [4]; deployed agents often exhibit non-trivial failure rates and inconsistency across tasks, as highlighted by Pan et al. [16]; and with 78% of firms overlooking AI compliance controls [17], the industry faces critical risks, illustrated by incidents like the "EchoLeak" vulnerability [20] and legal hallucinations [18]. These are not just early adoption issues but signs of a deeper problem: the lack of a standardized framework for building reliable agents.

Building production-ready agents is complex due to the probabilistic nature of foundation models. Key challenges include Complexity and Orchestration, which involves managing conflicts between multiple agents, and Control and Governance to ensure agents operate within defined boundaries. Explainability is critical for tracing decisions in regulated sectors, while Security focuses on mitigating data leaks and system disruptions. Reliability is essential to prevent failures in mission-critical applications, and Bias and Ethics must be managed to avoid legal complications. Organizations must also address Resource Management to balance costs, redesign Agent Operations for probabilistic systems, navigate complex Integration landscapes, and establish new Observability capabilities beyond traditional monitoring tools.

Contributions

To address this, this paper introduces the Well-Architected Framework (WAF) for AI Agents. Directly responding to the reliability challenges identified by Pan et al. (2025) [16] and taking cues from established cloud architecture frameworks, the WAF provides a set of design principles and pillars to guide the lifecycle of agentic systems. It offers concrete, testable checkpoints for Security, Reliability, Operational Excellence, and more. Critically, the framework places data foundation and data governance at the core of agent architecture, recognizing that the quality, integrity, and security of enterprise data directly determine the reliability and trustworthiness of autonomous agent behavior.

The framework addresses these complexities by providing Prescriptive Guidance through actionable best practices across the agent lifecycle. It focuses on Risk Mitigation by proactively identifying security and ethical risks, while ensuring Scalability and Resilience for reliable operations. The framework also emphasizes Cost Optimization strategies and Accelerated Time-to-Value to help organizations realize benefits faster. By establishing Consistency for maintainability and enabling Responsible Innovation, the framework serves as a blueprint for confident enterprise adoption.

This framework serves as a guide for stakeholders involved in the lifecycle of AI agents, to support successful and responsible adoption.

- **Strategic Leadership (CTOs, CIOs, Business Leaders):** For aligning AI initiatives with business strategy, governance, and IT compliance.
- **Technical Architects & Engineers:** For best practices in designing, developing, and deploying robust, scalable, and secure agent solutions.
- **Governance & Operations (Security, Compliance, Ops):** For insights into security, data privacy, infrastructure management, and operational monitoring.

II. LITERATURE REVIEW

Agentic AI Foundations and System Components

The literature describes AI agents as systems that use planning and reasoning capabilities from generative AI foundation models to perform tasks while maintaining their own memory to further learn and improve [1]. Across these descriptions, agents differ from earlier AI systems because they can be fully autonomous or semi-automated, with a human in the loop [15]. A common theme is that agentic systems are not defined only by a model, but by the interaction of sensing, reasoning, planning, memory, action, and feedback capabilities [12]. The core architecture therefore enables the agent to listen to its environment, sense changes, reason about them, plan actions, and execute tasks accordingly.

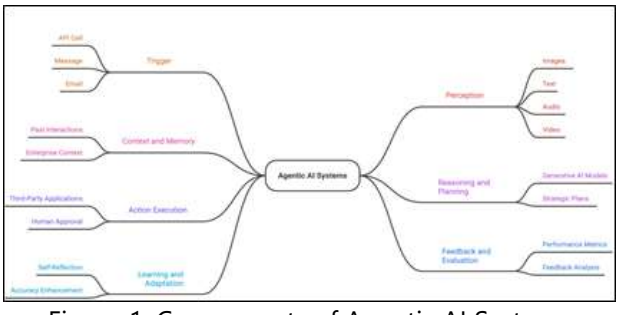

Figure 1. Components of Agentic AI Systems

As shown in Figure 1, prior descriptions converge on reasoning and acting as foundational features of AI agents, while also showing that multiple components must be integrated to create a modern, enterprise-ready agent [15]. The component view indicates that every agent begins with a Trigger, such as an API call, a WhatsApp message, or an incoming email. Once invoked, the agent relies on Perception to understand the environmental context required for informed decision-making. This is supported by Context and Memory, which preserves insights from past interactions and retrieves relevant enterprise data. The core of the agent lies in Reasoning and Planning, where advanced generative AI models formulate strategic plans. These plans are realized through Action Execution, allowing the agent to interact with third-party applications. To ensure reliability, a Feedback and Evaluation layer continuously monitors performance, enabling the final component, Learning and Adaptation, which allows the agent to self-reflect and improve its accuracy over time [6]. Taken together, this shows a trend toward composite agent architectures in which reliability depends on coordinated components rather than isolated model capability.

Evolution from Traditional ML to Autonomous Agents

Past research and system evolution indicate a shift from traditional Machine Learning (ML) models to autonomous AI agents, moving from systems that primarily predict or classify to entities that can understand, plan, act, and adapt within complex environments [15]. Figure 2 summarizes this evolution and provides the basis for comparing earlier task-specific models with newer agentic systems.

Characteristic	Traditional ML	Generative AI	Autonomous AI Agents
Focus	Pattern Recognition & Prediction	Content Generation	Goal-Oriented Autonomy
Autonomy	Limited	Bridging to Agents	High
Reasoning	Limited	Understanding & Nuance	Multi-Step
Tool Usage	No	Yes	Yes
Learning	Data-Dependent	Pre-trained	Adaptive
Architecture	Simple Models	Foundation Models	Robust Architectures

Figure 2. Evolution of AI Systems

Traditional ML models are described as Reactive and Task-Specific, designed to solve well-defined problems like classification or regression based on

historical data. Their performance is heavily Data-Dependent, relying on the quality and quantity of training sets, with capabilities that remain fixed post-training. These models operate with Limited Autonomy, functioning within predefined constraints that require human intervention for new scenarios. They are also Stateless, making predictions independently without retaining memory of past interactions. In contrast, Generative AI represents a shift toward Content Generation, moving beyond analysis to create novel text, images, and code. It demonstrates capacity for Understanding and Nuance, grasping context, style, and subjective subtleties. By leveraging Pre-trained Power from vast datasets, GenAI provides the "brain" for agents, enabling the reasoning and planning capabilities required for autonomy [15]. The synthesis of these developments shows that earlier approaches solve bounded prediction tasks, while agentic systems require autonomy, memory, adaptive behavior, and coordinated execution.

The emergence of autonomous AI agents is characterized by key evolutionary leaps: autonomy, multi-step reasoning and execution, tool usage, memory and statefulness, adaptive learning, and robust architectures [15]. Together, these trends show that the field has moved from task-specific models toward systems that can coordinate actions across workflows. At the same time, these capabilities increase the need for structured architectural guidance because agent behavior depends on multiple interacting layers rather than a single model output [1].

Production Reality and Enterprise Deployment Gaps

Although the potential of AI agents is clear, studies of agents in production [16] highlight a gap between research capabilities and enterprise reality. This gap is central to the literature because it shows that agentic AI is not only a modeling problem but also an architectural, operational, and governance problem. The literature on production agents challenges the vision of fully autonomous "set-and-forget" systems. Successful real-world deployments are instead defined by constraint and oversight. Production agents typically exhibit Bounded Complexity, with 68% executing fewer than 10 steps per task. There is also significant Human Dependence, as 74% of deployments rely primarily on human evaluation to ensure correctness. Organizations also prioritize Simplicity over Training, with 70% choosing to prompt off-the-shelf models rather than investing in complex fine-tuning. These findings reveal a trend toward controlled, bounded, and human-evaluated deployments rather than unrestricted autonomy.

Challenges in Adopting AI Agents

Prior work and current deployment experience show that building production-ready agents is complex due to the probabilistic nature of foundation models [16]. This topic links the evolution of agent capabilities to the practical barriers that appear when those capabilities are deployed in enterprise settings.

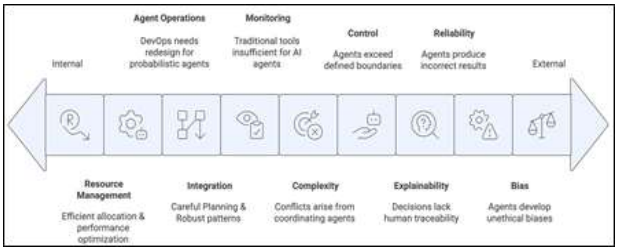

Figure 3. Challenges of Adopting AI Agents

As summarized in Figure 3, key challenges include Complexity and Orchestration, which involves managing conflicts between multiple agents, and Control and Governance to ensure agents operate within defined boundaries [13]. Explainability is critical for tracing decisions in regulated sectors, while Security focuses on mitigating data leaks and system disruptions [7]. Reliability is essential to prevent failures in mission-critical applications [5], and Bias and Ethics must be managed to avoid legal complications [13]. Organizations must also address Resource Management to balance costs, redesign Agent Operations for probabilistic systems, navigate complex Integration landscapes, and establish new Observability capabilities beyond traditional monitoring tools. Synthesizing these challenges shows that the main gap is not the absence of agent capabilities, but the lack of a structured way to

evaluate and govern those capabilities across the full lifecycle.

Positioning of the Current Study

The reviewed literature and reported deployment patterns indicate that AI agents require more than model capability: they require prescriptive architectural guidance across the agent lifecycle [1]. The Well-Architected Framework addresses these complexities by providing Prescriptive Guidance through actionable best practices across the agent lifecycle. It focuses on Risk Mitigation by proactively identifying security and ethical risks, while ensuring Scalability and Resilience for reliable operations. The framework also emphasizes Cost Optimization strategies and Accelerated Time-to-Value to help organizations realize benefits faster. By establishing Consistency for maintainability and enabling Responsible Innovation, the framework serves as a blueprint for confident enterprise adoption. In this broader field, the current study fits by translating the identified gaps—bounded complexity, human evaluation, reliability, security, governance, integration, observability, and lifecycle control—into a structured framework and an executable Auditor Agent for evaluating agentic AI systems [16].

III. THE NINE PILLARS OF A WELL-ARCHITECTED AI AGENT SYSTEM

Figure 4. Nine Pillars of Agentic AI Systems

The framework defines 157 checkpoints distributed across these nine pillars, with particular focus on Sustainability, Adaptability, and Responsibility. These checkpoints serve as the evaluation criteria for the WAF Auditor Agent, which automates the assessment of agent designs against the framework's standards.

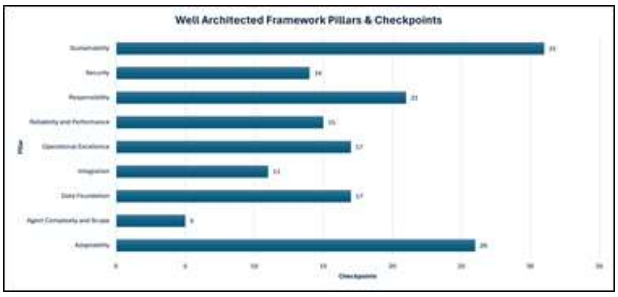

Figure 5. Well Architected Framework Pillars & Checkpoints

Pillar 1: Agent Complexity and Scope

The Agent Complexity and Scope pillar focuses on defining the architectural boundaries, reasoning limits, and evaluation strategies required to ensure an agent is viable, manageable, and safe for production. Unlike traditional software, AI agents can exhibit non-deterministic behavior; this pillar ensures that complexity is introduced intentionally and is always bounded by strict controls. Unbounded agents are expensive and unreliable, as they can enter "infinite loops" of reasoning without strict step limits. Furthermore, complexity kills viability; agents that attempt to "do everything" often fail to perform any task well. Simplicity reduces error propagation, whereas complex chains of 20+ steps have a high probability of failure. Consequently, human validation is essential, as automated benchmarks alone are insufficient for initial agent deployments. Recent research [16] highlights the critical challenges in deploying agents.

The evolution of agent architectures [15] and the need for meaningful agents [12] underscore the importance of scoping. Comprehensive frameworks are required to create and manage these agents effectively [1].

Table 1 outlines the major categories of challenges reported by participants in the study.

Table 1. Major categories of challenges reported by participants

Challenge Category	Representative Issues And Focus Areas
Core Technical Performance	Robustness And Reliability—Ensuring Consistent, Correct Behavior In Diverse And Unpredictable Environments; Scalability; Real-Time

	Responsiveness; Resource Constraints.
Data And Model Integrity	Data Quality And Availability; Model And Concept Drift; Versioning And Reproducibility.
System Integration And Validation	Integration With Legacy Systems; Testing And Validation; Security And Adversarial Robustness.
Transparency And Governance	Explainability And Interpretability; Bias And Fairness; Accountability And Responsibility.
Compliance And User Trust	Privacy And Data Protection; User Trust And Adoption; Regulatory Compliance.

Core Technical Performance issues were the most common, indicating that teams spend significant effort ensuring systems perform consistently under real-world conditions. Data and Model Integrity and System Integration and Validation followed closely. To manage complexity, organizations should Define Narrow Boundaries, clearly specifying what the agent cannot do and restricting it to a specialized domain. It is critical to Enforce Execution Limits by implementing strict "time-to-live" or "max-step" counters. Architects should Prefer Atomic Tools that perform single, deterministic actions rather than complex, multi-step functions. Initial deployments must mandate Human-First Evaluation of agent traces, and the system must be designed to Fail Gracefully with a defined fallback path if complexity limits are reached.

Pillar 2: Data Foundation

The Data Foundation layer emphasizes the importance of AI agents having access to high-quality data for better decision-making. High-quality data is regularly validated, kept up to date, clearly defined with metadata, and governed by appropriate data management rules. More than 52% of Chief Data Officers believe their data foundation layer is not ready for Generative AI implementation [2]. Furthermore, agents will hallucinate if not grounded with proper enterprise data. Finally, personalization and adaptability require a trusted data source with rich user profiles built in. Organizations must establish a single source of truth for AI agents and ensure data is normalized into a consistent format. It is essential to provide semantic data by tagging raw information with business context and to apply quality gates that ensure accuracy and completeness. A centralized repository of metadata is required to support discovery. Security should be enforced by implementing RBAC and ensuring agents always access up-to-date data. Additionally, teams should organize and tag unstructured data, implement contextual grounding with user history, and regularly audit data for bias to ensure fairness.

Pillar 3: Integration

Integration refers to the seamless connectivity of AI agents with the existing enterprise ecosystem, including applications, workflows, and human stakeholders. Integration allows AI to move from simple smart assistants to real contributors by embedding them directly within business workflows. This connectivity enables agents to adapt in real time, reading changed data and writing actions back to live environments, thereby driving tangible business value. To achieve seamless integration, organizations should Treat Agents as First-Class Services, exposing them via an Internal API Gateway. Agents must be equipped with a broad set of tools to access core systems, utilizing shared frameworks like the Model Context Protocol (MCP) [14]. It is vital to Standardize A2A Protocols for agent interaction and Integrate Agents Natively into business workflows. A Shared Enterprise Knowledge Graph should be leveraged to provide context. Communication should rely on Message Schemas with explicit intent and status fields. Finally, workflows must Enable HITL and HOTL Mechanisms for oversight and Implement Unified Observability to monitor end-to-end performance.

Pillar 4: Operational Excellence

This pillar focuses on the deployment, monitoring, and continuous improvement of AI agents that can scale in a cost-effective manner. Scaling agents to serve thousands of concurrent users requires careful architectural considerations. Agents should automatically scale up during high-load scenarios and scale down to save costs. Furthermore, in case of failures, complete traceability on LLM calls and

tool calls is needed. Operational excellence begins with the decision to Containerize & Orchestrate agents using tools like Kubernetes. Security is paramount, so teams must always have agents behind an authentication layer. To maintain visibility, Unified Observability with standardized logging is required. Efficiency is driven by the goal to Automate everything feasible via CI/CD and Infrastructure as Code. Architectural principles drive this transformation [3]. Architectures should be Designed for distributed deployment to handle varying loads, while FinOps practices help to Continuously monitor and optimize costs. Finally, systems should be Built for failure with early failure loops and governed by Clear Service Level Objectives (SLOs).

Pillar 5: Reliability and Performance

Reliability refers to the ability of AI systems to consistently deliver correct and predictable output under varying conditions. Reliability is a primary concern, as hallucination remains a core problem where agents can fabricate data. Over 61% of leaders cite accuracy as a major barrier to production [4] and legal risks from hallucinations are significant [18]. Reliability strategies are further detailed in recent guides ([5], [11]). Most real-world failures occur at the edge cases, making robust testing essential. To ensure reliability, teams must Build an evaluation suite specific to their agents, ensuring they Include all edge cases and test cases for tool usage. Continuous Evaluation should be standard practice before and during deployment. Feedback Loops allow for improvement based on user input. Trust is built by Adding citations and explanations to agent answers. Architectures should include self-critic loops and Log comprehensive agent traces. Finally, systems must Implement error management & graceful degradation to handle unexpected conditions safely.

Pillar 6: Adaptability

Adaptability refers to the ability of the agent to learn from interactions and improve over time. Static agents quickly become obsolete. In contrast, adaptive AI systems can outperform traditional models by adjusting to real-time data [10]. Designing agents with cognitive frameworks allows for better behavioral adaptation [6]. Personalization is a key driver of adoption, and continuous feedback loops are critical for long-term improvement. Adaptability requires systems to Collect continuous feedback from end users and Learn user preferences with every interaction. Reciprocal learning enables both agents and humans to learn from each other. Agents should Build episodic memory to retain user context and Optimize prompts based on historical feedback. Finally, it is essential to Detect drift by monitoring inputs and outputs for shifts in user behavior.

Pillar 7: Security

Security refers to the protection of the AI system, its data, and its operations from unauthorized access, modification, or destruction against cyber threats. Security is a critical concern, with Gartner predicting that 40% of data breaches will arise from AI agent misuse [7]. Data leakage is a growing risk [20], especially as agents gain access to confidential data and modification tools. Furthermore, new attack vectors such as prompt injection and memory poisoning are emerging. To secure agents, organizations must enforce Identity & Access Management with granular RBAC and Preserve user context throughout the execution cycle. Secure Tool Execution ensures tools are invoked on behalf of the user. Data security requires row level security, Data Protection via encryption, and PII masking. Input/Output Guardrails should validate against attacks. Network security must whitelist only relevant ports. Critical actions require an approval gate. Finally, teams must Sanitize Training Data, Monitor 3rd party plugins, employ proactive threat detection, and conduct Red Teaming exercises.

Pillar 8: Responsibility

Responsibility refers to the ethical and legal aspects of agent behaviour. It requires agents to act within a constrained environment, conduct themselves fairly, and operate under appropriate human oversight. Trust requires oversight. Accountability is mandatory, yet 78% of firms overlook AI compliance despite rising adoption [17]. Governance frameworks are essential for managing these risks [13]. Without responsible design, bias can be amplified, leading to ethical and legal risks. Responsible AI requires keeping a Human-in-the-Loop (HITL) for high-stakes

decisions and ensuring agents Explain Agent Decisions clearly. Organizations must Implement bias detection to monitor fairness and define a Clear Scope and Role for every agent. Borderline cases should be routed to humans, and every agent must have a defined Agent Owner to ensure accountability.

Pillar 9: Sustainability

Sustainability ensures that Agentic AI systems are designed and operated with a long-term perspective, combining Environmental Responsibility with Financial Sustainability. Cost is a viability killer for AI projects. The energy impact is also real, as large language models consume significant water and energy resources [19]. Ultimately, demonstrating ROI is essential to prove the worth of agentic systems. To ensure sustainability, organizations should Measure Cost Per Decision and Right-Size the Model to the task. It is important to Monitor ROI by comparing expenditures against value [9]. Efficiency can be improved by Batching and Caching queries. Finally, teams should create a Sustainability Impact Assessment and Enforce Budgets to manage financial and environmental costs.

IV. THE WAF AUDITOR AGENT

To put the Well-Architected Framework into practice, we developed the WAF Auditor Agent—an executable evaluation system that automates the assessment of AI agents against the 157 checkpoints defined in the framework. This section details the agent's architecture, workflow, and evaluation performance.

WAF Evaluation Architecture & Workflow

The WAF Auditor Agent is designed to integrate into the development lifecycle, allowing architects to trigger evaluations on-demand. The workflow begins when the user opens the WAF UI and selects the agent to be evaluated. The user then completes the survey across the WAF pillars by providing responses, notes, and descriptive context. After clicking Generate WAF Report, the frontend builds the evaluation request by mapping the survey answers and attaching the relevant agent profile. This request is sent to the AI Platform endpoint using the configured WAF agent space name. During execution, the WAF Auditor Agent applies the system prompt, retrieves relevant knowledge through RAG, scores the agent's maturity, and generates both the narrative assessment and structured JSON output. The API response returns these narrative and JSON markers to the frontend, where the parser extracts the report text and interprets the structured data. Finally, the dashboard renders the overall maturity score, pillar-wise breakdown, identified gaps, and remediation roadmap for review.

RAG Corpus Construction

The WAF Auditor Agent leverages a Retrieval-Augmented Generation (RAG) architecture to ground its evaluations in the framework's 157 checkpoints.

Corpus Construction: The knowledge base consists of structured representations of all WAF checkpoints, enriched with pillar-wise categorization and best-practice descriptions and few references related to agentic evaluation.

Retrieval and Grounding Mechanism

During evaluation, user inputs—including survey responses, pillar-specific notes, and agent descriptions are transformed into a structured request. The WAF Auditor Agent then performs retrieval over the WAF knowledge base to identify relevant checkpoints.

As implemented in the system workflow, the agent applies a predefined system prompt, uses RAG to retrieve relevant WAF knowledge, aligns retrieved checkpoints with the input context, and generates recommendations and maturity assessments grounded in these checkpoints.

This retrieval step ensures that outputs are not purely generative but anchored to predefined architectural standards.

Model and Execution Configuration

The Auditor Agent is deployed as a specialized agent within an AI platform and is invoked via an API

endpoint (chatservice/chat) with a dedicated agent space configuration .

The execution flow includes: System prompt enforcement to ensure structured evaluation Context injection from retrieved WAF checkpoints Generation of both narrative explanations and structured JSON outputs

The structured output includes pillar-wise maturity scores, identified gaps, and recommended actions, enabling downstream processing and visualization.

Integration into Development Workflows

The WAF Auditor Agent is tightly integrated into the development lifecycle through a UI-driven and API-enabled workflow:

Design-Time Evaluation Architects define agent characteristics and complete a WAF survey via the UI. Structured Request Generation The frontend constructs a 'waf-evaluation-request' by mapping responses to pillars, flattening notes, and attaching agent metadata. Automated Evaluation via API The request is sent to the WAF agent through the platform API, triggering RAG-based evaluation. Grounded Assessment and Scoring The agent retrieves relevant checkpoints, evaluates the design, and produces both narrative insights and structured outputs. Visualization and Iteration The frontend parses the response and renders a dashboard showing maturity scores, gap analysis, and remediation roadmap. Users can iteratively refine inputs and regenerate evaluations.

V. EVALUATION SCENARIOS AND RESULTS

To validate the effectiveness of the Well-Architected Framework Auditor Agent, we conducted an evaluation using seven diverse industry scenarios. For each scenario, we compared the Auditor Agent's generated report against a ground-truth expert assessment. The evaluation metrics focus on Semantic Precision, Semantic Recall, and Semantic F1 Score to measure how accurately the agent identifies gaps and provides recommendations aligned with the WAF standards.

Performance Metrics

To evaluate the alignment between the Auditor Agent outputs and ground-truth references, we adopt a semantic similarity–based evaluation approach rather than exact string matching.

Ground-truth recommendations are derived from the structured WAF checkpoint repository and grouped by pillar, while agent-generated recommendations are extracted from evaluation outputs. For each pillar, semantic similarity is computed between the set of agent-generated recommendations and the corresponding ground-truth recommendations.

We use a transformer-based sentence embedding model (SentenceTransformer: all-MiniLM-L6-v2) to encode both agent and ground-truth recommendations into a shared vector space. A pairwise cosine similarity matrix is then computed between all generated and ground-truth recommendations.

- Semantic Precision is computed as the average of the maximum similarity scores for each agent-generated recommendation. This measures how closely each generated recommendation aligns with the most relevant ground-truth item.
- Semantic Recall is computed as the average of the maximum similarity scores for each ground-truth recommendation. This captures how well the agent covers the full set of expected recommendations.
- Semantic F1 Score is computed as the harmonic mean of semantic precision and recall.

In addition to continuous semantic scores, we compute threshold-based binary metrics. A recommendation is considered a match if its maximum similarity exceeds a predefined threshold ($\tau = 0.7$). Binary precision and recall are then calculated based on matched pairs. This dual evaluation approach allows us to capture both: Quality of recommendations (via continuous semantic similarity), and Coverage of expected gaps (via threshold-based matching)

The methodology is particularly suited for evaluating natural language outputs where multiple valid phrasings may exist for the same architectural

recommendation. Figure 6 summarizes the performance of the Auditor Agent across the seven scenarios.

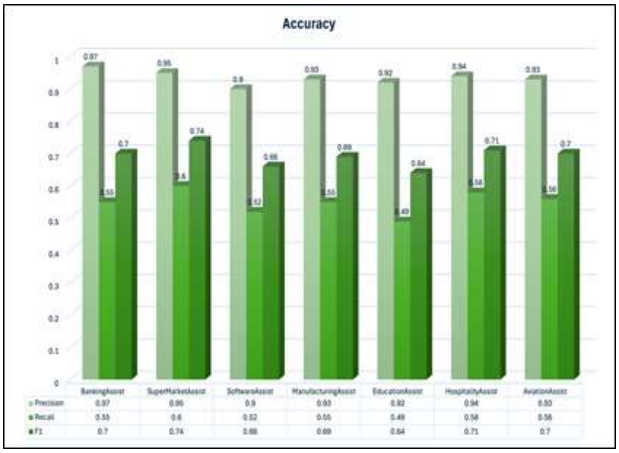

Figure 6. Auditor Agent Performance Metrics

Analysis of Results:

- **High Precision (Avg 0.93):** The agent is accurate in identifying correct recommendations, ensuring that reported gaps are valid and trustworthy.
- **Moderate Recall (Avg 0.55):** The agent tends to miss several relevant items, suggesting a need to improve coverage and retrieval sensitivity. While this reflects a trade-off for prioritization, it highlights an area for future improvement to ensure comprehensive assessment.

State of Agent Maturity

Beyond validating the Auditor Agent itself, the evaluation data provides a snapshot of the current state of enterprise agent maturity. By aggregating the scores assigned by the Auditor Agent across all scenarios, we can identify common strengths and weaknesses in current agent architectures. Figure 7 presents the average maturity scores across the nine pillars.

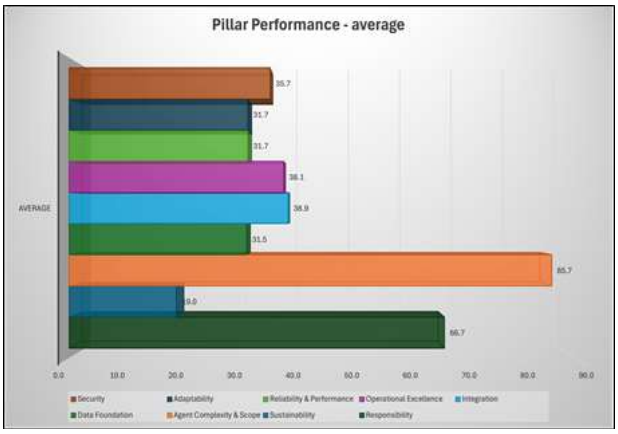

Figure 7. Average Maturity Scores by Pillar (Scale 0-100)

The data reveals strong performance in Agent Complexity & Scope and Responsibility, indicating that organizations are successfully defining agent boundaries and ethical guidelines. However, significant weaknesses persist across Data Foundation, Reliability, Adaptability, Security, and especially Sustainability. This suggests that while the initial agent design is often well-defined, there are gaps in reliability and long-term efficiency.

Gap Analysis

The Auditor Agent identified a total of 88 critical gaps across the 7 scenarios. While the Auditor Agent demonstrates high precision, recall remains moderate (Avg 0.55), indicating that some relevant gaps are not identified. A detailed analysis reveals several contributing factors:

1. **Implicit or Under-Specified Gaps:** Certain architectural issues are not explicitly described in the input but are implied (e.g., missing observability or incomplete fallback strategies). The agent tends to prioritize explicitly stated signals, leading to missed detections.
2. **Cross-Pillar Dependencies:** Some gaps span multiple pillars, such as data governance issues affecting reliability or security. Since evaluation is performed with pillar-level structuring, these dependencies may not always be fully captured.
3. **Retrieval Coverage Limitations:** The RAG-based retrieval step may not always surface all relevant checkpoints, particularly when the input description is sparse or ambiguous, leading to incomplete grounding.

4. **Granularity Mismatch:** In some cases, the Auditor Agent produces higher-level recommendations, whereas ground-truth annotations include fine-grained implementation gaps. This results in semantically correct but unmatched outputs under strict evaluation.

These observations suggest that improving retrieval coverage and enabling cross-pillar reasoning are key areas for future enhancement. Figure 8 breaks down the average number of critical gaps found per pillar.

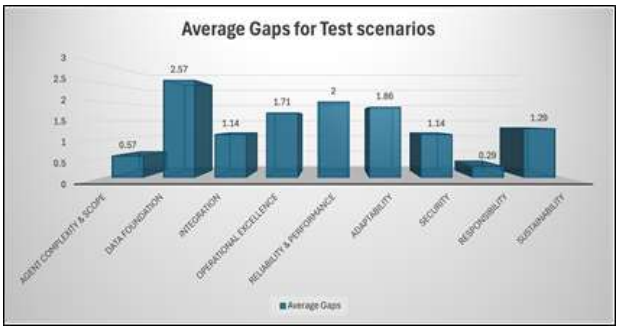

Figure 8. Average Critical Gaps Identified per Pillar

The pattern suggests that data quality, reliability, and adaptability remain the most critical improvement priorities for current agent deployments. Additionally, Figure 9 illustrates a strong inverse relationship between pillar scores and the number of identified gaps. Lower pillar scores consistently align with higher gap counts, indicating that reduced maturity directly drives structural weaknesses.

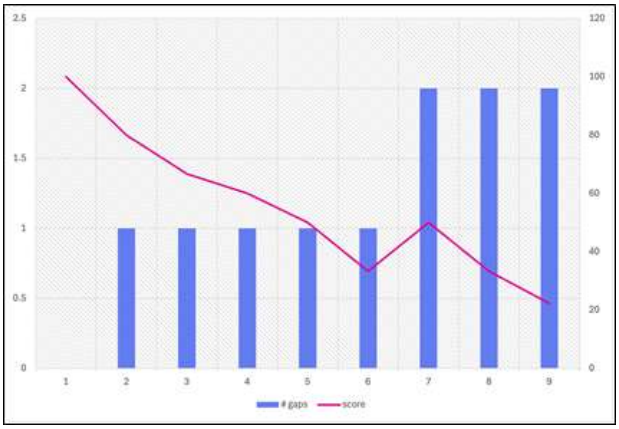

Figure 9. Relationship between Pillar Scores and Identified Gaps

VI. CONCLUSION

The move from simple chatbots to autonomous AI agents is changing how enterprises operate. While the potential for efficiency and innovation is significant, so are the risks associated with complexity, reliability, and security. The "Well-Architected Framework for Agentic AI Systems" provides a structured approach to managing these risks. By following the nine pillars—Agent Complexity and Scope, Data Foundation, Integration, Operational Excellence, Reliability and Performance, Adaptability, Security, Responsibility, and Sustainability—organizations can build agentic systems that are trustworthy, scalable, and aligned with business goals.

Central to this is the recognition that data is the lifeblood of agentic AI: without a strong data foundation and rigorous data governance, even the most sophisticated agents will fail to deliver reliable, trustworthy outcomes. This framework is not a one-time checklist but a guide for the full lifecycle of AI agents. The WAF Auditor Agent supports this by enabling teams to evaluate their agents against these standards at every stage. As the technology evolves, so too must the architectural practices that support it. We encourage architects, developers, and business leaders to use both this framework and the Auditor Agent when building autonomous enterprise systems.

REFERENCES

1. Aitelhaj, R., & Tay, R. (2025). Framework to create and manage AI agents : Building a comprehensive framework for creating and managing AI agents. Theseus. https://www.theseus.fi/bitstream/handle/10024/889695/Aitelhaj_Tay.pdf?sequence=2
2. AWS. (2025). CDO agenda 2025: Scaling generative AI for value. https://d1.awsstatic.com/psc-digital/2024/gc-600/cdo-biz-value/CDO-Agenda-2025-ScalingGenerativeAIforValue.pdf
3. Cornforth, M. (2025). Driving agentic transformation with architecture principles. MuleSoft Blog.

https://blogs.mulesoft.com/automation/driving-agentic-transformation-with-architecture-principles/

4. Edstellar. (2025). AI agent reliability challenges. https://www.edstellar.com/blog/ai-agent-reliability-challenges
5. Galileo. (2025). A guide to AI agent reliability for mission critical systems. https://galileo.ai/blog/ai-agent-reliability-strategies
6. Garg, V. (2025). Designing the mind: How agentic frameworks are shaping the future of AI behavior. Journal of Computer Science and Technology Studies, 7(5), 182–193. https://al-kindipublishers.org/index.php/jcsts/article/view/9781/8450
7. Gartner. (2025a). Gartner predicts forty percent of AI data breaches will arise from cross-border GenAI misuse by 2027. https://www.gartner.com/en/newsroom/press-releases/2025-02-17-gartner-predicts-forty-percent-of-ai-data-breaches-will-arise-from-cross-border-genai-misuse-by-2027
8. Gartner. (2025b). Gartner predicts over 40 percent of agentic AI projects will be canceled by end of 2027. https://www.gartner.com/en/newsroom/press-releases/2025-06-25-gartner-predicts-over-40-percent-of-agentic-ai-projects-will-be-canceled-by-end-of-2027
9. Gartner. (2025c). Take this view to assess ROI for generative AI. https://www.gartner.com/en/articles/take-this-view-to-assess-roi-for-generative-ai
10. Gartner. (2025d). Why adaptive AI should matter to your business. https://www.gartner.com/en/articles/why-adaptive-ai-should-matter-to-your-business
11. Kaminski, A. (2025). The secret to reliable AI agents isn't activity. https://www.linkedin.com/posts/avinoamkaminski_the-secret-to-reliable-ai-agents-isnt-activity-7386664301631090689-7UKp
12. Kapoor, S., Stroebl, B., Siegel, Z. S., Nadgir, N., & Narayanan, A. (2024). AI agents that matter. arXiv Preprint arXiv:2407.01502. https://arxiv.org/pdf/2407.01502
13. Kraprayoon, J., Williams, Z., & Fayyaz, R. (2025). AI agent governance: A field guide. arXiv Preprint arXiv:2505.21808. https://arxiv.org/pdf/2505.21808
14. Krishnan, N. (2025a). Advancing multi-agent systems through model context protocol: Architecture, implementation, and applications. arXiv Preprint arXiv:2504.21030. https://arxiv.org/pdf/2504.21030
15. Krishnan, N. (2025b). AI agents: Evolution, architecture, and real-world applications. arXiv Preprint arXiv:2503.12687. https://arxiv.org/pdf/2503.12687
16. Pan, M. Z., Arabzadeh, N., et al. (2025). Measuring agents in production. arXiv Preprint arXiv:2512.04123. https://arxiv.org/pdf/2512.04123v1
17. Security Brief. (2025). 78% of UK firms overlook AI compliance despite rising adoption. https://securitybrief.co.uk/story/78-of-uk-firms-overlook-ai-compliance-despite-rising-adoption
18. The New York Times. (2023). Avianca airline lawsuit ChatGPT. https://www.nytimes.com/2023/05/27/nyregion/avianca-airline-lawsuit-chatgpt.html
19. The Times. (2025). Thirsty ChatGPT uses four times more water than previously thought. https://www.thetimes.com/uk/technology-uk/article/thirsty-chatgpt-uses-four-times-more-water-than-previously-thought-bc0pqswdr
20. Varonis. (2025). EchoLeak. https://www.varonis.com/blog/echoleak