BITSUMMIT

EXPERTISE THAT DELIVERS

WHITE PAPER · MICROSOFT AI GOVERNANCE ARCHITECTURE · V2.0

Claude Code Security Governance with Microsoft Agent 365

An evangelism-grade architecture blueprint for governing Claude Code at enterprise scale on the Microsoft 365 platform. Authored by BITSUMMIT, a Microsoft Solutions Partner for Data and AI plus Digital and App Innovation, with deeper technical implementation detail, expanded Agent 365 capability coverage, and the supplementary controls required to bridge the Claude Code coverage gap between today and Microsoft's roadmap.

3
Governance stages from baseline policy to full sandbox isolation

12
Distinct capability gaps identified, each with prescriptive remediation

2
Microsoft Solutions Partner designations behind this architecture

90d
Recommended pilot-to-production timeline from Stage 1 kickoff

DOCUMENT ID
BS-RPT-2026-014

VERSION / DATE
v2.0 · 02 May 2026

CLASSIFICATION
Confidential

EXECUTIVE SUMMARY

The Microsoft Path to Enterprise Claude Code Governance

Microsoft Agent 365 went generally available on 01 May 2026. It is the strategic control plane for AI agents inside the Microsoft 365 estate, and it is on a published roadmap to cover Claude Code natively. This paper describes how to pair Agent 365 with the supplementary controls required today, in a Microsoft-native architecture, to govern Claude Code at enterprise scale right now while the platform matures.

Situation

Microsoft's strategic position on enterprise AI agent governance is now codified in a single product. Agent 365 brings together a centralized agent registry in Microsoft Entra, Tools Management for tenant-wide MCP allowlisting, Policy Templates that bundle Entra, Purview, Defender, and SharePoint controls, Entra ID Governance for agents with sponsor lifecycle workflows, and integration with Defender for AI agents detection. The same platform investment that produced Microsoft 365 Copilot governance now extends to third-party agents, including Anthropic's Claude Code.

Anthropic's Claude Code, in parallel, is the agentic coding tool of choice for an increasing number of enterprise development organizations. It ships with managed-settings, a rich hooks model, native sandboxing on Linux and macOS, and an emerging Compliance API. The two product lines, Microsoft Agent 365 and Anthropic Claude Code, are converging.

Complication

The convergence is not yet complete. Five gaps prevent Microsoft Agent 365 alone from delivering full enterprise-grade governance for Claude Code as of May 2026:

- **Coverage timing.** Agent 365's Shadow AI page detects OpenClaw at GA. Claude Code, GitHub Copilot CLI, and other local agents are on the Microsoft roadmap, with public preview through June 2026 and no committed Claude Code GA date.
- **Endpoint enforcement gaps.** Claude Code's `--dangerously-skip-permissions` flag silently bypasses every PreToolUse and PostToolUse hook. A confirmed Anthropic bug (issue #26923) also lets the Task subagent tool execute despite a PreToolUse exit-2 block.
- **Audit completeness.** Anthropic's Compliance API records control-plane events (admin actions, file activity) but excludes prompt and completion content. Cowork activity is excluded from the Compliance API, audit logs, and data exports entirely.
- **Sandboxing reach.** Native Claude Code sandboxing requires bubblewrap (Linux) or Seatbelt (macOS). Native Windows has no sandbox. Microsoft Intune does not orchestrate either bubblewrap or Seatbelt directly.

- **MCP server enforcement.** Agent 365's Tools Management is tenant-wide but does not yet auto-push `allowManagedMcpServersOnly: true` to Claude Code endpoints. Endpoint-side reinforcement is required to make the tenant policy binding.

Resolution

This paper provides a 90-day, three-stage Microsoft-native implementation plan. Every stage is built on Microsoft platform components — Entra, Intune, Defender XDR, Sentinel, Purview, and (where available) Agent 365 — with three open-source supplements (cco wrapper for native sandboxing, Docker Sandboxes for cross-platform microVM isolation, and an OpenTelemetry collector for tool-level audit) deployed inside the same Microsoft estate. Stage 1 (weeks 1 to 4) establishes a global managed-settings baseline delivered through Intune, paired with six Defender XDR custom detection rules that close the Shadow AI coverage gap until Microsoft ships native Claude Code support. Stage 2 (weeks 5 to 9) introduces role-based profiles via Entra group-driven Intune assignments, plus a centralized HTTP hooks server and an MCP gateway that survive the dangerously-skip-permissions bypass. Stage 3 (weeks 10 to 13) layers sandbox isolation: Windows 365 for Agents (Microsoft Cloud PC, public preview), Docker Sandboxes (cross-platform microVM), and an Intune-deployed bubblewrap wrapper for Linux and WSL2, with restrictive defaults for any session that runs outside a sandbox.

> **THE MICROSOFT × ANTHROPIC CONVERGENCE THESIS**
>
> **Agent 365 is the strategic destination; the supplementary controls in this paper are the bridge.** Microsoft has clearly committed to extending Agent 365 to cover Claude Code and other third-party agents. Until that coverage reaches GA, an enterprise that adopts Claude Code today needs compensating controls authored against the same Microsoft platform. Every supplementary control in this paper is designed to retire cleanly into Agent 365 functionality once Microsoft ships the corresponding native capability — the Defender XDR custom detections retire when Shadow AI covers Claude Code, the per-persona Entra group structure plugs directly into Tools Management and Policy Templates, and the Stage 3 isolation paths align with Windows 365 for Agents and the Defender for AI agents detection model.

> **BITSUMMIT MICROSOFT SOLUTIONS PARTNER CREDENTIALS**
>
> **This architecture is authored by BITSUMMIT under two active Microsoft Solutions Partner designations.** Solutions Partner for Data and AI (Azure) covers AI workload deployment and governance, Azure AI services integration, and AI security posture management — the disciplines that this paper exercises directly in Stages 1 and 2. Solutions Partner for Digital and App Innovation (Azure) covers modern application engineering, DevOps automation, and the developer-tooling integration disciplines that Stage 3 exercises. These designations require demonstrated technical capability across a defined competency framework, validated customer outcomes, and current Microsoft certification coverage maintained across the BITSUMMIT delivery practice.

CONTENTS

Table of Contents

SECTION 1

BITSUMMIT × Microsoft: Why This Architecture, Why This Partnership

This paper is published by BITSUMMIT, a Microsoft Solutions Partner for Data and AI plus Digital and App Innovation. The architecture, the tooling choices, and the implementation plan are all anchored in the Microsoft 365 platform. This section explains why that anchor is intentional and what it means for the reader.

BITSUMMIT's Microsoft Solutions Partner Designations

Under the Microsoft Cloud Partner Program, BITSUMMIT holds two active Solutions Partner designations directly relevant to AI agent governance work:

Designation	Relevance to This Architecture
Solutions Partner for Data and AI (Azure)	The competency that covers AI workload deployment and governance, Azure AI services integration, AI security posture management, and the practitioner skill required to operate the Microsoft AI security stack: Microsoft Purview, Defender for AI agents, Microsoft Entra ID Governance for agents, and the Tools Management surface in Agent 365. The Stage 1 detection authoring and the Stage 2 persona work in this paper exercise this designation directly.
Solutions Partner for Digital and App Innovation (Azure)	The competency that covers modern application engineering, DevOps automation, and developer-tooling integration. The Stage 3 sandbox isolation work in this paper — Windows 365 for Agents, Docker Sandboxes via Intune Win32 packaging, the cco bubblewrap wrapper for Linux and WSL2 — exercises this designation directly. The hooks server, MCP gateway, LLM gateway, and OpenTelemetry collector deployments are all developer-platform engineering disciplines.

These designations require demonstrated technical capability across a defined competency framework, validated customer outcomes, and current Microsoft certification coverage maintained across the BITSUMMIT delivery practice. They are reaccredited annually. They position BITSUMMIT to act as the implementation bridge between Anthropic's Claude Code tooling and an enterprise Microsoft 365 estate.

Why a Microsoft-Anchored Architecture for AI Agent Governance

The decision to anchor an enterprise Claude Code governance architecture in the Microsoft 365 platform is not a default; it is a considered strategic choice. There are three reasons it is the right choice for organizations whose primary identity, endpoint, and security tooling already lives in Microsoft 365.

Reason one — *Identity and policy convergence*

Microsoft Entra is now the de facto identity layer for both human users and AI agents in the Microsoft 365 estate. Agent 365's Agent ID provisions every agent in Entra. Entra ID Governance for agents brings agents into the same access package and sponsor lifecycle workflows used for people. Conditional Access can now apply to agent identities just as it applies to user identities. Building Claude Code governance on a non-Microsoft identity layer would require duplicating this entire surface area; building it on Entra means every agent identity governance investment Microsoft ships extends to Claude Code automatically.

Reason two — *Defender as the unified threat detection plane*

Microsoft Defender XDR's data tables — `DeviceProcessEvents`, `DeviceNetworkEvents`, `DeviceFileEvents`, `CloudAppEvents`, `IdentityLogonEvents` — already collect the telemetry needed to detect Claude Code activity. The Defender for AI agents preview surfaces agent-specific signals into the same Defender XDR alert and incident model. Microsoft Sentinel ingests Defender XDR alerts, Compliance API events, and OpenTelemetry tool events into a single workspace where the SOC operates today. The supplementary controls in this paper feed into existing Microsoft surfaces; nothing requires a new SIEM or a new EDR.

Reason three — *Intune as the universal endpoint policy plane*

Microsoft Intune already manages the Windows, macOS, and (with the Linux preview) Linux endpoints in scope for Claude Code governance at most enterprises. The Stage 1 Win32 app for Windows, the configuration profile for macOS, and the script-based deployment for Linux all ride on infrastructure that already exists. Per-persona Stage 2 profiles distribute through Intune assignments scoped to Entra groups. Stage 3 sandbox tooling (Docker Desktop, the cco wrapper, the centralized hook log forwarder) all package as Intune Win32 apps.

How Agent 365 Maps to Each Stage of the Plan

Microsoft Agent 365 is not a passive backdrop in this architecture. Each stage of the implementation plan exercises a specific Agent 365 surface — either at GA today or at the public preview window opening in June 2026. The table below shows the mapping.

Stage	Agent 365 Surface Used	Today (GA) vs Preview / Roadmap
Stage 1	Defender XDR custom detections (existing Defender XDR engine; no new license). Anthropic Compliance API activation. Conditional Access advisory.	Defender XDR detection authoring is GA today. Agent 365 Shadow AI for Claude Code is on the public preview / roadmap; the Defender XDR detections retire cleanly into Shadow AI when that coverage ships.
Stage 2	Entra ID Governance for agents (access packages and sponsor workflow). Agent 365 Tools Management for tenant-wide MCP allowlisting. Agent 365 Policy Templates as the strategic policy bundle once Claude Code-specific controls reach GA.	Entra ID Governance is GA. Tools Management is GA but does not yet auto-push `allowManagedMcpServersOnly` to Claude Code endpoints — the Stage 2 endpoint reinforcement bridges this. Policy Templates for Claude Code are on the roadmap.
Stage 3	Windows 365 for Agents (Microsoft Cloud PC for elevated path). Microsoft Defender for AI agents detection (preview through Defender for Cloud Apps).	Windows 365 for Agents is in public preview, US-only at preview window. Defender for AI agents is preview. Both are positioned for the elevated profile path; Docker Sandboxes and bubblewrap are the cross-platform interim and complement.

What This Paper Adds to Microsoft's Native Story

Microsoft has published authoritative documentation on Agent 365, Defender for AI agents, and the broader AI agent governance posture. This paper does not replace that documentation; it complements it in three specific ways:

- **Claude Code-specific implementation detail.** The paper provides KQL detection queries, managed-settings.json fragments, hook server reference implementations, and per-OS sandbox launchers tuned to Claude Code's specific behaviour, including the `--dangerously-skip-permissions` flag and the Task subagent exit-2 bypass. These are not in Microsoft documentation because they are Claude Code-specific.
- **Bridge architecture for the GA gap.** Microsoft is honest that Claude Code coverage in Shadow AI, Tools Management endpoint enforcement, and Defender for AI agents Claude Code support are not all GA today. The paper provides the supplementary control set that closes the gap until each piece reaches GA.
- **Honest validation discipline.** Every technical claim is sourced. Every preview is labeled. Every estimate is marked. The paper avoids promising Microsoft capabilities that have not shipped and avoids overselling supplementary controls that should retire when the platform matures.

> **FOR THE MICROSOFT READER**
>
> If you are reading this paper from a Microsoft account team or partner readiness perspective, the implementation plan is designed to be a customer-ready deliverable for organizations adopting Claude Code. BITSUMMIT can be brought into the customer engagement directly through the Microsoft Cloud Partner Program. Engagement-grade scoping for individual customer deployments lives in the companion document, BITDC-040 v4.0 (the Statement of Work).

SECTION 2

Scope and Context

This paper defines what is in scope, what is out, and the operating context that shaped the recommendations.

What This Paper Covers

This paper provides a prescriptive technical implementation plan for governing the Claude Code command-line agent within an enterprise Microsoft 365 tenant. It addresses three governance stages: a global managed-settings baseline (Stage 1), role-based profiles distributed via Microsoft Intune (Stage 2), and sandbox isolation for elevated-permission sessions (Stage 3). For each stage it identifies the concrete capability gaps left unresolved by Microsoft Agent 365 alone today, and prescribes tools, configuration steps, and validation procedures to close them.

The architecture targets developer endpoints running Windows 11 (with optional WSL2), macOS, and Linux, all enrolled in Entra ID and Microsoft Intune. It assumes a tenant on Microsoft 365 E5 or E7, with Anthropic Claude Code accessed through either a Claude Enterprise plan or an Anthropic API key issued to the organization.

What This Paper Does Not Cover

The following are explicitly out of scope and are flagged where they intersect Claude Code governance:

- Claude Cowork governance, except where it overlaps with Claude Code controls. Cowork activity is excluded from the Compliance API, audit logs, and data exports as of the date of this paper, and warrants its own architecture review.
- Governance of agents built on AWS Bedrock, Google Vertex AI, or non-Anthropic platforms. Agent 365 registry sync with Bedrock and Google Cloud is in public preview and is mentioned only where relevant to multi-cloud posture.
- End-user productivity Copilot governance (Microsoft 365 Copilot Chat, Copilot in Word / Excel / Outlook). These are governed natively by Microsoft and do not require the additional tooling described here.
- Procurement, licensing negotiation, or commercial terms with Anthropic or Microsoft.

The Operating Context

Claude Code is now a primary developer tool inside the organization. Developers run it interactively for hours per day, and increasingly invoke it headlessly through scheduled jobs and CI pipelines. Each session has the same OS-level permissions as the developer's user account, can read source repositories and credential stores, can execute arbitrary Bash commands, can call MCP servers that reach internal databases and APIs, and can fetch resources from the open internet. Without enforced policy, every Claude Code endpoint is a potential exfiltration path and a potential lateral movement vector if compromised by prompt injection or supply-chain attack.

Microsoft Agent 365 became generally available on 01 May 2026. Its feature set is genuinely useful for tenant-side control: a central agent registry, policy templates that bundle Entra, Purview, Defender, and SharePoint controls, Tools Management for tenant-wide MCP server allowlisting, Entra ID Governance for agents, and the new Shadow AI page powered by Defender and Intune. The roadmap to bring Claude Code into Shadow AI detection has been published but no GA date is committed. Stage 1 of this implementation plan must therefore close that visibility gap with custom detection rules until Microsoft delivers the native experience.

SECTION 3

Methodology and Source Validation

Every technical claim in this paper is sourced from primary documentation current as of 02 May 2026. Where vendor capability is in preview or on a roadmap, this is stated explicitly.

Data Sources

The technical content in this paper was assembled from the following categories of primary sources, each cross-checked against at least one secondary source where conflicts existed:

Source Category	Specific Sources	Use in This Paper
Microsoft Official Documentation	learn.microsoft.com (Agent 365 overview, Intune Win32 apps, Defender XDR custom detections, AI agent detection and protection, Entra ID Governance for agents, Purview AI policies)	Agent 365 capabilities, Intune deployment patterns, KQL detection authoring, Entra access package design
Microsoft Security Blog and Tech Community	microsoft.com/security/blog (Agent 365 GA announcement 01 May 2026), techcommunity.microsoft.com (What's New in Agent 365 May 2026)	Currently-shipping features, roadmap items, Shadow AI coverage timing, Defender for AI agents preview
Anthropic Official Documentation	code.claude.com (settings, permissions, hooks, sandboxing, security), support.claude.com (Compliance API, OpenTelemetry, enterprise configuration)	Claude Code configuration mechanics, permission semantics, hook event model, sandbox modes
Anthropic Code Repositories	github.com/anthropics/claude-code (deployment templates), github.com/anthropics/claude-code/issues (bug reports)	Reference deployment templates for Jamf, Intune, Group Policy; the Task tool exit-2 bypass bug (#26923)
Independent Practitioner Sources	Vendor blogs (Docker, TrueFoundry, MintMCP), security research (Blue Cycle, Harmonic, Agentic Control Plane)	Real-world deployment patterns and known-issue corroboration; only used where consistent with primary sources

Validation Discipline

The following discipline applies throughout this paper:

- Every named feature, file path, and command-line flag was checked against at least one Anthropic or Microsoft primary source dated within the last 90 days.

- Every preview, roadmap, or in-development capability is labeled as such inline. Generally-available capabilities are stated without qualification.
- Every quoted product limit, pricing figure, or release date carries a source citation in Appendix C.
- Where two sources disagreed (for example on Windows managed-settings paths), the most recent vendor-authored source was preferred, and any remaining ambiguity is flagged in the body text.

Known Limitations

This paper has the following known limitations, each of which should be revisited as the platforms mature:

- **Microsoft Agent 365 capabilities will continue to expand.** Microsoft has signaled that Claude Code Shadow AI detection, additional policy controls, and broader cross-cloud agent registry coverage are coming. Where this paper recommends compensating tooling for a specific gap, that recommendation should be re-evaluated each time Microsoft publishes an Agent 365 release note.
- **Anthropic is iterating on Claude Code's hook and sandbox model.** The Task tool PreToolUse exit-2 bypass is an active issue (anthropics/claude-code#26923). The fix may land before this paper's recommendations are fully implemented; verify status before deploying compensating controls.
- **Windows 365 for Agents was in public preview at the time of writing** and is US-only at the preview window. Its applicability to non-US deployments will need to be reconfirmed when GA arrives.
- **Pricing has not been validated.** Where this paper references license tiers (Agent 365 standalone, Microsoft 365 E7), no commercial validation has been performed. `[Insert: validated pricing per Anthropic / Microsoft account team quote]`.

SECTION 4

Current State: Three-Stage Plan and Agent 365 Capabilities

This section establishes baseline facts about the requesting organization's three-stage governance plan and the capabilities Microsoft Agent 365 provides at general availability. The Agent 365 capability detail is intentionally rich because Agent 365 is the strategic Microsoft platform on which the long-term governance posture is built.

The Three-Stage Governance Plan

The requesting organization defined a three-stage progression for Claude Code governance. The stages are intentionally sequential — each builds on the prior stage's foundation, and all three together are required to fully constrain elevated-permission sessions.

Stage	Required Capabilities	Out of Scope at This Stage
STAGE 1	Settings and policies that prevent the highest-impact risks. Global settings applied through enterprise-managed `managed-settings.json`. Selected users participate on demand for the trial. EDR-side detection authoring closes the Shadow AI coverage gap until Microsoft ships native Claude Code support.	Role-based access control. Differentiated profiles per workflow. Sandbox enforcement. Tool-by-tool allowlisting beyond the global baseline.
STAGE 2	Role-based access control. Each user assigned a profile suitable for their workflow. `managed-settings.json` and hooks distributed by MDM (Microsoft Intune). MCP allowlisting reinforced at the endpoint to make tenant-side Tools Management binding.	Sandbox enforcement. Cloud-hosted agent isolation. Forced isolation of elevated-permission sessions.
STAGE 3	Elevated-permission Claude Code agents sandboxed for complete control of access. Hybrid sandbox path: Windows 365 for Agents (cloud), Docker Sandboxes (microVM, cross-platform), bubblewrap via cco wrapper (Linux native, WSL2 on Windows). Centralized hook log forwarder captures session evidence including subagent invocations.	Agents not authored by the organization (Cowork, third-party SaaS agents). Multi-cloud agent governance.

What Microsoft Agent 365 Provides at GA (01 May 2026)

Agent 365 is built around three pillars: **Observe**, **Govern**, and **Secure**. Each pillar maps to a specific set of administrative surfaces in the Microsoft 365 admin center. The capabilities below are generally available unless

explicitly marked as preview. This section is intentionally exhaustive because the strength of the proposed architecture depends on accurately understanding what Agent 365 already provides versus what the supplementary controls bridge.

Observe — Visibility and Telemetry

- **Centralized agent registry.** One inventory of every agent in the tenant, including agents built on Microsoft AI platforms and ecosystem partner agents. Agents register through the Agent 365 SDK or CLI. The registry is the canonical source of truth for who owns each agent, when it was created, what permissions it has, and what its lifecycle state is. Surfaced in the Microsoft 365 admin center under `Agent 365 > Registry`.
- **Overview dashboard.** Real-time view of registered agents, active users, growth trends, connected platforms, runtime hours, and risk signals. Drill-down available to per-agent detail pages. Surfaced under `Agent 365 > Overview`.
- **Shadow AI page.** Identifies local agent activity on Windows devices using Defender and Intune signals. Initial coverage is OpenClaw only at GA. Microsoft has stated that Claude Code, GitHub Copilot CLI, and other widely used local agents will be added on a published roadmap. Until that coverage ships, the Defender XDR custom detections in Stage 1 of this paper close the gap.
- **Cross-cloud registry sync.** Public preview of registry sync with AWS Bedrock and Google Cloud Vertex AI. Microsoft 365 admin center can connect and sync agent inventories across cloud providers. Basic lifecycle governance (start, stop, delete) is being added in subsequent updates.
- **Advanced Hunting integration.** Defender XDR exposes Agent 365 observability data alongside other security data, queryable with KQL. The same hunting surface used by SOC teams for endpoint and identity investigations now covers agent activity. Tables include agent identity events, agent tool invocations (where instrumented), and agent policy decisions.
- **OpenTelemetry instrumentation hooks.** Agent 365 publishes a standard OTel schema for agent activity. Agents that emit OTel events on this schema appear automatically in the Agent 365 telemetry surface. The Stage 2 OpenTelemetry collector in this paper produces compatible output.

Govern — Lifecycle and Policy

- **Policy Templates.** Reusable groupings of Entra, Purview, Defender, and SharePoint policies. Apply standardized settings during agent approval or onboarding instead of configuring policies per agent. Templates can be authored at the tenant level by security architects and applied automatically when an agent is registered. The Stage 2 persona work in this paper aligns with Policy Templates as the strategic destination once Claude Code-specific Policy Template support reaches GA.
- **Tools Management.** Tenant-wide allowlist and blocklist for tools agents can use, including Microsoft MCP servers. Centralized control point for MCP server approval. Surfaced under `Agent 365 > Tools Management`. The list is enforced for native Microsoft agents at GA; for third-party agents like Claude Code, endpoint-side reinforcement (the `allowManagedMcpServersOnly: true` setting plus the corporate MCP gateway) makes the policy binding.
- **Rules-based governance.** Automated lifecycle rules including expiring inactive agents, auto-blocking risky agents, and auto-reassigning ownerless agents. Configurable thresholds: an agent that has not been used in N days can be

automatically retired; an agent whose owner has left the organization can be automatically reassigned to a new sponsor.

- **Microsoft Entra ID Governance for agents.** Access packages and sponsor lifecycle workflows that bring agents into the same identity governance model used for people. Each agent identity has an accountable user. Quarterly access reviews can be scheduled. Sponsor approval can be required for agent creation. The Stage 2 Elevated profile in this paper uses Entra access packages as the gating mechanism for the elevated grant.
- **Lifecycle actions.** Install, publish, block, unblock, delete, or reassign owner directly from the Agent 365 registry. These actions propagate through Defender, Intune, and Entra automatically when integration is configured.
- **Purview integration for agent activity classification.** Microsoft Purview sensitivity labels apply to agent-accessed data. Agents that touch labeled data are subject to the same DLP, retention, and information protection policies as users.

Secure — Identity, Data, and Threat Protection

- **Microsoft Entra enforcement.** Risk-based access controls for users and agents, including agents acting on behalf of users with delegated access. Conditional Access policies can target agent identities, including blocking agent access from untrusted devices, requiring compliant device posture, and enforcing session policies via Microsoft Defender for Cloud Apps.
- **Microsoft Purview.** Information protection, DLP, and risk safeguards extended to agent activity. Purview AI policies (preview at the time of writing) provide specific guardrails for AI tooling access to sensitive data. Purview communication compliance can flag agent-generated content for review where applicable.
- **Microsoft Defender for AI agents (preview).** Detects suspicious and malicious agent behavior, blocks dangerous actions in real time, alerts in near-real-time. Onboarding currently goes through Defender for Cloud Apps. Preview detection categories include: anomalous tool invocation patterns, large file fetches outside normal hours, prompt injection signature matches, and credential exposure indicators in agent context windows. Coverage extends to OpenClaw at preview launch with Claude Code on the roadmap.
- **Intune endpoint policies.** Block unsanctioned agent execution at the endpoint. Combined with Shadow AI detection on Windows. Intune device configuration profiles can deploy the managed-settings.json baseline for Claude Code as Win32 apps with SYSTEM-context installation.
- **Windows 365 for Agents (public preview, US only).** Cloud PC purpose-built for agent workloads, with built-in Intune integration and Defender visibility. Provisioned through the Microsoft 365 admin center under `Cloud PC > Provisioning policies`. Custom images can be authored with Claude Code, the cco wrapper, the OpenTelemetry SDK, and the Stage 2 managed-settings.json pre-installed. Network egress is restrictable to a defined allowlist.
- **Microsoft Sentinel integration.** Sentinel ingests Defender XDR alerts, Agent 365 lifecycle events, Compliance API events (where activated), and OpenTelemetry tool events. Workbooks can correlate agent activity with user activity, identity risk signals, and other security telemetry.

Licensing

Agent 365 is generally available for the Commercial segment as of 01 May 2026, licensed on a per-user basis. The recommended prerequisites are Entra P1, Entra P2, or Entra Suite, plus Purview Data Loss Prevention. Agent 365 is

also available as part of Microsoft 365 E7 (\"Frontier Suite\"). `[Insert: validated per-user pricing for Agent 365 standalone and Microsoft 365 E7 bundle from account team quote]`.

> **STRATEGIC IMPLICATION**
>
> **Every supplementary control in this paper has a defined retirement path into Agent 365.** The Defender XDR custom detections retire when Shadow AI covers Claude Code. The Stage 2 Entra group structure plugs into Tools Management and Policy Templates. The Stage 3 Windows 365 for Agents elevated path is a Microsoft-native primary destination once GA reaches the customer's region. The MCP gateway aligns with Tools Management's tenant policy. The OpenTelemetry collector aligns with the Agent 365 OTel schema. The architecture is designed to converge with Agent 365, not run parallel to it.

SECTION 5

Requirements and Constraints

The architecture must satisfy six business requirements, three regulatory constraints, and four operational constraints.

Business Requirements

ID	Requirement	Rationale
BR-1	Prevent exfiltration of source code, credentials, and customer data through Claude Code sessions	Claude Code has full user-equivalent permissions on the endpoint and can reach internal systems through MCP servers and external services through WebFetch. Without enforced controls, a single prompt injection or compromised dependency can become a breach.
BR-2	Maintain a complete audit trail of agent activity sufficient for SOC 2 and ISO 27001 evidence	Auditors will increasingly require evidence of agent governance. The trail must cover both control-plane events and tool-level activity.
BR-3	Permit different Claude Code permission profiles for different developer roles	A platform engineer needs broader Bash access than a junior developer; both need less than a security researcher running elevated investigations. One-size-fits-all settings either over-restrict productivity or under-restrict risk.
BR-4	Allow elevated-permission sessions only inside an isolated execution environment	Elevated profiles by definition relax restrictions. The risk of doing so unbounded is unacceptable; sandboxing is the compensating control.
BR-5	Enable rollout to a 10-person trial group without affecting other developers	Stage 1 must be deployable to a defined cohort. Rolling managed-settings changes to the entire developer fleet on day one is operationally unsafe.
BR-6	Avoid lock-in to a single AI vendor or single MCP gateway provider	The agent landscape will evolve. The architecture must use platform-agnostic interfaces (MCP, OpenTelemetry, OIDC) wherever possible.

Regulatory and Compliance Constraints

- **SOC 2 Type II evidence retention.** Audit logs must be retained for at least 90 days and ideally 12 months, in a tamper-evident store accessible to auditors.
- **Data residency.** Where regulated data is in scope, traffic to Anthropic must be evaluated against a Zero Data Retention addendum, and consideration must be given to AWS Bedrock or Google Vertex AI deployment options that keep traffic in private networks.
- **Privacy.** User prompt content and tool parameters may include personally identifiable information. Telemetry routed to a SIEM must be filterable or redactable at the collector before persisted storage.

Operational Constraints

- **Microsoft 365 tenant alignment.** The architecture must use Entra ID for identity, Microsoft Intune for endpoint policy delivery, Microsoft Defender for endpoint and process telemetry, and Microsoft Sentinel as the central SIEM. BITSUMMIT's Solutions Partner designations in Data and AI plus Digital and App Innovation are aligned with this constraint.
- **Multi-OS support.** Windows 11, macOS, and Linux endpoints are all in scope. Bubblewrap-based sandboxing on Linux requires WSL2 on Windows.
- **Headless and CI usage.** Claude Code is used both interactively and in scheduled or CI contexts. Settings must enforce regardless of invocation mode.
- **Skills capacity.** The implementing team includes Entra, Intune, and Defender practitioners but not specialist Claude Code or MCP gateway operators. Tooling that requires significant new operator skill is acceptable only where the alternative is a known security gap.

SECTION 6

Findings: Twelve Capability Gaps Across the Three Stages

This section enumerates the twelve specific capability gaps that the architecture must close. Each finding follows a Driver, Evidence, Implication, and Action pattern. Every finding has a defined retirement path back into Microsoft Agent 365 functionality where one exists.

F-1 Microsoft Agent 365 does not cover Claude Code in Shadow AI at general availability

CRITICAL | STAGE 1

Driver. The Agent 365 Shadow AI page launched on 01 May 2026 with detection coverage for OpenClaw, Microsoft's first-party local agent. Claude Code, GitHub Copilot CLI, and other widely deployed local agents are on a published roadmap with public preview through June 2026 and no committed GA date for Claude Code.

Evidence. Microsoft's GA announcement documentation explicitly lists Shadow AI's initial coverage as OpenClaw and references Claude Code under the "expanding soon" roadmap. The May 2026 What's New post in the Microsoft Agent 365 Tech Community confirms preview availability for additional local agent detection but does not commit a Claude Code GA date.

Implication. Until Microsoft ships native Claude Code coverage in Shadow AI, the Microsoft 365 admin center will not surface Claude Code endpoint activity in the Shadow AI page. SOC and security architecture teams have no Microsoft-native answer to the question "who is running Claude Code on which endpoint, with what flag set, and what process tree did it invoke."

Action. Author six Defender XDR custom detection rules that close the Shadow AI gap using existing Defender XDR data tables (`DeviceProcessEvents` , `DeviceNetworkEvents` , `DeviceFileEvents`). Detection logic is given in Section 9 (Stage 1). These detections retire when Microsoft ships Claude Code coverage in Shadow AI, with no migration cost beyond pruning the custom rules.

F-2 --dangerously-skip-permissions silently disables every PreToolUse and PostToolUse hook

CRITICAL STAGE 1

Driver. Claude Code ships an explicit escape hatch flag that bypasses every PreToolUse and PostToolUse hook for the session. Anthropic documents this flag intentionally for cases where the operator chooses to accept the lack of safety gates. In an enterprise context, with managed-settings deployed centrally, the existence of this flag is the single largest enforcement gap.

Evidence. Anthropic's Claude Code documentation at `code.claude.com/docs/en/permissions` describes the `bypassPermissions` permission mode and the `--dangerously-skip-permissions` CLI flag that activates it. The same documentation lists `permissions.disableBypassPermissionsMode: "disable"` as the managed-settings entry that prevents the mode from being activated. Without this setting, every hook authored in Stages 1 through 3 is defeatable by 32 characters of typing.

Implication. An audit log shows nothing because the hooks that would have written to it never fire. A managed deny-list shows nothing because the hooks that would have enforced it never fire. The entire Stage 1 baseline can be bypassed silently, and the bypass leaves no trace in the hook-generated logs.

Action. Layer three controls. First, set `permissions.disableBypassPermissionsMode: "disable"` in the global managed-settings.json baseline at every endpoint. Second, author Defender XDR custom detection rule D-2 to alert in near-real-time on any process command line that contains `--dangerously-skip-permissions`. Third, deliver a Conditional Access advisory brief covering Defender Network Protection custom indicators that allow `api.anthropic.com` only from sessions authenticated to the corporate Anthropic tenant, blocking any escape-hatch session that uses a personal-tenant Anthropic account.

F-3 Task subagent tool executes despite a PreToolUse exit-2 block (anthropics/claude-code#26923)

CRITICAL STAGE 3

Driver. Claude Code's Task tool spawns subagents that perform their own tool calls. Each subagent inherits the parent session's permissions but its tool calls are evaluated by the parent's hook chain. An active Anthropic bug allows a subagent's tool call to execute despite a PreToolUse hook returning exit code 2, which is the documented mechanism for blocking a tool call.

Evidence. GitHub issue `anthropics/claude-code#26923` filed in February 2026, with verified reproduction across 19 of 19 test sessions. The hook's BLOCKED audit entry is recorded, but the subagent runs to completion anyway. The bug is acknowledged by Anthropic and is on the active backlog at the time of writing. No fix release date is committed.

Implication. Hook-based blocking of subagent activity cannot be relied on. An attacker who induces Claude Code to spawn a subagent (through prompt injection in a fetched URL, a poisoned MCP server response, or a compromised Skill) can execute tool calls that the parent's hooks correctly identify as malicious and correctly attempt to block. The audit trail shows BLOCKED entries; the subagent shows COMPLETED. This is not a theoretical concern; the issue is reproducible today.

Action. Two compensating controls. First, network-layer enforcement at the MCP gateway and the LLM gateway, where subagent tool calls are denied at the wire independent of any client-side hook state. The MCP gateway evaluates each MCP request against tenant policy before forwarding; the LLM gateway can deny model invocations that match prompt-injection signature patterns. Second, the centralized hook log forwarder (Stage 3, see Section 9) records Task tool entry points specifically, providing forensic evidence that a subagent was invoked, even when the hook's BLOCKED determination did not stop execution. The forwarder produces a structured-JSON record per Task invocation suitable for downstream SIEM correlation and SOC alerting.