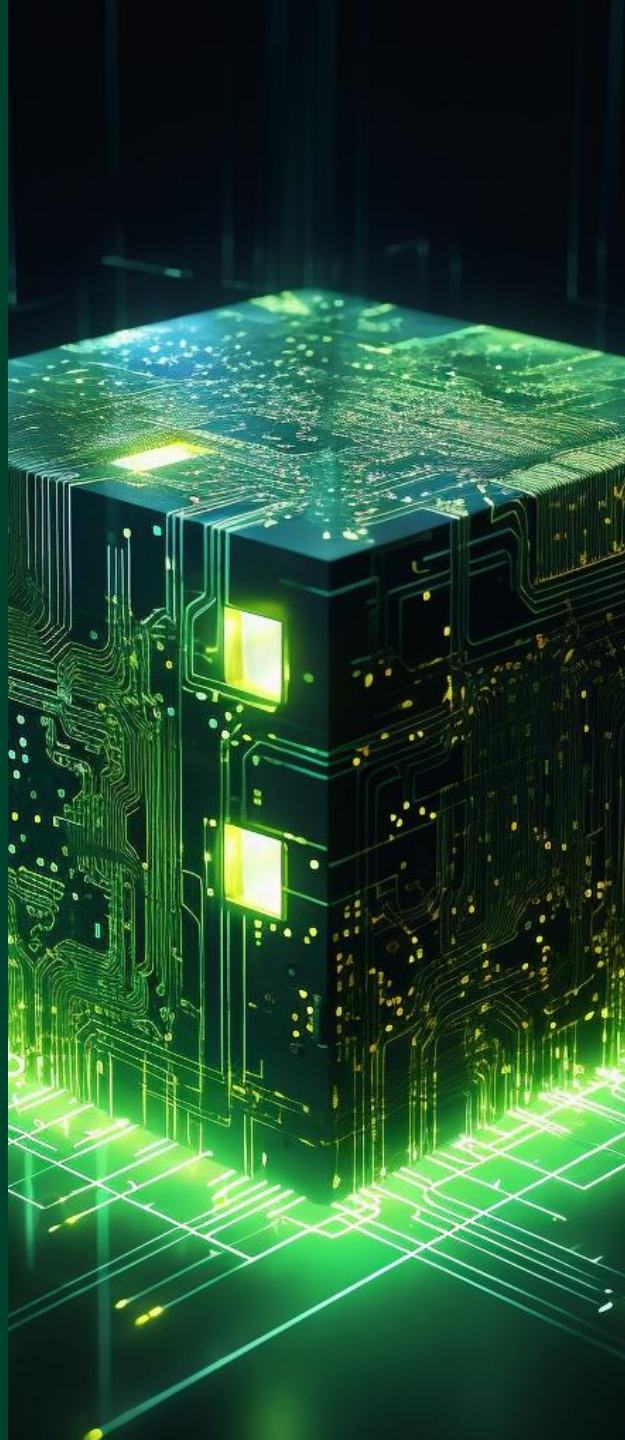
Leash your LLMs: Securing Claude & Copilot

Sean Metcalf, Scott Blake, & Paul Sems

TRUSTEDSEC

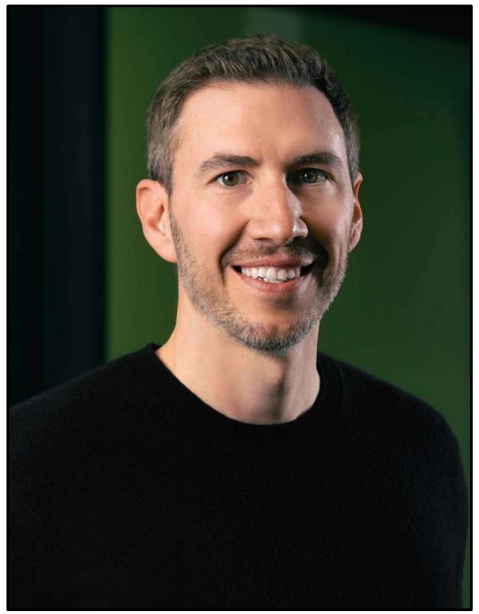


TRUSTEDSEC

Presenting Today



Sean Metcalf
Identity Security Architect



Scott Blake
Senior Security Consultant



Paul Sems
Managing Director of Remediation Services

“Life moves pretty fast. If you don't stop and look around once in a while, you could miss it.”



March 18th, 2026

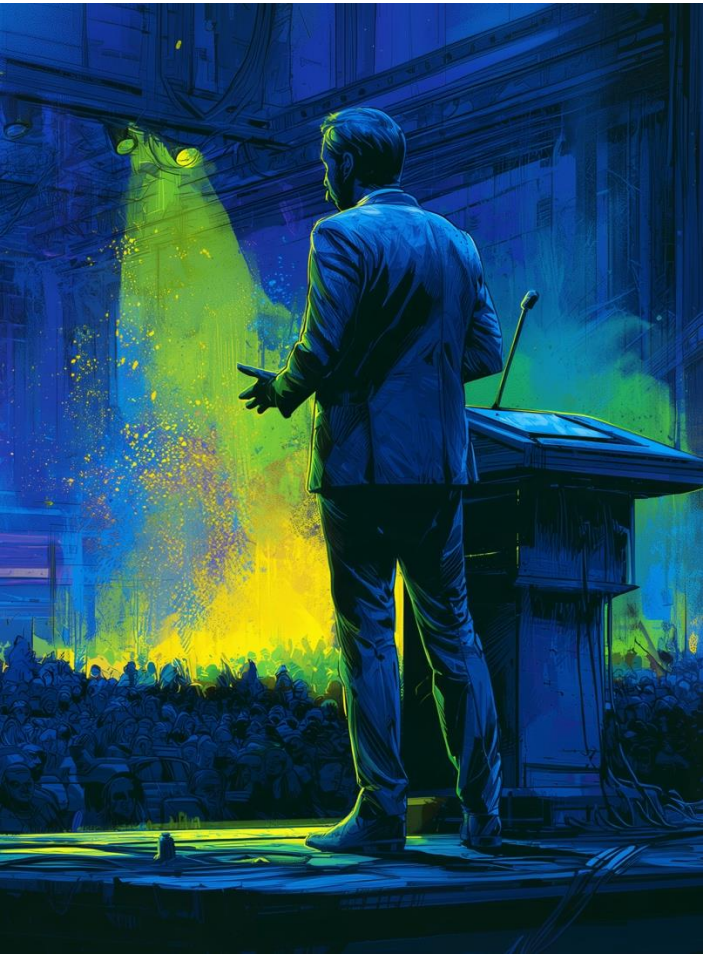
Copilot Security: What to Know before You Go

Sean Metcalf & Paul Sems

TRUSTEDSEC



Agenda



- Anthropic Claude
  - Latest & Greatest
  - Licensing
  - Permissions & controls
  - Hooks & tenant restrictions
  - Connectors, plugins, & skills
  - Reporting & logging
  - Before deploying
- Microsoft Copilot
  - Copilot expansion
  - Permissions & controls
  - Entra ID user settings
  - Logging
  - Agents
  - Licensing
  - Purview Data Security Posture Management (DSPM)
  - SharePoint Advanced Management
  - Prompt Injection
  - ToDos

Anthropic Claude

TRUSTEDSEC



Claude's Latest & Greatest



CHAT & COWORK MERGED INTO ONE CLAUDE FOR ENTERPRISE; AUTO-ENABLED NOVEMBER 16 UNLESS AN ADMIN OPTS OUT



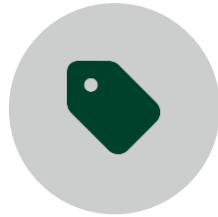
COMPLIANCE API NOW CAPTURES COWORK & CLAUDE CODE SESSION TRANSCRIPTS (GA AUGUST)



COWORK IS ON BY DEFAULT FOR TEAM AND ENTERPRISE ORGS



CLAUDE NOW WORKS INSIDE CHROME, MICROSOFT 365, GOOGLE WORKSPACE, & SLACK



ENTERPRISE IS NOW SELF-SERVE: $20/SEAT PLUS ALL USAGE AT API RATES



NEW ENTERPRISE COST CONTROLS (OCT 28): POOLED GROUP BUDGETS AND SMART REPORTS

Claude Is Not Just Chat Anymore

Claude + Cowork	Claude Code	Claude in Chrome	Claude for Microsoft 365	Claude in Slack	Dispatch & computer use
One conversation for questions and multi-step tasks	Agentic coding in terminal, IDE, desktop and web	Browsing agent that clicks, reads and fills forms	Add-ins for Excel, Word, PowerPoint and Outlook	Tag @Claude in a channel to delegate a task	Assign tasks from a phone that run on the desktop
Runs in Anthropic's cloud by default; reaches local files via Claude Desktop.	Reads files and runs commands	Prompt injection is the top risk: it acts inside logged-in sessions	Compliance API coverage in beta	Enterprise Admin API for channel and workspace settings (beta)	Claude can drive the mouse and keyboard
Uses connectors, plugins and local files	Governed by managed-settings.json	Admin site allow and block lists			Research preview
Cowork on by default (off for HIPAA orgs)	Included with Team and Enterprise	Not available to HIPAA orgs			



Licensing: Claude Team vs Enterprise

	Team	Enterprise
Price	Standard $20 / Premium $100 per seat (annual); $25 / $125 monthly	$20 per seat (annual), plus all usage billed at API rates
Usage	Included, with per-member limits; Premium = 5x Standard	No seat limits; org, group and user spend limits
Seats	2 to 150	20+ self-serve (prepaid), 50+ sales-assisted
SSO and domain capture	Yes	Yes
SCIM provisioning	No	**Yes**
Roles	Built-in roles only	**Groups and custom roles**
Audit logs and Compliance API	No	**Yes**, incl. Cowork, Claude Code and M365 add-in transcripts
Custom data retention	No (content kept indefinitely)	**Yes**
CMEK and US-only inference	No	**Yes** (US-only inference billed at 1.1x)
HIPAA-ready with BAA	No	**Yes**, self-serve or sales-assisted
Inference hooks and Tenant Restrictions	No	**Yes**
Cowork OpenTelemetry	Yes	Yes

Prices from claude.com/pricing and Anthropic support articles, November 2026. Subject to change.

When Team Isn't Enough

You need to investigate an incident

→ Audit logs and the Compliance API are Enterprise-only

Departing employees must lose access automatically

→ SCIM provisioning

Data must be deleted on a schedule

→ Custom retention; the default keeps content indefinitely

Teams need different access

→ Groups and custom roles, e.g. Cowork for one department only

You handle PHI

→ HIPAA-ready configuration and a signed BAA

Keys or inference location are mandated

→ Customer-managed keys, US-only inference

You need inline DLP or to block personal accounts

→ Inference hooks and Tenant Restrictions



Permissions & Controls

- Enforce SSO and domain capture so work accounts can't live outside the organization
- Cowork is on by default and unified Claude turns on Nov 16: decide, or scope by group
- Turn off 'Automatically approve' (Team starts new tasks in it), Remote Control and browser use
- Lock Claude Code with managed settings: forceLoginOrgUUID, allowManagedPermissionRulesOnly, allowManagedMcpServersOnly
- Network egress rules don't cover web search, web fetch, MCP connectors or Claude in Chrome; control those separately
- Enterprise: set org, group and per-user spend limits, since usage is uncapped by default

Enterprise-Only Guardrails

Inference hooks

Enterprise only

- Every governed prompt goes to your AI security server for an allow or deny before inference
- Covers claude.ai, Cowork, Claude Code, M365, Chrome and Claude Tag
- Not covered: Claude Platform API, Bedrock, Google Cloud; voice only partly
- Decide fail-open vs. fail-closed; start in shadow mode

Tenant Restrictions

Enterprise and Console orgs

- Your proxy injects the anthropic-allowed-org-ids header
- Anthropic blocks any Claude org not on your list, including personal accounts
- Works with Zscaler, Netskope, Prisma Access, Cloudflare, Cato; needs TLS inspection
- Proxy must overwrite the header, not append it

Neither exists on Team: if you need either one, it belongs in your Team vs Enterprise decision.

Connectors, Plugins & Skills

What to know

Owners (or Enterprise custom roles) add connectors; each user then connects with their own sign-in

Claude reaches custom MCP servers from Anthropic's cloud, so they must be internet-reachable

Connector tools can write to other systems, not just read from them

Plugins and skills can run code and call connectors; a default marketplace and admin-provisioned skills are on for everyone

What to do

- ✓ Turn off connector requests (on by default on Team)
- ✓ Approve each connector before it touches company data
- ✓ Set write tools to Needs approval or Blocked under Connectors
- ✓ Review OAuth scopes at consent
- ✓ Enterprise: confirm plugin scanning is on; curate your marketplace
- ✓ Inventory connectors, plugins and skills quarterly

Claude Reporting & Logging

Activity Feed

Enterprise (and Console orgs)

- Admin and user activity events
- Kept 6 years
- Queryable within about a minute
- No prompts or content

Compliance API content

Enterprise only

- Chats, files and projects
- Cowork, Claude Code and M365 add-in transcripts (Chrome and Google add-ins in beta)
- Shows what Claude was asked and returned, not what happened on the device
- Not Claude Code cloud sessions

OpenTelemetry

Team and Enterprise

- Real-time Cowork and Claude Code events to your SIEM
- Admin sets the endpoint
- Metadata only by default; prompt capture is an MDM policy, and older Desktop versions send it anyway

While the Compliance API is off, no activity is recorded. Turn it on before rollout and send it to your SIEM.

Interesting Claude Notes

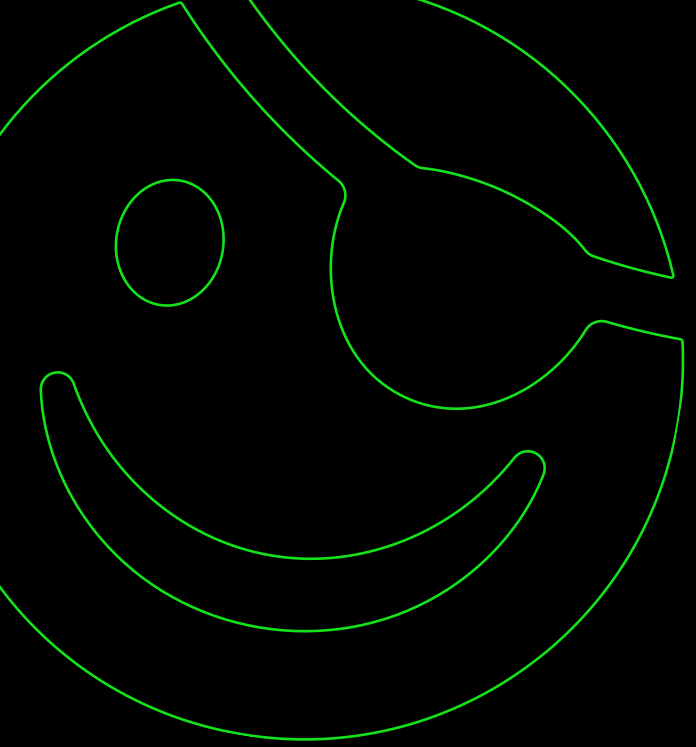
- Incognito chats still show up in the Compliance API
- Cowork cloud sessions are invisible to your EDR
- Smart reports: Claude reads staff chats; names can be shown
- The Compliance API key can read and delete every conversation

https://platform.claude.com/docs/en/manage-claude/compliance-faq



ToDos before Deploying Claude

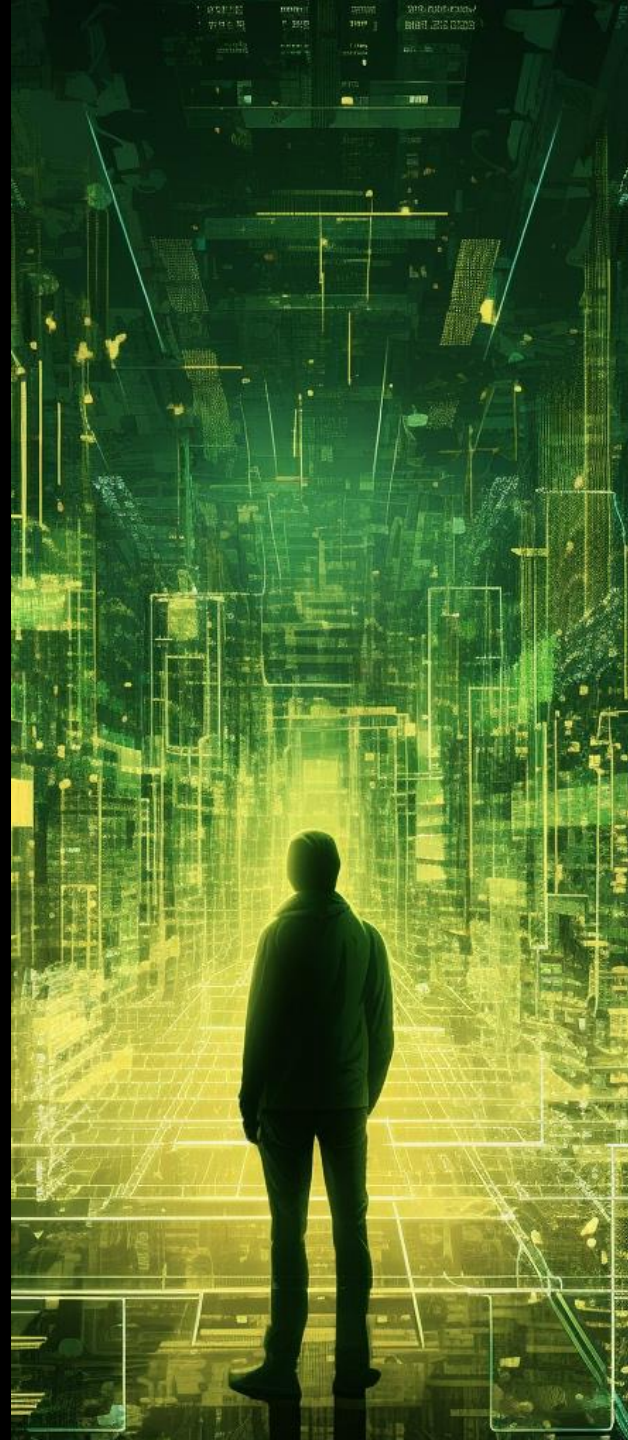
- AI acceptable use policy that covers agents, not just chat
- Pick Team or Enterprise by the controls you need
- Enforce SSO and domain capture; SCIM on Enterprise
- Decide on Cowork and unified Claude before Nov 16
- Keep auto-approve, Remote Control and browser use off
- Lock down Claude Code with managed settings
- Default connectors and plugins to off; approve each one
- Turn on the Compliance API and OTel; send both to your SIEM
- Set a retention window and spend limits
- Enterprise: add inference hooks; block personal accounts



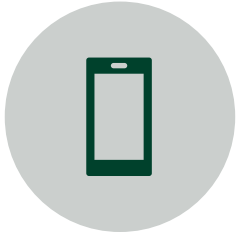
Microsoft Copilot



TRUSTEDSEC



Microsoft Copilot Updates



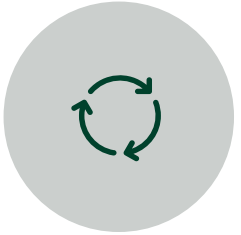
CONSUMER AND COMMERCIAL COPILOT MOVED UNDER ONE ORGANIZATION IN MARCH AND INTO ONE CLIENT APP RECENTLY



CLIENT APP ACCEPTS BOTH MICROSOFT ACCOUNTS AND ENTRA ID WORK ACCOUNTS, WITH DATA KEPT SEPARATE BETWEEN THE TWO



AGENT PLATFORM WITH ITS OWN IDENTITY MODEL, DATA-ACCESS LAYER, RUNTIME AND METERING



NOW EXPLICITLY MULTI-MODEL ANTHROPIC AND OPENAI MODELS AVAILABLE THROUGHOUT



>30M PAID M365 COPILOT SEATS

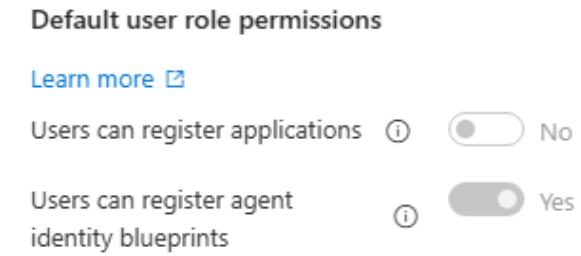
Copilot, Now with More Copilot!

Home	Cowork	Code	Autopilot	Today	Security copilot	Copilot Studio
The new landing environment Combines Chat & Cowork	Built with Anthropic tech Off by default (requires a spending policy) Multi-step work across M365 data & external sources Browser tasks run in the user's own Edge tabs Requires a full M365 Copilot license Billed by usage	Natural language to cloud-hosted internal apps Sandboxed; runs on Copilot Managed Runtime Same technology as GitHub Copilot Billed by usage	Formerly Scout Persistent agent given a name, role and goal Has its own metered agent identity Runs recurring work without per-task prompts Billed through usage-based billing In Private Preview	Proactive command center across mail, calendar, Teams Flags urgent items, drafts replies In Private Preview	Is an agentic AI that can be leveraged across Microsoft Defender, Entra, Intune, and Purview Priced based on security compute units (SCUs) - some available with Microsoft 365 E5 licensing Billed by usage	Platform for creating and managing custom agents, workflows, and apps with limited code

Permissions & Controls

- **Federated Copilot connectors**: real-time MCP access using the user's identity, not stored or indexed by Microsoft
- ServiceNow Knowledge/Catalog connectors enforce ServiceNow role permissions
- **Synced connectors (e.g., ServiceNow):** Admins can now edit the indexing query filter and the identity mapping in place
- SharePoint authoritative sites prioritized in Copilot Search
- PowerShell opt-outs (*KnowledgeAgentScope* & *KnowledgeAgentSelectedSitesList*) are no longer honored
  Use SharePoint Site AI settings or Restricted Content Discovery (can take >1 week to apply for large sites)

Entra ID User Settings: Agent Identity Blueprints



Entra ID users can create agent identity blueprints starting today (Nov 4th) with global rollout expected to complete in late November

Recommended to disable this capability

Configure User Settings: *"Users can register agent blueprints"*- Set to "No"

Agent identity blueprints are templates that define identity configuration and credentials for AI agents

Allowing all users to create blueprints has similar risk to allowing all users to register applications (default is enabled as well)

When a user creates an agent identity blueprint, they become the owner of it until removed - provides full control access to the blueprint and applied agent identities

Agent ID Developer and Agent ID Administrator can still create them

Inventory existing blueprints & owners

Copilot > Settings > AI providers operating as Microsoft subprocessors > Anthropic

- It’s off by default in commercial tenants in the Europe & the UK, and on by default in other areas for Anthropic
- Anthropic models are excluded from EU Data Boundary and in-country processing commitments
- Access can be scoped to Entra groups, and that scoping applies across Copilot and Copilot Studio
- “Anthropic models with Data Retention” is a separate opt-in. With it, Anthropic is an independent processor that keeps data up to 30 days (for flagged content up to 2 & classifier scores up to 7 years), outside Microsoft’s DPA
- Some features depend on Anthropic models

Copilot Studio and Power Platform Governance

- Every licensed user is a maker in the default environment, and any M365 Copilot license includes Copilot Studio access
- Agents set to “no authentication” or using the maker’s credentials overshares through the creator’s access
- Set new connectors to Blocked by default
- Block channel connectors in a data policy and use sharing limits to stop sharing with the whole organization
- Agents published to Microsoft 365 Copilot or Teams can require admin approval before users see them (M365 admin center > Agents)

Copilot Reporting

Microsoft 365 Copilot Chat

This report applies to Microsoft 365 Copilot Chat users who don't have a Microsoft 365 Copilot license assigned to them. Learn more about Copilot Chat

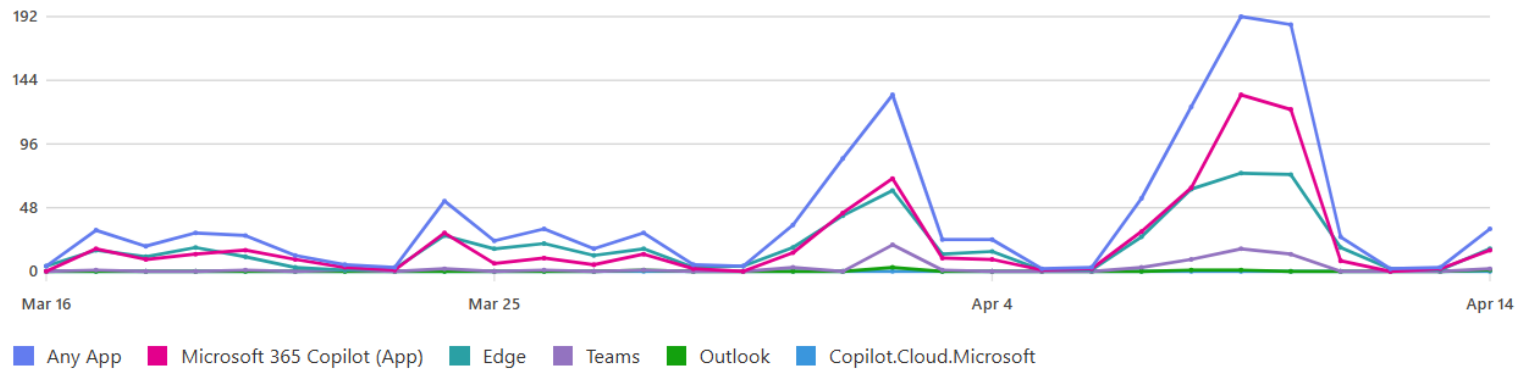
Filters: Periods: **Past 28 days (Jun 2, 2026 - Jun 29, 2026)**

Active users	Average daily active users	Total prompts submitted	Average prompts submitted per user
908	512	40,860	45

Adoption trend by app

Summary | Trend | Apps: **All**

The number of active users of Microsoft 365 Copilot Chat for the selected period. See metric definitions



TRUSTEDSEC

https://go.microsoft.com/fwlink/p/?linkid=2024339

Copilot Logging

Audit: create a custom retention policy for CopilotInteraction if you want logs longer than the default (180 days)

Retention: conversations are stored in each user's Exchange mailbox (in a hidden folder) which inherits retention policies

eDiscovery: searching with the "Copilot activity" condition includes transcripts, created files (saved in OneDrive), scheduled prompts, and conversations

Capture: Send the logs to Sentinel or Defender XDR (or the SIEM of your choice)

Copilot Pages & Notebooks: live in SharePoint embedded containers so they have their own retention, eDiscovery, and offboarding behavior.

Note that a transcript shows what the agent said, not what it did



Agents

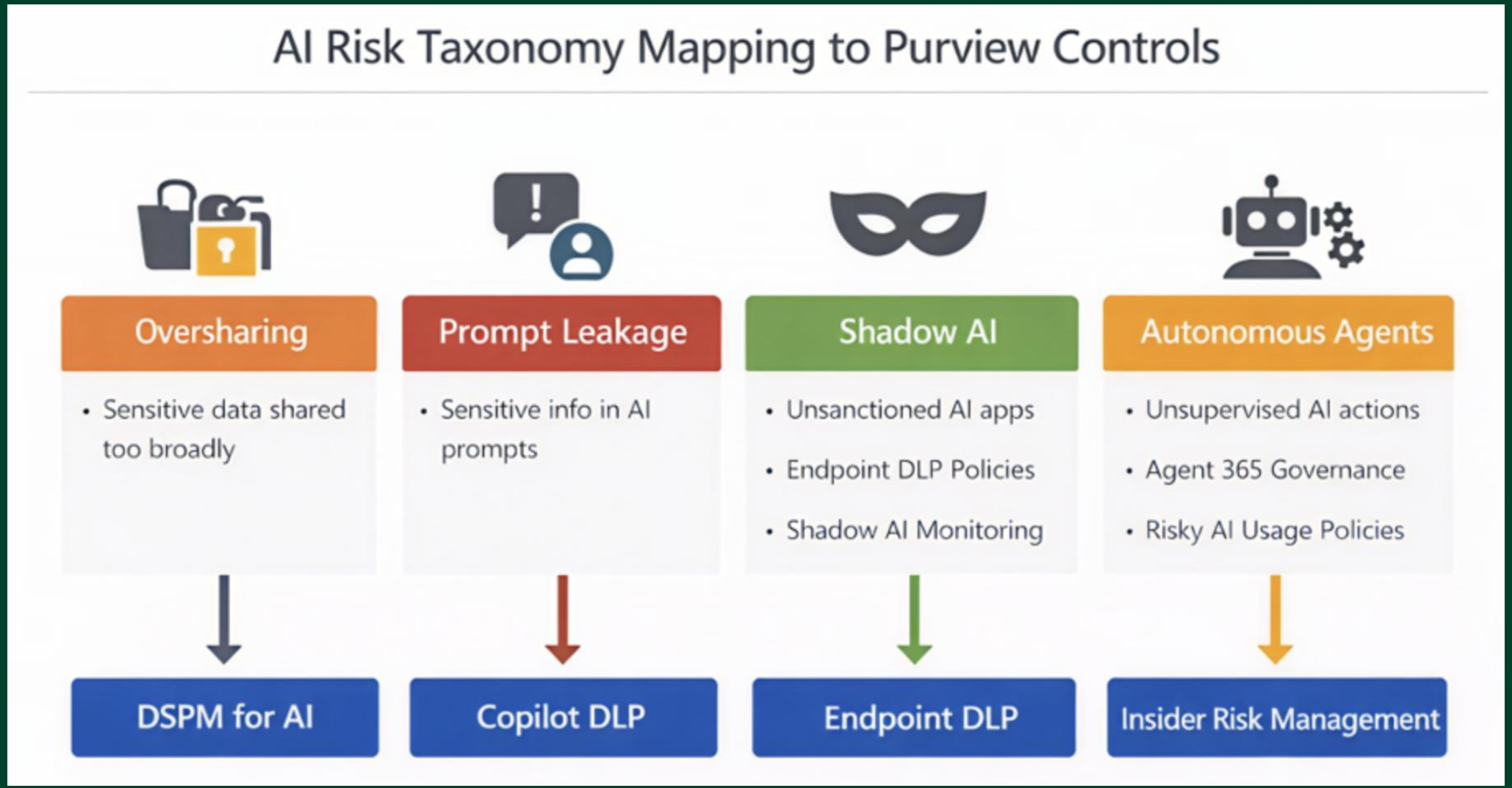
- Entra Agent ID per agent
- Inventory & manage using Agent 365
- $15/user/month requires E5-class base license) licensed to the human

Licensing

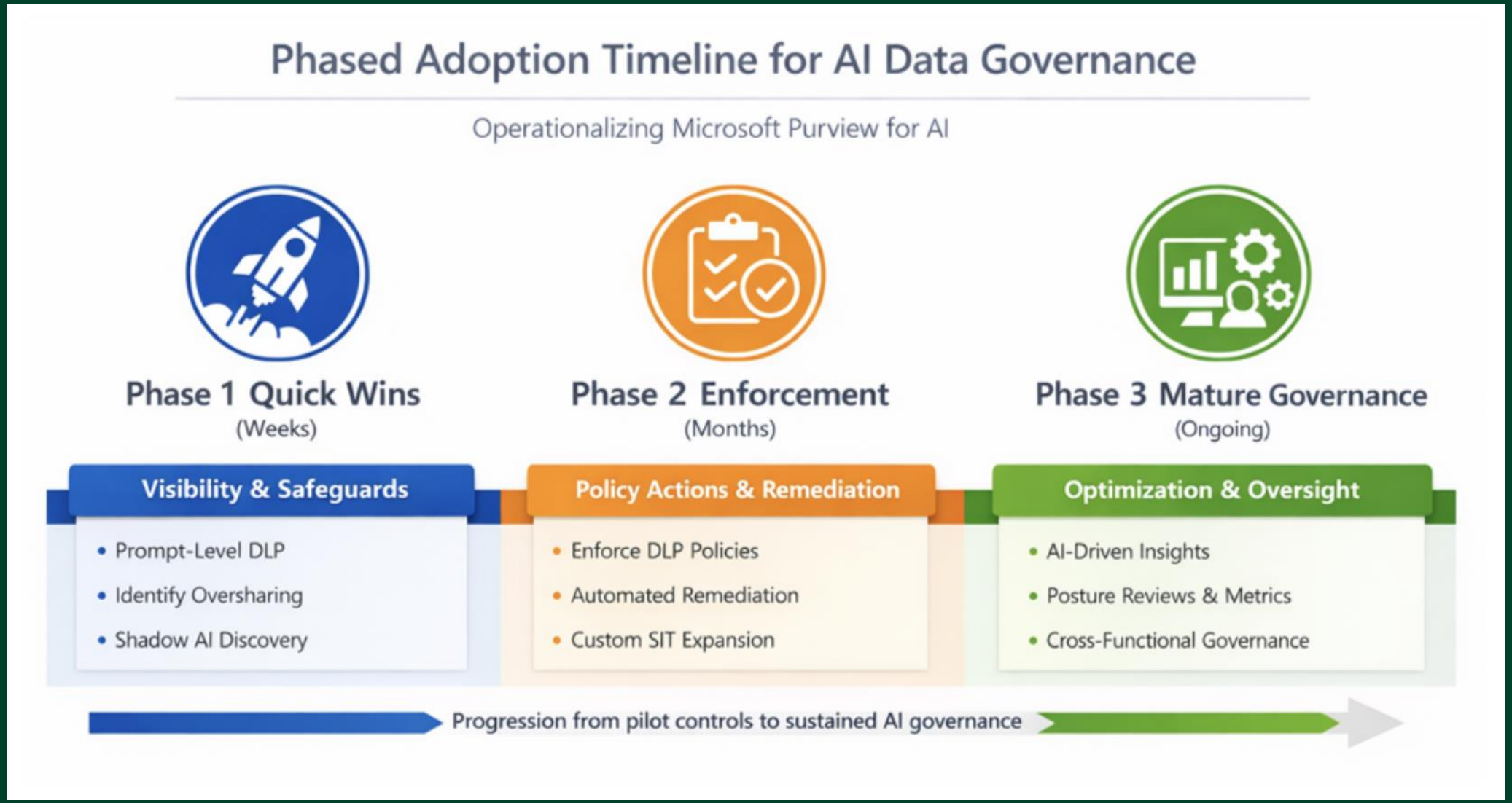


Copilot Credits are $0.01 pay-as-you-go or $200/month per 25k-credit pack billed annually

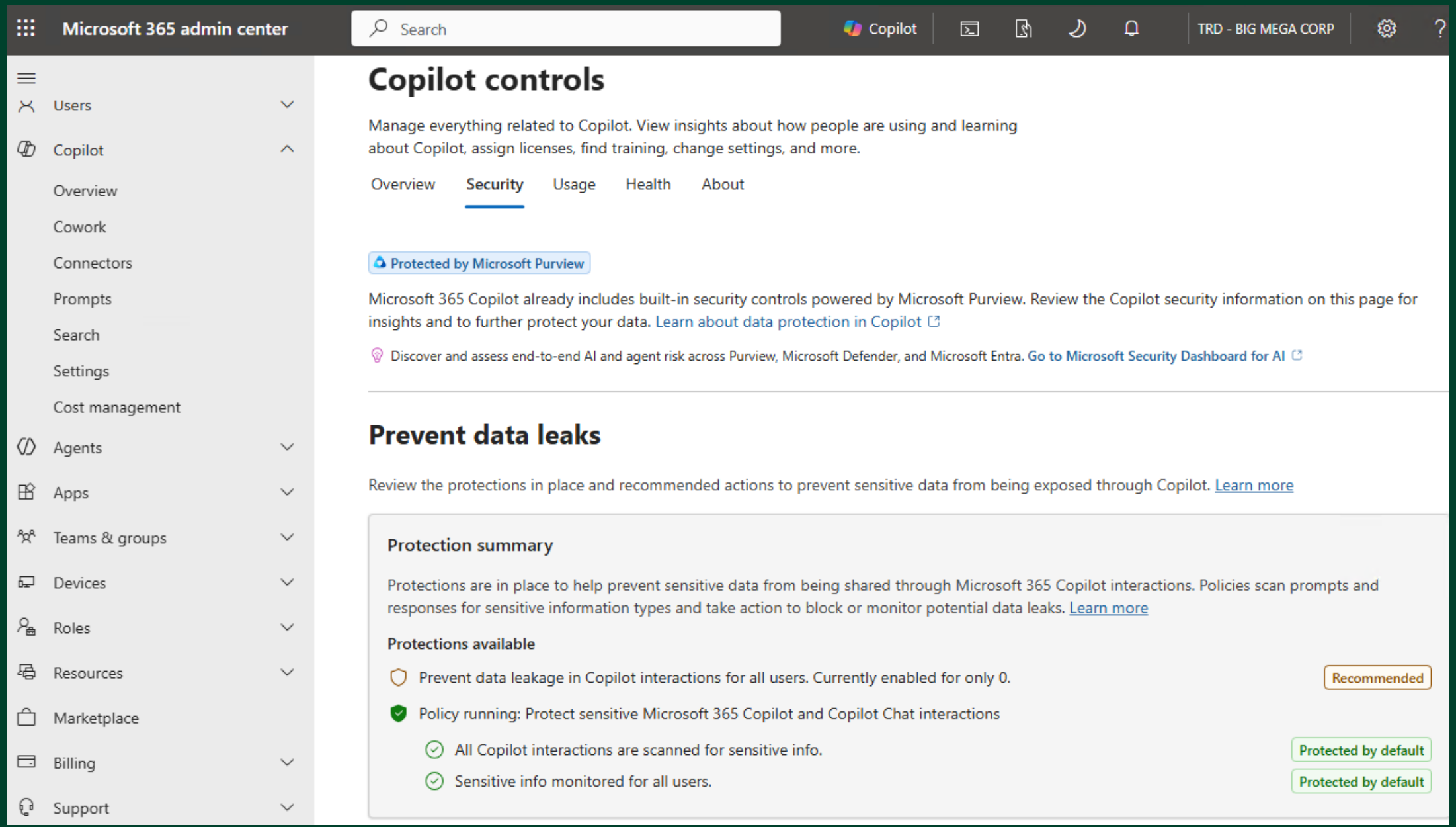
AI Risk Taxonomy Mapping to Purview Controls



https://techcommunity.microsoft.com/blog/microsoft-purview-blog/from-oversharing-to-enforcement-a-practical-guide-to-ai-data-security-with-micro/4513727



https://techcommunity.microsoft.com/blog/microsoft-purview-blog/from-oversharing-to-enforcement-a-practical-guide-to-ai-data-security-with-micro/4513727



TRUSTEDSEC

https://admin.cloud.microsoft/#/copilot/scc



Secure and govern Microsoft Copilot: Foundational deployment guidance

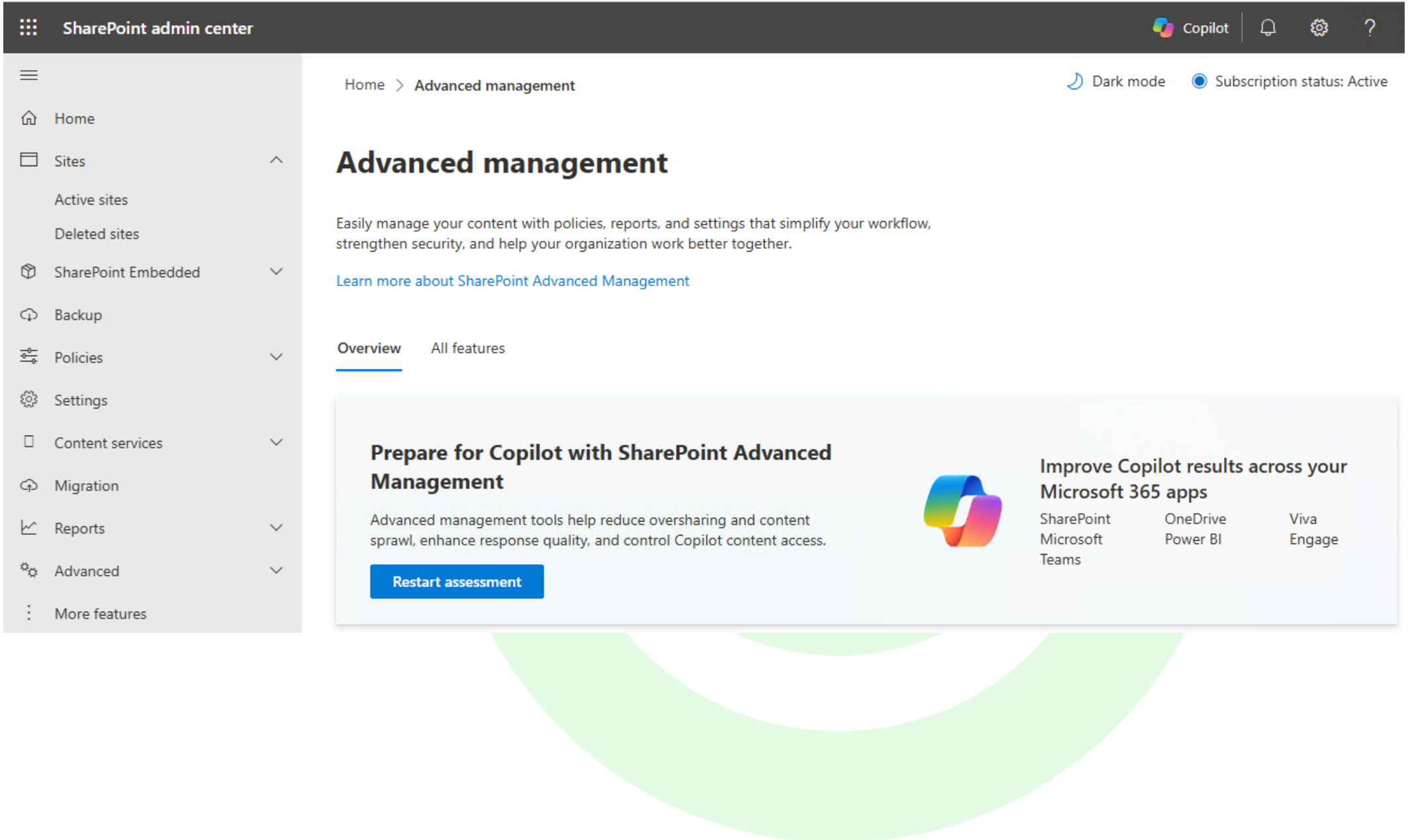
https://learn.microsoft.com/en-us/microsoft-365/copilot/secure-govern-copilot-foundational-deployment-guidance

Purview DSPM for AI: Data Risk Assessments

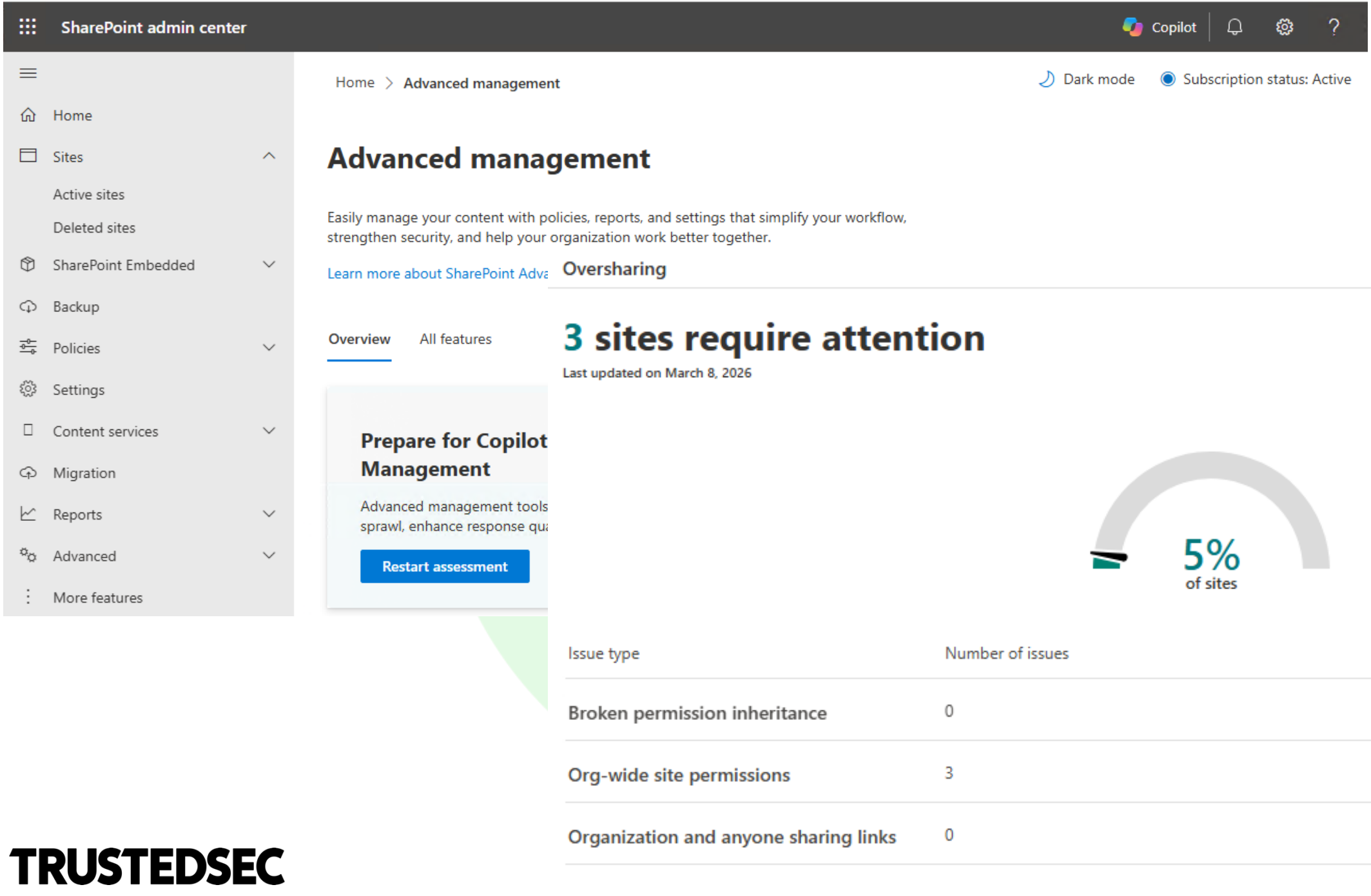
- **Default assessment:** runs automatically every week against the top 100 SharePoint sites by usage
- **Custom assessments:** point them at specific sites or users
- **What it reports per site:**
  - How many items hold sensitive content, by sensitive info type or label
  - Unlabeled files
  - Sharing links that are too broad: Anyone links and links that open to the whole organization
  - Which items Copilot or agents have actually accessed
- **Remediation options:**
  - Turn on Restricted Content Discovery or Restricted Access Control (SharePoint Advanced Management)
  - Start a site access review with the site owners
  - Create auto-labeling policies or retention and deletion policies
  - Newer builds also investigate individual items and can remove sharing links

SharePoint Advanced Management (SAM): Content Management Assessment




SharePoint admin center
Copilot
Home
Sites
Active sites
Deleted sites
SharePoint Embedded
Backup
Policies
Settings
Content services
Migration
Reports
Advanced
More features
Home > Advanced management
Dark mode
Subscription status: Active
Advanced management
Easily manage your content with policies, reports, and settings that simplify your workflow, strengthen security, and help your organization work better together.
Learn more about SharePoint Advanced Management
Overview
All features
Prepare for Copilot with SharePoint Advanced Management
Advanced management tools help reduce oversharing and content sprawl, enhance response quality, and control Copilot content access.
Restart assessment
Improve Copilot results across your Microsoft 365 apps
SharePoint
OneDrive
Viva
Microsoft Teams
Power BI
Engage

TRUSTEDSEC


SharePoint admin center
Copilot
Home
Sites
Active sites
Deleted sites
SharePoint Embedded
Backup
Policies
Settings
Content services
Migration
Reports
Advanced
More features
Home > Advanced management
Dark mode
Subscription status: Active
Advanced management
Easily manage your content with policies, reports, and settings that simplify your workflow, strengthen security, and help your organization work better together.
Learn more about SharePoint Adva
Overview
All features
Oversharing
3 sites require attention
Last updated on March 8, 2026
Prepare for Copilot Management
Advanced management tools
sprawl, enhance response qua
Restart assessment
5%
of sites
Issue type
Number of issues
Recommendations
Broken permission inheritance
0
View recommendations
Org-wide site permissions
3
View recommendations
Organization and anyone sharing links
0
View recommendations

TRUSTEDSEC

Implement Prompt Injection Detection & Mitigation



Use Insider Risk's "Risky AI usage" template, which detects prompt injection



Turn on Defender real-time protection for Copilot Studio agents



Avoid email- or event-triggered autonomous agents that also have write actions and web grounding



Cowork browser tasks run in Edge with the user's sign-ins - prompt injection within the browser session is a risk

ToDos before Deploying Copilot

- AI acceptable use policy for AI cowork and agents
- Identify allowed agent types – Microsoft, your org, and/or external publishers
- Create a process for agent lifecycle, especially when a sponsor leaves the organization
- Decide on if/how frontier models will be leveraged (both Microsoft's "Frontier" and OpenAI/Anthropic)
- Discover and inventory agents using Agent 365
- Configure "Allow web search in Copilot" in the Cloud Policy service
- Apply Intune App Protection policies to the Copilot app
- Keep Cowork browser use and Autopilot limited to a pilot group
- Restrict user ability to create AI agent blueprints & inventory existing ones
- Limit membership in the AI Administrator & Agent ID Administrator roles & manage with PIM
- Set Copilot Credit spending policies
- Restrict user consent for agents, MCP, and federated connectors (which run on OAuth)

ToDos before Deploying Copilot (continued)

- Use Conditional Access to control the Enterprise Copilot Platform (lookup the AppID in your tenant) – doesn't govern Autopilot, browser, or MCP actions – they need endpoint controls
- Set Conditional Access for agent identities and ID Protection for agents (requires Agent 365 + Entra P1 or M365 E3, or M365 E7) – Review at least annually
- Perform Purview DSPM for AI Data Risk Assessments
- Perform SharePoint Advanced Management Content Management Assessment
- Fix "everyone" permission grants throughout
- Leverage Purview to protect data with labels
- Configure Defender for Cloud Apps: mark unapproved AI apps as unsanctioned
- Leverage Endpoint DLP and Edge for Business inline DLP for monitoring paste and upload unapproved AI sites such as Gemini, etc.
- DSPM for AI collection policies for third-party AI sites
- Evaluate Tenant Restrictions v2 to block personal Microsoft accounts and other tenants (proxy or Global Secure Access for network coverage; Windows policy for managed devices)



TrustedSec Proactive Security Assessments

- Active Directory Security Assessment (ADSA)
- AWS Infrastructure
- Azure Infrastructure
- Backup Infrastructure
- Certificate Services
- Endpoints & Intune
- Microsoft 365 Security Assessment (M365)
- Professional Services

https://trustedsec.com/services/harden



TrustedSec Proactive Security Assessments

- Active Directory Security Assessment (ADSA)
- **Anthropic Claude Readiness Assessment**
- AWS Infrastructure
- Azure Infrastructure
- Backup Infrastructure
- Certificate Services
- Endpoints & Intune
- Microsoft 365 Security Assessment (M365)
- **Microsoft Copilot Readiness Assessment**
- Professional Services



Conclusion

Life comes at you fast, AI is faster

There are key decisions that need to be made before deploying AI

Get your house in order before inviting in the AI party crashers!



TRUSTEDSEC

Questions?





TRUSTEDSEC

FOLLOW US @TRUSTEDSEC

The information security industry is constantly evolving – keep up by following TrustedSec's active social media, blogs, podcasts and webinars.



TRUSTEDSEC

THANK YOU!

