

Four Ways Your Company Data Meets AI

Work out which one you are dealing with. They are four different problems.

1. Copilot

Your own files, in your own systems

FOR EXAMPLE,

Someone asks Copilot for last quarter's numbers. It finds the board pack that was shared with 'anyone in the company' three years ago.

WHAT TO DO

Fix the sharing links before you turn Copilot on.

REAL GAP

No good way to label a decade of old files automatically.

2. Paid AI accounts

Claude or ChatGPT, bought by the company

FOR EXAMPLE,

An analyst pastes a client contract in for a summary. The data left the building — legitimately, under a contract you signed.

WHAT TO DO

Check the admin console says training is off. Screenshot it next to the contract.

REAL GAP

You cannot prove the vendor honours that setting, or that deleted data is really gone.

3. Personal accounts

Free logins, used for work

FOR EXAMPLE,

A manager uses their own ChatGPT login on a work laptop to rewrite a proposal. No contract, no record, no way to delete it.

WHAT TO DO

Give people a paid account first. Then block the free sites.

REAL GAP

Personal phones stay invisible. That risk is cultural, not technical.

4. Data you sell

Your data, inside someone else's AI

FOR EXAMPLE,

A customer loads the data feed you license to them into their own model. You cannot charge for that, and you cannot get it back.

WHAT TO DO

Ban training in the licence. Deliver by API, not bulk files.

REAL GAP

No dependable watermark for tabular data. Markers can be found and stripped.

What to Set Up, and Which Tools Do It

The same four situations. What to change, and the kind of product that changes it.

1 Copilot

WHAT TO CHANGE

1. Stop company-wide sharing links being the default.
2. Turn on Restricted SharePoint Search until the clean-up is done.
3. Put sensitivity labels on HR, finance, legal and deal sites.

TOOLS THAT DO IT, AND WHAT THEY DO

SharePoint Advanced Management — lists every over-shared file and lets you fence off search while you fix them
Microsoft Purview — labels whole sites, so you are not tagging files one at a time
Entra Access Reviews — makes site owners re-approve access, so clean-ups do not quietly undo themselves

2 Paid AI accounts

WHAT TO CHANGE

1. Business plan only.
2. Confirm in the admin console that training is off.
3. Choose the country your data sits in before anyone starts.
4. Approve connectors one at a time.

TOOLS THAT DO IT, AND WHAT THEY DO

Claude / OpenAI enterprise admin — shows the training, retention and region settings in one place
Azure OpenAI — runs in your own subscription and does not train on your data by default
Entra consent policies — stops staff approving third-party apps on their own

3 Personal accounts

WHAT TO CHANGE

1. Provide the paid version first and tell everyone.
2. Then block free AI sites at the web filter.
3. Turn off users' ability to approve apps.
4. Watch copy-paste on work laptops.

TOOLS THAT DO IT, AND WHAT THEY DO

Defender for Cloud Apps / Netskope / Zscaler — show which AI sites staff visit, and block the ones you choose
Purview Endpoint DLP — spots company text being pasted into sites you do not manage
Entra ID — shows who signed in where and what they connected to

4 Data you sell

WHAT TO CHANGE

1. Say in the licence they may not train on it, and make that carry to their AI suppliers.
2. Deliver through an interface, not bulk files.
3. Seed marker records.
4. Offer a clean room.

TOOLS THAT DO IT, AND WHAT THEY DO

API gateway (Kong, Apigee, AWS) — enforces per-customer limits and records exactly what was taken
AWS / Snowflake Clean Rooms — let customers ask questions without ever holding your records
Confidential computing — runs the calculation sealed, so even the host cannot read the inputs

Selling Data Without It Ending Up Inside Someone's AI

For data brokers and licensors — the contract matters, but on its own it is not enough

1 Used as training or fine-tuning corpus

Annual customer attestation on AI use, plus audit rights you are willing to exercise.

WHAT TO CHANGE

Explicit prohibition on training, fine-tuning and derived-model development. Define 'derived data' and state who owns weights influenced by your data. Flow the obligation down to the customer's AI vendors.

TOOLS THAT HELP

Licence template · DPA annex · licence register

REAL GAP Audit rights over a customer's training are near-impossible to exercise in practice.

2 Bulk extraction to build a corpus

Volume and velocity anomaly detection against a per-customer baseline set at onboarding.

WHAT TO CHANGE

API delivery with per-record entitlement checks rather than bulk file drop. Per-key quotas and rate limits. Short-lived signed URLs. Bulk export off by default.

TOOLS THAT HELP

Kong · Apigee · AWS API Gateway · WAF

REAL GAP A determined customer can still scrape slowly enough to look normal.

3 No way to prove your data reached a model

Probe major public and customer-hosted models on a schedule. Retain the probe log as the evidence pack.

WHAT TO CHANGE

Seed canary records — plausible but fictitious, unique per customer. Document the scheme and have counsel review it before deployment.

TOOLS THAT HELP

Canary generation · scheduled model probing · provenance watermarking

REAL GAP No dependable watermark exists for tabular data. Markers can be found and stripped.

4 Customer needs the answer, not the data

Review the query log for patterns that reconstruct rows from aggregates.

WHAT TO CHANGE

Clean room or confidential computing enclave. Query-only access, aggregate outputs with minimum cell sizes, raw export disabled at platform level.

TOOLS THAT HELP

AWS Clean Rooms · Snowflake Data Clean Rooms · Nitro Enclaves · Intel SGX

REAL GAP Clean rooms only work where the customer will accept not holding the data.

5 Receiving side: you licensed it, then indexed it

Quarterly index provenance audit — can you name the licence behind every corpus?

WHAT TO CHANGE

Licensed data in its own scoped index with its own retention. Provenance register mapping every licensed source to the AI systems that touch it. Verify the no-training setting before ingesting.

TOOLS THAT HELP

Vector DB namespaces · data catalogue · licence register

REAL GAP Nothing stops an engineer loading licensed data into an index outside the register.

A Four-Week Runbook

In order, with a test at the end that gives an unambiguous answer

WEEK 1

Measure before you enable

- Oversharing report across SharePoint and OneDrive
- 90 days of proxy or CASB data for consumer AI domains
- List every AI tool in use, including ones bundled into existing licences

*A number you can show leadership.
Expect thousands of over-shared files.*

WEEK 2

Remediate the crown jewels

- Remove org-wide and anonymous links on HR, finance, legal and M&A sites
- Apply sensitivity labels to those sites, not to individual files
- Name a data owner per site — remediation without an owner does not stick

The files that would end a career are no longer reachable by search.

WEEK 3

Fence the rest, give people a path

- Restricted SharePoint Search scoped to remediated sites
- Provision the sanctioned tier and tell everyone it exists
- Verify the vendor setting matches the contract clause; screenshot both

A sanctioned route exists, so blocking the unsanctioned one becomes reasonable.

WEEK 4

Try to break it

- As a pilot user with ordinary permissions, ask for 'salary bands' and 'termination letters'
- If either returns, stop and go back to week 2
- Only then block consumer AI domains at the proxy

A pass or fail a non-technical executive can understand.

THEN KEEP IT CLOSED

Monthly delta

Re-run the oversharing report. Track the trend, not the absolute number.

Alerts to the owner

Route DLP alerts to the named data owner. IT cannot judge whether a file is sensitive.

Annual re-verification

Re-check every vendor setting against the contract, and after any migration.

One written rule

What may never leave: regulated data, privileged material, source code, unreleased financials.

Records coverage

Retention, legal hold and eDiscovery extended to prompts and responses.

Do not block first. Blocking consumer AI before a sanctioned tool exists moves the traffic to phones and home laptops, where you cannot see it at all. The risk does not go down — it goes dark. Sanction, then fence, then block.