AI inputs: how do you limit prompt injection and keep sensitive data out?

Before real data goes into AI, give staff one approved tool with clear rules on what may go in, and make every AI app retrieve only what the person asking can already see. Judge every other input by who could have written it and what the app could leak or do if it carried instructions.

Reviewed	Next review	Version	Written for	Latest version
2 Nov 2026	**2 Dec 2026**	**1.0**	**Teams that build or run LLM apps and teams that govern staff use of AI**	**sheets.einblick.dev/ai-input-security/**

If you only do five things

1. **Give staff an approved AI tool and clear rules on what goes in.**
   The Dutch data protection authority treats personal data entered into a chatbot against company rules as a data breach[1]. Choose a business plan under a processor contract[2] that keeps your data out of model training, and name the data that must never go in.
2. **Fix oversharing before an assistant indexes your files.**
   Microsoft 365 Copilot can surface any content a user has at least view access to[3], so old files shared with the whole organization can show up in any employee's answers. OWASP names oversharing through unscoped drives and legacy permissions as one of two failures behind most disclosure incidents[4].
3. **Enforce access in code, never in the prompt.**
   Check the user's permissions before content reaches the model, and give a customer-facing bot no more access than the customer has. The NCSC says prompt injection may never be fully mitigated[5], so assume users can get the model to reveal anything it can reach.
4. **Sanitize untrusted content and mark it as data.**
   Strip invisible Unicode characters, hidden text and spoofed role tags from uploads and retrieved documents, then wrap what remains in delimiters that mark it as data[6,4]. This makes attacks harder but does not stop them all[6].
5. **Block the links and images that carry data out.**
   Injected instructions can make a chatbot output an image URL that sends chat data to an attacker's server when the image loads[6]. Render links and images only from allowlisted domains[6], and log full inputs and outputs so you can spot suspicious activity[5].

Before you connect a new data source to an AI app

1. **Who can add or change content in it?**
   Anyone who can write to an indexed source can plant instructions or false answers. In research OWASP cites, five poisoned documents in a knowledge base of millions steered answers about 90% of the time[4].
2. **Does the index enforce each user's access?**
   OWASP says to enforce permissions per document and chunk inside the index query, because filtering the answer afterward can't undo what the model has already read[4]. A connector that indexes everything through one service account, without per-user permissions, fails this test.
3. **Does it hold more than the use case needs?**
   Index specific folders instead of the whole tenant, and leave out personal data the task doesn't use. OWASP lists sending external providers only the fields a task needs as a foundational control[4].
4. **Can you remove content from it again?**
   OWASP says to delete a document's embeddings within a set time after the source is deleted, and to verify this with audits[4]. Test removal before launch, including backups and relabeled files.

What changed in 2026

- **OWASP updated its LLM Top 10** on 3 August 2026. Prompt injection remains LLM01 and now covers instructions hidden in images and audio. System prompt leakage became Hidden Context Exposure, which also covers tool schemas and retrieved policy text[4].
- **Germany's BSI published a baseline against indirect prompt injection** in September 2026. The German-language guide, for LLM chats that work with uploaded documents, lists five measures drawn from red-teaming: sanitize the input, separate trust zones, harden the system prompt, choose a resistant model with reasoning and filter the output[6].
- **Microsoft added AI controls to Purview DLP** in 2026. A preview condition from June keeps external email out of Copilot's grounding to reduce prompt injection; since September, network-level DLP that covers generative AI platforms is generally available[7].

What you can skip

- **Keeping the system prompt secret.** OWASP says to assume hidden context is discoverable[4]. Write the prompt as if users will read it, and don't spend time on extraction-proof wording.
- **Products that promise to stop prompt injection.** The NCSC tells buyers to be wary of that claim[5]. In a study OWASP cites, adaptive attacks beat most of 12 published defenses more than 90% of the time[4]. Judge a product by what it limits once an injection succeeds.

Mistakes that cost the most

- **Secrets in the system prompt.** Credentials, connection strings and tokens placed there are exposed to anyone who extracts the prompt[4]. Keep them in a secrets store on the server, where the model can't reach them.

- **Classifying all your data before the first pilot.** Start with the sites and mailboxes the assistant will index and the few data types on your forbidden list.
- **A long AI policy.** Start with a one-page list of approved tools and forbidden data.

- **Grounding on external email.** In a 2025 demonstration OWASP describes, one crafted email made Microsoft 365 Copilot leak organizational data with no user interaction, bypassing its injection classifier and link filter[4]. Keep external email out of grounding where your platform allows it.
- **Checking typed prompts but not uploads.** Purview DLP for Microsoft 365 Copilot checks the text typed into a prompt, not files uploaded with it[8]. Test which inputs your own DLP inspects before you rely on it.
- **Treating embeddings as anonymous.** Embedding inversion can recover much of the source text, so OWASP says to treat a leaked vector database or backup as a leak of the documents themselves[4].
- **Unprotected prompt logs.** Observability platforms such as Langfuse and LangSmith log full prompts, responses and retrieved chunks by default[4]. Give those logs the same access controls and retention as the source data.

What this looks like at your size

Small team, no security staff

- Write down the data that never goes into AI tools: customer and patient records, HR files, passwords and keys.
- Remove organization-wide sharing from files that don't need it before you turn on an assistant that searches your files.
- Give a public chatbot only what you would publish on your website.

Growing, with IT or security staff

- Keep an inventory of AI apps: data sources, who can write to each source and who can query the app.
- Use browser or endpoint DLP to warn when sensitive data is pasted or uploaded into unapproved AI sites.
- Review permissions on the sites and mailboxes an assistant will index, largest and oldest first.
- Test each app with poisoned documents and hidden-text uploads before launch and after each new source.

Enterprise or regulated

- In Microsoft 365, use Purview DLP to keep labeled files out of Copilot and, in preview, to block prompts that contain sensitive information types[8].
- Keep external content in a separate index from confidential data[4].
- Monitor retrieval logs for unusual query volumes and cross-tenant hits, which can signal probing[4].
- Red-team against the BSI baseline[6], and share your defense design with testers so they can run adaptive attacks[4].

GDPR duties when AI tools process personal data

Before an approved tool goes live

An AI provider that processes personal data on your behalf is your processor, and processing likely to be high risk needs an impact assessment first[2].

Before use
Processor contract covering the data, purpose, duration and the provider's obligations

If high risk
Data protection impact assessment (DPIA) before processing starts

After personal data reaches an unapproved tool

This is often a reportable breach[1]. The clock runs from when you become aware of it[2]. The Council backed a longer deadline on 21 October 2026, but it is not yet law[9].

72 hours
Notify the supervisory authority, unless a risk to people is unlikely

If high risk
Tell the people affected without undue delay

Always
Document the facts, effects and remedial action

Version history

2 Nov 2026, version 1.0	Published after a copy edit and a source-by-source fact check, which narrowed 11 statements to what OWASP, the BSI and the NCSC actually say.

Sources

1. Caution: use of AI chatbot may lead to data breaches. Autoriteit Persoonsgegevens (Dutch DPA), 6 Aug 2024. autoriteitpersoonsgegevens.nl
2. Regulation (EU) 2016/679 (GDPR), Articles 4, 28 and 33 to 35. EUR-Lex, 2016. eur-lex.europa.eu
3. Data, Privacy, and Security for Microsoft Copilot. Microsoft Learn, 9 Jul 2026. learn.microsoft.com
4. OWASP GenAI LLM Top 10 2026 (LLM01, LLM02, LLM08 and LLM09). OWASP Gen AI Security Project, 3 Aug 2026. genai.owasp.org
5. Prompt injection is not SQL injection (it may be worse). NCSC (UK), 8 Dec 2025. ncsc.gov.uk
6. Basisschutz gegen Indirect Prompt Injection in dokumentbasierten LLM-Chats. BSI (Germany), 24 Sep 2026. bsi.bund.de
7. What's new in Microsoft Purview (June and September 2026). Microsoft Learn, 16 Oct 2026. learn.microsoft.com
8. Microsoft Purview DLP for Microsoft 365 Copilot and Cowork. Microsoft Learn, 16 Oct 2026. learn.microsoft.com
9. Digital rulebook: Council agrees position to simplify data protection, privacy and cyber reporting rules. Council of the EU, 21 Oct 2026. consilium.europa.eu