

CHEAT SHEET · TOMORIS TOOLBOX

Using Generative AI Safely: Cheat Sheet for Employees

What may go into a prompt, how to check results and recognize prompt injection, what the AI Act requires, and when to report

UPDATED

Oct 25, 2026

LENGTH

PDF, 7 pages

ONLINE

tomoris.com/en/downloads/generative-ki-sicher-nutzen-spickzettel

AT A GLANCE

4

Data classification levels, public to off-limits

LLM01

Prompt injection, number 1 in the OWASP Top 10 for LLM 2026

Art. 50

AI Act: disclose deepfakes, applies since Aug 2, 2026

10

Questions before sending and before using

WHO IT IS FOR

- All employees who use chat assistants, copilots, or writing, translation, or image tools
- Employees who use AI assistants with access to their mailbox, files, or browser
- Managers who share the cheat sheet, post it, or use it in training
- IT, data protection, and information security teams who define approvals and reporting channels

WHAT THIS COVERS

AI assistants save time, but they also bring risks: confidential data in the wrong tool, convincingly worded errors, and hidden instructions in third-party content. This cheat sheet shows what may go into a prompt, how to check results and recognize prompt injection, what the AI Act requires, and when to report.

An overview, not legal advice; the legal text in its current version is authoritative, in particular the AI Act including Regulation (EU) 2026/1744 and the GDPR. The traffic light and checklists are TOMORIS recommendations; your company's rules take precedence. Names and functions of the tools may change.

CONTENTS

01	What May Go into a Prompt: the Data Classification Traffic Light	2
02	Approved Tools Only: Avoid Shadow AI	3
03	Check Results: Facts, Sources, Figures, Code	3
04	Recognize Prompt Injection	4
05	Labeling and AI Literacy under the AI Act	5
06	Report an Incident: What, When, How	5
07	Checklists to Post	6
08	Sources	7

01 —

What May Go into a Prompt: the Data Classification Traffic Light

The more confidential the data, the narrower the range of tools. Never credentials.

Data classification traffic light (TOMORIS recommendation, adapt to your company's data classes)

LEVEL	EXAMPLES	WHERE IT MAY GO
Green: public	Published texts, general technical questions without reference to individuals or customers	All approved tools
Yellow: internal	Processes, templates, drafts without reference to individuals or customers	Approved tools with a work account
Red: confidential	Contracts, cost calculations, source code, trade secrets (§ 26b para. 1 UWG, Austrian Unfair Competition Act), customer and personnel data (Art. 4(1) GDPR)	Only tools explicitly approved for this purpose, otherwise placeholders

Off-limits

Passwords, PINs, one-time codes, API keys; health data and other special categories (Art. 9(1) GDPR)

Never credentials; special categories only if an exception under Art. 9(2) GDPR applies and data protection has approved the tool

- Enter only what the task needs: personal data must be limited to what is necessary (Art. 5(1)(c) GDPR); the BSI advises handling sensitive data sparingly (M18).
- Use placeholders such as [Customer A] instead of names. If the link can be restored with additional information, the data remains personal data (Recital 26 GDPR).
- Uploads count as input (BSI R2); the model also reads metadata, annotations, and comments (BSI, baseline protection), text in screenshots, and text under redaction bars in PDFs (OWASP LLM02:2026).

02

Approved Tools Only: Avoid Shadow AI

Private accounts, free web services, and extensions without approval are shadow AI.

The BSI recommends defining in binding terms which AI applications may be used for what, what may be entered, and how results may be reused, and blocking or enabling services selectively (M16). This is because operators of internet services can access inputs and outputs and, depending on the terms of use, use them for training (BSI R2).

- Work account only: private accounts and consumer versions are off-limits for company data unless your company rules otherwise.
- Extensions, plug-ins, and agents only after approval: if third parties offer add-on functions such as managing email, data may leak to them (BSI R2).
- Image generators: some publish images together with the prompt and user name by default (BSI R2).
- Microsoft Copilot (formerly Microsoft 365 Copilot, with a work account) only shows data for which you have at least view permission (Microsoft Learn). If you see something that is not meant for you, report it.

03

Check Results: Facts, Sources, Figures, Code

AI answers sound certain even when they are wrong: check them and ask better questions.

Generative AI produces statistically likely outputs that can be worded confidently and still be wrong; NIST calls this confabulation, colloquially known as hallucination (NIST AI 600-1, 2.2). Because the texts seem coherent, they are easily adopted without checking (automation bias, BSI R8; NIST 2.7).

Check before using (BSI, NIST, OWASP)

RESULT	TYPICAL ERROR	HOW TO CHECK
Facts, current events	Invented content; without real-time data, the model does not know about new events (BSI R4)	Compare with a reliable source (BSI M20)

Sources, quotes	Completely invented references and justifications (BSI R4, NIST 2.2)	Open every source: does it exist, and does it say that?
Figures, calculations	Incorrect or internally contradictory information (NIST 2.2)	Recalculate and compare with the original
Summaries	Limitations, exceptions, time references, or risks are missing (OWASP LLM07:2026)	Read the key passages in the original
Code, scripts	Security vulnerabilities, outdated libraries (BSI R6), invented package names (OWASP LLM07:2026)	Review and test, never run without checking (BSI M20, OWASP LLM10:2026)

- Ask better questions: allow “I don't know,” have the model use only the documents provided, and require a quote and source for each statement; differing answers to the same question can indicate hallucinations (Anthropic).
- Even this does not eliminate hallucinations completely: always check critical information (Anthropic), and check outputs before passing them on to others (Microsoft).

04 —

Recognize Prompt Injection

Hidden instructions in third-party content: warning signs and rules for AI agents.

In an indirect prompt injection, instructions to the AI are hidden in content that the model processes, such as a web page, a document, an email, or an image; you neither wrote nor saw them (OWASP LLM01:2026, NIST 2.9). Typical hiding places are white text on a white background, invisible Unicode characters, metadata, and comments. According to the BSI, there is no reliable countermeasure that eliminates the problem completely (BSI, baseline protection 2026).

Warning signs (BSI, OWASP, Anthropic)

WARNING SIGN	EXAMPLE
A request that does not fit the task	A web page summary asks for your email address (BSI R28, example 7).
Unexpected links or images	An embedded image loads from an external server and exfiltrates data in the process, without a click (BSI, baseline protection 2.5).
An action you did not request	The assistant wants to send or delete email, or change or delete documents (BSI R3, OWASP LLM03:2026).
A conspicuously one-sided recommendation	A hidden advertising instruction in a data sheet makes the model recommend a paid app (BSI R28, example 6).
A notice about instructions in the content	The assistant reports them instead of following them, as Anthropic recommends to developers. Take it seriously and report it.

- Few permissions: limit the access and execution rights of AI applications to the minimum (BSI M17). If you have email summarized, the assistant does not need permission to send (OWASP LLM03:2026).

- Confirm yourself: privileged, irreversible, or externally visible actions only after human confirmation; check the action itself, not just its summary (OWASP LLM01:2026, measure 7; BSI M17).
- Dangerous combination of three: access to private data, reading third-party content, and external communication or changes. In that case, OWASP recommends having a human approve every single action (LLM01:2026, measure 8).

05 —

Labeling and AI Literacy under the AI Act

Art. 50 and Art. 4 of the AI Act as amended by Regulation (EU) 2026/1744.

- Disclose deepfakes (Art. 50(4), first subparagraph): deployers who use AI to generate or manipulate image, audio, or video content that resembles real persons, places, or events and would falsely appear authentic (Art. 3(60)) must disclose this; for evidently artistic or satirical works, in a limited form. Applies since August 2, 2026 (Art. 113).
- Disclose AI texts (Art. 50(4), second subparagraph): published AI texts that inform the public on matters of public interest must be disclosed, unless they have undergone human review or editorial control and someone holds editorial responsibility.
- Provider obligations (Art. 50(1) and (2)): AI notice for chatbots, machine-readable marking of synthetic content; for systems placed on the market before August 2, 2026, Art. 50(2) must be met by December 2, 2026 (Art. 111(4)).
- Promote AI literacy (Art. 4): providers and deployers, which includes your employer if it uses AI under its own authority (Art. 3(4)), take measures to support the development of their staff's AI literacy; a specific level does not have to be guaranteed. Amended since July 27, 2026; previously, the text required ensuring "to their best extent" a "sufficient level" (Regulation (EU) 2026/1744).

What this means for you (TOMORIS recommendation): always label deepfakes, route publications through the approval process, and attend training. According to the European Commission, a certificate is not required; companies can document training internally (AI Literacy Q&A).

06 —

Report an Incident: What, When, How

What to do immediately, depending on the situation. Reporting is not an admission of guilt.

Act immediately (TOMORIS recommendation; your company's rules take precedence)

WHAT HAPPENED	WHAT YOU DO IMMEDIATELY
Suspected prompt injection	Don't confirm anything, don't open any link, don't enter anything; take a screenshot and report it to IT along with the source.
Unwanted action by the assistant	Stop the assistant and report what was sent, changed, shared, or deleted.

Confidential or personal data in the wrong tool

Note the tool, account, data, and time, and inform IT and data protection immediately. This may be a personal data breach (Art. 4(12) GDPR), which the controller reports to the supervisory authority without undue delay and, where feasible, within 72 hours of becoming aware of it, unless it is unlikely to result in a risk (Art. 33(1) GDPR).

Incorrect result already used

Inform the responsible person so the error can be corrected.

An answer shows data that is not meant for you

Don't pass it on, don't save it, and report it so permissions can be corrected.

Template: our contacts (fill in before posting)

REASON	CONTACT PERSON	PHONE OR EMAIL
Wrong data entered, data protection	_____	_____
Prompt injection, unwanted action	_____	_____
Approval of new AI tools	_____	_____

07 —

Checklists to Post

Five questions each before sending and before using, as a TOMORIS recommendation.

Before sending a prompt, an upload, or a connection to your mailbox and files:

- ☐ Am I using an approved tool with my work account?
- ☐ Which traffic light level does the data have, is the tool approved for it, and are placeholders enough?
- ☐ Does it contain credentials, comments, metadata, or redacted passages that must not leave the company?
- ☐ Does the content come from an unknown source? Then watch out for hidden instructions.
- ☐ Does the assistant really need access to my mailbox, files, or browser for this?

Before using a result, confirming an action, or publishing:

- ☐ Have I checked facts, figures, and sources: does every source exist, and does it really say that?
- ☐ Is anything important missing from the summary, such as exceptions, time references, or risks?
- ☐ Has the code been reviewed and tested, and do the suggested packages really exist?
- ☐ Do requests, links, and actions fit the task, and have I read the recipient, content, and destination?

☐ Does the result need to be labeled or approved before publication?

For managers: define tools, data classes, and the reporting channel (AI usage policy template), document training, and clean up permissions, for example in SharePoint (Microsoft Learn).

NEXT STEP

Check It Yourself or Put It into Practice Together

DISCUSS IMPLEMENTATION

Get in Touch

tomoris.com/en/contact

contact@tomoris.com

08

Sources

All statements are supported by sources. Legislation applies in its current version.

- AI Act, Regulation (EU) 2024/1689, consolidated July 27, 2026**
Art. 3(4) and (60); Art. 4, 50, 111(4), 113
eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX:02024R1689-20...
- Regulation (EU) 2026/1744, Digital Omnibus on AI**
Art. 1(5) (new Art. 4), Art. 4 (entry into force)
eur-lex.europa.eu/eli/reg/2026/1744/oj/deu
- European Commission: AI Literacy – Questions & Answers**
As of October 20, 2026; question on documentation
digital-strategy.ec.europa.eu/en/faqs/ai-literacy-questions-ans...
- Regulation (EU) 2016/679 (GDPR), EUR-Lex**
Recital 26; Art. 4(1) and (12), Art. 5, 9, 33
eur-lex.europa.eu/eli/reg/2016/679/oj/deu
- UWG (Austrian Unfair Competition Act), § 26b (trade secret), RIS**
§ 26b para. 1, version BGBl. I No. 109/2018
www.ris.bka.gv.at/NormDokument.wxe?Abfrage=Bundesnormen&Gesetze...
- OWASP Top 10 for LLM Applications 2026**
Numbers according to the 2026 PDF; individual pages on genai.owasp.org still use the 2025 numbers (LLM06, 09, 05)
genai.owasp.org/resource/owasp-genai-llm-top-10-2026
- NIST AI 600-1: Generative AI Profile (July 2024)**
Sections 2.2, 2.7, and 2.9
nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf
- BSI: Generative AI Models – Opportunities and Risks for Industry and Authorities, version 2.0**
Version 2.0 of January 17, 2025; R2 to R4, R6, R8, R28; M16 to M18, M20
www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/KI/Generative_KI-Mo...
- BSI: Baseline Protection against Indirect Prompt Injections in Document-Based LLM Chats, version 1.0**
Version 1.0 of September 14, 2026; introduction, section 2.5
www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/KI/Basisschutz_Indi...
- Anthropic: Reduce hallucinations**
Basic strategies, best-of-N verification
platform.claude.com/docs/de/test-and-evaluate/strengthen-guardr...
- Anthropic: Mitigate jailbreaks and prompt injections**
Section on indirect prompt injection
platform.claude.com/docs/de/test-and-evaluate/strengthen-guardr...
- Microsoft Learn: Data, privacy, and security for Microsoft Copilot**
Organizational data; content that Copilot creates
learn.microsoft.com/de-de/microsoft-365/copilot/microsoft-365-c...