

SPICKZETTEL · TOMORIS TOOLBOX

Generative KI sicher nutzen: Spickzettel für Mitarbeitende

Was in einen Prompt darf, wie Sie Ergebnisse prüfen und Prompt Injection erkennen, was die KI-Verordnung verlangt und wann Sie melden

STAND

25.10.2026

UMFANG

PDF, 7 Seiten

ONLINE

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

AUF EINEN BLICK

4

Stufen in der
Datenklassen-Ampel,
von öffentlich bis tabu

LLM01

Prompt Injection, Platz 1
der OWASP Top 10 für
LLM 2026

Art. 50

KI-Verordnung:
Deepfakes offenlegen,
gilt seit 2.8.2026

10

Fragen vor dem
Absenden und vor dem
Verwenden

FÜR WEN

- Alle Mitarbeitenden, die Chat-Assistenten, Copilots, Schreib-, Übersetzungs- oder Bildwerkzeuge nutzen
- Mitarbeitende, die KI-Assistenten mit Zugriff auf Postfach, Dateien oder Browser einsetzen
- Führungskräfte, die den Spickzettel weitergeben, aushängen oder in Unterweisungen nutzen
- IT, Datenschutz und Informationssicherheit, die Freigaben und Meldewege festlegen

WORUM ES GEHT

KI-Assistenten sparen Zeit, bringen aber Risiken: vertrauliche Daten im falschen Werkzeug, überzeugend formulierte Fehler und versteckte Anweisungen in fremden Inhalten. Dieser Spickzettel zeigt, was in einen Prompt darf, wie Sie Ergebnisse prüfen und Prompt Injection erkennen, was die KI-Verordnung verlangt und wann Sie melden.

Überblick ohne Rechtsberatung; maßgeblich ist der Gesetzestext in der geltenden Fassung, insbesondere der KI-Verordnung samt Verordnung (EU) 2026/1744 und der DSGVO. Ampel und Checklisten sind Empfehlungen von TOMORIS; die Vorgaben Ihres Unternehmens gehen vor. Bezeichnungen und Funktionen der Werkzeuge können sich ändern.

INHALT

01	Was in einen Prompt darf: die Datenklassen-Ampel	2
02	Nur freigegebene Werkzeuge: Schatten-KI vermeiden	3
03	Ergebnisse prüfen: Fakten, Quellen, Zahlen, Code	3
04	Prompt Injection erkennen	4
05	Kennzeichnung und KI-Kompetenz nach der KI-Verordnung	5
06	Vorfall melden: was, wann, wie	5
07	Checklisten zum Aushängen	6
08	Quellen	7

01

Was in einen Prompt darf: die Datenklassen-Ampel

Je vertraulicher die Daten, desto enger der Kreis der Werkzeuge. Zugangsdaten nie.

Datenklassen-Ampel (Empfehlung von TOMORIS, an die Datenklassen Ihres Unternehmens anpassen)

STUFE	BEISPIELE	WOHIN DAMIT
Grün: öffentlich	Veröffentlichte Texte, allgemeine Fachfragen ohne Personen- oder Kundenbezug	Alle freigegebenen Werkzeuge
Gelb: intern	Abläufe, Vorlagen, Entwürfe ohne Personen- oder Kundenbezug	Freigegebene Werkzeuge mit Arbeitskonto
Rot: vertraulich	Verträge, Kalkulationen, Quellcode, Geschäftsgeheimnisse (§ 26b Abs. 1 UWG), Kunden- und Personaldaten (Art. 4 Nr. 1 DSGVO)	Nur ausdrücklich dafür freigegebene Werkzeuge, sonst Platzhalter
Tabu	Passwörter, PINs, Einmalcodes, API-Schlüssel; Gesundheitsdaten und andere besondere Kategorien (Art. 9 Abs. 1 DSGVO)	Zugangsdaten nie; besondere Kategorien nur, wenn eine Ausnahme nach Art. 9 Abs. 2 DSGVO greift und der Datenschutz das Werkzeug freigegeben hat

- Nur eingeben, was die Aufgabe braucht: Personenbezogene Daten sind auf das notwendige Maß zu beschränken (Art. 5 Abs. 1 lit. c DSGVO); das BSI rät zu sparsamem Umgang mit sensiblen Daten (M18).
- Platzhalter wie [Kunde A] statt Namen. Lässt sich der Bezug mit zusätzlichen Informationen wiederherstellen, bleiben die Daten personenbezogen (Erwägungsgrund 26 DSGVO).
- Uploads zählen wie Eingaben (BSI R2); das Modell liest auch Metadaten, Anmerkungen und Kommentare (BSI, Basisschutz), Text in Screenshots und unter Schwärzungsbalken im PDF (OWASP LLM02:2026).

02

Nur freigegebene Werkzeuge: Schatten-KI vermeiden

Private Konten, kostenlose Webdienste und Erweiterungen ohne Freigabe sind Schatten-KI.

Das BSI empfiehlt, verbindlich festzulegen, welche KI-Anwendungen wofür genutzt, was eingegeben und wie Ergebnisse weiterverwendet werden dürfen, und Dienste gezielt zu sperren oder freizuschalten (M16). Denn Betreiber von Internetdiensten können auf Ein- und Ausgaben zugreifen und sie je nach Nutzungsbedingungen zum Training verwenden (BSI R2).

- Nur mit dem Arbeitskonto: Private Konten und Verbraucherversionen sind für Unternehmensdaten tabu, solange Ihr Unternehmen nichts anderes regelt.
- Erweiterungen, Plug-ins und Agenten nur nach Freigabe: Bieten Drittanbieter Zusatzfunktionen wie das Verwalten von E-Mails an, ist ein Datenabfluss an sie denkbar (BSI R2).
- Bildgeneratoren: Manche veröffentlichen Bilder samt Eingabe und Benutzername standardmäßig (BSI R2).
- Microsoft Copilot (vormals Microsoft 365 Copilot, mit Arbeitskonto) zeigt nur Daten, für die Sie mindestens Anzeigeberechtigung haben (Microsoft Learn). Sehen Sie etwas, das nicht für Sie bestimmt ist, melden Sie es.

03

Ergebnisse prüfen: Fakten, Quellen, Zahlen, Code

KI-Antworten klingen sicher, auch wenn sie falsch sind: prüfen und besser fragen.

Generative KI erzeugt statistisch wahrscheinliche Ausgaben, die selbstsicher formuliert und trotzdem falsch sein können; NIST nennt das Konfabulation, umgangssprachlich Halluzination (NIST AI 600-1, 2.2). Weil die Texte schlüssig wirken, werden sie leicht ungeprüft übernommen (Automation Bias, BSI R8; NIST 2.7).

Vor dem Verwenden prüfen (BSI, NIST, OWASP)

ERGEBNIS	TYPISCHER FEHLER	SO PRÜFEN SIE
Fakten, Aktuelles	Erfundene Inhalte; neue Ereignisse kennt das Modell ohne Echtzeitdaten nicht (BSI R4)	Mit einer verlässlichen Quelle abgleichen (BSI M20)
Quellen, Zitate	Frei erfundene Referenzen und Begründungen (BSI R4, NIST 2.2)	Jede Quelle öffnen: Gibt es sie, und sagt sie das?

Zahlen, Rechnungen	Falsche oder in sich widersprüchliche Angaben (NIST 2.2)	Nachrechnen, mit dem Original vergleichen
Zusammenfassungen	Einschränkungen, Ausnahmen, Zeitangaben oder Risiken fehlen (OWASP LLM07:2026)	Entscheidende Stellen im Original lesen
Code, Skripte	Sicherheitslücken, veraltete Bibliotheken (BSI R6), erfundene Paketnamen (OWASP LLM07:2026)	Prüfen und testen, nie ungeprüft ausführen (BSI M20, OWASP LLM10:2026)

- Besser fragen: „Ich weiß es nicht“ erlauben, nur bereitgestellte Dokumente verwenden lassen, Zitat und Quelle je Aussage verlangen; abweichende Antworten auf dieselbe Frage können auf Halluzinationen hindeuten (Anthropic).
- Auch das beseitigt Halluzinationen nicht vollständig: Kritische Informationen immer prüfen (Anthropic), Ausgaben vor der Weitergabe an andere prüfen (Microsoft).

04 —

Prompt Injection erkennen

Versteckte Anweisungen in fremden Inhalten: Warnzeichen und Regeln für KI-Agenten.

Bei einer indirekten Prompt Injection stecken Anweisungen an die KI in Inhalten, die das Modell verarbeitet, etwa in einer Webseite, einem Dokument, einer E-Mail oder einem Bild; Sie haben sie weder geschrieben noch gesehen (OWASP LLM01:2026, NIST 2.9). Typische Verstecke sind weiße Schrift auf weißem Grund, unsichtbare Unicode-Zeichen, Metadaten und Kommentare. Eine zuverlässige Gegenmaßnahme, die das Problem vollständig beseitigt, gibt es laut BSI nicht (BSI, Basisschutz 2026).

Warnzeichen (BSI, OWASP, Anthropic)

WARNZEICHEN	BEISPIEL
Aufforderung, die nicht zur Aufgabe passt	Eine Webseiten-Zusammenfassung bittet um Ihre E-Mail-Adresse (BSI R28, Beispiel 7).
Unerwartete Links oder Bilder	Ein eingebettetes Bild lädt von einem fremden Server und leitet dabei Daten aus, ohne Klick (BSI, Basisschutz 2.5).
Aktion, die Sie nicht verlangt haben	Der Assistent will E-Mails senden oder löschen, Dokumente ändern oder löschen (BSI R3, OWASP LLM03:2026).
Auffällig einseitige Empfehlung	Eine versteckte Werbeanweisung in einem Datenblatt lässt das Modell eine kostenpflichtige App empfehlen (BSI R28, Beispiel 6).
Hinweis auf Anweisungen im Inhalt	Der Assistent meldet sie, statt sie zu befolgen, wie Anthropic es Entwicklern empfiehlt. Ernst nehmen und melden.

- Wenig Rechte: Zugriffs- und Ausführungsrechte von KI-Anwendungen auf das Minimum beschränken (BSI M17). Wer E-Mails zusammenfassen lässt, braucht kein Recht zum Senden (OWASP LLM03:2026).

- Selbst bestätigen: Privilegierte, unumkehrbare oder nach außen sichtbare Aktionen nur nach menschlicher Bestätigung, dabei die Aktion selbst prüfen, nicht nur ihre Zusammenfassung (OWASP LLM01:2026, Maßnahme 7; BSI M17).
- Gefährliche Dreierkombination: Zugriff auf private Daten, Lesen fremder Inhalte und Kommunikation nach außen oder Änderungen. Dann empfiehlt OWASP, jede einzelne Aktion von einem Menschen freigeben zu lassen (LLM01:2026, Maßnahme 8).

05

Kennzeichnung und KI-Kompetenz nach der KI-Verordnung

Art. 50 und Art. 4 der KI-Verordnung in der Fassung der VO (EU) 2026/1744.

- Deepfakes offenlegen (Art. 50 Abs. 4 UAbs. 1): Wer als Betreiber mit KI Bild-, Ton- oder Videoinhalte erzeugt oder manipuliert, die wirklichen Personen, Orten oder Ereignissen ähneln und fälschlich echt wirken (Art. 3 Nr. 60), muss das offenlegen, bei offensichtlich künstlerischen oder satirischen Werken in eingeschränkter Form. Gilt seit 2. August 2026 (Art. 113).
- KI-Texte offenlegen (Art. 50 Abs. 4 UAbs. 2): Veröffentlichte KI-Texte, die über Angelegenheiten von öffentlichem Interesse informieren, sind offenzulegen, außer sie wurden menschlich überprüft oder redaktionell kontrolliert und jemand trägt die redaktionelle Verantwortung.
- Pflichten der Anbieter (Art. 50 Abs. 1 und 2): KI-Hinweis bei Chatbots, maschinenlesbare Kennzeichnung synthetischer Inhalte; für Systeme, die vor dem 2. August 2026 in Verkehr gebracht wurden, ist Abs. 2 bis 2. Dezember 2026 zu erfüllen (Art. 111 Abs. 4).
- KI-Kompetenz fördern (Art. 4): Anbieter und Betreiber, also auch Ihr Arbeitgeber, wenn er KI in eigener Verantwortung einsetzt (Art. 3 Nr. 4), ergreifen Maßnahmen, um die Entwicklung der KI-Kompetenz ihres Personals zu unterstützen; ein Niveau ist nicht zu garantieren. Neu gefasst seit 27. Juli 2026; davor hieß es, „nach besten Kräften“ ein „ausreichendes Maß“ sicherzustellen (VO (EU) 2026/1744).

Für Sie heißt das (Empfehlung von TOMORIS): Deepfakes immer kennzeichnen, Veröffentlichungen über den Freigabeprozess, Schulungen wahrnehmen. Ein Zertifikat ist laut EU-Kommission nicht nötig; Unternehmen können Schulungen intern dokumentieren (AI Literacy Q&A).

06

Vorfall melden: was, wann, wie

Was Sie je nach Lage sofort tun. Melden ist kein Schuldeingeständnis.

Sofort handeln (Empfehlung von TOMORIS; die Vorgaben Ihres Unternehmens gehen vor)

WAS PASSIERT IST	WAS SIE SOFORT TUN
Verdacht auf Prompt Injection	Nichts bestätigen, keinen Link öffnen, nichts eingeben; Screenshot machen und mit der Quelle der IT melden.

Ungewollte Aktion des Assistenten	Assistent stoppen und melden, was gesendet, geändert, geteilt oder gelöscht wurde.
Vertrauliche oder personenbezogene Daten im falschen Werkzeug	Werkzeug, Konto, Daten und Zeitpunkt notieren, sofort IT und Datenschutz informieren. Das kann eine Datenschutzverletzung sein (Art. 4 Nr. 12 DSGVO), die der Verantwortliche unverzüglich und möglichst binnen 72 Stunden nach Bekanntwerden der Aufsichtsbehörde meldet, außer sie führt voraussichtlich zu keinem Risiko (Art. 33 Abs. 1 DSGVO).
Fehlerhaftes Ergebnis schon verwendet	Die verantwortliche Person informieren, damit der Fehler korrigiert wird.
Antwort zeigt Daten, die nicht für Sie bestimmt sind	Nicht weitergeben, nicht speichern, melden, damit Berechtigungen korrigiert werden.

Vorlage: unsere Ansprechpersonen (vor dem Aushängen ausfüllen)

ANLASS	ANSPRECHPERSON	TELEFON ODER E-MAIL
Falsche Daten eingegeben, Datenschutz	_____	_____
Prompt Injection, ungewollte Aktion	_____	_____
Freigabe neuer KI-Werkzeuge	_____	_____

07 —

Checklisten zum Aushängen

Je fünf Fragen vor dem Absenden und vor dem Verwenden, als Empfehlung von TOMORIS.

Vor dem Absenden eines Prompts, eines Uploads oder einer Verbindung zu Postfach und Dateien:

☐ Nutze ich ein freigegebenes Werkzeug mit meinem Arbeitskonto?

☐ Welche Ampelstufe haben die Daten, ist das Werkzeug dafür freigegeben, und genügen Platzhalter?

☐ Sind Zugangsdaten, Kommentare, Metadaten oder geschwärzte Stellen enthalten, die nicht hinausdürfen?

☐ Stammt der Inhalt aus einer unbekannten Quelle? Dann auf versteckte Anweisungen achten.

☐ Braucht der Assistent dafür wirklich Zugriff auf Postfach, Dateien oder Browser?

Vor dem Verwenden eines Ergebnisses, dem Bestätigen einer Aktion oder einer Veröffentlichung:

☐ Habe ich Fakten, Zahlen und Quellen geprüft: Gibt es jede Quelle, und sagt sie das wirklich?

☐ Fehlt in der Zusammenfassung etwas Wichtiges, etwa Ausnahmen, Zeitangaben oder Risiken?

- ☐ Ist Code geprüft und getestet, und gibt es die vorgeschlagenen Pakete wirklich?
- ☐ Passen Aufforderungen, Links und Aktionen zur Aufgabe, und habe ich Empfänger, Inhalt und Ziel gelesen?
- ☐ Muss das Ergebnis gekennzeichnet oder vor der Veröffentlichung freigegeben werden?

Für Führungskräfte: Werkzeuge, Datenklassen und Meldeweg festlegen (Vorlage KI-Nutzungsrichtlinie), Schulungen dokumentieren, Berechtigungen etwa in SharePoint bereinigen (Microsoft Learn).

NÄCHSTER SCHRITT

Selbst prüfen oder gemeinsam umsetzen

UMSETZUNG BESPRECHEN

Kontakt aufnehmen

tomoris.com/contact

contact@tomoris.com

08

Quellen

Alle Aussagen sind belegt. Gesetzestexte gelten in der jeweils geltenden Fassung.

- 1 **KI-Verordnung (EU) 2024/1689, konsolidiert 27.07.2026**
Art. 3 Nr. 4 und 60; Art. 4, 50, 111 Abs. 4, 113
eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX:02024R1689-20...
- 2 **VO (EU) 2026/1744, Digital-Omnibus zur KI**
Art. 1 Nr. 5 (neuer Art. 4), Art. 4 (Inkrafttreten)
eur-lex.europa.eu/eli/reg/2026/1744/oj/deu
- 3 **EU-Kommission: AI Literacy – Questions & Answers**
Stand 20.10.2026; Frage zur Dokumentation
digital-strategy.ec.europa.eu/en/faqs/ai-literacy-questions-ans...
- 4 **Verordnung (EU) 2016/679 (DSGVO), EUR-Lex**
ErwG 26; Art. 4 Nr. 1 und 12, Art. 5, 9, 33
eur-lex.europa.eu/eli/reg/2016/679/oj/deu
- 5 **UWG, § 26b (Geschäftsgeheimnis), RIS**
§ 26b Abs. 1, Fassung BGBl. I Nr. 109/2018
www.ris.bka.gv.at/NormDokument.wxe?Abfrage=Bundesnormen&Gesetze...
- 6 **OWASP Top 10 for LLM Applications 2026**
Nummern laut PDF 2026; Einzelseiten auf genai.owasp.org noch mit Nummern 2025 (LLM06, 09, 05)
genai.owasp.org/resource/owasp-genai-llm-top-10-2026
- 7 **NIST AI 600-1: Generative AI Profile (Juli 2024)**
Abschnitte 2.2, 2.7 und 2.9
nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf
- 8 **BSI: Generative KI-Modelle – Chancen und Risiken für Industrie und Behörden, Version 2.0**
Version 2.0 vom 17.01.2025; R2 bis R4, R6, R8, R28; M16 bis M18, M20
www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/KI/Generative_KI-Mo...
- 9 **BSI: Basisschutz gegen Indirect Prompt Injections in dokument-basierten LLM-Chats, Version 1.0**
Version 1.0 vom 14.09.2026; Einleitung, Abschnitt 2.5
www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/KI/Basisschutz_Indi...
- 10 **Anthropic: Halluzinationen reduzieren**
Grundlegende Strategien, Best-of-N-Verifizierung
platform.claude.com/docs/de/test-and-evaluate/strengthen-guardr...
- 11 **Anthropic: Jailbreaks und Prompt-Injections abwehren**
Abschnitt Indirekte Prompt-Injection
platform.claude.com/docs/de/test-and-evaluate/strengthen-guardr...
- 12 **Microsoft Learn: Daten, Datenschutz und Sicherheit für Microsoft Copilot**
Organisationsdaten; Inhalte, die Copilot erstellt
learn.microsoft.com/de-de/microsoft-365/copilot/microsoft-365-c...