FÜR KUNDEN · PARTNER · ENTSCHEIDER

LEHRDOKUMENT · AGENTIC AI

Agentic AI, kontrolliert — wie aus Sprachmodellen belastbare Fachabläufe werden

Erst das Fundament: was Sprachmodelle, Agenten, Workflows und Orchestrierung wirklich sind. Dann die Architektur, die daraus ein Arbeitsmittel macht: Belege statt Behauptungen, Prüfung statt Vertrauen, menschliche Freigabe statt Automatik. Reifegrade werden in diesem Dokument offen benannt — Ehrlichkeit ist unser Arbeitsprinzip, nicht unsere Schwäche.

Neon Arc · Stand 16. Juli 2026

LESEART: Fett = Begriff. „Bild“ = die Metapher zum Merken. „Auf den Punkt“ = der Satz zum Mitnehmen. Grundregel des Hauses: im Zweifel abschwächen, nie aufhübschen.

1 · Was wir bauen — und was bewusst nicht

Die wichtigste Klarstellung zuerst: Wir verkaufen weder einen Chatbot noch bloße ‚LLM-Orchestrierung'. Wir bauen kontrollierte Fachabläufe, in denen KI nur ein Baustein ist.

DER EINE EHRLICHE SATZ

Wir nehmen einen echten Fachablauf, bauen die verlässlichen Teile als normale Software und setzen ein Sprachmodell nur für Sprache, Einordnung und Entwürfe ein. Der Mensch behält die Entscheidung; Code, Rechte und Protokolle behalten die Kontrolle.

In 30 Sekunden: Die meisten kennen KI als Chat — Frage rein, Antwort raus. Wir gehen weiter: ein abgestimmtes Team spezialisierter Agenten mit einem Koordinator, einer unabhängigen Prüf-Instanz und einem nachprüfbaren Ergebnis. Die Entscheidung bleibt immer beim Menschen, und jede Aussage ist belegt.

Vier Begriffe, die nie vermischt werden dürfen

LLM	Das Sprachmodell: ein sehr gutes Muster- und Textsystem. Es liest, formuliert, strukturiert und kann Vorschläge machen. Es ist nicht automatisch ein Agent und besitzt keine Rechte oder Hände.
AGENT	Ein Auftragsbearbeiter: ein Programm mit Ziel, Zustand, erlaubten Werkzeugen und Stopp-Regeln. Er kann ein LLM nutzen, kann aber auch reine, feste Logik sein.
WORKFLOW	Der fachliche Ablauf: Wer gibt was hinein? Welche Prüfung folgt? Was darf automatisch passieren? Was muss freigegeben werden? Der Workflow existiert auch ohne KI.
ORCHESTRIERUNG	Die Verkehrsleitung: Sie startet passende Schritte, gibt nur den nötigen Kontext weiter, wartet auf Ergebnisse, setzt Grenzen und protokolliert alles.

Was ein Kunde tatsächlich bekommt

01 **Einen klaren Fachablauf.** Eingang, Rollen, Entscheidung, Ausnahme und Rückfallweg werden gemeinsam präzisiert.

02 **Ein selbst gebautes Softwarepaket.** Oberfläche, Datenmodell, Rechte, Schnittstellen, Prüfregeln, Tests und Audit-Spur sind normaler, versionierter Code.

03 **Einen gezielten KI-Baustein.** Das Sprachmodell liest unstrukturierte Sprache, extrahiert, klassifiziert oder formuliert einen Entwurf — keine unsichtbare Endentscheidung.

04 **Betrieb und Nachweis.** Monitoring, Freigaben, Fehlerpfade, Kostenlimits und eine messbare Pilotfrage machen daraus ein Arbeitsmittel statt einer Demo.

2 · Der Maschinenraum — wer macht was

So sieht ein robustes Agenten-System aus: nicht als magischer Schwarm, sondern als klarer Fluss mit zuständigen Komponenten.

Der robuste Standardablauf

01 **Eingang.** Ein Mensch oder ein anderes System liefert etwa ein PDF, eine Anfrage oder einen Datensatz.

02 **Feste Schranken.** Server-Code prüft Anmeldung, Rolle, Mandant, Dateityp, Vollständigkeit, Frist und Kostenlimit. Das braucht kein LLM.

03 **KI-Arbeit, nur wenn sie Mehrwert hat.** Ein begrenzter Auftrag plus freigegebener Kontext geht an das Modell: zum Beispiel „fasse zusammen“ oder „finde diese Angabe im Text“.

04 **Mechanische Prüfung.** Code verifiziert Zitate, Werte, Quellen, Regeln und Format. Eine zweite Prüfroutine darf den Entwurf verwerfen.

05 **Menschliche Freigabe und Aktion.** Erst danach wird gespeichert, versendet oder eine Aufgabe ausgelöst. Jede relevante Aktion hat eine Spur.

Vier Arten von „Agenten“ — präzise statt Marketing

STATISCHER WORKER	Ein kleiner Dienst oder Job: Fristen prüfen, PDFs validieren, Daten abgleichen. Er folgt festem Code und ruft kein LLM.
KI-WORKER	Ein begrenzter Job mit LLM-Aufruf: Text extrahieren, kategorisieren, Entwurf erzeugen. Er liefert Struktur, keine freie Handlung.
ORCHESTRATOR	Normales Programm für Zustände und Reihenfolge: welcher Schritt jetzt, welcher parallel, wann abbrechen, wer muss freigeben?
MENSCH	Die letzte Instanz für Folgen, die Recht, Geld, Kundenbeziehung oder Risiko betreffen. Kein Modell ersetzt diese Verantwortung.

Bau-Ebene und Produkt-Ebene sind zwei verschiedene Dinge

Entwickelt wird bei uns mit modernen KI-Entwicklungswerkzeugen (etwa Claude Code) — das sind Werkzeuge unserer Werkbank, keine unsichtbaren Mitarbeiter im Kundensystem. Im **laufenden Produkt** übernimmt ein passend konfiguriertes Sprachmodell (für datenschutzkritische Anwendungen ein EU-Anbieter wie Mistral) einen eng begrenzten Denk- oder Sprachschritt. Der **Orchestrator** selbst ist überwiegend unser eigener Code — keine LLM-Plauderrunde.

> AUF DEN PUNKT Beide Ebenen nutzen dieselben Prinzipien — aber unterschiedliche Werkzeuge. Was uns beim Bauen hilft, läuft deshalb noch lange nicht beim Kunden.

3 · Wie KI und Agenten funktionieren — einfach erklärt

Damit Sie einordnen können, wie das alles wirklich läuft — einfach, aber richtig.

LLM (Sprachmodell) — das „Gehirn“

Ein **LLM** (Large Language Model, z. B. Claude oder Mistral) ist ein Programm, das Sprache versteht und erzeugt. Es hat kein Bewusstsein — es sagt sehr gut das nächste Wort voraus. Ein **Token** ist ein Sprach-Häppchen (ca. ein halbes Wort); das **Kontextfenster** ist sein Kurzzeitgedächtnis (wie viel es auf einmal „im Kopf“ hat). Ein **Prompt** ist die Anweisung, die man ihm gibt.

BILD *Ein extrem belesener Praktikant, der alles schon mal gelesen hat, blitzschnell formuliert — aber nur weiß, was auf seinem Schreibtisch (Kontextfenster) liegt, und Dinge erfindet, wenn man ihn lässt.*

Chatbot vs. Agent — der entscheidende Unterschied

Ein **Chatbot** antwortet nur (Text rein, Text raus). Ein **KI-Agent** kann *handeln*: Er liest Informationen, bedient vorhandene Software über **Werkzeuge** und arbeitet eine Aufgabe in mehreren Schritten ab — unter Aufsicht.

> AUF DEN PUNKT Ein Chatbot redet. Ein Agent handelt — in Schritten, mit Werkzeugen, und mit menschlicher Freigabe für alles, was etwas verändert.

Werkzeuge / Function-Calling — wie ein Agent etwas TUT

Das Modell selbst kann nichts anfassen. Man gibt ihm eine Liste klar beschrifteter **Werkzeuge** („Termin anlegen“, „Kunde suchen“) — das nennt man **Function-Calling**. Das Modell entscheidet nur, *welches* Werkzeug es aufrufen will; ausführen tut es unser Programm — und bei verändernden Werkzeugen erst nach menschlicher Freigabe.

BILD *Ein Werkzeugkasten mit beschrifteten Werkzeugen. Der Agent zeigt auf eines („nimm den Schraubenzieher“) — aber ein Mensch dreht bei allem Riskanten selbst.*

Orchestrator + Spezial-Agenten — wie ein Team

Ein **Orchestrator** (Koordinator) zerlegt eine Aufgabe, verteilt die Teile an **Spezial-Agenten** mit je engem Aufgabengebiet und führt die Ergebnisse geprüft zusammen. **Recherche-Agenten** lesen nur Quellen und folgen keinen fremden Weisungen. So „kommunizieren“ sie: Der Koordinator delegiert und sammelt ein — bei uns über eine gemeinsame Datenablage, nicht als Agent-zu-Agent-Chat.

BILD *Ein Dirigent und sein Orchester: Der Dirigent gibt den Einsatz, jeder Musiker spielt seinen Part, zusammen wird es ein Stück. Kein Musiker spielt allein das ganze Werk.*

Erzeuger ist nicht Prüfer — warum das der Trick ist

Der Agent, der ein Ergebnis *erzeugt*, darf es nicht selbst freigeben. Eine **zweite, unabhängige Instanz** versucht, es zu widerlegen (adversariale Prüfung) — im Zweifel gilt es als falsch.

BILD *Vier-Augen-Prinzip in der Buchhaltung: Wer bucht, gibt nicht frei. Und der Prüfer ist bewusst ein Miesepeter.*

RAG & Zitat-Zwang — „belegt statt behauptet“

Halluzination heißt: Das Modell erfindet etwas plausibel Klingendes. Gegenmittel: **RAG** (Retrieval-Augmented Generation) — der Agent schlägt die Antwort erst in freigegebenen Quellen nach und darf nur mit **Fundstelle** antworten. Ohne Beleg wird die Aussage blockiert und als „offen“ ausgewiesen — nicht erfunden ausgegeben.

BILD *Kein Gedächtnis-Zitieren aus dem Kopf — der Agent muss in die Bibliothek gehen und die Seite mit dem Finger zeigen, sonst gilt es nicht.*

Welche Aufgaben Agenten übernehmen können

- **Wissen & Dokumente:** in Handbüchern/Richtlinien suchen, mit Zitat antworten.
- **Proaktiv im Hintergrund:** Morgenbriefing, Übergabe, Monitoring, Erinnerungen.
- **Prüfen & Sorgfalt:** Unterlagen gegen Regeln checken, Prüfvermerk entwerfen.
- **Vorgänge & Triage:** Anfragen einsortieren, vorbereiten, zur Freigabe vorlegen.
- **Verdichten:** Berichte, Protokolle, Entscheidungsvorlagen entwerfen — Mensch gibt frei.

4 · Tiefer: Wie ein Sprachmodell wirklich tickt

Für alle, die es genauer wissen wollen — was unter der Haube passiert.

Training vs. Inference — zwei verschiedene Dinge

Ein Modell wird **einmal trainiert** (Monate, riesige Textmengen, sehr teuer). Dabei lernt es Muster und speichert sie in Milliarden Zahlen — den **Gewichten** (Parametern). Danach ist das Modell fix. Jede Antwort im Betrieb heißt **Inference**: Es rechnet mit den fixen Gewichten und lernt dabei *nicht* in diese Gewichte hinein. Ob ein Anbieter Eingaben speichert oder für Verbesserungen nutzt, ist eine separate Produkt-, Vertrags- und Konto-Einstellung; das wird vor einem Kundeneinsatz konkret geprüft.

BILD *Ausbildung vs. Arbeitstag: Nach der Ausbildung (Training) wendet der Mitarbeiter sein Wissen an (Inference) — er lernt im Job aber nicht automatisch dazu.*

Next-Token-Vorhersage, Transformer, Attention

Im Kern macht ein LLM **eine** Sache: das nächste Sprach-Häppchen vorhersagen — wieder und wieder. Die Architektur heißt **Transformer**, ihr Trick **Attention** („Aufmerksamkeit"): Das Modell gewichtet, welche früheren Wörter für das nächste wichtig sind. Kein Regelwerk, kein Nachschlagen im Kopf — reine Mustervorhersage. Genau deshalb ist es großartig im Formulieren und schlecht im Faktentreue-Garantieren.

BILD *Ein extrem gutes Autovervollständigen, das den ganzen Satzzusammenhang mitgewichtet — nicht nur das letzte Wort.*

Warum es halluziniert — und der Temperatur-Regler

Es ist auf **„plausibel klingend"** optimiert, nicht auf „wahr". Fehlt Wissen, füllt es die Lücke selbstbewusst. Die **Temperatur** ist ein Regler für Zufall/Kreativität: niedrig (Richtung 0) = fast immer die wahrscheinlichste, konsistente Antwort (für Fakten und Prüfung); hoch = kreativer, aber unzuverlässiger. Für Prüf-Aufgaben fährt man bewusst niedrig.

BILD *Temperatur = vorsichtiger Sachbearbeiter (niedrig) gegenüber Brainstorming-Kollege (hoch).*

Das Kontextfenster und seine Grenzen

Das Arbeitsgedächtnis ist begrenzt (in Häppchen/Tokens gemessen). Zu viel Text auf einmal: Wichtiges in der Mitte geht unter („lost in the middle"). Deshalb geben wir Agenten nur das Nötige, sauber portioniert — nicht alles auf einmal. Das ist auch ein Kostenhebel: mehr Tokens = mehr Geld und mehr Latenz.

GEWICHTE / PARAMETER	Die Milliarden gelernten Zahlen, in denen das Wissen steckt. Größer ist oft klüger, aber teurer und langsamer.
TOKEN	Sprach-Häppchen (~ein halbes Wort). Kosten und Kontextgröße rechnet man in Tokens.
KNOWLEDGE CUTOFF	Trainings-Stichtag: danach weiß das Modell nichts — für Aktuelles braucht es Werkzeuge/Quellen (RAG).
SYSTEM-PROMPT	Die fixe Grundanweisung (Auftrag, Rolle, Regeln), die vor jeder Nutzerfrage steht.

5 · Was ein Agent wirklich ist — und „fixe“ Agenten ohne LLM

Der Begriff wird inflationär benutzt. Hier die saubere Definition — und warum gute ‚Agenten‘ oft gar keine KI brauchen.

Die saubere Definition

Ein **KI-Agent** ist im engeren Sinn ein Sprachmodell mit Ziel, Schleife, Werkzeugen und Stopp-Bedingung (optional mit Gedächtnis). Die Schleife: denken, Werkzeug wählen, Ergebnis beobachten, weiterdenken — bis das Ziel erreicht ist oder eine Grenze/Freigabe greift. Eine reine, regelbasierte Automation kann ebenfalls Aufträge abarbeiten, braucht dafür aber kein LLM. Ein einfacher Chatbot antwortet meist nur einmal.

BILD *Ein Chatbot beantwortet eine Frage. Ein Agent bearbeitet einen Auftrag — er darf zwischendurch nachschlagen, rechnen, anlegen, und hört auf, wenn er fertig ist.*

Die sechs Schrauben, an denen man einen Agenten „einstellt“

01 **Modell** — das Gehirn (welches, wie stark).

02 **System-Prompt** — der Auftrag + die Regeln + die Rolle.

03 **Werkzeuge** — was es tun darf (Hände).

04 **Guardrails / Tore** — was verboten ist, Freigabe-Pflicht, Kostengrenzen.

05 **Wissen / Kontext** — freigegebene Quellen (RAG), geladene Anleitungen.

06 **Orchestrierung** — wie viele Agenten, wie koordiniert.

Steuern heißt: an diesen sechs Schrauben drehen — nicht „die KI umprogrammieren“.

„Fixe“ Agenten ohne LLM — ja, und bewusst

Nicht jeder „Agent“ braucht ein Sprachmodell. **Regelbasierte Automationen** (wenn-dann, Zeitpläne, Schwellenwerte) sind vorhersehbar: gleiche Eingabe, gleiche Ausgabe, kein Halluzinations-Risiko, günstig und gut prüfbar. Ein großer Teil guter Hintergrund-Jobs ist reine Logik ohne KI; auch ein Fristen- oder Compliance-Motor kann bewusst als Zustandsmaschine ohne Modell-Aufruf gebaut werden.

BILD *Ein Thermostat ist ein perfekter „Agent“ ohne Intelligenz — Regel: zu kalt, dann heizen. Kein Gehirn nötig.*

> AUF DEN PUNKT Wir setzen KI nur dort ein, wo echtes Urteilsvermögen nötig ist. Alles Verlässliche läuft als reine Logik — sicherer, günstiger, nachprüfbar. Oft ist der beste ‚KI-Agent‘ gar keine KI.

Hybrid — unser Standard

Das Beste aus beidem: **KI entscheidet und strukturiert vor, Code führt aus und prüft.** Die KI schlägt vor, mechanische Tore setzen um. So bekommt man Flexibilität *und* Verlässlichkeit — und das Riskante passiert nie ungeprüft.

6 · Schwarmintelligenz — und die Linearitätsfalle

Mehr Agenten sind nicht automatisch besser — die Kunst ist die Koordination. Und die häufigste Falle hat mit Agenten gar nichts zu tun.

Die Linearitätsfalle: Warum der Chat allein nicht reicht

Wer eine KI **seriell** befragt — Frage, Antwort, nächste Frage — bekommt flüssige Antworten, aber jede hängt am Gesprächsverlauf davor: Die **Reihenfolge der Fragen färbt das Ergebnis**, und niemand prüft gegen. Das Ergebnis fühlt sich wie Gründlichkeit an, ist aber vom Zufall der Befragungsreihenfolge mitbestimmt. Der Chat ist dabei nicht das falsche Werkzeug — er ist das unstrukturierte: Für eine schnelle Auskunft reicht er, für eine Fach- oder Führungsentscheidung braucht es Struktur, Belege und Gegenprüfung.

> AUF DEN PUNKT Gegen die Linearitätsfalle helfen drei Dinge: parallele Perspektiven statt einer Fragenkette, Beleg-Pflicht statt Eloquenz, und eine unabhängige Instanz, die widerlegen darf.

Können Agenten „miteinander reden wie ein Team“?

Technisch **können** mehrere Agenten kommunizieren (Multi-Agenten-Systeme). Aber freies Agent-zu-Agent-Geplauder ist oft **kontraproduktiv**: Der Kontext bläht auf, Fehler pflanzen sich fort, sie drehen sich im Kreis, die Kosten explodieren. Die robuste Lösung ist **strukturierte Koordination**: Ein Koordinator zerlegt, schickt Spezial-Agenten mit eigenem, sauberem Kontext los, jeder liefert ein Ergebnis zurück — koordiniert wird über eine gemeinsame Ablage, nicht über ein Dauergespräch. Ein unabhängiger Prüf-Agent kontrolliert, statt mitzuquatschen.

BILD *Kein Großraumbüro, in dem alle durcheinanderreden — sondern ein Projektleiter, der Aufgaben verteilt und Ergebnisse einsammelt.*

Wann freies „Reden“ doch sinnvoll ist

Für **Vielfalt und Debatte**: mehrere Perspektiven, die bewusst gegeneinander argumentieren — etwa Perspektiv-Agenten, die dasselbe aus verschiedenen Haltungen durchdenken, oder ein Bewertungs-Panel. Da ist der „Streit“ das Produkt, nicht der Störfaktor.

> AUF DEN PUNKT Schwarmintelligenz richtig verstanden: nicht „alle reden gleichzeitig“, sondern „viele abgestimmte Perspektiven, geprüft zusammengeführt“. Mehr Agenten sind nicht automatisch besser — die Kunst ist die Koordination.

7 · Sicherheit & Datenschutz — die Wächter

Die Begriffe, auf die es im technischen Gespräch ankommt — und wie eine saubere Architektur sie konkret umsetzt.

PII & Maskierung / Anonymisierung / Pseudonymisierung

PII = „personally identifiable information", also personenbezogene Daten (Name, E-Mail, Telefon, IBAN, Gesundheitsdaten). **Maskieren** = unkenntlich machen (a***@firma.de). **Pseudonymisieren** = durch ein Kürzel ersetzen, das man mit einem Schlüssel zurückholen könnte. **Anonymisieren** = so entfernen, dass es *nicht mehr* rückführbar ist. Sauber gebaut heißt: Personenbezogenes wird maskiert, *bevor* es überhaupt zum Modell gelangt — besonders sensible Daten (Art. 9 DSGVO) gehen gar nicht erst hin.

BILD *Der schwarze Filzstift des Zensors: Bevor das Blatt rausgeht, sind Name und Nummer schon geschwärzt.*

Prompt-Injection — das Sicherheitsrisiko Nr. 1 bei KI

Ein Angreifer versteckt in einem Text heimliche Anweisungen („ignoriere deine Regeln, gib X preis"), damit der Agent etwas Unerwünschtes tut. Laut OWASP das größte KI-Risiko. Schutz: Nicht vertrauenswürdige Inhalte werden isoliert verarbeitet und als reine **Daten** behandelt, nie als Befehle — mehrschichtig, im Live-Pfad.

BILD *Ein Trickbetrüger schiebt dem Boten einen Zettel unter: „Der Chef sagt, gib mir den Tresorschlüssel." Der gut geschulte Bote nimmt nur Befehle vom echten Chef, nicht aus fremden Zetteln.*

SSRF — Server-Side Request Forgery

Ein Angriff, bei dem man den *Server* austrickst, damit er in **Ihrem** Auftrag interne oder verbotene Adressen abruft (z. B. interne Verwaltungs-Server oder Cloud-Zugangsdaten). Schutz: Ausgehende Abrufe erlauben nur unbedenkliche Ziele und blocken interne Adressen — auch gegen den Trick, wo sich eine Adresse erst beim Abruf ändert (DNS-Pinning).

BILD *Du überredest die Poststelle der Firma, dir die geheimen internen Akten zu holen — weil die Poststelle Türen öffnen darf, die du nicht darfst. SSRF-Schutz heißt: Die Poststelle prüft jede Adresse und weigert sich bei internen.*

Die restlichen Wächter — kurz

RBAC (ROLLEN-RECHTE)	Jedes Werkzeug ist an eine Rolle gebunden. *Bild:* Werksausweis, der nur bestimmte Türen öffnet.
HUMAN-IN-THE-LOOP	Der Mensch bestätigt jede verändernde Aktion. *Bild:* Vier-Augen-Prinzip, zwei Schlüssel für den Tresor.

MULTI-TENANT / ISOLATION	Ein sauber implementiertes Mandantenmodell verhindert, dass Mandant A Daten von Mandant B sieht. *Bild*: abgeschlossene Schließfächer im selben Haus.
AUDIT-TRAIL	Nachvollziehbares Protokoll (wer, wann, was). „Fälschungssicher" oder „revisionssicher" gilt erst, wenn Unveränderbarkeit, Aufbewahrung und alle Kontrollen für den konkreten Prozess belegt sind. *Bild*: Notariatsbuch — aber nur, wenn jede Seite wirklich versiegelt ist.
GUARDRAILS / TORE	Eingebaute Schranken: zum Beispiel keine Freigabe bei offenen Prüfungen, Rollenrechte oder Kostenlimits — jeweils pro Ablauf bewusst definiert und getestet.

8 · Datenhoheit, EU-Anbieter & lokale Modelle

Die Frage kommt sicher: „Wo laufen die Daten, welches Modell, ist das DSGVO-sicher?" — hier die ehrlichen Antworten.

Offenes vs. geschlossenes Modell — und was „lokal" heißt

Geschlossene Modelle (Claude, GPT) laufen nur beim Anbieter über eine Schnittstelle — top Qualität, aber die Daten gehen dorthin. **Offene Modelle** (die „Gewichte" sind einsehbar: z. B. Llama, Mistral, Qwen) kann man **selbst betreiben** — auf einem gemieteten oder eigenen GPU-Server. Das ist mit „lokal / self-hosted / on-premise" gemeint: Das Modell läuft in Ihrer Kontrolle. Das erhöht die technische Kontrolle, ersetzt aber weder Zugriffsrechte, Verschlüsselung, Backups noch eine Datenschutzprüfung.

BILD *Geschlossenes Modell = Taxi (bequem, aber ein Fremder fährt und weiß, wohin). Offenes Modell selbst betrieben = eigenes Auto (mehr Aufwand, aber du fährst und niemand sieht mit).*

Souveränität in drei Stufen

01 **Stufe 1 — EU-Anbieter:** KI-Dienst mit vertraglich und technisch passender EU-Region. Schneller Start, aber Region, Aufbewahrung, Unterauftragsverarbeiter und Produktmodus werden konkret geprüft.

02 **Stufe 2 — dedizierter EU-Server:** ein reservierter Server im EU-Rechenzentrum, kein geteilter Dienst.

03 **Stufe 3 — self-hosted / on-premise:** offenes Modell im eigenen Haus, bis zum abgeschotteten Netz (air-gapped). Höchster Schutz, höchster Aufwand.

„Ist Cloud-Hosting DSGVO-sicher?“ — die ehrliche Antwort

Ein Cloud-Dienst kann **Teil einer DSGVO-konformen Architektur** sein, aber nicht „automatisch DSGVO-sicher“. EU-Regionen und EU-Datengrenzen gelten nur für passende, korrekt konfigurierte Dienste und sehen begrenzte Ausnahmen vor. Bei US-Anbietern ist die Rechtslage ein zusätzlicher Risikofaktor; das konkrete Risiko hängt von Dienst, Region, Vertrag, Datenart, Schlüsseln und Zugriffskonzept ab. Das ist keine Rechtsberatung — diese Punkte werden mit Datenschutz und IT des Kunden geprüft.

> AUF DEN PUNKT Wir behandeln Datenstandort und Zugriff nicht als Marketing-Label: Dienst, Region, Vertrag, Datenfluss und Schlüsselkonzept werden für den konkreten Fall geprüft. Je höher der Schutzbedarf, desto mehr Kontrolle planen wir ein — bis self-hosted.

Was wir nutzen — offen gesagt

Entwickelt wird mit starken Denk-Modellen über moderne Entwicklungswerkzeuge. Für datenschutzkritische Anwendungen wird ein EU-Anbieter wie **Mistral** geprüft und passend konfiguriert. Wir **staffeln** die Modelle nach Aufgabe: starke Modelle für Orchestrierung und Kernlogik, mittlere fürs Prüfen, günstige und schnelle fürs mechanische Lesen und Ausführen. Der Anbieter ist per Konfiguration umschaltbar — kein Programmwechsel.

INFERENCE	Der laufende Betrieb des Modells (eine Anfrage beantworten) — im Gegensatz zum Training.
FINE-TUNING VS. PROMPTING	Fine-Tuning = das Modell nachtrainieren (teuer). Prompting = gutes Anweisen ohne Umbau (unser Weg).
EMBEDDING / VEKTOR-DATENBANK	Texte werden in Zahlen-Fingerabdrücke übersetzt, um Ähnliches zu finden — die Basis von Dokumenten-Suche/RAG.
AVV / SCC	Auftragsverarbeitungs-Vertrag / EU-Standardvertragsklauseln — wichtige Vertragsbausteine, aber kein Ersatz für eine passende technische Architektur.
CLOUD ACT	US-Gesetz, das bei US-Anbietern ein zusätzlicher Rechts- und Risikoaspekt sein kann. Konkrete Bewertung immer mit Datenschutz und Vertrag prüfen.

9 · Der EU AI Act — Checkliste statt Schreckgespenst

Regulierung ist kein Verkaufsargument und kein Schreckgespenst. Sie ist eine Checkliste dafür, wo Verantwortung, Transparenz und Dokumentation hingehören.

DIE KORREKTE HALTUNG

Nicht „wir sind automatisch compliant", sondern: Wir klassifizieren jeden echten Einsatz, dokumentieren Rolle, Risiko, Daten und menschliche Kontrolle — und holen bei Grenzfällen rechtliche Prüfung dazu.

Die Zeitlinie in Klartext (Stand Juli 2026)

SEIT 02.02.2025	Verbot bestimmter KI-Praktiken und Pflicht zu angemessener KI-Kompetenz: Wer KI einsetzt, muss die beteiligten Menschen passend schulen.
SEIT 02.08.2025	Governance und Pflichten für Anbieter allgemeiner KI-Modelle (GPAI). Wer ein Modell nur per Schnittstelle nutzt, ist nicht dadurch selbst GPAI-Modellanbieter.
SEIT 02.08.2026	Der AI Act gilt grundsätzlich. Transparenzpflichten können je Einsatz greifen; der genaue Umfang hängt von der Rolle und dem Produkt ab.
HOCHRISIKO	Nicht das Wort „KI" entscheidet, sondern der konkrete Zweck und die Folgen. Personal, Bildung, Kredit, Leistungen, Biometrie oder kritische Infrastruktur brauchen besondere Prüfung.

Was für jeden Kundenfall schriftlich geklärt wird

01 **Rolle und Zweck.** Sind wir Anbieter des Systems, der Kunde Betreiber oder beides? Welche Entscheidung wird vorbereitet — und welche nicht?

02 **Daten und Folgen.** Welche personenbezogenen oder vertraulichen Daten gehen hinein? Hat das Ergebnis Einfluss auf Menschen, Geld, Zugang oder Sicherheit?

03 **Kontrolle und Transparenz.** Wo sieht der Mensch, dass KI beteiligt war? Wo kann er korrigieren, stoppen und eine Entscheidung begründen?

04 **Nachweis.** Modell, Prompt, Quellen, Werkzeuge, Tests, Fehlerfälle, Freigaben und Änderungen werden nachvollziehbar dokumentiert.

05 **Rechtliche Schicht.** Der AI Act ersetzt weder DSGVO noch Vertrag, Auftragsverarbeitung, Löschkonzept oder gegebenenfalls Datenschutz-Folgenabschätzung.

Warum Unternehmen trotzdem so stark in KI investieren

Weil in fast jedem Betrieb unstrukturierte Arbeit steckt: E-Mails lesen, Dokumente verstehen, Informationen suchen, Fälle vorsortieren, Texte vorbereiten. Ein gut begrenzter KI-Schritt spart Zeit und macht Wissen schneller verfügbar. Der Wert entsteht aber erst, wenn er in einen echten Prozess, verlässliche Daten und verantwortliche Freigaben eingebettet ist — nicht durch möglichst viele „Agenten".

Was Selbst-Hosting nicht automatisch löst

Ein eigenes Modell kann Datenwege verkürzen, ersetzt aber weder Risiko-Einstufung noch Transparenz, Rollenrechte, menschliche Aufsicht, Tests oder Datenschutz. Es ist eine Hosting-Entscheidung — kein Compliance-Stempel.

> AUF DEN PUNKT Unser Gegenmittel unabhängig von der Risikoklasse: Zweck begrenzen, Daten minimieren, Rollen und Freigaben erzwingen, KI-Ausgaben prüfen, Entscheidungen protokollieren, Tests gegen Fehler fahren und einen Rückfallweg ohne KI vorsehen. Bei Personal-, Kredit-, Zugangs- oder sicherheitsrelevanten Entscheidungen folgt vor Einsatz ein gesondertes Legal-/Risk-Review.

10 · Vom Konzept zum belastbaren System

Der Ablauf hinter einer guten Demo — wie aus einer Idee ein belastbares System wird, ohne so zu tun, als sei KI Magie.

UNSER ARBEITSPRINZIP

Wir starten nicht mit „Welches Modell nehmen wir?", sondern mit einem konkreten Vorgang: Wer gibt was hinein, welche Entscheidung soll vorbereitet werden, was darf automatisch passieren — und was muss ein Mensch freigeben? KI ist dabei ein Baustein. Die Verlässlichkeit entsteht durch Regeln, Daten, Tests und Verantwortung.

Der Ablauf in fünf klaren Schritten

01 **Einen engen, echten Fall wählen.** Zum Beispiel: „Ein Lieferant lädt einen Nachweis hoch; der Prüfer soll nur die kritischen Punkte zuerst sehen." Nicht „wir automatisieren alles".

02 **Eingang, Entscheidung und Ergebnis festlegen.** Welche Daten kommen hinein? Welche Regel oder Quelle gilt? Was sieht der Mensch am Ende? Das wird als Ablauf und Akzeptanzkriterium aufgeschrieben.

03 **Den sicheren Teil in Code bauen.** Rechte, Datenzugriff, Berechnungen, Fristen, Freigaben und Protokolle sind normale Softwarelogik. Sie muss vorhersehbar bleiben.

04 **KI nur für den unstrukturierten Teil einsetzen.** Lesen, Zusammenfassen, Einordnen oder einen Entwurf machen. Quellen und erlaubte Werkzeuge werden eng begrenzt.

05 **Beweisen, dann ausweiten.** Mit Testfällen, Gegenbeispielen, Review und einer kleinen Demo. Erst wenn der enge Fall nachvollziehbar funktioniert, kommt der nächste dazu.

BILD *Wie beim Hausbau: Erst der Plan, dann das Fundament und die Leitungen; erst danach kommt die schöne Fassade. Das Sprachmodell ist ein sehr schneller Facharbeiter — nicht das Fundament.*

Womit wir praktisch arbeiten

FACHLOGIK	Die Regeln des Geschäfts: Fristen, Rollen, Freigaben, Status und Sonderfälle. Sie bestimmen, was das System überhaupt tun darf.
ANWENDUNGSCODE	Die normale Software, die Oberflächen, Datenfluss, Zugriffe und Integrationen verbindet. Versioniert, überprüfbar und testbar.
KI / SPRACHMODELL	Liest und strukturiert Sprache, schlägt nächste Schritte oder Entwürfe vor. Sie erhält nur den nötigen Kontext und keine unbegrenzte Handlungsfreiheit.
QUELLEN & DATEN	Freigegebene Dokumente, Datenbankeinträge und fachliche Regeln. Bei Aussagen mit Risiko zählen Fundstelle und Datenqualität mehr als eloquente Formulierungen.
KONTROLLEN	Rollenrechte, Maskierung, Tests, Gegenprüfungen, Freigabe und Protokolle. Sie machen aus einer KI-Demo ein kontrollierbares Arbeitsmittel.

Reifegrad: das Wort, das Vertrauen schafft

Wir benennen für jedes System offen, wo es steht — mit einem festen Vokabular. Das ist kein Kleingedrucktes, sondern unser Qualitätsversprechen: Sie wissen immer, was ein Nachweis belegt und was nicht.

KONZEPT	Das Problem und der Ablauf sind beschrieben. Noch kein Beweis, dass es technisch oder wirtschaftlich funktioniert.
TECHNISCH GEBAUT	Der Ablauf läuft im Code und auf Beispieldaten. Belegt technische Machbarkeit — nicht automatisch Betriebssicherheit.
IN HÄRTUNG	Fehlerfälle, Rechte, Sicherheit, Bedienung und Wiederholbarkeit werden systematisch geprüft.
KONTROLLIERTER PILOT	Ein klar abgegrenzter echter Fall mit benannten Nutzern, Messgröße und Rückfallebene. Erst hier entsteht Nutzennachweis.

BETRIEB	Wiederholbar, mit Verantwortlichen, Support, Monitoring und echten Nutzungsdaten. Nicht mit „fertig programmiert“ verwechseln.

> AUF DEN PUNKT Wir bauen keine KI-Show. Wir nehmen einen konkreten Arbeitsablauf, sichern die verlässlichen Teile in Code ab und setzen KI dort ein, wo sie wirklich hilft. Dann belegen wir den Reifegrad offen.

Und ehrlich bleibt ehrlich: Ein grüner Test beweist nicht, dass Kunden zahlen. Eine schöne Demo beweist nicht, dass ein System sicher oder produktionsreif ist. Beides sind unterschiedliche Nachweise — wir verwechseln sie nicht.

Belegtes Wissen und generierter Vorschlag — der wichtige Unterschied

Die Belegpflicht sichert Aussagen **über den Fall**: Jede Zuschreibung braucht ein wörtliches Zitat, sonst wird sie blockiert. Ein **Vorschlag** dagegen — ein Aktionsplan, eine Frist, eine Zielzahl — beschreibt etwas Zukünftiges und ist deshalb *nicht* belegbar: Es gibt nichts zu zitieren, was noch nicht existiert. Ein Sprachmodell formuliert solche Zahlen plausibel, ohne den Fachkontext zu kennen (realistische Laufzeiten, Aufsichtsfristen, Benchmarks). Das ist kein Fehler der Kontrollschicht, sondern die Grenze jeder generierenden KI — und wir benennen sie offen, statt sie zu kaschieren.

Im echten Einsatz wird aus „geraten“ „hergeleitet und geprüft“ — auf drei Ebenen:

01 **Der Mensch entscheidet.** Jeder Vorschlagswert wird übernommen, geändert oder verworfen — protokolliert und namentlich zugerechnet. Bereits eingebaut.

02 **Fach-Korpus als zweite Quelle.** Der Generator prüft Werte gegen echte Projektdaten, Benchmarks und regulatorische Fristen des Kunden — dann sind sie hergeleitet, nicht geschätzt.

03 **Deterministische Guardrails.** Harte Regeln im Code, die kein Modell überreden kann: keine Kennzahl ohne hinterlegte Berechnungsbasis, Fristen nur gegen den echten Projektkalender, definierte Wertebereiche.

> AUF DEN PUNKT Was wir nicht versprechen, ist eine Garantie ohne diese Schichten — sie entstehen erst im Pilot mit den Daten des Kunden. Was wir zusichern: Eine generierte Zahl wird nie unbemerkt zur Entscheidung. Der Vorschlag ist als Vorschlag markiert, die Prüfung liegt beim Menschen und, im Ausbau, bei belegbaren Quellen und festen Regeln.

11 · Referenz-Systeme — Funktion und Reifegrad

Woran wir arbeiten — nach Funktion benannt, mit ehrlichem Reifegrad. Interne Projektnamen und Kundennamen nennen wir grundsätzlich nicht.

KI-COMPLIANCE-AUDITOR	Ein Prüf-Agent für deutsches und EU-Recht: gegnerische Prüf-Rollen nacheinander (Aufspüren, Widerlegen, Konsolidieren) plus eine automatische Sperre, die jede Rechts-Aussage ohne belegte Quelle blockiert. Arbeitet gegen primärquellen-geprüfte Belege plus eine angebundene offene Rechts-Bibliothek als Nachschlage-Index. Kein Rechtsurteil, keine Rechtsberatung. Reifegrad: technisch gebaut, in Härtung.
KI-BETRIEBSASSISTENT	Eine mandantenfähige Betriebssoftware, deren KI-Assistent abgegrenzte Werkzeuge bedient — jede verändernde Aktion nur mit menschlicher Freigabe (erst Vorschau, dann Klick); jedes Werkzeug an eine Rolle gebunden; personenbezogene Daten werden vor der KI maskiert; der Anbieter ist per Konfiguration auf einen EU-Dienst umschaltbar. Reifegrad: einsatzbereit, in Härtung.
PROAKTIVE AUTOMATIONEN + BAUKASTEN	Eine Management-Plattform mit proaktiven Hintergrund-Automationen (Briefings, Übergaben, Stimmungsanalyse) — koordiniert über eine gemeinsame Datenablage — plus ein visueller Baukasten, mit dem man Abläufe ohne Programmierung zusammensteckt. Reifegrad: technisch gebaut; die Ausführung einzelner Bausteine ist noch in Umsetzung.
SICHERER FACH-CHATBOT	Ein öffentlich erreichbarer Fach-Chat im Live-Betrieb: mehrere Schutzschichten gegen manipulierte Eingaben (Prompt-Injection) und aktive Maskierung personenbezogener Daten direkt im Antwortstrom. Reifegrad: live, gehärtet aus geprüftem Vorgänger.
LIEFERKETTEN-PRÜFSYSTEM	Dokumentenprüfung mit belegtem Entwurf, menschlichem Review, unveränderlichem Snapshot und Audit-Kette. Reifegrad: arbeitender Prototyp auf Beispieldaten — die greifbare Demo für einen Pilot.
FÜHRUNGSFALL-KOMPASS	Digitalisierte Führungs-Fallarbeit: Perspektiv-Agenten denken parallel, eine unabhängige Prüf-Instanz falsifiziert, mechanische Tore erzwingen wörtliche Belege, der Mensch wählt und gibt frei. Reifegrad: arbeitender Prototyp auf Beispieldaten.

Warum so zurückhaltend formuliert? Weil jede dieser Beschreibungen beim Nachfassen halten muss. Konkrete Zahlen, Testberichte und Live-Einblicke zeigen wir im technischen Review — belegt statt behauptet.

12 · Die vier Sorgen — und die Architektur-Antworten

Die vier Fragen, die in jedem Unternehmen zuerst kommen — und die Architektur-Antworten darauf.

F **Halluzination — „Erfindet die KI nicht Dinge?“**

A Das Modell kann erfinden — genau deshalb: Zitat-Zwang gegen einen echten Korpus, Unbelegtes wird blockiert und sichtbar als „offen“ markiert, und eine unabhängige zweite Instanz falsifiziert. Erzeuger ist nicht Prüfer.

F **Kontrollverlust — „Macht die KI dann alles allein?“**

A Nein. Der Agent entscheidet nichts, er legt einen belegten Entwurf vor. Verändernde Aktionen brauchen menschliche Freigabe. Dazu mechanische Tore und Kostengrenzen.

F **Datenabfluss — „Wo landen unsere Daten?“**

A Personenbezogene Daten werden vor der KI maskiert, europäische Anbieter sind konfigurierbar, Außen-Abrufe sind gehärtet. Ein Pilot läuft in einer Sandbox mit anonymisierten Daten; Datenklasse, Vertrag und Region werden pro Fall schriftlich geklärt.

F **Prompt-Injection — „Kann man den Agenten austricksen?“**

A Nicht vertrauenswürdige Inhalte werden isoliert verarbeitet und als Daten behandelt, nie als Befehle — mehrschichtig im Live-Pfad. Versteckte Anweisungen werden zitiert und analysiert, aber nicht ausgeführt.

Fünf kluge Fragen, die Sie jedem Anbieter stellen sollten

01 Wo würden unsere Daten und das KI-Modell verarbeitet — in der EU, oder im eigenen Haus?

02 Welche Aktionen sollen automatisch laufen, welche brauchen zwingend eine menschliche Freigabe?

03 Wie wird protokolliert, wer wann welche Aktion ausgelöst hat — und wie lange aufbewahrt?

04 Wie ist der Agent gegen manipulierte Eingaben geschützt?

05 Mit welchem einzelnen, klar umrissenen Fall starten wir — hoher Nutzen, geringes Risiko?

Wer auf diese fünf Fragen keine konkreten, belegbaren Antworten bekommt, sollte kein Projekt starten — bei keinem Anbieter, auch nicht bei uns.

Worte, die Sie von uns nicht hören werden

Ehrlichkeit ist bei uns keine Notlösung, sondern die Kernbotschaft. Deshalb verzichten wir bewusst auf Formulierungen, die technisch nicht haltbar sind:

- „Unsere KI halluziniert nicht.“ — Kein Sprachmodell kann das garantieren. Haltbar ist: Unbelegtes wird ausgewiesen oder blockiert statt erfunden ausgegeben.
- „Das System ist deterministisch.“ — Die Code-Teile sind es (gleiche Eingabe, gleiche Ausgabe); der KI-Schritt bleibt probabilistisch und wird kontrolliert.

- „Automatisch DSGVO-konform / zertifiziert sicher." — Konformität entsteht pro Einsatz aus Architektur, Vertrag und Prüfung, nicht aus einem Label.
- „Unhackbar", „military grade", Erfolgszahlen ohne Messung. — Wenn dafür kein konkreter, aktueller Nachweis vorliegt, sagen wir es nicht.

> AUF DEN PUNKT Wer selbst sagt, was das System NICHT kann, dem können Sie glauben, was es kann.

13 · Häufige technische Fragen

Wenn Ihre IT uns auf den Zahn fühlt — die Fragen und unsere Antworten, vorab schriftlich.

F **„Welches Modell, welches Kontextfenster?"**

A Im Produkt wählen wir Modelle nach Aufgabe, Schutzbedarf, Qualität und Kosten — gestaffelt statt pauschal. Das konkrete Modell, die Region und das Kontextfenster sind Teil der technischen Architekturentscheidung, nicht ein Marketing-Slogan.

F **„Wie verhindert ihr Halluzinationen konkret?"**

A Drei Schichten: Zitat-Zwang gegen echte Quellen (RAG), Erzeuger ist nicht Prüfer (unabhängige Falsifikation), mechanische Tore im Code. Was keinen Beleg hat, kommt nicht durch — es wird sichtbar als offen markiert.

F **„Deterministisch oder KI? Wie koordiniert ihr mehrere Agenten?"**

A Wo Verlässlichkeit zählt, deterministische Logik ohne Modell; KI nur fürs Urteil. Koordination über einen Koordinator plus gemeinsame Ablage, isolierte Kontexte — kein freies Agenten-Geplauder.

F **„Wie testet/evaluiert ihr die Qualität?"**

A An kritischen Stellen mit automatisierten Test-Suiten, Gegenbeispielen, Sicherheits-/Angriffs-Tests und Regressions-Prüfungen. Für eine konkrete Aussage zeigen wir Testfall, Ergebnis und aktuellen Bericht — keine Zahl aus dem Bauch.

F **„Kosten & Latenz?"**

A Wir staffeln Modelle nach Aufgabe: teuer nur, wo ein starkes Urteil nötig ist. Budgetrahmen, Grenzen und Wartezeit werden pro Ablauf gemessen und bewusst festgelegt; reine Logik ist meist günstiger und schneller.

F **„Vendor-Lock-in — was, wenn der Anbieter ausfällt?"**

A Wir planen Schnittstellen und Modell-Konfiguration so, dass ein Anbieterwechsel möglich bleibt. Ob ein konkreter Ablauf ohne Umbau wechseln kann, zeigen wir im technischen Review — „kein Lock-in" ist kein pauschales Versprechen.

F **„Fine-Tuning oder Prompting?"**

A Prompting, geladene Anleitungen und RAG — kein Nachtrainieren. Günstiger, schneller anpassbar, und wir behalten die Kontrolle über die Quellen. Fine-Tuning nur, wenn ein Fall es wirklich verlangt.

F **„Observability / Protokolle?“**

A Wir definieren pro Ablauf, welche Ereignisse, Entscheidungen und Freigaben nachvollziehbar sein müssen und welche personenbezogenen Daten dabei maskiert werden. Revisionssicherheit wird nur behauptet, wenn sie für den konkreten Prozess belegt ist.

F **„Auf welchem Betriebssystem läuft das? Bei uns oder auf einem Server?“**

A Auf der Nutzerseite ist es eine Web-Anwendung: kein Installer, keine lokale Datenbank — moderne Browser genügen; Windows, macOS und Linux sind nur unterschiedliche Türen zum selben System. Die geschützte Verarbeitung läuft serverseitig: Geheimnisse, Datenbankzugriffe und Modell-Schlüssel gehören in den Maschinenraum, nie in den Browser.

F **„Ist das sicher genug für Produktion?“**

A Eine Demo beweist Design und einzelne Kontrollen, nicht Produktionsreife. Vor echten Kundendaten gehören Datenschutzprüfung, Rechte-Matrix, Backup/Restore-Test, Monitoring, Incident-Prozess, Löschkonzept, Schwachstellenmanagement und je Schutzbedarf ein unabhängiger Penetrationstest dazu — das sagen wir unaufgefordert.

> AUF DEN PUNKT Die stärkste Antwort auf jede Testfrage: ehrlich, mit Prinzip — und „die konkrete Umsetzung zeigen wir offen im technischen Review“. Ein kritischer Technik-Mensch vertraut dem mehr als einem, der auf alles sofort eine glatte Antwort hat.

14 · Die Begriffe in Bildern

Ein starkes Bild pro Begriff — Bilder bleiben hängen, Definitionen nicht.

AGENTEN-NETZWERK	Dirigent + Orchester — jeder spielt seinen Part, zusammen wird es ein Stück.
ERZEUGER ≠ PRÜFER	Vier-Augen-Prinzip: Wer bucht, gibt nicht frei — und der Prüfer ist ein Miesepeter.
WERKZEUGE / FUNCTION-CALLING	Beschrifteter Werkzeugkasten: Der Agent zeigt, der Mensch dreht bei Riskantem selbst.
RAG / ZITAT-ZWANG	In die Bibliothek gehen und die Seite mit dem Finger zeigen — sonst gilt es nicht.
HALLUZINATION	Der Praktikant, der überzeugend Quatsch erzählt, wenn man ihn lässt.
PII-MASKIERUNG	Der schwarze Filzstift des Zensors, bevor das Blatt rausgeht.
SSRF	Die Poststelle überreden, geheime interne Akten zu holen — Schutz = sie weigert sich.

PROMPT-INJECTION	Der Zettel des Trickbetrügers: „Der Chef sagt…" — der Bote gehorcht nur dem echten Chef.
RBAC	Werksausweis, der nur bestimmte Türen öffnet.
HUMAN-IN-THE-LOOP	Zwei Schlüssel für den Tresor.
MULTI-TENANT	Abgeschlossene Schließfächer im selben Haus.
AUDIT-TRAIL	Notariatsbuch, jede Seite versiegelt die vorige.
OFFEN VS. GESCHLOSSEN (MODELL)	Eigenes Auto vs. Taxi mit fremdem Fahrer.
SOUVERÄNITÄT IN STUFEN	Eine Leiter — EU-Anbieter, dedizierter Server, eigenes Haus.
LINEARITÄTSFALLE	Eine Kette von Zufallsabzweigungen, die sich wie ein gerader Weg anfühlt.

15 · Glossar

Jeder Begriff in einem Satz — zum schnellen Nachschlagen.

AGENTIC AI	KI, die nicht nur antwortet, sondern handelt: Ein Team von Agenten zerlegt, arbeitet parallel, führt geprüft zusammen.
KI-AGENT	Programm, das eine Aufgabe eigenständig in Schritten erledigt und Software bedient — unter Aufsicht.
ORCHESTRATOR	Der koordinierende Agent: zerlegt, verteilt, führt geprüft zusammen.
LLM / SPRACHMODELL	Das „Gehirn", das Sprache versteht und erzeugt (z. B. Claude, Mistral).
TOKEN / KONTEXTFENSTER	Sprach-Häppchen / Kurzzeitgedächtnis des Modells (wie viel es auf einmal sieht).
PROMPT	Die Anweisung an das Modell.
FUNCTION-CALLING / WERKZEUG	Das Modell darf abgegrenzte Funktionen aufrufen (Termin anlegen, suchen) — ausgeführt vom Programm.
API	Eine kontrollierte digitale Serviceschalter-Schnittstelle: Software bittet andere Software um genau definierte Daten oder Aktionen.

OAUTH	Ein zeitlich und fachlich begrenzter Besucherausweis für eine externe Anwendung — nicht das Passwort des Mitarbeiters weitergeben.
WEBHOOK	Eine digitale Türklingel: Ein Fremdsystem meldet sofort, dass etwas passiert ist, statt dass wir ständig nachfragen.
SFTP	Ein verschlossener digitaler Briefkasten für Dateiaustausch zwischen Systemen.
OBJEKT-SPEICHER / STORAGE	Das digitale Archiv für Originaldateien. Die Datenbank hält dazu nur den Index und die Rechte.
SIGNED URL	Ein kurz gültiger Besucherpass zu genau einer privaten Datei — keine öffentliche Daueradresse.
RAG	Antwort erst in freigegebenen Quellen nachschlagen und mit Fundstelle belegen.
HALLUZINATION	Das Modell erfindet etwas plausibel Klingendes — Gegenmittel: Zitat-Zwang und Blockade.
ERZEUGER ≠ PRÜFER	Wer ein Ergebnis erzeugt, gibt es nicht selbst frei; eine unabhängige Instanz falsifiziert.
PII	Personenbezogene Daten (Name, E-Mail, IBAN, Gesundheit).
MASKIERUNG / PSEUDONYM / ANONYM	Unkenntlich / durch Kürzel ersetzbar / nicht mehr rückführbar.
SSRF	Angriff: den Server austricksen, interne/verbotene Adressen abzurufen. Schutz: geprüfte Ausgangs-Abrufe.
PROMPT-INJECTION	Versteckte Befehle in Texten, die den Agenten kapern sollen. Schutz: Inhalte als Daten behandeln, isoliert verarbeiten.
RBAC	Rollenbasierte Rechte: jedes Werkzeug an eine Rolle gebunden.
RLS	Zeilenbasierte Rechte in der Datenbank: Selbst eine korrekte Abfrage sieht nur Daten des erlaubten Arbeitsbereichs.
SECRET / ENV	Ein geheimer Schlüssel in der Server-Umgebung. Er darf nie im Browser, Screenshot oder Quellcode landen.
HUMAN-IN-THE-LOOP	Der Mensch bestätigt/korrigiert — die Maschine handelt nicht allein.
MULTI-TENANT / ISOLATION	Ein sauberes Mandantenmodell verhindert, dass Mandant A Daten von Mandant B sieht.

AUDIT-TRAIL	Nachvollziehbares Protokoll: wer, wann, was. Revisionssicherheit braucht zusätzlich konkrete Unveränderbarkeits- und Aufbewahrungsnachweise.
DSGVO / EU AI ACT	EU-Datenschutz / EU-KI-Gesetz (Aufsicht, Protokolle, Doku für wichtige Systeme).
ON-PREMISE / SELF-HOSTED	Betrieb im eigenen Haus — die Daten verlassen es nicht.
OFFENE GEWICHTE (OPEN WEIGHTS)	Modell, dessen Innenleben einsehbar ist und das man selbst betreiben kann (Llama, Mistral, Qwen).
SOUVERÄNITÄT	Kontrolle darüber, wo Daten liegen und wer rechtlich Zugriff hat.
AVV / SCC / CLOUD ACT	Datenschutz-Verträge / EU-Standardklauseln / US-Zugriffsgesetz.
GUARDRAILS	Eingebaute Schranken, die verhindern, dass der Agent Unerlaubtes tut.
TEMPERATUR	Regler für Zufall/Kreativität des Modells — für Prüf-Aufgaben bewusst niedrig.
REIFEGRAD	Unser fünfstufiges Vokabular von Konzept bis Betrieb — damit ein Nachweis immer sagt, was er belegt.

DER MERKSATZ ZUM SCHLUSS

Nicht die klügste KI gewinnt, sondern die, der die Verantwortlichen vertrauen können. Der Schwarm liefert Perspektiven — der Mensch entscheidet — alles bleibt nachweisbar.

Stand 16.07.2026. Technische Aussagen sind aus dem aktuellen internen Review abgeleitet. Datenschutz-, Vertrags- und Compliance-Aussagen gelten nie pauschal: konkrete Architektur, Anbieter-Produkt, Vertrag und Einsatz werden je Kundenfall geprüft. Dieses Dokument ist ein Lehrdokument — keine Rechtsberatung und kein Angebot. Kontakt: Neon Arc · info@neonarc.com