✱ GENERATIVE AI FÜR DEN MITTELSTAND · 2026

GenAI als Wettbewerbs-vorteil.

Ohne Kosten- und Compliance-Fallen. Ein Whitepaper für Entscheider, die 2026 nicht mehr warten, sondern fundiert handeln wollen – mit einem klaren Framework, einem Team, das du heute schon hast, und in 90 Tagen.

Ein Whitepaper von _app:soluts/*

Der kürzeste Weg von Idee zu Trusted-AI-Release.

GenAI scheitert 2026 nicht an Ideen. Sondern an fehlender Risikoklassifizierung, unklarer Architektur und nicht berechneten Betriebskosten.

Wir hören in jedem zweiten Erstgespräch dieselbe Frage: „Wir wollen GenAI einsetzen – aber wir wissen nicht, wo wir anfangen sollen, was uns rechtlich auf die Füße fallen kann und was es am Ende kostet."

Das ist die richtige Frage. Antworten darauf gibt es zu wenige ehrliche und zu viele Hype-getriebene. Dieses Whitepaper schließt die Lücke zwischen Strategiepapier und Architekturzeichnung – für Entscheider, die ein GenAI-Feature wirklich in Produktion bringen müssen.

Wir betrachten GenAI nicht als isolierte Modellfrage. Für uns ist GenAI ein Produkt- und Betriebsfeature: UX, Datenzugriff, Architektur, Monitoring, Kosten und Compliance müssen zusammenpassen – sonst hält das Feature den ersten echten Tag im Betrieb nicht durch.

Drei Themen greifen ineinander: **Recht** (EU AI Act, BFSG, Store-Policies), **Architektur** (RAG, Tool-Calling, Guardrails, Observability) und **Wirtschaftlichkeit** (Unit Economics, FinOps, Release-Engineering). Wer eines davon weglässt, baut entweder ein Demo oder ein Risiko.

Am Ende steht eine 90-Tage-Roadmap, die du mit dem Team umsetzen kannst, das du heute hast. Kein „KI-Stab", keine Doppelhierarchien – nur Klarheit über Scope, Verantwortung, Reihenfolge und Prüfschritte.

Viel Erfolg.
Das appsoluts-Team

Tobias Suchy
CO CEO / CTO · Backend Development

Wer 2026 GenAI in Produktion bringt, gewinnt nicht durch das beste Modell – sondern durch klare Risiko-klassen, belastbare Architektur und Kosten, die er kennt. Dieses Whitepaper zeigt, wie das in 90 Tagen geht – mit dem Team, das du heute hast.

FÜR WEN

Produkt- und Technologieverantwortliche, die GenAI-Features in echte Apps bringen wollen – wirtschaftlich sauber, rechtlich belastbar und ohne neue Vollzeitstellen. Sekundär: Legal, Security & Geschäftsführung.

GenAI ist kein Experiment mehr. Es ist Infrastruktur.

Nach dem Lesen kannst du deinen Use Case einordnen, die richtigen Architekturfragen stellen, Compliance- & Kostenrisiken früher erkennen und intern auf einer Sprache mit Product, Tech, Legal und Business sprechen.

01 · CHANCE

Produktnah, nicht abstrakt.

Der wirtschaftliche Hebel liegt nicht im Modell, sondern in konkreten Prozessen: weniger manuelle Prüfung, schnellere Entscheidungen, bessere Self-Service-Flows in Service, Sales und produktnahen Teams.

02 · RISIKO

Nicht nur EU AI Act.

Bußgelder bis **7 % des Jahresumsatzes** sind die sichtbare Spitze. Darunter liegen Haftung, teure Nacharbeit, eskalierende Betriebskosten und Vertrauensverlust bei Kunden und Aufsicht.

03 · ZEIT

90 Tage – wenn fokussiert.

Ein erstes Trusted-AI-Release ist in einem Quartal machbar – wenn Scope, Risikoklasse, Architektur und Go-Live bewusst eng gehalten werden. Mit dem Team, das du heute hast.

KERNTHESE

Die Frage ist nicht mehr „Sollen wir GenAI einsetzen?" Sie lautet: Wie setzen wir es so ein, dass wir Kontrolle behalten, Vertrauen gewinnen & wirtschaftlich profitieren.

01

DIE NEUE REALITÄT

Warum 2026 kein Jahr für Sandboxes ist.

KURZFASSUNG

Drei Kräfte verändern die Spielregeln gleichzeitig: ein verbindlicher EU AI Act, ein nicht-deterministisches Verhalten von KI-Systemen und verschärfte Store-Richtlinien von Apple & Google. Kein Bereich davon ist optional.

Vor zwei Jahren war GenAI ein Proof of Concept. Heute ist es ein Produktbestandteil.

Mit echten Nutzern, echten Daten und echten Haftungsfragen.

01 · RECHT

EU AI Act ist in Kraft.

Er unterscheidet nicht zwischen „wir testen das noch" und „wir betreiben das produktiv". Wer ein KI-System einsetzt, das Entscheidungen beeinflusst, ist regulatorisch relevant. Hochrisiko-Systeme müssen bis zum 2. August 2026 vollständig konform sein.

02 · TECHNIK

KI verhält sich nicht wie Software.

Gleiche Eingabe, unterschiedliche Ausgabe. Halluzinationen, Prompt Injection, Data Leakage – keine theoretischen Szenarien, sondern Failure Modes, die in der Praxis auftreten. Ohne Architektur, die das abfängt, ist ein stabiler Betrieb nicht möglich.

03 · DISTRIBUTION

Stores ziehen Policies nach.

Apple & Google verschärfen ihre Store-Richtlinien. KI-Apps müssen On-Device-Processing, Content-Kennzeichnung und Datenschutzanforderungen aktiv nachweisen. Wer das ignoriert, riskiert die Distribution.

KEY TAKEAWAY

2026 ist kein Jahr mehr für Sandbox-Experimente. Wer GenAI produktiv denkt, muss Risiko, Architektur und Wirtschaftlichkeit von Anfang an gemeinsam bewerten.

Wer handeln muss – und warum jetzt.

Zwei Rollen sind unmittelbar betroffen. Entscheider-Teams in beiden Rollen müssen sicherstellen, dass die technologische Roadmap nicht von regulatorischen Fristen überholt wird.

ANBIETER · PROVIDER

Unternehmen, die KI-Anwendungen entwickeln oder unter ihrem Namen vertreiben. Auch dann, wenn sie außerhalb der EU sitzen – sobald EU-Nutzer betroffen sind, gilt das Gesetz.

BETREIBER · DEPLOYER

Alle Organisationen, die KI-Lösungen im Geschäftsalltag einsetzen – für Kundenanfragen, interne Auswertungen, Prozessautomatisierung.

AI vs. GenAI – kein Akademiker-Detail.

Die Unterscheidung hat direkte Auswirkungen auf Compliance & Architektur. Der EU AI Act behandelt beide. In diesem Whitepaper fokussieren wir auf GenAI – weil dort die größten Chancen und die komplexesten Governance-Anforderungen liegen.

AI · ARTIFICIAL INTELLIGENCE

Oberbegriff. Mustererkennung, Vorhersagen, Entscheidungs-unterstützung. In Apps oft unsichtbar im Hintergrund – Empfehlungen, Fraud Detection, Bildklassifizierung.

GENAI · GENERATIVE AI

Teilbereich: Modelle, die neue Inhalte erzeugen – Text, Bilder, Audio, Code. In Apps für Nutzer sichtbar – Chatbots, Zusammenfassungen, Natural Language Interfaces.

02

RISIKOKLASSEN

Die wichtigste Entscheidung vor der ersten Zeile Code.

KURZFASSUNG

Diese Klassifizierung bestimmt alles: Dokumentationstiefe, Time-to-Market, Compliance-Aufwand und Budgetbedarf. Eine falsche Einschätzung am Anfang erzeugt massive Korrekturschleifen.

Vier Stufen, eine Logik: Je höher das Risiko, desto strenger die Regeln.

STUFE 4 · VERBOTEN

Inakzeptables Risiko.

Systeme, die Grundrechte gefährden: Social Scoring, biometrische Massenüberwachung, Manipulation vulnerabler Gruppen. Marktzulassung ausgeschlossen – kein Spielraum.

STUFE 3 · MAXIMALE COMPLIANCE

Hohes Risiko.

KI in kritischer Infrastruktur, Bildung (Prüfungsbewertung), HR (CV-Screening), Kreditprüfung, Rechtspflege. CE-Kennzeichnung, Qualitätsmanagement, lückenlose Dokumentation. Deadline 2. August 2026.

STUFE 2 · TRANSPARENZPFLICHT

Begrenztes Risiko.

Chatbots, KI-Kundenberatung, Bildgeneratoren, Deepfakes. Nutzer müssen aktiv darüber informiert werden, dass sie mit einer KI interagieren. Technisch lösbar – aber nicht vergessen.

STUFE 1 · KEINE SONDERREGELN

Minimales Risiko.

Spam-Filter, KI-Suche, Übersetzungsdienste, Empfehlungssysteme. DSGVO gilt weiterhin. Freiwillige Verhaltenskodizes werden empfohlen.

Julian Bansen
Lead App Developer

Risikoklasse ist für uns keine Compliance-Frage, sondern ein Architekturmuster. Sie entscheidet, wie wir testen, deployen und monitoren – nicht erst beim Go-Live, sondern ab dem ersten Commit.

PRAXIS-TIPP

Wir integrieren die Risikoklassifizierung direkt in die Product Discovery – bevor die erste Architekturentscheidung getroffen wird. Manchmal lässt sich durch gezieltes „Context-Limiting" eine niedrigere, kosteneffizientere Risikoklasse erreichen.

Was der EU-AI-Act von Hochrisiko-Systemen verlangt.

Fällt euer Use-Case in die rote Zone, gelten zehn konkrete Pflichten – verankert in Artikel 9 bis 49 des EU-AI-Acts. Diese Bausteine müssen nachweisbar erfüllt sein, bevor das System produktiv geht. Keine Empfehlungen, sondern gesetzliche Voraussetzung für den Markteintritt.

ART. 9

Risikomanagementsystem

Identifikation & Minderung von Risiken über den gesamten Lebenszyklus.

ART. 10

Data Governance

Trainings-, Validierungs- und Testdaten auf Bias und Repräsentativität prüfen.

ART. 11

Technische Dokumentation

Funktionsweise und Architektur für Aufsichtsbehörden offenlegen.

ART. 12

Log-Aufbewahrung

Automatische Protokollierung für Rückverfolgbarkeit von Entscheidungen.

ART. 13

Gebrauchsanweisung

Klare Grenzen und Genauigkeits-angaben für Betreiber bereitstellen.

ART. 14

Human Oversight

Menschen können die KI jederzeit überwachen, korrigieren & abschalten.

ART. 15

Genauigkeit & Robustheit

Fehlertoleranz und Widerstand gegen Manipulationsversuche nachweisen.

ART. 15

Cybersicherheit

Schutz vor Datenverlust und externer Kontrolle.

ART. 43

Konformitätsbewertung

Offizielles Prüfverfahren vor Markteintritt.

ART. 49

EU-Registrierung

Eintragung in die öffentliche EU-Datenbank für Hochrisiko-KI.

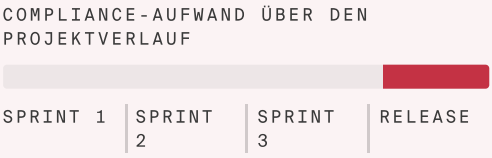
Audit-Readiness ist keine Einmal-Aktion. Sie ist ein Betriebsmodus.

Die EU-AI-Act-Anforderungen gehören in **Sprint 1**, nicht in die letzten zwei Wochen vor Release. Wer das Inversemodell fährt, spart bei jedem Folge-Release einen ganzen Compliance-Zyklus.

FALSCH GEDACHT

Compliance kommt am Ende.

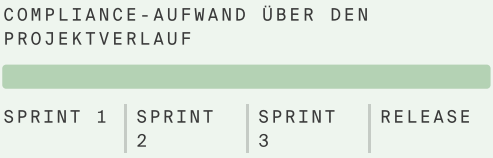
Die 10 Pflichten werden am Projektende „angeflanscht" – als nachträglicher Compliance-Layer, kurz vor Go-Live. Ergebnis: technische Schulden, teure Refactorings, verschobene Releases.



RICHTIG GEMACHT

Compliance ist eine Produktanforderung.

Die zehn Pflichten sind in **Architektur, Backlog und Definition of Done** verankert. Wer so baut, hat keinen separaten „Audit-Sprint" mehr – das System ist jederzeit prüfungsbereit.



Der Compliance-Flow: 5 Schritte vom Use-Case zum auditierbaren Release.

Dieser Ablauf ist kein einmaliges Projekt, sondern ein **wiederholbarer Prozess**. Jede neue KI-Funktion durchläuft ihn von Anfang an – und bei jedem Modell- oder Datenquellen-Update von vorn. Schritt 5 ist nicht das Ende, sondern der Beginn des nächsten Durchlaufs: Wer das System ändert, bewertet die Konformität neu.

01 **Use-Case identifizieren**
Welches Problem löst die KI – und wer ist betroffen?

02 **Risikoklasse einordnen**
Inakzeptabel · Hoch · Begrenzt · Minimal.

03 **Maßnahmen festlegen**
Pflichten je Klasse als konkrete Tasks definieren.

04 **Konformität bewerten**
Interne Prüfung, ggf. externes Audit.

05 **Release & Monitoring**
Soft Launch, Tracing, Re-Evaluation bei Updates.

Governance – ohne neue Vollzeitstellen.

Niemand erwartet fünf neue Headcounts für das erste GenAI-Projekt. Was zählt: klare Verantwortlichkeiten. In einem schlanken Setup trägt der CTO oft gleichzeitig den Hut des Technical Leads und des Security Leads. Governance entsteht nicht aus Org-Charts, sondern aus **Routinen**.

Vier Maßnahmen für den Start

01

AI-Stammtisch

30 Min/Woche, Produkt + Technik + Recht.

02

Schnittstellen-Check

Legal ab echten Nutzerdaten / Medium Risk.

03

Fehlerkultur

GenAI macht Fehler. Rückendeckung vom Business Lead.

04

Klein starten, groß dokumentieren

AI-System-Logbuch ab Tag 1.

Warte nicht auf das perfekte Setup. Bestimme jemanden mit Ownership – und wachse mit den Anforderungen. Governance soll Innovation beschleunigen, nicht blockieren.

03

DAS APPSOLUTS FRAMEWORK

Trusted AI by Design.

KURZFASSUNG

Eine produktive KI-Funktion zerfällt in drei Teile, die getrennt gedacht und getrennt verantwortet werden müssen: **das Modell** (was die KI kann), **die Datenquelle** (was die KI weiß) und **die Aktionslogik** (was die KI tun darf). Wer das vermischt, baut Blackboxes. Wer es trennt, baut Systeme, die du belegen, prüfen und verbessern kannst.

Deployment-Entscheidung.

Die erste Architekturentscheidung ist der Ort der Inferenz. Kein Ansatz ist pauschal besser – nur passend oder unpassend.

KRITERIUM	CLOUD (SAAS)	ON-DEVICE	HYBRID
Leistung	Maximale Reasoning-Power	Limitiert durch Chip	Beste aus beiden Welten
Datenschutz	Daten verlassen das Gerät	Höchster Standard	Kontextabhängiges Routing
Latenz	Netzabhängig	Nahezu null, offline-fähig	Flexibel steuerbar
Kosten	Variable Token-Kosten	Einmalige Entwicklung	Mischkalkulation

Copilot oder Autopilot?

COPILOT — STANDARD 2026

Human-in-the-Loop

Die KI bereitet vor, der Mensch entscheidet. Minimiert Haftungsrisiken und stärkt das Nutzervertrauen.

ENTSCHEIDER **Mensch**

AUTOPILOT — HIGH STAKES

Autonomous Agency

Das System handelt selbstständig in Backend-Systemen. Erfordert „Policy-Enforcement-Points" für Preislimits, Berechtigungen und Rollback.

ENTSCHEIDER **System**

RAG 2.0 – Halluzinationen technisch verhindern.

Anstatt das Modell aus dem Gedächtnis (Trainingsdaten) antworten zu lassen, versorgen wir es im Moment der Anfrage mit exakten Fakten aus kontrollierten Datenquellen.

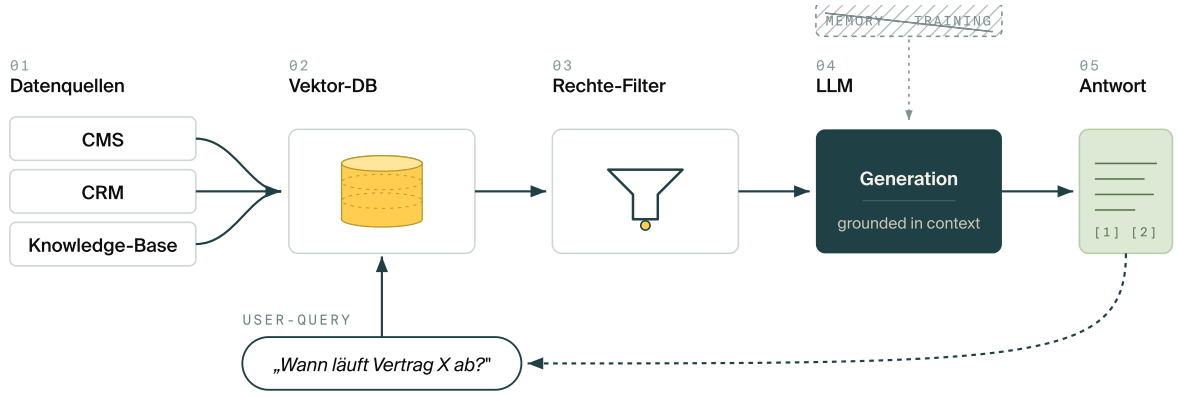


Abb. 1 – RAG-Architektur. Die Antwort entsteht ausschließlich aus kontrolliertem Kontext, nicht aus dem Modellgedächtnis.

Drei Voraussetzungen für ein „trusted" RAG

01

Retrieval-Qualität

Findet das System die Passage, die wirklich zählt? Falsche Dokumente sind kein Sicherheitsnetz – sie sind ein hübscher Umweg zur falschen Antwort.

02

Aktualität

Wie schnell landen Änderungen im Index? Bei wöchentlichen Releases muss das Wissen mitziehen – sonst antwortet die KI mit veralteten Informationen.

03

Rechte & Scopes

Inhalte vor dem Prompt filtern. Das Modell darf nur das wissen, was der jeweilige Nutzer sehen darf. OWASP adressiert genau das als Disclosure-Kategorie.

Tool-Calling: Von der Antwort zur Aktion.

Wenn die KI nicht nur informieren, sondern handeln soll – Adresse im CRM ändern, Ticket erstellen, Termin buchen – übersetzt sie natürliche Sprache in exakte technische Befehle (JSON), die das Backend ausführt.

01 · SCHEMA

Safe Defaults.

Jede API-Anbindung läuft über strikte Schemata. Die KI schlägt vor, das Backend validiert. Keine offenen Befehle, keine improvisierten Aufrufe – Aktionen werden nur ausgeführt, wenn sie typgeprüft sind.

02 · BESTÄTIGUNG

Human-in-the Loop.

Löschvorgänge, Zahlungen und externe Kommunikation brauchen eine **Bestätigungsschranke**. Die KI bereitet vor – der Mensch klickt „Bestätigen". Verantwortung bleibt dort, wo sie hingehört.

03 · FEHLER

Error Handling.

Wenn die API 404 meldet, muss die KI den Fehler verstehen und transparent kommunizieren – keine erfundenen Meldungen, keine stillen Fallbacks. Sichtbare Fehler sind besser als unsichtbare.

KERNTHESE

Tool-Calling ist die Brücke zwischen Sprache und System. Wer Schemata und Bestätigungsschranke früh definiert, baut Vertrauen in jede einzelne Aktion ein & muss später nicht zurückbauen.

Monitoring & Observability: Lückenlose Beweisspur.

Der EU AI Act fordert Rückverfolgbarkeit. Wir implementieren lückenloses Tracing – und machen es damit zum Wettbewerbsvorteil, nicht zur Bürde.

01 · AUDIT

Prompt-Logging.

DSGVO-konform dokumentieren, welcher Input zu welchem Output führte. Auditierbar, ohne personenbezogene Daten preiszugeben – mit definierten Retention-Fristen und rollenbasierten Zugriffen.

02 · QUALITÄT

Evals & Benchmarks.

Faithfulness unter **90 %**: System schlägt Alarm – bevor der Nutzer es bemerkt. Qualität wird messbar, nicht nur spürbar.

03 · KOSTEN

Cost-Monitoring.

Token-Kosten pro User-Session in Echtzeit. Ineffiziente Loops fressen sonst das Budget – sichtbar erst, wenn die Rechnung kommt.

KERNTHESE

Observability ist kein Compliance-Aufwand, sondern die Beweisspur, die jede produktive KI-Anwendung tragen muss. Wer sie erst nach dem Go-Live nachrüstet, baut zweimal.

Trust-Features: Vertrauen entsteht nicht durch Magie – sondern durch Transparenz.

Diese vier Trust-Features kosten wenig Engineering – und verändern messbar, wie Nutzer eine KI-Funktion erleben. Sie sind kein UX-Schmuck, sondern die **Schnittstelle**, an der ein KI-Produkt beweist, dass es Verantwortung an den Menschen zurückgibt – nicht abnimmt.

01 QUELLE

Source Citation.

Belege für KI-Aussagen mit Deep-Links zu Originalquellen.

„Ich kann das selbst prüfen."

02 SICHERHEIT

Confidence Scores.

Anzeige, wie sicher sich die KI bei einer Aussage ist – inklusive Unsicherheits-bereiche.

„Ich weiß, wann ich vorsichtig sein muss."

03 KONTROLLE

Undo-Buttons.

KI-Aktionen rückgängig machen – vor allem bei Tool-Calling-Resultaten in CRM, Ticket- und Mail-Systemen.

„Ich behalte die Kontrolle."

04 LERNEN

Feedback-Loops.

Daumen hoch/runter mit kurzem Kontextfeld. Speist Evals, schärft Prompts, identifiziert systematische Schwächen.

„Ich verbessere das System aktiv mit."

04

BUSINESS OPERATIONS

Wirtschaftlichkeit ohne Illusionen.

KURZFASSUNG

Ein exzellentes KI-Feature ist wertlos, wenn die Betriebskosten den Customer Lifetime Value auffressen. Hier verlassen wir die technologische Euphorie und schauen auf die Zahlen.

Was kostet ein „Successful Task"?

GenAI verursacht variable Kosten pro Interaktion. Der Erfolgsmaßstab ist kein API-Aufruf – sondern ein abgeschlossenes Nutzerziel. Wenn ein KI-System ein Support-Ticket verhindert, entsteht Business Value nur dann, wenn die Inferenzkosten deutlich unter den eingesparten Servicekosten liegen.

Eine Interaktion gilt erst dann als Successful Task, wenn drei Bedingungen zusammen erfüllt sind:

User-Ziel erreicht

Die ursprüngliche Aufgabe ist nachweislich gelöst.

Keine Eskalation

Kein Hand-off an menschlichen Support, kein Refund.

Auditierbar

Antwort und Trace sind nachvollziehbar dokumentiert.

WARNUNG · FINOPS

Ohne striktes FinOps-Management liegen GenAI-Projektkosten regelmäßig **um ein Vielfaches** über der ursprünglichen Planung. Das ist kein Randphänomen – das ist die Norm bei ungesteuertem Rollout.

Budgets gehören in die Produkt-anforderung – nicht ins Add-on.

GenAI-Kosten skalieren mit Nutzung, nicht mit Lizenzen. Ohne aktive Steuerung kippt die Unit Economics binnen Wochen. Zwei Hebel halten das Budget produktnah.

01 · CAPS

Harte Limits

Budgets auf **User- und Feature-Ebene**. Ein unkontrollierter Agent darf niemals das Cloud-Budget leerräumen – Caps gehören in die Architektur, nicht ins Monitoring.

02 · ATTRIBUTION

Cost Attribution

Welches Feature verursacht welche Kosten? Eine präzise Aufschlüsselung pro **Successful Task** identifiziert „teure Spielereien" ohne echten Impact – bevor sie die Marge angreifen. Benchmark: Ein vermiedenes Support-Ticket (Deflection) spart laut **Gartner 5–15 €** – die Inferenzkosten pro Task müssen deutlich darunter liegen.

KEY TAKEAWAY

Wer Cost Caps und Cost Attribution nicht ab Sprint 1 mitdenkt, finanziert Experimente, die das Geschäftsmodell nie tragen wird.

KI-Modelle verhalten sich nicht deterministisch.

Ein Provider-Update kann die Antwortqualität über Nacht verändern – ohne Ankündigung. Release-Discipline ist die einzige Versicherung gegen still degradierende Systeme im Produktivbetrieb.

01 · VERSIONIERUNG

Prompts wie Code behandeln.

Jede Prompt-Änderung wird versioniert und gegen die Vorgängerversion A/B-getestet. Was nicht messbar besser ist, geht nicht in Produktion.

02 · SAFE ROLLOUT

Neue Modelle via Feature Flag.

Provider-Wechsel und Modell-Updates erreichen zuerst **1 %** der Nutzer. Erst nach Qualitäts- und Kosten-Check skaliert der Rollout – mit jederzeit aktivem Kill-Switch.

03 · FALLBACK

Safe Defaults statt offener Flanke.

Bei Degradation, Timeout oder Guardrail-Verstoß automatischer Rückfall auf eine geprüfte Vorgängerversion oder eine deterministische Antwort – nie ein leeres Feature.

KEY TAKEAWAY

Ein KI-Release ist nie „fertig". Ohne Versionierung, Safe Rollout und Fallback ist jedes Provider-Update ein Glücksspiel mit der Antwortqualität.

05

DIE 90-TAGE-ROADMAP

Von der Idee zum Trusted-AI-Release.

KURZFASSUNG

Sprint in der Vorbereitung – Marathon im Betrieb. Der Plan ist modular: je nach Risikoklasse skalierst du die Intensität der Prüfschritte.

Drei Phasen, ein Quartal.

PHASE 01 TAG 01-30

Assessment

Verstehen, wo Risiko & Wert liegen.

ARBEITSPAKETE

- Use-Case-Inventur
- Risikoklassifizierung
- Vendor Due Diligence
- AI Literacy
- Business Case

DELIVERABLE

Use-Case-Register & freigegebener Business Case.

PHASE 02 TAG 31-60

Architecture

Bauen, was Phase 3 freigeben kann.

ARBEITSPAKETE

- RAG-Infrastruktur
- Guardrails & Tool-Calling
- BFSG-Design
- Observability-Setup
- FinOps-Setup

DELIVERABLE

Lauffähiger Prototyp mit Telemetrie & Kostenlimits.

PHASE 03 TAG 61-90

Go-Live

Belegen, dass das System sicher ist.

ARBEITSPAKETE

- Red Teaming
- Technische Dokumentation
- Transparenz-Check
- Soft Launch
- EU-Konformitäts-erklärung

DELIVERABLE

Trusted-AI-Release + signiertes Konformitätsdossier.

KEY TAKEAWAY

90 Tage reichen – wenn die Phasen nicht parallel laufen, sondern **aufeinander aufbauen**. Wer in Phase 1 klassifiziert, baut in Phase 2 nur das, was Phase 3 freigeben kann.

Drei Haken – bevor du den Release-Button drückst.

Wenn dein Team an diesen drei Stellen einen Haken setzen kann, bist du nicht „fertig" – aber ***release-fähig***.

Rechtssicherheit, rechtlich abgesichert.

→ KAP. 02 RISIKOKLASSEN

Wir wissen, in welche **EU-Risikoklasse** unser System fällt und haben die Transparenzpflichten sichtbar im UI erfüllt – inkl. BFSG-Barrierefreiheit für Endkunden.

> ***„Kennen wir unsere Klasse – und ist sie im UI ablesbar?"***

Architektur, technisch im Griff.

→ KAP. 03 TRUSTED AI

Halluzinationen sind durch **RAG** technisch abgesichert, jeder „Successful Task" wird mit Evals und Tracing überwacht. Failure Modes brechen sichtbar, nicht still.

> ***„Haben wir den Output im Griff – und sehen wir, wenn nicht?"***

Betrieb, zukunftssicher aufgestellt.

→ KAP. 04 BUSINESS OPS

Der **Modell-Provider** lässt sich austauschen, ohne das System neu zu bauen. Kosten pro Successful Task sind messbar, Budgets in der Produktanforderung verankert.

> ***„Können wir wechseln – und kennen wir den Preis?"***

RELEASE-REIFE

Drei Haken sind keine Garantie – aber die **Mindestschwelle**, unter der ein produktives KI-Release rechtlich, technisch und betrieblich nicht trägt. Alles darunter ist ein ***Prototyp mit Kundenkontakt***.

Lass deinen Use Case im GenAI Architektur-Check einordnen.

Bevor du Budget bindest, Modelle vergleichst oder einen Piloten startest: Wir schauen 60 Minuten gemeinsam auf deinen Use Case und sagen dir ehrlich, ob und wie er sich umsetzen lässt.

01

Risikoeinordnung

Grobe EU-AI-Act-Klasse für deinen Use Case.

02

Architektur-Skizze

Erste Einschätzung zu RAG, Tools, Guardrails, Daten.

03

Kosten- & Compliance-Risiken

Wo es teuer, langsam oder rechtlich heikel wird.

04

Drei nächste Schritte

Konkret, priorisiert, ohne Beratungsprosa.

Christian Blank

CEO & Business Development

SERVICE

GenAI Architektur-Check 60 Min. · kostenlos · remote oder vor Ort

E-MAIL

christian@appsoluts.de

Betreff „GenAI Architektur-Check"

TERMIN

Termin buchen

Direkt online

07

FAZIT & AUSBLICK

Trusted AI ist eine Haltung – keine Funktion.

KURZFASSUNG

Sechs Kapitel, ein roter Faden: **GenAI ist Infrastruktur, kein Feature**. Was daraus bleibt – und was als nächstes auf dich zukommt: autonome Agenten, schärfere EU-Pflichten und ein Markt, der nicht mehr das schnellste, sondern das prüfbarste Modell belohnt.

Fazit: Zwei Erkenntnisse fürs nächste Jahr.

Modelle wechseln, Frameworks veralten, Anbieter konsolidieren. Zwei Prinzipien aus diesem Whitepaper überdauern das alles – weil sie nicht am Modell hängen, sondern an der Architektur drumherum.

ARCHITEKTUR · TRENNEN

Trusted AI trennt was die KI **kann**, **weiß** und **tun darf**. RAG, Evals und Tracing machen jeden Output prüfbar – über jede Modellgeneration hinweg.

WIRTSCHAFTLICHKEIT · WECHSELBAR

Kosten pro Successful Task entscheiden über Tragfähigkeit. Wer den **Provider tauschen** kann, ohne das System neu zu bauen, bleibt rentabel und unabhängig.

Ausblick: Zwei Entwicklungen, die 2026 zählen.

Der nächste Schritt für Trusted AI passiert nicht im Modell, sondern drumherum: Systeme werden autonomer, Regulierung wird konkreter. Wer das jetzt einplant, baut keine Notlösungen unter Zeitdruck.

AGENTEN · VOM CHAT ZUR AKTION

Der Markt verschiebt sich von Antwort-Bots zu **autonomen Agenten**, die mehrere Tools nacheinander auslösen. Human Oversight und sichtbare Failure-Modes werden Pflicht – nicht Kür.

REGULIERUNG · AUGUST 2026

Der EU AI Act erreicht volle Anwendung für **Hochrisiko-Systeme**. Wer Risikoklasse, Dokumentation und Monitoring jetzt aufsetzt, ist rechtzeitig – nicht zu spät.

Über _app:soluts/*

Wir entwickeln seit April 2017 Apps, Web- und Backend-Systeme für KMU, Konzerne und Start-ups. Unser Fokus liegt auf Software, die langlebig trägt – nicht auf Demos, die nach dem ersten Sprint vergessen sind.

Wir verbinden Beratung, UX/UI-Design, Entwicklung, Testing und Betrieb in einem Team. Das spart Übergaben und macht Verantwortung sichtbar: Wer baut, betreut auch. Wer berät, kennt den Code. Das Ergebnis sind Produkte, die du in fünf Jahren noch erweitern und betreiben kannst – ökonomisch wie ökologisch.

Trusted AI ist für uns kein Trend, sondern die logische Fortsetzung dieser Haltung: Komplexität gehört unter Kontrolle, bevor sie produktiv geht. Genau dort setzen wir an.

GEGRÜNDET

April 2017 · 8+ Jahre kontinuierlich.

TEAM

8+ Köpfe aus Engineering, Design & Produkt.

PROJEKTE

100+ ausgeliefert, 50+ Partner aus KMU, Konzern & Start-up.

SCHWERPUNKTE

Apps · Web · Backend · Trusted AI.

WERTE

Qualität · Innovation · Verantwortung · Code Green.

KONTAKT

appsoluts.de

VISION

Wir verfolgen die Vision, **nachhaltige, innovative & gesellschaftlich relevante Software** zu entwickeln.

08

GLOSSAR & QUELLEN

Begriffe, mit denen alles steht und fällt.

KURZFASSUNG

EU AI Act, RAG, Evals, Human-in-the-Loop – Trusted AI lebt von **präzisen Begriffen**. Auf der folgenden Seite stehen die wichtigsten kurz erklärt, plus die Quellen, auf denen dieses Whitepaper aufbaut.

Glossar I.

Agentic AI	KI-Systeme, die mehrstufige Aufgaben autonom planen und über Tool-Calls ausführen – zentrale Entwicklung 2026.
BFSG	Barrierefreiheitsstärkungsgesetz. Verpflichtend für digitale Produkte ab Juni 2025.
EU AI Act	EU-Verordnung zur Regulierung von KI-Systemen. In Kraft seit 2024, stufenweise anwendbar bis 2026/2027.
Evals	Automatisierte Testsuiten, die Qualität, Faktentreue und Sicherheit von KI-Antworten kontinuierlich messen.
FinOps	Methodik zur Steuerung und Optimierung von Cloud- und KI-Kosten – technisch und betriebswirtschaftlich.
Guardrails	Technische/regelbasierte Mechanismen, die Eingaben und Ausgaben eines KI-Systems begrenzen.
Halluzination	KI-Verhalten, bei dem das Modell faktisch falsche, aber plausibel klingende Inhalte erzeugt.
Hochrisiko-System	KI-Anwendung mit verpflichtenden Auflagen nach EU AI Act – etwa in Personal, Bildung, Kreditvergabe, Justiz oder kritischer Infrastruktur.
Human-in-the-Loop	Architekturprinzip, bei dem Menschen kritische Aktionen explizit bestätigen – bevor die KI handelt.

Glossar II.

LLM	Large Language Model. Statistisches Sprachmodell, das aus Trainingsdaten Wahrscheinlichkeiten für Wortfolgen ableitet – Grundbaustein generativer KI.
OWASP LLM Top 10	Referenzliste der häufigsten Sicherheitsrisiken in LLM-Anwendungen – u.a. Prompt Injection, Data Leakage.
Prompt Injection	Angriffsmethode, bei der manipulierte Eingaben das Verhalten eines LLMs verändern.
RAG	Retrieval-Augmented Generation. LLM wird im Moment der Anfrage mit Fakten aus kontrollierten Quellen versorgt.
Successful Task	Betriebsmetrik: nicht Tokens oder API-Calls, sondern der konkrete erfolgreich abgeschlossene Nutzervorgang zählt.
Tool-Calling	Architekturmuster, in dem das LLM strukturierte Werkzeuge (APIs, Datenbanken) gezielt aufruft – statt nur frei zu antworten.
Trusted AI by Design	Architekturprinzip: Modell, Datenquelle und Aktionslogik strikt trennen – damit Output prüfbar und Provider austauschbar bleiben.

Quellen & Belege.

[1] **EU AI Act · Art. 113 lit. b**	Anwendungsdaten und Übergangsfristen für Hochrisiko-Systeme. · artificialintelligenceact.eu
[2] **EU AI Act · Art. 99 Abs. 3**	Sanktionen und Bußgeldhöhen für Verstöße – bis 7 % des weltweiten Jahresumsatzes. · artificialintelligenceact.eu
[3] **EU AI Act · Art. 9 – 15, 43, 49**	Pflichten für Hochrisiko-Systeme: Risikomanagement, Daten-Governance, Doku, Logs, Human Oversight, Konformitätsbewertung, EU-Registrierung. · artificialintelligenceact.eu
[4] **EU AI Act · Art. 50**	Transparenzpflichten für GenAI: Kennzeichnung KI-generierter Inhalte und Interaktionen. · artificialintelligenceact.eu
[5] **§ 38 BFSG**	Barrierefreiheitsstärkungsgesetz. Verpflichtende Geltung für digitale Produkte ab 28.06.2025. · gesetze-im-internet.de/bfsg
[6] **DSGVO · Art. 5, 22, 35**	Grundsätze der Verarbeitung, automatisierte Einzelentscheidung und Datenschutz-Folgenabschätzung – relevant für jeden KI-Use-Case mit Personenbezug. · dsgvo-gesetz.de
[7] **NIST AI RMF 1.0**	AI Risk Management Framework des US-amerikanischen NIST. De-facto-Referenz für vertrauenswürdige KI. · nist.gov/itl/ai-risk-management-framework
[8] **ISO/IEC 42001:2023**	Internationale Norm für KI-Managementsysteme – Pendant zu ISO 27001 für AI-Governance. · iso.org/standard/81230
[9] **OWASP LLM Top 10**	Referenzliste der häufigsten Sicherheitsrisiken in LLM-Anwendungen. · owasp.org/llm-top-10
[10] **Gartner · Customer Service**	Benchmark zur Ticket-Deflection: ein vermiedener Support-Kontakt spart im Schnitt 5–15 €. · Gartner Research, Customer Service & Support.

HINWEIS

Alle Verweise spiegeln den Stand der Veröffentlichung. Aktuelle Fassungen, Konkretisierungen und Leitlinien der Aufsichtsbehörden bitte den Originalquellen entnehmen.

Impressum & rechtliche Hinweise.

HERAUSGEBER

appsoluts GmbH
Kronprinzenstr. 97
40217 Düsseldorf · Deutschland

Geschäftsführung Christian Blank · Tobias Suchy
Registergericht Amtsgericht Düsseldorf · HRB 80393
USt-IdNr. DE311609132

T +49 211 972 653 50
E info@appsoluts.de
W www.appsoluts.de

REDAKTION

Redaktion: appsoluts GmbH
Konzept & Text: appsoluts GmbH
Gestaltung: appsoluts GmbH

BILDNACHWEIS

Sämtliche Abbildungen, Portraits und Grafiken: © appsoluts GmbH.

VERÖFFENTLICHUNG

Titel GenAI als Wettbewerbsvorteil 2026
Untertitel Recht, Architektur & Wirtschaftlichkeit
Edition Mai 2026
Sprache Deutsch
Format Whitepaper (A4 Hochformat)
Umfang 39 Seiten

RECHTLICHER HINWEIS

Die Inhalte dieses Whitepapers wurden mit größtmöglicher Sorgfalt erstellt. Sie stellen keine Rechtsberatung dar und ersetzen insbesondere keine individuelle rechtliche Beratung. Für die Richtigkeit, Vollständigkeit und Aktualität wird keine Haftung übernommen.

_app:soluts/*