TECHNICAL ARCHITECTURE WHITEPAPER

KI-Shield

Trust as a Service

Security Architecture Whitepaper

Version 1.0 · 17. April 2026

Datenschutzkonforme Nutzung generativer KI
mit Zero-Knowledge-Pseudonymisierung, Post-Quantum-signierter Audit-Kette
und Blockchain-verankerter Beweisführung

Publication Date	2026-04-17
Document Version	1.0
Status	Published — öffentliche Architektur-Dokumentation
Author	KI-Shield UG (haftungsbeschränkt)
Sitz	Greußen, Thüringen, Deutschland · HRB 524511 Amtsgericht Jena
Kontakt	info@ki-shield.de · ki-shield.de
Bezogene Schutzrechte	DE 20 2026 ___ ___ U1 (Basis-Gebrauchsmuster, 13.03.2026) DE 20 2026 ___ ___ U1 (Zero-Knowledge-Browser, 09.04.2026)
Lizenz	Architektur-Beschreibung: CC BY 4.0 International Geschützte Verfahren: Gebrauchsmusterrechte vorbehalten

Executive Summary

KI-Shield ist eine kryptografische Sicherheitsarchitektur, die Unternehmen die datenschutzkonforme Nutzung generativer KI-Sprachmodelle (ChatGPT, Claude, Gemini, Mistral AI, Llama und weitere) ermöglicht. Kern des Systems ist ein mehrschichtiges Pseudonymisierungsverfahren, das personenbezogene Daten (PII) bereits auf dem Endgerät des Nutzers im Webbrowser erkennt und durch typerhaltende Stellvertreterzeichenketten ersetzt, bevor sie an einen externen Sprachverarbeitungsdienst übertragen werden. Das System ist so konstruiert, dass der Betreiber von KI-Shield zu keinem Zeitpunkt Klartextzugriff auf schutzwürdige Inhalte erhält.

Ergänzend implementiert KI-Shield eine manipulationssichere Audit-Kette mit hybrider Signatur-Architektur: jeder Audit-Eintrag wird gleichzeitig mit einem klassischen Ed25519-Schlüssel und mit dem NIST-standardisierten Post-Quantum-Verfahren ML-DSA-65 (FIPS 204) signiert und vor dem Speichern unmittelbar gegen-verifiziert (Write-Time-Verifikation). Die kryptografische Integritätskette bleibt im Klartext und ist ohne Entschlüsselung öffentlich prüfbar, während die eigentlichen Inhalte mit einem aus dem Nutzerpasswort abgeleiteten Schlüssel verschlüsselt werden. Eine tägliche Verankerung in der Polygon-Blockchain sichert die Beweiskraft öffentlich ab.

Dieses Whitepaper dokumentiert die technische Architektur, die in den beiden eingetragenen deutschen Gebrauchsmustern GM-1 (Basis-Proxy-System, 13.03.2026) und GM-2 (Zero-Knowledge-Browser-System, 09.04.2026) geschützten Verfahren sowie deren Zusammenspiel mit den einschlägigen regulatorischen Anforderungen (DSGVO, EU AI Act, eIDAS 2.0, § 203 StGB, BSI IT-Grundschutz). Es richtet sich an IT-Verantwortliche, Datenschutzbeauftragte, Compliance-Verantwortliche und Auditoren.

1. Regulatorischer Kontext

Die Nutzung externer KI-Sprachmodelle für Geschäftsprozesse steht in einem wachsenden regulatorischen Spannungsfeld. KI-Shield adressiert durch seine Architektur — nicht durch vertragliche Zusagen — die folgenden Rahmenwerke:

Rahmen	Pflicht	Technische Antwort von KI-Shield
DSGVO Art. 25	Datenschutz durch Technikgestaltung	Client-seitige Pseudonymisierung vor jeder Datenübertragung.
DSGVO Art. 32	Angemessenheit technischer Maßnahmen	Zero-Knowledge, AES-256-GCM, Argon2id ≥ 600.000 Iterationen.
DSGVO Art. 9	Besondere Kategorien (Gesundheitsdaten etc.)	Dedizierte Erkennungsschicht für medizinische Begriffe.
§ 203 StGB	Schweigepflicht für Berufsgeheimnisträger	Kein Klartext verlässt den Nutzer-Rechner. Kein Geheimnis offenbart.
EU AI Act	Transparenz + Governance bei KI-Einsatz	Vollständig protokollierter, signierter Audit-Trail je Anfrage.
eIDAS 2.0	Qualifizierte Zeitstempel / PQC-Pflicht	RFC-3161-Zeitstempel + ML-DSA-65 Post-Quantum-Signatur.
BSI IT-Grundschutz	Basisschutz kritischer Prozesse	EU-Hosting, mTLS zwischen Diensten, Step-CA-PKI.
NIST FIPS 204	Standard für quantensichere Signaturen	ML-DSA-65 (Dilithium) in jeder Audit-Signatur.

2. Architektur-Übersicht

KI-Shield besteht aus zwei strikt getrennten Domänen: einer **Nutzer-Domäne** (Webbrowser auf dem Endgerät) und einer **Betreiber-Domäne** (Server-Infrastruktur). Zwischen beiden Domänen wird zu keinem Zeitpunkt unverschlüsselter personenbezogener Inhalt übertragen.

2.1 Module in der Nutzer-Domäne (Webbrowser, lokal)

• **Dokumentenextraktionsmodul (10):** PDF-, DOCX- und OCR-Verarbeitung (WebAssembly-basiert). Dokumente verlassen das Endgerät nicht.

• **Mustererkennungsmodul (20):** vier parallele Erkennungsschichten (siehe §3).

• **Substitutionsmodul (30):** Ersetzt Tokens durch typerhaltende Stellvertreterketten; verwaltet die bidirektionale Zuordnungstabelle (33) ausschließlich im flüchtigen RAM.

• **Kryptografisches Sicherungsmodul (40):** Web-Crypto-API-basiert. Argon2id-Schlüsselableitung (≥ 600.000 Iterationen), AES-256-GCM-Verschlüsselung.

• **Re-Substitutionsmodul (70):** Ersetzt in der KI-Antwort die Stellvertreterketten zurück durch die Originaltokens anhand der RAM-lokalen Zuordnungstabelle.

• **Orchestrierungsmodul (80):** Wählt zwischen Betriebsmodi (Nur-RAM vs. verschlüsselter Chiffretext-Storage für Mehr-Sitzungs-Kontinuität).

2.2 Module in der Betreiber-Domäne (Server)

• **Empfangsschnittstelle (51) des Vermittlungsmoduls (50):** Nimmt ausschließlich den bereits substituierten Anfragetext und Metadaten der erkannten Datenklassen entgegen.

• **Weiterleitung an externen Sprachverarbeitungsdienst (60):** Standardisiertes HTTPS-Routing an gewählten LLM-Anbieter (BYOK-Unterstützung für ChatGPT, Claude, Gemini, Mistral AI, Llama etc.).

• **Speichereinheit (53):** Speichert ausschließlich Chiffretextblöcke des Nutzers — ohne kryptografische Mittel zur Entschlüsselung. Entspricht dem Zero-Knowledge-Prinzip.

• **Audit-Kette:** Jede Transaktion wird mit Ed25519 + ML-DSA-65 dual-signiert und Write-Time-verifiziert (siehe §5).

• **Tägliche Blockchain-Verankerung:** Merkle-Root der Audit-Kette wird in der Polygon-Blockchain festgeschrieben.

3. Vier-Schichten-PII-Erkennung

Das Mustererkennungsmodul (20) kombiniert vier parallele Erkennungsschichten. Jede Schicht liefert Kandidaten; ein Konfliktauflösungs-Verfahren nach dem Prinzip der längsten Übereinstimmung sowie kontextbasierter Unterscheidung konsolidiert die Ergebnisse.

3.1 Namensbasierte Erkennungsschicht (21)

Diese Schicht prüft Tokens gegen eine Sammlung von Eigennamen mittels einer Datenstruktur mit konstanter Lookup-Zeit (Set-basierte Hash-Suche). Kontextuelle Hinweise wie Anreden (Herr, Frau, Dr., Prof.) und akademische Titel werden berücksichtigt, um die Präzision bei Mehrdeutigkeiten (z.B. Müller als Nachname vs. allgemeine Berufsbezeichnung) zu erhöhen.

3.2 Musterbasierte Erkennungsschicht (22) mit Prüfziffer-Validierung

Strukturierte Identifikatoren werden mittels regulärer Ausdrücke erkannt und durch validierende Prüfziffernalgorithmen verifiziert:

- **MOD-97-Validierung** für die Internationale Bankkontonummer (IBAN) nach ISO 13616.
- **Luhn-Prüfziffer** für Kreditkartennummern (ISO/IEC 7812).
- **ISO/IEC 7064 Mod-11,10** für Krankenversicherungsnummern.
- **ICAO-9303-Prüfziffern** für maschinenlesbare Identitätsdokumente (Pass, Personalausweis).

Die Prüfziffer-Validierung reduziert Fehlalarme gegen Null: nur Tokens mit mathematisch korrekter Prüfziffer werden als PII-Kandidaten markiert.

3.3 Schlüsselwortbasierte Erkennungsschicht (23)

Kuratierte Schlüsselwortlisten für 42 PII-Kategorien (Gesundheitsdaten, Religionszugehörigkeit, politische Meinung, biometrische Daten etc.). Die Auswertung erfolgt kontextuell — eine medizinische Diagnose wird nur bei passendem umgebenden Textkontext erkannt, um Treffer in unbeteiligten Nennungen zu vermeiden.

3.4 Kontextbasierte Konfliktauflösung

Bei mehrdeutigen Erkennungen (z.B. Geräteseriennummer vs. Telefonnummer, Kreditkartennummer vs. anderer numerischer Identifikator) entscheidet eine kontextbasierte Analyse des umgebenden Textes. Dies reduziert Präzisionsverluste an typischen Schnittstellen zwischen den Kategorien.

3.5 Neuronale Erkennungsschicht (Server-Variante)

Die Server-Variante (KI-Shield Proxy, `ki-shield.de/chat`) ergänzt die drei Browser-Schichten durch eine neuronale Erkennungsschicht mit einem spaCy-basierten Named-Entity-Recognition-Modell. In der reinen Browser-Variante (GM-2-Schutzumfang) arbeiten ausschließlich die drei deterministischen Schichten §3.1–§3.3, die ohne ML-Inferenz auskommen.

4. Substitution und Zuordnungstabelle

4.1 Typerhaltende Pseudonyme

Erkannte Tokens werden nicht durch einen generischen Platzhalter ersetzt, sondern durch typspezifische Stellvertreterketten aus einer Vorlagenbibliothek (31). Beispiele:

```
Müller → [PERSON_001]
15.03.1978 → [DOB_001]
DE89 3704... → [IBAN_001]
Diabetes Typ 2 → [MEDICAL_001]
```

Der Typerhalt sichert die semantische Kohärenz des Textes: die KI behandelt [IBAN_001] weiterhin als Bankkontonummer und erzeugt sinnvolle Antworten. Eine generische Schwärzung würde die Nutzbarkeit der KI-Antwort drastisch reduzieren.

4.2 Initialisierungsoffset gegen Informationslecks

Zu Beginn jeder Sitzung wählt das Substitutionsmodul einen pseudozufälligen Initialisierungsoffset (32) aus einem vorgegebenen Wertebereich. Die Nummerierung der Pseudonyme startet nicht bei 001, sondern bei einer zufälligen Zahl. Damit kann aus dem numerischen Anteil einer Stellvertreterkette nicht rückgeschlossen werden, wie viele Tokens des jeweiligen Typs in der Sitzung erkannt wurden — eine Schwäche naiver Substitutionsverfahren wird geschlossen.

4.3 Bidirektionale Zuordnungstabelle — flüchtig im RAM

Die Zuordnungstabelle (33) verknüpft Originaltokens mit Stellvertreterketten und ermöglicht die spätere Re-Substitution. Sie wird **ausschließlich im flüchtigen Arbeitsspeicher des Webbrowsers** gehalten. Sie wird zu keinem Zeitpunkt an den Server übermittelt. Der Server hat keinerlei Möglichkeit, die Pseudonyme wieder auf die ursprünglichen Tokens abzubilden.

5. Web-Crypto-Flow und Betriebsmodi

KI-Shield unterstützt zwei Betriebsmodi für die Mehr-Sitzungs-Kontinuität der Zuordnungstabelle:

5.1 Modus A — Nur-RAM (maximaler Zero-Knowledge-Schutz)

Die Zuordnungstabelle existiert ausschließlich im flüchtigen Arbeitsspeicher des Webbrowsers und wird mit dem Tab-Schließen vernichtet. Keine Persistenz. Auch KI-Shield selbst kann eine beendete Sitzung nicht rekonstruieren.

5.2 Modus B — Verschlüsselter Chiffretext-Storage

Für Nutzer, die Sitzungen über Tage hinweg fortsetzen möchten, wird die Zuordnungstabelle kryptografisch gesichert und als Chiffretextblock auf dem Server vorgehalten — jedoch ohne dass der Server über Mittel zur Entschlüsselung verfügt:

- Schlüsselableitung aus dem Nutzerpasswort mittels **Argon2id mit mindestens 600.000 Iterationen** und einem kryptografisch zufälligen Salt (Web-Crypto-API).
- Verschlüsselung der Zuordnungstabelle mit einem **Authenticated-Encryption-Algorithmus** (AES-256-GCM) mit zufälligem Initialisierungsvektor und Authentifizierungsanhang.
- Übermittlung des resultierenden Chiffretextblocks an die Speichereinheit (53) des Servers.
- In späterer Sitzung: Abruf des Chiffretextblocks vom Server, Entschlüsselung im Browser mit dem erneut aus dem Passwort abgeleiteten Schlüssel.

Der Server sieht ausschließlich den verschlüsselten Block. Der Entschlüsselungsschlüssel verlässt zu keinem Zeitpunkt den Browser des Nutzers.

5.3 Browser-Standards, kein Browser-Plugin

Das System verwendet ausschließlich standardisierte Browser-Schnittstellen — die Web Crypto API, die File API und die Canvas API. Eine Installation zusätzlicher Software, einer Browser-Erweiterung oder eines Plugins ist nicht erforderlich. Damit ist KI-Shield auf jedem handelsüblichen Browser unmittelbar einsatzfähig.

6. Hybride Post-Quantum-Audit-Kette

6.1 Warum hybrid?

Klassische Signaturverfahren (ECDSA, RSA, Ed25519) gelten gegen aktuelle klassische Angreifer als sicher, sind aber anfällig gegen zukünftige kryptografisch relevante Quantencomputer. Post-Quantum-Verfahren (ML-DSA-65) sind dagegen als quantensicher konzipiert, haben aber weniger langjährige Analyseerfahrung. KI-Shield signiert daher **jeden Audit-Eintrag gleichzeitig mit beiden Verfahren**. Ein Angreifer müsste beide Verfahren gleichzeitig brechen, um einen Eintrag zu fälschen.

6.2 Write-Time-Verifikation

Nach Erzeugung der beiden Signaturen und vor dem Schreiben in die Audit-Kette werden beide Signaturen unmittelbar gegen den gerade signierten Payload und die jeweiligen Public Keys verifiziert. Schlägt die Verifikation fehl, wird der Eintrag nicht geschrieben. Dieses Verfahren verhindert, dass korrupte Signaturen durch Speicherfehler oder Implementierungsfehler in die Kette gelangen.

6.3 Hash-Chain mit öffentlich prüfbarer Integrität

Jeder Audit-Eintrag enthält den Hash des vorherigen Eintrags (previous_hash) sowie seinen Chain-Index. Dies verhindert das nachträgliche Einfügen, Entfernen oder Umordnen von Einträgen — jede Manipulation zerstört die Verkettung. Die kryptografische Integritätskette bleibt dabei **im Klartext**, während die inhaltlichen Nutzlasten (z.B. Pseudonymisierungs-Metadaten) mit dem nutzerspezifischen Schlüssel verschlüsselt sind. Ein Auditor kann damit die Integrität der Kette verifizieren, ohne die Inhalte zu sehen.

6.4 Qualifizierter RFC-3161-Zeitstempel

Der Hash jedes Audit-Eintrags wird zusätzlich durch einen qualifizierten Zeitstempel nach RFC 3161 versiegelt. Der Zeitstempel stammt von einer Time-Stamp Authority (TSA) mit Zertifikatsbindung an eine öffentliche CA. Damit ist der Existenzzeitpunkt eines Eintrags unabhängig vom KI-Shield-System beweisbar — ein Gericht kann den Zeitpunkt auch dann bestätigen, wenn KI-Shield nicht mehr existiert.

6.5 Tägliche Verankerung in der Polygon-Blockchain

Einmal täglich wird der Merkle-Root aller Audit-Einträge des Tages in einer Transaktion der Polygon-Proof-of-Stake-Blockchain festgeschrieben. Dies schafft eine dritte, von KI-Shield unabhängige Beweisebene: selbst wenn TSA und Audit-Kette gleichzeitig kompromittiert würden, liefert die unveränderliche Blockchain-Verankerung weiterhin den Beweis des Existenzzeitpunkts.

7. Compliance-Mapping

Die folgende Übersicht ordnet jede regulatorische Kernanforderung einer konkreten technischen Maßnahme aus der KI-Shield-Architektur zu. Die Tabelle ist so gestaltet, dass sie direkt in ein Datenschutz-Folgenabschätzungs-Dokument (DSFA) übernommen werden kann.

Anforderung	Quelle	KI-Shield-Maßnahme
Datensparsamkeit	DSGVO Art. 5(1)(c)	Server erhält nur substituierte Pseudonyme.
Privacy by Design	DSGVO Art. 25	Pseudonymisierung architektonisch erzwungen, nicht optional.
Technische Sicherheitsmaßnahmen	DSGVO Art. 32	AES-256-GCM, Argon2id ≥ 600k, Ed25519 + ML-DSA-65.
Besondere Kategorien	DSGVO Art. 9	Dedizierte Erkennungsschicht für Gesundheits-/Biometrie-Daten.
Transparenz ggü. Betroffenen	DSGVO Art. 13/14	Audit-Kette mit Nachweis jeder KI-Anfrage.
Betroffenenrechte (Art. 15ff)	DSGVO Art. 15-22	Export von Nutzerdaten als signierte Paketdatei.
Löschpflicht	DSGVO Art. 17	Chiffretext-Löschung wirkt wie kryptografische Vernichtung.
Berufsgeheimnis	§ 203 StGB	Kein Klartext verlässt Nutzer-Rechner — kein Offenbaren.
AI Act (Transparenz)	EU AI Act Art. 13/14	Lückenlose Audit-Kette mit mathematischer Integrität.
eIDAS 2.0 PQC	eIDAS 2.0	ML-DSA-65 (FIPS 204) für zukünftige PQC-Pflicht.

8. Sicherheitsbetrachtungen

Angriffsfläche des Betreibers: Selbst bei vollständiger Kompromittierung des KI-Shield-Servers besitzt der Angreifer ausschließlich verschlüsselte Chiffretextblöcke und Hashes. Ohne das Nutzerpasswort sind diese Daten kryptografisch unbrauchbar. Ein erfolgreicher Angriff auf den Server leckt **keine** personenbezogenen Daten der Nutzer.

Angriffsfläche des LLM-Anbieters: Der externe Sprachverarbeitungsdienst erhält ausschließlich den substituierten Anfragetext mit Pseudonymen. Er sieht keine realen Namen, IBANs, Adressen oder Diagnosen. Auch ein US Cloud Act-Zugriff auf OpenAI, Anthropic oder Google liefert keine identifizierenden Nutzerdaten.

Post-Quantum-Widerstandsfähigkeit: Die gleichzeitige Signatur mit Ed25519 und ML-DSA-65 gewährleistet, dass beim Auftreten kryptografisch relevanter Quantencomputer die ML-DSA-65-Signatur die Beweiskraft der Audit-Kette fortführt. Eine separate Neu-Signierung ist nicht erforderlich.

Restrisiko Nutzer-Endgerät: Die Zero-Knowledge-Architektur verlagert einen Teil des Vertrauens auf die Integrität des Nutzer-Endgeräts und -Browsers. Ein kompromittiertes Endgerät (Trojaner, Keylogger) kann den Schutz unterlaufen. KI-Shield empfiehlt daher kombinierten Einsatz mit Endpoint-Security.

9. Bewusste Grenzen

Im Sinne wissenschaftlicher Redlichkeit: KI-Shield ist keine Antwort auf jede DSGVO-Frage, und nicht jede in diesem Dokument genannte Komponente ersetzt eine rechtliche Prüfung im Einzelfall. Folgende Einschränkungen sind bekannt und werden offen kommuniziert:

- **Auftragsverarbeitung bleibt erforderlich.** Die Zero-Knowledge-Architektur reduziert die Eingriffsintensität, hebt aber nicht die AVV-Pflicht nach Art. 28 DSGVO auf.
- **Erkennungsrate ist nicht 100%.** Kein PII-Erkennungssystem ist perfekt. Nutzer sollten bei hochsensiblen Inhalten zusätzlich manuell prüfen.

• **Sprach- und Domänenabhängigkeit.** Die Erkennungsschichten sind primär auf deutsche und englische Texte kalibriert. Andere Sprachen sind in Entwicklung.

• **Externe LLM-Qualität.** Wir haben keinen Einfluss darauf, ob ChatGPT oder Claude zufällig personenbezogene Phantasie-Antworten generiert. Unser Re-PII-Check der Antwort filtert erkennbare Fälle.

10. Referenzen und Standards

[NIST FIPS 204] Module-Lattice-Based Digital Signature Standard (ML-DSA). NIST, 2024.

[NIST FIPS 180-4] Secure Hash Standard (SHS): SHA-256. NIST, 2015.

[NIST SP 800-38D] Recommendation for Block Cipher Modes of Operation: Galois/Counter Mode. AES-256-GCM.

[RFC 3161] Internet X.509 PKI Time-Stamp Protocol. IETF, 2001.

[RFC 8032] Edwards-Curve Digital Signature Algorithm (EdDSA) — Ed25519. IETF, 2017.

[RFC 9106] Argon2 Memory-Hard Function for Password Hashing and Proof-of-Work Applications. IETF, 2021.

[ISO 13616] IBAN Structure (MOD-97).

[ISO/IEC 7812] Identification cards — Luhn Check-Digit.

[ISO/IEC 7064] Information technology — Security techniques — Check character systems (Mod-11,10).

[ICAO Doc 9303] Machine Readable Travel Documents.

[DSGVO] Verordnung (EU) 2016/679.

[EU AI Act] Verordnung (EU) 2024/1689 über harmonisierte KI-Vorschriften.

[eIDAS 2.0] Verordnung (EU) 2024/1183.

[§ 203 StGB] Verletzung von Privatgeheimnissen (Berufsgeheimnisträger).

[BSI IT-GS] BSI IT-Grundschutz-Kompendium.

Anhang: Prior-Art-Statement und Lizenzierung

Die in diesem Dokument beschriebene Kernarchitektur ist in den folgenden eingetragenen deutschen Gebrauchsmustern geschützt:

Schutzrecht	Gegenstand	Eintragungsdatum
DE 20 2026 ___ ___ U1	Datenschutzkonformes Proxy-System für LLMs mit hybrider PQ-Audit-Kette und Zero-Knowledge-Verschl	13.03.2026
DE 20 2026 ___ ___ U1	Zero-Knowledge-Browser-System für sprachmodellbasierte Anfragen mit lokaler Zuordnungstabelle	09.04.2026

Lizenzierung: Die in diesem Dokument *zusätzlich* beschriebenen Architektur-Elemente, die nicht Gegenstand der oben genannten Gebrauchsmuster sind (insbesondere regulatorische Interpretationen, Compliance-Mappings und ergänzende Erkennungsalgorithmen), werden unter der Creative Commons Attribution 4.0 International Lizenz (CC BY 4.0) veröffentlicht.

Nachbauhinweis: Die Implementierung der Verfahren, die durch die oben genannten Gebrauchsmuster geschützt sind, erfordert eine Lizenz der Schutzrechtsinhaberin KI-Shield UG. Die Beschreibung in diesem Whitepaper dient der Transparenz gegenüber Kunden, Auditoren und der wissenschaftlichen Öffentlichkeit und stellt keine Lizenzgewährung dar.

Zweck der Veröffentlichung: Dieses Dokument dient primär der technischen Dokumentation gegenüber Geschäftskunden, Datenschutzbeauftragten und Auditoren. Es ergänzt — nicht ersetzt — die amtliche Gebrauchsmuster-Beschreibung.

Ende des Whitepapers Version 1.0

KI-Shield UG (haftungsbeschränkt) · HRB 524511 Amtsgericht Jena · Greußen, Thüringen
info@ki-shield.de · ki-shield.de · ki-shield.eu
Veröffentlicht am 17. April 2026