# 1_Defensive-Publication_Der_Prior-Art_Schutz_NWM_v3.2

# Defensive Publication: Datensouveränes Wissensmanagement mit hybrider KI-Architektur in Einzelorganisationen und Netzwerken

**Projekt-Code:** #NWM (Netz-Werkzeug.Macherei)
**Status:** Öffentlich – v3.2
**Datum:** 260422 (22.04.2026)
**Verfasser:** Netz-Werkzeug.Macherei
**Lizenz:** CC BY-NC-SA 4.0
**Vorgängerversionen:** v3.0 (15.04.2026, Zenodo + eigene Domain) → v3.1 (22.04.2026)

---

## Begriffsverzeichnis

| Begriff | Definition |
|---|---|
| **iKi (interne KI)** | Lokal ausgeführte KI-Modelle (On-Premise) zur datenschutzkritischen Verarbeitung |
| **cKi (Cloud-KI)** | Externe KI-Dienste über API-Gateways für hohe Rechenleistung |
| **HITL (Human-in-the-Loop)** | Manuelle Freigabeinstanz vor Cloud-Datenübermittlung |
| **HITL-Freigabe** | Nutzerbestätigung vor externer Datenübermittlung |
| **HITL-Logging** | Protokollierung aller Nutzer-Freigaben zur Auditierbarkeit (revisionssicheres Append-only Log) |
| **RAG (Retrieval-Augmented Generation)** | Anreicherung von KI-Anfragen mit kontextuellen Dokumenten aus lokaler Wissensdatenbank |
| **SSOT (Single Source of Truth)** | Zentrale, autoritative Datenplattform |
| **SOP (Standard Operating Procedure)** | Dokumentierte Standardarbeitsanweisungen für Prozesskonformität |
| **ZDR (Zero-Data-Retention)** | Richtlinie zur Nicht-Speicherung von Anfragen durch Gateway-Anbieter |
| **RBAC (Role-Based Access Control)** | Rollenbasierte Zugriffssteuerung für Nutzerberechtigungen |
| **PII (Personally Identifiable Information)** | Personenbezogene Daten, die besonderen Schutz erfordern |

| Begriff | Definition |
|---|---|
| **NER (Named Entity Recognition)** | Verfahren zur automatischen Erkennung von Entitäten (Personen, Orte, Organisationen) |
| **Föderation** | Sichere Verbindung mehrerer unabhängiger Instanzen zu dezentralem Netzwerk |
| **Fine-Tuning** | Anpassung vortrainierter KI-Modelle auf domänenspezifische Daten |

---

# 1. Technisches Feld

Die vorliegende Offenlegung betrifft ein informationstechnisches System zur Integration von Künstlicher Intelligenz (KI) in die Wissens-Infrastruktur von souveränen Organisationen (KMUs, Vereine, NGOs, Gemeinden, Interessensgemeinschaften) sowie datenschutzbewussten Einzelanwendern (Heim-Server-Betreiber).

Insbesondere beschreibt sie eine **auditierbare Orchestrierungsarchitektur mit deterministischen Entscheidungs-Checkpoints und revisionssicherer Protokollierung**, die lokale Datenverarbeitung (On-Premise) mit cloud-basierten KI-Diensten unter Wahrung der Datensouveränität verbindet. Das System umfasst:

- **Hybride KI-Orchestrierung** (Interne iKi + Cloud cKi)
- **RAG-Anbindung** (Retrieval-Augmented Generation) an selbstgehostete Wissensplattformen
- **Human-in-the-Loop** (HITL) als finale Freigabeinstanz mit technischem Trigger-Mechanismus
- **Multimodale Verarbeitung** (Text, Bild, Audio, Video)
- **Agenten-Funktionen** (ausführende KI, z.B. Referenz-Implementierungen wie LangChain, AutoGen, CrewAI, OpenClaw)
- **Auditierbares Qualitätsmanagement:** Die Architektur ermöglicht prozesskonforme Abläufe durch revisionssicheres HITL-Logging, konsistente Wissensanwendung (SOP-Anbindung via RAG) und auditierbare Datenflüsse. Dies unterstützt Qualitätsmanagement-Anforderungen (z.B. ISO 9001) auf technischer Ebene (ersetzt jedoch keine Zertifizierung).
- **Dezentrale Netzwerkstruktur:** Das System unterstützt verteilte Zugriffsszenarien über verschlüsselte VPN-Verbindungen, wodurch mehrere Standorte sicher miteinander verbunden werden können, ohne dass die Infrastruktur im öffentlichen Internet erreichbar sein muss.

**Zweck dieser Veröffentlichung:** Sicherung des Standes der Technik/Wissenschaft (Prior Art) gemäß §3 PatG (DE) bzw. Art. 54 EPÜ zur Verhinderung von Patentansprüchen Dritter auf das beschriebene Systemdesign.

**Bezug zu Vorgängerversionen:** Diese Veröffentlichung baut auf der Defensive Publication NWM v3.0 vom 15.04.2026 (DOI/Zeitstempel via Zenodo + eigene Domain) und v3.1 vom 22.04.2026 auf. Die Kernarchitektur (hybride iKi/cKi-Orchestrierung mit Routing-Logik, HITL, RAG, Multi-Level-Guardrails, Föderation) ist bereits seit dem 15.04.2026 als Prior Art etabliert. Die vorliegende Version v3.2 erweitert und präzisiert die technische Beschreibung.

**Hinweise zu Produkten:**

- Alle im Dokument genannten Produkte, Dienste und Technologien (z.B. Nextcloud, Ollama, OpenRouter) dienen als Referenz-Implementierungen. Das System ist technologie-agnostisch und übertragbar auf vergleichbare Lösungen mit ähnlichen Spezifikationen.
- Die Einzelkomponenten sind nach Vorlieben einsetzbar.
- **Hinweis zur Aktualität:** Die Entwicklung im Bereich der Künstlichen Intelligenz schreitet mit hoher Geschwindigkeit voran. Produktbezeichnungen, Modellversionen und technische Spezifikationen, die in diesem Dokument genannt werden, können bereits wenige Wochen nach Veröffentlichungsdatum veraltet sein.

**Haftungsausschluss und rechtliche Hinweise:**

1. Diese Veröffentlichung beschreibt ein Architekturkonzept und stellt keine fertige Software-Lösung dar. Die tatsächliche Sicherheit im Betrieb hängt von der korrekten Implementierung, der Einhaltung von Sicherheitsrichtlinien durch den Betreiber sowie der fortlaufenden Anpassung an aktuelle Bedrohungslagen und regulatorische Anforderungen ab.
2. Alle genannten Produkte, Dienstleistungen und Technologien dienen als Referenz-Implementierungen und sind Warenzeichen oder eingetragene Warenzeichen ihrer jeweiligen Eigentümer. Die Nennung stellt keine Produktempfehlung dar.
3. Diese Veröffentlichung dient ausschließlich der Sicherung des Standes der Technik (Prior Art) und stellt keine Patentlizenz dar. Die CC BY-NC-SA 4.0 Lizenz bezieht sich auf das Urheberrecht an dieser Dokumentation, nicht auf eventuelle Patentrechte, die Dritte an den beschriebenen Verfahren oder Systemen halten könnten.
4. Keine Gewährleistung für Vollständigkeit, Richtigkeit oder Aktualität der genannten Preise, Modellversionen oder regulatorischen Verweise.
5. Die rechtliche Zulässigkeit konkreter Implementierungen – insbesondere bezüglich DSGVO, EU AI Act und der Allgemeinen Geschäftsbedingungen externer KI-Dienstleister – ist im Einzelfall durch den Implementierer zu prüfen.

---

# 2. Hintergrund & Problemstellung

## 2.1 Technisches Problem

Bestehende Lösungen zur KI-Integration neigen dazu, entweder vollständig cloud-basiert zu sein (mit Risiken hinsichtlich Datenschutz, Datenabfluss und Vendor-Lock-in) oder vollständig lokal zu operieren (mit Einschränkungen hinsichtlich Rechenleistung und Modellkapazität).

Für Organisationen und Einzelanwender mit hohen Datenschutzanforderungen (z.B. aufgrund von DSGVO, Berufsgeheimnissen oder persönlicher Präferenz) besteht ein technisches Problem darin, leistungsfähige KI-Modelle zu nutzen, ohne sensible Daten unkontrolliert an externe Anbieter zu übermitteln.

## 2.2 Zielgruppen & Anforderungsprofil

| Zielgruppe | Bisherige Lösungen | Architektureller Ansatz |
|---|---|---|
| **Enterprise** | Kong AI Gateway, MuleSoft (teuer, komplex) | Nicht primärer Fokus |
| **KMU (<500 MA)** | Keine spezifischen integrierten Lösungen | **Fokus: Hybride Orchestrierung** |
| **Einzelanwender** | PrivateGPT, Ollama (lokal, kein Team) | **Fokus: Team-Funktion & Sovereignty** |

**Technischer Kern:** Es geht nicht nur um KI. Es geht um **eigenes Wissen** (Termine, Kontakte, Projekte, Dateien, SOPs). Sicher gespeichert. KI-gestützt durchsuch- und bearbeitbar. Im Team nutzbar, ohne dass externe Dritte mitlesen.

---

# 3. Charakteristische Merkmalskombination (Anspruchskette)

Die nachfolgend beschriebene Architektur zeichnet sich durch die zwingende Kombination folgender Merkmale aus, die in Sequenz und Wechselwirkung das Schutzziel der Datensouveränität bei gleichzeitiger Nutzung externer Hochleistungs-KI sicherstellen:

**M1:** Eine lokale Verarbeitungsebene (iKi) mit mindestens einem Open-Weight-Sprachmodell (≥8B Parameter), das ALLE eingehenden Nutzeranfragen vor jeglicher externer Übermittlung verarbeitet.

**M2:** Eine vier-stufige PII-Filter-Pipeline mit regelbasierter Erkennung, LLM-basierter semantischer Umformulierung, reversiblem Mapping und Verifikation (siehe Abschnitt 6.1).

**M3:** Ein deterministisches HITL-Gate mit konfigurierbarem Trigger-Mechanismus, das zwischen iKi-Vorverarbeitung und cKi-Übermittlung als nicht-umgehbarer Checkpoint operiert (siehe Abschnitt 6.2).

**M4:** Eine RAG-Anbindung an eine selbstgehostete SSOT-Plattform mit RBAC-Berücksichtigung bereits beim Retrieval (siehe Abschnitt 4.3).

**M5:** Eine Multi-Provider-Routing-Schicht mit ZDR-Policy-Filterung und Anbieter-Diversifizierung (siehe Abschnitt 6.3 und 6.4).

**M6:** Eine optionale Föderationsschicht über VPN für Inter-Instanz-Kommunikation ohne öffentlichen Internet-Zugang (siehe Abschnitt 6.5).

**M7:** Eine optionale Multimodal-Erweiterung für Bild-, Audio- und Video-Verarbeitung unter Beibehaltung der M1-M5-Sicherheitskette (siehe Abschnitt 4.4).

Die Merkmale M1–M5 sind obligatorisch, M6 und M7 sind optional. Jede Implementierung, die alle Merkmale M1–M5 in der beschriebenen sequenziellen Anordnung umfasst, ist von dieser Offenlegung erfasst.

---

# 4. Systemarchitektur

## 4.1 Zentrale Wissensplattform (Single Source of Truth)

Als zentrale Ablage- und Kollaborationsplattform dient eine selbstgehostete oder Managed-Hosting-Cloud-Instanz (beispielsweise basierend auf Nextcloud-Technologie).

**Funktionen:**

- Dokumentenablage (Dateien, PDFs, Bilder, Videos)
- Kalender, Kontakte, Aufgaben (Kanban-Boards, z.B. Deck)
- Wiki-Funktionalität (Markdown-basiert)
- Kommunikation (z.B. Nextcloud Talk)
- **SSOT:** Single Source of Truth – alle Daten an einem Ort

**Strukturierung:**

- **Geteiltes Wissen:** Team-Zugriff über zentrale Plattform
- **Persönliches Wissen:** Internes Denken, Entwurf, persönliche Notizen (z.B. synchronisierte Vaults)

**Hosting-Optionen:**

- Initial: Managed Hosting für schnellen Start
- Langfristig: Eigener Server (Proxmox/Linux) für maximale Kontrolle
- **Cloud-agnostisch:** Nextcloud ist Referenz, aber übertragbar auf OwnCloud, Seafile, etc.

**Netzwerk-Zugriff:**

Das System unterstützt föderierte Zugriffsszenarien: Nutzer können über verschlüsselte VPN-Verbindungen (z.B. Tailscale, WireGuard) auf die Plattform zugreifen, ohne dass diese im öffentlichen Internet erreichbar sein muss. Die Architektur erlaubt die Trennung von Plattform (Speicher) und KI-Ressourcen (Rechenleistung), die sicher über das interne Netzwerk verbunden werden.

## 4.2 Hybrider KI-Orchestrator

Ein zentrales Merkmal des Systems ist eine Orchestrierungsschicht, die eingehende Anfragen analysiert und routet.

**Sequenzieller Datenfluss:**

```
Eingabe → iKi (Datensammlung + RAG-Retrieval + 4-stufige PII-Filterung)
       → HITL-Gate (Trigger-basierte Nutzer-Validierung)
       → cKi (Cloud-Verarbeitung, falls erforderlich)
       → iKi (Re-Identifikation + Formatierung)
       → Ausgabe an Nutzer
```

**Gateway-agnostisch:** Das System nutzt referenziell Multi-Model-Gateways (z.B. OpenRouter.ai), ist aber nicht darauf beschränkt.

### 4.2.1 Lokale Verarbeitungsebene (iKi – interne KI)

| Aspekt | Spezifikation (Beispiele) |
|---|---|
| **Software** | Ollama, LM Studio, Msty Studio (Referenz-Implementierungen) |
| **Modelle** | 8B–13B Parameter (Llama-3-8B, Mistral, Qwen) |
| **Hardware** | GPU ≥16GB VRAM (z.B. RTX 4060 Ti 16GB) |
| **Einsatz** | Datenschutzkritische Anfragen, PII-Anonymisierung (intern), RAG-Retrieval |

### 4.2.2 Cloud-Verarbeitungsebene (cKi – cloudbasierte KI)

| Aspekt | Spezifikation |
|---|---|
| **Gateway** | OpenRouter.ai (Referenz), direkt API (Anthropic, OpenAI, etc.) |
| **Modelle** | 70B+ Parameter (Qwen-72B, GPT-4, Grok, Gemini) |
| **Datenschutz** | EU-Only wo möglich, ZDR (Zero-Data-Retention), PII-Filter (zusätzlich extern) vor Sendung an Rechendienstleister |
| **Einsatz** | Komplexe Tasks, hohe Leistung, wenn iKi nicht ausreicht |

### 4.2.3 Routing-Logik (Orchestrator-gesteuert)

Die Routing-Entscheidung erfolgt anhand eines zweidimensionalen Klassifikationsschemas:

| Fall | Sensibilität | Leistung | Lösung |
|---|---|---|---|
| **A** | – | – | → iKi/cKi (lokale Ressourcen bevorzugt) |
| **B** | – | + | → cKi |
| **C** | + | – | → iKi |
| **D** | + | + | → L1 / L2 / L3 (siehe unten) |

**Lösungsoptionen für Fall D (Sensibel UND Hochleistung):**

- **L1:** Lokale Anonymisierung (iKi) → dann Cloud (cKi); **Standard-Fall**
- **L2:** In mehrere Fragen aufteilen, nur über iKi (Verzicht auf Hochleistung)
- **L3:** Direkt mit cKi (Akzeptanz des reduzierten, aber nicht vollständig eliminierbaren Restrisikos via Human-in-the-Loop)

**Sensibilitäts-Klassifikation:** Die Sensibilität wird durch den lokalen PII-Detector quantifiziert (Confidence-Score, siehe Abschnitt 6.1). Liegt der Score über einem konfigurierbaren Schwellwert, wird die Anfrage als sensibel klassifiziert.

**Leistungs-Klassifikation:** Die erforderliche Leistung wird durch Komplexitäts-Heuristiken (Token-Länge, semantische Komplexität, Aufgabentyp) abgeschätzt.

## 4.3 RAG-Implementierung (Retrieval-Augmented Generation)

| Komponente | Technologie (Beispiele) | Funktion |
|---|---|---|
| **Vektor-Datenbank** | ChromaDB, Qdrant, LlamaIndex | Speichert Dokumenten-Embeddings |
| **Indexierung** | Plattform-Inhalte (PDF, MD, DOCX, etc.) | Automatische Extraktion bei Upload |
| **Retrieval** | iKi fragt Vektor-DB vor Generierung der cKi-Anfrage | Kontext wird lokal angereichert |
| **Orchestrierung** | Python/FastAPI + litellm, n8n, Node-RED | Workflow-Automatisierung |

**Spezifischer Datenfluss:**

1. Dokument-Upload → automatisches Chunking (z.B. RecursiveCharacterTextSplitter, chunk_size=512, overlap=50)
2. Embedding-Generierung **lokal** via iKi (z.B. nomic-embed-text, bge-m3) – KEINE Cloud-Embeddings
3. Speicherung in Vektor-DB mit Metadaten (Quelle, RBAC-Tag, Zeitstempel)
4. Bei Query: lokale Embedding-Generierung → Top-k Retrieval (k=5, Cosine-Similarity > 0.7)

5. RBAC-Filterung der retrieved chunks **vor** Kontext-Injection
6. PII-Filter prüft auch retrieved chunks (nicht nur User-Input!)
7. Kontext-Injection in Prompt-Template **vor** HITL-Gate

## 4.4 Multimodale Architektur

Das System ist erweiterbar auf multimodale Verarbeitung unter Beibehaltung der gleichen Sicherheitskette (M1–M5):

| Modalität | Lokal (iKi) | Cloud (cKi) |
|---|---|---|
| **Text** | Llama-3-8B, Mistral | GPT-4, Qwen-72B, Grok |
| **Bild** | Stable Diffusion (lokal) | DALL-E 3, Midjourney (via API) |
| **Audio** | Whisper (Transkription) | ElevenLabs (TTS), Whisper API |
| **Video** | Frame-Extraktion + Whisper | Sora, Runway (via API) |

**Hinweis:** Multimodale Funktionen sind optional und können nach Bedarf aktiviert werden. Die PII-Filterung wird modalitätsspezifisch erweitert (z.B. Gesichts-Anonymisierung in Bildern, Stimm-Anonymisierung in Audio).

## 4.5 Agenten-Funktionen

Das System ist kompatibel mit ausführenden KI-Agenten:

| Agenten-Framework | Zweck | Integration |
|---|---|---|
| **LangChain** | Workflow-Framework | Orchestrator-Integration |
| **AutoGen** | Multi-Agenten-Systeme | Für komplexe Tasks |
| **CrewAI** | Rollen-basierte Agenten-Teams | Für Team-Automatisierung |
| **OpenClaw** | Lokaler Agent mit Aktionen (E-Mails, CLI, etc.) | Eine Referenz unter vielen |
| **Nextcloud LLM2** | Offizielle Nextcloud-KI-App | Basis-Integration |

**Hinweis:** Agenten-Frameworks demonstrieren lokal-first Architekturen mit Multi-Channel-Integration. Die Integration in den Orchestrator erfolgt unter Beibehaltung der M1–M5-Sicherheitskette: Auch Agenten-Aktionen, die externe Dienste aufrufen, durchlaufen das HITL-Gate.

---

# 5. Hardware- & Software-Referenz

## 5.1 Hardware-Klassifikation

| Kategorie | Beispiele | Einsatz | Kosten (ca.) (Stand Q1/2026) | Empfehlung |
|---|---|---|---|---|
| **Consumer GPU** | RTX 4060 Ti 16GB | Lokale Inferenz (8B–13B) | 500–700€ | iKi-Standard |
| **Enthusiast GPU** | RTX 4090 24GB | Lokale Inferenz (bis 35B) | 1.500–2.000€ | Upgrade-Option |
| **Datacenter GPU** | A100 40/80GB, H100 | Training, große Inferenz | 10.000–30.000€ | Cloud mieten empfohlen |
| **Cloud-GPU** | RunPod, Vast.ai | Fine-Tuning, Training | 1–5€/Stunde | Für Fine-Tuning-Szenarien |
| **Mac Studio** | M2/M3 Ultra | Lokale KI (macOS) | 4.000–8.000€ | macOS-Option |
| **Intel NUC** | Mini-PC | Kompakte iKi-Hosting | 800–1.500€ | KMU-Setup |

**Fazit:** Für KMU-Infrastrukturen ergeben sich hardwareseitige Grenzen für lokale Inferenz (16–36GB VRAM). Für hochperformante Cloud-KI (Grok-4, GPT-5, Qwen-397B) ist Cloud-Nutzung erforderlich. 16GB VRAM ermöglichen lokale Inferenz für Modelle im Bereich 8B–13B Parameter, was für Standardaufgaben ausreichend ist.

## 5.2 Automatisiertes Deployment-Szenario

Das System ist als **automatisiertes Deployment** implementierbar:

1. **Installieren** (Setup-Skript auf Server/PC)
2. **Serverplatz freigeben** (Plattform-Instanz)
3. **Plattform mitsamt KI-Orchestrator + vorkonfigurierte Prompts/Skripte installieren**
4. **Internes Modell herunterladen** (z.B. Llama-3.1)
5. **Interne Dateien einbeziehen und indexieren lassen** (RAG)
6. **RAG automatisch konfigurieren lassen**
7. **Weitere Mitglieder einbinden** (Berechtigungen via RBAC)
8. **Auf weiteren Rechnern installieren** (Sync-Clients)

**Hinweis:** Diese Systematik ist durch diese Defensive Publication als Prior Art gesichert. Die Implementierung ist automatisierbar; der konkrete Aufwand hängt von der individuellen Infrastruktur ab.

---

# 6. Verfahren zur Datensicherheit (Defense-in-Depth)

## 6.1 Lokale Vorfilterung (iKi-Ebene) – 4-Stufen-Pipeline

Bevor Daten das lokale Netzwerk verlassen, durchlaufen sie eine interne KI-Schicht (iKi) mit folgender 4-Stufen-Pipeline:

**Stufe 1 – Regelbasierte Erkennung (NER/Regex):**

- Erkennung von E-Mail-Adressen, Telefonnummern, IBAN, Postadressen, Personennamen via NER-Modell (z.B. spaCy de_core_news_lg) und regulären Ausdrücken
- Ersetzung durch typisierte Platzhalter-Token: [PERSON_1], [ORG_1], [LOC_1], [EMAIL_1], etc.

**Stufe 2 – LLM-basierte semantische Umformulierung:**

- Die iKi (lokales Modell, ≥8B Parameter) erhält die Anfrage mit System-Prompt zur Abstrahierung von Beziehungsstrukturen und Kontext-Hinweisen
- Beispiel-Transformation:
  - Input: "Mein Bruder Hans Müller hatte folgenden Traum..."
  - Output: "[PERSON_1] aus dem persönlichen Umfeld der anfragenden Person berichtete folgenden Traum..."
- Ziel: Erhalt des semantischen Gehalts bei gleichzeitiger Trennung des Bezugs zur ursprünglichen Identität und zu Beziehungsstrukturen

**Stufe 3 – Reversibles Mapping:**

- Platzhalter werden in lokaler Lookup-Tabelle (RAM-only, session-scoped) gespeichert
- Nach cKi-Antwort erfolgt Re-Identifikation durch Rückersetzung der Platzhalter
- Lookup-Tabelle wird nach Session-Ende gelöscht (kein persistentes Speichern)

**Stufe 4 – Verifikation:**

- Sekundäres lokales Modell prüft anonymisierten Output auf Reststellen erkennbarer PII (Cross-Check)
- Bei Detektion von Reststellen: Rückführung zu Stufe 1 oder Eskalation an HITL-Gate

**Konfigurierbarer Schwellwert:** Der PII-Confidence-Score $\theta_{PII}$ ist konfigurierbar (Default: 0.3). Anfragen mit Score > $\theta_{PII}$ werden zwingend dem HITL-Gate zugeführt.

## 6.2 Human-in-the-Loop (HITL) – Technische Spezifikation

Der Nutzer behält die finale Entscheidungsgewalt über jeden Datenfluss. Es geschieht keine automatische Weiterleitung an externe Dienste ohne Nutzerbestätigung.

**Trigger-Bedingungen für das HITL-Gate:**

Das HITL-Gate wird aktiviert, sobald mindestens eine der folgenden Bedingungen erfüllt ist:

(i) Eine Anfrage zur Übermittlung an cKi vorgesehen ist (zwingend)
(ii) Der lokale PII-Detector einen Confidence-Score > θ_PII zurückgibt (konfigurierbarer Schwellwert)
(iii) Der RAG-Retrieval Dokumente mit erhöhter Sensitivitäts-Markierung zurückgibt
(iv) Der Nutzer manuelle Validierung in den Einstellungen aktiviert hat

**Diff-Anzeige:**

Dem Nutzer werden präsentiert:

- Die Original-Anfrage (lokal)
- Die anonymisierte cKi-Anfrage nach 4-Stufen-Pipeline
- Die retrieved RAG-Chunks (mit RBAC-Filterung)
- Visuelles Diff (Inline-Markup): Entfernte/ersetzte Tokens werden farblich hervorgehoben

**Approval-Modi:**

- **Single:** Einzelfreigabe pro Anfrage (Default für sensible Kontexte)
- **Session:** Freigabe für definierte Zeitspanne (max. 60 min, konfigurierbar)
- **Pattern:** Freigabe für strukturell ähnliche Anfragen (Embedding-Distanz < ε)

**Bearbeitungsoptionen:**

Nach Anzeige der anonymisierten Anfrage kann der Nutzer:

- Die Anfrage **direkt bestätigen** und versenden
- Die Anfrage **manuell editieren** (korrigieren, präzisieren, weiter anonymisieren)
- Die Anfrage **ablehnen** (kein Versand an cKi; ggf. Fallback auf reine iKi-Verarbeitung)

**Audit-Log (HITL-Logging):**

Jede HITL-Entscheidung wird protokolliert:

- Hash der Original-Anfrage (SHA-256)
- Hash der anonymisierten Anfrage (SHA-256)
- Zeitstempel
- Nutzer-ID
- Entscheidung (Bestätigt / Editiert / Abgelehnt)
- Modell-Routing (Ziel-cKi)

Persistierung in append-only Log (revisionssicher), Aufbewahrung gemäß organisatorischer Vorgaben. Das Log dient als Grundlage für QM-Auditierbarkeit (z.B. ISO 9001) und Datenschutz-Nachweispflichten.

**Ziel:** Der Nutzer hat die finale Kontrolle über jeden Datenabfluss an externe Dienste; das System gewährleistet, dass keine Daten ohne explizite Nutzeraktion übertragen werden.

## 6.3 Anbieter-Souveränität & Routing-Strategie

Das System vermeidet Abhängigkeiten von einzelnen Anbietern durch aktive Diversifizierung.

**Modell-Wahl:** Der Nutzer definiert die Regeln für Modell-Hersteller (Google, Microsoft, Europäische/Asiatische Anbieter). Der Orchestrator wendet diese Regeln an.

**Infrastruktur-Diversifizierung:** Nutzung von Gateways, die Anfragen über mehrere Rechen-Dienstleister nach konfigurierbaren Regeln verteilen. Durch die Verteilung der Anfragen auf verschiedene Rechenstandorte und Anbieter wird die Korrelation einzelner Anfragen zu einem Nutzerprofil statistisch erschwert.

**ZDR-Auswahlkriterium:** Primäre Auswahl von Anbietern mit Zero-Data-Retention (ZDR)-Richtlinien.

## 6.4 Multi-Level-Guardrails (Richtlinien-Ebenen)

| Ebene | Maßnahme | Ziel |
|---|---|---|
| **1 (ZDR-Policy)** | Gateway speichert keine Anfragen zu Trainingszwecken | Inhaltsschutz |
| **2 (Regionale Compliance)** | Rechenstandorte innerhalb Jurisdiktionen (z.B. EU-Only) | Datenschutz |
| **3 (Inhaltsfilter)** | Zusätzliche PII-Kontrolle auf Gateway-Ebene | Moderation |
| **4a (Policy-Kontrolle pro Schlüssel)** | Pro API-Schlüssel können individuelle Guardrails (ZDR-Policy, Region, Inhaltsfilter) konfiguriert werden, sodass unterschiedliche Themenbereiche oder Teams getrennten Sicherheitsrichtlinien unterliegen können. | Differenzierte Sicherheitspolitik |
| **4b (Mandantenfähige Key-Verwaltung)** | Soweit von den jeweiligen Anbieter-AGBs gedeckt, kann eine organisationsweite Key-Verwaltung mit anbieterseitigem Multi-User-Support (z.B. Anthropic Workspaces, OpenAI Organizations) genutzt werden, um Endnutzer-Identitäten gegenüber dem Inferenz-Anbieter zu abstrahieren. | Identitäts-Abstraktion |

**Hinweis zu 4b:** Diese Maßnahme setzt die ausdrückliche Zulässigkeit gemäß den Allgemeinen Geschäftsbedingungen des jeweiligen Anbieters voraus. Die Verwendung anbieterseitiger Multi-User-Funktionen ist von einfachem Key-Sharing zu unterscheiden, welches in der Regel gegen Nutzungsbedingungen verstößt.

## 6.5 Netzwerk-Sicherheit (Perimeter)

| Maßnahme | Umsetzung |
|---|---|
| **Keine offenen Ports für allgemeinen Internet-Zugriff** | Keine Ports werden für unbefugten Zugriff aus dem öffentlichen Internet geöffnet |
| **VPN-Tunnel** | Zugriff über verschlüsselte VPN-Tunnel (z.B. Tailscale, WireGuard, oder vergleichbare Lösungen) |
| **VLAN-Trennung** | Zwischen KI-Systemen, Nutzer-Clients und Server-Infrastruktur |

**Zugriffskontrolle:** Die Authentifizierung und Autorisierung erfolgt zentral über die Plattform. Für den Zugriff auf KI-Ressourcen und sensible Bereiche wird **RBAC (Role-Based Access Control / Rollenbasierte Zugriffssteuerung)** eingesetzt. Dies stellt sicher, dass Nutzer nur die Ressourcen sehen und nutzen können, die ihrer Rolle entsprechen (z.B. Admin, Nutzer, Gast).

## 6.6 Admin-Ethik & Datenschutz

**Problem:** Als Administrator der Plattform hat der Systembetreiber technisch Zugriff auf alle Nutzerdaten.

**Lösungsansätze:**

| Ansatz | Beschreibung | Priorität |
|---|---|---|
| **Transparente Kommunikation** | - Nutzer werden informiert, dass Admin technischen Zugriff haben *könnte*<br>- Zugriff erfolgt nur im Notfall oder bei technischem Problem, mit gleichzeitiger Information des Nutzers<br>- Logs über Datenzugriffe können dem Nutzer auf Wunsch ausgegeben werden | Hoch |
| **Vertragliche Absicherung** | AGB: Admin verpflichtet sich vertraglich zur Vertraulichkeit; Zugriff ausschließlich im technischen Notfall | Hoch |
| **Verschlüsselung (Client-Side)** | Nutzer nutzen Client-side Verschlüsselung für sensible Daten (z.B. Cryptomator) | Mittel |
| **Getrennte Instanzen** | Nutzer betreiben eigene Instanz via Föderation | Mittel |
| **Logging & Audit** | Plattform-Logging aktiviert, regelmäßige Checks | Mittel |

## 6.7 Backup-Konzept (3-2-1-Regel)

| Ebene | Maßnahme |
|---|---|
| **3 Kopien** | Original + 2 Backups |
| **2 Medien** | SSD + HDD (oder Cloud + Lokal) |

| Ebene | Maßnahme |
|---|---|
| **1 Offsite** | Externer Standort (z.B. Managed Hosting + lokale SSD) |

**Datenarten:**

- **Plattform-Daten:** Managed Hosting Backup + wöchentlich auf lokale HDD
- **Persönliche Vaults:** Sync mit Plattform + separates Backup auf zweiter SSD
- **Vektor-Datenbanken:** Export und Sicherung in Backup-Routine
- **HITL-Audit-Logs:** Revisionssichere Aufbewahrung gemäß QM-Vorgaben

---

# 7. Zukunftsszenarien & Erweiterbarkeit

## 7.1 Mobile Roboter & Privacy

Das System ist erweiterbar auf mobile Endgeräte (Roboter, Drohnen, ...) mit iKi+cKi-Orchestrierung. Privacy-Anforderungen sind dabei identisch: PII-Filter vor Cloud-Sendung, HITL für sensible Tasks.

## 7.2 AGI-Szenario

Bei fundamentalem KI-Fortschritt (AGI) bleibt die Kernarchitektur (lokal + Cloud, HITL, SSOT, Multi-Level-Guardrails) anwendbar. Die Modell-Ebene ist austauschbar.

## 7.3 Skalierung

| Größe | Architektur |
|---|---|
| **Einzelnutzer** | iKi auf lokalem PC, cKi via Gateway |
| **KMU (<500 MA)** | Zentraler Server, mehrere Clients, Team-Berechtigungen |
| **Föderation** | Instanzen verbinden sich sicher miteinander → Dezentrales Netzwerk |

## 7.4 Modell-Anpassung und Fine-Tuning

Das System ermöglicht die Integration von feinabgestimmten Modellen (Fine-Tuning) auf Basis lokaler Daten, sofern entsprechende Hardware-Ressourcen verfügbar sind. Dies umfasst:

- **Lokales Fine-Tuning:** Anpassung von Open-Weight-Modellen (z.B. Llama, Mistral) auf domänenspezifische Daten unter Beibehaltung der Datensouveränität
- **Cloud-basiertes Training:** Nutzung externer Trainingsdienste mit vorheriger Anonymisierung sensibler Daten gemäß Abschnitt 6.1

- **Modell-Versionierung:** Nachverfolgbarkeit von Modell-Updates und deren Auswirkungen auf Systemverhalten

Diese Funktionalität erweitert die Architektur um die Möglichkeit, KI-Modelle kontinuierlich an organisationsspezifische Anforderungen anzupassen, ohne die Kernprinzipien der Datensouveränität zu verletzen.

---

## 8. Lizenzierung & Patentschutz

| Aspekt | Spezifikation |
|---|---|
| **Lizenz für Dokumentation** | CC BY-NC-SA 4.0 |
| **Wirkung** | Dritte müssen Urheber nennen, keine kommerzielle Nutzung des Dokuments, gleiche Lizenz für Abwandlungen |
| **Patentschutz (Prior Art)** | Defensive Publication (Zeitstempel via Zenodo + GitHub + eigene Domain) |
| **GitHub-Strategie** | Code-Referenz (Setup-Skripte), synchron mit eigener Homepage |
| **Vorgängerversion v3.0** | Veröffentlicht 15.04.2026 (Zenodo + eigene Domain) – sichert die Kernarchitektur als Prior Art |

**Wichtig:** Nicht die Lizenz schützt vor Patenten, sondern die **Veröffentlichung selbst** (Zeitstempel). Die CC BY-NC-SA 4.0 Lizenz gewährt keine Patentlizenzen.

---

## 9. Schlussfolgerung

Durch diese Veröffentlichung wird der Stand der Technik/Wissenschaft für eine Wissensmanagement-Systemarchitektur mit hybrider KI-Integration für datensouveräne Organisationen und Einzelanwender mit dezentraler Netzwerk-Funktionalität etabliert.

Das spezifische, sequenzielle Zusammenspiel der Merkmale M1–M7 (siehe Abschnitt 3) – insbesondere:

- **Lokaler Orchestrierung (iKi) mit 4-stufiger PII-Filter-Pipeline (M1, M2)**
- **Deterministisches HITL-Gate mit Trigger-Mechanismus, Diff-Visualisierung und Audit-Logging (M3)**
- **RAG-Anbindung an selbstgehostete SSOT-Plattform mit RBAC-Berücksichtigung beim Retrieval (M4)**
- **Multi-Provider-Routing mit ZDR-Policy und mandantenfähiger Key-Verwaltung (M5)**

- **Optionaler Föderationsschicht über VPN (M6)**
- **Optionaler Multimodal-Erweiterung unter Beibehaltung der Sicherheitskette (M7)**

stellt eine Lösung für das Problem der Datensouveränität bei gleichzeitiger Nutzung moderner KI-Leistung dar.

Als auditierbare Orchestrierungsarchitektur unterstützt das System Qualitätsmanagement-Anforderungen (Auditierbarkeit, Konsistenz, Dokumentation) und ermöglicht die technische Abbildung von Prozessstandards (z.B. ISO 9001) auf technischer Ebene.

---

## 10. Quellen & Referenzen


| ID | Quelle | Beschreibung | Link |
|---|---|---|---|
| [1] | OpenRouter State of AI 2025/2026 | Marktpräferenzen für Multi-Model-Strategien | openrouter.ai/state-of-ai |
| [2] | Red Hat: Autonome KI-Agenten | Trend zu modularen Agenten-Architekturen | redhat.com/de/resources/autonomous-ai-agents |
| [3] | Paperclip (selbstgehostet) | DSGVO-konforme KI-Agenteninfrastruktur | github.com/paperclip-ai/paperclip |
| [4] | DSK Orientierungshilfe RAG | Datenschutzrechtliche Besonderheiten generativer KI-Systeme (Okt 2025) | dsk.bund.de |
| [5] | Praxisleitfaden KI & Datenschutz v2.0 | Datenschutz-Richtlinien für KI-Systeme | ki-datenschutz.de |
| [6] | EU-Kommission FAQ AI Literacy | Art. 4 KI-VO (07.05.2025) | commission.europa.eu |
| [7] | EU KI-Omnibus | Fristen für Hochrisiko-KI-Systeme (ab 2. August 2026) | consilium.europa.eu |
| [8] | Ollama | Lokale Modellausführung | ollama.com |
| [9] | Nextcloud LLM2 | Offizielle Nextcloud-KI-App | nextcloud.com/ai |
| [10] | LangChain | Workflow-Framework für KI-Agenten | langchain.com |
| [11] | AutoGen | Multi-Agenten-Systeme | microsoft.github.io/autogen |

| ID | Quelle | Beschreibung | Link |
|---|---|---|---|
| [12] | CrewAI | Rollen-basierte Agenten-Teams | crewai.com |
| [13] | OpenClaw | Lokaler Agent (Referenz unter vielen) | github.com/openclaw |
| [14] | ChromaDB | Vektor-Datenbank für RAG | trychroma.com |
| [15] | Qdrant | Vektor-Datenbank für RAG | qdrant.tech |
| [16] | Tailscale | VPN für sichere Verbindungen | tailscale.com |
| [17] | WireGuard | VPN-Protokoll | wireguard.com |
| [18] | Cryptomator | Client-Side-Verschlüsselung | cryptomator.org |
| [19] | spaCy | NER-Bibliothek für PII-Erkennung | spacy.io |
| [20] | NWM v3.0 (Vorgänger) | Defensive Publication v3.0 vom 15.04.2026 | (eigene Domain + Zenodo) |
| [21] | NWM v3.1 (Vorgänger) | Defensive Publication v3.1 vom 22.04.2026 | (eigene Domain + Zenodo) |

*... mit dem Sinn für Verbindungen und Zusammenhänge.*