



GEM CORTEX AI

– TECHNISCHE UND BETRIEBLICHE SPEZIFIKATION

Grundprinzip: Generative AI ist in Cortex ein kontrollierter Bestandteil eines Unternehmenssystems: Sie erhält eine Identität, Berechtigungen, Quellen, Limits, Messverfahren und die sichere Möglichkeit, abzulehnen oder zu eskalieren.



Technische Zusammenfassung für das Management

- ✓ Jede Anfrage wird im Kontext der Identität, des Zwecks, der Datenklassifizierung, der zugelassenen Tools und des Budgets bewertet.
- ✓ Der deterministische Verarbeitungsweg hat Vorrang, wenn eine Datenbank, eine Regel oder eine API ein genaueres, kostengünstigeres und besser auditierbares Ergebnis liefert als ein LLM.
- ✓ Wissen, Modelle, Tools und die Steuerungsschicht sind so voneinander getrennt, dass sie unabhängig geändert, getestet und jeweils unter eigener Betriebsverantwortung betrieben werden können.
- ✓ Aktionen mit Auswirkungen auf ein Unternehmenssystem unterliegen expliziten Berechtigungen, Limits, Protokollierung und je nach Risiko einer menschlichen Freigabe.
- ✓ Die Qualität wird anhand des akzeptierten Ergebnisses, der Quellenangaben und der Kosten pro Einheit bewertet, nicht anhand des Sprachflusses.

Minimaler Produktkontrakt

- ✓ **Eingabe:** Identität oder Workload, Zweck, Datenklassifizierung, Korrelations-ID und Kostenkontext.
- ✓ **Entscheidung:** Gewählter Laufzeitpfad, Modell/Tool, angewendete Policy.
- ✓ **Ergebnis:** Antwort, strukturierter Entwurf oder Aktionsstatus einschließlich Quellen.
- ✓ **Kontrolle:** Ablehnung, menschliche Freigabe, Tool-Limits, Timeout, Rollback.
- ✓ **Nachweis:** Audit-Trail, Telemetrie, Evaluationsmetrik und Kostenzuordnung.



1. Referenzarchitektur und Verteilung der Verantwortlichkeiten

Die Architektur ist logisch und in Schichten aufgebaut. Sie dient der Trennung von Verantwortlichkeiten, damit Identität, die Entscheidung über den Verarbeitungsweg, die Arbeit mit Wissen, Tool-Aufrufe, die Modelllaufzeitumgebung und die Betriebskontrolle unabhängig gesteuert werden können. Dies erleichtert Sicherheitsreviews, das Testen von Änderungen und die vertragliche Aufteilung des Betriebs.

- **Kanäle und Konsumenten:** Workspace, Portale, API, Batch. Der Kunde muss keinen weiteren Chat einführen.
- **Access:** SSO, Rollen, Nutzer- und Workload-Identität, Datenprofil. AI erbt den Berechtigungskontext.
- **Core & Flow:** Klassifizierung der Absicht, Auswahl des Laufzeitpfads, Policy und Workflow. Dieselbe Eingabe wird durch eine Regel, RAG, ein Modell oder ein Tool bearbeitet.
- **Knowledge & Tools:** Ingest, Index, Retrieval, Quellenangaben, Konnektoren und Tool-Register. Wissen hat einen eigenen Lebenszyklus.
- **Model Execution:** Anbieteradapter, lokale Inferenz, Routing und Fallback. Das Modell ist eine austauschbare Komponente.
- **Control:** Audit, Telemetrie, Evaluation, Kosten und Budgets. Gemeinsamer Nachweis für CIO, CISO und die Finanzabteilung.

Daten- und Steuerungsebene: Die Datenebene überträgt die konkrete Anfrage und das Ergebnis. Die Steuerungsebene verwaltet Konfigurationen, Kataloge, Policies und Budgets. Die Inhalte von Prompts, Dokumenten und Antworten eignen sich nicht automatisch für reguläre Anwendungslogs; die Telemetrie wird minimiert und klassifiziert.

2. Laufzeitpfade und Entscheidungslogik

Der Laufzeitpfad beschreibt, durch welchen Mechanismus das Ergebnis entsteht. Die Auswahl soll sich nicht an der Attraktivität des Modells aus Marketingsicht orientieren, sondern an der Art der Aufgabe, der Verfügbarkeit der Informationsgrundlagen, der erforderlichen Genauigkeit, der Berechtigung, der Latenz und der Wirtschaftlichkeit pro Einheit. Der Router liefert neben dem Ergebnis auch eine maschinenlesbare Begründung für den gewählten Pfad.

- **A – Direct / no LLM:** Für einen bekannten Lookup, eine Regel oder eine exakte Berechnung. Ausgabe aus DB/API. Kontrolle: Schema, Berechtigungen, Timeout. Geringste Variabilität, testbar und auditierbar.
- **B – RAG + LLM:** Für eine Antwort, die internes Wissen erfordert. Text mit Quellenangaben. Kontrolle: ACL, Retrieval-Evaluation, Abstention. RAG ist kein Heilmittel gegen Halluzinationen; die Qualität wird durch die ACL und den Evaluationsdatensatz bestimmt.
- **C – General model:** Für eine allgemeine Transformation ohne interne Daten. Textausgabe. Kontrolle: Datenprofil, Filter. Verringert das Risiko, eine allgemeine Antwort mit einer Stellungnahme des Unternehmens zu verwechseln.

- **D – Tool / Workflow:** Für das Lesen aus einem Unternehmenssystem oder dessen Änderung. Tool-Ergebnis. Kontrolle: Scope, Freigabe, Idempotenz, Audit. Schreibvorgänge und wesentliche Aktionen erfolgen mit einem Limit, einer Vorschau und menschlicher Freigabe.
- **E – Batch / Document:** Für große Dateimengen oder lang laufende Aufgaben. Datensatz, Dokumententwurf. Kontrolle: Warteschlange, Checkpoint, menschliche Prüfung. Skalierung von Dokumentenprozessen außerhalb des interaktiven Chats.

3. AI-Typen und Model Execution

Cortex trennt die erforderliche Fähigkeit vom konkreten Modell. Das größte oder teuerste Modell ist nicht automatisch das beste; Klassifizierung, Extraktion, Suche und Synthese stellen unterschiedliche Anforderungen.

- **LLM:** Synthese, Zusammenfassung, Dialog. Flexible, aber variable Ausgabe.
- **Kleineres Sprach- oder Klassifizierungsmodell:** Klassifizierung, Feldextraktion. Geringere Latenz.
- **Embedding + reranker:** Semantische Suche. Höhere Relevanz, jedoch empfindlich gegenüber ACL und Poisoning.
- **OCR / multimodales Modell:** Scans, Formulare, Tabellen. Die Qualität hängt vom Bild ab.
- **Speech:** Sprachtranskription. Erfordert Einwilligung und einen sicheren Handoff.
- **Agentensystem:** Mehrstufiger Plan und Tool-Nutzung. Ausschließlich bounded autonomy: begrenztes Ziel, begrenzte Schritte, Zeit, Kosten, Tools und Freigaben.

Modellregister und Routing: Jeder Endpunkt hat einen Verantwortlichen, eine Version, eine Region, einen Zweck, eine zulässige Datenklassifizierung und ein Preisprofil. Das Routing bewertet die mindestens erforderliche Fähigkeit, das Sicherheitsprofil, die Qualität auf einem konkreten Datensatz und die Kosten. Ein Fallback darf den Datenverarbeitungsmodus nicht stillschweigend ändern. Ein Modellwechsel ist eine Änderung im Produktivbetrieb und unterliegt einer Regressionsevaluation.

4. Daten und Wissensschicht (RAG)

RAG im Produktivbetrieb ist kein einmaliges Hochladen von Dokumenten in eine Vektordatenbank. Es handelt sich um einen kontrollierten Lebenszyklus der Quelle, von der Freigabe über die Indexierung bis zur Aktualisierung und Löschung.

- **Onboarding:** Verantwortlicher, Zweck, Klassifizierung, Rechtsgrundlage und ACL. Eine Quelle ohne Verantwortlichen wird nicht in den produktiven Index aufgenommen.
- **Ingest:** Antimalware, Parsing, OCR. Eingabefehler werden von Modellfehlern getrennt.
- **Chunking:** Aufteilung nach Struktur, Bedeutung und Revisionsständen. Die Chunk-Größe beeinflusst sowohl den Recall als auch den Preis.

- **Index:** Lexikalischer und Vektorindex. Hybrides Retrieval kombiniert exakte Begriffe mit semantischer Ähnlichkeit.
- **Retrieval:** Filter für ACL, Tenant, Rolle und Dokumentenrevisionsstand. Nicht zugelassene Inhalte dürfen nicht in den Modellkontext gelangen.
- **Context build:** Deduplizierung, Längenbegrenzung, Reihenfolge und Quellenangaben.
- **Antwort:** Grounded Prompt, strukturierte Ausgabe, Quellenangaben und Abstention.
- **Freshness/delete:** Neuindexierung von Änderungen, Löschung und Weitergabe von ACL-Änderungen.

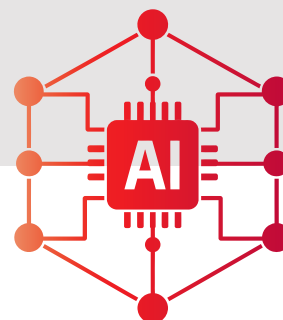
5. Integration, API und Tool-Kontrakt



Die Integration ist der wesentliche Unterschied zwischen einer isolierten AI-Demo und einem Unternehmensdienst. Cortex unterstützt mehrere Muster in Abhängigkeit von Latenz, Volumen, Transaktionalität und Betriebsverantwortung.

- **REST + OpenAPI:** Synchrones Lesen, validiertes Schreiben. Verbindliche Rahmenbedingungen: Timeout, Retry, Version, Autorisierung.
- **Event / message:** Lange Workflows, Batch-Verarbeitung. Verbindliche Rahmenbedingungen: Deduplizierung, Reihenfolge, Dead-Letter Queue.
- **Batch / file:** Große Dokumentensätze. Verbindliche Rahmenbedingungen: Manifest, Prüfsumme, Checkpoint.
- **MCP (Model Context Protocol):** Standardisierte Bereitstellung von Resources und Tools. Protokollversion, Scopes, Token-Audience, Schema und Tool-Risiko.
- **Database read model:** Exakte Lookups. Read-only-Konto, row-level security, Allowlist für Abfragen.
- **Webhook:** Benachrichtigungen über den Abschluss. Signatur, Replay-Schutz, Retry.

MCP-Nuancen: MCP kann den Kontrakt für Ressourcen und Tools vereinheitlichen, ist jedoch für sich genommen weder ein Identity Provider noch ein Datenkatalog oder ein Sicherheitsgateway. GEM schreibt die unterstützte Version fest, prüft die Token-Audience und trennt Read- und Write-Tools.



6. Identitäten, AI-Konten und Secret-Management

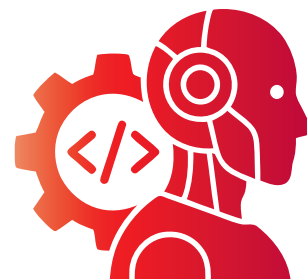
Das Identity-first-Modell bedeutet, dass Cortex die Sicherheit nicht erst nach der Generierung einer Antwort berücksichtigt. Der Nutzer oder Workload wird bereits am Eingang erkannt; sein Kontext wird an die Policy, das Retrieval und das Tool weitergegeben. Dadurch wird die Entstehung gemeinsam genutzter Konten, unbekannter API-Schlüssel und von Zugriffsrechten reduziert, die über einen Rollenwechsel oder das Ausscheiden eines Mitarbeiters hinaus bestehen bleiben.

- **Nutzer:** SSO über den unternehmensinternen IdP. Eigenes Unternehmenskonto, kein gemeinsam genutztes persönliches Passwort.
- **Workload identity:** Eigenständige Identität eines Backends, Agenten oder Konnektors. Trennung von Mensch und Dienst.
- **Provider-Tenant/Projekt:** Erfassung von Organisation, Projekt, Region, Verantwortlichem und Abrechnung.
- **API-Schlüssel / Secret:** Referenz auf den Vault, Rotation, Verbot der Speicherung im Code oder Prompt.
- **Servicekonto des Zielsystems:** Minimaler Scope, Ablaufdatum, Verantwortlicher.

Joiner-Mover-Leaver (JML) und Rezertifizierung: Beim Eintritt werden nur die erforderlichen Rollen angelegt, bei einem Positionswechsel werden die Berechtigungen neu berechnet, beim Austritt werden Sessions beendet und Delegationen entzogen. Die Automatisierung darf keine persönlichen Passwörter konsolidieren. Die Rezertifizierung erfolgt risikobasiert.

7. Sicherheit, Datenschutz und Governance

Das Sicherheitsmodell kombiniert herkömmliche Anwendungskontrollen mit Risiken, die für generative AI spezifisch sind. Das Referenz-Bedrohungsmodell basiert auf den OWASP Top 10 für LLM.



- **Prompt injection:** Trennung von Anweisungen und Daten, Bereinigung, Tool-Policy, minimale Scopes.
- **Sensitive information disclosure:** ACL vor dem Retrieval, Datenprofil, DLP/Schwärzung.
- **Supply chain:** Inventar, SBOM, Scanning und Lieferantenfreigabe.
- **Improper output handling:** Schema, Escaping, Parametrisierung und Validierung.
- **Excessive agency:** Bounded autonomy, Budget, Freigabe, Kill Switch.
- **Vector/embedding weaknesses:** Quellenfreigabe, Metadaten, Lineage, Neuindexierung.
- **Misinformation:** Quellenangaben, Abstention, kritische Validierung, menschliche Prüfung.
- **Unbounded consumption:** Rate Limit, Token-Budget, Timeout, Quoten.

Datenprofil und Betriebsmodus des Anbieters: Jeder Endpunkt verfügt über ein freigegebenes Anbieter-/Datenprofil: Zweck, Region, Datenklassifizierung, Aufbewahrungsdauer, Nutzung für das Training und Unterauftragnehmer. Eine Aufbewahrungsdauer von null oder ein Verbot der Trainingsnutzung darf nur für einen konkreten Dienst und eine konkrete vertragliche Konfiguration angegeben werden.

Governance und regulatorischer Kontext: Control führt ein AI-System-/Anwendungsfallregister mit Verantwortlichem, Zweck, Modellen, Datensätzen, Tools, Risiken, Änderungen und Abnahmenachweisen. Die Dokumentation kann die Steuerung nach NIST AI RMF, ISO/IEC 42001, GDPR, NIS2 und AI Act unterstützen. Die Plattform allein zertifiziert den Kunden nicht automatisch; die rechtliche Einstufung wird von der verantwortlichen Organisation bestätigt.

8. *Deployment-Modelle und ihre Zielkonflikte*

Das Deployment wird anhand des Datenprofils, der Netzwerkbeschränkungen, der Betriebskompetenzen, der Latenz und der Wirtschaftlichkeit gewählt. On-premise ist nicht automatisch sicherer oder kostengünstiger.

- **Customer cloud:** Infrastruktur und Dienste im Tenant des Kunden. Kontrolle über Netzwerk, Identität und Kosten.
- **GEM Managed cloud:** Von GEM betriebener Dienst. Schnellerer Start und gebündelte Betriebsverantwortung.
- **Hybrid:** Daten oder Tools intern, Inferenz in der Cloud. Kombination interner Systeme mit flexibler Kapazität.
- **On-prem / restricted:** Lokale Laufzeitumgebung, Modelle und Daten. Geeignet für eingeschränkte Konnektivität und einen spezifischen Datenverarbeitungsmodus.

Wirtschaftlicher Zusammenhang: Bei lokaler Inferenz fallen keine Kosten für Cloud-Token an; sie beansprucht jedoch Hardware, Energie, Teamkapazität, Lizenzen, Aktualisierungen und Reserven für Lastspitzen. Die TCO werden daher je akzeptiertem Ergebnis und Servicelevel verglichen.

9. *AI FinOps, Budgets und wirtschaftliche Telemetrie*



AI FinOps verknüpft den technischen Verbrauch mit dem Verantwortlichen, dem Anwendungsfall und dem akzeptierten Ergebnis. Ein günstiges Modell kann teure Nacharbeit verursachen; ein qualitativ besseres Modell kann wirtschaftlicher sein, wenn seine Ausgabe die Prüfung häufiger besteht.

- **Zuordnungsschlüssel:** Tenant, Projekt, Kostenstelle, Anwendungsfall, Umgebung, Modell, Team.
- **Budget:** Monatliches Soft Budget, Alarmierung, Freigabeschwelle und Hard Stop. Ein Hard Stop ist dort sicher, wo Cortex den gesamten Betrieb steuert.
- **Wirtschaftlichkeit pro Einheit:** Preis pro Dokument, Anfrage, Workflow oder akzeptiertem Entwurf.
- **Value ledger:** Baseline, genutztes Volumen, Akzeptanzquote, Veränderung des Zeitaufwands, Nacharbeit, TCO.

Budgetsteuerung: Cortex kann vor dem Aufruf den Umfang der Eingabe schätzen und das Budget prüfen. Schätzung und endgültige Rechnung können aufgrund von Caching, Batch-Rabatten oder verzögert verfügbaren Daten voneinander abweichen; im Rahmen des Finanzabschlusses erfolgt ein Abgleich. Ein Hard Limit weist weitere Anfragen zurück; ein Rate Limit schützt die Kapazität.

Optimierungshebel: Deterministischer Verarbeitungsweg, Modellrouting, Kontext, Cache, Batch-/asynchroner Modus. Eine Optimierung wird nur übernommen, wenn sie die Abnahmemetrik nicht unter den festgelegten Schwellenwert senkt.

10. Nichtfunktionale Anforderungen, SLO und Resilienz

Eine einheitliche Verfügbarkeit für Cortex insgesamt wäre irreführend. Ein direkter Datenbank-Lookup, interaktives RAG und Batch-Verarbeitung haben unterschiedliche Eigenschaften. Das SLO wird je Laufzeitpfad und Kritikalität festgelegt.

- **Verfügbarkeit:** Anteil erfolgreich bearbeiteter Anfragen.
- **Latenz:** P50/P95/P99, time-to-first-token. Retrieval, Modell und Tool getrennt messen.
- **Durchsatz:** Gleichzeitige Anfragen, Dokumente pro Stunde.
- **Qualität:** Rate kritischer Fehler, Groundedness, Akzeptanz des Entwurfs.
- **Freshness:** Verzögerung bei der Synchronisierung und Löschung der Quelle.
- **Wiederherstellung:** RTO, RPO, Sicherung der Konfiguration und des Indexes.
- **Kapazitätsbedingte Degradation:** Fallback-Modell, Warteschlange. Die Degradation darf das Datenprofil nicht verändern.



Portabilität und Exit: Das Konzept für den Produktivbetrieb enthält einen geregelten Ausstieg: Export von Konfiguration, Prompts, Evaluationsdatensätzen, Audit-Aufzeichnungen und Kostenzuordnungen. Zugleich beschreibt es, was aufgrund der Lizenz des Anbieters nicht exportiert werden kann.

11. *Observability, Änderungsmanagement und Betrieb*

Der Betrieb eines AI-Dienstes erfordert die Verknüpfung klassischer Observability mit der Qualität der Ausgaben und den Kosten. Eine einzige Korrelations-ID verfolgt die Anfrage über Gateway, Retrieval, Modell, Tool und Zielanwendung hinweg.

- **Metriken:** Anfragen, Fehler, Latenz, Token, Kosten, Cache, Warteschlange, Tool-Erfolg.
- **Trace:** Orchestrierungsschritte, Modell-/Prompt-Version, Router-Entscheidungen.
- **Log:** Audit von Zugriffen, Freigaben und Konfigurationsänderungen.
- **Eval:** Online-Feedback, Golden Set, kritische Fehler.
- **Business event:** Akzeptierter Entwurf, abgeschlossener Fall, Nacharbeit.

Änderungsmanagement: Eine Änderung des Modells, des Prompts, der Daten oder des Tools erfordert einen Verantwortlichen, eine Offline-Evaluation, ein Sicherheitsreview, eine schrittweise Bereitstellung, eine Freigabe und Monitoring nach der Bereitstellung.

Incidents und Runbooks: Verfügbarkeits-, Qualitäts-, Zugriffs-, Agenten- und Finanzvorfälle.

Referenz-RACI: Business outcome und Prozess: Kunde (Accountable), GEM (Responsible).
Datenklassifizierung: Kunde. Anwendungsschicht: GEM. Modellendpunkt: Vertraglicher Betreiber. Qualität: GEM. Incident: Service owner.

12. *Implementierungsweg und Abnahme*

Die Implementierung reicht von der Entscheidung, ob ein Anwendungsfall sinnvoll ist, bis zum Nachweis, dass er sicher betrieben werden kann.

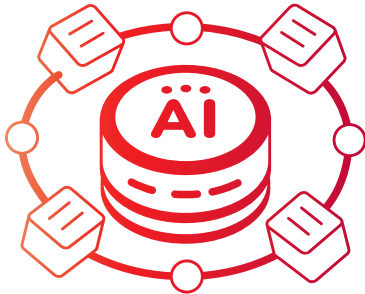
- **0. Qualification:** Geschäftliches Problem, Sponsor, Datenverfügbarkeit, Risiko.
- **1. Discovery:** Baseline, Datenfluss, Integration, Evaluationsdatensatz.
- **2. Pilot:** Begrenzter Scope, draft/read-only, Messung gegenüber der Baseline.
- **3. Limited production:** Reale Identitäten, Monitoring, Support.
- **4. Scale:** Weitere Anwendungsfälle, Automatisierung der Governance.



Abnahmepaket: Funktionale, qualitative, sicherheitsbezogene, betriebliche, wirtschaftliche und übergabebezogene Abnahme. Der Autonomiegrad wird gesondert anhand der Auswirkungen freigegeben, nicht anhand der durchschnittlichen Genauigkeit.

13. Grenzen, Voraussetzungen und Entscheidungscheckliste

- **Eingabequalität:** Ein Scan von geringer Qualität und widersprüchliche Dokumente begrenzen die Genauigkeit.
- **Quellberechtigungen:** Ein Konnektor ohne ACL darf nicht als permission-aware dargestellt werden.
- **Provider/API:** Funktionen, Limits und Preise ändern sich; Adapter verringern die Auswirkungen, beseitigen sie jedoch nicht.
- **Modellvariabilität:** Dieselbe Eingabe führt nicht zwingend immer zu einem identischen Text.
- **Autonomie:** Irreversible und rechtlich relevante Schritte bleiben unter menschlicher Kontrolle.
- **Legacy-Integration:** Ohne Dokumentation kann die Integration den Großteil des Aufwands bestimmen.
- **Offline/on-prem:** Erfordert Kapazität, Patching und einen Modelllebenszyklus.
- **ROI:** Die Zeitersparnis lässt sich erst durch eine Änderung des Prozesses und der Kapazität realisieren.



Solution design checklist:

Sind die Aufgabe und die Baseline definiert? S

ind der Datenfluss und die Klassifizierung bekannt?

Sind Schreib- und Lesevorgänge voneinander getrennt?

Gibt es Negativtests?

Sind das SLO-Profil und der Exit bekannt?

