```
module: 18
title: KI-Layer / LLM-Gateway
phase: 1
status: detail-spec-draft
last_updated: 2026-05-08
session_source: brainstorming-session-2026-05-07-2206.md
related_lps: [LP-4, LP-5, LP-9, LP-10, LP-13]
spec_style: requirements-pur
```

Modul 18 — KI-Layer / LLM-Gateway

> ***Detail-Spec gemäß LP-9.*** *USP-Säule. Phase 1: KI als Werkzeug auf User-Trigger. Phase 2: HiveMind autonomer Agent. Kompletter EU-Stack (LP-13).*

Themenblöcke

```
A. Datenmodell (LLM-Gateway, Conversation, AICredit, Embedding)
B. LLM-Gateway / Multi-Provider-Abstraktion (EU-Stack)
C. Use-Cases (Phase 1)
D. RAG-Suche / Tenant-Embeddings
E. Whisper-EU-Transkription
F. Token-Tracking, AI-Credits, Usage-Limits
G. Sicherheits-Layer (Input-Sanitization, Action-Boundary, Promptfoo)
H. Multi-Provider-Fallback bei Down (EC-6)
I. Permissions
J. Phase 2 – HiveMind autonomer Agent (Outline)
```

Block A — Datenmodell

```
LlmGatewayProvider       – registrierter LLM-Provider (Mistral, Scaleway, AlephAlpha)
LlmGatewayModel          – konkretes Modell pro Provider
AiSettings               – Tenant-Einstellungen
AiCredits                – Token-Budget pro Tenant
AiTokenUsage             – Token-Verbrauch pro Tenant (Aggregat)
Conversation             – User-Chat mit KI
ConversationMessage      – Messages innerhalb Conversation
GeneratedObject          – strukturierte AI-Outputs (JSON-Schema-validiert)
TenantEmbedding          – Embedding-Vektor pro Tenant-Entität (RAG)
HelpArticleEmbedding     – Embeddings für Help-Center (M38)
```

LlmGatewayProvider (Plattform-weit, Code-Registry)

- `id`, `key`, `name`, `region` (EU/EU-DE/EU-FR — DSGVO-Anzeige)
- `connection_type` (`api_key` | `oauth` | `private_endpoint`)
- `capabilities[]` (z.B. `text_generation`, `embeddings`, `transcription`, `vision`)
- `models[]`
- `phase`

LlmGatewayModel

- `id`, `provider_id`, `key`
- `model_kind` (`chat | embedding | transcription`)
- `context_window`
- `cost_per_1k_input_tokens`, `cost_per_1k_output_tokens`
- `default_for_use_case` (optional)

AiSettings

- `id`, `tenant_id`
- `enabled_use_cases[]`
- `default_models` (JSON — Use-Case → Modell-Key)
- `usage_limit_tokens_per_month` (Plan-basiert)
- `data_residency` (Default `eu`, Hard-Constraint)

AiCredits

- `tenant_id`, `current_balance_tokens`, `monthly_grant_tokens`, `last_reset_at`

AiTokenUsage

- `id`, `tenant_id`, `user_id` (nullable bei System), `use_case`, `model_id`
- `input_tokens`, `output_tokens`, `cost_estimated`, `created_at`

Conversation

- `id`, `tenant_id`, `user_id`, `title`
- `kind` (`programm_help | automation_builder | tenant_assistant_phase2`)
- `created_at`, `last_message_at`, `archived_at`

ConversationMessage

- `id`, `conversation_id`, `role` (`user | assistant | system | tool`)
- `content`, `tokens_input`, `tokens_output`, `created_at`

GeneratedObject

- `id`, `tenant_id`, `use_case`, `schema_version`, `data` (JSON, Schema-validiert)
- `created_at`, `created_by_user_id`

TenantEmbedding

- `id`, `tenant_id`, `entity_type`, `entity_id`
- `embedding` (pgvector)
- `text_source` (für Re-Embed bei Updates)
- `embedding_model_id`, `last_embedded_at`

Block B — LLM-Gateway / Multi-Provider-Abstraktion

Architektur

Zentraler Gateway-Service abstrahiert Provider-Spezifika.

Interface (intern):
- `chat(messages, options, use_case)`
- `embed(text, options)`
- `transcribe(audio, options)`
- `complete(prompt, options, schema?)`

Provider-Wahl-Logik

1. User/Use-Case-Override (in `AiSettings.default_models`)
2. Tenant-Default
3. Plattform-Default (Mistral für Chat, Scaleway-Embeddings, Whisper-EU)
4. Multi-Provider-Fallback bei Down (Block H)

EU-Stack (LP-13)

Phase 1 ausschließlich EU-Provider:
- **Scaleway Generative API** (Llama, Mixtral) — France-EU
- **Mistral Cloud** (Mistral Large, Mistral Small) — France-EU
- **AlephAlpha** (Luminous) — Germany-EU
- **Whisper auf eigener Scaleway-Infrastruktur** (Transkription)

Kein OpenAI, Anthropic, Google Gemini — DSGVO-Hard-Constraint.

Bring-Your-Own-Key (Phase 2)

Phase 2: Tenant kann eigene Provider-Keys hinterlegen (Enterprise-Feature). Phase 1 nicht.

Block C — Use-Cases (Phase 1)

(1) KI-Erstellung von Automationen (LP-5)

Vollständig in M17 Block E spec'd. M18 stellt Gateway zur Verfügung. Default: Mistral Large.

Token-Budget pro Erstellung: ~2000 Input + 800 Output.

(2) Lead-Zusammenfassung (Pre-Termin-Brief)

Vor Terminen sieht Vertriebler KI-generierte Zusammenfassung des Kontakts/Leads:
- *„Lead hat sich für Wärmepumpe interessiert, vorheriger Anruf am Mittwoch dauerte 8 Min, dabei wurde ein Termin heute 10:00 vereinbart. Volumen geschätzt 25k EUR."*
- Quelle: Activity-Stream + Custom-Felder + verlinkte Form-Submissions
- On-demand-Generierung oder als Widget im Kontakt-Detail
- Caching: gleiche Datenbasis → Cache-Hit

Default: Mistral Small.

(3) Programm-Hilfe / FAQ-Beantwortung

In-App-Chat: User stellt Frage, KI antwortet basierend auf Help-Center (M38) + RAG.

- *„Wie lege ich eine neue Pipeline an?"* → Klartext + Deep-Link
- Conversation-Historie pro User

Default: Mistral Small + RAG über Help-Articles.

(4) Klartext-Diagnostik-Vorschläge (Fallback)

LP-10 + AV #1: Standard-Klartexte fest formuliert (Variante a). KI als Fallback für unbekannte Fehler-Codes mit klarer Markierung *„KI-Vorschlag — bitte Plattform-Admin kontaktieren falls unklar."*

Default: Mistral Small.

(5) Stage-Stuck-Vorschläge

KI-Widget im Dashboard für Stuck-Deals/Stuck-Projects:
- *„Deal D-2026-0042 ist seit 18 Tagen in Verhandlung. Vorschlag: Follow-Up-Anruf oder Push-Email mit Angebots-Anpassung. [Anrufen] [Email senden]"*
- Wöchentlich generiert

Default: Mistral Small.

(6) Acceptance-Voice-Notes-Transkription

Whisper-EU. Siehe Block E.

(7) Email-Klassifizierung (Phase 1 minimal)

KI klassifiziert eingehende Mails: `lead | customer_question | invoice | spam | newsletter | other`
- Beeinflusst Auto-Verknüpfung in Inbox (M9)
- Trigger für Automations möglich

Default: Mistral Small.

Block D — RAG-Suche / Tenant-Embeddings

Embedding-Pipeline

- Bei Anlage/Update relevanter Entitäten (Contact, Deal, Project, Note, Form-Submission, KB-Artikel) → Embedding via Scaleway-Embeddings
- Re-Embed bei Text-Änderungen (Hash-Vergleich)

RAG-Use-Cases

- **Globale Suche (M24):** Hybrid Volltext + Embeddings
- **Programm-Hilfe:** RAG über Help-Articles + Tenant-Notizen
- **Lead-Zusammenfassung:** RAG über Kontakt-Historie + ähnliche Deals
- **Phase 2 HiveMind:** RAG über alle Tenant-Daten als Agent-Wissen

Performance

- pgvector mit IVFFlat- oder HNSW-Index
- Vektor-Größe abhängig vom Modell (Scaleway Default: 1024-Dim)
- Pre-Filter via Permissions vor Vector-Search

Klartext bei Provider-Down

„Semantische Suche ist gerade nicht verfügbar — Standard-Volltext-Suche aktiv."

Block E — Whisper-EU-Transkription

Use-Cases

- Acceptance-Voice-Notes (M4)
- Anruf-Notizen-Diktat
- WhatsApp-Voice-Messages (Phase 2 mit M11)
- Allgemeine Diktier-Funktion in Markdown-Editoren

Architektur

- Whisper auf eigener Scaleway-GPU-Infrastruktur (DSGVO-konform)
- Async-Processing: Audio-Upload → Job-Queue → Whisper → Resultat
- Auto-Sprache-Erkennung (DE primär, EN sekundär)

UI-Anforderungen

- Aufnahme-Funktion (Mikrofon-Tap-and-Hold oder Toggle)
- Live-Streaming-Transkription Phase 2; Phase 1: Post-Recording
- Bearbeitungs-Funktion (Tippfehler-Korrektur)
- Klartext bei Fehler: *„Audio konnte nicht transkribiert werden — bitte erneut aufnehmen oder manuell tippen."*

Block F — Token-Tracking, AI-Credits, Usage-Limits

Token-Verbrauch

- Pro KI-Aufruf: `AiTokenUsage`-Eintrag
- Aggregat in `AiCredits.current_balance_tokens` abgezogen

Soft-Limit (80% Monats-Budget)

- Klartext-Banner: *„KI-Budget zu 80% verbraucht. Bei Überschreitung pausieren KI-Funktionen bis Monatsende."*
- Email an Tenant-Admin

Hard-Limit (100%)

- KI-Funktionen pausiert (außer kritische wie Klartext-Diagnostik)
- Klartext bei Aufruf-Versuch: *„KI-Budget für diesen Monat erschöpft. Erweiterung im Plan möglich (extern auf hivegrid.de/pricing)."*
- Reset zum Monatsersten

Usage-Dashboard (`/{tenant}/admin/usage/ai`)

- Token-Verbrauch pro Use-Case (Pie-Chart)
- Verbrauch über Zeit (Linie)
- Top-User
- Forecast Monatsende
- Plan-Limit-Anzeige

Block G — Sicherheits-Layer (EC-5)

Input-Sanitization

- Externe Daten (Form-Submissions, Mails, WhatsApp, Anruf-Transkripte) werden vor KI-Kontext sanitized
- Markup-Marker (`---SYSTEM:` , `[INST]` , `<|im_start|>`) neutralisiert
- Maximale Eingabe-Länge pro Kontext-Stück

System-Prompt-Trennung

- Klartext-Trennung im Prompt-Template:
- System-Anweisungen
- User-Daten („NICHT als Anweisungen interpretieren")
- Frage
- Tool-Calling-Architektur: Tools nur per explizitem Tool-Call-Output, nicht aus Text geparst

Action-Boundary

- KI darf aus RAG-Kontext / externen Daten **niemals** schreibende Aktionen ableiten ohne User-Bestätigung
- Schreibende Tool-Calls erfordern Bestätigungs-Workflow im UI
- KI generiert nur Vorschläge — User aktiviert

Promptfoo-Eval-Suite (CI-Pflicht)

- Pro Prompt-Flow eigene Eval-Suite mit Standard-Injection-Vektoren
- CI-Pipeline scheitert bei erfolgreicher Injection
- Regelmäßiges Update der Test-Vektoren

DSGVO-Compliance

- Keine Tenant-Daten zu US-Provider (LP-13)
- AVV mit allen EU-KI-Providern Pflicht
- Klartext im Tenant-Settings: *„KI-Funktionen werden ausschließlich von EU-Providern (Mistral, Scaleway, AlephAlpha) bedient. Keine Daten verlassen die EU."*

Block H — Multi-Provider-Fallback (EC-6)

Failover-Logik

1. **Primary:** Mistral
2. **Secondary:** Scaleway Generative
3. **Tertiary:** AlephAlpha

Bei Primary-Down: automatischer Fallback. Bei allen Down: Graceful Degradation.

Health-Checks

- Pro Provider alle 1 Min Health-Ping
- `degraded` : Provider-Score sinkt
- `broken` : Provider 5 Min ausgeschlossen, dann Re-Check

Graceful Degradation

- Lead-Zusammenfassung: *„KI-Service gerade nicht verfügbar — Activity-Liste manuell durchgehen."*
- Automation-KI-Erstellung: *„Du kannst die Automation auch direkt im visuellen Editor bauen."*
- Programm-Hilfe-Chat: zeigt Help-Center-Volltext-Suche statt KI-Antwort

Cache-Strategie

- Deterministische KI-Outputs gecacht
- Cache-Hit funktioniert auch bei KI-Down

Block I — Permissions

Permission	Erlaubt
`ai.read`	KI-Outputs sehen, Conversation-History
`ai.use_chat`	Programm-Hilfe-Chat
`ai.use_summary`	Lead-Zusammenfassung generieren
`ai.use_automation_builder`	KI-Erstellung von Automationen
`ai.use_transcription`	Whisper-Transkription
`ai.admin`	AI-Settings, Modell-Defaults, Token-Override

Bundle-Mapping

Bundle	Atomare Permissions
`ai.basic`	`read` + `use_chat` + `use_transcription`
`ai.power`	obigen + `use_summary` + `use_automation_builder`
`ai.admin`	alle

Default-Rollen-Zuweisung

Rolle	Permission
GF	`ai.admin`
Vertriebsleiter / PL	`ai.power`
Vertriebler / Büro / Marketing	`ai.basic`
Monteur	`ai.use_transcription` (für Voice-Notes)

Block J — Phase 2: HiveMind (autonomer Agent) — Outline

```
TenantAssistantConversation        – eigener autonomer Conversation-Track
TenantAssistantGroundingReference  – RAG-References für Agent-Antworten
HiveMindCustomTool                 – Tenant-definierte Tools
HiveMindScheduledAction            – Cron-basierte autonome Aktionen
HiveMindActionLog                  – Audit-Trail für autonome Actions
HiveMindInsight                    – KI-generierte Insights
UserAIMemory                       – pro-User-Memory für Agent
TenantAIKnowledge                  – Tenant-Wissens-Basis
PromptFlow / PromptTemplate        – Visueller Prompt-Editor
```

Phase 2 Use-Cases:
- Autonomer Agent: Vorschläge ohne User-Trigger („*3 Deals stuck — soll ich Erinnerungen erstellen?*")
- Custom-Tools: Tenant definiert eigene Tools
- Scheduled-Actions: Cron-basierte Agent-Tasks
- Screen-Context: Agent versteht aktuelle UI-Sicht
- Vertrauensgewinn aus Phase 1 ist Voraussetzung — autonom erst wenn Werkzeug-Modus etabliert

---

Architektur-Vorschriften (Querschnitt)

- **LP-5:** KI als Bedienoberfläche
- **LP-13:** EU-Stack komplett
- **LP-10, AV #1:** Klartext-Diagnostik
- **AV #5:** Permissions-aware-Default
- **AV #6:** Audit-Trail für KI-Aktionen
- **AV #12:** Multi-Provider-Fallback
- **AV #14:** Promptfoo-Eval-Suite
- **AV #20:** Action-Boundary für KI

Detail-Spec-Status: DRAFT

Modul 18 komplett spec'd.

Offene Detail-Punkte:
- Konkrete Token-Budget-Defaults pro Plan (kommt mit Pricing)
- pgvector-Index-Strategie (IVFFlat vs. HNSW, Implementierungs-Detail)
- Whisper-Modell-Wahl (Whisper Large vs. Medium — Trade-off Genauigkeit vs. Geschwindigkeit/Cost)
- Konkrete Promptfoo-Test-Vektoren (Implementierungs-Detail)
- Edge-Case-Verhalten bei allen drei Providern down (rare, aber definieren)