1. Allgemeine Übersicht

VOGO eBS AI Agentic Solutions ist das nativ in die VOGO eBS-Plattform integrierte KI-Modul, das entwickelt wurde, um Kommunikation zu automatisieren, Benutzer in Echtzeit zu unterstützen und komplexe Workflows in Enterprise-Umgebungen zu orchestrieren. Die Lösung kombiniert modernste LLMs mit agentischen KI-Architekturen — RAG, LangChain, LangGraph und MCP — und bietet Organisationen erweiterte KI-Fähigkeiten ohne Abhängigkeit von Drittanbieter-Plattformen.

2. Technische Architektur & LLM-Stack

Unterstützte LLM-Anbieter

Die Plattform unterstützt Konfiguration und Echtzeit-Umschaltung zwischen mehreren LLM-Anbietern ohne Code-Änderungen:

- Anthropic (Claude) — erweitertes Schlussfolgern, erweiterter Kontext, Enterprise-Sicherheit
- OpenAI (GPT-4o, GPT-4 Turbo) — Referenzmodell für konversationale Anwendungen
- Google Gemini — native Integration mit dem Google Workspace-Ökosystem
- Ollama — Open-Source-Modelle (LLaMA, Mistral, Qwen) lokal, on-premise ausführen, ohne Token-Kosten
- Groq — ultraschnelle Inferenz für latenzkritische Szenarien
- xAI (Grok) — alternatives Modell für spezifische Anwendungsfälle
- Mistral — europäisches Modell, relevant für strenge DSGVO-Konformität
- OpenRouter — einheitliches Gateway für Zugang zu Dutzenden von Modellen über eine einzige API

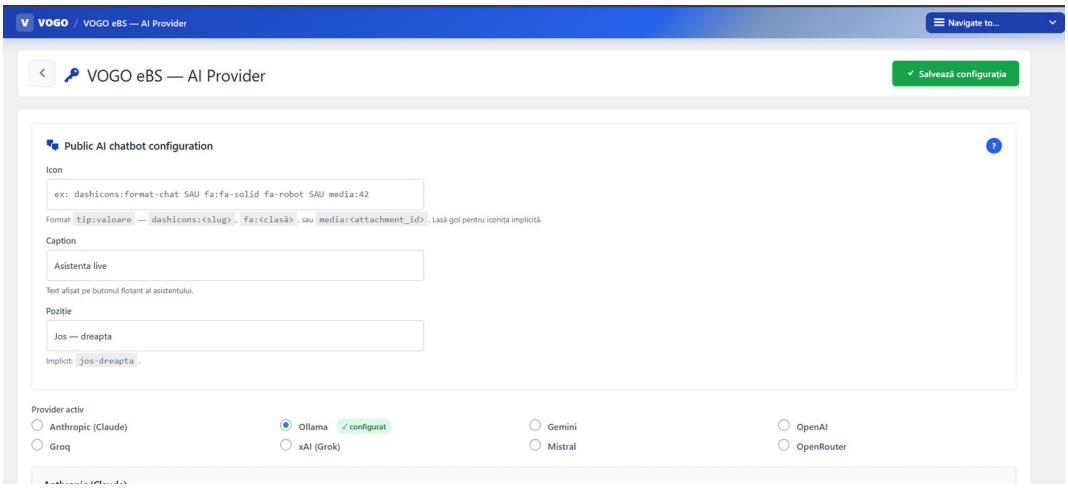

Abb. 1 — Konfiguration des aktiven KI-Anbieters über die VOGO eBS AI Provider Oberfläche

Architekturkomponenten

- LangChain — LLM-Anbieter-Abstraktion, RAG/Retrievers, Output Parsers, Tool Use
- LangGraph — Orchestrierung zustandsbehafteter agentischer Workflows (State Machines), Schleifen, Bedingungen, Retry-Logik, Human-in-the-Loop
- LangSmith — vollständige Observability: Call Tracing, Testing/Evals, Produktions-Monitoring, Prompt-Versionierung
- RAG (Retrieval-Augmented Generation) — eigene Wissensdatenbanken ohne Re-Training; Abruf der Top-K-Chunks → präzise kontextuelle Antwort
- MCP (Model Context Protocol) — offener Standard zur Verbindung von Agenten mit Tools und externen Quellen (ERP, CRM, DMS, Kalender); Embeddings + Vektordatenbank — semantische Indexierung von Dokumenten (Pinecone, ChromaDB, pgvector) für sofortige kontextuelle Suche

3. Funktionale Fähigkeiten

KI Live-Assistent — In Formulare Integrierter Assistent

Die Plattform umfasst einen nativ in das Formular-Modul integrierten KI-Assistenten, der Benutzer in Echtzeit bei der Erstellung und Ausfüllung digitaler Formulare führt.

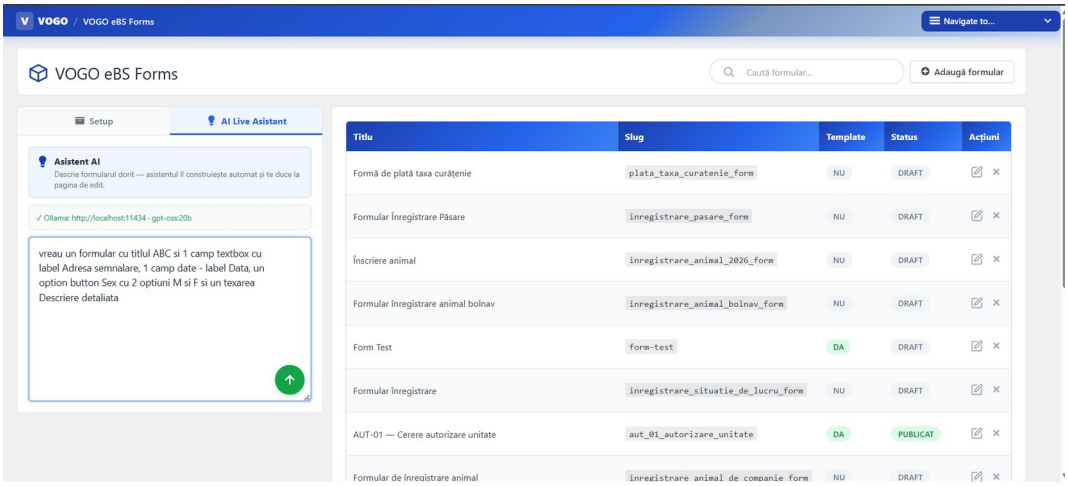

Abb. 2 — KI Live-Assistent aktiv im VOGO eBS Forms-Modul

KI Live-Support — Intelligenter Support in Diensten

Das Dienste-Modul profitiert von einer Live-KI-Support-Schicht, die Betreiber und Bürger bei der Navigation in digitalen Service-Workflows unterstützt und Lösungszeiten sowie das Volumen manueller Anfragen reduziert.

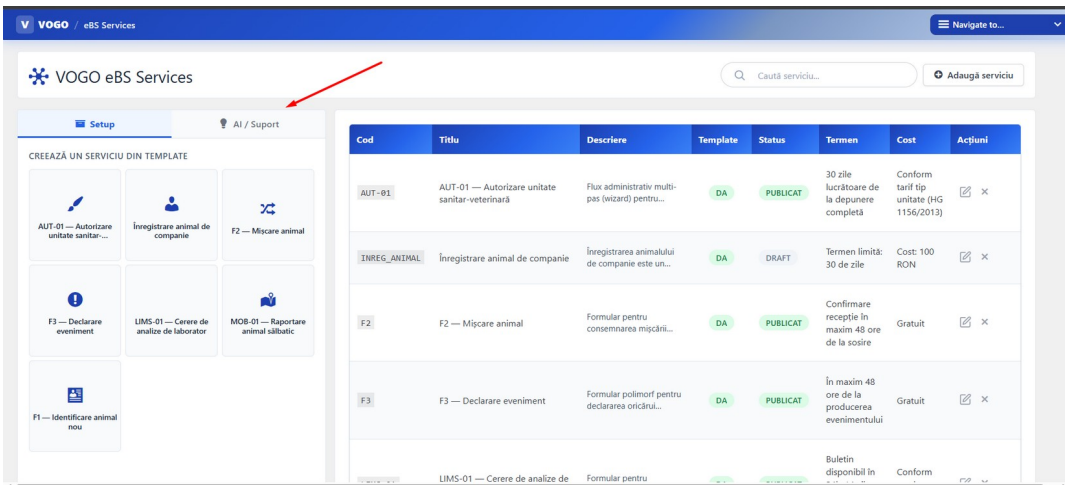

Abb. 3 — KI Live-Support aktiv im VOGO eBS Services-Modul

Chatbot & Kommunikationsautomatisierung

- Konversationeller KI-basierter Chatbot mit Text- und Spracherkennung, menschenähnliches Verhalten
- Vollständiger On-Premise-Betrieb ohne Abhängigkeit von externen Cloud-Diensten
- Proprietäres NLU-Modul (Natural Language Understanding) zur Erkennung von Absichten und Entitäten
- Q&A-Modul von Administratoren ohne Code konfigurierbar
- Personalisierte Begrüßungsnachricht wird automatisch beim Öffnen der Chat-Sitzung gesendet
- Webanwendung für den Chat-Kanal, in jede HTML-Seite einbettbar
- Human-in-the-Loop (HITL)-Workflows: Der KI-Agent schlägt vor, der menschliche Operator genehmigt oder greift ein

Unterstützte Kommunikationskanäle

- Web-Chat — einbettbares Widget in jeder HTML-Seite
- E-Mail — automatische Verarbeitung und Antwort
- SMS — Benachrichtigungen und bidirektionale Interaktionen
- Social Media — administratorkonfigurierbare Integration (Facebook, WhatsApp und andere)
- Erweiterbare Multi-Kanal-Architektur für zusätzliche Kanäle

Benutzerverwaltung & SSO

- Differenzierte Interaktionsmodi für interne und externe Benutzer
- Anonymes externes Konto (Browser-Cookie-basiert) oder authentifiziert mit Benutzername/Passwort; Synchronisierung mit dem bestehenden SSO-System — interne Benutzer verwalten keine mehrfachen Konten; Erstellung neuer Chatbot-Projekte für den Hauptadministrator der Plattform verfügbar
- Der Projektadministrator konfiguriert Integrationen mit externen Systemen und sozialen Kanälen

4. Erweiterte KI-Fähigkeiten

Dokumentenverarbeitung & Extraktion

- Automatische Ingestion von PDF, Word, HTML, CSV, Notion, Confluence über LangChain Document Loaders
- OCR für gescannte Dokumente (Tesseract, AWS Textract, Google Document AI)
- Intelligentes Chunking: RecursiveCharacterTextSplitter, TokenTextSplitter, SemanticChunker
- Metadaten-Extraktion für Filterung und kontextuellen Abruf in Enterprise-Systemen mit Tausenden von Dokumenten

Speicher & Cache

- Conversation Memory (Buffer, Window, Summary) — Kontext-Persistenz in Chatbots
- Semantisches Gedächtnis — relevante Erinnerungen aus früheren Sitzungen zur Personalisierung
- Prompt Caching — 90% Rabatt (Anthropic) / 75% (Google) auf wiederholte System-Prompts
- LangGraph Checkpointing — persistentes Gedächtnis zwischen Sitzungen ohne benutzerdefinierten Code

Kosten- & Leistungsoptimierung

- Router-Muster — einfache Anfragen → leichte Modelle; komplexe Anfragen → Premium-Modelle (60-80% Kosteneinsparung)
- Async & Batching — parallele Verarbeitung, Anthropic Batch API (50% Rabatt gegenüber Echtzeit)
- Quantisierung (GGUF, GPTQ, AWQ) — große Modelle auf Standard-Hardware ausführen
- Structured Output / Pydantic — automatisch validierte JSON-Antworten ohne fragiles Parsing

5. Sicherheit & Compliance

- PII Detection & Redaction (Presidio, AWS Comprehend) — automatische Maskierung persönlicher Daten vor Übertragung an Cloud-LLM
- Guardrails (NeMo Guardrails, LlamaGuard) — konfigurierbare Verhaltensregeln: verbotene Themen, Erkennung toxischer Inhalte
- Prompt-Injection-Schutz — Eingabe-Bereinigung, Ausgabe-Validierung, robuster System-Prompt
- Rate Limiting & Auth — Benutzerauthentifizierung, RPM/TPM-Begrenzung, vollständiges Audit-Log
- Ausgabe-Validierungs-Pipelines — Konfidenz-Scoring, Halluzinations-Erkennung, Quellenverifizierung
- DSGVO-Konformität — Daten lokal verarbeitet (Ollama on-premise), ohne Daten-Exfiltration

6. Deployment-Optionen

Cloud (SaaS)	Managed Deployment auf Cloud-Infrastruktur; automatische Skalierung, kein Server-Management
On-Premise	Installation auf eigener Infrastruktur; Daten verlassen die Organisation nicht; ideal für strenge DSGVO
Hybrid	Lokales LLM (Ollama) + Cloud-Orchestrierung; Kosten-Compliance-Balance
SageMaker (AWS)	Deployment feinabgestimmter Modelle auf AWS; Autoscaling, Spot Instances, optimierte Kosten
Docker / Lambda	Serverlose Architektur; Lambda + SageMaker-Endpoint; keine Server zu verwalten

7. Unterstützte Integrationen

- VOGO Portal — KI-Assistent direkt im Portal der Organisation verfügbar
- VOGO Forms — KI Live-Assistent zur Erstellung und Ausfüllung von Formularen
- VOGO Services — KI Live-Support für digitale Service-Workflows
- DMS (Dokumentenmanagement-System) — Dokumentenzugang über Web-Services

- CRM, ERP, Ticketing — Integration über MCP oder REST-API
- SAP, Oracle eBS, Microsoft D365 — Enterprise-Konnektoren für Betriebsdaten
- E-Mail, SMS, Social Media — Multi-Kanal-Kommunikationskanäle

8. Weitere Funktionalitäten

- Das Modul enthält ein Bürger-Kommunikations-Automatisierungssystem — ein verfügbarer KI-Assistent, der sowohl direkt im Portal der Organisation als auch in der entsprechenden mobilen Anwendung (Android oder iOS) angezeigt werden kann, die auf Google Play und dem Apple Store veröffentlicht werden kann
- Chatbot, basierend auf künstlicher Intelligenz, programmiert, um Gespräche so nah wie möglich am menschlichen Verhalten zu führen, durch Text- und Spracherkennung.
- Kann vollständig on-premise betrieben werden, ohne Abhängigkeit von externen Diensten.
- Bietet die Fähigkeit zur Integration mit mehreren Kommunikationskanälen.
- Für den Web-Chat-Kanal wird eine Webanwendung bereitgestellt, die in jede HTML-Seite eingebettet werden kann.
- Ermöglicht dem Projektadministrator, Integrationen mit externen Systemen und Social-Media-Kanälen zu aktivieren und zu konfigurieren.
- Ermöglicht einfache Konfiguration und Administration des Assistenten-Inhalts ohne Code-Schreiben.
- Bietet vollständige Werkzeuge für Chatbot-Tests: Testen des Konversationsflusses, der Absichten und erkannten Entitäten.
- Ermöglicht die Interaktion mit der DMS-Lösung über ihre Web-Services.
- Ermöglicht Integration mit E-Mail- und SMS-Lösungen, mit Erweiterungsmöglichkeit für zusätzliche Kanäle.
- Ermöglicht unterschiedliche Interaktionsmodi für interne und externe Benutzer.
- Ermöglicht Human-in-the-Loop (HITL)-Workflows.
- Ermöglicht Synchronisierung mit dem SSO-System und eliminiert die Notwendigkeit der Mehrfach-Kontoverwaltung für interne Benutzer.
- Enthält ein dediziertes LLM-Modul.
- Enthält ein proprietäres Natural Language Understanding (NLU)-Modul, basierend auf künstlicher Intelligenz, das Absichten und Entitäten erkennt, die von Benutzern frei ausgedrückt werden.
- Ermöglicht die Konfiguration einer Begrüßungsnachricht, die beim Öffnen des Chat-Fensters automatisch gesendet wird. Ermöglicht die Konfiguration vordefinierter, hierarchischer Fragen und Antworten nach einem Konversations-Skript-Modell.
- Ermöglicht Integration mit etablierten LLM-Lösungen wie, aber nicht beschränkt auf: LLaMA, OpenAI, Gemini.
- Ermöglicht jedem externen Benutzer, ein anonymes Konto (basierend auf einem Browser-Cookie) oder ein mit eigenem Benutzernamen und Passwort authentifiziertes Konto zu erstellen.