# 1. General Overview

VOGO eBS AI Agentic Solutions is the artificial intelligence module natively integrated into the VOGO eBS platform, designed to automate communication, assist users in real time, and orchestrate complex workflows in enterprise environments. The solution combines state-of-the-art LLMs with agentic AI architectures — RAG, LangChain, LangGraph and MCP — providing organizations with advanced AI capabilities without dependency on third-party platforms.

# 2. Technical Architecture & LLM Stack

## Supported LLM Providers

The platform supports configuration and real-time switching between multiple LLM providers, without code changes:

- Anthropic (Claude) — advanced reasoning, extended context, enterprise security
- OpenAI (GPT-4o, GPT-4 Turbo) — reference model for conversational applications
- Google Gemini — native integration with the Google Workspace ecosystem
- Ollama — run open-source models (LLaMA, Mistral, Qwen) locally, on-premise, with no per-token costs
- Groq — ultra-fast inference for latency-critical scenarios
- xAI (Grok) — alternative model for specific use cases
- Mistral — European model, relevant for strict GDPR compliance
- OpenRouter — unified gateway for access to dozens of models through a single API

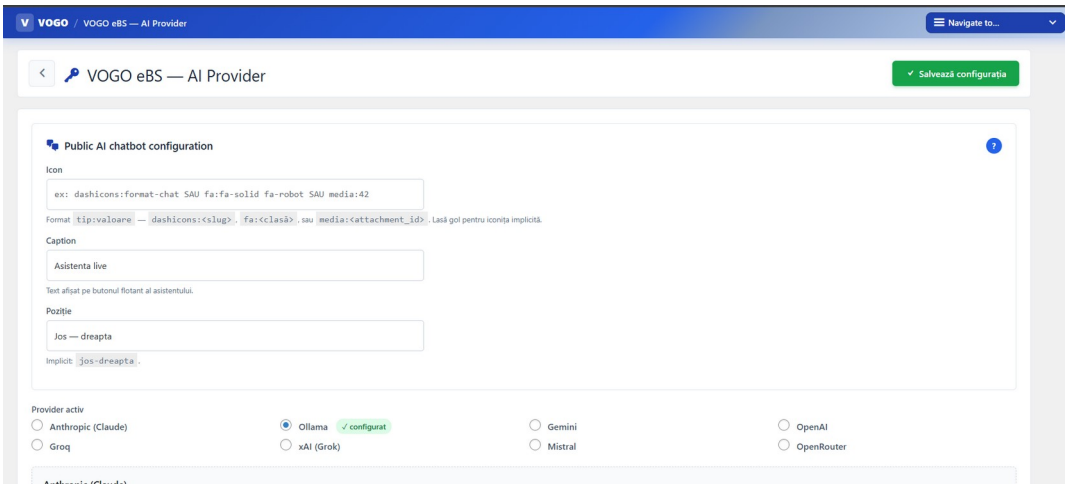


*Fig. 1 — Active AI provider configuration from the VOGO eBS AI Provider interface*

## Architectural Components

- LangChain — LLM provider abstraction, RAG/Retrievers, Output Parsers, Tool Use
- LangGraph — orchestration of stateful agentic workflows (state machines), loops, conditions, retry logic, Human-in-the-Loop
- LangSmith — complete observability: call tracing, testing/evals, production monitoring, prompt versioning
- RAG (Retrieval-Augmented Generation) — proprietary knowledge bases without re-training; retrieval of top-K chunks → precise contextual response
- MCP (Model Context Protocol) — open standard for connecting agents to tools and external sources (ERP, CRM, DMS, calendars); Embeddings + Vector Database — semantic indexing of documents (Pinecone, ChromaDB, pgvector) for instant contextual search

# 3. Functional Capabilities

## AI Live Assistant — Assistant Integrated in Forms

The platform includes an AI assistant natively integrated into the Forms module that guides users in creating and completing digital forms in real time.

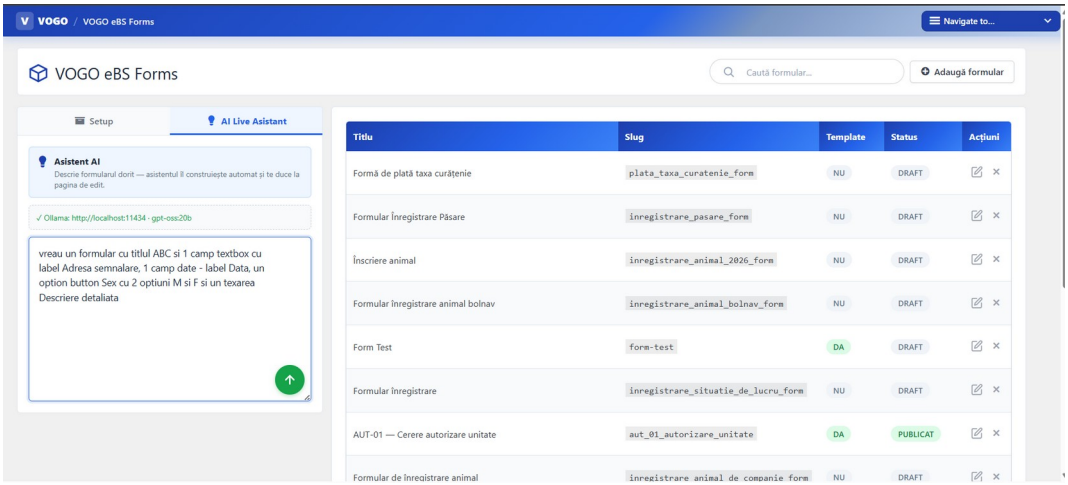


*Fig. 2 — AI Live Assistant active in the VOGO eBS Forms module*

## AI Live Support — Intelligent Support in Services

The services module benefits from a live AI support layer that assists operators and citizens in navigating digital service workflows, reducing resolution times and the volume of manual requests.

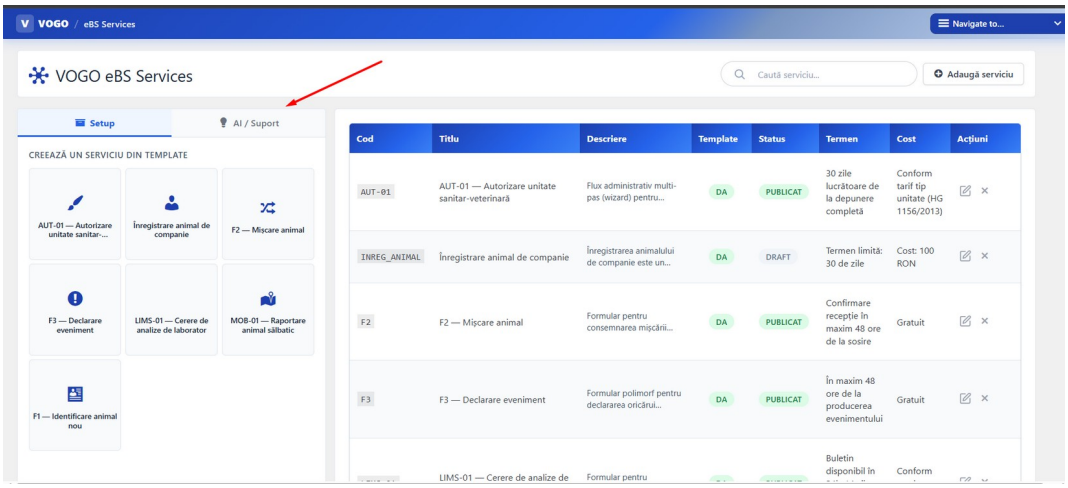


*Fig. 3 — AI Live Support active in the VOGO eBS Services module*

## Chatbot & Communication Automation

- Conversational AI-based chatbot with text and voice recognition, human-like behavior
- Full on-premise operation, with no dependency on external cloud services
- Proprietary NLU (Natural Language Understanding) module for intent and entity recognition
- Q&A module configurable by administrators without writing code
- Personalized welcome message automatically sent when opening the chat session
- Web application for the chat channel, embeddable in any HTML page
- Human-in-the-Loop (HITL) workflows: the AI agent proposes, the human operator approves or intervenes

## Supported Communication Channels

- Web chat — embeddable widget in any HTML page
- E-mail — automatic processing and response
- SMS — notifications and two-way interactions
- Social media — administrator-configurable integration (Facebook, WhatsApp and others)
- Extensible multi-channel architecture for additional channels

## User Management & SSO

- Differentiated interaction modes for internal and external users

- Anonymous external account (browser cookie-based) or authenticated with username/password; Synchronization with the existing SSO system — internal users do not manage multiple accounts; New chatbot project creation available to the platform's main administrator
- The project administrator configures integrations with external systems and social channels

## 4. Advanced AI Capabilities

### Document Processing & Extraction

- Automatic ingestion of PDF, Word, HTML, CSV, Notion, Confluence via LangChain Document Loaders
- OCR for scanned documents (Tesseract, AWS Textract, Google Document AI)
- Intelligent chunking: RecursiveCharacterTextSplitter, TokenTextSplitter, SemanticChunker
- Metadata extraction for filtering and contextual retrieval in enterprise systems with thousands of documents

### Memory & Cache

- Conversation Memory (Buffer, Window, Summary) — context persistence in chatbots
- Semantic Memory — relevant memories from previous sessions for personalization
- Prompt Caching — 90% discount (Anthropic) / 75% (Google) on repeated system prompts
- LangGraph Checkpointing — persistent memory between sessions without custom code

### Cost & Performance Optimization

- Router patterns — simple queries → lightweight models; complex queries → premium models (60-80% cost reduction)
- Async & Batching — parallel processing, Anthropic Batch API (50% discount vs real-time)
- Quantization (GGUF, GPTQ, AWQ) — running large models on standard hardware
- Structured Output / Pydantic — automatically validated JSON responses, without fragile parsing

## 5. Security & Compliance

- PII Detection & Redaction (Presidio, AWS Comprehend) — automatic masking of personal data before sending to cloud LLM
- Guardrails (NeMo Guardrails, LlamaGuard) — configurable behavior rules: prohibited topics, toxic content detection
- Prompt Injection Protection — input sanitization, output validation, robust system prompt
- Rate Limiting & Auth — per-user authentication, RPM/TPM limiting, complete audit log
- Output validation pipelines — confidence scoring, hallucination detection, source verification
- GDPR Compliance — data processed locally (Ollama on-premise), without data exfiltration

## 6. Deployment Options

| | |
|---|---|
| **Cloud (SaaS)** | Managed deploy on cloud infrastructure; automatic scaling, zero server management |
| **On-Premise** | Installation on own infrastructure; data does not leave the organization; ideal for strict GDPR |
| **Hybrid** | Local LLM (Ollama) + cloud orchestration; cost-compliance balance |
| **SageMaker (AWS)** | Deployment of fine-tuned models on AWS; autoscaling, Spot Instances, optimized cost |
| **Docker / Lambda** | Serverless architecture; Lambda + SageMaker endpoint; no servers to manage |

## 7. Supported Integrations

- VOGO Portal — AI assistant available directly in the organization's portal
- VOGO Forms — AI Live Assistant for creating and completing forms
- VOGO Services — AI Live Support for digital service workflows
- DMS (Document Management System) — document access via web services
- CRM, ERP, Ticketing — integration via MCP or REST API
- SAP, Oracle eBS, Microsoft D365 — enterprise connectors for operational data

- E-mail, SMS, Social Media — multi-channel communication channels

## 8. Additional Features

- The module includes a citizen communication automation system — an AI assistant available that can be displayed both directly in the organization's portal and in the corresponding mobile application (Android or iOS) which can be published on Google Play and the Apple Store
- Chatbot, based on artificial intelligence, programmed to conduct conversations as close as possible to human behavior, through text and voice recognition.
- Can operate entirely on-premise, with no dependency on external services.
- Provides the capability to integrate with multiple communication channels.
- For the web chat channel, provides a web application embeddable in any HTML page.
- Allows the project administrator to activate and configure integrations with external systems and social media channels.
- Allows easy configuration and administration of the assistant's content, without writing code.
- Provides complete tools for chatbot testing: conversational flow testing, intent and entity recognition testing.
- Allows interaction with the DMS solution through its web services.
- Allows integration with e-mail and SMS solutions, with the possibility of extension for additional channels.
- Allows different interaction modes for internal and external users.
- Allows Human-in-the-Loop (HITL) workflows.
- Allows synchronization with the SSO system, eliminating the need for multiple account management for internal users.
- Includes a dedicated LLM module.
- Contains a proprietary Natural Language Understanding (NLU) module, based on artificial intelligence, which recognizes intents and entities freely expressed by users.
- Allows configuration of a welcome message automatically sent when opening the chat window. Allows configuration of predefined, hierarchical questions and answers, on a conversational script model.
- Allows integration with established LLM solutions such as but not limited to: LLaMA, OpenAI, Gemini.
- Allows each external user to create an anonymous account (based on a browser cookie) or authenticated with their own username and password.