

1. Visão Geral

O VOGO eBS AI Agentic Solutions é o módulo de inteligência artificial integrado nativamente na plataforma VOGO eBS, concebido para automatizar a comunicação, apoiar os utilizadores em tempo real e orquestrar fluxos de trabalho complexos em ambientes empresariais. A solução combina LLM de última geração com arquiteturas de IA agêntica — RAG, LangChain, LangGraph e MCP — oferecendo às organizações capacidades avançadas de IA sem dependência de plataformas de terceiros.

2. Arquitetura Técnica e Stack LLM

Fornecedores de LLM Suportados

A plataforma suporta a configuração e a alternância em tempo real entre vários fornecedores de LLM, sem alterações de código:

- ▶ Anthropic (Claude) — raciocínio avançado, contexto alargado, segurança empresarial
- ▶ OpenAI (GPT-4o, GPT-4 Turbo) — modelo de referência para aplicações conversacionais
- ▶ Google Gemini — integração nativa com o ecossistema Google Workspace
- ▶ Ollama — execução de modelos open source (LLaMA, Mistral, Qwen) localmente, on-premise, sem custos por token
- ▶ Groq — inferência ultrarrápida para cenários críticos em latência
- ▶ xAI (Grok) — modelo alternativo para casos de utilização específicos
- ▶ Mistral — modelo europeu, relevante para uma conformidade estrita com o RGPD
- ▶ OpenRouter — gateway unificado de acesso a dezenas de modelos através de uma única API

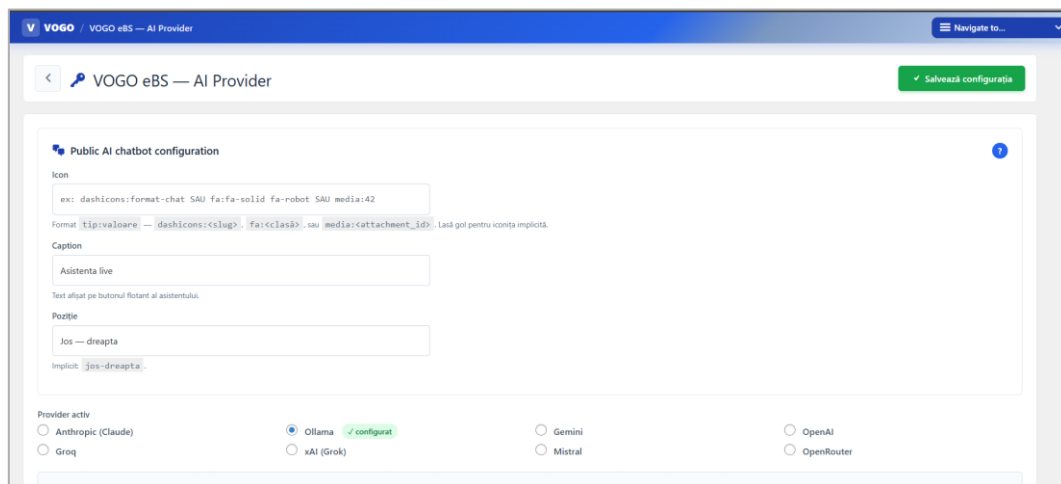


Fig. 1 — Configuração do fornecedor de IA ativo na interface VOGO eBS AI Provider

Componentes Arquiteturais

- ▶ LangChain — abstração de fornecedores de LLM, RAG/Retrievers, Output Parsers, Tool Use
- ▶ LangGraph — orquestração de fluxos agênticos com estado (máquinas de estados), ciclos, condições, lógica de nova tentativa, Human-in-the-Loop
- ▶ LangSmith — observabilidade completa: rastreio de chamadas, testes/avaliações, monitorização em produção, versionamento de prompts
- ▶ RAG (Retrieval-Augmented Generation) — bases de conhecimento próprias sem retreino; recuperação dos top-K fragmentos → resposta contextual precisa
- ▶ MCP (Model Context Protocol) — norma aberta para ligar agentes a ferramentas e fontes externas (ERP, CRM, DMS, calendários); Embeddings + Base de Dados Vetorial — indexação semântica de documentos (Pinecone, ChromaDB, pgvector) para pesquisa contextual instantânea

3. Capacidades Funcionais

AI Live Assistant — Assistente Integrado nos Formulários

A plataforma inclui um assistente de IA integrado nativamente no módulo Forms, que orienta os utilizadores na criação e no preenchimento de formulários digitais em tempo real.

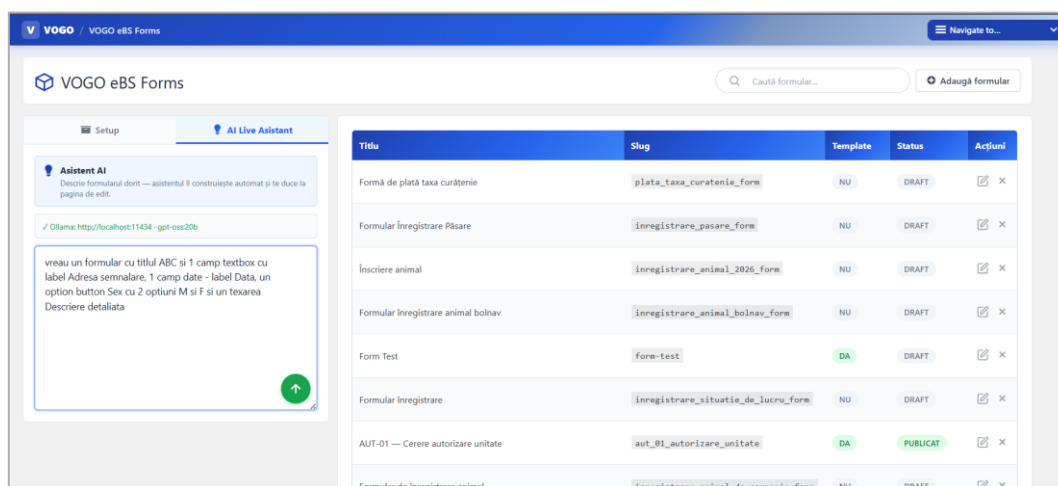


Fig. 2 — AI Live Assistant ativo no módulo VOGO eBS Forms

AI Live Support — Suporte Inteligente nos Serviços

O módulo de serviços beneficia de uma camada de suporte de IA em tempo real que apoia operadores e cidadãos nos fluxos de serviços digitais, reduzindo os tempos de resolução e o volume de pedidos manuais.

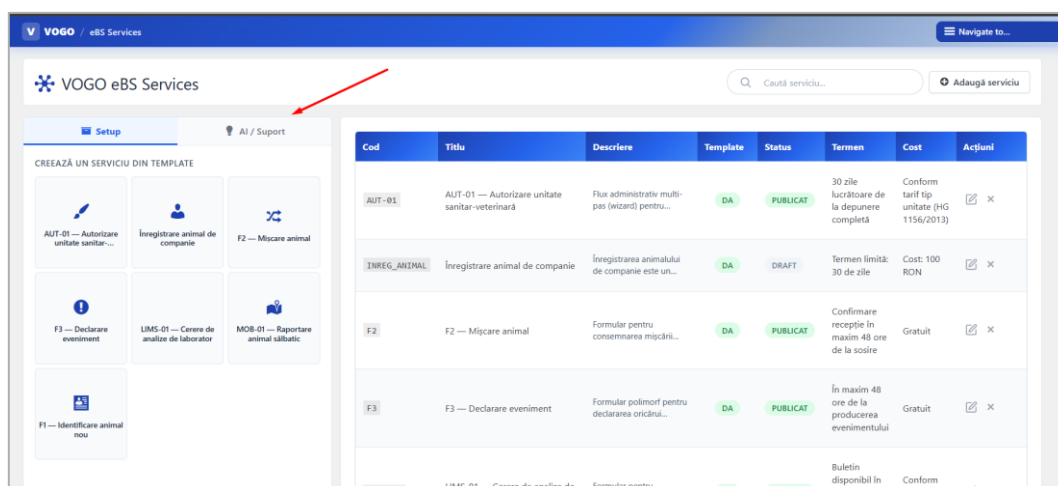


Fig. 3 — AI Live Support ativo no módulo VOGO eBS Services

Chatbot e Automação da Comunicação

- ▶ Chatbot baseado em IA conversacional, com reconhecimento de texto e voz e comportamento semelhante ao humano
- ▶ Funcionamento totalmente on-premise, sem dependência de serviços cloud externos
- ▶ Módulo NLU (Natural Language Understanding) proprietário para reconhecimento de intenções e entidades
- ▶ Módulo de perguntas e respostas configurável pelos administradores sem escrever código
- ▶ Mensagem de boas-vindas personalizada enviada automaticamente ao abrir a sessão de chat
- ▶ Aplicação web para o canal de chat, incorporável em qualquer página HTML
- ▶ Fluxos Human-in-the-Loop (HITL): o agente de IA propõe, o operador humano aprova ou intervém

Canais de Comunicação Suportados

- ▶ Web chat — widget incorporável em qualquer página HTML
- ▶ E-mail — processamento e resposta automáticos
- ▶ SMS — notificações e interações bidirecionais
- ▶ Redes sociais — integração configurável pelo administrador (Facebook, WhatsApp e outras)
- ▶ Arquitetura multicanal extensível para canais adicionais

Gestão de Utilizadores e SSO

- ▶ Modos de interação diferenciados para utilizadores internos e externos
- ▶ Conta externa anónima (baseada em cookie do browser) ou autenticada com nome de utilizador/palavra-passe; Sincronização com o sistema SSO existente — os utilizadores internos não gerem várias contas; criação de novos projetos de chatbot disponível para o administrador principal da plataforma
- ▶ O administrador do projeto configura as integrações com sistemas externos e canais sociais

4. Capacidades Avançadas de IA

Processamento e Extração de Documentos

- ▶ Ingestão automática de PDF, Word, HTML, CSV, Notion, Confluence através de LangChain Document Loaders
- ▶ OCR para documentos digitalizados (Tesseract, AWS Textract, Google Document AI)
- ▶ Fragmentação inteligente: RecursiveCharacterTextSplitter, TokenTextSplitter, SemanticChunker
- ▶ Extração de metadados para filtragem e recuperação contextual em sistemas empresariais com milhares de documentos

Memória e Cache

- ▶ Conversation Memory (Buffer, Window, Summary) — persistência do contexto nos chatbots
- ▶ Semantic Memory — memórias relevantes de sessões anteriores para personalização
- ▶ Prompt Caching — desconto de 90% (Anthropic) / 75% (Google) em system prompts repetidos
- ▶ LangGraph Checkpointing — memória persistente entre sessões sem código personalizado

Otimização de Custos e Desempenho

- ▶ Padrões de encaminhamento — consultas simples → modelos ligeiros; consultas complexas → modelos premium (redução de custos de 60-80%)
- ▶ Async e Batching — processamento paralelo, Anthropic Batch API (desconto de 50% face ao tempo real)
- ▶ Quantização (GGUF, GPTQ, AWQ) — execução de modelos grandes em hardware standard
- ▶ Structured Output / Pydantic — respostas JSON validadas automaticamente, sem parsing frágil

5. Segurança e Conformidade

- ▶ Detecção e Ocultação de PII (Presidio, AWS Comprehend) — mascaramento automático de dados pessoais antes do envio para LLM na cloud
- ▶ Guardrails (NeMo Guardrails, LlamaGuard) — regras de comportamento configuráveis: temas proibidos, deteção de conteúdo tóxico
- ▶ Proteção contra Prompt Injection — sanitização de entradas, validação de saídas, system prompt robusto
- ▶ Rate Limiting e Autenticação — autenticação por utilizador, limitação RPM/TPM, registo de auditoria completo
- ▶ Pipelines de validação de saídas — pontuação de confiança, deteção de alucinações, verificação de fontes
- ▶ Conformidade com o RGPD — dados processados localmente (Ollama on-premise), sem exfiltração de dados

6. Opções de Implementação

Cloud (SaaS)	Implementação gerida em infraestrutura cloud; escalamento automático, zero gestão de servidores
On-Premise	Instalação em infraestrutura própria; os dados não saem da organização; ideal para RGPD estrito
Híbrido	LLM local (Ollama) + orquestração na cloud; equilíbrio entre custo e conformidade
SageMaker (AWS)	Implementação de modelos fine-tuned na AWS; escalamento automático, Spot Instances, custo otimizado
Docker / Lambda	Arquitetura serverless; Lambda + endpoint SageMaker; sem servidores a gerir

7. Integrações Suportadas

- ▶ VOGO Portal — assistente de IA disponível diretamente no portal da organização
- ▶ VOGO Forms — AI Live Assistant para criação e preenchimento de formulários
- ▶ VOGO Services — AI Live Support para fluxos de serviços digitais

- ▶ DMS (Document Management System) — acesso a documentos através de serviços web
- ▶ CRM, ERP, Ticketing — integração através de MCP ou REST API
- ▶ SAP, Oracle eBS, Microsoft D365 — conectores empresariais para dados operacionais
- ▶ E-mail, SMS, Redes Sociais — canais de comunicação multicanal

8. Funcionalidades Adicionais

- ▶ O módulo inclui um sistema de automação da comunicação com os cidadãos — um assistente de IA que pode ser apresentado tanto diretamente no portal da organização como na respetiva aplicação móvel (Android ou iOS), que pode ser publicada no Google Play e na Apple Store
 - ▶ Chatbot baseado em inteligência artificial, programado para conduzir conversas o mais próximas possível do comportamento humano, através de reconhecimento de texto e voz.
 - ▶ Pode funcionar inteiramente on-premise, sem dependência de serviços externos.
 - ▶ Oferece a capacidade de integração com múltiplos canais de comunicação.
 - ▶ Para o canal de web chat, disponibiliza uma aplicação web incorporável em qualquer página HTML.
 - ▶ Permite ao administrador do projeto ativar e configurar integrações com sistemas externos e canais de redes sociais.
 - ▶ Permite a configuração e administração simples do conteúdo do assistente, sem escrever código.
 - ▶ Disponibiliza ferramentas completas de teste do chatbot: teste dos fluxos conversacionais, teste do reconhecimento de intenções e entidades.
 - ▶ Permite a interação com a solução DMS através dos respetivos serviços web.
 - ▶ Permite a integração com soluções de e-mail e SMS, com possibilidade de extensão a canais adicionais.
 - ▶ Permite modos de interação diferentes para utilizadores internos e externos.
 - ▶ Permite fluxos de trabalho Human-in-the-Loop (HITL).
 - ▶ Permite a sincronização com o sistema SSO, eliminando a necessidade de gerir várias contas para os utilizadores internos.
 - ▶ Inclui um módulo LLM dedicado.
 - ▶ Contém um módulo proprietário de Natural Language Understanding (NLU), baseado em inteligência artificial, que reconhece intenções e entidades expressas livremente pelos utilizadores.
 - ▶ Permite configurar uma mensagem de boas-vindas enviada automaticamente ao abrir a janela de chat. Permite configurar perguntas e respostas predefinidas e hierárquicas, segundo um modelo de guião conversacional.
 - ▶ Permite a integração com soluções LLM consolidadas, tais como, entre outras: LLaMA, OpenAI, Gemini.
 - ▶ Permite a cada utilizador externo criar uma conta anónima (baseada num cookie do browser) ou autenticada com o seu próprio nome de utilizador e palavra-passe.