1. Presentación General

VOGO eBS AI Agentic Solutions es el módulo de inteligencia artificial integrado de forma nativa en la plataforma VOGO eBS, diseñado para automatizar la comunicación, asistir a los usuarios en tiempo real y orquestar flujos de trabajo complejos en entornos enterprise. La solución combina LLMs de última generación con arquitecturas de IA agéntica — RAG, LangChain, LangGraph y MCP — proporcionando a las organizaciones capacidades de IA avanzadas sin dependencia de plataformas de terceros.

2. Arquitectura Técnica & Stack LLM

Proveedores LLM Soportados

La plataforma soporta la configuración y conmutación en tiempo real entre múltiples proveedores LLM, sin cambios de código:

- Anthropic (Claude) — razonamiento avanzado, contexto extendido, seguridad enterprise
- OpenAI (GPT-4o, GPT-4 Turbo) — modelo de referencia para aplicaciones conversacionales
- Google Gemini — integración nativa con el ecosistema Google Workspace
- Ollama — ejecutar modelos open-source (LLaMA, Mistral, Qwen) localmente, on-premise, sin costos por token
- Groq — inferencia ultrarrápida para escenarios con latencia crítica
- xAI (Grok) — modelo alternativo para casos de uso específicos
- Mistral — modelo europeo, relevante para cumplimiento RGPD estricto
- OpenRouter — pasarela unificada para acceso a docenas de modelos a través de una única API

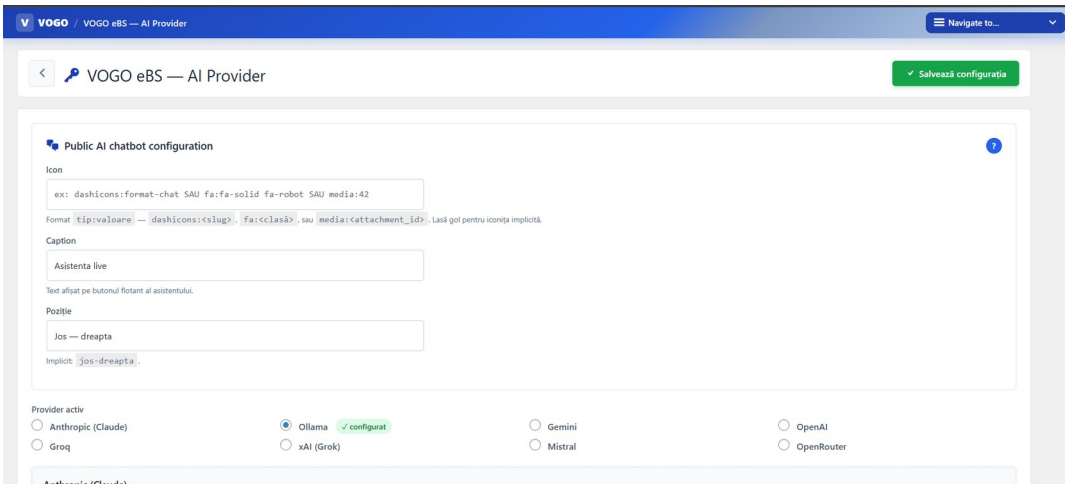

Fig. 1 — Configuración del proveedor IA activo desde la interfaz VOGO eBS AI Provider

Componentes Arquitectónicos

- LangChain — abstracción de proveedor LLM, RAG/Retrievers, Output Parsers, Tool Use
- LangGraph — orquestación de flujos agénticos con estado (máquinas de estado), bucles, condiciones, retry logic, Human-in-the-Loop
- LangSmith — observabilidad completa: trazado de llamadas, testing/evals, monitoreo de producción, versionado de prompts
- RAG (Retrieval-Augmented Generation) — bases de conocimiento propias sin re-entrenamiento; recuperación de top-K chunks → respuesta contextual precisa
- MCP (Model Context Protocol) — estándar abierto para conectar agentes a herramientas y fuentes externas (ERP, CRM, DMS, calendarios); Embeddings + Base de Datos Vectorial — indexación semántica de documentos (Pinecone, ChromaDB, pgvector) para búsqueda contextual instantánea

3. Capacidades Funcionales

AI Live Assistant — Asistente Integrado en Formularios

La plataforma incluye un asistente IA integrado de forma nativa en el módulo de Formularios que guía a los usuarios en la creación y cumplimentación de formularios digitales en tiempo real.

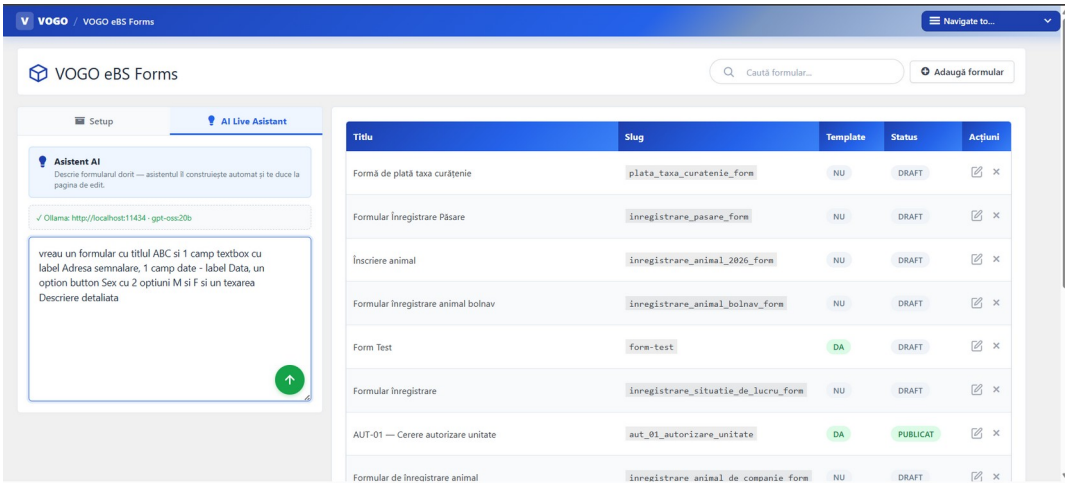

Fig. 2 — AI Live Assistant activo en el módulo VOGO eBS Forms

AI Live Support — Soporte Inteligente en Servicios

El módulo de servicios se beneficia de una capa de soporte IA en vivo que asiste a operadores y ciudadanos en la navegación de flujos de servicios digitales, reduciendo los tiempos de resolución y el volumen de solicitudes manuales.

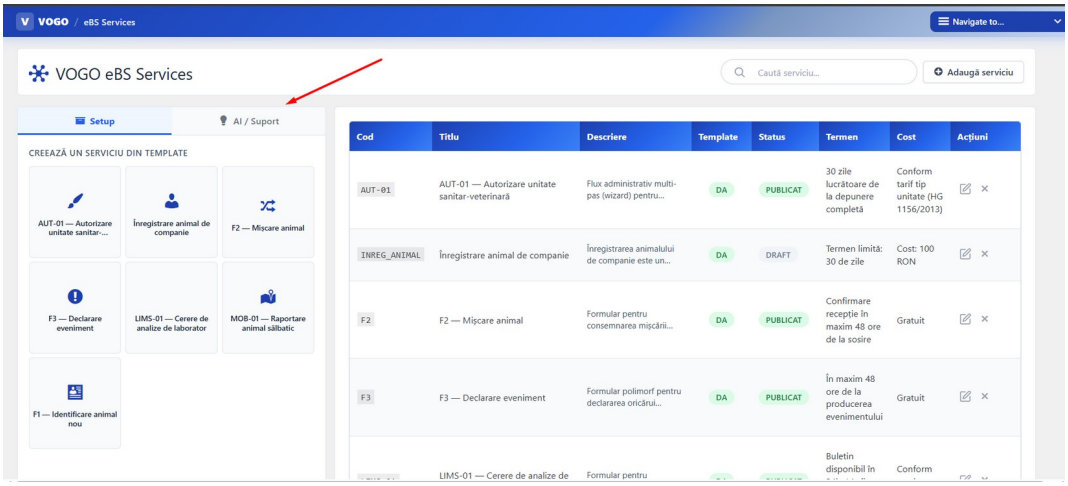

Fig. 3 — AI Live Support activo en el módulo VOGO eBS Services

Chatbot & Automatización de la Comunicación

- Chatbot basado en IA conversacional con reconocimiento de texto y voz, comportamiento similar al humano
- Funcionamiento íntegro on-premise, sin dependencia de servicios cloud externos
- Módulo NLU (Natural Language Understanding) propietario para reconocimiento de intenciones y entidades
- Módulo Q&A configurable por administradores sin escribir código
- Mensaje de bienvenida personalizado enviado automáticamente al abrir la sesión de chat
- Aplicación web para el canal chat, integrable en cualquier página HTML
- Flujos Human-in-the-Loop (HITL): el agente IA propone, el operador humano aprueba o interviene

Canales de Comunicación Soportados

- Chat web — widget integrable en cualquier página HTML
- E-mail — procesamiento y respuesta automáticos
- SMS — notificaciones e interacciones bidireccionales
- Redes sociales — integración configurable por administrador (Facebook, WhatsApp y otros)
- Arquitectura multi-canal extensible para canales adicionales

Gestión de Usuarios & SSO

- Modos de interacción diferenciados para usuarios internos y externos
- Cuenta externa anónima (basada en cookie de navegador) o autenticada con usuario/contraseña; Sincronización con el sistema SSO existente — los usuarios internos no gestionan múltiples cuentas; Creación de nuevo proyecto chatbot disponible para el administrador principal de la plataforma
- El administrador de proyecto configura las integraciones con sistemas externos y canales sociales

4. Capacidades IA Avanzadas

Procesamiento & Extracción de Documentos

- Ingestión automática de PDF, Word, HTML, CSV, Notion, Confluence mediante LangChain Document Loaders
- OCR para documentos escaneados (Tesseract, AWS Textract, Google Document AI)
- Fragmentación inteligente: RecursiveCharacterTextSplitter, TokenTextSplitter, SemanticChunker
- Extracción de metadatos para filtrado y recuperación contextual en sistemas enterprise con miles de documentos

Memoria & Caché

- Conversation Memory (Buffer, Window, Summary) — persistencia del contexto en chatbots
- Memoria Semántica — recuerdos relevantes de sesiones anteriores para personalización
- Prompt Caching — descuento del 90% (Anthropic) / 75% (Google) en system prompts repetidos
- LangGraph Checkpointing — memoria persistente entre sesiones sin código personalizado

Optimización de Costos & Rendimiento

- Patrones Router — consultas simples → modelos ligeros; consultas complejas → modelos premium (reducción de costos 60-80%)
- Async & Batching — procesamiento paralelo, Anthropic Batch API (50% descuento vs tiempo real)
- Cuantización (GGUF, GPTQ, AWQ) — ejecución de modelos grandes en hardware estándar
- Structured Output / Pydantic — respuestas JSON validadas automáticamente, sin parsing frágil

5. Seguridad & Cumplimiento

- PII Detection & Redaction (Presidio, AWS Comprehend) — enmascaramiento automático de datos personales antes del envío al LLM cloud
- Guardrails (NeMo Guardrails, LlamaGuard) — reglas de comportamiento configurables: temas prohibidos, detección de contenido tóxico
- Protección contra Prompt Injection — saneamiento de entrada, validación de salida, system prompt robusto
- Rate Limiting & Auth — autenticación por usuario, limitación RPM/TPM, registro de auditoría completo
- Pipelines de validación de salida — scoring de confianza, detección de alucinaciones, verificación de fuentes
- Cumplimiento RGPD — datos procesados localmente (Ollama on-premise), sin exfiltración de datos

6. Opciones de Despliegue

Cloud (SaaS)	Despliegue gestionado en infraestructura cloud; escalado automático, cero gestión de servidores
On-Premise	Instalación en infraestructura propia; los datos no salen de la organización; ideal para RGPD estricto
Hybrid	LLM local (Ollama) + orquestación cloud; equilibrio costo-cumplimiento
SageMaker (AWS)	Despliegue de modelos fine-tuned en AWS; autoscaling, Spot Instances, costo optimizado
Docker / Lambda	Arquitectura serverless; Lambda + endpoint SageMaker; sin servidores que gestionar

7. Integraciones Soportadas

- VOGO Portal — asistente IA disponible directamente en el portal de la organización
- VOGO Forms — AI Live Assistant para la creación y cumplimentación de formularios
- VOGO Services — AI Live Support para flujos de servicios digitales

- DMS (Sistema de Gestión Documental) — acceso a documentos mediante servicios web
- CRM, ERP, Ticketing — integración mediante MCP o API REST
- SAP, Oracle eBS, Microsoft D365 — conectores enterprise para datos operativos
- E-mail, SMS, Redes Sociales — canales de comunicación multi-canal

8. Funcionalidades Adicionales

- El módulo incluye un sistema de automatización de la comunicación con ciudadanos — un asistente IA disponible que puede mostrarse tanto directamente en el portal de la organización como en la aplicación móvil correspondiente (Android o iOS) que puede publicarse en Google Play y la App Store
- Chatbot, basado en inteligencia artificial, programado para mantener conversaciones lo más parecidas posible al comportamiento humano, mediante reconocimiento de texto y voz.
- Puede funcionar íntegramente on-premise, sin dependencia de servicios externos.
- Proporciona la capacidad de integración con múltiples canales de comunicación.
- Para el canal de chat web, proporciona una aplicación web integrable en cualquier página HTML.
- Permite al administrador de proyecto activar y configurar integraciones con sistemas externos y canales de redes sociales.
- Permite la configuración y administración fácil del contenido del asistente, sin escribir código.
- Proporciona herramientas completas para pruebas de chatbots: prueba del flujo conversacional, de intenciones y entidades reconocidas.
- Permite la interacción con la solución DMS a través de sus servicios web.
- Permite la integración con soluciones de e-mail y SMS, con posibilidad de extensión para canales adicionales.
- Permite diferentes modos de interacción para usuarios internos y externos.
- Permite flujos Human-in-the-Loop (HITL).
- Permite la sincronización con el sistema SSO, eliminando la necesidad de gestión múltiple de cuentas para usuarios internos.
- Incluye un módulo LLM dedicado.
- Contiene un módulo NLU (Natural Language Understanding) propietario, basado en inteligencia artificial, que reconoce intenciones y entidades expresadas libremente por los usuarios.
- Permite la configuración de un mensaje de bienvenida enviado automáticamente al abrir la ventana de chat. Permite la configuración de preguntas y respuestas predefinidas, jerárquicas, en un modelo de script conversacional.
- Permite la integración con soluciones LLM consolidadas como, pero sin limitarse a: LLaMA, OpenAI, Gemini.
- Permite a cada usuario externo crear una cuenta anónima (basada en una cookie de navegador) o autenticada con su propio nombre de usuario y contraseña.