1. Présentation Générale

VOGO eBS AI Agentic Solutions est le module d'intelligence artificielle intégré nativement dans la plateforme VOGO eBS, conçu pour automatiser la communication, assister les utilisateurs en temps réel et orchestrer des flux de travail complexes dans des environnements enterprise. La solution combine des LLMs de dernière génération avec des architectures IA agentiques — RAG, LangChain, LangGraph et MCP — offrant aux organisations des capacités IA avancées sans dépendance vis-à-vis de plateformes tierces.

2. Architecture Technique & Stack LLM

Fournisseurs LLM Supportés

La plateforme supporte la configuration et la commutation en temps réel entre plusieurs fournisseurs LLM, sans modification de code :

- Anthropic (Claude) — raisonnement avancé, contexte étendu, sécurité enterprise
- OpenAI (GPT-4o, GPT-4 Turbo) — modèle de référence pour les applications conversationnelles
- Google Gemini — intégration native avec l'écosystème Google Workspace
- Ollama — exécution de modèles open-source (LLaMA, Mistral, Qwen) localement, on-premise, sans coûts par token
- Groq — inférence ultra-rapide pour les scénarios à latence critique
- xAI (Grok) — modèle alternatif pour des cas d'usage spécifiques
- Mistral — modèle européen, pertinent pour la conformité RGPD stricte
- OpenRouter — passerelle unifiée pour accéder à des dizaines de modèles via une seule API

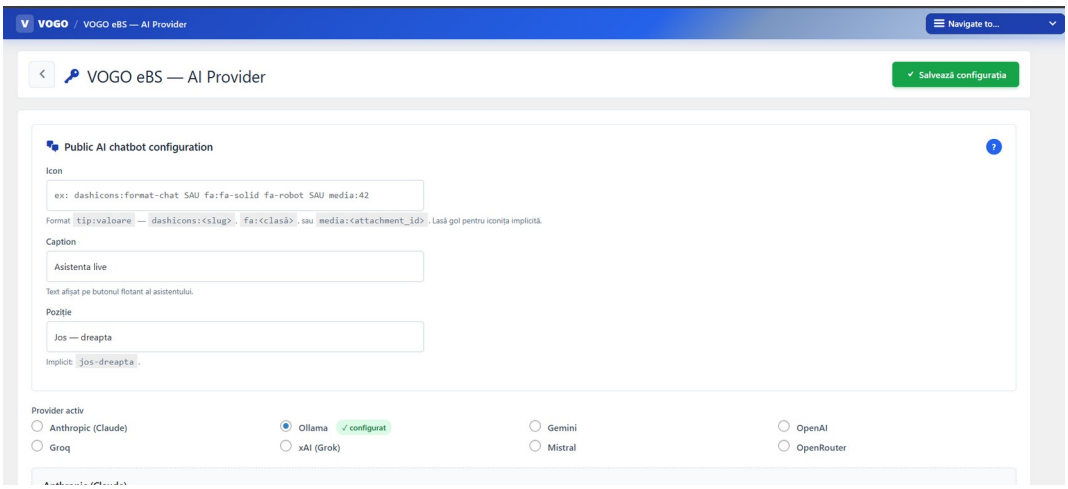

Fig. 1 — Configuration du fournisseur IA actif depuis l'interface VOGO eBS AI Provider

Composants Architecturaux

- LangChain — abstraction fournisseur LLM, RAG/Retrievers, Output Parsers, Tool Use
- LangGraph — orchestration de flux agentiques avec état (machines d'état), boucles, conditions, retry logic, Human-in-the-Loop
- LangSmith — observabilité complète : traçage des appels, testing/evals, monitoring production, versionnage des prompts
- RAG (Retrieval-Augmented Generation) — bases de connaissances propriétaires sans ré-entraînement ; récupération des top-K chunks → réponse contextuelle précise
- MCP (Model Context Protocol) — standard ouvert pour connecter les agents à des outils et sources externes (ERP, CRM, DMS, calendriers) ; Embeddings + Base Vectorielle — indexation sémantique des documents (Pinecone, ChromaDB, pgvector) pour une recherche contextuelle instantanée

3. Capacités Fonctionnelles

AI Live Assistant — Assistant Intégré dans les Formulaires

La plateforme inclut un assistant IA intégré nativement dans le module Formulaires qui guide les utilisateurs dans la création et la complétion de formulaires numériques en temps réel.

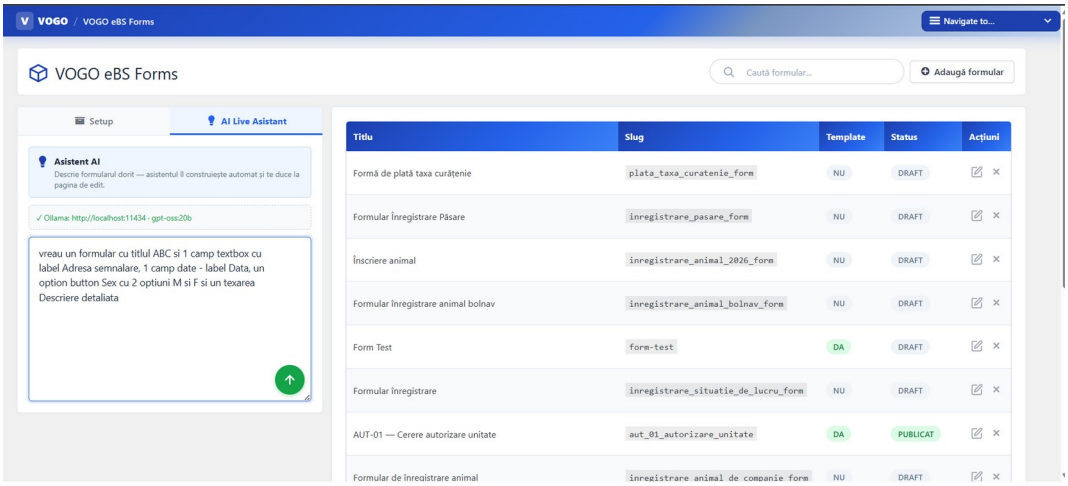

Fig. 2 — AI Live Assistant actif dans le module VOGO eBS Forms

AI Live Support — Support Intelligent dans les Services

Le module de services bénéficie d'une couche de support IA live qui assiste les opérateurs et les citoyens dans la navigation des flux de services numériques, réduisant les délais de résolution et le volume de demandes manuelles.

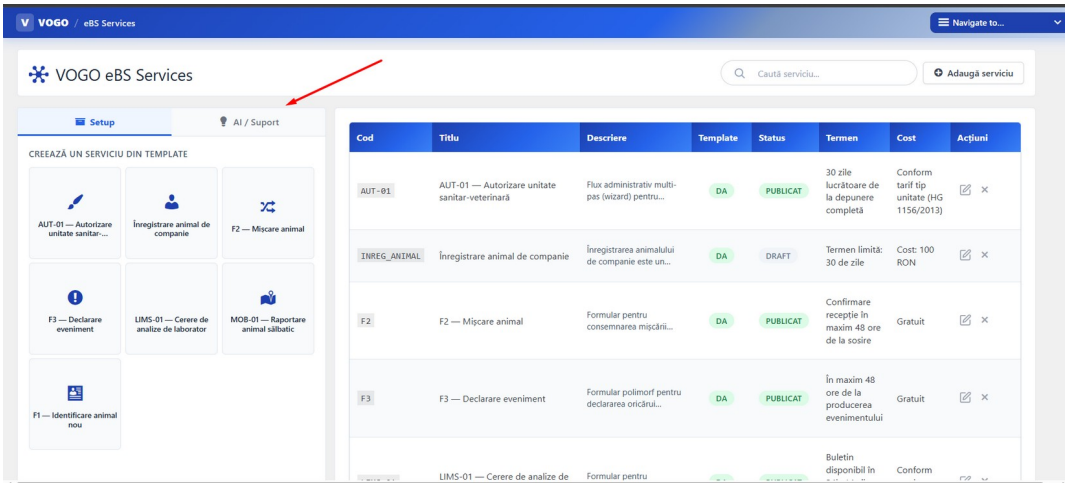

Fig. 3 — AI Live Support actif dans le module VOGO eBS Services

Chatbot & Automatisation de la Communication

- Chatbot basé sur l'IA conversationnelle avec reconnaissance vocale et textuelle, comportement similaire à l'humain
- Fonctionnement intégral on-premise, sans dépendance vis-à-vis des services cloud externes
- Module NLU (Natural Language Understanding) propriétaire pour la reconnaissance des intentions et entités
- Module Q&R configurable par les administrateurs sans écriture de code
- Message de bienvenue personnalisé envoyé automatiquement à l'ouverture de la session de chat
- Application web pour le canal chat, intégrable dans toute page HTML
- Flux Human-in-the-Loop (HITL) : l'agent IA propose, l'opérateur humain approuve ou intervient

Canaux de Communication Supportés

- Chat web — widget intégrable dans toute page HTML
- E-mail — traitement et réponse automatiques
- SMS — notifications et interactions bidirectionnelles
- Réseaux sociaux — intégration configurable par l'administrateur (Facebook, WhatsApp et autres)
- Architecture multi-canal extensible pour des canaux additionnels

Gestion des Utilisateurs & SSO

- Modes d'interaction différenciés pour les utilisateurs internes et externes
- Compte externe anonyme (basé sur cookie navigateur) ou authentifié avec identifiant/mot de passe ; Synchronisation avec le système SSO existant — les utilisateurs internes ne gèrent pas plusieurs comptes ; Création d'un nouveau projet chatbot disponible pour l'administrateur principal de la plateforme
- L'administrateur de projet configure les intégrations avec les systèmes externes et les canaux sociaux

4. Capacités IA Avancées

Traitement & Extraction de Documents

- Ingestion automatique de PDF, Word, HTML, CSV, Notion, Confluence via LangChain Document Loaders
- OCR pour les documents numérisés (Tesseract, AWS Textract, Google Document AI)
- Découpage intelligent : RecursiveCharacterTextSplitter, TokenTextSplitter, SemanticChunker
- Extraction de métadonnées pour le filtrage et la récupération contextuelle dans des systèmes enterprise avec des milliers de documents

Mémoire & Cache

- Conversation Memory (Buffer, Window, Summary) — persistance du contexte dans les chatbots
- Mémoire Sémantique — souvenirs pertinents des sessions précédentes pour la personnalisation
- Prompt Caching — remise de 90% (Anthropic) / 75% (Google) sur les system prompts répétés
- LangGraph Checkpointing — mémoire persistante entre les sessions sans code personnalisé

Optimisation des Coûts & des Performances

- Patterns Router — requêtes simples → modèles légers ; requêtes complexes → modèles premium (réduction des coûts de 60-80%)
- Async & Batching — traitement parallèle, Anthropic Batch API (remise de 50% vs temps réel)
- Quantification (GGUF, GPTQ, AWQ) — exécution de grands modèles sur du matériel standard
- Structured Output / Pydantic — réponses JSON validées automatiquement, sans parsing fragile

5. Sécurité & Conformité

- PII Detection & Redaction (Presidio, AWS Comprehend) — masquage automatique des données personnelles avant envoi au LLM cloud
- Guardrails (NeMo Guardrails, LlamaGuard) — règles de comportement configurables : sujets interdits, détection de contenu toxique
- Protection contre l'Injection de Prompt — assainissement des entrées, validation des sorties, system prompt robuste
- Rate Limiting & Auth — authentification par utilisateur, limitation RPM/TPM, journal d'audit complet
- Pipelines de validation des sorties — scoring de confiance, détection des hallucinations, vérification des sources
- Conformité RGPD — données traitées localement (Ollama on-premise), sans exfiltration de données

6. Options de Déploiement

Cloud (SaaS)	Déploiement managé sur infrastructure cloud ; mise à l'échelle automatique, zéro gestion de serveur
On-Premise	Installation sur l'infrastructure propre ; les données ne quittent pas l'organisation ; idéal pour le RGPD strict
Hybrid	LLM local (Ollama) + orchestration cloud ; équilibre coût-conformité
SageMaker (AWS)	Déploiement de modèles fine-tunés sur AWS ; autoscaling, Spot Instances, coût optimisé
Docker / Lambda	Architecture serverless ; Lambda + endpoint SageMaker ; aucun serveur à gérer

7. Intégrations Supportées

- VOGO Portal — assistant IA disponible directement dans le portail de l'organisation
- VOGO Forms — AI Live Assistant pour la création et la complétion de formulaires

- VOGO Services — AI Live Support pour les flux de services numériques
- DMS (Système de Gestion Documentaire) — accès aux documents via services web
- CRM, ERP, Ticketing — intégration via MCP ou API REST
- SAP, Oracle eBS, Microsoft D365 — connecteurs enterprise pour les données opérationnelles
- E-mail, SMS, Réseaux Sociaux — canaux de communication multi-canal

8. Fonctionnalités Supplémentaires

- Le module inclut un système d'automatisation de la communication avec les citoyens — un assistant IA disponible qui peut être affiché directement dans le portail de l'organisation et dans l'application mobile correspondante (Android ou iOS) publiable sur Google Play et l'Apple Store
- Chatbot basé sur l'intelligence artificielle, programmé pour mener des conversations aussi proches que possible du comportement humain, par reconnaissance vocale et textuelle.
- Peut fonctionner intégralement on-premise, sans dépendance vis-à-vis des services externes.
- Offre la capacité d'intégration avec plusieurs canaux de communication.
- Pour le canal de chat web, fournit une application web intégrable dans toute page HTML.
- Permet à l'administrateur de projet d'activer et de configurer des intégrations avec des systèmes externes et des canaux de médias sociaux.
- Permet une configuration et une administration faciles du contenu de l'assistant, sans écriture de code.
- Fournit des outils complets pour les tests des chatbots : test du flux conversationnel, des intentions et des entités reconnues.
- Permet l'interaction avec la solution DMS via ses services web.
- Permet l'intégration avec des solutions e-mail et SMS, avec possibilité d'extension pour des canaux additionnels.
- Permet des modes d'interaction différents pour les utilisateurs internes et externes.
- Permet des flux Human-in-the-Loop (HITL).
- Permet la synchronisation avec le système SSO, éliminant la nécessité de gérer plusieurs comptes pour les utilisateurs internes.
- Inclut un module LLM dédié.
- Contient un module NLU (Natural Language Understanding) propriétaire, basé sur l'intelligence artificielle, qui reconnaît les intentions et entités exprimées librement par les utilisateurs.
- Permet la configuration d'un message de bienvenue envoyé automatiquement à l'ouverture de la fenêtre de chat. Permet la configuration de questions et réponses prédéfinies, hiérarchiques, sur un modèle de script conversationnel.
- Permet l'intégration avec des solutions LLM reconnues telles que, mais sans s'y limiter : LLaMA, OpenAI, Gemini.
- Permet à chaque utilisateur externe de créer un compte anonyme (basé sur un cookie navigateur) ou authentifié avec son propre identifiant et mot de passe.