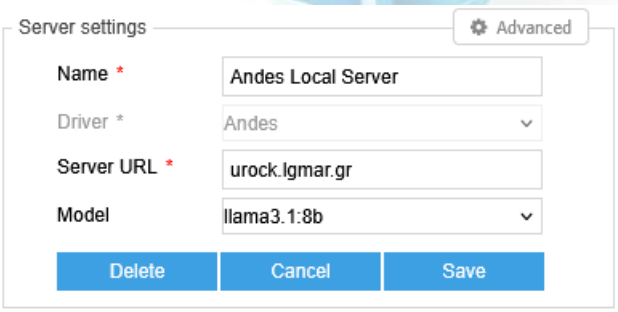
OMNi
AI assistant setup for OLLAMA local server

With Omni AI you are free to select from many LLM's (Large Language Models) including Openai Chat-GPT, Google Gemini, Meta Llama, gpt-oss, DeepSeek, Google Gemma, Mistral Nemo, MS Phi, Alibaba Qwen, Anthropic Claude et.al. and a variety of AI services.

Ollama is an open-source platform designed to run large language models locally. This guide will help you to setup a local server as your provider. Select **Tools | AI Providers | Servers +Add**. You will need the following settings:



- **Driver**: Andes (used for a local OLLAMA server)
- Server **URL**: The same with your Omni server i.e. omni.company.gr
- **Model**: Added automatically, you may change after save.

Open-source models are different in speed, reasoning/complexity. We suggest to start with a fast text model that fits your servers capacity. Below a brief description of some open-source models:

- **LLaMA** Meta's latest collection of multimodal models
- **DeepSeek** Harmonizes high computational efficiency with superior reasoning.
- **gpt-oss** OpenAI's open-weight models designed for powerful reasoning
- **Gemma** Lightweight, state-of-the-art open models built from the same research and technology used to create the Gemini models
- **Mistral** Family of small and large models by Mistral AI SAS.
- **MS-Phi** Small language model that offers high quality results at a small size
- **Qwen** Transformer-based large language models by Alibaba Cloud

For a full list of open-source models running with OLLAMA visit the following link: https://ollama.com/search

In case you require further assistance or guidance, please do not hesitate to contact our support team at support@lgmar.gr

LgMAR Software
Tel.: (+30) 2117708711
info@lgmar.gr, www.lgmar.gr