Callisto AI

Callisto AI integrates artificial intelligence into Callisto. The basic component added to Callisto for AI integration is the Callisto AI Gateway, a separate Ubuntu server that runs the speech services. Some speech services can run in the cloud or on premise. In addition to the gateway, a separate LLM (Large Language Model) server is used, which can run either in the cloud or on premise. When the LLM is run on premise, one of the recommended models can be used; in this case the model is included as part of the Callisto AI Gateway machine. Alternatively, the LLM can run in the cloud, or customers can use their own LLM.

All AI components are configured centrally. A new Callisto AI entry is available on the navigation bar.

Settings

On the navigation bar, click on Callisto AI > Settings.

The Settings page is divided into three areas: AI Gateway, LLM Server, and Speech Services. After changing any value click Save to apply the configuration, or Cancel to discard your changes.

AI Gateway

The Callisto AI Gateway is a separate server that handles the communication with the speech services. This server has to be configured so that Callisto can reach it and route calls to it. Configure the following parameters:

- IP Address — IP address of the AI Gateway.
- SIP Port — SIP signalling port used for call setup (default 8060).
- MRCP Port — port used for the MRCP speech sessions (default 1544).
- RTP Port Range — range of UDP ports used for the RTP media streams (for example 5000–6000).
- Routing Number — internal number used to route calls to the AI Gateway (for example 8800).

LLM Server

The LLM Server provides the language model that generates the conversational responses. Select the Provider from the drop-down list.

The following providers are available: On Premise, Google, and OpenAI.

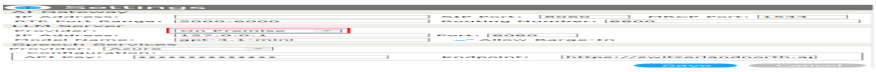
Cloud Provider (OpenAI or Google)

For a cloud provider, enter the credentials of your account:

- API Key — access key of your LLM provider account.
- Model Name — name of the language model to use (for example gpt-4.1-mini).

On Premise Model

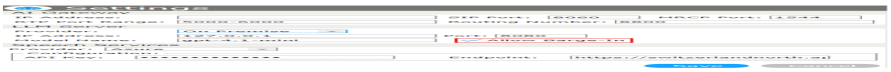
For an on-premise model, specify the address of your local LLM server instead:



- IP Address — IP address of the on-premise LLM server.
- Port — port on which the LLM server is reachable.
- Model Name — name of the language model served by the on-premise server.

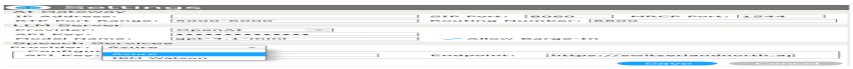
Allow Barge-In

Enable Allow Barge-In to let the caller interrupt the AI bot by speaking. When this option is disabled, the caller must wait until the bot has finished before speaking.



Speech Services

The Speech Services provider handles speech recognition (ASR) and speech synthesis (TTS). Select the Provider from the drop-down list.



The following providers are available: Azure and IBM Watson.

Azure

For Azure, a single set of credentials is used for both recognition and synthesis:

- API Key — access key of your Azure Speech resource.
- Endpoint — endpoint URL of your Azure Speech resource (region-specific).

IBM Watson

For IBM Watson, the speech-to-text (ASR) and text-to-speech (TTS) services are configured separately:

- ASR Configuration — API Key and URL of the IBM Watson Speech to Text service.
- TTS Configuration — API Key and URL of the IBM Watson Text to Speech service.