

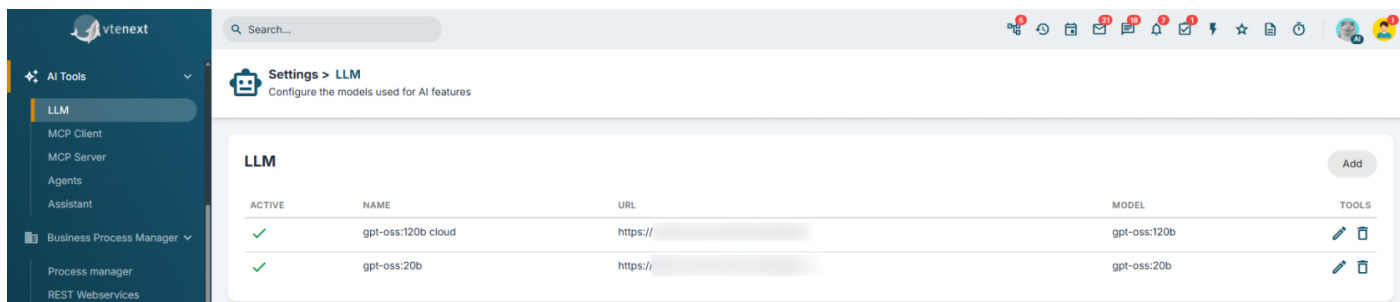
22.1 LLM

The **LLM** section of the settings allows you to register the **Large Language Models (LLMs)** that vtenext can use for AI agents and other AI-powered features. Here, you can define the model's connection details, assign it a name for identification, and configure parameters that influence the model's response behavior.

vtenext does not include a built-in AI model within the CRM. To configure an LLM, you must have either a remote model accessible through **OpenAI-compatible APIs** or a locally installed model that is reachable by the system. In other words, this configuration is used to connect vtenext to an external or local AI service—the model itself is not built into the CRM.

The List View

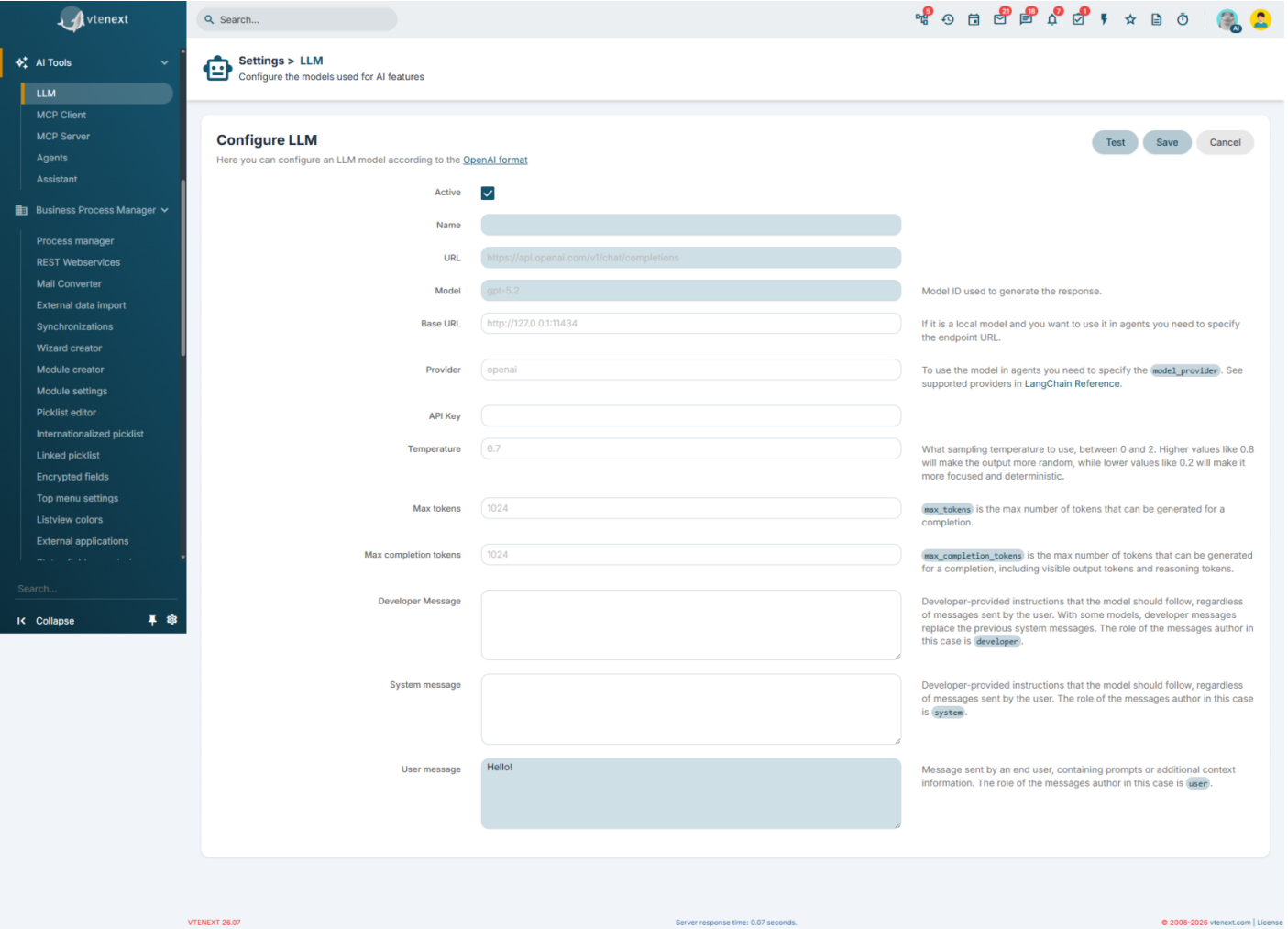
The list displays all previously saved configurations. For each entry, you can view:



- Active or inactive status
- Configuration name
- Service URL
- Model used

From the list, you can create a new configuration, edit an existing entry, delete it, or quickly enable and disable it.

Creating a new LLM



1. Open the LLM section.
2. Click Add.
3. Fill in the required fields.
4. If necessary, run a configuration test.
5. Save.

Base URL

http://127.0.0.1:11434

Provider

openai

API Key

Temperature

0.7

Max tokens

1024

Max completion tokens

1024

If it is a local model and you want to use it in agents you need to specify the endpoint URL.

To use the model in agents you need to specify the `model_provider`. See supported providers in [LangChain Reference](#).

What sampling temperature to use, between 0 and 2. Higher values like 0.8 will make the output more random, while lower values like 0.2 will make it more focused and deterministic.

`max_tokens` is the max number of tokens that can be generated for a completion.

`max_completion_tokens` is the max number of tokens that can be generated for a completion, including visible output tokens and reasoning tokens.

Active	makes the model available for use
Name	descriptive name of the model displayed in vtenext. Required field

URL	API endpoint to which vtenext sends requests to the model (for example, https://api.openai.com/v1/chat/completions). Required field
Model	identifier of the model to be used (for example, gpt-5.2). Required field
Base URL	the base address of the service hosting the model. In practice, it tells AI agents and the Python orchestrator where the model service is located. It becomes particularly important when using a local model or an internal service within the infrastructure, for example an endpoint such as http://127.0.0.1:11434
Provider	indicates to vtenext and the Python orchestrator what type of service is behind the model, such as OpenAI or Ollama
API Key	authentication key provided by the provider, required to authorize API requests
Temperature	controls the level of randomness in generated responses. Lower values (for example, 0.2) produce more consistent, predictable, and repeatable responses; higher values (for example, 0.8 or above) encourage more varied and creative responses
Maximum Tokens	defines the maximum total number of tokens that the model can use to process the request, including both the messages sent and the generated response (if supported by the provider)
Maximum Completion Tokens	limits the maximum number of tokens that the model can use exclusively for the generated response
Developer Message	instructions addressed to the model with the developer role, used to define behavioral rules or application constraints
System Message	general instructions that define the model's behavior during the conversation
User Message	test message sent to the model to verify its operation. Required field

How the test works

The TEST button sends an actual request to the model using the configured parameters and displays:

Configura LLM

Qui puoi configurare un modello LLM compatibile con il [formato OpenAI](#)

Attivo ☒

Nome

URL

Modello ID del modello utilizzato per generare la risposta.

Prova

Salva

Annulla

- Operation result
- Response code
- Returned headers
- Full response body

Compatibile con il [formato OpenAI](#)



Risultato della chiamata

RISULTATO

HEADERS

RISPOSTA

Stato **Success**

Codice di ritorno 200 OK

Call Result: RESULT tab



Risultato della chiamata

RISULTATO

HEADERS

RISPOSTA

Content-Type: application/json

Set-Cookie: aid=4dc0d1aa-5357-4285-a355-c9015add5705; Max-Age=31536000; Path=/; Secure; H

x-build-commit: 63edbbb37578481383768a373789def4d0644e45

x-build-time: 2026-07-09T17:42:29-07:00

x-frame-options: DENY

x-request-id: e890cb3b-bf6c-47ee-bace-ee1d81b829d7

Content-Length: 839

Date: Fri, 10 Jul 2026 15:08:14 GMT

Server: Google Frontend

traceparent: 00-1e4ef07affab53e9a548327743036561-c9c9f3982d1fbb43-00

x-cloud-trace-context: 1e4ef07affab53e9a548327743036561/14540420706859989827

Via: 1.1 google

Alt-Svc: h3=":443"; ma=2592000,h3-29=":443"; ma=2592000

Connection: close

Call Result: HEADERS tab



Risultato della chiamata

RISULTATO

HEADERS

RISPOSTA

Formatta la risposta come: JSON

```
{
  "id": "chatcmpl-363",
  "object": "chat.completion",
  "created": 1783696094,
  "model": "gpt-oss:120b",
  "system_fingerprint": "fp_ollama",
  "choices": [
    {
      "index": 0,
      "message": {
        "role": "assistant",
        "content": "Hello! 🙌 How can I assist you today with VTENEXT? Feel free to let me know what you need—whether it's help with account settings, navigation tips, troubleshooting, or anything else. I'm here to help!",
        "reasoning": "We need to respond as assistant for VTENEXT app. Probably"
      }
    }
  ]
}
```

Call Result: RESPONSE tab

The test is used to verify the configuration, but

The test sends an actual request to the remote or local model. If the model cannot be reached via API, the test cannot be performed.

Practical examples

Example 1: OpenAI-compatible model

Use this configuration when the model is available as a remote service through an external API. Set the name, URL, model, and API Key, then send a simple test message.

Example 2: Local model

If the model runs locally or on an internal infrastructure, also fill in the **Base URL** and **Provider** fields. In this case as well, the model must already be installed, active, and reachable over the network.

Revision #6

Created 2026-07-10 15:21:19 UTC by Admin

Updated 2026-07-13 07:55:10 UTC by Admin