

22.5.1 Configurazione passo-passo KITT

In questa sessione, vedremo come configurare il nostro personale Assistente AI, basandoci sulla piattaforma Ollama. Per fare un po' di chiarezza, Ollama è un programma gratuito ed open source che permette di gestire ed eseguire modelli linguistici di intelligenza artificiale (LLM), mettendo a disposizione appunto dei modelli preconfigurati e potendo poi personalizzarli secondo necessità. In vte, utilizzeremo i modelli di Ollama, in modo da poter esplorare la potenza dello strumento, senza dover per forza partire da zero.

L'esempio che faremo, è quello di creare una persona virtuale. La situazione è che nell'azienda del nostro cliente, un dipendente storico che ha prestato servizio in quella specifica azienda per più di 40 anni, sta per andare in pensione. Molte informazioni sono nella memoria storica di questa persona. Perchè questo tesoro non vada perduto, configureremo lo strumento per trasferire la memoria del futuro pensionato, all'interno di un Assistente Virtuale.

Come prima cosa, andiamo su Ollama che ci richiede la creazione di un account (scegliamo di crearne uno gratuito), inserendo i dati richiesti nell'apposita finestra di registrazione:



Registrati

Email

Il tuo indirizzo email

Continua

OPPURE

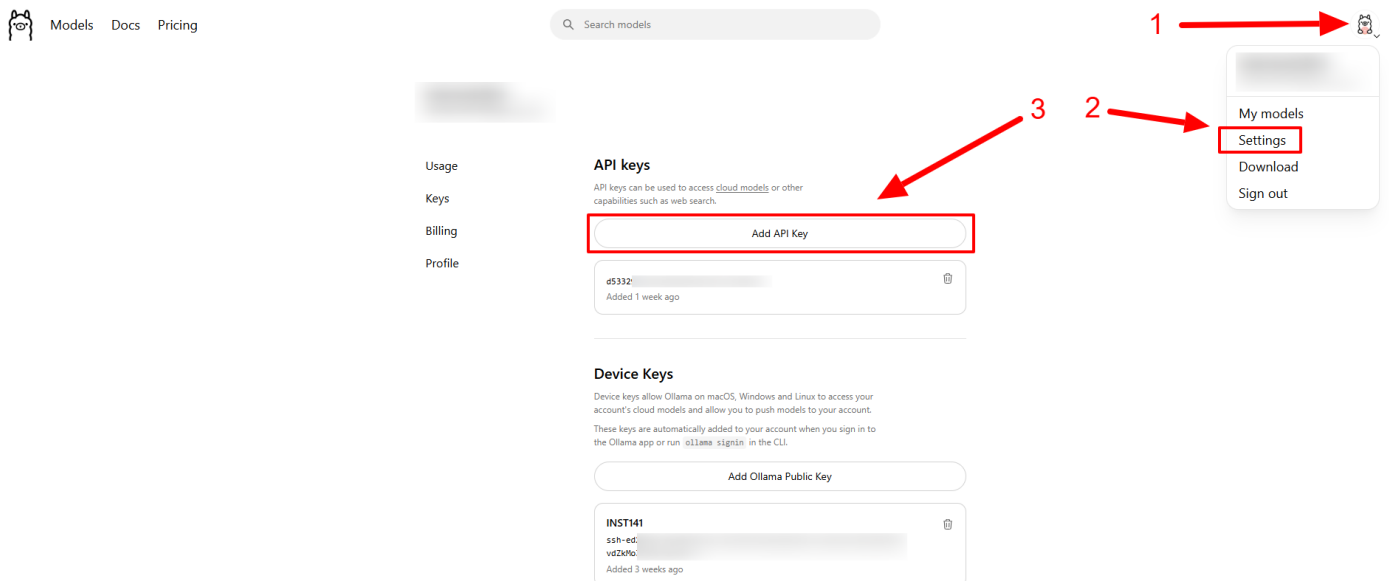
 Continua con Google

 Continua con GitHub

Hai già un account? [Accedi](#)

Schermata di registrazione ad Ollama

Una volta registrati, sarà possibile, cliccando sul proprio avatar in alto a destra, e poi cliccando su Settings, creare un'API Key, fondamentale per iniziare la configurazione su vtenext:

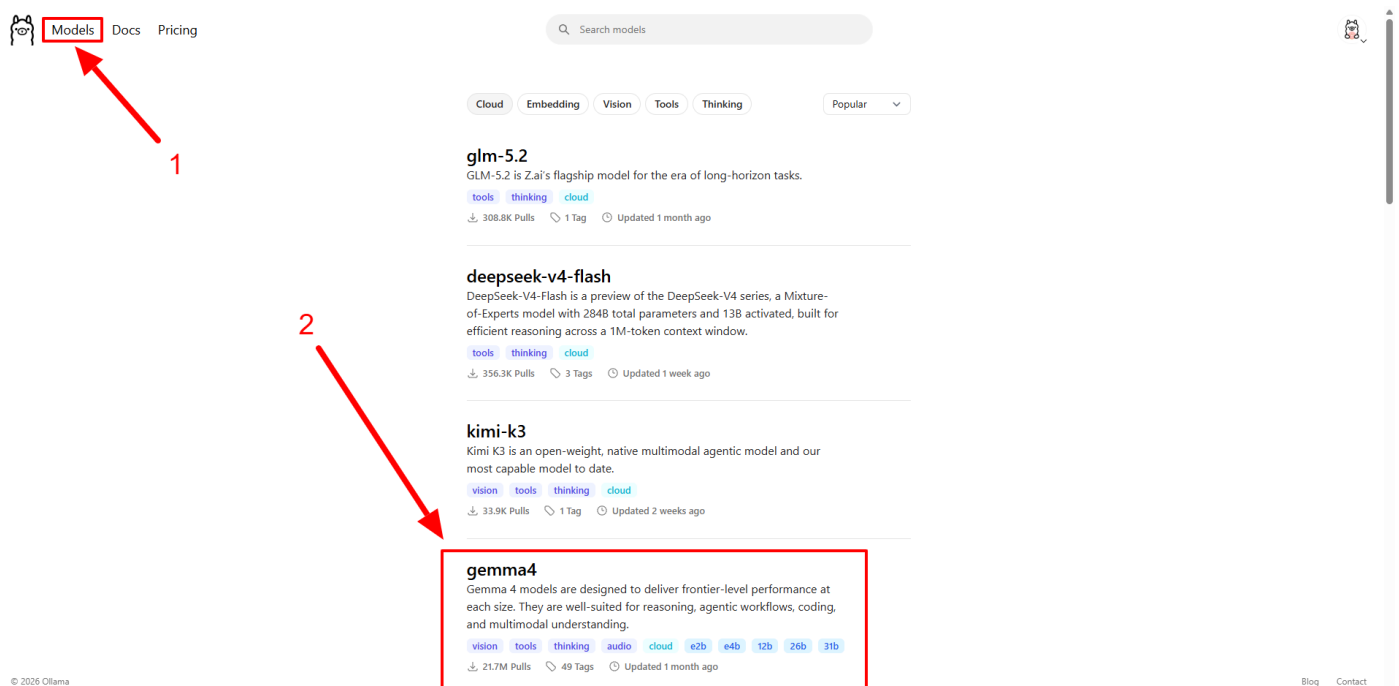


The screenshot shows the Ollama web interface. At the top left, there is a navigation bar with the Ollama logo, 'Models', 'Docs', and 'Pricing'. A search bar labeled 'Search models' is at the top center. On the left side, there is a sidebar menu with 'Usage', 'Keys', 'Billing', and 'Profile'. The main content area is titled 'API keys' and contains a section for adding a new API key. A red box highlights the 'Add API Key' button. Below this, there is a list of existing API keys, with one key shown: 'd5332' added 1 week ago. Below the API keys section, there is a section for 'Device Keys' with an 'Add Ollama Public Key' button. A dropdown menu is open in the top right corner, showing 'My models', 'Settings' (highlighted with a red box), 'Download', and 'Sign out'. Red arrows and numbers indicate the steps to create an API key: 1. Click the user avatar in the top right. 2. Click 'Settings' in the dropdown menu. 3. Click 'Add API Key' in the API keys section.

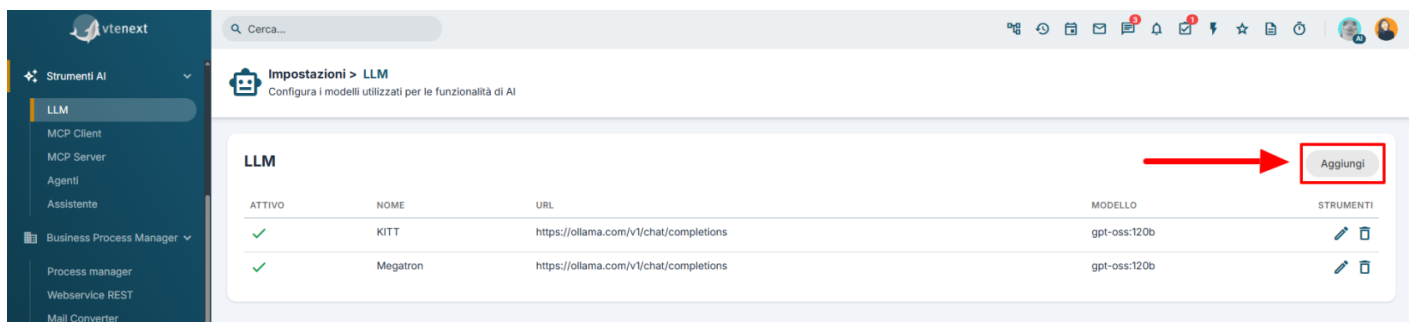
Schermata di generazione/configurazione dell'API Key di Ollama

Come ultimo step da fare sul sito Ollama, è quello di cercare un modello adatto alle nostre esigenze. Si noti che i modelli free ed a pagamento sono inseriti tutti nella stessa lista, quindi è facile scegliere il modello non desiderato. Per continuare il percorso gratuito che vi può aiutare a fare i dovuti test, vi suggeriamo di cercare questi modelli:

- gemma4
- nemotron-3-super
- minimax-m3
- gpt-oss



E' giunta l'ora di spostarci su vtenext e preparare la configurazione per l'**LLM**, ovvero l'interprete del linguaggio naturale. Andando in **IMPOSTAZIONI > STRUMENTI AI > LLM**, vedremo la lista di quelli già configurati e possiamo crearne di nuovi, cliccando sul pulsante **AGGIUNGI** presente in alto a destra:



Supponendo di aver scelto come modello "gemma4", configureremo tutti i campi necessari per agganciare tale modello. Ogni campo avrà una sua precisa configurazione e nello specifico, sarà la seguente (tutti i campi lasciati in bianco, sono opzionali):

Attivo	Spunta attiva per dire che questo interprete può essere utilizzato
Nome	Franco: stiamo cercando di creare una persona virtuale
URL	https://ollama.com/v1/chat/completions
Modello	gemma4:latest
Base URL	https://ollama.com
Provider	ollama
API Key	l'API Key generata su Ollama, copiata ed incollata direttamente in questo campo
Temperatura	0.1 per avere risposte più deterministiche e senza deduzioni o libere interpretazioni
User Message	Io sono Franco, come ti posso aiutare?

vtnext

Strumenti AI

LLM

MCP Client

MCP Server

Agenti

Assistente

Business Process Manager

Process manager

Webservice REST

Mail Converter

Sincronizzazioni

Gestore Cambio di stato

Campi Condizionali

Creazione wizard

Creazione moduli

Gestore Moduli

Editor di Picklist Standard

Editor di Picklist Multi-Ling...

Editor di Picklist Collegato

Editor di campi cifrati

Impostazione voci di menù

Colorazione Viste per Lista

Cerca...

Nascondi

Cerca...

Impostazioni > LLM

Configura i modelli utilizzati per le funzionalità di AI

Modifica LLM

Qui puoi configurare un modello LLM compatibile con il formato OpenAI

Attivo

Nome

URL

Modello

Base URL

Provider

API Key

Temperatura

Token massimi

Token di completamento massimi

Developer Message

System message

User message

Prova

Salva

Annulla

Se è un modello locale e lo vuoi utilizzare negli agenti è necessario indicare l'endpoint.

Per utilizzare il modello negli agenti è necessario indicare il model_provider. Vedi i provider supportati in LangChain Reference.

Temperatura di campionamento da utilizzare, tra 0 e 2. Valori più alti come 0,8 renderanno l'output più casuale, mentre valori più bassi come 0,2 lo renderanno più mirato e deterministico.

max_tokens è il massimo numero di token che possono essere generati per un completamento.

max_completion_tokens è il massimo numero di token che possono essere generati per un completamento, inclusi i token di output visibili e i token di ragionamento.

Istruzioni fornite dallo sviluppatore che il modello deve seguire, indipendentemente dai messaggi inviati dall'utente. Il ruolo dell'autore del messaggio in questo caso è developer. In alcuni modelli, i messaggi developer sostituiscono i precedenti messaggi system.

Istruzioni fornite dallo sviluppatore che il modello deve seguire, indipendentemente dai messaggi inviati dall'utente. Il ruolo dell'autore del messaggio in questo caso è system.

Messaggio inviato da un utente finale, contenente il prompt o informazioni contestuali aggiuntive. Il ruolo dell'autore del messaggio in questo caso è user.

Schermata con le configurazioni fatte per l'LLM di Franco

Una volta inseriti i parametri, cliccare sul pulsante PROVA in alto a destra, per poter verificare che il collegamento funzioni correttamente. Se tutto è andato a buon fine, la schermata con le varie tab, avrà questo aspetto:



Compatibile con il [formato OpenAI](#)

Risultato della chiamata

RISULTATO

HEADERS

RISPOSTA

Stato

Success

Codice di ritorno

200 OK

TAB RISULTATO



Risultato della chiamata

RISULTATO HEADERS RISPOSTA

Content-Type: application/json

Set-Cookie: aid=5013efb1-28e9-4087-b3a6-02aa83299c13; Max-Age=31536000; Path=/; Secure; HttpOnly

x-build-commit: a083b15ea635ff93836cee498abde05672929e5d

x-build-time: 2026-08-12T14:41:44-07:00

x-frame-options: DENY

x-request-id: 2ed16831-4ecc-472d-a99a-f106ab0585b8

Content-Length: 1096

Date: Thu, 13 Aug 2026 14:27:03 GMT

Server: Google Frontend

traceparent: 00-d3e91d550b3fc1a5d97ede36b7510a43-b1f7cd320435a0fb-00

x-cloud-trace-context: d3e91d550b3fc1a5d97ede36b7510a43/12823944078663459067

Via: 1.1 google

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

Connection: close

TAB HEADERS



Risultato della chiamata

RISULTATO HEADERS RISPOSTA

Formatta la risposta come:

JSON



```
{
  "id": "chatcmpl-996",
  "object": "chat.completion",
  "created": 1786631223,
  "model": "gemma4:latest",
  "system_fingerprint": "fp_ollama",
  "choices": [
    {
      "index": 0,
      "message": {
        "role": "assistant",
        "content": "Ciao Franco! Piacere di conoscerti.\n\nIn realtà, essendo io un'intelligenza artificiale, **sono io qui per aiutare te!** 😊\n\nPuoi chiedermi di fare moltissime cose, ad esempio:\n*   **Scrivere o revisionare testi** (email, lettere, articoli, racconti).\n*   **Rispondere a domande** su qualsiasi argomento"
```

TAB RISPOSTA

Impostazioni > LLM
Configura i modelli utilizzati per le funzionalità di AI

LLM

ATTIVO	NOME	URL	MODELLO	STRUMENTI
✓	KITT	https://ollama.com/v1/chat/completions	gpt-oss:120b	
✓	Megatron	https://ollama.com/v1/chat/completions	gpt-oss:120b	
✓	Franco	https://ollama.com/v1/chat/completions	gemma4:latest	

Aggiungi

Cliccando infine sul pulsante SALVA, potremo vedere la lista di tutti gli LLM configurati, tra cui il nostro Franco appena creato.

Andando ora in **IMPOSTAZIONI > STRUMENTI AI > MCP SERVER**, sarà possibile vedere la lista degli MPC già configurati ed aggiungerne di nuovi, cliccando sul pulsante AGGIUNGI in alto a destra.

Impostazioni > MCP Server
Pubblica i server MCP con i tool desiderati

Server MCP

ATTIVO	ENDPOINT	DESCRIZIONE	TOOL PUBBLICATI	CLIENT MCP	STRUMENTI
✓	vte-mcp	Default VTE MCP Server	16	Default VTE MCP Client 13-08-2026 15:44:12	
✓	Megatron	Test MCP server 2	20	mammadu-mcp 13-08-2026 16:32:55	

Aggiungi

Vista per lista degli MCP Server

vtnext

Strumenti AI

LLM

MCP Client

MCP Server

Agenti

Assistente

Business Process Manager

Process manager

Webservice REST

Mail Converter

Sincronizzazioni

Gestore Cambio di stato

Campi Condizionali

Creazione wizard

Creazione moduli

Gestore Moduli

Editor di Picklist Standard

Editor di Picklist Multi-Ling...

Editor di Picklist Collegato

Editor di campi cifrati

Impostazione voci di menù

Colorazione Viste per Lista

Cerca...

Nascondi

Cerca...

Impostazioni > MCP Server

Pubblica i server MCP con i tool desiderati

Configura Server MCP

Salva

Annula

Attivo

Base URL

Endpoint

Descrizione

Crea MCP client

Identificatore univoco usato nell'URL dell'endpoint del server.

Crea automaticamente un MCP client che punta a questo server. Il client viene mantenuto sincronizzato e verrà eliminato quando il server viene rimosso.

Tool pubblicati

Ricerca...

(5 / 24 selected)

Selezione

Base

create

delete

describe

listtypes

processes.processes_relat

query

relate

retrieve

retrieveInventory

touch.get_associated_prod

touch.get_currencies

touch.get_notifications

touch.get_ticket_comments

touch.write_ticket_comments

updateRecord

Processes

esempio_processo_tool

tool_process_sending_email_to_a_customer

ocr_document

Tool SDK

convert_lead

get_current_context

process_text

summarize

tools_manual

translate

VTENEXT 26.07

© 2008-2026 vtnext.com | Licenza

Vista in dettaglio della creazione dell'MCP Server per il nostro Franco

Le opzioni scelte nei TOOL PUBBLICATI, ci permetteranno di fare delle ricerche interne, di avere delle risposte relative solo a quello che vogliamo che il nostro Assistente sia veramente esperto. Inizialmente ne scegliamo pochi, in modo da sperimentare e vedere poi le reazioni che l'Assistente avrà alle nostre domande. Si fa veramente presto a tornare in questa schermata e cambiare le opzioni impostate.

Cliccando infine sul pulsante SALVA in alto a destra, potremmo tornare nella vista per lista di tutti i nostri MCP Server, e troveremo in lista anche quello creato per il nostro Assistente Franco.

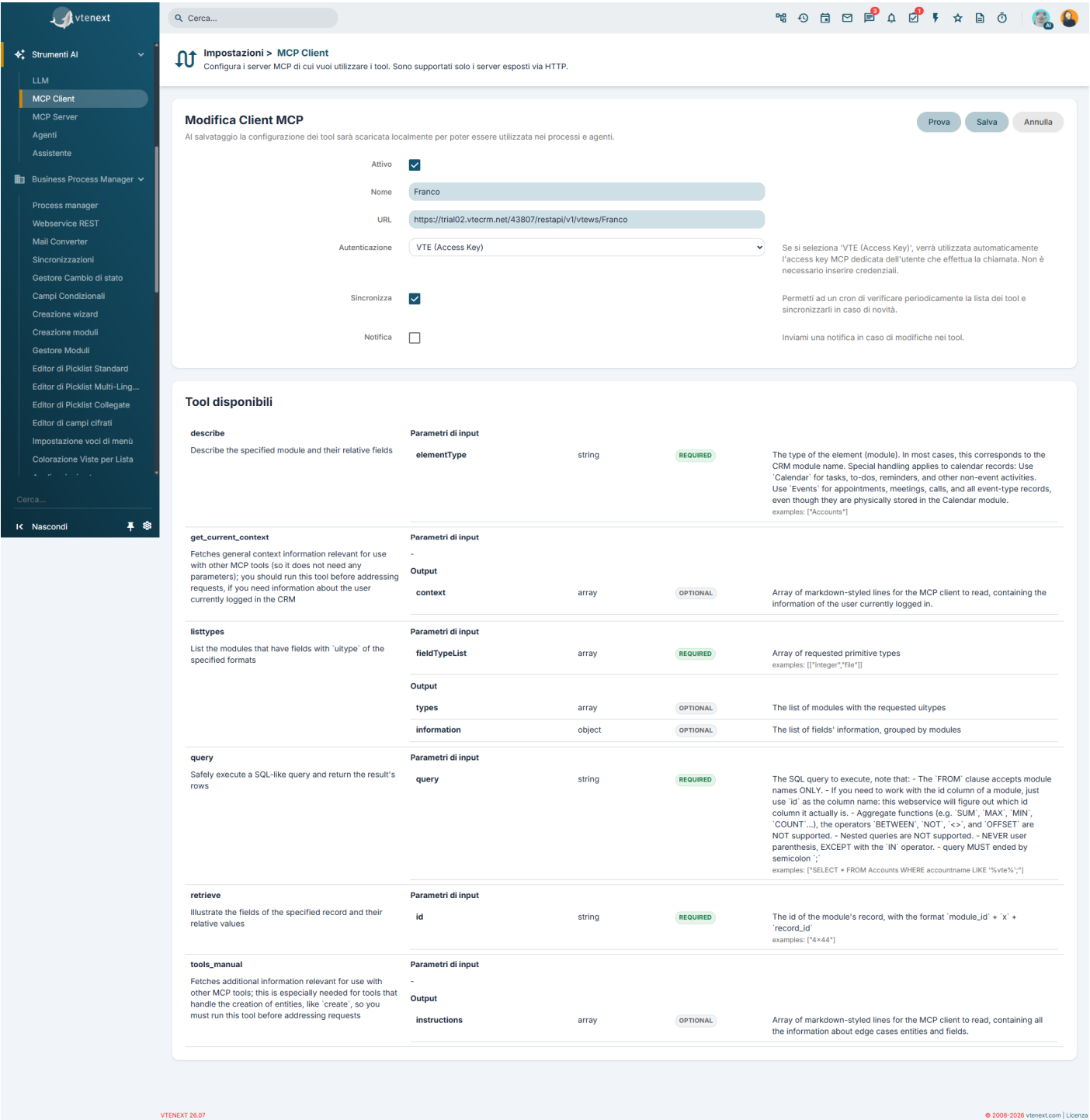
Server MCP						Aggiungi
ATTIVO	ENDPOINT	DESCRIZIONE	TOOL PUBBLICATI	CLIENT MCP	STRUMENTI	
✓	vte-mcp	Default VTE MCP Server	16	Default VTE MCP Client 13-08-2026 15:44:12		
✓	Megatron	Test MCP server 2	20	mammadu-mcp 13-08-2026 16:32:55		
✓	Franco	Franco conosce tutto quello che è inserito nelle FAQ	6	Franco 13-08-2026 16:43:04		

Vista per lista degli MCP Server

Ora serve configurare l'MCP Client, cliccando su IMPOSTAZIONI > STRUMENTI AI > MCP CLIENT, noteremo che esiste già la configurazione fatta per il nostro Assistente Franco. Cliccando sulla matita presente sulla destra, avremmo modo di verificare se è tutto in ordine e possiamo infine testarne il funzionamento cliccando sul pulsante PROVA sulla destra.

Client MCP					Aggiungi
ATTIVO	NOME	URL	ULTIMA SINCRONIZZAZIONE	STRUMENTI	
✓	Default VTE MCP Client	https://trial02.vtecrm.net/43807/restapi/v1/vtews/vte-mcp	13-08-2026 16:50:13		
✓	Web Search MCP - Exa	https://mcp.exa.ai/mcp	13-08-2026 16:56:13		
✓	Megatron	https://trial02.vtecrm.net/43807/restapi/v1/vtews/Megatron	13-08-2026 17:00:17		
✓	Franco	https://trial02.vtecrm.net/43807/restapi/v1/vtews/Franco	13-08-2026 17:00:50		

Vista per lista degli MCP Client



Vista in dettaglio dell'MCP Client per l'Assistente Franco

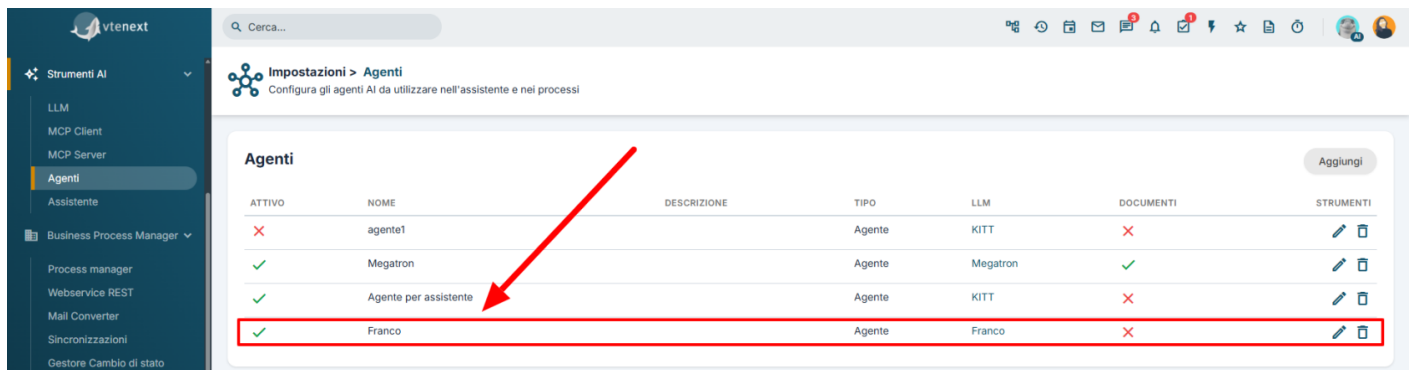
Andando ora in IMPOSTAZIONI > STRUMENTI AI > AGENTI, sarà possibile creare l'agente collegato all'LLM creato inizialmente e decidere quali strumenti utilizzerà per fornirci le sue risposte. Da notare che nella configurazione, abbiamo scelto di selezionare tutte le tools inerenti a Franco che avevamo preimpostato, non abbiamo scelto che Franco possa attingere dal web perchè vogliamo che sia più preciso possibile, facendo in modo che cerchi solamente nelle FAQ e che infine avremmo anche la possibilità di scegliere altre tools provenienti da altri LLM configurati, come Megatron che si vede nell'esempio.

web_search_exa

Search the web for any topic and get clean, ready-to-use content. Best for: Finding current information, news,...

Vista della configurazione dell'Agente per l'assistente Franco

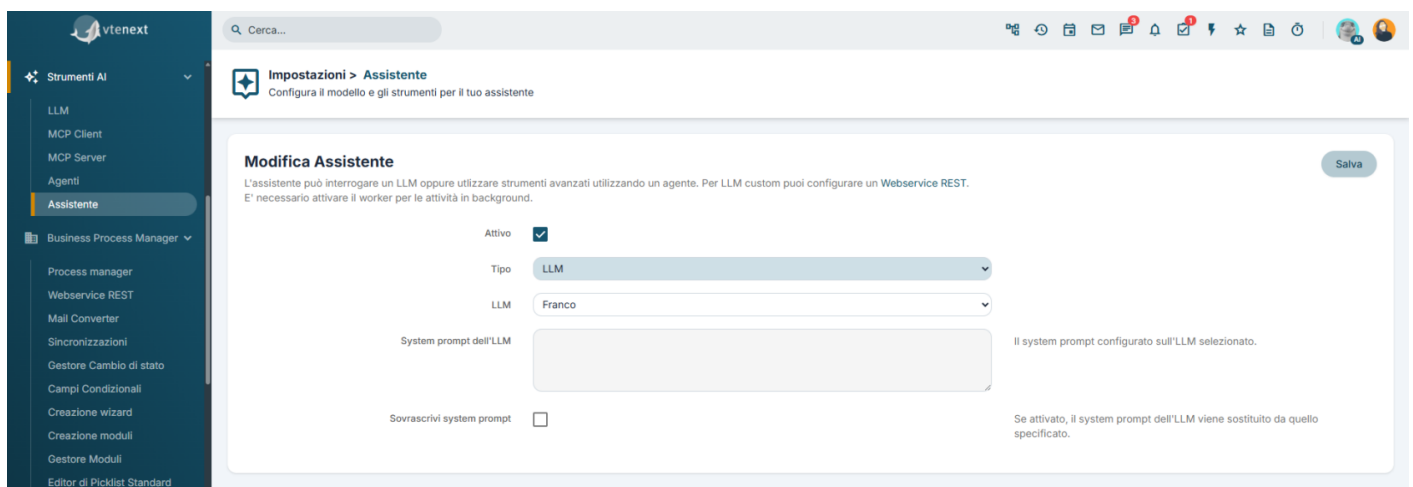
Cliccando infine sul pulsante salva presente in alto a destra, torneremo nella vista per lista, ove troveremo il nostro Agente Franco appena creato e configurato.



ATTIVO	NOME	DESCRIZIONE	TIPO	LLM	DOCUMENTI	STRUMENTI
✗	agente1		Agente	KITT	✗	 
✓	Megatron		Agente	Megatron	✓	 
✓	Agente per assistente		Agente	KITT	✗	 
✓	Franco		Agente	Franco	✗	 

Vista per lista degli Agenti

Cliccando infine su IMPOSTAZIONI > STRUMENTI AI > ASSISTENTE, avremo la possibilità di configurare la fonte da dove il nostro Assistente vero e proprio, attingerà tutte le informazioni e risposte che si appresterà a fornirci. In questo esempio, abbiamo scelto LLM in modo da utilizzare il modello che avevamo pensato sin dall'inizio.



Modifica Assistente

L'assistente può interrogare un LLM oppure utilizzare strumenti avanzati utilizzando un agente. Per LLM custom puoi configurare un Webservice REST. E' necessario attivare il worker per le attività in background.

Attivo ☒

Tipo **LLM**

LLM **Franco**

System prompt dell'LLM

Sovrascrivi system prompt ☐

Il system prompt configurato sull'LLM selezionato.

Se attivato, il system prompt dell'LLM viene sostituito da quello specificato.

Vista in dettaglio della configurazione dell'Assistente

Andando ora nel modulo FAQ del crm ed iniziando a scrivere tutte le possibili "domande e risposte" riguardo tutti gli argomenti che ipoteticamente il collega Franco conosce, possiamo creare una sorta di database della sua conoscenza.

vt
next

Moduli

Anagrafiche

Marketing

Progetti

Vendite

Inventario

Collaboration

Tutti

Assistenza Clienti

Aziende

Calendario

Campagne

Casi

Cestino

Collaboratori

Contatti

Conversazioni

Cerca...

LISTA

+ Crea

Altro

Filtro

Tutti

Visualizzando da 1 a 19 di 19

Azione	Numero Faq	Domanda	Risposta	Categoria	Stato	Data Creazione
☐						
☐		FAQ19	Si possono saldare le flange?	Generale	Pubblicato	06-08-2026 15:39:41
☐		FAQ18	Quale gelato preferisci?	Generale	Pubblicato	06-08-2026 15:34:45
☐		FAQ17	Quale è la procedura di tenuta e pressio...	Pensionato	Pubblicato	06-08-2026 15:13:27
☐		FAQ16	Come ci si prepara per la saldatura di f...	Pensionato	Pubblicato	06-08-2026 15:12:00
☐		FAQ15	Come si controlla la dimensione del corp...	Pensionato	Pubblicato	06-08-2026 15:10:03
☐		FAQ14	Quale è la procedura di saldatura del co...	Pensionato	Pubblicato	06-08-2026 15:09:30
☐		FAQ13	Quale è la sequenza di assemblaggio del ...	Pensionato	Pubblicato	06-08-2026 15:08:56

Revision #34

Created 2026-08-06 12:56:15 UTC by Admin

Updated 2026-09-08 10:14:32 UTC by Alberto