*Paolo Dalprato*

# Ollama Easy GUI

*Ultimo aggiornamento: 17 gennaio 2026*

**Un'interfaccia grafica per usare modelli AI sul tuo computer, senza abbonamenti e senza inviare dati a servizi esterni.**

---

## Perché usare AI in locale

Quando usi ChatGPT, Claude o altri assistenti AI online, le tue conversazioni viaggiano verso server remoti. Per molti usi va benissimo, ma ci sono situazioni in cui preferiresti che i tuoi dati restassero sul tuo computer: documenti aziendali riservati, riflessioni personali, progetti che non vuoi condividere con nessuno.

Ollama Easy GUI risolve questo problema: è un'applicazione che ti permette di conversare con modelli AI che funzionano interamente sul tuo PC. Nessun dato esce dal tuo computer, nessun abbonamento da pagare, nessuna limitazione sul numero di messaggi.

Il repository su Github è ollama-easy-gui (https://github.com/paolodalprato/ollama-easy-gui)

## Cosa puoi fare con questa applicazione

L'applicazione offre le funzionalità essenziali per lavorare con AI in locale:

- **Conversare con modelli AI** scegliendo tra decine di opzioni gratuite
- **Allegare documenti** per farli analizzare: PDF, immagini, file Word e testo
- **Salvare ed esportare** le conversazioni in vari formati
- **Personalizzare il comportamento** dei modelli con istruzioni permanenti
- **Estendere le capacità** dell'AI permettendole di leggere file o cercare informazioni

## Requisiti essenziali

Per usare Ollama Easy GUI servono:

- Un computer con Windows, macOS o Linux

- Almeno 8 GB di RAM (16 GB consigliati)
- Spazio su disco per i modelli AI (da 2 a 50 GB a seconda del modello)
- Una connessione internet per il download iniziale

> **Serve una GPU?**
>
> No, non è obbligatoria. I modelli funzionano anche solo con la CPU, anche se più lentamente, ma una scheda grafica NVIDIA adatta ai modelli che si intendono usare accelera notevolmente le risposte.

> **Questo manuale è per Windows**
>
> L'applicazione è stata sviluppata e testata su Windows. Dovrebbe funzionare anche su macOS e Linux, ma questi sistemi non sono ancora stati testati. Le istruzioni in questo manuale fanno riferimento a Windows; se usi un altro sistema operativo, consulta l'appendice Note per macOS e Linux (https://docs.ai-know.pro/ollama-easy-gui/altri-sistemi/) per le differenze principali.

## Come è organizzato questo manuale

Il manuale ti guida passo dopo passo dall'installazione all'uso avanzato:

1. **Preparare il computer**: installare i programmi necessari
2. **Installazione**: scaricare e avviare Ollama Easy GUI
3. **La tua prima chat**: creare una conversazione e scegliere un modello
4. **Gestire i modelli**: scaricare nuovi modelli e capire quale scegliere
5. **Personalizzare le risposte**: configurare il comportamento dell'AI
6. **Allegati e export**: lavorare con documenti e salvare le conversazioni
7. **MCP: estendere l'AI**: permettere all'AI di usare strumenti esterni
8. **Risoluzione problemi**: cosa fare quando qualcosa non funziona

## A chi è rivolto

Questo manuale è pensato per chi:

- Ha familiarità con l'uso del computer ma non con la programmazione
- Vuole provare AI locali senza dover imparare comandi da terminale
- Cerca un'alternativa ai servizi cloud per motivi di privacy o costi
- È curioso di capire come funzionano i modelli AI "dietro le quinte"

Non servono competenze tecniche avanzate: dove necessario, le procedure sono spiegate passo per passo con immagini.

# Indice

## Guida completa



## Preparare il computer



## Installazione



## L'interfaccia

- Nascondere le sidebar
- Chiudere l'applicazione

## La tua prima chat

- Creare una nuova conversazione
- Inviare un messaggio
- Gestire le conversazioni
- Esportare una risposta
- Cambiare modello durante la conversazione
- Il primo esperimento
- Suggerimenti per prompt efficaci

## Gestire i modelli

- Modelli locali vs Hub
- I modelli installati
- Scaricare nuovi modelli
- Capire i nomi dei modelli
- Quale modello scegliere
- Rimuovere un modello
- Modelli e MCP

## Personalizzare le risposte

- Cos'è un system prompt
- Base prompt: personalità per modello
- Global system prompt: istruzioni universali
- Come funzionano insieme
- Buone pratiche

## Allegati e export

- Allegare file
- Limiti degli allegati

- Esportare le conversazioni
- Dove vengono salvati i dati
- Formati di export a confronto

## MCP: estendere l'AI

- Cos'è MCP in parole semplici
- Modelli compatibili
- Attivare MCP
- Configurare gli strumenti
- MCP in azione
- Repository
- Privacy e MCP
- Risoluzione problemi MCP

## Risoluzione problemi

- Problemi di installazione
- Problemi di avvio
- Problemi con i modelli
- Problemi con gli allegati
- Problemi con MCP
- Consultare i log
- Ottenere supporto

## Note per macOS e Linux

- Installazione dei prerequisiti
- Differenze operative
- Supporto GPU

# Preparare il computer

Prima di installare Ollama Easy GUI, devi preparare il tuo computer con tre programmi di supporto. Non preoccuparti se non li conosci: ti guiderò passo per passo.

## Panoramica dei prerequisiti

| Programma | A cosa serve | Tempo stimato |
|---|---|---|
| **Ollama** | Fa funzionare i modelli AI sul tuo computer | 5 minuti |
| **Git** | Scarica l'applicazione da internet | 3 minuti |
| **Node.js** | Esegue l'interfaccia grafica | 3 minuti |

**Controlla prima**

Potresti già avere alcuni di questi programmi installati. Per verificare, apri il Prompt dei comandi (cerca "cmd" nel menu Start) e digita i comandi indicati in ogni sezione.

---

## Installare Ollama

Ollama è il "motore" che fa funzionare i modelli AI. Senza Ollama, l'interfaccia grafica non avrebbe nulla da far girare.

### Procedura per Windows

1. Vai su ollama.ai (https://ollama.ai)
2. Clicca sul pulsante **Download for Windows**
3. Esegui il file scaricato e segui le istruzioni di installazione
4. Al termine, Ollama si avvierà automaticamente (vedrai una piccola icona nell'area di notifica)

### Verifica l'installazione

Apri il Prompt dei comandi e digita:

```
ollama --version
```

Se vedi un numero di versione (es. "ollama version 0.5.4"), l'installazione è riuscita.

### Scarica un primo modello

Ollama da solo non include modelli AI: devi scaricarne almeno uno. Per iniziare, consiglio **llama3.2** che è leggero e funziona bene sulla maggior parte dei computer.

Nel Prompt dei comandi digita:

```
ollama pull llama3.2
```

Il download richiede qualche minuto (il modello pesa circa 2 GB). Una volta completato, il modello sarà disponibile per sempre sul tuo computer.

> **Modelli alternativi**
>
> Puoi scaricare altri modelli in seguito direttamente dall'interfaccia di Ollama Easy GUI, senza usare il Prompt dei comandi.

---

## Installare Git

Git è uno strumento che gli sviluppatori usano per condividere software. Nel nostro caso, serve per scaricare Ollama Easy GUI da GitHub, il sito dove è pubblicato il codice.

### Procedura per Windows

1. Vai su git-scm.com (https://git-scm.com)
2. Clicca su **Download for Windows**
3. Esegui il file scaricato
4. Durante l'installazione, puoi lasciare tutte le opzioni predefinite (clicca "Next" fino alla fine)

### Verifica l'installazione

Chiudi e riapri il Prompt dei comandi (importante: deve essere una nuova finestra), poi digita:

```
git --version
```

Se vedi qualcosa come "git version 2.47.1", Git è installato correttamente.

## Installare Node.js

Node.js è un ambiente che permette di eseguire applicazioni scritte in JavaScript. Ollama Easy GUI usa Node.js per far funzionare la sua interfaccia grafica.

### Procedura per Windows

1. Vai su nodejs.org (https://nodejs.org)
2. Scarica la versione **LTS** (Long Term Support), quella raccomandata per la maggior parte degli utenti
3. Esegui il file scaricato e segui le istruzioni

### Verifica l'installazione

Chiudi e riapri il Prompt dei comandi, poi digita:

```
node --version
```

Dovresti vedere un numero di versione (es. "v22.12.0"). Se il numero inizia con 16 o superiore, sei a posto.

---

## Riepilogo

Se hai seguito tutti i passaggi, ora hai:

- [x] Ollama installato e funzionante
- [x] Almeno un modello AI scaricato (llama3.2)
- [x] Git pronto per scaricare software
- [x] Node.js pronto per eseguire l'applicazione

Sei pronto per installare Ollama Easy GUI. Passa al capitolo successivo.

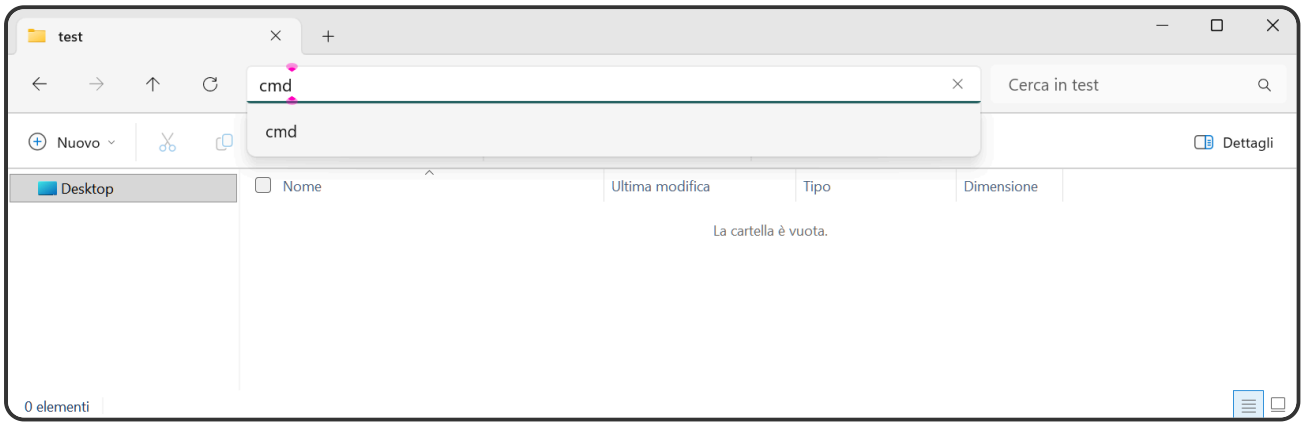
> ⚠ **Problemi?**
>
> Se qualcosa non funziona, controlla la sezione Risoluzione problemi (https://docs.ai-know.pro/ollama-easy-gui/risoluzione-problemi/). I problemi più comuni riguardano i permessi di amministratore durante l'installazione o la necessità di riavviare il computer.

# Installazione

Con i prerequisiti pronti, puoi ora scaricare e avviare Ollama Easy GUI. La procedura richiede pochi minuti.

## Scaricare l'applicazione

Scegli una cartella in cui installare l'applicazione: può essere una cartella già esistente oppure puoi crearne una nuova. Apri la cartella con Esplora risorse, in alto nella riga del percorso scrivi **cmd** e dai il return, si apre il Prompt dei comandi già posizionato sulla cartella di installazione.



(https://docs.ai-know.pro/ollama-easy-gui/img/cmd.png)

Ora scarica l'applicazione con Git:

```
git clone https://github.com/paolodalprato/ollama-easy-gui.git
```

Git creerà una cartella chiamata `ollama-easy-gui` con tutti i file necessari.

## Installare le dipendenze

Entra nella cartella appena creata:

```
cd ollama-easy-gui
```

Installa le librerie necessarie all'applicazione:

```
npm install
```

Questo comando scarica alcuni componenti aggiuntivi. L'operazione richiede circa un minuto e devi eseguirla solo la prima volta.

# Avviare l'applicazione

Hai due modi per avviare Ollama Easy GUI: il metodo rapido e il metodo con configurazione.

## Metodo rapido (npm start)

Il modo più semplice: apri il Prompt dei comandi nella cartella dell'applicazione e digita:

```
npm start
```

L'applicazione si avvia e mostra l'indirizzo dove raggiungerla, tipicamente `http://localhost:3003`. Apri questo indirizzo nel tuo browser per usare l'interfaccia.

## Metodo con configurazione (file .bat)

Su Windows puoi usare un file batch che offre funzionalità aggiuntive: verifica degli aggiornamenti e configurazione ottimale per la tua scheda grafica.

Prima crea il file di configurazione: nella cartella dell'applicazione trovi il file `start-ollama-easy-gui.bat.example`. Copialo e rinominalo in `start-ollama-easy-gui.bat` (puoi anche dargli un altro nome, l'importante è che l'estensione sia `.bat`).

Poi modifica il file con un editor di testo (il Blocco Note va benissimo) per adattarlo al tuo hardware:

```
:: Numero di layer da eseguire sulla GPU (più alto = più veloce, ma serve più VRAM)
set "OLLAMA_GPU_LAYERS=18"

:: Abilita Flash Attention per risposte più veloci
set "OLLAMA_FLASH_ATTENTION=1"

:: Thread della CPU da usare per l'inferenza
set "OLLAMA_NUM_THREADS=24"
```

**Quali valori usare?**

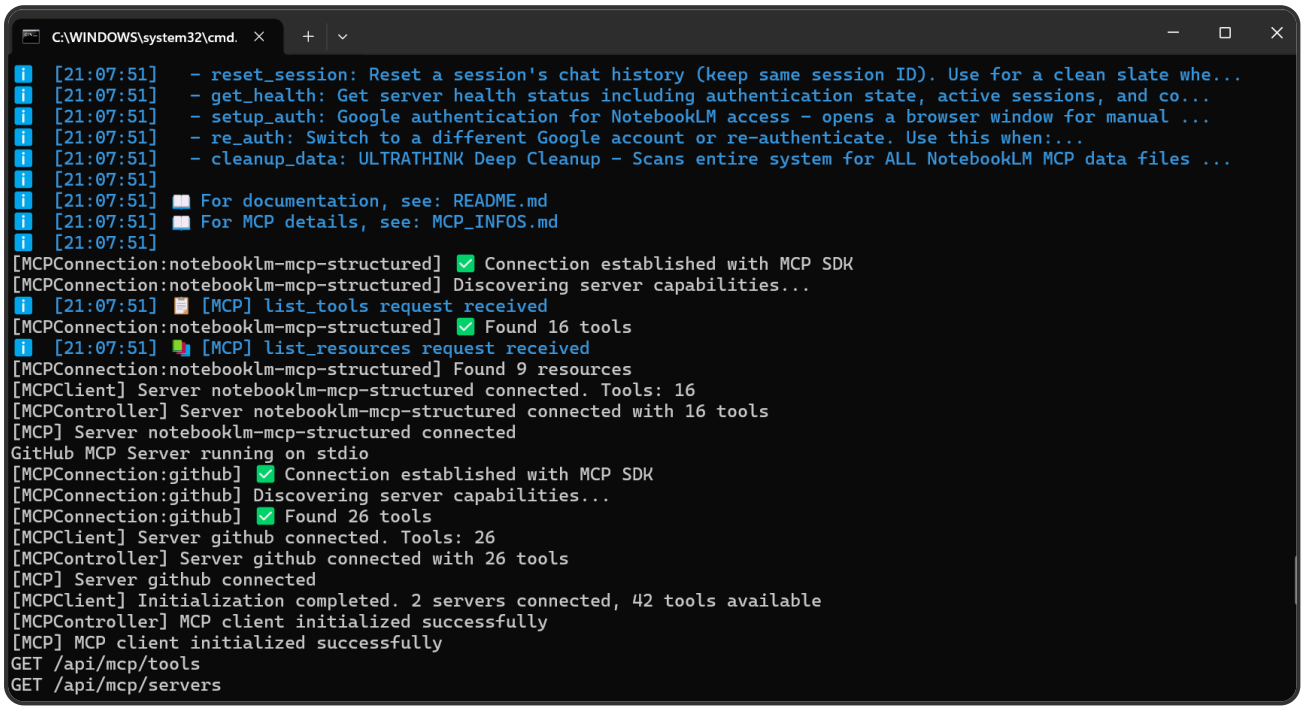
- **OLLAMA_GPU_LAYERS**: con 6 GB di VRAM puoi provare 20-25, con 8 GB prova 30-35
- **OLLAMA_NUM_THREADS**: imposta un numero uguale ai core della tua CPU (es. 8 per un quad-core con hyperthreading)

D'ora in poi, avvia l'applicazione facendo doppio clic sul file `start-ollama-easy-gui.bat`. A ogni avvio il file controlla automaticamente se ci sono aggiornamenti e li installa.

## Accedere all'interfaccia

Qualunque metodo tu abbia usato, l'interfaccia sarà disponibile nel browser all'indirizzo:

```
http://localhost:3003
```



(https://docs.ai-know.pro/ollama-easy-gui/img/terminale-avvio.png)

Se nella finestra del terminale non ci sono messaggi di errore (come nello screenshot), l'applicazione è avviata correttamente. Apri l'indirizzo nel browser per accedere all'interfaccia grafica.

## Aggiornare l'applicazione

Se usi il file `.bat`, gli aggiornamenti vengono scaricati e installati automaticamente a ogni avvio.

Se invece avvii con `npm start`, puoi aggiornare manualmente:

1. Apri il Prompt dei comandi nella cartella dell'applicazione
2. Scarica gli aggiornamenti:

   ```
   git pull
   ```

3. Aggiorna le dipendenze:

   ```
   npm install
   ```

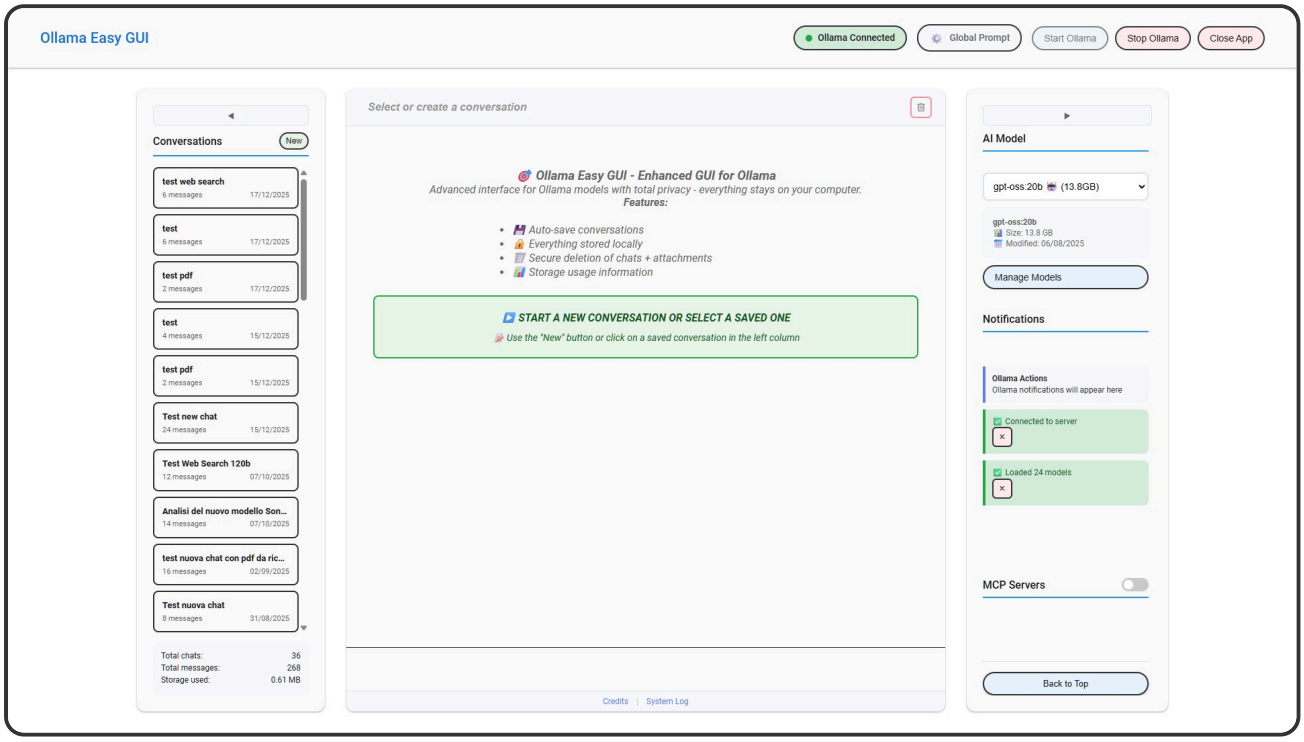
**Le tue impostazioni sono al sicuro**

Gli aggiornamenti non sovrascrivono le tue conversazioni, i prompt salvati o le configurazioni personalizzate. Tutti i tuoi dati sono nella cartella `app/data` che non viene toccata dagli aggiornamenti.

# L'interfaccia

Prima di iniziare a usare l'applicazione, prendiamoci un momento per conoscere l'interfaccia e le sue componenti.

## Panoramica generale




(https://docs.ai-know.pro/ollama-easy-gui/img/start.jpg)

L'interfaccia è organizzata in quattro aree principali:

- **Barra superiore**: titolo, stato di Ollama e controlli globali
- **Sidebar sinistra**: elenco delle conversazioni
- **Area centrale**: la chat vera e propria
- **Sidebar destra**: selezione modello e impostazioni

## Barra superiore

La barra in alto contiene i controlli principali dell'applicazione:

- **Titolo**: "Ollama Easy GUI" - cliccandolo si aprono i crediti con le informazioni su versione, autore e licenza

- **Stato Ollama**: indicatore visivo che mostra se il servizio Ollama è attivo o meno
- **Global Prompt**: apre l'editor per le istruzioni che si applicano a tutti i modelli (approfondito nel capitolo Personalizzazione)
- **Start Ollama**: avvia il servizio Ollama se non è attivo
- **Stop Ollama**: ferma il servizio Ollama
- **Close App**: chiude completamente l'applicazione

## Sidebar sinistra

La colonna di sinistra gestisce le conversazioni:

- **Pulsante New**: crea una nuova conversazione
- **Lista conversazioni**: tutte le chat salvate, ordinate per data

Cliccando su una conversazione la apri nell'area centrale.

In fondo alla sidebar trovi alcune **statistiche**:

- Numero totale di chat
- Numero totale di messaggi
- Spazio occupato su disco

## Area centrale

È lo spazio principale dove avviene la conversazione:

- **Titolo della chat**: in alto, mostra il nome della conversazione attuale. È un testo editabile: cliccaci sopra per rinominare la chat
- **Messaggi**: i tuoi messaggi e le risposte del modello, in ordine cronologico
- **Area di input**: in basso, dove scrivi i messaggi
- **Pulsante allega** (graffetta): per aggiungere file alla conversazione
- **Pulsante invia**: invia il messaggio (oppure premi Invio)

Su ogni risposta del modello trovi delle icone per:

- **Copiare** il testo negli appunti
- **Esportare** la risposta in vari formati (Markdown, testo, Word)

### Footer dell'area centrale

In fondo all'area centrale trovi due elementi:

- **Log**: apre il visualizzatore dei log dell'applicazione, utile per diagnosticare problemi. I log sono divisi per categoria (App, Chat, MCP, Models). Per i dettagli consulta la sezione "Consultare i log" nel capitolo Risoluzione problemi
- **Credits**: mostra le informazioni sull'applicazione (versione, autore, licenza)

## Sidebar destra

La colonna di destra contiene le impostazioni principali:

### Selezione modello

Il menu a tendina **AI Model** mostra tutti i modelli disponibili. Il modello selezionato qui diventa quello predefinito per le nuove conversazioni.

I modelli con un asterisco (*) dopo il nome hanno un prompt di sistema personalizzato.

### Pulsante Manage Models

Apre la finestra di gestione modelli con due schede:

- **Local Models**: modelli già scaricati sul computer
- **Download from Hub**: catalogo online per scaricare nuovi modelli

### Sezione MCP

Quando attivi l'interruttore MCP, compare il pannello dei server MCP configurati. Ogni server mostra:

- **Nome e descrizione**
- **Stato di connessione**: 🟢 connesso, 🔴 disconnesso, ⚪ disabilitato
- **Pulsante Enable/Disable**: per attivare o disattivare il server

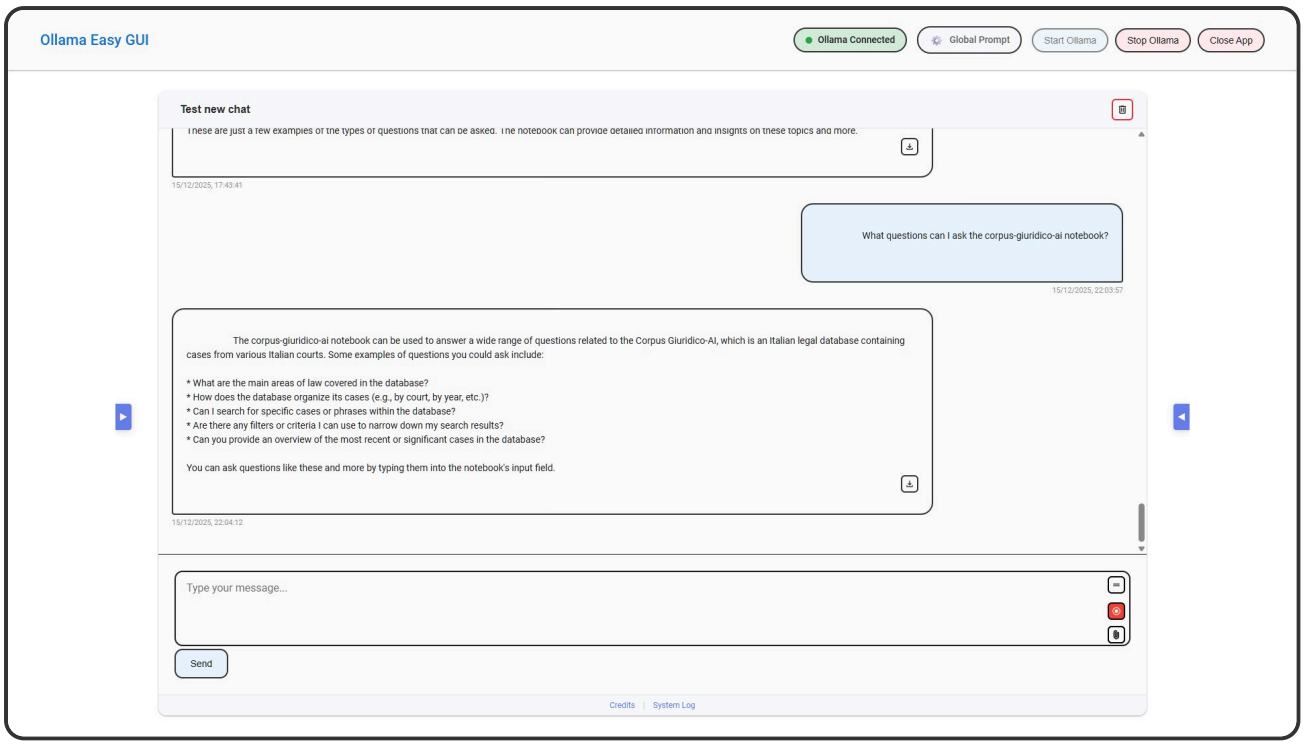
L'argomento MCP è approfondito nel capitolo dedicato.

### Notifiche

In basso nella sidebar compaiono le notifiche dell'applicazione: conferme di operazioni completate, avvisi ed eventuali errori.

## Nascondere le sidebar

Su schermi piccoli o se preferisci più spazio per la chat, puoi nascondere entrambe le sidebar cliccando sulle frecce ai bordi. L'interfaccia si adatta mostrando solo l'area centrale.



(https://docs.ai-know.pro/ollama-easy-gui/img/sidebar-collassata.jpg)

Per riaprire una sidebar, clicca nuovamente sulla freccia corrispondente.

## Chiudere l'applicazione

Per chiudere Ollama Easy GUI hai due opzioni:

- **Close App** nella barra superiore: chiude l'applicazione in modo pulito, fermando tutti i processi
- **Chiudere la finestra del Prompt dei comandi**: se hai avviato l'app da terminale, chiudere quella finestra termina l'applicazione

**Ollama continua a funzionare**

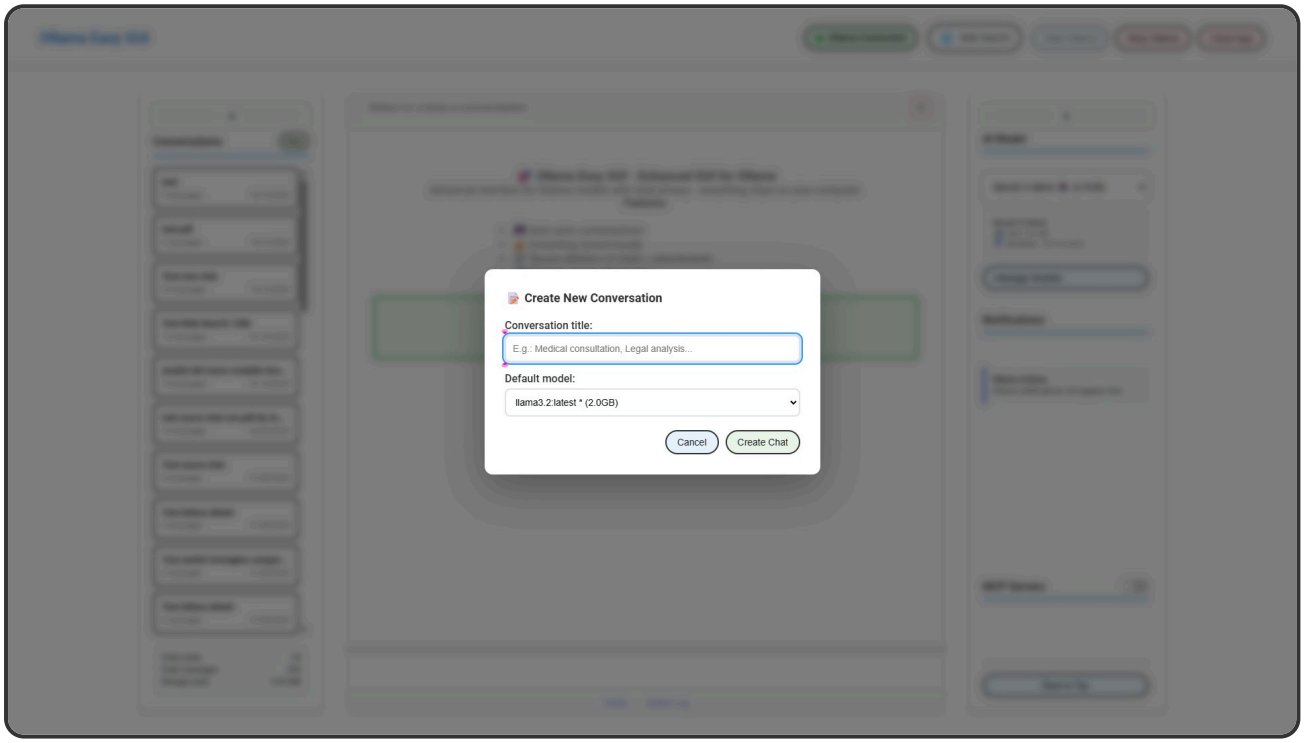
Chiudere Ollama Easy GUI non ferma il servizio Ollama, che continua a girare in background. Se vuoi fermare anche Ollama, usa il pulsante **Stop Ollama** prima di chiudere l'app, oppure chiudi Ollama dall'icona nell'area di notifica di Windows.

# La tua prima chat

Ora che conosci l'interfaccia, vediamo come usarla per conversare con un modello AI.

## Creare una nuova conversazione

Per iniziare una chat, clicca il pulsante **New** nella parte alta della sidebar sinistra. Si apre un popup che ti permette di selezionare il modello da usare:




(https://docs.ai-know.pro/ollama-easy-gui/img/nuova-chat.jpg)

Il popup mostra il modello di default (quello selezionato nella sidebar destra), ma puoi cambiarlo per questa specifica conversazione. Conferma per aprire una nuova chat pronta per ricevere il tuo primo messaggio.

**Modello di default**

Il modello selezionato nella sidebar destra diventa il default per le nuove conversazioni. Se usi sempre lo stesso modello, impostalo lì per non doverlo scegliere ogni volta.

## Inviare un messaggio

Scrivi il tuo messaggio nell'area di testo in basso e premi **Invio** oppure clicca il pulsante di invio. Il modello inizierà a rispondere immediatamente, con il testo che appare progressivamente come in una chat normale.

> **Risposte in streaming**
>
> Le risposte appaiono parola per parola mentre vengono generate. Questo ti permette di leggere l'inizio della risposta senza aspettare che sia completa.

## Gestire le conversazioni

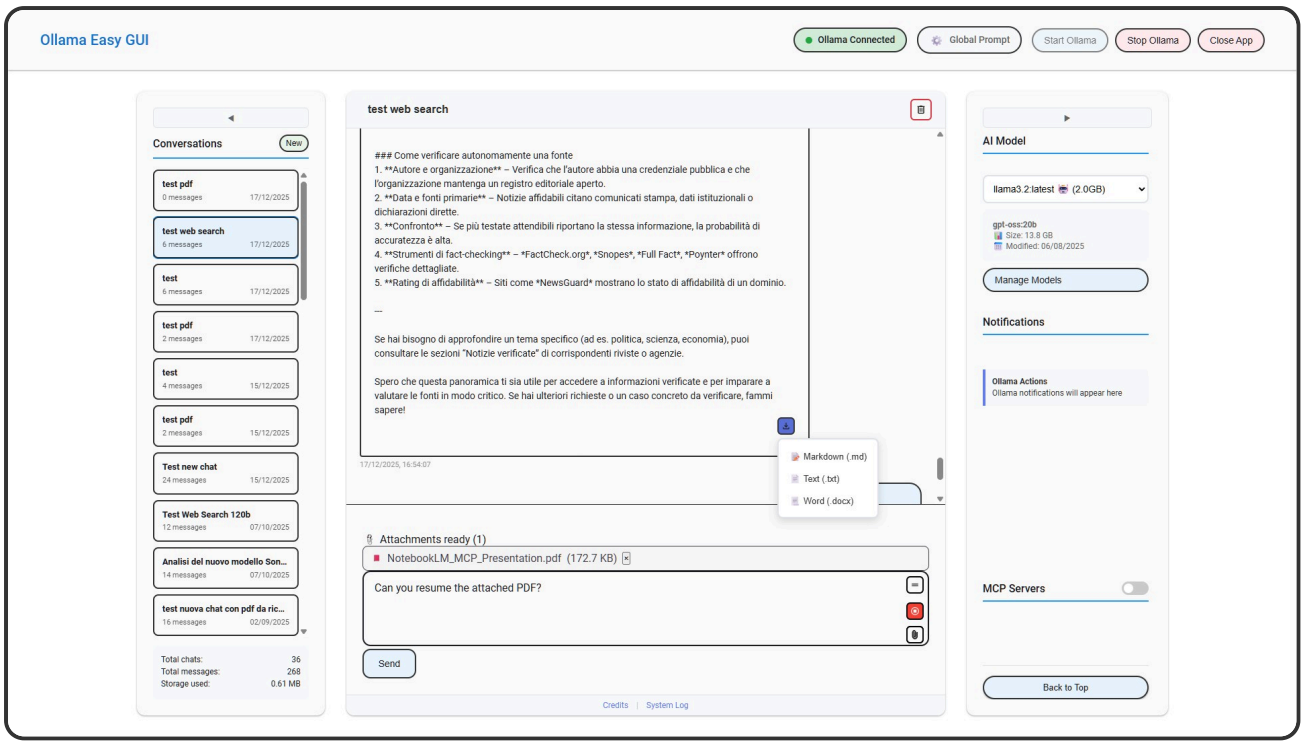
Ogni conversazione viene salvata automaticamente nella sidebar sinistra. Puoi:

- **Rinominare** una conversazione cliccando sul suo titolo (quando è aperta nell'area centrale)
- **Riprendere** una conversazione passata selezionandola dalla lista
- **Eliminare** una conversazione con il pulsante cestino

Le conversazioni vengono salvate sul tuo computer nella cartella `app/data/conversations`. Ogni conversazione ha la sua sottocartella che contiene sia i messaggi che eventuali allegati.

## Esportare una risposta

Puoi salvare una singola risposta in diversi formati cliccando sull'icona di download in fondo a destra di ogni risposta:



(https://docs.ai-know.pro/ollama-easy-gui/img/chat-con-download.jpg)

I formati disponibili sono:

- **Markdown** (.md): ideale se usi editor che lo supportano
- **Testo** (.txt): il formato più universale
- **Word** (.docx): per documenti formali

## Cambiare modello durante la conversazione

Puoi cambiare modello in qualsiasi momento selezionandone uno diverso dal menu a tendina nella sidebar destra. Il nuovo modello continuerà la conversazione da dove l'hai lasciata, ma avrà una "personalità" diversa in base alle sue caratteristiche.

> **Comportamento dei modelli**
>
> Modelli diversi rispondono in modo diverso. Alcuni sono più creativi, altri più precisi, altri ancora più adatti a compiti specifici come la programmazione. Sperimenta per trovare quello che preferisci.

## Il primo esperimento

Prova a inviare un messaggio semplice per verificare che tutto funzioni:

```
Ciao! Puoi presentarti brevemente?
```

Se ricevi una risposta, complimenti: hai appena eseguito la tua prima inferenza AI locale. Tutto è avvenuto sul tuo computer, senza che nessun dato sia stato inviato a internet.

## Suggerimenti per prompt efficaci

Per ottenere risposte migliori:

- **Sii specifico**: invece di "parlami della storia", chiedi "riassumi le cause della Prima Guerra Mondiale in 5 punti"
- **Dai contesto**: spiega cosa vuoi ottenere e perché
- **Chiedi un formato**: specifica se vuoi una lista, un paragrafo, una tabella
- **Itera**: se la prima risposta non è perfetta, chiedi di modificarla o approfondire

I modelli locali funzionano meglio con istruzioni chiare. Nel prossimo capitolo vedremo come scaricare modelli aggiuntivi e scegliere quello più adatto alle tue esigenze.

# Gestire i modelli

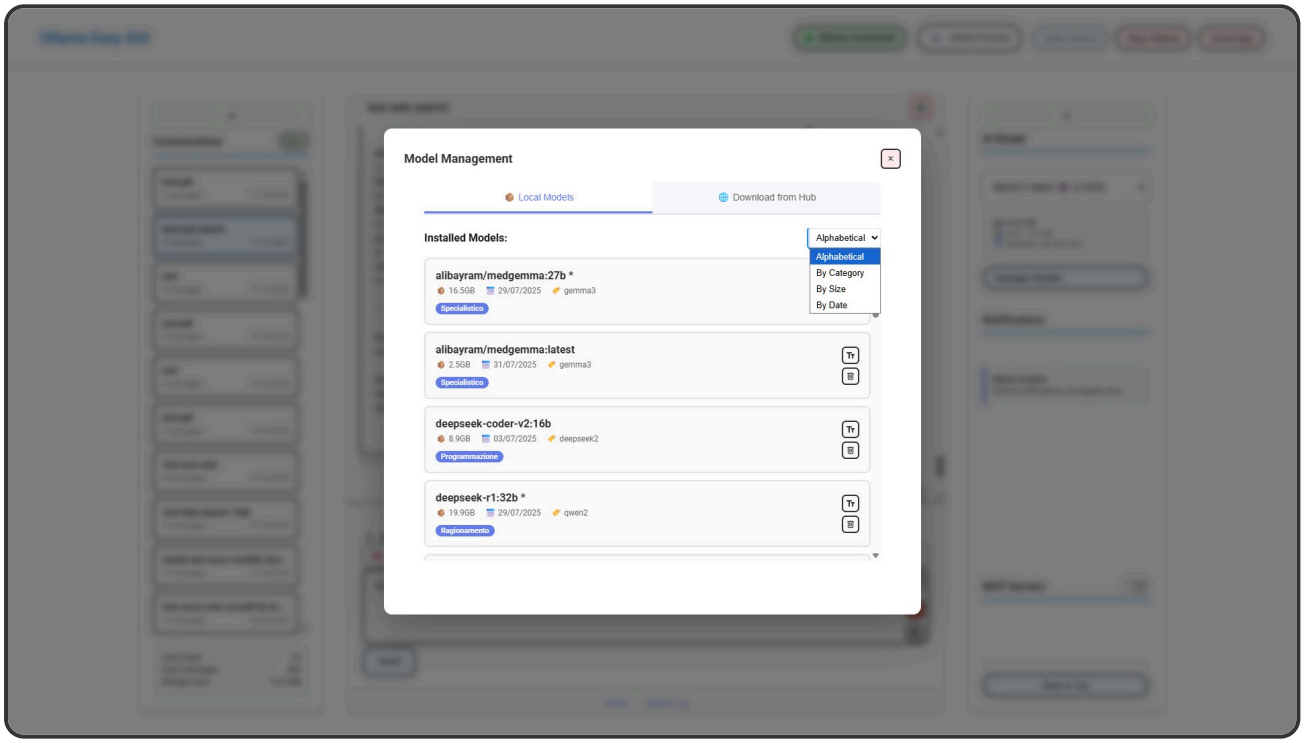
I modelli AI sono il cuore dell'applicazione. In questo capitolo vedremo come scaricare nuovi modelli, capire le differenze tra loro e scegliere quello più adatto alle tue esigenze.

## Modelli locali vs Hub

L'interfaccia di gestione modelli ha due schede:

- **Local Models**: modelli già scaricati sul tuo computer, pronti all'uso
- **Hub Search**: catalogo online dei modelli disponibili per il download

Per accedere alla gestione modelli, clicca il pulsante **Manage Models** nella sidebar destra.



(https://docs.ai-know.pro/ollama-easy-gui/img/modelli-locali.jpg)

## I modelli installati

Nella scheda dei modelli locali vedi tutti i modelli disponibili sul tuo computer. Per ogni modello sono mostrati:

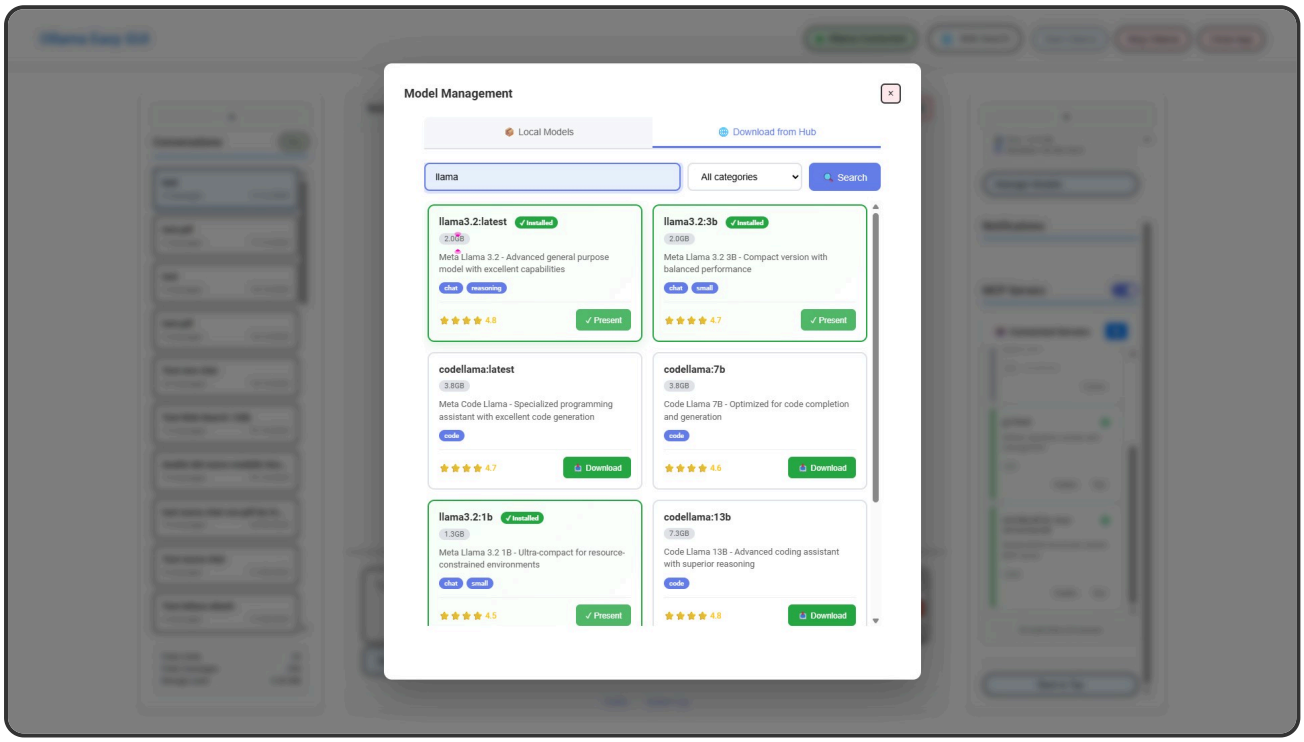
- **Nome e versione**: per esempio "llama3.2:3b"
- **Dimensione**: quanto spazio occupa su disco

- **Data**: quando è stato scaricato
- **Categoria**: Chat, Code, Reasoning, Multimodal

Puoi ordinare l'elenco per nome, dimensione, data o categoria usando i pulsanti in alto.

## Scaricare nuovi modelli

Passa alla scheda **Hub Search** per cercare modelli nel catalogo Ollama:

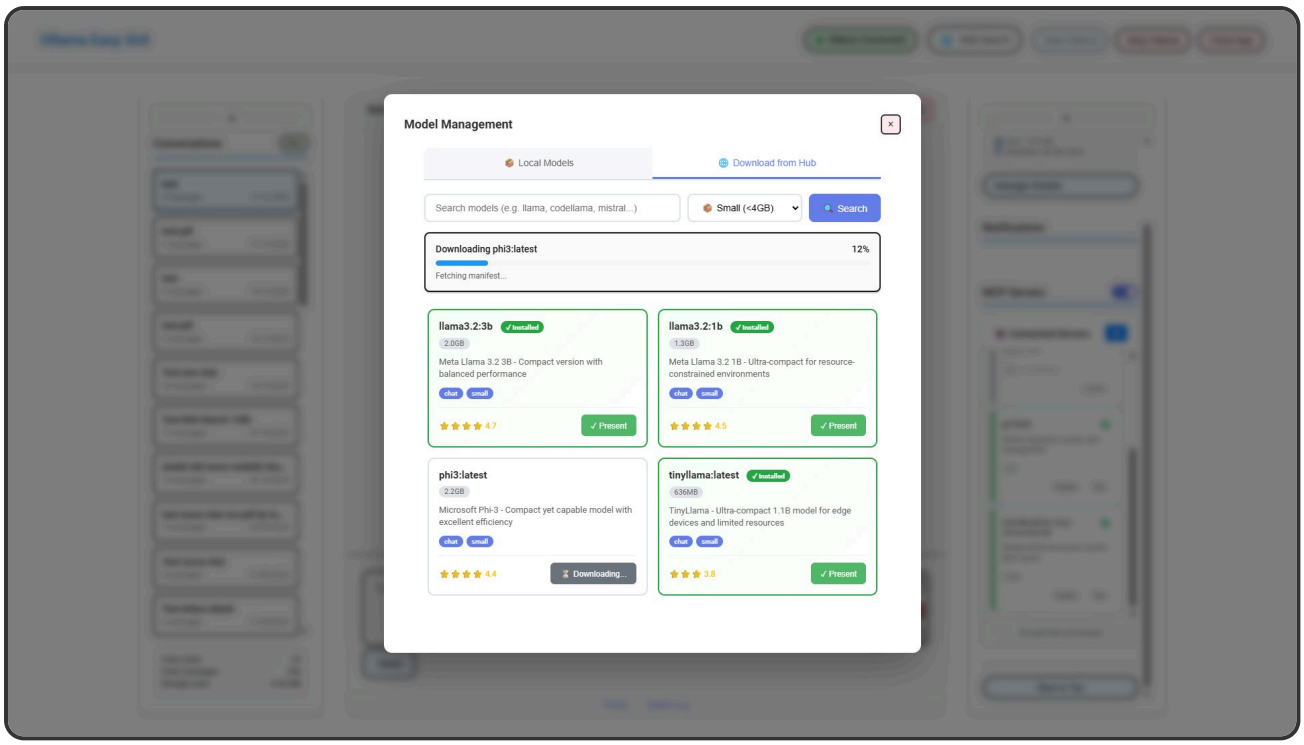


(https://docs.ai-know.pro/ollama-easy-gui/img/hub-modelli.jpg)

Puoi filtrare per categoria:

- **Chat**: modelli generici per conversazione
- **Code**: specializzati nella programmazione
- **Reasoning**: ottimizzati per ragionamento logico
- **Multimodal**: capaci di analizzare anche immagini

Trova un modello interessante e clicca su **Download**. Vedrai il progresso del download in tempo reale:



(https://docs.ai-know.pro/ollama-easy-gui/img/download-modello.jpg)

> **Prima di scaricare**
>
> I modelli possono essere molto pesanti: un modello da 70 miliardi di parametri richiede oltre 40 GB di spazio su disco.

Inoltre modelli grandi richiedono più RAM e una GPU con sufficiente memoria per funzionare a velocità accettabile. Consulta la sezione "Quale modello scegliere" per capire cosa può gestire il tuo hardware.

## Capire i nomi dei modelli

I nomi dei modelli seguono uno schema preciso:

```
nome:variante
```

Per esempio: `llama3.2:3b`, `qwen2.5:7b-instruct`, `codellama:13b`

Il numero dopo i due punti indica tipicamente la dimensione: - **1b-3b**: modelli leggeri, veloci, adatti a computer meno potenti - **7b-8b**: buon compromesso tra qualità e velocità - **13b-14b**: risposte più accurate, servono 16 GB di RAM - **70b e oltre**: massima qualità, richiedono hardware potente

## Quale modello scegliere

La scelta dipende dal tuo hardware e da cosa vuoi fare:

### Per computer con 8 GB di RAM

| Modello | Uso consigliato |
|---|---|
| llama3.2:3b | Conversazione generale, veloce |
| qwen2.5:3b | Buono per testi in più lingue |
| phi3:3.8b | Ragionamento e logica |

### Per computer con 16 GB di RAM

| Modello | Uso consigliato |
|---|---|
| llama3.1:8b | Uso generale, ottime risposte |
| qwen2.5:7b | Multilingue, anche italiano |
| mistral:7b | Veloce e affidabile |
| codellama:7b | Programmazione |

### Per computer con GPU (8+ GB VRAM)

| Modello | Uso consigliato |
|---|---|
| llama3.3:70b | Massima qualità |
| qwen2.5-coder:32b | Programmazione avanzata |
| command-r:35b | Ricerca e analisi documenti |

**Inizia in piccolo**

Se non sai quale scegliere, inizia con llama3.2:3b per testare che tutto funzioni, poi passa a modelli più grandi se il tuo hardware lo permette.

## Rimuovere un modello

Per liberare spazio su disco, puoi rimuovere i modelli che non usi più. Nella scheda dei modelli locali, clicca l'icona del cestino accanto al modello da eliminare.

> **Riscaricare è sempre possibile**
>
> Rimuovere un modello non è irreversibile, potrai sempre riscaricarlo dall'Hub se ne dovessi avere bisogno.

## Modelli e MCP

Non tutti i modelli supportano MCP (Model Context Protocol), la funzionalità che permette all'AI di usare strumenti esterni. Se prevedi di usare MCP, scegli modelli che supportano il "function calling":

- llama3.1, llama3.2, llama3.3
- qwen2.5 (tutte le varianti)
- mistral, mistral-nemo
- command-r, command-r-plus

I modelli più vecchi come llama2 o codellama non supportano MCP. Approfondiremo questo argomento nel capitolo dedicato.

# Personalizzare le risposte

I modelli AI rispondono in base alle loro caratteristiche predefinite, ma puoi modificare il loro comportamento usando due strumenti: il base prompt (per singolo modello) e il global system prompt (per tutti i modelli).

## Cos'è un system prompt

Un system prompt è un'istruzione che viene inviata al modello prima di ogni tua domanda. Il modello la considera come un contesto permanente che guida le sue risposte.

Per esempio, se imposti come system prompt:

```
Rispondi sempre in italiano, anche se la domanda è in un'altra lingua.
Usa un tono professionale ma accessibile.
Limita le risposte a massimo 200 parole.
```

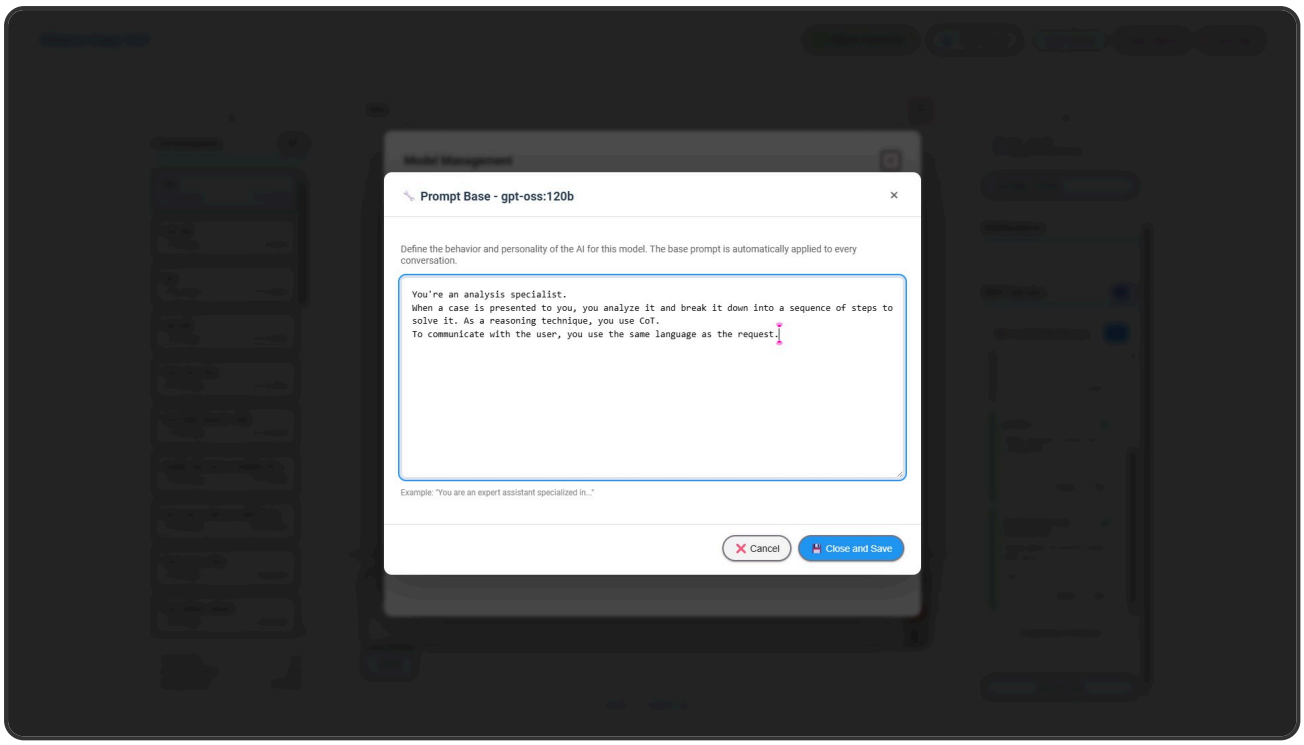
Il modello seguirà queste istruzioni in ogni conversazione.

## Base prompt: personalità per modello

Il base prompt è specifico per ogni modello. Ti permette di dare una "personalità" diversa a ciascuno.

Per impostarlo:

1. Apri la gestione modelli (**Manage Models**)
2. Trova il modello che vuoi personalizzare
3. Clicca l'icona del prompt (l'icona con le due T)
4. Si apre la finestra di inserimento del base prompt
5. Una volta inserito il testo, per salvarlo e uscire occorre cliccare si **Close and Save**, se invece non si vogliono salvare le modifiche basta cliccare sul tasto **Cancel**
6. Per annullare un base prompt personalizzato occorre aprire il base prompt (passi 1-2-3-4), cancellare il testo del base prompt ed uscire salvando



(https://docs.ai-know.pro/ollama-easy-gui/img/base-prompt.jpg)

## Esempi di base prompt utili

**Per un assistente di scrittura:**

```
Sei un editor esperto. Quando ti viene chiesto di revisionare un testo:
- Correggi errori grammaticali e di punteggiatura
- Suggerisci miglioramenti stilistici
- Mantieni il tono originale dell'autore
- Spiega brevemente le modifiche proposte
```

**Per un tutor di programmazione:**

```
Sei un tutor di programmazione paziente. Quando spieghi codice:
- Usa esempi semplici e concreti
- Spiega ogni passaggio, non dare nulla per scontato
- Se l'utente fa errori, correggili con gentilezza
- Suggerisci sempre le best practice
```

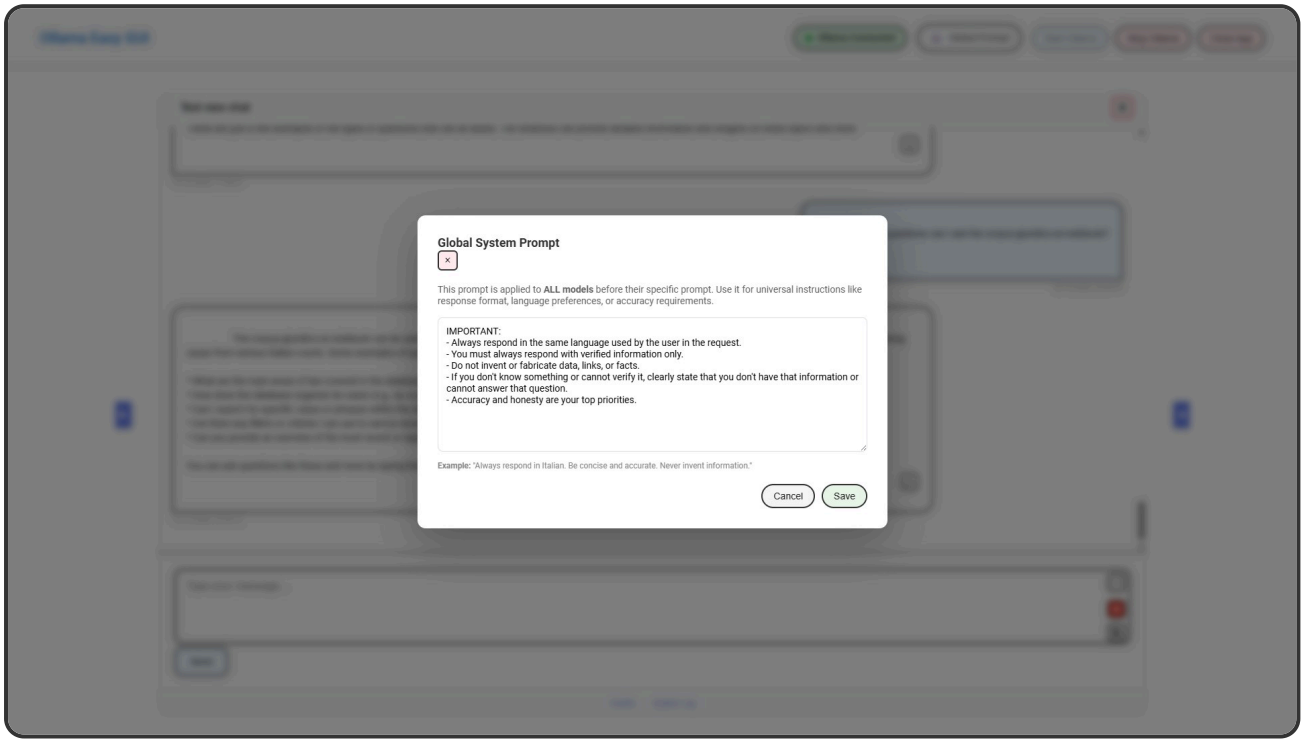
**Per analisi di documenti:**

```
Quando analizzi documenti:
- Estrai i punti chiave in forma di elenco
- Identifica eventuali contraddizioni o lacune
- Mantieni un approccio oggettivo
- Cita sempre le parti rilevanti del documento
```

## Global system prompt: istruzioni universali

Il global system prompt si applica a tutti i modelli, prima del loro base prompt specifico. È utile per impostare preferenze generali che vuoi mantenere sempre.

Per accedervi, clicca il pulsante **Global Prompt** nella barra superiore dell'applicazione, la gestione è simile a quella del base prompt.



(https://docs.ai-know.pro/ollama-easy-gui/img/global-prompt.jpg)

### Quando usare il global prompt

Il global prompt è ideale per:

- **Preferenze linguistiche**: "Rispondi sempre in italiano"
- **Formato delle risposte**: "Usa elenchi puntati quando possibile"
- **Regole di accuratezza**: "Se non sei sicuro di qualcosa, dillo chiaramente"
- **Stile di comunicazione**: "Evita introduzioni lunghe, vai dritto al punto"

### Esempio pratico

Un global prompt efficace potrebbe essere:

```
Regole generali:
1. Rispondi nella stessa lingua della domanda
2. Sii conciso: preferisci risposte brevi ma complete
3. Se non sai qualcosa, ammettilo invece di inventare
4. Quando fornisci istruzioni, numerale in passaggi chiari

Formato:
- Usa grassetto per i termini importanti
- Usa elenchi per più di 3 elementi
- Includi esempi quando aiutano la comprensione
```

## Come funzionano insieme

Quando invii un messaggio, il modello riceve le istruzioni in questo ordine:

1. **Global system prompt** (se impostato)
2. **Base prompt del modello** (se impostato)
3. **Il tuo messaggio**

Il modello considera tutto questo contesto per formulare la risposta. Se le istruzioni sono in conflitto, le ultime applicate (il base prompt) tendono a prevalere.

## Buone pratiche

Per ottenere risultati migliori:

- **Sii specifico**: istruzioni vaghe producono risultati vaghi
- **Testa le modifiche**: dopo aver cambiato un prompt, fai qualche domanda di prova
- **Non esagerare**: prompt troppo lunghi possono confondere il modello
- **Itera**: affina i prompt nel tempo in base ai risultati

# Allegati e export

Ollama Easy GUI permette di allegare documenti alle conversazioni e di esportare le risposte in vari formati. Vediamo come usare queste funzionalità.

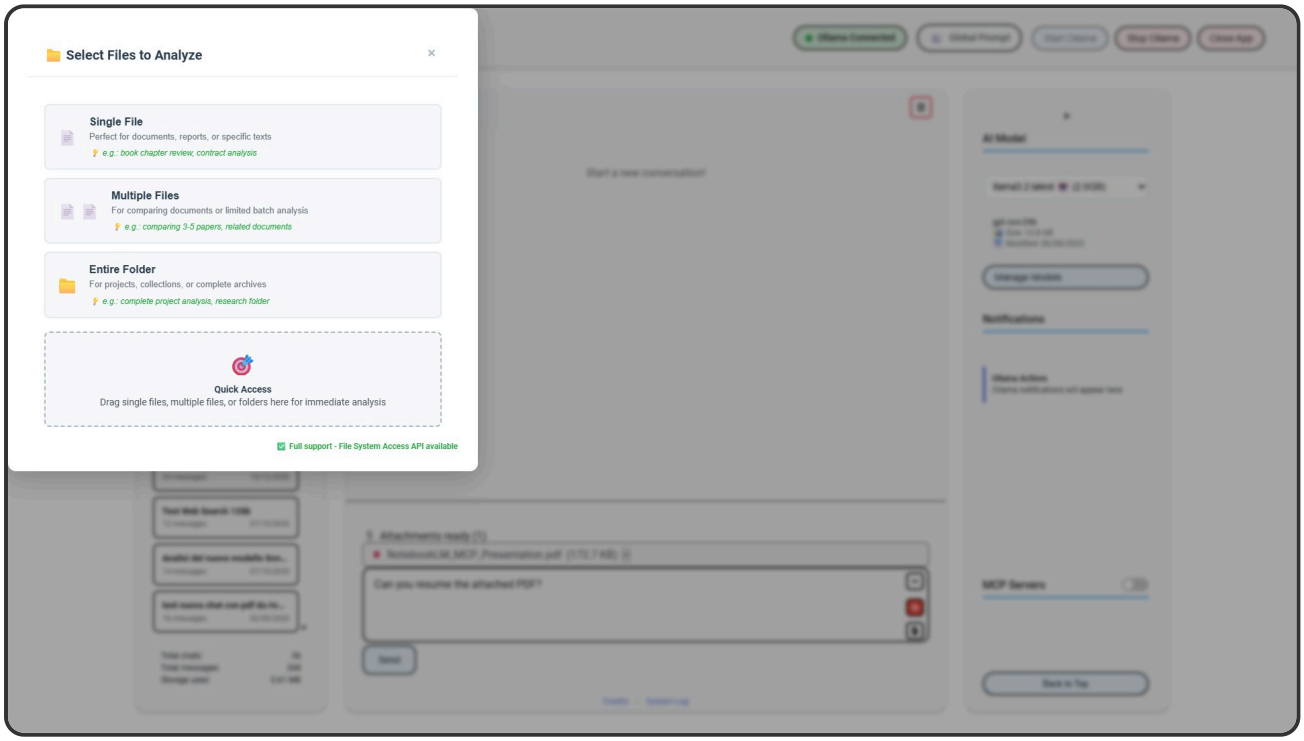
## Allegare file

Puoi allegare file per farli analizzare dal modello. I formati supportati sono:

- **Documenti**: PDF, Word (.docx), file di testo (.txt, .md)
- **Immagini**: JPG, PNG, GIF, WebP
- **Codice**: qualsiasi file di testo semplice

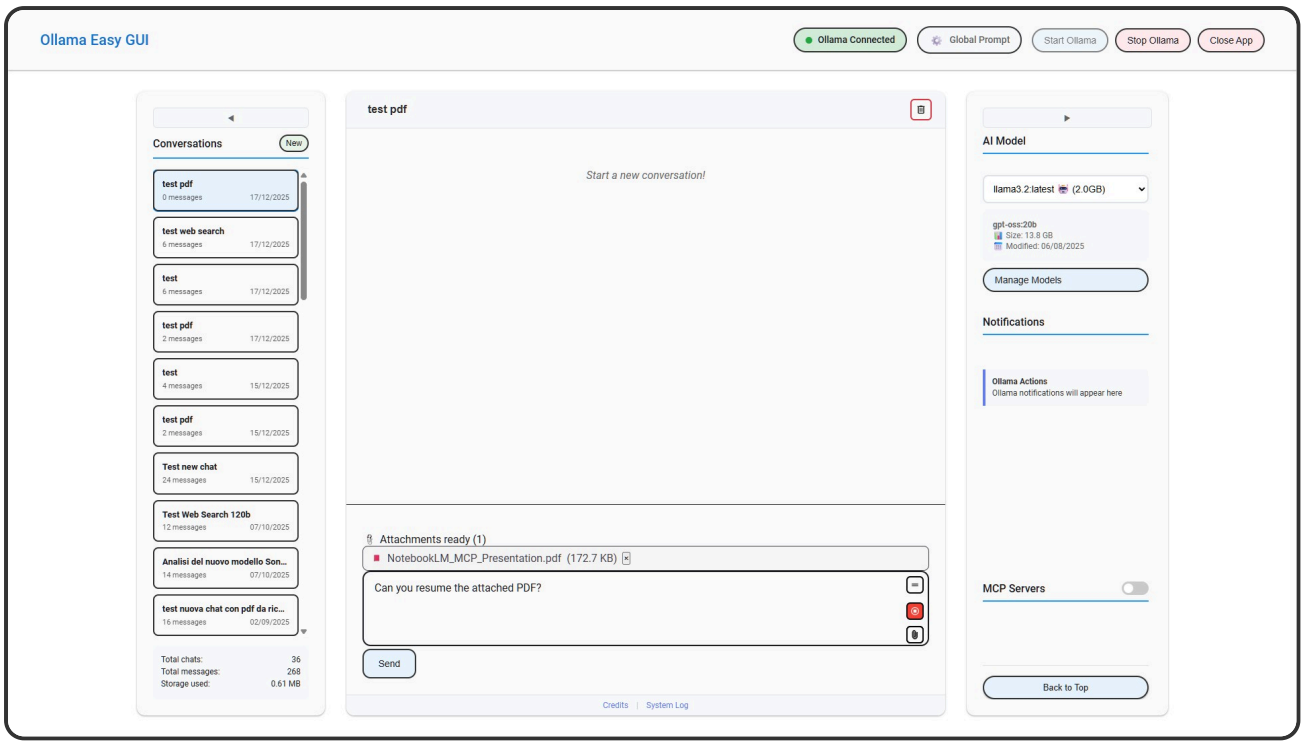
### Come allegare un file

Clicca l'icona della graffetta nell'area di input oppure trascina direttamente i file nella finestra:




(https://docs.ai-know.pro/ollama-easy-gui/img/allega-file.jpg)

Puoi selezionare uno o più file, oppure un'intera cartella. I file selezionati appariranno sotto l'area di testo:



(https://docs.ai-know.pro/ollama-easy-gui/img/file-allegato.jpg)

Scrivi la tua domanda e invia. Il modello riceverà sia il tuo messaggio che il contenuto dei file allegati.

## Esempi di utilizzo

**Analisi di un documento:**

> Allega un PDF di un contratto e chiedi: "Riassumi le clausole principali di questo contratto e segnala eventuali punti critici"

**Revisione di codice:**

> Allega un file di codice e chiedi: "Controlla questo codice Python e suggerisci miglioramenti"

**Descrizione di un'immagine:**

> Allega un'immagine e chiedi: "Descrivi cosa vedi in questa immagine" (richiede un modello multimodale)

**Comportamento linguistico**

Quando alleghi documenti, il modello tende a rispondere nella lingua del documento, anche se la tua domanda è in un'altra lingua. Per forzare una lingua specifica, indicalo esplicitamente nella domanda o nel system prompt.

## Limiti degli allegati

Ci sono alcune limitazioni da considerare:

- **Dimensione massima**: 50 MB per file
- **Immagini**: richiedono modelli multimodali (come llava o bakllava)
- **PDF complessi**: tabelle e layout elaborati potrebbero non essere interpretati perfettamente

Per documenti molto lunghi, considera di spezzare l'analisi in parti: chiedi prima un riassunto generale, poi approfondisci sezioni specifiche.

## Esportare le conversazioni

Puoi salvare singoli messaggi o intere conversazioni in diversi formati.

### Esportare un singolo messaggio

Passa il mouse su una risposta del modello e clicca l'icona di download. Scegli il formato:

- **Markdown** (.md): conserva la formattazione, ideale per documentazione
- **Testo** (.txt): formato universale, leggibile ovunque
- **Word** (.docx): per documenti formali o da condividere

### Esportare un'intera conversazione

Per esportare tutta la conversazione, usa il menu della conversazione nella sidebar sinistra. Troverai le stesse opzioni di formato.

## Dove vengono salvati i dati

Tutti i tuoi dati restano sul computer locale:

| Tipo di dato | Posizione |
|---|---|
| Conversazioni | `app/data/conversations/` |
| Configurazioni | `app/data/` |
| File esportati | Cartella download del browser |

I file delle conversazioni sono semplici JSON leggibili. Puoi farne backup copiando l'intera cartella `app/data`.

## Formati di export a confronto

| Formato | Pro | Contro |
|---|---|---|
| **Markdown** | Mantiene formattazione, codice evidenziato, intestazioni | Richiede un editor che lo supporti |
| **Testo** | Universalmente leggibile, leggero | Perde la formattazione |
| **Word** | Facile da condividere, modificabile | File più pesante |

**Per documentazione tecnica**

Se esporti conversazioni che contengono codice, il formato Markdown è la scelta migliore: mantiene l'evidenziazione della sintassi e la formattazione dei blocchi di codice.

# MCP: estendere l'AI

MCP (Model Context Protocol) è una funzionalità avanzata che permette ai modelli AI di usare strumenti esterni, come leggere file dal tuo computer o accedere a servizi online. In questo capitolo spieghiamo cos'è, quando usarlo e come configurarlo.

## Cos'è MCP in parole semplici

Normalmente, un modello AI può solo leggere quello che gli scrivi e rispondere. Non può "fare" nulla nel mondo reale: non può aprire file, non può cercare su internet, non può controllare cosa c'è in una cartella.

MCP cambia questo: permette di collegare al modello degli "strumenti" che può usare durante la conversazione. Per esempio:

- **Strumento filesystem**: il modello può leggere e scrivere file sul tuo computer nelle cartelle abilitate
- **Strumento GitHub**: il modello può cercare informazioni nei repository
- **Strumenti personalizzati**: puoi aggiungerne altri a seconda delle tue esigenze, per connetterti ad applicativi sul tuo pc oppure per connetterti a service in rete

Quando MCP è attivo e chiedi al modello "**Usando filesystem leggi il file relazione.txt nella cartella X**", il modello:

1. Se è attivo capisce che deve usare lo strumento filesystem
2. Chiama lo strumento con il percorso del file
3. Riceve il contenuto del file
4. Lo usa per rispondere alla tua domanda

Tutto questo avviene automaticamente, senza che tu debba fare nulla di diverso.

## Modelli compatibili

Non tutti i modelli supportano MCP. Serve una capacità chiamata "function calling" che permette al modello di generare chiamate strutturate agli strumenti.

### Modelli che funzionano con MCP

| Modello | Qualità MCP | Note |
|---|---|---|
| llama3.1, llama3.2, llama3.3 | Eccellente | Supporto nativo |
| qwen2.5 (tutte le varianti) | Eccellente | Ottimo anche per italiano |
| mistral, mistral-nemo | Buona | Veloce e affidabile |
| command-r, command-r-plus | Eccellente | Ottimo per ricerche |

### Modelli che NON funzionano con MCP

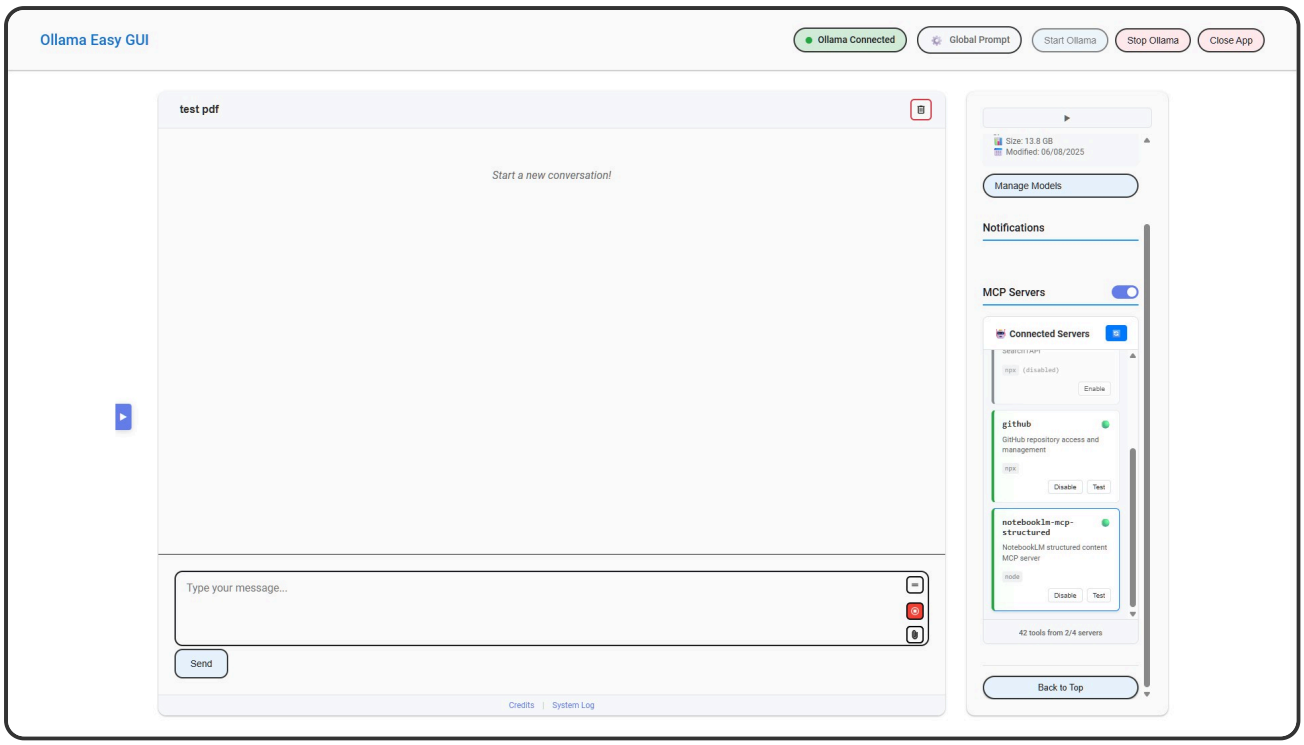
- llama2 (tutte le versioni)
- codellama
- phi, phi2
- gemma (versione 1)

Se usi un modello non compatibile con MCP attivo, semplicemente ignorerà gli strumenti disponibili.

## Attivare MCP

MCP è disattivato di default. Per attivarlo:

1. Nella sidebar destra, trova l'interruttore **MCP**
2. Attivalo (diventa blu)
3. Clicca sull'icona delle impostazioni MCP per vedere gli strumenti disponibili



(https://docs.ai-know.pro/ollama-easy-gui/img/mcp-pannello.jpg)

Gli strumenti configurati appaiono nell'elenco. Attiva quelli che vuoi rendere disponibili al modello.

## Configurare gli strumenti

La configurazione degli strumenti MCP avviene nel file `app/data/mcp-config.json`. La prima volta devi crearlo copiando l'esempio:

```
copy app\data\mcp-config.json.example app\data\mcp-config.json
```

Poi modificalo con un editor di testo:

```
{
  "mcpServers": {
    "filesystem": {
      "command": "npx",
      "args": ["-y", "@modelcontextprotocol/server-filesystem", "D:/Documenti"],
      "enabled": false,
      "description": "Accesso ai file"
    },
    "github": {
      "command": "npx",
      "args": ["-y", "@modelcontextprotocol/server-github"],
      "env": { "GITHUB_TOKEN": "il-tuo-token-qui" },
      "enabled": false,
      "description": "Accesso a GitHub"
    }
  }
}
```

### Strumento filesystem

Permette al modello di leggere e scrivere file. Devi specificare quali cartelle può accedere:

```
"args": ["-y", "@modelcontextprotocol/server-filesystem", "D:/Documenti", "D:/Progetti"]
```

> **Sicurezza**
>
> Limita l'accesso solo alle cartelle necessarie. Non dare accesso all'intero disco.

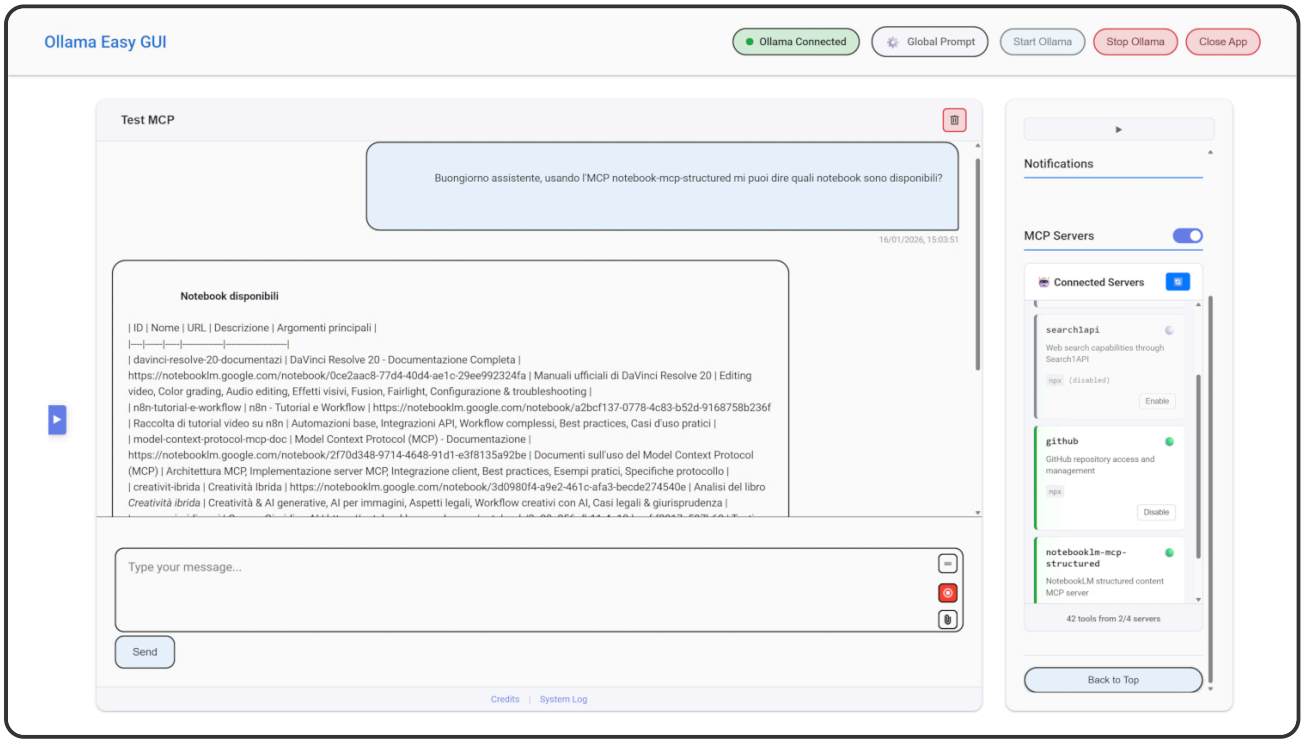
### Strumento GitHub

Permette al modello di cercare nei repository GitHub. Richiede un token di accesso personale:

1. Vai su GitHub → Settings → Developer settings → Personal access tokens
2. Crea un nuovo token con i permessi necessari
3. Inseriscilo nel file di configurazione

## MCP in azione

Una volta attivati i server MCP si può fare la richiesta, meglio se citando il server MCP che si vuole utilizzare




(https://docs.ai-know.pro/ollama-easy-gui/img/mcp-query.jpg)

## Repository

Inizia ad esserci un'ampia disponibilità di server MCP e di repository. Uno di questi è github.com/modelcontextprotocol/servers (https://github.com/modelcontextprotocol/servers)

## Privacy e MCP

I server MCP possono funzionare anche localmente sul tuo computer, tuttavia anche in questi casi alcuni MCP/strumenti potrebbero connettersi a servizi esterni:

- **filesystem**: completamente locale, nessun dato esce
- **github**: si connette ai server GitHub

Prima di usare un MCP/strumento, considera quali dati potrebbe trasmettere.

## Risoluzione problemi MCP

**"Il modello non usa gli strumenti"** - Verifica che MCP sia attivo (interruttore blu) - Verifica che lo strumento specifico sia attivato nel pannello - Prova un modello diverso (potrebbe non supportare function calling)

**"Errore durante l'esecuzione dello strumento"** - Controlla i log dell'applicazione per dettagli - Verifica che i percorsi nel file di configurazione siano corretti - Per GitHub, verifica che il token sia valido

**"Lo strumento è lento"** - È normale per operazioni su molti file - I modelli più piccoli potrebbero impiegare più tempo a generare le chiamate corrette

# Risoluzione problemi

Questa sezione raccoglie i problemi più comuni e le relative soluzioni. Se non trovi risposta qui, puoi chiedere aiuto nei canali della comunità.

## Problemi di installazione

### "npm" non è riconosciuto come comando

**Causa**: Node.js non è installato o il terminale non è stato riavviato dopo l'installazione.

**Soluzione**: 1. Verifica che Node.js sia installato: cerca "Node.js" nel menu Start 2. Chiudi e riapri il Prompt dei comandi 3. Se il problema persiste, reinstalla Node.js

### "git" non è riconosciuto come comando

**Causa**: Git non è installato o non è nel PATH di sistema.

**Soluzione**: 1. Reinstalla Git da git-scm.com (https://git-scm.com) 2. Durante l'installazione, assicurati che l'opzione "Git from the command line" sia selezionata 3. Riavvia il computer

### Errore durante npm install

**Causa**: problemi di rete o permessi.

**Soluzione**: 1. Prova a eseguire il Prompt dei comandi come Amministratore 2. Se sei dietro un proxy aziendale, configura npm:

```
npm config set proxy http://proxy.azienda.it:8080
```

3. Prova a cancellare la cache e riprovare:

```
npm cache clean --force
npm install
```

## Problemi di avvio

### L'applicazione non si avvia

**Causa possibile 1**: Ollama non è in esecuzione.

**Verifica**: cerca l'icona di Ollama nell'area di notifica (vicino all'orologio). Se non c'è: 1. Cerca "Ollama" nel menu Start e avvialo 2. Attendi qualche secondo che si avvii completamente 3. Riprova ad avviare Ollama Easy GUI

**Causa possibile 2**: la porta 3003 è già in uso.

**Soluzione**: chiudi altre istanze dell'applicazione.

### Pagina bianca nel browser

**Causa**: il server è avviato ma c'è un errore nel frontend.

**Soluzione**: 1. Apri gli strumenti sviluppatore del browser (F12) 2. Controlla la tab "Console" per messaggi di errore 3. Prova a svuotare la cache del browser e ricaricare

## Problemi con i modelli

### Nessun modello disponibile nella lista

**Causa**: Ollama non ha modelli installati o non è raggiungibile.

**Soluzione**: 1. Verifica che Ollama sia in esecuzione 2. Scarica almeno un modello:

```
ollama pull llama3.2
```

3. Ricarica l'interfaccia web

### Il modello risponde molto lentamente

**Causa**: risorse hardware insufficienti o modello troppo grande.

**Soluzioni**: - Prova un modello più piccolo (es. passa da 8b a 3b) - Chiudi altre applicazioni per liberare RAM - Se hai una GPU, verifica che Ollama la stia usando - Aumenta `OLLAMA_NUM_THREADS` nel file `.bat`

### Il modello non capisce l'italiano

**Causa**: alcuni modelli sono addestrati principalmente su testi inglesi.

**Soluzioni**: - Usa modelli multilingue come `qwen2.5` o `mistral` - Aggiungi nel system prompt: "Rispondi sempre in italiano" - Considera che modelli piccoli (<3b) possono avere capacità limitate in lingue diverse dall'inglese

# Problemi con gli allegati

## Il PDF non viene letto correttamente

**Causa**: PDF con formattazione complessa o scansionato.

**Soluzioni**: - Se il PDF è una scansione, il testo non è estraibile: usa un OCR prima - Per PDF con molte tabelle, considera di estrarre il testo manualmente - Prova ad allegare una versione più semplice del documento

## Le immagini non vengono analizzate

**Causa**: il modello non è multimodale.

**Soluzione**: usa un modello che supporta le immagini: - `llava` - `bakllava` - `llama3.2-vision`

# Problemi con MCP

## Gli strumenti MCP non funzionano

**Checklist**: 1. MCP è attivato? (interruttore nella sidebar) 2. Lo strumento specifico è attivato nel pannello MCP? 3. Il modello supporta function calling? (usa llama3.1, qwen2.5, mistral) 4. Il file mcp-config.json esiste ed è configurato correttamente?

## Errore "strumento non trovato"

**Causa**: il server MCP non si è avviato correttamente.

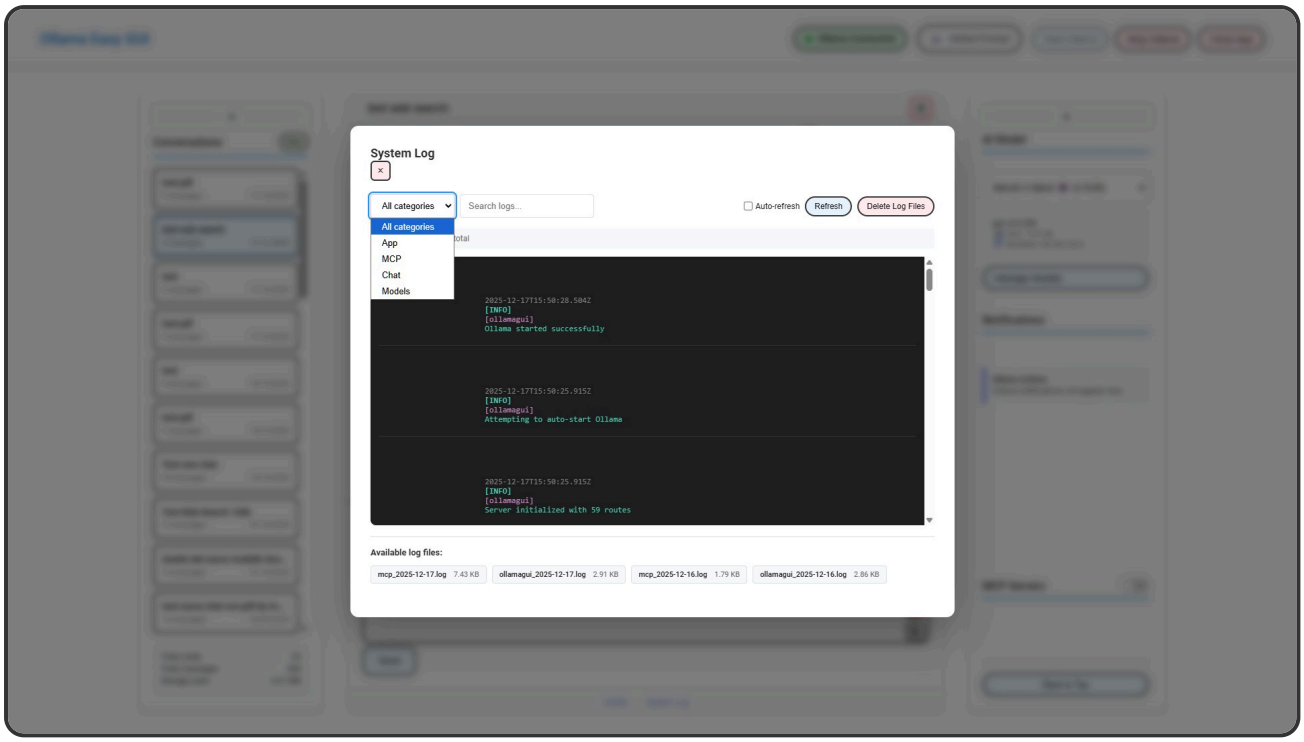
**Soluzione**: 1. Verifica che Node.js sia installato (i server MCP lo richiedono) 2. Controlla i log per errori specifici 3. Prova a riavviare l'applicazione

# Consultare i log

L'applicazione tiene traccia degli errori nei file di log. Per visualizzarli:

1. Clicca il pulsante **Log** nel footer dell'interfaccia
2. Seleziona la categoria da visualizzare:
3. **App**: errori generali dell'applicazione
4. **Chat**: problemi durante le conversazioni
5. **MCP**: errori degli strumenti esterni

6. **Models**: problemi con download o caricamento modelli



(https://docs.ai-know.pro/ollama-easy-gui/img/log-viewer.jpg)

Puoi cercare testo specifico e filtrare per data.

## Ottenere supporto

Se il problema persiste:

- **GitHub Issues**: github.com/paolodalprato/ollama-easy-gui/issues (https://github.com/paolodalprato/ollama-easy-gui/issues) per segnalare bug o chiedere aiuto
- **GitHub Discussions**: per domande generali e discussioni

Quando chiedi aiuto, includi: - Descrizione del problema - Sistema operativo e versione - Messaggi di errore (dai log o dalla console) - Passaggi per riprodurre il problema

# Note per macOS e Linux

Questa appendice raccoglie le differenze principali per chi vuole usare Ollama Easy GUI su sistemi diversi da Windows.

> ⚠ **Piattaforme non testate**
>
> L'applicazione è stata sviluppata e testata solo su Windows. Su macOS e Linux dovrebbe funzionare, ma potrebbero esserci problemi non documentati. Feedback e segnalazioni sono benvenuti sul repository GitHub (https://github.com/paolodalprato/ollama-easy-gui).

## Installazione dei prerequisiti

### Ollama

Su macOS e Linux, Ollama si installa da terminale:

```
curl -fsSL https://ollama.com/install.sh | sh
```

Su macOS è disponibile anche un'applicazione scaricabile dal sito ufficiale.

### Node.js

Su macOS puoi usare Homebrew:

```
brew install node
```

Su Linux (Ubuntu/Debian):

```
sudo apt update
sudo apt install nodejs npm
```

### Git

Su macOS:

```
brew install git
```

Su Linux (Ubuntu/Debian):

```
sudo apt install git
```

# Differenze operative

## File .bat

Il file `start-ollama-easy-gui.bat` è specifico per Windows e non funziona su altri sistemi. Su macOS e Linux avvia l'applicazione con:

```
cd percorso/ollama-easy-gui
npm start
```

Se vuoi configurare le variabili d'ambiente per la GPU (come fa il file .bat su Windows), esportale prima di avviare:

```
export OLLAMA_GPU_LAYERS=18
export OLLAMA_FLASH_ATTENTION=1
export OLLAMA_NUM_THREADS=8
npm start
```

## Percorsi delle cartelle

Nel manuale si fa riferimento a percorsi Windows come `%USERPROFILE%\Documents`. Gli equivalenti sono:

| Windows | macOS/Linux |
|---|---|
| `%USERPROFILE%\Documents` | `~/Documents` |
| `app\data\conversations` | `app/data/conversations` |

## Prompt dei comandi

Dove il manuale dice "Prompt dei comandi", su macOS usa Terminale e su Linux usa il terminale della tua distribuzione.

# Supporto GPU

Su Linux, il supporto GPU NVIDIA richiede l'installazione dei driver CUDA. Su macOS con chip Apple Silicon (M1, M2, M3), Ollama utilizza automaticamente la GPU integrata.