CFAI Reference Guide v3.1

Cognitive FlowAI — Version 3.1 | English / Italiano

ENGLISH

Why CFAI?

CFAI is not another AI chatbot. It is a real-time cognitive assistant designed to support you during complex mental processes — while you speak, write, present, or meet — without replacing your critical thinking.

Imagine an assistant that follows you as you act, helping you stay focused, capture what matters, and act on it in real time. It does not do everything for you: it works alongside you.

Who Can Benefit?

- Developers during technical discussions, design sessions, or brainstorming
- Job candidates preparing for interviews or complex technical assessments
- Recruiters evaluating profiles in real time
- Sales professionals during calls, client demos, and consulting sessions
- Content creators preparing pitches, scripts, and presentations
- Vocal coaches, trainers, educators
- Anyone working with meetings, documents, or complex workflows

Subscription Plans

CFAI uses a credit-based system. One credit covers approximately 1,000 AI tokens (≈ a short message exchange). Local features that do not call cloud AI – such as Whisper voice transcription — do not consume credits.

Plan	Price	Credits	Best for
Trial	Free	Limited free usage	Trying out the app, no card required
Starter	€7.99 / month	500 credits	Light usage, occasional AI requests
Popular	€12.99 / month	1500 credits	Regular daily use, ongoing workflows
Best Value	€24.99 / month	3000 credits	Power users, heavy daily use

All paid plans auto-renew monthly. You can change or cancel anytime from your account page.

What the Trial includes

- Single AI chat session
- Microphone input for voice transcription
- General and Advanced Settings
- Two Whisper transcription models (tiny + base)
- Snipping Tool
- Image attachment

Paid plans unlock the full feature set described in this guide.

How credits are consumed

Action	Approximate cost
AI Chat (1 message)	~0.5-3 credits
Voice transcription (local Whisper)	Free – fully local, no credits
Screenshot + Prompt	~1-3 credits
In-Place Replace (text transformation)	~0.5-2 credits
Cognitive Map generation	~1-5 credits
Web Search	~1-3 credits per request
Meeting Mode (situational analysis)	~1-3 credits per analysis (transcription is free)
Document RAG (chat with a document)	~2-8 credits per message
Agentic AI loop	~1-3 credits per step

Actual cost depends on prompt and response length. Credits accumulate by token usage and are deducted only once a full credit's worth has been consumed. You will not be charged in fractions.

What CFAI Does — The Feature Set

CFAI offers four main modes and a wide set ofsupporting tools.

Main Modes

1. **AI Chat (with TTS)** – voice-driven conversational AI
2. **Free Write (Voice Transcription)** – voice dictation into a rich text editor
3. **Screenshot + Prompt** – capture your screen and ask the AI
4. **Cognitive Predictive Map** – real-time semantic compass while you speak

Supporting Features

- Dual Audio Capture – record your voice and system audio at the same time
- CRC (Clipboard Response Copy) – auto-copyAI replies for use in other apps
- Real-Time Translation – translate spoken text on the fly
- Local LLM Support – connect to your own AI server (Ollama, LM Studio, Jan.ai)
- Document RAG – chat with attached documents
- Speaker Diarization – label who is speaking in meetings
- 101 Languages Supported – Whisper-powered multilingual transcription
- Local OCR – read text from images or screen captures
- Agentic AI – multi-step planning and tool use
- Lecture Mode – AI-generated lecture scripts with TTS narration
- Meeting Mode – speaker-aware live meeting assistance
- Multi-Session – multiple parallel chat sessions
- Context Orientation – automatic awareness ofopen windows and clipboard
- Semantic Cache + Response Cache – avoid repeating identical AI calls
- In-Place Replace (IPR) – transform any selected text in any app
- Web Search – live search integrated into chat
- Conversation Context – persistent memory across messages with 4 distinct configuration modes
- Quick Launch – full-screen launcher accessible from anywhere
- Always on Top – keep CFAI floating over other windows
- Tray Icon – silent background presence in the Windows tray
- Suggested Counter-Question – AI suggests follow-up questions automatically
- Auto-Apply Interval – automatic re-application ofcontext at configurable intervals
- Offline TTS (Piper) – neural text-to-speech fully local

Modes — Detailed Reference

AI Chat (Voice Chatbot)

The main conversational interface. Works like ChatGPT or Claude, with full voice integration.

- Speak or type your prompt
- Replies are displayed with rich formatting (code blocks, diagrams, charts, tables)
- Optional Text-to-Speech reads the answer aloud
- Attach images or files for context
- Streaming responses – see the answer as it is generated
- Stop, retry, or expand any reply
- Lock the context to stay on a specific topic

- Auto-copy reply to clipboard (CRC) for use in other apps

Multi-Session: Create multiple parallel chats. Each session keeps its own history and can have its own attached document.

Free Write (Voice Transcription)

A rich text editor with built-in voice dictation, similar to a lightweight Word processor.

- Full editorial toolbar: Bold, Italic, Strikethrough, Code, H1-H6, lists, quotes, undo/redo
- Voice dictation – speak and your words appear at the cursor
- Real-time translation ofspoken text
- Optional AI correction oftranscribed text
- Automatic generation ofthe Cognitive Predictive Map from what you write
- Context binding: lock CFAI to a specific topic for the session

Screenshot + Prompt

Capture your screen and send the image with a written or spoken prompt to receive a targeted AI response.

- Full screen capture – grab the entire display instantly
- Snipping Tool – draw a rectangle to capture only a portion
- Attach additional files to the prompt
- Voice prompt: speak your question after capturing
- AI response is auto-copied to clipboard (CRC) for use anywhere

Cognitive Predictive Map

CFAI's most distinctive feature. Starting from a defined context, the system builds an interactive semantic map and tracks where you are cognitively as you speak.

How it works:

1. Define a context (e.g., "system design interview: distributed caching")
2. Start speaking – CFAI transcribes in real time
3. The map highlights the active concept node as you move through topics
4. The system calculates cognitive drift – how far you have moved from the main topic
5. Smart suggestions appear to help you stay on track or expand coherently

It works like a real-time semantic compass – designed for moments when clarity, focus, and impact matter most.

Core Features

Dual Audio Capture

CFAI can listen to two audio sources simultaneously:

- **Microphone** – your voice in real time
- **System audio** – anything playing on your computer: calls, videos, meetings

Both sources are transcribed and labeled separately, so CFAI captures both sides ofa conversation – you and your interlocutor – with no external bot, browser extension, or recording tool.

Ideal for:

- Live meetings on Zoom, Google Meet, Microsoft Teams
- Job interviews with real-time AI support
- Sales calls with full bidirectional context
- Watching tutorials or online courses with AI commentary

Note: System audio capture requires a loopback audio device (Stereo Mix, VoiceMeeter, VB-Audio Cable, etc.) enabled in Windows Sound settings.

Voice Transcription (Whisper, fully local)

CFAI uses OpenAI Whisper running 100% on your machine for voice recognition. Audio is never sent to any external service. Voice transcription does not consume credits.

All models are provided in **CT2 INT8** format (CTranslate2 quantized) for optimized performance.

Model	Size	Characteristics
Whisper Tiny (CT2 INT8)	77 MB	Lowest quality – very fast, low RAM
Whisper Base (CT2 INT8)	145 MB	Balanced – good default
Whisper Small (CT2 INT8)	480 MB	Better quality – noticeably slower on CPU
Whisper Medium (CT2 INT8)	1530 MB	High quality – GPU strongly recommended
Whisper Large-v3 (CT2 INT8)	3090 MB	Best quality – GPU required for real-time

Download the model that fits your hardware from **Advanced Settings → Transcription**.

GPU Acceleration: NVIDIA only — requires CUDA libraries (~600 MB). A **Test GPU support** button is available to verify your setup before downloading the libraries.

CRC — Clipboard Response Copy

CRC is CFAI's bridge to any other application. Press the CRC hotkey and the AI response is automatically copied to your clipboard — ready to paste into ChatGPT, Claude, Notion, Word, email, Slack, VS Code, or anywhere else.

In-Place Replace (IPR)

Select any text in any application, press the IPR hotkey, and CFAI replaces it with an AI-processed version.

Default presets:

- **Correggi grammatica** — "correggi la grammatica e l'ortografia, restituisci solo il testo corretto"
- **Traduci in inglese** – "traduci il testo in inglese, restituisci solo il testo tradotto"
- **Rendi formale** – "riscrivi il testo in tono formale e professionale, restituisci solo il testo riscritto"

You can add, edit, and remove **fully custom presets** with arbitrary prompts. Each preset has a name + a complete prompt. Choose which one is active for the default hotkey, or use **IPR Quick Pick** to open a small picker and choose the preset on the fly.

Real-Time Translation

While you dictate or transcribe, CFAI can translate spoken text on the fly into your chosen output language – useful for meetings, interviews, and content creation across languages.

101 Languages Supported

Voice transcription works in over a hundred languages thanks to Whisper. Select your speaking language from the Settings panel.

Local OCR

CFAI can read text directly from screenshots or your screen, then feed that text into a prompt or chat. **100% local processing** — no images are uploaded.

Document RAG

Attach a document to a chat session and CFAI will use its contents to answer your questions. Supported formats: `.txt`, `.md`, `.pdf`, `.docx`, `.csv`, `.json`.

The document is processed locally — its contents never leave your machine. **Passage embeddings are computed locally** and the number ofindexed passages is displayed in the UI.

Meeting Mode

For browser-based meetings (Google Meet, Zoom Web, Teams Web), CFAI integrates with a companion browser extension to capture speaker-labeled audio. Each speaker's words appear separately in the transcript.

Speaker Diarization

In meetings or recordings with multiple voices, CFAI automatically labels each speaker, so transcripts read like a clean dialogue.

Lecture Mode

CFAI can generate a structured lecture from a document or topic description, broken into pages, with speaker notes and optional diagrams. TTS narration reads the lecture aloud, while you can navigate, pause, and even have CFAI transcribe what the student says alongside the lecture.

Agentic AI

For complex requests, CFAI runs a multi-step agentic loop: think → plan → use a tool → observe → continue. You see each step live in the chat. Stop the loop at any time.

Web Search

CFAI can search the web inside a chat. Ask for current information, news, prices, references – CFAI fetches and synthesizes the answer.

Conversation Context — Full Reference

The Conversation Context is one of CFAI's most powerful features. It is configurable via 4 distinct tabs:

Tab 1 — Instructions

A free-form text area to add a custom context or instruction that guides the entire conversation.

- Add custom prompts that shape everyAI response
- **Attached document support:** drop a file directly here for RAG within the conversation
- Supported formats: `.txt`, `.md`, `.pdf`, `.docx`, `.csv`, `.json`
- The document stays on your device – only passage embeddings are computed locally
- The UI shows the number ofindexed passages (e.g., "39 passages indexed")

Tab 2 — Guided

Pre-configured awareness modes for technical and structured contexts.

Collection Depth – three levels:

- **Instant** – minimal context, lowest latency
- **Moderate** – extended context (clipboard, tabs, files), ~500ms latency
- **Deep** – full system awareness, deepest analysis

What CFAI should focus on – selectable focus areas (checkboxes):

- **Code & Technical** – Focus on code, algorithms, technical details
- **UI / UX Structure** – Focus on interface, layout, user experience

- **Business Logic** – Focus on rules, workflows, domain logic
- **Documentation & Specs** – Focus on docs, specifications, requirements

Tab 3 — Active Contexts

Real-time visibility into all open applications and windows on your system. CFAI can include selected windows in the AI context.

- Columns shown: **APP**, **WINDOW** (title), **TYPE** (DOCS, OTHER, etc.)
- Automatic type recognition based on application and content
- Multi-select with **Select all** / **Clear** buttons
- Refresh button to update the window list in real time
- You decide which windows feed into the AI context – privacy by design

Tab 4 — Clipboard

Optional automatic clipboard history capture for inclusion in AI context.

- Enable/disable via Advanced Settings → Clipboard History
- Configurable: how many recent entries to send to the LLM (default 10)
- Status indicator: "Listening" / "Not listening"
- Only captures while active – no retroactive history sent

Conversation Memory (Mirror Memory)

Separate from Conversation Context. Conversation Memory persists messages within the current chat session.

- **Default state: DISABLED**
- When disabled: "Each message will be sent without previous context. The AI won't remember what you discussed earlier."
- When enabled: the last N messages are automatically included in every new request
- Number ofmessages to include is configurable
- Must be explicitly enabled – privacy-first default

Response Cache & Semantic Cache

CFAI maintains two distinct local caches:

Response Cache

Stores literal AI responses. Ifyou ask the **exact same question** twice, the cached answer is returned instantly.

Semantic Cache

Stores semantically similar AI responses. If you ask **a similar question** (different wording but same intent), the cached answer is returned.

Both are toggled independently in Advanced Settings → Cache.

Cache management actions:

- View entries
- Export cache
- Clear this chat
- Clear All

Both caches save credits and time.

Local LLM Support

Beyond CFAI's own AI, you can connect any OpenAI-compatible LLM server – local or cloud. Run models entirely on your own machine for zero per-token cost and full data privacy.

Compatible servers:

- Ollama → `http://localhost:11434/v1`
- LM Studio → `http://localhost:1234/v1`
- Jan.ai → `http://localhost:1337/v1`
- Any server implementing the OpenAI API spec

Connection status (CONNECTED / DISCONNECTED) is visible in App Preferences.

Offline Text-to-Speech (Piper)

CFAI uses **Piper**, an open-source neural TTS engine, running fully offline.

- Engine size: ~7 MB
- Voice model size: ~60 MB per language
- Languages available: Italian (Paola), English (Lessac), and more
- Install individual voices on demand from Advanced Settings → Text-to-Speech
- Test voices before committing to a download
- 100% offline after initial download

Quick Launch

Press the Quick Launch hotkey from anywhere on your desktop to open a full-screen menu where you can jump directly to AI Chat, Free Write, Screenshot + Prompt, or the Cognitive Map.

Always on Top

Keep CFAI floating above all other windows so it stays visible during meetings, presentations, or coding sessions. Window transparency is configurable from 0% to 100%.

Tray Icon

CFAI runs silently in the Windows system tray. Click the icon to show or hide the app at any time.

Hotkeys

All hotkeys are global — they work even when CFAI is not in focus (e.g., while you are in Zoom or VS Code).

Action	Default Hotkey	Status
CRC (Clipboard Response Copy)	Ctrl+Shift+Space	DEFAULT
Quick Launch Menu	CapsLock+Shift+Ctrl	DEFAULT
Snipping Tool	Ctrl+Alt+N	DEFAULT
In-Place Replace (IPR)	Ctrl+Shift+Alt+Space	Not configured by default
IPR Quick Pick	Ctrl+Shift+Alt+Q	Not configured by default
Capture Screen (full screen)	—	Not configured by default

All hotkeys can be remapped in **Advanced Settings → Keyboard Shortcuts**.

Physical Controller (Bluetooth)

CFAI supports hands-free operation via a Bluetooth button controller for real-time workflow management.

Command mapping:

Action	Command
1 short click	Toggle microphone on/off
2 quick clicks	Capture screenshot and open prompt modal
1 long press	Expand last AI response

How to connect:

1. Turn on the controller
2. Open Windows Bluetooth settings: Start → Bluetooth & other devices → Add device
3. Select the device (BT1818 or similar)
4. Turn the controller off
5. Launch CFAI and log in
6. Turn the controller back on
7. A green indicator in the Settings bar confirms the connection

The controller is sold separately.

Installation & First Launch

Step 1 — Install CFAI Run the installer. Windows will register CFAI in the Start Menu.

Step 2 — Create an account or log in On first launch, the login screen appears. Click **Sign up** to create a free Trial account, or **Log in** ifyou already have one.

Step 3 — Select your audio source From the Settings panel (gear icon), choose your microphone and/or system audio source.

Step 4 — Download a transcription model Open **Advanced Settings** → **Transcription** and download a Whisper model. The Base model is recommended for most users; faster machines can use larger models for better accuracy.

Step 5 — (Optional) Connect a local LLM Ifyou want to use your own AI model, configure it in Settings → API.

Step 6 — Start using CFAI Press **CapsLock+Shift+Ctrl** to open the Quick Launch Menu from anywhere on your desktop, then pick a mode.

Settings — Reference

Basic Settings (Settings Drawer)

Setting	Description
Audio Source	Microphone or system audio loopback device
Speaking Language	Language for voice transcription (101+ supported)
Output Language	Language for transcriptions and AI replies
Corrected Transcription	AI-powered post-correction oftranscribed text
Auto-copy to Clipboard (CRC)	Auto-copyAI responses to clipboard
Transparency	Window opacity from 0% to 100%
Theme	Light / Dark mode
Always on Top	Keep CFAI above all other windows
Bluetooth Controller	Controller connection status

Advanced Settings — App Preferences

Setting	Description
Theme	Light / Dark mode toggle
Corrected transcription	AI post-correction oftranscribed text
Auto-copy prompt to clipboard	Automatic clipboard copy ofgenerated prompts
Audio input mode	Configure audio capture behavior
Suggested counter-question	AI suggests follow-up questions automatically
App Language	Interface language (separate from transcription language)
Local Model	Status indicator (CONNECTED / DISCONNECTED) for connected LLM
Visible response sections	Granular control over response section display
Response length (max tokens)	Maximum token limit for AI responses
Auto-apply interval (seconds)	Interval for automatic context re-application
Enhanced knowledge search	Advanced knowledge base search
Ricerca Web (Web Search)	Enable/disable integrated web search

Advanced Settings — Other Sections

- **Keyboard Shortcuts** – remap any global hotkey
- **Memory (Conversation Context)** – toggle, set number ofmessages to include (disabled by default)
- **Cache** – Response Cache + Semantic Cache (separate toggles), view, export, clear
- **Transcription** – download Whisper models, enable GPU acceleration, switch model, Test GPU support
- **In-Place Replace** – add, edit, delete custom presets; set active preset
- **Text-to-Speech** – install Piper voices per language, test voices
- **Clipboard History** – enable capture, set entries sent to LLM (default 10), view status

---

Account & Credits

Check your plan: Click **About** (bottom-right ofthe main window) to view your current plan and remaining credits.

Upgrade or change your plan:

- From the in-app Upgrade modal (shown when relevant)
- From your account page on the CFAI website
- Stripe-powered secure checkout
- Auto-renews monthly – cancel anytime

Out ofcredits: Ifyou run out ofcredits before the end ofthe month, an in-app modal lets you top up immediately or upgrade to a larger plan.

Password recovery: Available from the login screen — click "Forgot password" and enter your email.

PrivacyArchitecture — What Runs Where

CFAI is built privacy-first. Here is exactly what runs locally vs in the cloud.

Always 100% local (never cloud)

- Voice transcription (all Whisper models)
- OCR (text recognition from images)
- TTS (Piper neural synthesis)
- Document RAG embeddings (passage indexing)
- Semantic Cache + Response Cache
- Clipboard History capture

- Active Contexts detection (open windows scanning)
- Cognitive Map computation

Cloud optional (configurable)

- Main AI Chat → either CFAI's endpoint or your local LLM (Ollama / LM Studio / Jan.ai)
- Web Search → requires internet by definition
- Translation engine ifnot in local-only mode

Never sent to cloud

- Raw audio (never)
- Document content for RAG (only relevant tokens as query, never the full document)
- Clipboard content
- Raw screenshots

See the full Privacy Policy at **cfai.io** for GDPR details.

ITALIANO

Perché CFAI?

CFAI non è l'ennesimo chatbot AI. È un assistente cognitivo in tempo reale progettato per supportarti nei processi mentali complessi – mentre parli, scrivi, presenti o partecipi a una riunione — senza sostituire il tuo pensiero critico.

Immagina un assistente che ti segue mentre agisci, ti aiuta a rimanere focalizzato, a catturare ciò che conta e ad agire in tempo reale. Non fa tutto al posto tuo: lavora accanto a te.

Chi può beneficiarne?

- Developer durante discussioni tecniche, sessioni di design o brainstorming
- Candidati che si preparano per colloqui o assessment complessi
- Recruiter che valutano profili in tempo reale
- Professionisti delle vendite durante call, demo clienti e consulenze
- Content creator che preparano pitch, script e presentazioni
- Vocal coach, formatori, educatori
- Chiunque lavori con riunioni, documenti o flussi complessi

Piani di Abbonamento

CFAI usa un sistema a crediti. Un credito copre circa 1.000 token AI (≈ un breve scambio di messaggi). Le funzionalità locali che non chiamano l'AI cloud – come la trascrizione vocale Whisper — non consumano crediti.

Piano	Prezzo	Crediti	Indicato per
Trial	Gratis	Uso gratuito limitato	Provare l'app, nessuna carta richiesta
Starter	€7,99 / mese	500 crediti	Uso leggero, richieste AI occasionali
Popular	€12,99 / mese	1500 crediti	Uso quotidiano regolare
Best Value	€24,99 / mese	3000 crediti	Power user, uso intensivo

Tutti i piani a pagamento si rinnovano automaticamente ogni mese. Puoi cambiare o disdire in qualsiasi momento dalla pagina del tuo account.

Cosa fa CFAI — Le funzionalità

CFAI offre quattro modalità principali e un ampio set di strumenti di supporto.

Modalità principali

1. **Chat AI (con TTS)** – AI conversazionale a voce
2. **Scrittura Libera (Trascrizione Vocale)** – dettatura in un editor di testo
3. **Screenshot + Prompt** – cattura lo schermo e chiedi all'AI
4. **Mappa Predittiva Cognitiva** – bussola semantica in tempo reale

Funzionalità di supporto

- Dual Audio Capture – registra la tua voce e l'audio di sistema contemporaneamente
- CRC (Clipboard Response Copy) – copia automatica delle risposte AI
- Traduzione in tempo reale – traduci il parlato al volo
- Local LLM Support – collega il tuo server AI personale (Ollama, LM Studio, Jan.ai)
- Document RAG – chatta con documenti allegati
- Speaker Diarization – identifica chi sta parlando nelle riunioni
- 101 lingue supportate – trascrizione multilingue via Whisper
- OCR Locale – leggi testo da immagini o schermo
- Agentic AI – pianificazione e uso di strumenti multi-step
- Lecture Mode – script di lezione generati dall'AI con narrazione TTS
- Meeting Mode – assistenza live alle riunioni con riconoscimento dei parlanti
- Multi-Session – più sessioni di chat in parallelo
- Context Orientation – consapevolezza automatica di finestre aperte e clipboard
- Response Cache + Cache Semantica – evita di ripetere chiamate AI identiche o simili
- In-Place Replace (IPR) – trasforma testo selezionato in qualsiasi app
- Web Search – ricerca web integrata nella chat

- Conversation Context – memoria persistente tra i messaggi con 4 modalità di configurazione
- Quick Launch – launcher full-screen accessibile da ovunque
- Always on Top – mantieni CFAI sopra le altre finestre
- Tray Icon – presenza silenziosa nel tray di Windows
- Suggested Counter-Question – l'AI suggerisce domande di follow-up automaticamente
- Auto-Apply Interval – riapplicazione automatica del contesto a intervalli configurabili
- TTS Offline (Piper) – sintesi vocale neurale totalmente locale

---

Conversation Context — Riferimento Completo

Il Conversation Context è una delle funzionalità più potenti di CFAI. È configurabile tramite 4 tab distinti:

Tab 1 — Instructions

Un'area di testo libera per aggiungere un contesto o un'istruzione personalizzata che guida l'intera conversazione.

- Aggiungi prompt custom che modellano ogni risposta AI
- **Supporto documento allegato:** trascina un file direttamente qui per RAG nella conversazione
- Formati supportati: `.txt`, `.md`, `.pdf`, `.docx`, `.csv`, `.json`
- Il documento resta sul tuo dispositivo – solo i passage embeddings vengono calcolati localmente
- La UI mostra il numero di passaggi indicizzati (es. "39 passages indexed")

Tab 2 — Guided

Modalità di consapevolezza pre-configurate per contesti tecnici e strutturati.

Collection Depth – tre livelli:

- **Instant** – contesto minimo, latenza più bassa
- **Moderate** – contesto esteso (clipboard, tab, file), latenza ~500ms
- **Deep** – consapevolezza completa del sistema, analisi più profonda

Su cosa CFAI deve focalizzarsi – aree selezionabili (checkbox):

- **Code & Technical** – Focus su codice, algoritmi, dettagli tecnici
- **UI / UX Structure** – Focus su interfaccia, layout, esperienza utente
- **Business Logic** – Focus su regole, workflow, logica di dominio
- **Documentation & Specs** – Focus su documentazione, specifiche, requisiti

Tab 3 — Active Contexts

Visibilità in tempo reale di tutte le applicazioni e finestre aperte sul sistema. CFAI può includere finestre selezionate nel contesto AI.

- Colonne mostrate: **APP**, **WINDOW** (titolo), **TYPE** (DOCS, OTHER, ecc.)
- Riconoscimento automatico del tipo in base ad applicazione e contenuto
- Selezione multipla con **Select all** / **Clear**
- Pulsante di refresh per aggiornare la lista delle finestre in tempo reale
- Decidi tu quali finestre vengono incluse nel contesto AI – privacy by design

Tab 4 — Clipboard

Cattura cronologia appunti automatica opzionale per inclusione nel contesto AI.

- Abilita/disabilita tramite Impostazioni Avanzate → Clipboard History
- Configurabile: quante entries recenti inviare all'LLM (default 10)
- Indicatore di stato: "Listening" / "Not listening"
- Cattura solo mentre attiva – nessuna cronologia retroattiva inviata

---

Conversation Memory (Mirror Memory)

Separata dal Conversation Context. La Conversation Memory persiste i messaggi all'interno della sessione di chat corrente.

- **Stato di default: DISABILITATA**
- Quando disabilitata: "Each message will be sent without previous context. The AI won't remember what you discussed earlier."
- Quando abilitata: gli ultimi N messaggi vengono inclusi automaticamente in ogni nuova richiesta
- Il numero di messaggi da includere è configurabile
- Va abilitata esplicitamente – privacy-first di default

Response Cache & Cache Semantica

CFAI mantiene due cache locali distinte:

Response Cache

Memorizza risposte AI letterali. Se fai la **stessa identica domanda** due volte, la risposta in cache viene restituita istantaneamente.

Cache Semantica

Memorizza risposte AI semanticamente simili. Se fai **una domanda simile** (formulazione diversa ma stessa intenzione), la risposta in cache viene restituita.

Entrambe attivabili indipendentemente in Impostazioni Avanzate → Cache.

Azioni di gestione cache:

- View entries
- Export
- Clear this chat
- Clear All

Entrambe le cache risparmiano crediti e tempo.

Modelli Whisper — Lista Completa

Tutti i modelli sono forniti in formato **CT2 INT8** (CTranslate2 quantized) per performance ottimizzate.

Modello	Dimensione	Caratteristiche
Whisper Tiny (CT2 INT8)	77 MB	Qualità minore – molto veloce, poca RAM
Whisper Base (CT2 INT8)	145 MB	Bilanciato – default consigliato
Whisper Small (CT2 INT8)	480 MB	Qualità migliore – più lento su CPU
Whisper Medium (CT2 INT8)	1530 MB	Alta qualità – GPU fortemente consigliata
Whisper Large-v3 (CT2 INT8)	3090 MB	Qualità massima – GPU richiesta per real-time

GPU Acceleration: solo NVIDIA — richiede librerie CUDA (~600 MB). Pulsante **Test GPU support** disponibile per verificare il setup prima del download.

TTS Offline (Piper)

CFAI usa **Piper**, un motore TTS neurale open-source, in esecuzione totalmente offline.

- Dimensione motore: ~7 MB
- Dimensione modello voce: ~60 MB per lingua
- Lingue disponibili: Italiano (Paola), Inglese (Lessac) e altre
- Installa singole voci on demand da Impostazioni Avanzate → Text-to-Speech
- Testa le voci prima di confermare il download
- 100% offline dopo il download iniziale

In-Place Replace (IPR) — Preset

Preset di default forniti:

- **Correggi grammatica** — "correggi la grammatica e l'ortografia, restituisci solo il testo corretto"
- **Traduci in inglese** – "traduci il testo in inglese, restituisci solo il testo tradotto"
- **Rendi formale** – "riscrivi il testo in tono formale e professionale, restituisci solo il testo riscritto"

Puoi creare **preset completamente custom** con prompt arbitrari, non solo modificare quelli esistenti. Ogni preset ha un nome + un prompt completo. **IPR Quick Pick** apre un selettore per scegliere il preset al volo.

Hotkey

Tutti i tasti rapidi sono globali – funzionano anche quando CFAI non è in primo piano.

Azione	Tasto Rapido Default	Stato
CRC (Clipboard Response Copy)	Ctrl+Shift+Space	DEFAULT
Quick Launch Menu	CapsLock+Shift+Ctrl	DEFAULT
Snipping Tool	Ctrl+Alt+N	DEFAULT
In-Place Replace (IPR)	Ctrl+Shift+Alt+Space	Non configurato di default
IPR Quick Pick	Ctrl+Shift+Alt+Q	Non configurato di default
Capture Screen (full screen)	—	Non configurato di default

Tutti gli hotkey possono essere rimappati in **Impostazioni Avanzate → Scorciatoie Tastiera**.

Impostazioni Avanzate — App Preferences

Impostazione	Descrizione
Tema	Toggle Chiaro / Scuro
Trascrizione corretta	Post-correzione AI del testo trascritto

Impostazione	Descrizione
Auto-copia prompt negli appunti	Copia automatica dei prompt generati
Audio input mode	Configura comportamento cattura audio
Suggested counter-question	L'AI suggerisce domande di follow-up automaticamente
App Language	Lingua interfaccia (separata dalla lingua trascrizione)
Local Model	Indicatore di stato (CONNECTED / DISCONNECTED) per LLM collegato
Visible response sections	Controllo granulare sulla visualizzazione delle sezioni di risposta
Response length (max tokens)	Limite token massimo per le risposte AI
Auto-apply interval (seconds)	Intervallo per la riapplicazione automatica del contesto
Enhanced knowledge search	Ricerca avanzata nella knowledge base
Ricerca Web	Abilita/disabilita la ricerca web integrata

Architettura Privacy — Cosa Gira Dove

CFAI è costruito privacy-first. Ecco esattamente cosa gira in locale vs in cloud.

Sempre 100% locale (mai cloud)

- Trascrizione vocale (tutti i modelli Whisper)
- OCR (riconoscimento testo da immagini)
- TTS (sintesi neurale Piper)
- Document RAG embeddings (indicizzazione passaggi)
- Response Cache + Cache Semantica
- Cattura Clipboard History
- Active Contexts detection (scansione finestre aperte)
- Calcolo Cognitive Map

Cloud opzionale (configurabile)

- AI Chat principale → endpoint CFAI o LLM locale (Ollama / LM Studio / Jan.ai)
- Web Search → richiede internet per definizione
- Motore di traduzione se non in modalità local-only

Mai inviato in cloud

- Audio raw (mai)
- Contenuto documenti per RAG (solo token rilevanti come query, mai il documento completo)
- Contenuto clipboard
- Screenshot raw

Vedi la Privacy Policy completa su **cfai.io** per i dettagli GDPR.
