VADEMECUM

MarcoFLY Framework

Il prontuario completo dell'app e del framework: ogni scelta, cosa fa e cosa implica.

Marco Motta (MarcoFLY)

marcofly.app · 13 agosto 2026

Su questo documento

Questo vademecum è generato dalla stessa fonte della guida pubblicata su marcofly.app/guide e di quella dentro l'app: non esiste un testo separato tenuto allineato a mano. È uno snapshot — la data in copertina dice quando è stato composto.

marcofly.app/guide

Indice

01

👋 Benvenuto

MFF (MarcoFLY Framework) è uno strumento di controllo epistemico sopra l'AI generativa: rende visibile quanto un modello è sicuro di ciò che dice, affermazione per affermazione, e ti aiuta a verificarlo. Questa app applica il framework automaticamente a ogni messaggio, su decine di modelli di provider diversi, con le TUE chiavi API (BYOK).

Le due modalità: «MFF Framework» inietta il protocollo epistemico (etichette, scudi, carta dello stato) e consuma crediti MFF solo quando il framework viene applicato davvero. «Chat semplice» (PLAIN) è conversazione libera senza framework e senza crediti — ma con TUTTE le funzioni della piattaforma: allegati, ricerca web, peer review, lettura vocale, export.

Cosa MFF NON è

- Non garantisce verità: le etichette riflettono ciò che il modello dichiara di sapere; la verifica finale resta tua.
- Non sostituisce esperti: per decisioni mediche, legali, fiscali o finanziarie serve un professionista.
- Non azzera le allucinazioni: le riduce e le rende riconoscibili, ma il modello può sbagliare anche con bollino verde.

02

🚀 Avvio rapido — 3 passi

1. Configura una chiave API (o usa i crediti MFF inclusi). Vai in Impostazioni → Chiavi provider AI: incolla una chiave — la via più rapida è «Collega OpenRouter» via OAuth, una chiave sola per centinaia di modelli, gratuiti inclusi.
2. Crea una sessione dalla Dashboard: scegli modalità (MFF o PLAIN), dominio, modello — e se vuoi i moduli di difesa. Ogni scelta è spiegata nelle sezioni di questo vademecum.
3. Scrivi la domanda: la risposta arriva in streaming, con le etichette. Con 📎 alleghi immagini e documenti, con 🌐 chiedi una ricerca web prima della risposta. Ogni messaggio ha 📋 copia, 🔊 lettura, ⚖️ peer review e la rigenerazione.

> 💡 Per le prime prove usa un modello :free (gratuito) — perfetto per imparare l'interfaccia. Poi passa a un modello a pagamento con la tua chiave per contesto pieno e ricerca web.

03

🔓 Chiavi gratuite, provider per provider

Ogni capitolo qui sotto è linkabile (icona 🔗): dalle Impostazioni della PWA arrivi dritto al provider che stai configurando. La regola di lettura: 🆓 si parte senza carta · 🌱 costo d'ingresso minimo · 💳 serve credito — dichiarato, mai nascosto.

> 💡 Se è il tuo primo giorno: fai SOLO il primo capitolo (OpenRouter, 2 minuti con l'automatismo). Con quella sola utenza hai subito più LLM gratuiti e puoi usare tutta l'app; il resto lo aggiungi quando ti serve.

OpenRouter FREE

`openrouter.ai` · `sk-or-…`

La porta d'ingresso consigliata: UNA chiave, centinaia di modelli — e i :free si usano SENZA depositare nulla.

1. In Impostazioni → Chiavi provider AI premi «Collega OpenRouter»: è l'automatismo MFF — si apre il sito OpenRouter, fai login (o crei l'account in 30 secondi), autorizzi, e la chiave arriva a MFF da sola. Niente copia-incolla, niente errori di battitura.
2. Da quel momento, con la STESSA utenza, hai subito diversi LLM gratuiti (i modelli col suffisso :free) per iniziare a usare il sistema: crei una sessione, scegli OpenRouter, e il selettore ti mostra i :free disponibili.
3. In alternativa manuale: openrouter.ai → Keys → «Create key» → incolla in MFF. Se un giorno vorrai i modelli a pagamento, aggiungi credito sul TUO account OpenRouter: MFF non c'entra nulla col pagamento.

Google AI Studio FREE

`aistudio.google.com` · `AIza…`

Free tier generoso sui Gemini: spesso non serve nemmeno la carta per iniziare.

1. Vai su aistudio.google.com col tuo account Google → «Get API key» → «Create API key».
2. Copia la chiave (inizia con AIza) e incollala in Impostazioni → card Google AI: MFF la valida con una query di prova e la cifra.

Groq FREE

`console.groq.com` · `gsk_…`

Free tier con limiti mensili: velocità di risposta 5-10× i provider classici — ottimo per prove e compiti rapidi.

1. console.groq.com → registrati → API Keys → «Create API Key».
2. Copia la chiave gsk_… e incollala nella card Groq di MFF.

Cerebras FREE

`cloud.cerebras.ai` · `csk-…`

Free tier molto generoso su chip wafer-scale: modelli open (Llama, Qwen) a velocità altissima.

1. cloud.cerebras.ai → account gratuito → API Keys → crea la chiave csk-…
2. Incollala nella card Cerebras di MFF.

NVIDIA NIM FREE

`build.nvidia.com` · `nvapi-…`

Free tier permanente (40 richieste/minuto): catalogo ampio di modelli open serviti da NVIDIA.

1. build.nvidia.com → account (gratuito) → API Catalog → «Get API Key».
2. Incolla la chiave nvapi-… nella card NVIDIA NIM.

SambaNova FREE

`cloud.sambanova.ai` ·

Credito iniziale gratuito + un tier persistente (20 richieste/min): buono per modelli open veloci.

1. cloud.sambanova.ai → registrati → API Keys → crea la chiave.
2. Incollala nella card SambaNova di MFF.

Cloudflare Workers AI FREE

`dash.cloudflare.com` ·

Free tier integrato in ogni account Cloudflare: modelli open sull'edge CF.

1. dash.cloudflare.com → My Profile → API Tokens → crea un token con permesso «Workers AI Read».
2. In MFF la card Cloudflare chiede DUE valori: il token e l'Account ID (lo trovi nella home del dashboard CF).

Mistral FREE

`console.mistral.ai` ·

Free tier per esperimenti sui modelli Mistral; piani a consumo per l'uso serio.

1. console.mistral.ai → registrati → API Keys → crea la chiave.
2. Incollala nella card Mistral di MFF.

Cohere FREE

`dashboard.cohere.com` ·

Trial key gratuita generosa (rate-limitata); la chiave production è a parte.

1. dashboard.cohere.com → account → API Keys → usa la Trial key.
2. Incollala nella card Cohere di MFF.

Hyperbolic LOW

`app.hyperbolic.xyz` ·

Piccolo credito iniziale gratuito, poi prezzi bassissimi (fino a ~10× sotto i listini classici) sui modelli open.

1. app.hyperbolic.xyz → registrati → Settings → API Key.
2. Incollala nella card Hyperbolic di MFF.

OrcaRouter LOW

`orcarouter.ai` · `sk-orca-…`

Aggregatore zero-markup con chiave propria e indipendente: paghi i listini dei vendor senza ricarico. Non usa il tuo credito OpenRouter.

1. orcarouter.ai → account → API Keys → crea la chiave sk-orca-…
2. Incollala nella card OrcaRouter di MFF.

DeepSeek PAID

`platform.deepseek.com` · `sk-…`

Niente free tier, ma il costo d'ingresso è tra i più bassi (ricarica minima di pochi dollari) e i listini sono molto aggressivi.

1. platform.deepseek.com → account → ricarica minima → API Keys.
2. Incolla la chiave sk-… nella card DeepSeek di MFF.

OpenAI PAID

`platform.openai.com · sk-…`

Niente free tier API: serve un minimo di credito sull'account. (I GPT si provano gratis anche via OpenRouter :free quando disponibili.)

1. platform.openai.com → API Keys → «Create new secret key» → aggiungi credito minimo nel Billing.
2. Incolla la chiave sk-… nella card OpenAI di MFF.

Anthropic PAID

`console.anthropic.com · sk-ant-…`

Niente free tier API: credito minimo nel Billing. In cambio: i Claude con ricerca web nativa e reasoning esteso.

1. console.anthropic.com → API Keys → «Create Key» → ricarica nel Billing.
2. Incolla la chiave sk-ant-… nella card Anthropic di MFF.

xAI (Grok) PAID

`console.x.ai · xai-…`

Richiede credito iniziale sull'account.

1. console.x.ai → API Keys → crea la chiave xai-… → aggiungi credito.
2. Incollala nella card xAI di MFF.

Perplexity PAID

`perplexity.ai/settings/api · pplx-…`

Serve credito, ma i Sonar hanno la ricerca web NATIVA: le citazioni compaiono direttamente tra le fonti in MFF.

1. perplexity.ai/settings/api → genera la API key pplx-... → ricarica.
2. Incollala nella card Perplexity di MFF.

Z.AI (GLM) PAID

`z.ai ·`

I modelli GLM di Zhipu; serve credito sull'account.

1. z.ai → account → API Keys → crea la chiave.
2. Incollala nella card Z.AI di MFF.

04

🏷️ Le etichette epistemiche

Ogni affermazione porta il suo livello di impegno dichiarato. Non sono decorazione: sono un contratto — e restano nell'export. La regola aurea: il tag PRECEDE il contenuto, un tag per blocco.

🟢 **SICURO**

Fonte diretta in sessione, nessun dubbio. Senza fonte citabile questa etichetta è VIETATA: declassa da sola a PROBABILE.

🔵 **PROBABILE**

Deduzione solida, ma la fonte esatta non è in sessione. Fiducia alta, non un fatto provato.

🟡 **FORSE**

Stima con margine significativo. Buon punto di partenza, pericoloso come punto di arrivo.

🟠 **DIPENDE**

Vero solo a certe condizioni: leggi le premesse prima di agire.

🔴 **NON SO**

Dati insufficienti o contraddittori. Invece di inventare, MFF si ferma.

⚫ **NON POSSO**

Fuori dal training o oltre il cutoff. Onestà strutturale, non reticenza.

🌐 **CONFERMATO:WEB**

Ammessa SOLO se una ricerca web reale è avvenuta in quel turno, con la fonte citata.

🌐 VERIFICATO:WEB

La ricerca corrobora senza conferma esclusiva. Stesse regole rigide della gemella.

I suffissi: da dove viene e quanto ci crede

Dopo l'etichetta possono comparire suffissi di fonte: [doc] documento ufficiale · [test] testimonianza · [comm] comunicato · [log] registro tecnico · [emp] dato empirico · [lit] letteratura scientifica · [leg] fonte legale · [cert] certificazione verificabile. Poi [mem] = memoria del training non verificata in sessione (sempre da controllare) e [p:75%] = probabilità soggettiva stimata dal modello. Il glossario in fondo li spiega tutti.

05

🛡️ Gli scudi — moduli di difesa

Nella modalità MFF puoi attivare fino a 9 moduli epistemici che guidano il comportamento del modello. L1 è sempre attivo; il set predefinito L1+L2+L3 è ottimo per la maggior parte dei compiti. Le schede qui sotto sono le STESSE che vedi nel wizard di creazione sessione.

📌 L1 · Citazione Forzata

Sempre attivo e non disattivabile. Ogni affermazione del modello deve avere una base epistemica esplicita: il modello non può produrre dichiarazioni fattuali prive di una fonte o di un ragionamento verificabile.

🔎 L2 · ZVE — Zona Verifica Esterna

Crea una Zona di Verifica Esterna: il modello distingue attivamente tra ciò che è certo e ciò che richiede verifica indipendente, segnalando i contenuti a rischio di imprecisione.

⚖️ L3 · Falsificazione — condizioni di fallimento

Applica il principio popperiano di falsificabilità: il modello deve indicare esplicitamente le condizioni che potrebbero smentire le proprie affermazioni.

🌐 L4 · Grounding Web — ricerca web obbligatoria

Attiva la ricerca web in tempo reale prima di rispondere su argomenti che richiedono dati aggiornati. Riduce drasticamente le allucinazioni su eventi recenti o informazioni variabili nel tempo. Le etichette 🌐 CONFERMATO:WEB / 🌐 VERIFICATO:WEB sono ammesse solo se un blocco di grounding è realmente iniettato nel turno (v1.5.3 anti-hallucination).

L4 si attiva con il bottone 🌐 nella chat ed è universale tramite MWAL (Tavily lato MFF o BYOK Serper/Tavily/SerpAPI/Brave). I provider con web_search nativo (Anthropic, Perplexity, OpenRouter paid) lo usano direttamente. Disabilitato sui modelli :free per cost gating.

🧠 L5 · Peer Review — reasoning esteso

Simula il processo di peer review accademica e abilita i token di ragionamento esteso sui modelli che li supportano (DeepSeek-R1, Claude, Qwen Thinking, ecc.). Aumenta la precisione e riduce gli errori.

Configura lo sforzo di reasoning (basso / medio / alto / massimo) nelle Impostazioni avanzate per bilanciare costo e qualità della risposta.

🔄 L6 · Deriva Epistemica — anti-allucinazione

Monitora e corregge la deriva cognitiva nella conversazione: riconosce quando le risposte successive contraddicono le precedenti o si discostano dal contesto originale.

✅ L7 · VMS — validazione multi-sorgente

Validazione Multi-Sorgente: la seconda fonte la scegli tu selezionando un peer model di un provider diverso. MFF gira automaticamente la stessa domanda al secondo modello e produce un blocco ── VMS ── con CONVERGENZA ALTA/MEDIA/BASSA.

🛡️ PAVA · Protocollo PAVA — anti-amnesia

Protocollo Scudo. Mantiene un registro strutturato del contesto conversazionale per prevenire la perdita di coerenza nelle sessioni lunghe. Il modello viene periodicamente ri-ancorato al contesto e agli obiettivi della sessione.

⚙️ NIST · NIST-RMF — risk management

Protocollo Scudo. Allinea il processo di risposta al NIST Risk Management Framework, classificando e gestendo il rischio epistemico in ogni fase della risposta.

> 💡 Aggiungi L4 per fatti recenti (si usa poi col bottone 🌐 in chat), L5 per ragionamento complesso, L7 se hai configurato un peer model. PAVA e NIST sono per sessioni molto lunghe o domini critici. Implicazione: più scudi = più rigore, ma risposte più lunghe e più token.

> ⚠️ Sui modelli :free gli scudi che costano (L4 ricerca, L5 reasoning esteso) sono disattivati per contenimento costi.

06

Sessioni: ogni scelta e cosa implica

Una sessione è una conversazione completa con un modello: ha titolo, modello, modalità, dominio, lingua e scudi propri, fissati alla creazione. Ogni scelta del wizard cambia qualcosa di concreto nel motore.

Il dominio fa due cose

Primo: definisce cosa significano / / in quel contesto — in dominio scientifico «SICURO» vuol dire risultato replicato e pubblicato; in dominio creativo la soglia è un'altra. Secondo: imposta la temperatura predefinita, da 0.2 (Legale: conservativo e ripetibile) a 1.0 (Creativo: massima variabilità). Nelle impostazioni avanzate puoi forzarla.

La modalità operativa

MFF-E: analisi epistemica rigorosa · MFF-G: riduce rumore e ridondanza · MFF-X: interpreta e spiega senza alterare · MFF-EX: le combina (la modalità centrale) · MFF-EGX: aggiunge la disciplina della sintesi · AUTO: sceglie il protocollo turno per turno. PLAIN: nessun framework, nessun credito.

La lingua della sessione è una scelta, e resta

È separata dalla lingua dell'interfaccia. Cambiare lingua UI NON traduce le risposte già generate: l'etichetta epistemica appartiene al testo che quel modello ha prodotto — tradurlo sotto le etichette le falsificherebbe.

Dentro la conversazione

- Ogni messaggio successivo porta il contesto della sessione: fai follow-up senza ripetere. Per argomenti nuovi apri una sessione nuova

(contesto pulito, costi minori).

- Sotto ogni risposta: 📋 copia · 🌐 fonti · 🔊 lettura vocale · ⚖️ peer review · 🔁 rigenera. I tuoi messaggi si possono modificare e reinviare.
- «▸ Continua» prosegue una risposta troncata dal budget; nelle sessioni oltre i 100 messaggi «⤒ Carica messaggi precedenti» risale la storia.
- Tutte le sessioni sono nell'Archivio: riapri e continui da dove eri. Da lì si eliminano anche, singolarmente.

07

🧠 Scegliere il modello

Il selettore mostra sempre il listino completo del provider, marcando quali modelli risultano serviti in questo momento: la verità viene dal catalogo live, non da una lista scritta a mano. Un modello non servito è visibile e marcato, non nascosto.

Anche le capacità vengono dal catalogo: visione → allegati immagine abilitati; reasoning → pensiero esteso con L5; generazione immagini → il modello entra nelle rotte di generazione. Il nome del modello non decide nulla: decide ciò che il provider dichiara.

Gratuiti contro a pagamento

I modelli :free non costano nulla, con tre compromessi: contesto e budget documenti più stretti, niente ricerca web 🌐, e quote giornaliere del provider (un 429 su un :free spesso significa solo «riprova più tardi o cambia :free»). I modelli a pagamento girano sulla TUA chiave: paghi il listino del provider, MFF non aggiunge nulla.

Se il provider fallisce: il failover dichiarato

Su errori recuperabili (rate limit, sovraccarico, modello ritirato) il motore prova le riserve: prima i fallback che indichi tu nelle impostazioni avanzate (fino a 3, in ordine), poi la catena di policy sui tuoi provider con chiave. Tutto è DICHIARATO: risposta, etichette e carta riportano chi ha davvero servito. Il turno riparte da zero sul modello di riserva — quello che leggi è sempre di un solo modello. Errori non recuperabili (chiave non valida, saldo esaurito) non fanno failover: vengono mostrati, e il credito MFF torna indietro.

> 💡 Riepilogo di scelta: compiti generici → un medio veloce ed economico · analisi profonda → un modello con reasoning (più lento, più costoso) · fatti recenti → modello a pagamento + 🌐 · bozze ed esperimenti → :free.

08

Provider e routing

MFF parla con 17 provider, in due famiglie. Il provider della sessione lo scegli TU alla creazione — non c'è un instradamento nascosto: quello che scegli è quello che risponde, e se serve un ripiego viene dichiarato (v. failover).

Aggregatori: una chiave, tanti vendor

- OpenRouter (chiave sk-or-…): centinaia di modelli, inclusi i :free; si collega anche in un click via OAuth. Sui modelli a pagamento usa il suo web_search nativo. Piccolo markup (~5%) rispetto al diretto.
- OrcaRouter (chiave sk-orca-…): aggregatore zero-markup con chiave propria e indipendente — non usa il tuo credito OpenRouter.

Diretti: la strada più corta e le capacità native

OpenAI · Anthropic · Google · Groq · Cerebras · Mistral · DeepSeek · xAI · Perplexity · Z.AI · NVIDIA NIM · SambaNova · Hyperbolic · Cohere · Cloudflare Workers AI. Nessun markup, e le funzioni native del vendor: la ricerca web integrata di Anthropic, le citazioni native di Perplexity (mostrate tra le fonti), i free tier ampi di Groq, Cerebras, NIM e SambaNova.

> 💡 Configurazione minima: una chiave OpenRouter basta per tutto (gratuiti inclusi). Aggiungi chiavi dirette per i provider che usi di più: elimini il markup e sblocchi le capacità native. La peer review, per progetto, pesca dai tuoi provider DIRETTI: il revisore è indipendente dal canale del primo modello.

La generazione di immagini è una capacità della piattaforma, non del modello di chat: se il modello di sessione non genera immagini, il motore instrada sul primo dei tuoi provider con chiave che può farlo, e lo dichiara.

09

🔑 Configurare le chiavi API

1. Vai in Impostazioni → Chiavi provider AI e apri la card del provider.
2. Incolla la chiave: MFF la valida SUBITO con una query di test reale — una chiave che non passa il test non viene salvata.
3. La chiave viene cifrata lato server (key-ring AES con rotazione) e non viene mai più mostrata in chiaro, mai messa in un URL, mai mandata al browser. Il modello non la vede MAI: le richieste partono dal server.
4. Ogni chiave mostra un semaforo (attiva · non valida · in pausa) e, in vista Esperto, la whitelist 🎯 dei modelli consentiti: la sessione e il failover non usciranno mai dalla lista.

Dove creare la chiave, provider per provider

PROVIDER	DOVE	NOTE
OpenAI	platform.openai.com → API Keys	sk-… · serve credito minimo
Anthropic	console.anthropic.com → API Keys	sk-ant-… · credito nel Billing
Google AI Studio	aistudio.google.com → Get API key	AIza… · free tier generoso
Groq	console.groq.com → API Keys	gsk_… · free tier, ottimo per test
Perplexity	perplexity.ai/settings/api	pplx-… · ricerca nativa con citazioni

OpenRouter	openrouter.ai → Keys (o OAuth da MFF)	sk-or-… · anche senza deposito coi :free
DeepSeek	platform.deepseek.com → API Keys	sk-… · ricarica minima bassa
Cerebras	cloud.cerebras.ai → API Keys	csk-… · free tier molto generoso
xAI (Grok)	console.x.ai → API Keys	xai-… · richiede credito iniziale
Mistral	console.mistral.ai → API Keys	free tier per esperimenti
Cohere	dashboard.cohere.com → API Keys	trial key gratuita generosa
Cloudflare Workers AI	dash.cloudflare.com → API Tokens	token Workers AI + Account ID
NVIDIA NIM	build.nvidia.com → Get API Key	nvapi-… · free tier permanente
SambaNova	cloud.sambanova.ai → API Keys	credito iniziale + tier persistente
Hyperbolic	app.hyperbolic.xyz → Settings → API Key	prezzi molto bassi
Z.AI (GLM)	z.ai → API Keys	modelli GLM

Le chiavi di RICERCA (per il 🌐)

Separate dalle chiavi dei modelli: Serper, Tavily, SerpAPI, Brave — sempre in Impostazioni, sezione Ricerca web. Anche queste validate con una query di prova. Eleggi la predefinita con la ⭐: il motore prova prima quella, le altre restano riserve nell'ordine fisso Serper → Tavily → SerpAPI → Brave. Con una tua chiave paghi solo il tuo provider; senza, usi la quota inclusa di MFF (limitata e condivisa).

> ⚠️ Non condividere mai le chiavi: non incollarle in chat, email o documenti. Se sospetti una compromissione, revocala subito dal dashboard del provider — poi inserisci la nuova in MFF.

10

🌐 Ricerca web e fonti

1. Premi 🌐 accanto al composer: un mini-modello estrae la query di ricerca proposta e te la MOSTRA. Puoi correggerla o riscriverla — la query è una decisione tua, non un automatismo cieco.
2. La ricerca gira sulla catena MWAL: la tua chiave predefinita per prima, poi le riserve (Serper → Tavily → SerpAPI → Brave). Se un motore fallisce, il passaggio al successivo è dichiarato.
3. I risultati entrano nel contesto come blocco di grounding: la risposta può citarli, e le fonti restano nel chip 🌐 del messaggio — sopravvivono al reload e finiscono nell'export.

> ⚠️ Regola anti-allucinazione: le etichette 🌐 CONFERMATO:WEB e 🌐 VERIFICATO:WEB sono ammesse SOLO se un blocco di grounding reale è stato iniettato in quel turno (o il provider ha ricerca nativa: Anthropic, Perplexity). Senza, il modello deve dichiarare che non ha accesso live e ripiegare su 🔵 PROBABILE[mem] — «ho cercato sul web» detto a parole non è una prova.

- Limite onesto n.1: gli snippet dei motori sono per costruzione una cache — per dati che cambiano al minuto (quotazioni, l'ora esatta) la fonte può essere pertinente ma stantia.
- Limite onesto n.2: sui modelli :free il 🌐 è disattivato per contenimento costi.

11

⚖️ Peer review fra modelli

Un SECONDO modello, di un provider diverso, rilegge la risposta e ti mostra il giudizio accanto all'originale. È il livello L5-B del protocollo: la convergenza di due agenti indipendenti può promuovere un 🔵 PROBABILE a 🟢 SICURO; una divergenza forza 🟡 o 🔴. Due modelli che non condividono né pesi né canale, d'accordo sulla stessa affermazione, valgono più di uno.

1. Alla creazione sessione, in Impostazioni avanzate, scegli il peer model: la lista pesca dai tuoi provider DIRETTI con chiave, deliberatamente diversi dal provider di sessione.
2. In chat premi ⚖️ sotto la risposta che vuoi far rileggere: la review arriva in streaming e resta agganciata a quel messaggio, col suo verdetto di convergenza.
3. In alternativa, l'auto-trigger (checkbox in Avanzate) lancia la review su OGNI risposta: modalità per sessioni critiche — raddoppia il consumo del peer provider.

Le review sono persistite: le ritrovi al reload e nell'export; quelle ancorate a messaggi più vecchi della pagina caricata restano raggiungibili in testa al transcript. Se il ⚖️ appare spento, la sessione non ha un peer model: si sceglie alla creazione. Funziona anche in PLAIN — e in PLAIN non consuma crediti MFF: paghi solo il tuo provider del peer.

Implicazione epistemica: se rigeneri una risposta, le review restano agganciate al messaggio a cui si riferivano — una review non migra mai su un testo che non ha mai letto.

1 2

🗂️ Carta dello stato e memoria trasferibile

La sessione raccoglie in una carta 🗺️ ciò che risulta stabilito e a che titolo, estratto dalle etichette già dichiarate — non un riassunto scritto da un modello. Ogni affermazione porta il suo dichiarante: dopo un failover leggi CHI ha davvero dichiarato cosa, mai un attribuzione di comodo.

Serve a due cose: guardarla, e portarla altrove. «Copia come memoria» produce un testo da allegare a una sessione nuova — magari con un altro modello — senza trascinare il transcript intero. Chi la riceve legge che quelle etichette le ha dichiarate un altro: affermazioni da discutere, non verdetti da ereditare. Questo è anche il modo corretto di «cambiare modello»: non a sessione viva, ma con la carta che viaggia.

13

Crediti MFF

> I crediti MFF non pagano i token AI: quelli si pagano al provider con la tua chiave (BYOK). I crediti sono il costo separato del framework epistemico — e si spendono SOLO quando viene applicato davvero.

SCENARIO	CREDITI
Modalità MFF, framework applicato nella risposta	1 credito (riservato a inizio turno, confermato a fine)
Modalità MFF, framework NON applicato	0 — il credito riservato torna indietro
Errore del provider a metà risposta	0 — rimborso automatico
Modalità PLAIN, sempre	0
Ricerca web e peer review su chiavi tue	0 crediti MFF — solo il costo del tuo provider

Welcome bonus

Ogni nuovo account riceve 500 crediti in modo progressivo: 50 all'iscrizione, +450 alla prima risposta in cui il framework viene applicato davvero. La progressione è un'anti-frode minima; superato il primo messaggio MFF, hai i 500 pieni. Nella beta gratuita i crediti non si comprano; il saldo e ogni movimento sono in Impostazioni → Crediti.

> Se esaurisci i crediti, la modalità MFF si ferma ma PLAIN resta libera e completa: puoi continuare a lavorare a costo zero.

14

💰 Ottimizzare i costi

I modelli addebitano per token (~3-4 caratteri in italiano), in ingresso e in uscita. Ogni risposta in MFF mostra i suoi numeri nel footer tecnico: token, tempi e — quando il listino è noto — la stima di costo del messaggio e il cumulato della sessione. Se l'usage del provider non arriva, la stima è marcata con ~, mai spacciata per dato.

1. Usa i :free per bozze ed esplorazione (compromessi: contesto ridotto, niente 🌐).
2. Sessioni corte + carta trasferita battono le conversazioni chilometriche: la storia oltre il tetto viene comunque troncata dal motore — la carta è la memoria che viaggia senza pagare contesto.
3. Reasoning su Basso finché il compito non chiede di più: i token di pensiero si pagano come gli altri.
4. Whitelist 🎯 sulla chiave: nessun automatismo potrà scegliere un modello costoso al posto tuo.
5. Fallback economici nelle Avanzate: il failover rispetta la tua lista.
6. PLAIN per il casual: zero crediti, stesse funzioni.

> ℹ️ Garanzia di piattaforma: le tutele tecniche (cifratura chiavi, breaker, failover dichiarato, rate limit) sono identiche per utenti free e paganti.

15

⚙️ Impostazioni e diagnostica

La pagina Impostazioni ha due viste, selezionabili dal menu in testa e sincronizzate tra i tuoi dispositivi: «Utente base» mostra l'essenziale (profilo, chiavi, saldo, voce, privacy); «Utente esperto» aggiunge diagnostica, preferenze AI tecniche, whitelist 🎯 e storico movimenti. Se questo vademecum cita qualcosa che non vedi, quasi certamente sta nella vista Esperto.

La Diagnostica (vista Esperto): nessun numero inventato

- Probe BYOK: una query di test VERA su ogni tua chiave, con esito e latenza.
- Pannello immagini: dichiara QUALE delle tue chiavi genererebbe un'immagine ora, e con che modello.
- Salute modelli: telemetria di ciò che hai usato davvero (tempo di prima risposta, errori, 429).
- Censimento campi: ciò che le API dei provider espongono e MFF non traduce ancora. Un dato assente è mostrato come assente, mai spacciato per zero.

Cosa fare quando qualcosa è rosso

- Chiave «non valida»: revocata o scaduta sul sito del provider — rigenerala lì e reinseriscila. MFF non può ripararla.
- Chiave «in pausa»: il breaker l'ha fermata dopo errori ripetuti (tre «non valida» di fila = un quarto d'ora) per non bruciare richieste. Si riattiva da sola, o subito col bottone «Riattiva».

- Probe fallito su una chiave che ieri andava: guarda il codice prima di cancellare — un rate limit o un guasto del provider passano da soli.

16

Installare MFF (PWA)

MFF è una Progressive Web App: funziona nel browser, ma installata dà icona dedicata, schermo intero senza barra URL e avvio istantaneo. Le funzioni AI restano identiche (serve la connessione: non c'è inferenza offline).

Android (Chrome / Edge / Brave)

1. Apri l'app nel browser: appare il banner «Installa app» → Installa. Se non appare: menu ⋮ → «Installa app» / «Aggiungi a schermata Home».

iOS (Safari)

1. Apri l'app in Safari (su iOS le PWA richiedono Safari) → icona Condividi → «Aggiungi alla schermata Home» → Aggiungi.

Desktop (Chrome / Edge / Brave)

1. Nella barra indirizzi a destra appare l'icona di installazione (monitor con freccia) → Installa: l'app si apre in finestra dedicata.

17

🔧 Risoluzione problemi

- Errore 429 su un modello :free → è la quota giornaliera del provider, non un guasto: aspetta qualche minuto, prova un altro :free, o usa un modello a pagamento con la tua chiave.
- Risposta interrotta a metà → «▸ Continua» sotto il messaggio riprende da dove si era fermata. Se il provider è caduto, il failover dichiarato ha già provato le riserve; l'eventuale errore resta visibile e il credito è rimborsato.
- Risposta lenta → i modelli con reasoning esteso sono lenti per costruzione: per compiti semplici usa un modello fast (Groq, Cerebras) o abbassa lo sforzo di reasoning nelle Avanzate.
- Chiave rossa o in pausa → vedi «Impostazioni e diagnostica»: rigenera dal provider se non valida; aspetta o «Riattiva» se in pausa.
- La sessione non si crea → il messaggio sotto il bottone dice il perché (modello non disponibile, crediti esauriti in MFF, sessione scaduta). Ricarica la pagina e riprova; in PLAIN i crediti non servono.
- Interfaccia con elementi vecchi → aggiorna la PWA: hard reload (Ctrl+Shift+R / Cmd+Shift+R); da installata, chiudi e riapri l'app.

> ℹ️ Niente di tutto questo risolve? Scrivi al supporto indicando: cosa stavi facendo, modello e provider, e l'eventuale codice di errore mostrato. Più contesto = fix più veloce.

18

? Domande frequenti

MFF è gratuito?

Sì: beta pubblica gratuita, donationware. Porti la tua chiave AI (BYOK); MFF non fa pagare l'uso dell'IA.

Posso usare MFF senza chiavi API?

Per chiamare un modello serve una chiave (anche solo OpenRouter coi modelli gratuiti, senza deposito). I crediti MFF sono un sistema separato: pagano il framework, non i token.

MFF vede le mie conversazioni?

I messaggi transitano dal server MFF verso il provider AI che hai scelto, e le sessioni sono salvate per te nel tuo account; MFF non le analizza né le condivide, e le chiavi sono cifrate. Il provider applica la SUA privacy policy: leggila se il tema è sensibile.

Posso cambiare modello a metà sessione?

No, ed è una scelta: le etichette appartengono al modello che le ha dichiarate. Il flusso corretto è la Carta dello stato: copiala come memoria e allegala a una sessione nuova col modello che vuoi.

Cosa succede se esaurisco i crediti?

La modalità MFF si ferma; PLAIN resta libera e completa. L'app ti avvisa quando il saldo è basso.

Quanto contesto «ricorda» il modello?

Dipende dal modello, e MFF applica comunque un tetto alla storia inviata per proteggere i costi: le sessioni focalizzate rendono meglio. Per la memoria di lungo corso usa la Carta dello stato.

Posso scaricare o cancellare i miei dati?

Sì, entrambi da Impostazioni → area Privacy: export completo dei tuoi dati, ed eliminazione dell'account — immediata, totale e irreversibile, protetta da conferma scritta.

Un modello risponde male: che faccio?

In ordine: riformula in modo più specifico · aggiungi scudi (L3 falsificazione, L5 reasoning) · chiedi una ⚖ peer review · rigenera · prova un modello diverso · apri una sessione nuova se il contesto è «inquinato».

Che differenza c'è tra il sito e l'app?

Il sito spiega il framework e genera il prompt da incollare in qualsiasi chat AI. L'app applica il framework automaticamente a ogni messaggio, con router multi-provider, scudi selezionabili, cronologia e tutto ciò che descrive questo vademecum.

MFF sostituisce il fact-checking?

No. Rende visibile l'incertezza e ti dice cosa e dove verificare (blocchi ZVE), ma la verifica finale sui temi critici resta tua. È una bussola dell'affidabilità, non un oracolo.

19

📖 Dizionario dei termini

Ogni termine tecnico del framework, spiegato in lingua semplice. Fuso dal glossario della guida pubblica (74 voci in 7 categorie).

🤖 Termini sull'AI in generale

AI / Intelligenza Artificiale
Un programma capace di generare risposte simili a quelle umane analizzando enormi quantità di testo. *Non "pensa" come noi*: predice la parola più probabile, una dopo l'altra.

LLM
Large Language Model. Il "cervello" tecnico dietro ChatGPT, Claude, Gemini. Letteralmente: *"modello linguistico di grandi dimensioni"*.

Prompt
L'istruzione che dai all'AI. **Tutto quello che scrivi nella chat è un prompt**.

Token
L'unità minima di testo che l'AI elabora. Una parola lunga può essere 2-3 token. Le AI hanno un limite di token per conversazione.

Contesto
La memoria della conversazione corrente. Quando l'AI "dimentica" cose dette prima, è perché ha superato il suo limite di contesto.

Multimodale
Un'AI che capisce non solo testo, ma anche immagini, audio, video, documenti.

Allucinazione
Quando l'AI inventa informazioni che sembrano reali ma non lo sono. *Come sognare ad occhi aperti credendoci.*

Bias

Distorsione sistematica nelle risposte. L'AI può ereditare pregiudizi dai dati con cui è stata addestrata.

Training / Addestramento

La fase in cui l'AI "studia" miliardi di testi. Avviene una sola volta. **Per questo l'AI può non sapere cose recenti.**

Cut-off

La data limite oltre cui l'AI non sa nulla. Esempio: cut-off gennaio 2024 = non sa nulla di eventi post-gennaio 2024.

🎓 Termini filosofici e scientifici

Epistemico / Epistemologia

Dal greco *"episteme"* = conoscenza. Riguarda **quanto e come sappiamo qualcosa**. "Livello epistemico" = grado di sicurezza con cui affermi qualcosa. Esempio: "so che 2+2=4" (alto) vs "penso che domani piova" (basso).

Esegesi

Interpretazione critica di un testo. Non ripetere le parole, **ma spiegarne il senso vero**. Esempio: l'esegesi di una clausola contrattuale ne spiega le conseguenze pratiche.

Inferenza

Conclusione tratta da informazioni note. *"Il pavimento è bagnato → inferisco che è piovuto"*. Non è certezza, è deduzione probabile.

Speculazione

Ipotesi non dimostrata, basata su ragionamento ma non su prove. *"Forse l'azienda fallirà l'anno prossimo"*.

Stima

Calcolo approssimativo basato su dati parziali. **A metà strada tra certezza e speculazione.**

Eziologia
Lo studio delle cause di un fenomeno (soprattutto in medicina). *"Quale è la causa di questa malattia?"*

Validazione
Processo di verifica che conferma se qualcosa è corretto, attraverso prove o controlli.

Fonte primaria
La fonte originale: lo studio scientifico, la legge, il documento ufficiale. **Non una sintesi di terzi.**

Fonte secondaria
Una rielaborazione o sintesi di fonti primarie. *Wikipedia è una fonte secondaria.*

Cross-check
Confronto incrociato tra più fonti per verificare la coerenza dell'informazione.

Peer review
Revisione scientifica fatta da pari, esperti dello stesso campo. **Lo standard di qualità della ricerca seria.**

🛡️ Termini specifici del Framework

Framework
Insieme strutturato di regole. *Come una ricetta: non è il piatto, è il metodo per farlo bene.*

MFF
MarcoFLY Framework. Il nome del protocollo.

MFF-EX
EXecutable. La parte "operativa" del framework — **le istruzioni che l'AI esegue**.

MFF-EL

Epistemic Layer. Lo "strato epistemico" — **le regole su come gestire la certezza**.

APEX

Versione completa e ottimizzata del framework. *Come la "Pro" di un software*.

L1 → L7

I sette livelli (Layers) di difesa contro errori e allucinazioni. **L1 sempre attivo, gli altri opzionali.**

ZVE

Zona di Verifica Esterna. Il riquadro che l'AI crea per segnalarti **cosa va verificato altrove**.

PAVA

Il protocollo anti-degrado (regola R17): impedisce all'AI di "tagliare gli angoli" nei task e nelle conversazioni lunghe — checksum, niente troncamenti, validazione delle differenze.

Deriva Epistemica

Quando l'AI *"scivola"* gradualmente da fatti certi a supposizioni, senza accorgersene.

Dominio

L'ambito tematico della conversazione (medico, legale, scientifico, quotidiano...). **Il framework si adatta al dominio.**

Modulo / Layer / Livello

Sinonimi nel framework. Indicano i singoli "scudi" L1-L7.

Scudo

Termine evocativo per *"modulo difensivo"*. Ogni layer è uno scudo contro un tipo di errore.

VMS

Validazione Multi-Sorgente. Il modulo L7 che **confronta più fonti**.

Citazione Forzata
La regola L1: l'AI deve sempre dichiarare da dove viene un'informazione.

Esegesi Contestuale
La Fase X della pipeline: **separare i fatti dalle interpretazioni** dell'AI stessa. (L4 è invece il grounding web.)

MWAL
MarcoFLY Web Access Layer. Il modulo che porta la ricerca web dentro la chat, quando serve una fonte in tempo reale.

📋 I sei livelli di certezza (etichette MFF-EL)

🟢 **SICURO**
Fonte diretta, documento ufficiale, **nessun dubbio**. L'AI ha prove concrete in sessione (URL, log, documento verificato). Quello che leggi è reale.

🔵 **PROBABILE**
Logica solida, certezza non garantita. Il modello è fiducioso — ma la fonte esatta non è in sessione. *Deduzione ben strutturata, non un fatto provato.*

🟡 **FORSE**
Un'ipotesi, non una risposta. Il modello stima, non sa. Margine di errore significativo. **Utile come punto di partenza — pericoloso come punto di arrivo.**

🟠 **DIPENDE**
Vero solo a certe condizioni. **Leggi le premesse prima di agire.** L'affermazione regge — ma solo se le variabili in gioco si verificano.

🔴 **NON SO**
La risposta più onesta che un'AI possa darti. Dati insufficienti, fonti contraddittorie. *Invece di inventare, MFF si ferma.* Un'allucinazione confezionata bene è più pericolosa del silenzio.

⚫ NON POSSO

Non è reticenza, è **onestà strutturale**. La richiesta va oltre il cutoff, fuori dal set di addestramento o contro le policy di sicurezza. L'AI non improvvisa, non aggira, non inventa. Si ferma.

🏛️ Standard internazionali citati

NIST-RMF

National Institute of Standards and Technology — Risk Management Framework. L'ente americano degli standard scientifici. **Il loro framework è la guida globale per usare l'AI in sicurezza.**

TrAM

Trustworthiness Assessment Model. Modello scientifico (Schlicker et al. 2025) per valutare l'affidabilità dei sistemi automatici. *Spiega la differenza tra fidarsi a priori e fidarsi dopo verifica.*

AT / PT (nel TrAM)

Antecedent Trust / Post Trust. **AT** = la fiducia che hai prima di usare uno strumento. **PT** = la fiducia che hai dopo averlo usato e verificato.

OSF / PsyArXiv

Repository scientifici open-access dove vengono pubblicate ricerche. *Sono la "biblioteca pubblica" della scienza moderna.*

💻 Termini tecnici minori

Open Source

Codice pubblico, ispezionabile e modificabile da chiunque.

CC BY-NC-ND 4.0

La licenza con cui è distribuito il sito: consultabile gratuitamente, citando l'autore. **Non commerciale, non modificabile.**

PWA

Progressive Web App. *Il sito si comporta come un'app: si può "installare" sul telefono.*

Repository / Repo
Cartella pubblica online (es. GitHub) dove vive il codice di un progetto.

BYOK
Bring Your Own Key. Usi la tua chiave API personale, cifrata lato server e mai esposta: le sessioni restano nel tuo account, cancellabili in ogni momento.

SSE
Server-Sent Events. Lo streaming del testo in tempo reale: le parole arrivano mentre vengono scritte.

ORCID
Identificatore univoco per ricercatori — la carta d'identità del mondo scientifico, riconosciuta da università e riviste.

🏷️ Abbreviazioni e tag che usa il Framework

📎 Suffissi sulla fonte (appaiono dopo l'etichetta colorata)

[doc]
Documento ufficiale (legge, contratto, paper, manuale, atto pubblico).

[test]
Testimonianza diretta (intervista, dichiarazione, deposizione).

[comm]
Comunicato ufficiale di ente, azienda o istituzione.

[log]
Log o registro tecnico (eventi sistema, audit trail, monitoring).

[emp]
Dato empirico (esperimento controllato, misurazione, osservazione diretta).

[lit]
Letteratura scientifica (paper peer-reviewed, libro accademico, review).

[leg]
Fonte legale o normativa (sentenza, decreto, regolamento, direttiva).

[cert]
Certificazione formale verificabile (attestato, ISO, ente certificatore terzo).

🧠 Memoria e calibrazione

[mem]
Memoria non verificata: l'AI ricorda l'informazione dal training, ma non c'è prova in sessione corrente. *Sempre da verificare.*

[p:xx%]
Probabilità soggettiva stimata dal modello. Esempio: `[p:75%]` = 75% di confidenza interna.

⚙️ Tag delle fasi della pipeline

[FASE E]
Fase Epistemica: l'AI **produce** la risposta e la classifica con etichette di certezza.

[FASE X]
Fase Esegetica: l'AI **interpreta** e spiega quanto prodotto, separando fatti da inferenze.

🛡️ Tag dei moduli (livelli di difesa)

[L1]...[L7]
Riferimento al livello di difesa attivato per la singola affermazione.

[ATTIVO] / [SEMPRE ATTIVO]
Stato del modulo nella sessione corrente. **L1 è sempre attivo**, gli altri secondo configurazione.

[SIM] / [EXT]

Modalità del modulo VMS. **SIM** = simulato (l'AI valida autonomamente). **EXT** = esterno (richiede input dall'utente).

📦 Blocchi speciali nelle risposte

── ZVE ──

Zona di Verifica Esterna. Riquadro che elenca **cosa controllare fuori dalla chat** e con quale urgenza.

── VMS ──

Validazione Multi-Sorgente. Riquadro che mostra **il confronto tra fonti diverse** con score di convergenza.

⚠️ **DRIFT ALERT L6**

Allerta deriva epistemica: l'AI sta scivolando da fatti certi verso supposizioni. **Si ferma e ricalibra.**

Carta di Stato

Riepilogo periodico della sessione: progetto, dominio, livelli attivi, tesi principale, validazioni esterne.

IN ATTESA · R1–R20

IN ATTESA = etichetta sospesa fino a verifica esterna. **R1–R20** = regole operative interne (es. R12: nessun 🟢 SICURO senza fonte).