HANDBOOK

MarcoFLY Framework

The complete handbook of the app and the framework: every choice, what it does, and what it implies.

Marco Motta (MarcoFLY)

marcofly.app · 13 August 2026

About this document

This handbook is generated from the same source as the guide published at marcofly.app/guide and the one inside the app: there is no separate text kept in sync by hand. It is a snapshot — the date on the cover says when it was composed.

marcofly.app/guide

Contents

01

👋 Welcome

MFF (MarcoFLY Framework) is an epistemic control layer on top of generative AI: it makes visible how confident a model is about what it says, statement by statement, and helps you verify it. This app applies the framework automatically to every message, across dozens of models from different providers, with YOUR API keys (BYOK).

The two modes: "MFF Framework" injects the epistemic protocol (labels, shields, state card) and consumes MFF credits only when the framework is actually applied. "Simple chat" (PLAIN) is free conversation without the framework and without credits — but with ALL platform features: attachments, web search, peer review, read-aloud, export.

What MFF is NOT

- It does not guarantee truth: labels reflect what the model declares to know; the final check stays with you.
- It does not replace experts: medical, legal, tax or financial decisions need a professional.
- It does not zero out hallucinations: it reduces them and makes them recognisable, but the model can be wrong even with a green badge.

02

🚀 Quick start — 3 steps

1 Set up an API key (or use the included MFF credits). Go to Settings → AI provider keys: paste a key — the fastest path is "Connect OpenRouter" via OAuth, one key for hundreds of models, free ones included.

2 Create a session from the Dashboard: choose mode (MFF or PLAIN), domain, model — and the defense modules if you want. Every choice is explained in the sections of this vademecum.

3 Write your question: the answer streams in, with labels. Use 📎 to attach images and documents, 🌐 to request a web search before the answer. Every message has 📋 copy, 🔊 read-aloud, ⚖️ peer review and regeneration.

> 💡 For your first tries use a :free model — perfect for learning the interface. Then move to a paid model with your key for full context and web search.

03

🔓 Free keys, provider by provider

Each chapter below is linkable (🔗 icon): from the PWA Settings you land straight on the provider you are configuring. Reading key: 🆓 start without a card · 🌱 minimal entry cost · 💳 credit required — declared, never hidden.

> 💡 If it is your first day: do ONLY the first chapter (OpenRouter, 2 minutes with the automation). With that single account you immediately get several free LLMs and can use the whole app; add the rest when you need it.

OpenRouter FREE

`openrouter.ai` · `sk-or-…`

The recommended entry door: ONE key, hundreds of models — and the :free ones work WITHOUT depositing anything.

1. In Settings → AI provider keys press "Connect OpenRouter": it is the MFF automation — the OpenRouter site opens, you log in (or create the account in 30 seconds), authorize, and the key reaches MFF by itself. No copy-paste, no typos.
2. From that moment, with the SAME account, you immediately get several free LLMs (models with the :free suffix) to start using the system: create a session, pick OpenRouter, and the picker shows the available :free models.
3. Manual alternative: openrouter.ai → Keys → "Create key" → paste into MFF. If one day you want paid models, add credit to YOUR OpenRouter account: MFF has nothing to do with the payment.

Google AI Studio FREE

`aistudio.google.com` · `AIza…`

Generous free tier on Gemini models: often you don't even need a card to start.

1. Go to aistudio.google.com with your Google account → "Get API key" → "Create API key".
2. Copy the key (starts with AIza) and paste it in Settings → Google AI card: MFF validates it with a test query and encrypts it.

Groq FREE

`console.groq.com` · `gsk_…`

Free tier with monthly limits: 5-10× the response speed of classic providers — great for tests and quick tasks.

1. console.groq.com → sign up → API Keys → "Create API Key".
2. Copy the gsk_… key and paste it into MFF's Groq card.

Cerebras FREE

`cloud.cerebras.ai` · `csk-…`

Very generous free tier on wafer-scale chips: open models (Llama, Qwen) at very high speed.

1. cloud.cerebras.ai → free account → API Keys → create the csk-… key
2. Paste it into MFF's Cerebras card.

NVIDIA NIM FREE

`build.nvidia.com` · `nvapi-…`

Permanent free tier (40 requests/minute): a wide catalog of open models served by NVIDIA.

1. build.nvidia.com → account (free) → API Catalog → “Get API Key”.
2. Paste the nvapi-… key into the NVIDIA NIM card.

SambaNova FREE

`cloud.sambanova.ai` ·

Free initial credit + a persistent tier (20 requests/min): good for fast open models.

1. cloud.sambanova.ai → sign up → API Keys → create the key.
2. Paste it into MFF’s SambaNova card.

Cloudflare Workers AI FREE

`dash.cloudflare.com` ·

Free tier built into every Cloudflare account: open models on the CF edge.

1. dash.cloudflare.com → My Profile → API Tokens → create a token with the “Workers AI Read” permission.
2. In MFF the Cloudflare card asks for TWO values: the token and your Account ID (found on the CF dashboard home).

Mistral FREE

`console.mistral.ai` ·

Free tier for experimenting with Mistral models; pay-per-use plans for serious use.

1. console.mistral.ai → sign up → API Keys → create the key.
2. Paste it into MFF's Mistral card.

Cohere FREE

`dashboard.cohere.com` ·

Generous free trial key (rate-limited); the production key is separate.

1. dashboard.cohere.com → account → API Keys → use the Trial key.
2. Paste it into MFF's Cohere card.

Hyperbolic LOW

`app.hyperbolic.xyz` ·

Small free initial credit, then very low prices (down to ~10× below classic price lists) on open models.

1. app.hyperbolic.xyz → sign up → Settings → API Key.
2. Paste it into MFF's Hyperbolic card.

OrcaRouter LOW

`orcarouter.ai` · `sk-orca-…`

Zero-markup aggregator with its own independent key: you pay vendor list prices with no surcharge. It does not use your OpenRouter credit.

1. orcarouter.ai → account → API Keys → create the sk-orca-… key
2. Paste it into MFF's OrcaRouter card.

DeepSeek PAID

`platform.deepseek.com` · `sk-…`

No free tier, but the entry cost is among the lowest (minimum top-up of a few dollars) and prices are very aggressive.

1. platform.deepseek.com → account → minimum top-up → API Keys.
2. Paste the sk-… key into MFF's DeepSeek card.

OpenAI PAID

`platform.openai.com` · `sk-…`

No API free tier: a minimum of credit on the account is required. (GPT models can also be tried free via OpenRouter :free when available.)

1. platform.openai.com → API Keys → "Create new secret key" → add minimum credit in Billing.
2. Paste the sk-… key into MFF's OpenAI card.

Anthropic PAID

`console.anthropic.com` · `sk-ant-…`

No API free tier: minimum credit in Billing. In return: Claude models with native web search and extended reasoning.

1. console.anthropic.com → API Keys → "Create Key" → top up in Billing.
2. Paste the sk-ant-… key into MFF's Anthropic card.

xAI (Grok) PAID

`console.x.ai` · `xai-…`

Requires initial credit on the account.

1. console.x.ai → API Keys → create the xai-… key → add credit.

2. Paste it into MFF's xAI card.

Perplexity PAID

`perplexity.ai/settings/api` · `pplx-…`

Credit required, but Sonar models have NATIVE web search: citations appear directly among the sources in MFF.

1. perplexity.ai/settings/api → generate the pplx-… API key → top up.
2. Paste it into MFF's Perplexity card.

Z.AI (GLM) PAID

`z.ai` ·

Zhipu's GLM models; account credit required.

1. z.ai → account → API Keys → create the key.
2. Paste it into MFF's Z.AI card.

04

🏷️ The epistemic labels

Every statement carries its declared commitment level. They are not decoration: they are a contract — and they persist in exports. The golden rule: the tag PRECEDES the content, one tag per block.

🟢 CERTAIN

Direct source in session, no doubt. Without a citable source this label is FORBIDDEN: it self-downgrades to PROBABLE.

🔵 PROBABLE

Solid deduction, but the exact source is not in session. High confidence, not a proven fact.

🟡 MAYBE

Estimate with a significant margin. A good starting point, dangerous as a final answer.

🟠 DEPENDS

True only under certain conditions: read the premises before acting.

🔴 UNKNOWN

Insufficient or contradictory data. Instead of inventing, MFF stops.

⚫ CANNOT

Outside training or past the cutoff. Structural honesty, not reticence.

🌐 CONFIRMED:WEB

Allowed ONLY if a real web search happened in that turn, with the source cited.

🌐 VERIFIED:WEB

The search corroborates without exclusive confirmation. Same strict rules as its twin.

Suffixes: where it comes from and how confident

After the label, source suffixes may appear: [doc] official document · [test] testimony · [comm] official statement · [log] technical log · [emp] empirical data · [lit] scientific literature · [leg] legal source · [cert] verifiable certification. Then [mem] = training memory not verified in session (always to be checked) and [p:75%] = the model's estimated subjective probability. The glossary at the bottom explains them all.

05

🛡️ The shields — defense modules

In MFF mode you can enable up to 9 epistemic modules that steer the model's behaviour. L1 is always on; the default set L1+L2+L3 is great for most tasks. The cards below are the SAME ones you see in the session creation wizard.

📌 L1 · Forced Citation

Always active and cannot be disabled. Every model assertion must have an explicit epistemic basis: the model cannot produce factual claims without a verifiable source or reasoning chain.

🔎 L2 · EVZ — External Verification Zone

Creates an External Verification Zone: the model actively distinguishes certain knowledge from content requiring independent verification, flagging potentially inaccurate claims.

⚖️ L3 · Falsification — failure conditions

Applies the Popperian falsifiability principle: the model must explicitly identify the conditions that could refute its claims.

🌐 L4 · Web Grounding — mandatory web search

Activates real-time web search before responding on topics requiring current data. Drastically reduces hallucinations about recent events or time-sensitive information. The 🌐 CONFIRMED:WEB / 🌐 VERIFIED:WEB labels are allowed only if a grounding block is actually injected in this turn (v1.5.3 anti-hallucination).

L4 is activated via the 🌐 button in chat and works universally through MWAL (Tavily on MFF side or BYOK Serper/Tavily/SerpAPI/Brave). Providers with native web_search (Anthropic, Perplexity, OpenRouter paid) use it directly. Disabled on :free models for cost gating.

🧠 L5 · Peer Review — extended reasoning

Simulates academic peer review and enables extended reasoning tokens on supported models (DeepSeek-R1, Claude, Qwen Thinking, etc.). Increases accuracy and reduces errors.

Configure the reasoning effort (low / medium / high / max) in Advanced settings to balance response cost and quality.

🔄 L6 · Epistemic Drift — anti-hallucination

Monitors and corrects cognitive drift during the conversation: detects when later responses contradict earlier ones or deviate from the original context.

✅ L7 · MSV — multi-source validation

Multi-Source Validation: you choose the second source by selecting a peer model from a different provider. MFF automatically routes the same question to the second model and produces a —— MSV —— block with CONVERGENCE HIGH/MEDIUM/LOW.

🛡️ PAVA · PAVA Protocol — anti-amnesia

Shield Protocol. Maintains a structured log of conversational context to prevent coherence loss in long sessions. The model is periodically re-anchored to the original session context and objectives.

⚙️ NIST · NIST-RMF — risk management

Shield Protocol. Aligns the response process with the NIST Risk Management Framework, classifying and managing epistemic risk at every stage of the response.

> 💡 Add L4 for recent facts (then used via the 🌐 button in chat), L5 for complex reasoning, L7 if you configured a peer model. PAVA and NIST are for very long sessions or critical domains. Implication: more shields = more rigor, but longer answers and more tokens.

> ⚠️ On :free models the costly shields (L4 search, L5 extended reasoning) are disabled for cost containment.

06

💬 Sessions: every choice and what it implies

A session is a complete conversation with one model: it has its own title, model, mode, domain, language and shields, set at creation. Every wizard choice changes something concrete in the engine.

The domain does two things

First: it defines what 🟢 / 🔵 / 🟡 mean in that context — in the scientific domain "CERTAIN" means replicated, published results; in the creative domain the bar differs. Second: it sets the default temperature, from 0.2 (Legal: conservative, repeatable) to 1.0 (Creative: maximum variability). Advanced settings can override it.

The operating mode

MFF-E: rigorous epistemic analysis · MFF-G: cuts noise and redundancy · MFF-X: interprets and explains without altering · MFF-EX: combines them (the central mode) · MFF-EGX: adds the discipline of synthesis · AUTO: picks per turn. PLAIN: no framework, no credits.

The session language is a choice, and it stays

It is separate from the interface language. Switching the UI language does NOT translate answers already generated: the epistemic label belongs to the text that model produced — translating under the labels would falsify them.

Inside the conversation

- Each new message carries the session context: follow up without repeating. For new topics open a new session (clean context, lower costs).

- Under every answer: 📋 copy · 🌐 sources · 🔉 read-aloud · ⚖ peer review · 🔄 regenerate. Your own messages can be edited and resent.
- "▸ Continue" resumes an answer truncated by the budget; in sessions beyond 100 messages "⤒ Load earlier messages" walks back through history.
- All sessions live in the Archive: reopen and continue where you left off. From there you can also delete them, one by one.

07

🧠 Choosing the model

The picker always shows the provider's full lineup, marking which models are actually served right now: the truth comes from the live catalog, not a hand-written list. An unserved model is visible and marked, not hidden.

Capabilities come from the catalog too: vision → image attachments enabled; reasoning → extended thinking with L5; image generation → the model joins the generation routes. The model name decides nothing: what the provider declares does.

Free versus paid

:free models cost nothing, with three trade-offs: tighter context and document budget, no 🌐 web search, and provider daily quotas (a 429 on a :free often just means "retry later or switch :free"). Paid models run on YOUR key: you pay the provider's list price, MFF adds nothing.

If the provider fails: declared failover

On recoverable errors (rate limit, overload, retired model) the engine tries the reserves: first the fallbacks you set in advanced settings (up to 3, in order), then the policy chain across your keyed providers. Everything is DECLARED: answer, labels and card report who actually served. The turn restarts from scratch on the reserve model — what you read is always from a single model. Non-recoverable errors (invalid key, exhausted balance) do not fail over: they are shown, and the MFF credit comes back.

> 💡 Choice summary: generic tasks → a fast, cheap mid-tier · deep analysis → a reasoning model (slower, pricier) · recent facts → paid model + 🌐 · drafts and experiments → :free.

08

🔌 Providers and routing

MFF talks to 17 providers, in two families. YOU choose the session provider at creation — there is no hidden routing: what you choose is what answers, and any fallback is declared (see failover).

Aggregators: one key, many vendors

- OpenRouter (sk-or-… key): hundreds of models, :free included; also connects in one click via OAuth. On paid models it uses its native web_search. Small markup (~5%) versus direct.
- OrcaRouter (sk-orca-… key): zero-markup aggregator with its own independent key — it does not use your OpenRouter credit.

Direct: the shortest path and native capabilities

OpenAI · Anthropic · Google · Groq · Cerebras · Mistral · DeepSeek · xAI · Perplexity · Z.AI · NVIDIA NIM · SambaNova · Hyperbolic · Cohere · Cloudflare Workers AI. No markup, and the vendor's native features: Anthropic's built-in web search, Perplexity's native citations (shown among sources), the generous free tiers of Groq, Cerebras, NIM and SambaNova.

> 💡 Minimal setup: one OpenRouter key covers everything (free models included). Add direct keys for the providers you use most: you remove the markup and unlock native capabilities. Peer review, by design, draws from your DIRECT providers: the reviewer is independent from the first model's channel.

Image generation is a platform capability, not a chat-model one: if the session model cannot generate images, the engine routes to the first of your keyed providers that can, and declares it.

09

🔑 Configuring API keys

1. Go to Settings → AI provider keys and open the provider card.
2. Paste the key: MFF validates it IMMEDIATELY with a real test query — a key that fails the test is not saved.
3. The key is encrypted server-side (AES key-ring with rotation) and is never shown in clear again, never put in a URL, never sent to the browser. The model NEVER sees it: requests leave from the server.
4. Each key shows a traffic light (active · invalid · paused) and, in Expert view, the 🎯 allowed-models whitelist: session and failover will never leave the list.

Where to create the key, provider by provider

PROVIDER	WHERE	NOTES
OpenAI	platform.openai.com → API Keys	sk-... · needs minimum credit
Anthropic	console.anthropic.com → API Keys	sk-ant-... · credit in Billing
Google AI Studio	aistudio.google.com → Get API key	AIza... · generous free tier
Groq	console.groq.com → API Keys	gsk_... · free tier, great for testing
Perplexity	perplexity.ai/settings/api	pplx-... · native search with citations
OpenRouter	openrouter.ai → Keys (or OAuth from MFF)	sk-or-... · works without deposit on :free

DeepSeek	platform.deepseek.com → API Keys	sk-... · low minimum top-up
Cerebras	cloud.cerebras.ai → API Keys	csk-... · very generous free tier
xAI (Grok)	console.x.ai → API Keys	xai-... · needs initial credit
Mistral	console.mistral.ai → API Keys	free tier for experiments
Cohere	dashboard.cohere.com → API Keys	generous free trial key
Cloudflare Workers AI	dash.cloudflare.com → API Tokens	Workers AI token + Account ID
NVIDIA NIM	build.nvidia.com → Get API Key	nvapi-... · permanent free tier
SambaNova	cloud.sambanova.ai → API Keys	initial credit + persistent tier
Hyperbolic	app.hyperbolic.xyz → Settings → API Key	very low prices
Z.AI (GLM)	z.ai → API Keys	GLM models

SEARCH keys (for the 🌐)

Separate from model keys: Serper, Tavily, SerpAPI, Brave — in Settings, Web search section. These too are validated with a test query. Elect the default with the ⭐ : the engine tries it first, the others remain reserves in the fixed order Serper → Tavily → SerpAPI → Brave. With your own key you only pay your provider; without one, you use MFF's included quota (limited and shared).

> ⚠️ Never share keys: don’t paste them in chats, emails or documents. If you suspect a compromise, revoke it immediately on the provider’s dashboard — then add the new one in MFF.

10

🌐 Web search and sources

1 Press 🌐 next to the composer: a mini-model extracts the proposed search query and SHOWS it to you. You can fix or rewrite it — the query is your decision, not a blind automatism.

2 The search runs on the MWAL chain: your default key first, then the reserves (Serper → Tavily → SerpAPI → Brave). If an engine fails, the hand-off is declared.

3 Results enter the context as a grounding block: the answer can cite them, and sources stay in the message's 🌐 chip — they survive reloads and end up in exports.

> ⚠️ Anti-hallucination rule: the 🌐 CONFIRMED:WEB and 🌐 VERIFIED:WEB labels are allowed ONLY if a real grounding block was injected in that turn (or the provider has native search: Anthropic, Perplexity). Otherwise the model must declare it has no live access and fall back to 🔵 PROBABLE[mem] — saying "I searched the web" is not evidence.

- Honest limit #1: search snippets are by construction a cache — for minute-by-minute data (quotes, the exact time) the source can be pertinent yet stale.
- Honest limit #2: on :free models the 🌐 is disabled for cost containment.

11

⚖️ Peer review between models

A SECOND model, from a different provider, re-reads the answer and shows you its judgement next to the original. It is protocol level L5-B: convergence of two independent agents can promote a 🔵 PROBABLE to 🟢 CERTAIN; divergence forces 🟡 or 🔴. Two models sharing neither weights nor channel, agreeing on the same statement, are worth more than one.

1. At session creation, in Advanced settings, choose the peer model: the list draws from your keyed DIRECT providers, deliberately different from the session provider.
2. In chat press ⚖ under the answer you want re-read: the review streams in and stays anchored to that message, with its convergence verdict.
3. Alternatively, the auto-trigger (checkbox in Advanced) runs the review on EVERY answer: a mode for critical sessions — it doubles the peer provider's consumption.

Reviews are persisted: you find them on reload and in exports; those anchored to messages older than the loaded page remain reachable at the top of the transcript. If the ⚖ appears dimmed, the session has no peer model: it is chosen at creation. It works in PLAIN too — and in PLAIN it consumes no MFF credits: you only pay your peer provider.

Epistemic implication: if you regenerate an answer, reviews stay anchored to the message they referred to — a review never migrates onto a text it never read.

12

🗂️ State card and transferable memory

The session collects in a 🗂 card what has been established and on what grounds, extracted from the labels already declared — not a summary written by a model. Every statement carries its declarant: after a failover you read WHO actually declared what, never a convenient attribution.

It serves two purposes: reading it, and taking it elsewhere. “Copy as memory” produces a text to attach to a new session — perhaps with another model — without dragging the whole transcript. The recipient reads that those labels were declared by someone else: statements to discuss, not verdicts to inherit. This is also the correct way to “switch model”: not mid-session, but with the card travelling.

13

MFF credits

> MFF credits do not pay for AI tokens: those go to the provider via your key (BYOK). Credits are the separate cost of the epistemic framework — and are spent ONLY when it is actually applied.

SCENARIO	CREDITS
MFF mode, framework applied in the answer	1 credit (reserved at turn start, confirmed at end)
MFF mode, framework NOT applied	0 — the reserved credit comes back
Provider error mid-answer	0 — automatic refund
PLAIN mode, always	0
Web search and peer review on your keys	0 MFF credits — only your provider's cost

Welcome bonus

Every new account receives 500 credits progressively: 50 at signup, +450 at the first answer where the framework is actually applied. The progression is minimal anti-fraud; past your first MFF message, you have the full 500. In the free beta credits cannot be purchased; balance and every movement are in Settings → Credits.

> If you run out of credits, MFF mode stops but PLAIN stays free and full-featured: you can keep working at zero cost.

14

💰 Optimising costs

Models charge per token (~4 characters in English), both input and output. Every answer in MFF shows its numbers in the technical footer: tokens, timings and — when the price list is known — the message cost estimate and the session running total. If the provider's usage never arrives, the estimate is marked with ~, never passed off as data.

1. Use :free models for drafts and exploration (trade-offs: reduced context, no 🌐).
2. Short sessions + a transferred card beat mile-long conversations: history beyond the cap gets truncated by the engine anyway — the card is the memory that travels without paying context.
3. Keep reasoning on Low until the task demands more: thinking tokens are billed like any others.
4. 🎯 whitelist on your key: no automatism can pick an expensive model on your behalf.
5. Cheap fallbacks in Advanced: failover respects your list.
6. PLAIN for casual use: zero credits, same features.

> ℹ️ Platform guarantee: the technical protections (key encryption, breaker, declared failover, rate limits) are identical for free and paying users.

15

⚙️ Settings and diagnostics

The Settings page has two views, selectable from the menu at the top and synced across your devices: “Basic user” shows the essentials (profile, keys, balance, voice, privacy); “Expert user” adds diagnostics, technical AI preferences, 🎯 whitelist and movement history. If this vademecum mentions something you can’t see, it almost certainly lives in the Expert view.

Diagnostics (Expert view): no invented numbers

- BYOK probe: a REAL test query on each of your keys, with outcome and latency.
- Images panel: declares WHICH of your keys would generate an image right now, and with which model.
- Model health: telemetry of what you actually used (time to first token, errors, 429s).
- Field census: what provider APIs expose and MFF does not translate yet. Missing data is shown as missing, never passed off as zero.

What to do when something is red

- “Invalid” key: revoked or expired on the provider’s site — regenerate it there and paste it again. MFF cannot repair it.
- “Paused” key: the breaker stopped it after repeated errors (three “invalid” in a row = a quarter hour) to stop burning requests. It reactivates by itself, or immediately via the “Reactivate” button.
- Failed probe on a key that worked yesterday: look at the code before deleting — a rate limit or a provider outage passes on its own.

16

📱 Installing MFF (PWA)

MFF is a Progressive Web App: it works in the browser, but installed it gives you a dedicated icon, full screen without the URL bar and instant start. AI features stay identical (connection required: there is no offline inference).

Android (Chrome / Edge / Brave)

1. Open the app in the browser: the “Install app” banner appears → Install. If it doesn’t: ⋮ menu → “Install app” / “Add to Home screen”.

iOS (Safari)

1. Open the app in Safari (on iOS PWAs require Safari) → Share icon → “Add to Home Screen” → Add.

Desktop (Chrome / Edge / Brave)

1. On the right of the address bar the install icon appears (monitor with arrow) → Install: the app opens in its own window.

17

🔧 Troubleshooting

- 429 error on a :free model → it is the provider's daily quota, not a failure: wait a few minutes, try another :free, or use a paid model with your key.
- Answer interrupted midway → "▸ Continue" under the message resumes where it stopped. If the provider went down, declared failover already tried the reserves; any error stays visible and the credit is refunded.
- Slow answer → extended-reasoning models are slow by construction: for simple tasks use a fast model (Groq, Cerebras) or lower the reasoning effort in Advanced.
- Red or paused key → see "Settings and diagnostics": regenerate at the provider if invalid; wait or "Reactivate" if paused.
- Session won't create → the message under the button says why (model unavailable, credits exhausted in MFF mode, session expired). Reload and retry; PLAIN needs no credits.
- UI showing stale elements → refresh the PWA: hard reload (Ctrl+Shift+R / Cmd+Shift+R); if installed, close and reopen the app.

> ℹ️ Nothing here helps? Write to support with: what you were doing, model and provider, and any error code shown. More context = faster fix.

18

❓ FAQ

Is MFF free?

Yes: free public beta, donationware. You bring your own AI key (BYOK); MFF doesn't charge for AI usage.

Can I use MFF without API keys?

Calling a model needs a key (even just OpenRouter with free models, no deposit). MFF credits are a separate system: they pay for the framework, not the tokens.

Does MFF see my conversations?

Messages transit through the MFF server to the AI provider you chose, and sessions are stored for you in your account; MFF neither analyses nor shares them, and keys are encrypted. The provider applies ITS privacy policy: read it if the topic is sensitive.

Can I switch model mid-session?

No, by design: labels belong to the model that declared them. The correct flow is the State card: copy it as memory and attach it to a new session with the model you want.

What if I run out of credits?

MFF mode stops; PLAIN stays free and full-featured. The app warns you when the balance is low.

How much context does the model "remember"?

It depends on the model, and MFF caps the sent history anyway to protect costs: focused sessions work best. For long-term memory use the State card.

Can I download or delete my data?

Yes, both from Settings → Privacy area: full export of your data, and account deletion — immediate, total and irreversible, protected by written confirmation.

A model answers poorly: what do I do?

In order: rephrase more specifically · add shields (L3 falsification, L5 reasoning) · request a ⚖️ peer review · regenerate · try a different model · open a new session if the context is “polluted”.

What’s the difference between the site and the app?

The site explains the framework and generates the prompt to paste into any AI chat. The app applies the framework automatically to every message, with a multi-provider router, selectable shields, history and everything this vademecum describes.

Does MFF replace fact-checking?

No. It makes uncertainty visible and tells you what and where to verify (EVZ blocks), but the final check on critical topics stays with you. It is a reliability compass, not an oracle.

19

📖 Glossary

Every technical term of the framework, explained in plain language. Merged from the public guide's glossary (74 entries in 7 categories).

🤖 AI terms in general

AI / Artificial Intelligence

A program capable of generating human-like answers by analyzing huge amounts of text. *It doesn't "think" like us*: it predicts the most likely next word, one after another.

LLM

Large Language Model. The technical "brain" behind ChatGPT, Claude, Gemini.

Prompt

The instruction you give the AI. **Everything you type in the chat is a prompt.**

Token

The minimum unit of text the AI processes. A long word can be 2-3 tokens. AIs have a token limit per conversation.

Context

The memory of the current conversation. When AI "forgets" things said earlier, it's because it has exceeded its context limit.

Multimodal

An AI that understands not just text, but also images, audio, video, documents.

Hallucination

When AI invents information that looks real but isn't. *Like day-dreaming and believing it.*

Bias

Systematic distortion in answers. AI can inherit prejudices from the data it was trained on.

Training

The phase where AI "studies" billions of texts. It happens once. **That's why AI may not know recent things.**

Cut-off

The deadline beyond which AI knows nothing. Example: cut-off January 2024 = no knowledge of post-January 2024 events.

🎓 Philosophical and scientific terms

Epistemic / Epistemology

From Greek *"episteme"* = knowledge. About **how much and how we know something**. "Epistemic level" = how confident you are about a claim. Example: "I know 2+2=4" (high) vs "I think it'll rain tomorrow" (low).

Exegesis

Critical interpretation of a text. Not repeating the words, **but explaining the actual meaning**. Example: the exegesis of a contract clause explains its practical consequences.

Inference

A conclusion drawn from known information. *"The floor is wet → I infer it has rained."* Not certainty, just probable deduction.

Speculation

Unproven hypothesis, based on reasoning but not evidence. *"Maybe the company will fail next year."*

Estimate

Approximate calculation based on partial data. **Halfway between certainty and speculation.**

Etiology

The study of the causes of a phenomenon (especially in medicine). *"What is the cause of this disease?"*

Validation

A verification process that confirms whether something is correct, through evidence or checks.

Primary source

The original source: the scientific study, the law, the official document. **Not a third-party summary.**

Secondary source

A reworking or summary of primary sources. *Wikipedia is a secondary source.*

Cross-check

Cross-comparison between multiple sources to verify the consistency of the information.

Peer review

Scientific review done by peers, experts in the same field. **The quality standard of serious research.**

🛡️ Framework-specific terms

Framework

A structured set of rules. *Like a recipe: it isn't the dish, it's the method to make it well.*

MFF

MarcoFLY Framework. The protocol's name.

MFF-EX

EXecutable. The "operational" part of the framework — **the instructions AI executes**.

MFF-EL

Epistemic Layer. The "epistemic layer" — **the rules on how to handle certainty**.

APEX

Complete and optimized version of the framework. *Like the "Pro" version of a software.*

L1 → L7

The seven Layers of defense against errors and hallucinations. **L1 always on, the others optional.**

ZVE

External Verification Zone. The box AI creates to flag **what should be verified elsewhere**.

PAVA

The anti-degradation protocol (rule R17): it stops the AI from "cutting corners" in long tasks and conversations — checksum, no truncation, diff validation.

Epistemic Drift

When AI *"slides"* gradually from certain facts to assumptions, without noticing.

Domain

The thematic area of the conversation (medical, legal, scientific, daily...). **The framework adapts to the domain.**

Module / Layer / Level

Synonyms in the framework. They refer to the individual L1-L7 "shields".

Shield

Evocative term for *"defensive module"*. Each layer is a shield against a type of error.

VMS

Multi-Source Validation. The L7 module that **compares multiple sources**.

Forced Citation

The L1 rule: AI must always declare where information comes from.

Contextual Exegesis

Phase X of the pipeline: **separate facts from interpretations** made by the AI itself. (L4 is web grounding.)

MWAL

MarcoFLY Web Access Layer. The module that brings web search into the chat when a real-time source is needed.

📋 The six certainty levels (MFF-EL labels)

🟢 CERTAIN

Direct source, official document, **no doubt**. AI has concrete in-session evidence (URL, log, verified document). What you read is real.

🔵 PROBABLE

Solid logic, certainty not guaranteed. The model is confident — but the exact source isn't in session. *Well-structured deduction, not a proven fact.*

🟡 MAYBE

A hypothesis, not an answer. The model estimates, it doesn't know. Significant error margin. **Useful as a starting point — dangerous as a finish line.**

🟠 DEPENDS

True only under certain conditions. **Read the premises before acting.** The statement holds — but only if the variables occur.

🔴 **IDK**

The most honest answer an AI can give you. Insufficient data, contradictory sources. *Instead of inventing, MFF stops*. A well-packaged hallucination is more dangerous than silence.

⚫ **CANNOT**

It's not reticence, it's **structural honesty**. The request goes beyond the cutoff, outside the training set, or against safety policies. AI doesn't improvise, bypass, or invent. It stops.

🏛️ International standards cited

NIST-RMF

National Institute of Standards and Technology — Risk Management Framework. The U.S. agency for scientific standards. **Their framework is the global guide for using AI safely.**

TrAM

Trustworthiness Assessment Model. Scientific model (Schlicker et al. 2025) for evaluating the reliability of automated systems. *It explains the difference between trusting upfront and trusting after verification.*

AT / PT (in TrAM)

Antecedent Trust / Post Trust. **AT** = the trust you have before using a tool. **PT** = the trust you have after using and verifying it.

OSF / PsyArXiv

Open-access scientific repositories where research is published. *The "public library" of modern science.*

💻 Minor technical terms

Open Source

Public code, inspectable and modifiable by anyone.

CC BY-NC-ND 4.0
The license under which the site is distributed: free consultation, citing the author. **Non-commercial, no derivatives.**

PWA
Progressive Web App. *The site behaves like an app: it can be "installed" on the phone.*

Repository / Repo
Public online folder (e.g. GitHub) where a project's code lives.

BYOK
Bring Your Own Key. You use your own personal API key, encrypted server-side and never exposed: sessions stay in your account, deletable at any time.

SSE
Server-Sent Events. Real-time text streaming: the words arrive as they are written.

ORCID
A unique identifier for researchers — the ID card of the scientific world, recognised by universities and journals.

🏷️ Abbreviations and tags used by the Framework

📎 Source suffixes (appear after the colored label)

[doc]
Official document (law, contract, paper, manual, public document).

[test]
Direct testimony (interview, statement, deposition).

[comm]
Official communication from an institution, company, or agency.

[log]
Technical log or registry (system events, audit trail, monitoring).

[emp]

Empirical data (controlled experiment, measurement, direct observation).

[lit]

Scientific literature (peer-reviewed paper, academic book, review).

[leg]

Legal or regulatory source (ruling, decree, regulation, directive).

[cert]

Verifiable formal certification (attestation, ISO, third-party certifier).

🧠 Memory and calibration

[mem]

Unverified memory: AI remembers it from training, but no proof in the current session. *Always to verify*.

[p:xx%]

Subjective probability estimated by the model. Example: `[p:75%]` = 75% internal confidence.

⚙️ Pipeline phase tags

[PHASE E]

Epistemic Phase: AI **produces** the answer and labels it with certainty markers.

[PHASE X]

Exegetical Phase: AI **interprets** and explains, separating facts from inferences.

🛡️ Module tags (defense levels)

[L1]...[L7]

Reference to the defense level activated for the specific claim.

[ACTIVE] / [ALWAYS ACTIVE]

Module status in the current session. **L1 is always active**, others as configured.

[SIM] / [EXT]

VMS module mode. **SIM** = simulated (AI validates autonomously). **EXT** = external (requires user input).

📦 Special blocks in responses

— ZVE —

External Verification Zone. Box that lists **what to check outside the chat** and how urgently.

— VMS —

Multi-Source Validation. Box showing **the comparison between different sources** with convergence score.

⚠️ **DRIFT ALERT L6**

Epistemic drift alert: AI is sliding from facts to assumptions. **It stops and recalibrates.**

State Card

Periodic session summary: project, domain, active levels, main thesis, external validations.

PENDING · R1–R20

PENDING = label suspended pending external verification. **R1–R20** = internal operational rules (e.g. R12: no 🟢 CERTAIN without source).