Start here

What you are setting up

An OhWP Studios beginner guide: connect Claude Desktop to FreeLLMAPI, then enable the Cowork workspace for tools and files. This evaluation walkthrough follows the Windows setup successfully tested on 26 November 2026. Business deployments need an appropriate commercial provider plan.

Claude Desktop > FreeLLMAPI on your PC > your selected model provider

Opus, Sonnet and Haiku become routing slots. Choosing the Opus slot does not turn a free model into Claude Opus. The actual model is the one you select in FreeLLMAPI.

Before you install

1. Use an updated Windows 11 PC for this walkthrough. Cowork also supports eligible Windows 10 systems; use Anthropic's readiness checker if unsure.
2. Aim for at least 25 GB free on C: before starting. This is a practical buffer, not a universal minimum. Our installed build required 16.1 GB free and downloaded a 1.27 GB image; other builds can differ.
3. Keep the laptop plugged in and connected to the internet during workspace installation.
4. If this is a work-managed PC, have IT handle any installation or virtualization restrictions.

Install these two apps

1. Install FreeLLMAPI from its Microsoft Store listing. Open it and keep it running.
2. Install Claude Desktop from Anthropic's download page. For Cowork, use the current Windows package with workspace support; do not choose a legacy installation without Cowork.
3. Follow FreeLLMAPI's Open dashboard action. Its usual local address is http://127.0.0.1:31415. Use the address shown by your installation if its port differs.

FreeLLMAPI — Microsoft Store

Claude Desktop download

Cowork requirements and readiness checker

What “free” means: eligible provider quotas or trials. There is no guaranteed personal allowance of 7.4 billion tokens. Availability, country eligibility and limits vary. Use free/trial access within the provider's terms; enabling paid billing can incur charges.

Get one model working

Start with Google AI Studio

You only need one working provider to get started. Add NVIDIA and Cloudflare later if you want the three-provider configuration on page 5.

Create and add your Google key

1. Open Google AI Studio and sign in with your Google account.
2. Open the API Keys page. Choose Create API key, selecting or creating a project if requested.
3. Copy your key privately. A Gemini chat subscription is separate from API access.
4. In FreeLLMAPI, open Keys > Providers > Add key. Select Google / Google AI Studio, paste your Google key, give it a label and save.
5. Open Playground. Select Gemini 3.5 Flash Lite and send: Reply with SETUP_OK and the result of 2 + 2.
6. Continue only when the response includes SETUP_OK and 4. If that model is missing or unavailable for your account, try an available free Flash/Lite model and use its exact name in the following steps.

Google AI Studio API Keys

Google API-key instructions

Check current model pricing and free-tier eligibility

Use the working model for all four slots initially

1. In FreeLLMAPI, open Keys > Anthropic.
2. Set Default, Opus, Sonnet and Haiku to the model that just passed the Playground test.
3. Click Save mapping. Avoid Auto during first-time setup: a slower or unavailable model can cause the connection check to time out.

Copy the key Claude will use

Open Keys > Unified API key and copy that key. This is the FreeLLMAPI key, not your Google provider key. You will paste the unified key into Claude Desktop on page 3.

> Checkpoint: the provider answered a real prompt, all four slots point to that working model, and you have copied the unified key. A green provider-health indicator alone is not enough.

Keep API keys out of screenshots and shared guides. Your prompts and tool outputs go to whichever upstream provider you select; this setup is not offline AI.

Connect Claude Desktop

Enter these settings once

1. Open Claude Desktop. For a new third-party setup, start from the sign-in screen; you do not need to create an Anthropic account for this configuration.
2. Open the top-left application menu > Help > Troubleshooting > Enable Developer Mode. Confirm and reopen the menu if the app restarts.
3. Open Developer > Configure Third-Party Inference. If asked, create a configuration named FreeLLMAPI.
4. Open Connection, choose Gateway, and enter the values below. Labels can vary slightly by app version.

Field	Value
Inference provider / connection	Gateway
Credential kind	Static API key
Gateway base URL	http://127.0.0.1:31415
Gateway API key	Your FreeLLMAPI unified API key
Gateway auth scheme	x-api-key
Model discovery, if shown	Enabled

Use the Claude base URL shown under FreeLLMAPI > Keys > Anthropic. Do not add /v1 or /v1/messages. Do not paste a provider key into Claude's gateway-key field.

1. Choose Test connection. If a model list is discovered, keep the supplied Claude-style IDs.
2. If manual models are required, add the three entries below under Models or Identity & models. These are gateway aliases, not claims about the underlying model.
3. Choose Apply Changes / Apply locally, then Save & Restart / Relaunch now. On the welcome screen, select the option to use your third-party configuration.

Model ID	Display name	Tier alias
claude-opus-4-5	Opus slot	opus
claude-sonnet-4-5	Sonnet slot	sonnet
claude-haiku-4-5	Haiku slot	haiku

If Identity asks for a display name, use your preferred local name. Leave unrelated advanced settings at their defaults. A work-managed configuration may require IT rather than local edits.

Anthropic: in-app configuration

FreeLLMAPI: Claude Desktop connection

Checkpoint: Claude opens in Gateway mode and its connection check passes. FreeLLMAPI must remain running.

Verify chat and Cowork

A connection check is only the first test

Test an actual conversation

1. Start a new conversation or task in Claude Desktop. Choose the Sonnet slot.
2. Send the first test prompt below. Wait until the response is finished.
3. Repeat with the Opus slot. Both should give the correct marker and arithmetic.

```
Connection test only. Do not use files, tools or connectors.
Reply with DESKTOP_CHAT_OK and the result of 17 + 25.
```

Expected answer: DESKTOP_CHAT_OK 42. In FreeLLMAPI > Analytics, confirm the request succeeded and check which provider/model actually served it.

Let Cowork install its workspace

1. Open Cowork. In versions with a unified home screen, a new task can start Cowork automatically.
2. Allow the initial workspace download and setup to finish. Keep both apps open. This can take several minutes; downloading is not an error.
3. If setup is blocked by storage, follow page 6. If the readiness check reports virtualization problems, open “Turn Windows features on or off”, enable Virtual Machine Platform if required, and restart Windows. Hardware virtualization must also be enabled in firmware; use your PC maker's instructions or ask IT.
4. Once setup is complete, start a fresh task with the workspace prompt below. Starting after installation avoids repeated “workspace still starting” tool attempts.

```
Workspace verification only: use the Bash tool to run
uname -s && printf COWORK_WORKSPACE_OK
Report the actual tool output. Do not read personal files,
use connectors or change settings.
```

Expected tool response: Linux, followed by COWORK_WORKSPACE_OK. Expand the Bash entry and check its Response; the assistant merely saying “ready” is not proof.

Check file creation

Create an empty folder called Cowork Test, add it using Project or folder / Add folder, and ask: “Create cowork-test.txt in this test folder containing COWORK_FILE_OK.” Approve only that folder. Open the resulting file to confirm its contents.

Checkpoint: chat answers correctly, Bash returns real output, and the test file exists. Browser use and external connectors are separate features; add and test those individually when needed.

Add the stronger model options

The configuration verified on our PC

NVIDIA: an evaluation option for the Opus slot

1. Open NVIDIA Build, sign in and complete its developer onboarding.
2. Open API Keys or a model's Get API Key action and generate a key for hosted API access. Check the trial/free allowance available to your account.
3. In FreeLLMAPI > Keys > Providers > Add key, choose NVIDIA / NVIDIA NIM. Paste the NVIDIA key and save.
4. In Playground, select Nemotron-3 Ultra 550B (NV). Test a short prompt and a small coding task. Continue only if complete answers arrive.

NVIDIA Build

NVIDIA API quickstart

Cloudflare: for the Haiku connection probe

1. Sign in to Cloudflare. Open Workers AI > Use REST API.
2. Choose Create a Workers AI API Token, review its scoped permissions and create it. Copy both the API token and your Account ID.
3. In FreeLLMAPI > Keys > Providers > Add key, choose Cloudflare. Put the Account ID and API token into their separate fields, then save.
4. In Playground, select Llama 3.3 70B fp8-fast (CF) and test a short prompt.

Cloudflare: obtain the token and Account ID

Save the final mapping

FreeLLMAPI slot	Select this tested option
Default	Gemini 3.5 Flash Lite
Opus	Nemotron-3 Ultra 550B (NV)
Sonnet	Gemini 3.5 Flash Lite
Haiku	Llama 3.3 70B fp8-fast (CF)

Go to Keys > Anthropic, set these four choices and click Save mapping. Repeat the page-4 chat tests for all three slots. These are tested choices as of 26 November 2026, not a permanent ranking or an availability guarantee.

> If you cannot access NVIDIA or Cloudflare, keep using your tested Google model for those slots. You do not need every provider in FreeLLMAPI. Choose the strongest option that passes actual chat and tool tests on your account.

Our test rejected Qwen via OVH after a 429 rate limit. GLM 5.3 and Kimi timed out. NVIDIA hosted trial access is for testing, not production. For client business use, choose a commercial endpoint whose terms permit the intended work; retest compatibility.

NVIDIA API trial terms: sections 1.2 and 1.4

Fix the common errors

Match the symptom before changing anything

What you see	What to do
Cannot reach 127.0.0.1:31415	Open FreeLLMAPI and its dashboard. Confirm its current Claude base URL. If the dashboard opens, test the mapped model in Playground; an upstream timeout can look like a local connection failure.
Timeout; probedModel is claude-haiku-4-5	Set Haiku to a fast model that passes Playground. Save mapping, then click Check again in Claude. Changing Opus alone does not fix a Haiku probe.
401 / Unauthorized	Use the current FreeLLMAPI unified key in Claude and x-api-key authentication. Provider keys belong only in FreeLLMAPI.
404 / model unavailable	Select a model that is currently enabled and available to your provider account. Copy its exact catalog choice.
429 / rate limited	Use a different working provider or wait for the shown reset. Repeated retries do not create more quota.
503 / overloaded	Try another working model/provider, or retry later.
Workspace download fails	Keep a stable connection and retry the task. The app needs access to downloads.claude.ai. On a managed network, ask IT to investigate; do not disable endpoint protection.

“Not enough disk space to set up the workspace”

1. Open File Explorer > This PC and check free space on C:. A folder selected on another drive does not automatically relocate Cowork's VM.
2. Aim for 25 GB free before installation. Move files you recognize to a trusted drive, or review Windows Settings > System > Storage > Temporary files / Cleanup recommendations. Review every selected category, especially Downloads and Recycle Bin.
3. Keep projects, passwords, chat data and anything you cannot identify. Do not manually delete Claude's config, VM folders or system files to make space.
4. After reclaiming space, fully quit and reopen Claude, then start a fresh Cowork task. If it still fails, read the newest log or export a diagnostic report.

For advanced troubleshooting, Windows third-party logs are normally under %LOCALAPPDATA%\Claude-3p\logs\. main.log shows app/download activity; cowork_vm_node.log shows workspace startup. Review logs for private information before sharing.

A successful chat does not prove the workspace is ready. Our PC initially answered messages while its VM was blocked by storage. We confirmed the repair using a real Bash response.

Use it day to day

Keep this page as your quick reference

Every time you use it

1. Open FreeLLMAPI first. Leave it running in the tray.
2. Open Claude Desktop in your saved Gateway configuration.
3. Use Sonnet for quick tasks and Opus for harder work, according to your saved mapping.
4. Give Cowork a dedicated folder for the job. Start with a copy of important files and review the proposed changes.
5. If an answer fails, check FreeLLMAPI Analytics for the actual model, provider and error before changing settings.

The three things not to mix up

Item	Where it belongs
Provider API key	FreeLLMAPI > Keys > Providers
Unified API key	Claude Desktop gateway credentials
Model slot / actual model	Opus/Sonnet/Haiku are slots; FreeLLMAPI mapping selects the model

Share the guide, not your credentials

Each person should use their own provider accounts and keys. A Claude configuration export may include credentials. Do not distribute your exported configuration or copy your configLibrary folder to other people's PCs. Share these instructions and have each person enter their own values.

Pass/fail checklist

- FreeLLMAPI dashboard opens.
- At least one model answers correctly in Playground.
- Default, Opus, Sonnet and Haiku are explicitly mapped.
- Claude's gateway connection check passes.
- Opus and Sonnet each answer an actual chat prompt.
- Bash returns Linux and COWORK_WORKSPACE_OK.
- A test file is created in the dedicated test folder.

References and scope

FreeLLMAPI project and current install links

Anthropic: third-party Desktop overview

Anthropic: gateway contract

Anthropic: setup and troubleshooting

Prepared from real Windows requests and workspace execution verified on 26 November 2026, plus the linked documentation. Exact labels and provider/model availability may change. Free-model mappings are a community integration; they do not grant a Claude subscription or establish feature parity. The free-tier project recommends evaluation, with paid APIs for production.

Follow the video

A checked route through the supplied 20-minute tutorial

The supplied tutorial was reviewed using its narration and sampled on-screen frames on 26 November 2026. Pages 1–7 cover the core Windows setup and tested fixes. Pages 8–12 add the video-specific optional branches. Times below are approximate.

Video time	Action / guide location
01:03–02:20	Install router and tour Models, Playground, Keys, Agents — page 1
03:20–04:44	Google, NVIDIA and Cloudflare keys — pages 2 and 5; Z.ai — page 9
04:44–06:17	OpenRouter, keyless options, Image/Fusion — pages 9 and 12
06:49–08:55	NVIDIA Custom Endpoint, model discovery, routing and mappings — page 10
08:57–11:08	Claude gateway, Unified key, restart, Auto and Analytics — pages 3 and 10
11:08–14:14	Cowork organisation and expense-tracker demos — page 11
14:14–17:08	Antigravity troubleshooting and OpenCode demo — page 12
17:09–18:26	Alternative paths; an earlier Claude video is referenced, not included
18:27–20:01	Resource/course promotion and outro — no setup action required

Updates deliberately made to the tutorial

- Use the official Store distribution rather than routinely bypassing a Windows installer warning.
- Test pinned models first. Auto does not guarantee uninterrupted responses, and today's catalogue totals are not a personal token allowance.
- Use evaluation trials only for permitted evaluation. Business work requires suitable commercial access.
- Extra providers and coding/media demos are optional. They are not needed to make the basic Claude gateway work.

A model name in the video is an example from that recording. Use the exact ID in your current provider catalogue and prove both chat and tool execution. Do not assume the demonstrated model remains free or available.

The extra provider keys

Z.ai, OpenRouter and the video's model-selection dialog

The repeated Add key loop

1. Open FreeLLMAPI > Keys > Add key and select the provider. Create the key in your own account and paste it into the matching provider field. Cloudflare also needs a separate account ID.
2. If a Models on this endpoint dialog appears, tick the models you are entitled to use. The video uses Select All > Save; use an allowed subset where costs, terms or data handling differ.
3. Confirm the saved provider entry, select one model in Playground, send a fixed-response test and check the actual provider in Analytics. A green badge alone does not pass.

Z.ai — optional

1. Open docs.z.ai/guides/overview/quick-start and follow its official Open Platform / API Keys link. Sign in, create a Standard API Key and name it for this installation.
2. In FreeLLMAPI choose Z.ai, paste the key, choose entitled models and Save. Test ZAI_CHAT_OK 42 in Playground. Confirm allowance and billing; GLM models are not automatically all free. A Coding Plan may need its own documented endpoint.

OpenRouter — optional

1. Open openrouter.ai/settings/keys, sign in and create a labelled key. Choose an expiry you can maintain. The video's seven-day expiry is a demonstration choice; record a rotation date if you use it.
2. If supported by the current interface, set the evaluation key's credit limit to USD 0 and verify it saved. Select only exact free model variants. Blank/unlimited is not a verified zero limit; check current documentation if zero is handled differently.
3. Copy the key when shown, paste it under FreeLLMAPI's OpenRouter provider, choose the permitted models and Save. Test OPENROUTER_CHAT_OK 42, then inspect Analytics and OpenRouter Activity for route and cost. Do not enable billing/top-ups without approval.

Z.ai quick start

OpenRouter keys

OpenRouter activity

Custom models and Auto

The NVIDIA Fetch models step missing from the first edition

Add a custom NVIDIA evaluation endpoint

1. Use this when the built-in NVIDIA entry does not list the model you need. It does not grant extra entitlement or remove trial restrictions.
2. Open Keys > Add key > Custom Endpoint. Set Type to chat and give it a clear label.
3. Enter the upstream base https://integrate.api.nvidia.com/v1 and your NVIDIA provider key. This upstream includes /v1; Claude's local gateway base on page 3 does not. Do not append /chat/completions here.
4. Choose Fetch models. Select the desired models from the returned list, then Add models / Save. The video uses DeepSeek V4.1 Flash, GLM 5.3 Flash and MiniMax M3.1 Flash; use exact current catalogue IDs instead of guessing.
5. Find the saved models under Custom, test each in Playground and check Analytics. Prove tool compatibility before assigning a model to a Cowork slot.

Routing strategy and the optional Auto setting

1. On Models, review the routing strategies for speed, intelligence, reliability or manual ordering. The video chooses the intelligence strategy. A strategy name is not proof of best performance.
2. Keep explicit mappings until chat and tools pass. Then, if desired, go to Keys > Anthropic and change Default, Opus, Sonnet and Haiku to Auto > Save mapping.
3. Restrict the eligible pool to providers/models approved for your use, data and costs. Auto must not send business work to evaluation-only trials or introduce unapproved paid usage.
4. Start a new chat and a workspace tool/file task. In Analytics inspect which model and provider actually answered.
5. If the app provides a reversible provider disable control, temporarily exclude the first eligible route and test that another approved one answers; then restore it. Do not force a quota overrun or revoke a key.
6. If fallback cannot be verified, restore known-working pinned mappings. Auto cannot prevent errors when every eligible provider is unavailable.

NVIDIA official hosted endpoint reference

Reproduce the practical demos

Use disposable files and fictional transactions

Cowork: inventory first, then approved moves

1. First pass the shell and real host-folder tests on page 4. Create a backed-up disposable folder containing synthetic documents, pictures and sample duplicates. Grant access only to this folder.
2. Ask Cowork to inventory it without changing anything: count files/types, identify possible duplicates and propose a folder structure. Exclude hidden files, system folders and application data.
3. Review the actual results and request a move plan plus a log of original path, destination and reason. Approve only the intended batch before it writes.
4. Move suspected duplicates into a review folder; do not automatically delete them or assume the same filename means identical content.
5. Open the files in Windows, inspect the move log and test restoring one sample to its original path. A log helps reversal but does not replace a backup. Real client folder organisation is a separately approved task.

Playground: build an expense-tracker demonstration

1. Choose a tested model or your approved Auto pool. Request one HTML file showing a dark expense tracker, a fictional transaction form, a chart and a budget bar. The tutorial uses USD; GHS is a useful local adaptation.
2. Review the generated code before execution. Use Playground's HTML preview if available, or save the reviewed file in the test folder and open it locally. Never insert API keys, banking details or real client data.
3. Add fictional expenses and check totals and chart updates. Ask for a longer dashboard with overview cards and analysis, optionally switch to a stronger tested model as the video does, and retest.
4. Treat the result as a demo, not audited financial software. Visual polish is not evidence of correct arithmetic, secure code or production readiness.

> The tutorial's before/after file counts, response speed and model comparison are the creator's examples. Your acceptance evidence is the actual output and files on the PC you are setting up.

Optional tools and boundaries

The tutorial's repair suggestion, coding alternative and tour extras

Antigravity repair assistance — optional

1. Use the official Google Antigravity download/docs and your own eligible account only if you want this separate tool. Open a disposable diagnostic or approved project folder as its workspace.
2. Provide the exact redacted error, app/version and affected folder. Add the official FreeLLMAPI repository link for context. Do not paste whole private logs or exported credentials.
3. Ask for diagnosis and proposed changes first. Review commands and permissions; approve destructive, access or billing changes separately. Apply only the approved fix and repeat the original failed test. Keep a rollback note.

OpenCode coding alternative — optional

1. Use OpenCode's official documentation, choose Desktop and the correct Windows architecture, verify the source/publisher and install. Do not automatically bypass security warnings.
2. Open an empty BrewLab test folder and select an available, appropriately entitled model. The video's free built-in catalogue may have changed.
3. Request a small single-file coffee-shop website using sample content and optional 3D graphics. Review the files, open locally and check navigation/layout. Compare models with the same prompt and verify actual routes; do not infer general equality from one screenshot.

Other video mentions

Parallel and Kilo may appear as keyless options: review current entitlement/privacy, enable only with agreement, and test a public search or dummy chat. Image and Fusion are tour extras, not fully configured in the video; test their providers and allowances separately. Agents lists other client integrations, which each need their own configuration.

The video's reference to an earlier genuine-Claude tutorial is not a supplied walkthrough. For genuine Claude business use, the admin handbook documents the paid Claude API path. Its resource library and course promotion are not required to complete this setup.

Google Antigravity docs

OpenCode official docs/downloads

FreeLLMAPI official project