AI STACK SIGNAL

FREE PLAIN-ENGLISH GUIDE · OCT 2026

TRY AN AI THAT RUNS IN YOUR BROWSER

LFM3 Omni is a small AI from Liquid AI that understands text, pictures, video and speech, and can answer out loud. Liquid AI put a free demo on Hugging Face that runs it inside a Chrome tab, on your own computer. No app to install and no account needed to open it. This guide shows you how to try it, what to say to it, and what to expect.

01
What it is
One small model doing the job of four.

02
Try it, step by step
From opening Chrome to your first spoken answer.

03
Five things to say
Exact words to try, and what should happen.

04
Speed, privacy, limits
Honest expectations before you start.

01 THE SHORT VERSION WHAT IT IS

LFM3 Omni (full name LFM3-Omni-2B) is an AI model from the company **Liquid AI**. Liquid AI calls it "a single end-to-end model that understands text, images, video and audio, and answers in text and speech." It is built to run on your own device instead of on a company's servers.

1 Normally that takes four separate AIs

A talking assistant usually chains four tools: one turns your speech into text, one looks at pictures, one writes the answer, and one reads the answer out loud. Liquid AI says this one model "replaces a separate ASR model, vision-language model, LLM and TTS model." Those are the four jobs in that order.

2 It is small on purpose

It has 2.17 billion parameters (the numbers the AI learned). Big chat AIs in the cloud are far larger. Liquid AI says it is "designed to run efficiently on phones, laptops and other edge hardware." "Edge" just means your own device.

3 The demo runs it in a browser tab

Liquid AI's demo page on Hugging Face (a popular site for sharing AI models) describes itself as "Run LFM3-Omni-2B locally in the browser on WebGPU." WebGPU is a feature of modern browsers that lets a web page use your computer's graphics chip. That is what does the heavy lifting.

2.17B
parameters, per Liquid AI's model card

4 in, 2 out
In: text, images, video, audio. Out: text and speech

Oct 2026
Model posted Oct 21, browser demo Oct 29, on Hugging Face

! **A grain of salt on speed.** The developer who shared the demo on Reddit's r/LocalLLaMA reported about 90 ms to the first word, 140 ms to the first sound (about a seventh of a second) and about 280 tokens per second. That is **one person's test on a top-end Mac (M4 Max)**. Your computer will likely be slower.

02 FIVE MINUTES OF SETUP
BEFORE YOU START

You don't install anything. You do need the right browser, some patience on the first visit, and ideally a microphone.

WHAT YOU NEED

- **A computer with a recent browser.** The demo's own help message says to use "a recent version of Chrome, Edge or Safari." This guide uses Chrome.
- **A graphics chip WebGPU can use.** The page checks this for you as it loads.
- **A good internet connection** for the first load. The page downloads the whole model before it can answer.
- **A microphone** if you want to talk. Your laptop's built-in one is fine. No mic? You can type instead.
- **Speakers or headphones** to hear it answer.

MESSAGES YOU MIGHT SEE

- **"Checking your GPU"**: normal, the first step of loading.
- **"This browser doesn't support WebGPU."** Update your browser, or try a recent Chrome, Edge or Safari.
- **"This GPU can't run the model."** Your graphics chip isn't up to it. Try a different, newer computer.
- **"Check your connection, then try again."** The download was interrupted. Press Try again.
- **"I can't hear you. Allow microphone access, then try again."** See step 4 on the next page.

THE ADDRESS

Type this into Chrome's address bar, or search for "Liquid AI LFM3 Omni WebGPU" and pick the Hugging Face result from LiquidAI:

`huggingface.co/spaces/LiquidAI/LFM3-Omni-2B-WebGPU`

Check that the address starts with huggingface.co/spaces/LiquidAI. That is Liquid AI's own page. Copycat pages with similar names are common for popular AI tools.

✓ **Good to know:** opening the demo page needs no account and no sign-in. If any page asks for a payment or your password to "unlock" this demo, close it.

02 TRY IT IN CHROME

FROM ZERO TO A SPOKEN ANSWER

1 Open the demo page in Chrome

Go to `huggingface.co/spaces/LiquidAI/LFM3-Omni-2B-WebGPU`. You'll see "LFM3-Omni WebGPU" at the top and a glowing orb in the middle.

2 Wait for it to load, and keep the tab open

It first checks your graphics chip, then downloads the model. A percentage and the total download size show while it loads. How long it takes depends on your connection and your computer. Don't close or reload the tab, or it starts over.

3 Read the welcome

When it's ready, it says "Talk to me, or type." Under that are example questions you can click: "What time is it?", "Where am I right now?", "How much battery do I have?" and "Set a timer for one minute."

4 Talk to it

Hold the Space bar (or tap the orb) and speak. The first time, Chrome asks to use your microphone: click **Allow**. Let go of Space to send. Press **Esc** to cancel. One voice message can be up to 30 seconds; after that it sends what it has.

Prefer typing? Type in the message box and press **Enter**.

5 Listen to the answer

It answers out loud and in text at the same time. To stop it mid-answer, press **Esc** or tap the orb. You can also hold Space to talk over it.

6 Handy buttons

Spoken replies switches the voice off, so replies come as text only. **Transcript** opens the whole conversation. **New conversation** clears it and starts fresh.

03 FIVE THINGS TO TRY

EXACT WORDS TO SAY OR TYPE

Liquid AI's demo can check the time, tell where you are, set timers and more, all from inside the tab. Say these out loud (hold Space) or type them.

1 Kitchen timer

SAY OR TYPE

```
Set a timer for ten minutes for the pasta.
```

The page chimes and shows a notice when the time is up. Keep the tab open. To cancel, say "Cancel the pasta timer."

2 A phrase for your trip

SAY OR TYPE

```
How do I say "Where is the train station?" in Spanish?
```

Liquid AI lists Spanish, French, German, Italian, Japanese and more among its speech languages. Small models can slip on grammar, so double-check anything important.

3 Ask about your computer

SAY OR TYPE

```
How much battery do I have, and what kind of computer is this?
```

It reads this from your browser. Some browsers don't share battery levels; if so, it will say so.

4 The offline test

SAY OR TYPE

```
Am I online right now?
```

Once it has fully loaded, turn off your Wi-Fi and ask this. The demo's own built-in answer for an offline device says the assistant "runs entirely on this device, so it keeps working without a connection." Turn Wi-Fi back on afterwards.

5 A quick explainer

SAY OR TYPE

```
Explain in two short sentences what a graphics chip does.
```

Plain question and answer, like any chat AI.

! **"Where am I right now?"** makes Chrome ask for your location. You can click Block; it should then tell you it couldn't get your location. Only allow it if you're comfortable.

04 HONEST EXPECTATIONS
SPEED AND PRIVACY

ON A SLOWER OR OLDER COMPUTER

WHAT	WHAT TO EXPECT
First visit	The page downloads the whole model before it can answer. The loading screen shows the size and progress. On a slow connection, this is the longest wait.
Reply speed	The quick numbers you may have seen (140 ms to first sound) are one developer's test on an M4 Max Mac. Older laptops will be slower. You'll still get answers, just with a longer pause.
Won't start	If it says the browser doesn't support WebGPU or the GPU can't run the model, that computer can't run this demo. Try a newer one.
Helps a bit	Close other heavy tabs and apps while it loads and runs, and keep a laptop plugged in.

PRIVACY IN PLAIN WORDS

- The AI itself runs on your computer's graphics chip, inside the tab. The demo says it "runs on your device's graphics chip through WebGPU."
- So what you say or type is handled by your own computer, not sent to an AI company's server to be answered.
- That is why the offline test on page 5 works.

WHAT STILL LEAVES YOUR COMPUTER

- The page and the model are downloaded from Hugging Face, so Hugging Face sees your visit, like any website.
- Location: only if you click Allow when Chrome asks.
- Your browser and computer settings work as usual. Running locally is not the same as private browsing.

✓ **Why this matters:** an AI that runs on your own machine means your photos and voice notes can stay on your machine. That is the big idea behind small "on-device" models like this one.

BEFORE YOU RELY ON IT

LIMITS AND WHERE TO READ MORE

LIMITS

- **The demo is for talking and typing.** The model itself understands photos and video, per Liquid AI, but when we checked the demo page on October 30, 2026 it had no button to add a photo. Watch the page for updates.
- **It's a small model.** Liquid AI calls it a "non-reasoning" model: it answers straight away rather than working through hard problems step by step. Check facts, numbers and translations that matter.
- **Voice messages max out at 30 seconds** each.
- **It's a demo,** not a finished app. It may change or move.
- **License:** the model comes under Liquid AI's own "LFM Open License v1.0." Read it before using the model in a business.

SOURCES

1. Liquid AI, LFM3-Omni-2B model card: huggingface.co/LiquidAI/LFM3-Omni-2B
2. Liquid AI, LFM3 Omni 2B WebGPU demo: huggingface.co/spaces/LiquidAI/LFM3-Omni-2B-WebGPU
3. Speed figures: post by the developer who shared the demo on Reddit's r/LocalLLaMA, October 30, 2026 (one machine, an M4 Max)

WHAT WE DON'T KNOW

- Exactly how big the browser download is. The loading screen tells you.
- How fast it runs on an average laptop. Nobody has published wide tests yet.
- Whether the demo works on phones. Liquid AI hasn't said, so we don't claim it.

news.aistacksignal.com/omni

This guide is general information. AI Stack Signal is not connected to Liquid AI or Hugging Face. Facts checked against Liquid AI's model card and demo page on October 30, 2026; check them for later updates.