Pre-Workshop Setup Guide

OOSC 4.0 · IIIT ALLAHABAD

Session: Build Your Own Browser AI Agent

Speaker: Nishant Kumar Thakur

When: Saturday, 29 Aug · 2:30 PM – 3:55 PM

What you'll need for the workshop

1. A laptop with **Google Chrome** installed (latest version)
2. A code editor — we recommend **VS Code** (https://code.visualstudio.com)
3. The AI model pre-downloaded (instructions below)

Step 1: Install VS Code + Live Server

1. Download VS Code from https://code.visualstudio.com
2. Open VS Code
3. Go to Extensions (click the square icon on the left sidebar, or press `Ctrl+Shift+X` / `Cmd+Shift+X`)
4. Search for **"Live Server"** by Ritwick Dey
5. Click **Install**

That's it. We'll use this to run your code in the browser during the workshop.

Step 2: Check your browser supports WebGPU

1. Open **Google Chrome**
2. Type `chrome://gpu` in the address bar and press Enter
3. Look for the line: **WebGPU: Hardware accelerated**
4. If you see it — you're good. If not — update Chrome to the latest version and try again.

> **Don't have WebGPU?** Try Microsoft Edge (also works). If neither works, don't worry — we'll help you at the workshop.

Step 3: Download the workshop files

1. Download the workshop files from https://github.com/nkthakur48/webllm-workshop (click **Code** → **Download ZIP**)
2. Unzip it somewhere easy to find (e.g., your Desktop or Documents)
3. Open the folder in VS Code: **File** → **Open Folder** → **select the unzipped folder**

You should see a file called `precache.html` in the sidebar.

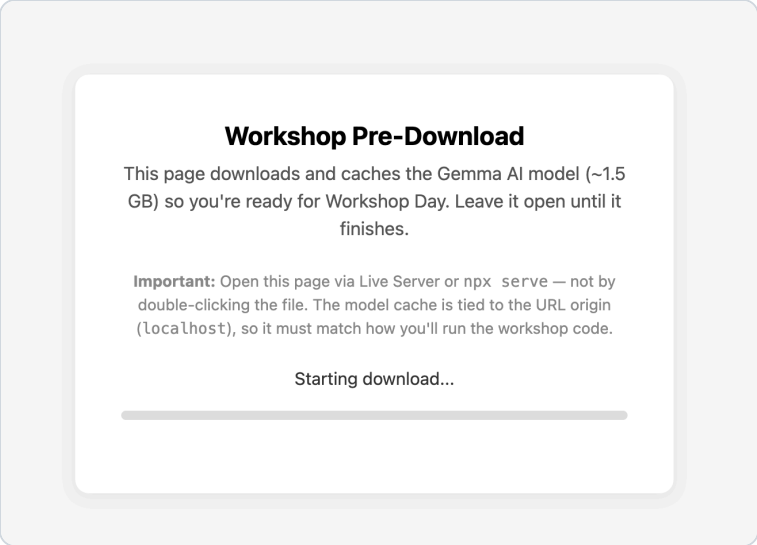
📌 Step 4: Pre-download the AI model (~1.5 GB) — most important step

This downloads the Gemma AI model into your browser's cache so the workshop starts fast. **Do this on good wifi — it's a one-time 1.5 GB download.**

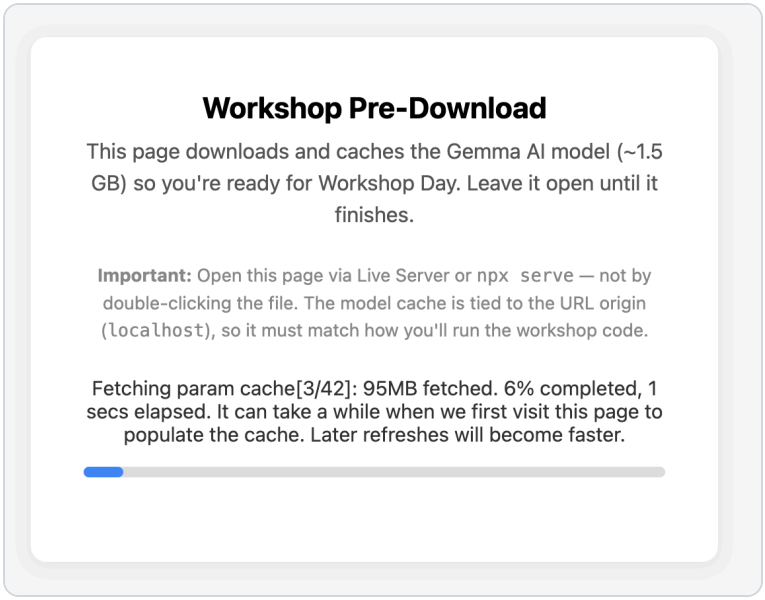
> **Why this matters:** The browser caches the model per website address (origin). If you download the model by double-clicking `precache.html` (which opens as `file://`), the cache won't carry over to `localhost://` when we run code during the workshop. That's why we use Live Server for this step too.

1. Open the workshop folder in VS Code (from Step 3)
2. Right-click `precache.html` in the sidebar → **Open with Live Server**
3. Chrome will open automatically — you'll see a progress bar
4. **Leave the tab open until it says "Model cached!"** (2–10 minutes depending on your wifi)
5. Once done, you'll see: **"You can close this tab. The model will load instantly at the workshop."**

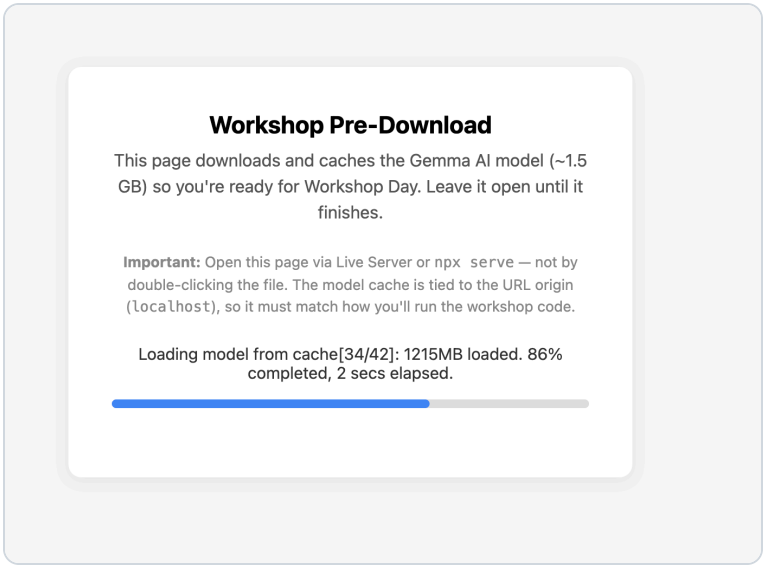
Right after you open it, you'll see:



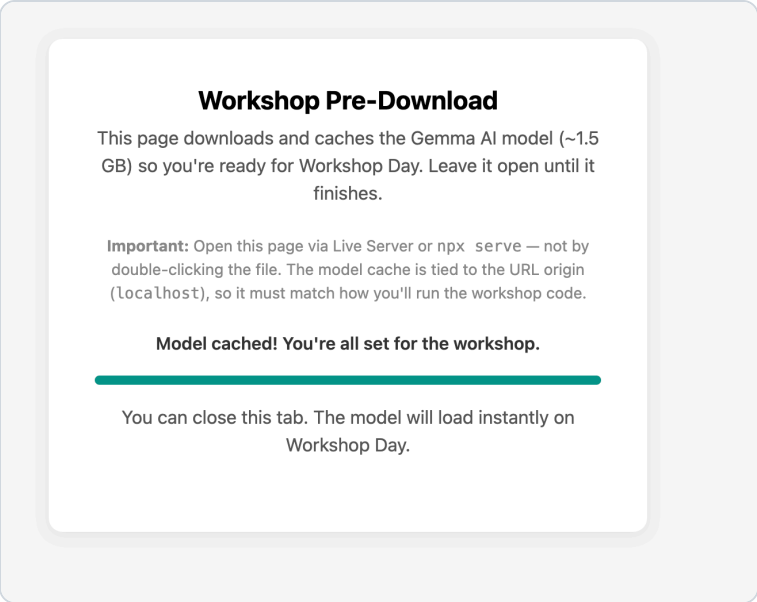
While it's downloading, you'll see a progress bar like this:



If you've already run this before, it'll load from cache instead of re-downloading:



When it's finished, the page will look like this:



Troubleshooting

Problem	Fix
Page shows "Download failed"	Make sure you're using Chrome or Edge. Check that `chrome://gpu` shows WebGPU support.
Download is stuck or very slow	Try a faster wifi network, or tether to your phone's data.
Tab crashes	Your laptop may not have enough GPU memory. Let us know at the workshop — we have a lighter model as backup.
"Open with Live Server" not showing	Make sure you installed the Live Server extension (Step 1 above). Alternatively, open a terminal in the folder and run `npx serve` then visit the URL it prints.

Important

- Do **not** clear your browser cache between now and the workshop.
- At the workshop, always open files using Live Server (same as above) — never by double-clicking the HTML file directly.

What to expect

In this 90-minute session, you'll build a working AI chat from scratch — load the model, send messages, get streaming responses, give your AI a personality — then wrap it in a polished chat UI, customize your AI agent, and submit it for the hackathon.

No prior AI or machine learning experience needed. Just bring your laptop and curiosity.

See you there! 🚀