

How to create server-side AI



AI Overview

Creating a server-side AI involves setting up a backend server with sufficient hardware, installing AI frameworks, and integrating large language models to handle requests and generate responses.

Hardware Requirements

To run AI models efficiently, especially large language models (LLMs), your server should include:

- High-performance CPU:** Multi-core processors like AMD Ryzen 9 or Intel Xeon for data preprocessing and request handling.
- GPU(s) with sufficient VRAM:** For LLMs, 12–16GB VRAM is needed for 7B–13B models, and 24–48GB for 30B–70B models. NVIDIA RTX 4090 or similar GPUs are recommended.
- RAM:** At least 96–128GB DDR5 for smooth multitasking and model inference.
- Storage:** Fast NVMe SSDs for model storage and caching.
- Cooling and power supply:** Effective thermal management and a robust PSU (e.g., 2000W) to support high-performance GPUs.

Software Stack

Operating System: Ubuntu 24.04 LTS or Pop!_OS for stability and performance.

AI Engine: Open-source frameworks like Ollama, Open WebUI, or PyTorch/TensorFlow for running LLMs locally.

Server Framework: Node.js or Python-based backend to handle API requests and manage sessions.

Containerization: Docker for packaging the server and AI models, optionally deployed on Kubernetes (AKS) for scalability.

Reverse Proxy & TLS: NGINX and cert-manager for secure HTTPS connections.

Deployment and Integration

Set up the server environment: Install OS, GPU drivers, and required libraries.

Containerize AI models: Use Docker to package the AI engine and dependencies.

Create API endpoints: Implement REST or GraphQL endpoints to accept user input and return AI responses.

Server-side rendering (SSR): For web applications, use frameworks like Next.js to render AI responses server-side for performance and SEO benefits.

Session and user management: Implement authentication and role-based access (Admin/User) to manage interactions securely.

Optimization and Best Practices

Batch processing: Optimize GPU usage by batching requests.

Memory management: Monitor...

Article Content

Build your first SharePoint client-side web part (Hello World part 1 ...

Preview the web part You can preview and test your client-side web part in the SharePoint hosted workbench without deploying your solution to SharePoint. This is done by starting a local web server the hosted workbench

Facebook Ads for Real Estate | Advanced 2026 Guide

Run Facebook ads for real estate by targeting location and interests, using lead forms or video, and tracking results with server-side data for better ROI.

Building my AI Server

The goal was simple: build a powerful AI inference server that could handle local LLM serving, fine-tuning experiments, and general ML workloads without breaking the bank.

How to Host Your Own Private AI on a Dedicated

In this guide, we will walk you through the exact hardware requirements and software steps to build your own private AI server using

Outlook Rules Not Working? How to Fix Them (2026)

Fix Outlook rules not working with this complete troubleshooting guide. Covers client-side vs server-side issues, migration fixes, and quota limits.

Building a Self-Hosted AI Server

A comprehensive guide to building a powerful self-hosted AI server with web-based chat interface, programmatic API access, and advanced document Q& A capabilities.

3 Ways to Build a Door in Minecraft

Doors are a basic essential to all buildings in Minecraft. Every house should have a door, including your first one. Without one, hostile mobs would be purging in to kill, even if you already have a strong weapon. Later in

How to build a high-performance AI server locally

Network Engineer and tech enthusiast NetworkChuck has provided a fantastic tutorial on how he built an AI server to run locally and provide large language model processing for affordable AI...

HTML Viewer

Instantly render HTML code online with live preview, console output and responsive mobile, tablet and desktop viewports. Paste, upload or share via link.

How to Build a Home AI Server for Local LLMs: The Complete 2026

Build a home AI server for local LLMs in 2026. Complete guide covering hardware tiers from RTX 3090 to DGX Spark, software setup with Ollama, and cost break-even analysis.

Top 5 AI Tools You Can Host on Your Own Server (and

As AI tools become more accessible and powerful, more developers, agencies, and small businesses are exploring how to self-host AI applications.

How to Build an Affordable Custom AI Server for AI

In this overview, Jun Yamog guides you through the essentials of building a high-performance AI server, from selecting the right GPUs to

Documentation | NestJS

Nest is a framework for building efficient, scalable Node.js server-side applications. It uses progressive JavaScript, is built with TypeScript and combines elements of

How to Accept Payments on a Website: Payment

FAQ How do I make a website accept payments? To make a website accept payments, choose payment methods, open a merchant or provider

How to run PHP programs

PHP (Hypertext Preprocessor) is a popular server-side scripting language primarily used for web development. It is designed to generate

Create organization-wide email signatures and disclaimers

Create organization-wide email signatures in Microsoft 365 to ensure consistent branding and legal compliance across all outgoing messages. You can set up email signatures to apply to all

Google Tag Manager Server Side Tagging (Complete

Explore server side tagging with Google Tag Manager to enhance data control, privacy, and site speed. Learn key benefits and setup in our guide.

The Ultimate Guide to build your Custom AI Server

An Ultimate Guide for AI Development and Hosting: Building your own Custom AI Server with Dashboard and embedded Client UI.

Building Your Own Local AI Powerhouse: A Complete Guide

A comprehensive guide to building fully open-source, local, and capable AI systems with complete privacy, customization, and offline capabilities.

Building My AI Home Lab: From Laptop to Dedicated

This isn't just an AI playground; it's practical, hands-on learning that makes me better at server administration, containerization, and systems thinking.

Get started with Microsoft 365 Copilot Notebooks

Note Copilot Notebooks is available to Microsoft 365 Copilot and Copilot Chat licensed users. Accounts must have a SharePoint or OneDrive license (service plan) to create notebooks. Copilot Notebooks

Claude can now create and edit files | Claude by Anthropic

Claude can now create and edit Excel spreadsheets, documents, PowerPoint slide decks, and PDFs directly in Claude.ai and the desktop app.

TCP Server-Client implementation in C

Prerequisites - Socket Programming in C/C++, TCP and UDP server using select, UDP Server-Client implementation in C If we are creating a

Demo Kit

Explore SAPUI5 SDK Demo Kit for comprehensive tools, documentation, and interactive demos to develop responsive web applications with SAP's UI framework.

Contact Us

For more information, pricing, or custom solutions, please contact us:

Website: <https://www.mr-security.co.za>

Email: info@mr-security.co.za

Phone: +33 1 45 23 67 81

Address: 10 Rue de la Paix, 75002 Paris, France

This document is for informational purposes only. Specifications subject to change without notice.

