

Local LLM, Agentic AI
Model Evaluation,
Performance Tuning, and
Web-Assisted Research
Using LM Studio



LM Studio

Table of Content

- 1. Executive Summary.....4**
- 2. Key Purposes.....4**
- 3. Local LLM Environment Setup (LM Studio)..... 5**
 - 3.1 System Requirements & Download..... 5
 - 3.2 Setup & Configuration.....6
- 4. Model Benchmarking and Evaluation.....7**
 - 4.1 Artificial Analysis..... 7
 - 4.2 LMArena..... 9
 - 4.3 SWE-Bench.....11
- 5. Runtime Configuration and Performance Tuning.....12**
- 6. Web-Assisted Research Setup..... 14**
 - 6.1 DuckDuckGo Plugin Integration..... 14
 - 6.2 Research Prompt Configuration..... 15
- 7. Speculative Decoding..... 17**
- 8. Configuring AnythingLLM with LM Studio as the LLM Provider..... 18**
- 9. MCP Integrations..... 24**
 - 9.1 Playwright MCP..... 24
 - 9.2 Google Maps MCP..... 25
 - 9.3 MCP.so.....26
- 10. Vision-Language Model: Bank Receipt Number Extraction..... 27**
- 11. Conclusion..... 28**



1. Executive Summary

- Designed and deployed a fully local LLM and agentic AI environment using LM Studio, ensuring cost efficiency, security, and operational control
- Evaluated and optimized open-weight foundation models using industry benchmarks, public leaderboards, and hands-on performance tuning
- Implemented advanced inference optimizations, including GPU offloading and speculative decoding, to improve latency and throughput
- Built agentic and web-assisted research workflows with tool and MCP integrations for scalable automation
- Demonstrated vision-language intelligence through accurate extraction of structured data from real-world financial documents

2. Key Purposes

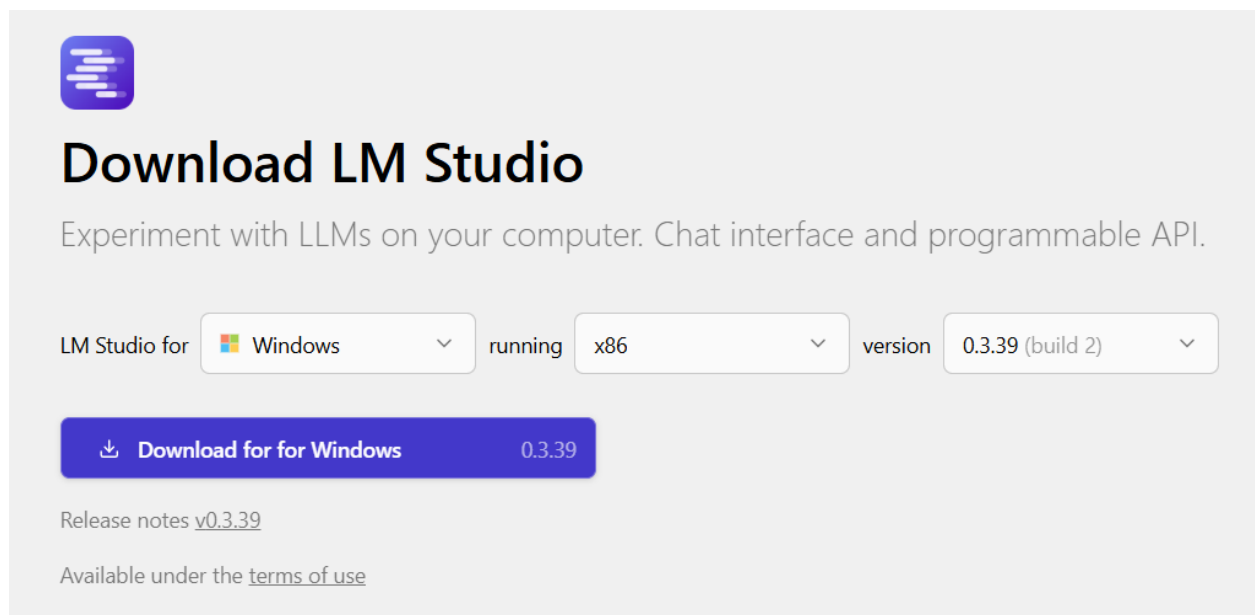
- Design and validate a fully local LLM and agentic AI environment using LM Studio
- Evaluate open-weight foundation models through benchmark-driven and hands-on analysis
- Optimize inference performance using runtime tuning and speculative decoding techniques
- Implement agentic and web-assisted research workflows with tool and MCP integrations
- Demonstrate vision-language capabilities through real-world document data extraction

3. Local LLM Environment Setup (LM Studio)

3.1 System Requirements & Download

Go to the downloads page and click on the appropriate Operating System version that is applicable to your case:

- **Source:** Download the official installer from <https://lmstudio.ai/download>
- **OS Support:** Available for macOS (Apple Silicon), Windows, and Linux.
- **System Requirements:** LM Studio runs on macOS (Apple Silicon), Windows (x64/ARM), and Linux (x64/ARM64) with 16GB RAM recommended.

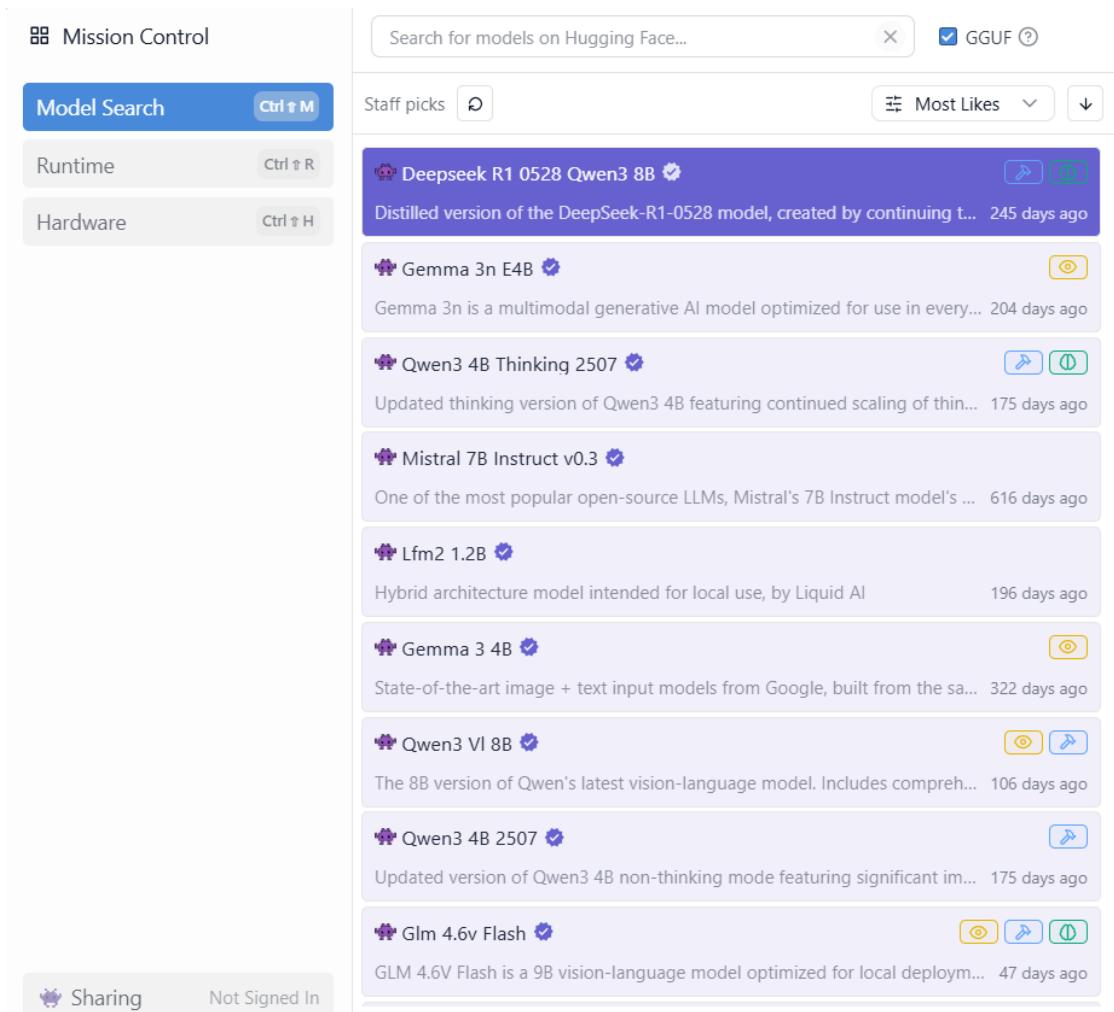


The screenshot shows the LM Studio download page. At the top left is the LM Studio logo, a purple square with white horizontal lines. Below it is the heading "Download LM Studio" in a large, bold, black font. Underneath the heading is a subtitle: "Experiment with LLMs on your computer. Chat interface and programmable API." Below this is a selection interface with three dropdown menus: "LM Studio for" (set to Windows), "running" (set to x86), and "version" (set to 0.3.39 (build 2)). Below these dropdowns is a large blue button with a white download icon, the text "Download for for Windows", and the version number "0.3.39". At the bottom of the page, there are two links: "Release notes v0.3.39" and "Available under the terms of use".

With the installation completed, LM Studio is now ready to be launched on the local system.

3.2 Setup & Configuration

After installation, the first step is to download a model. Use the search (Discover) icon on the left panel to browse and select available models.



The screenshot displays the Kasadara Mission Control interface. On the left, a sidebar contains a 'Mission Control' header and three main sections: 'Model Search' (highlighted in blue with a 'Ctrl + M' shortcut), 'Runtime' (with 'Ctrl + R'), and 'Hardware' (with 'Ctrl + H'). Below these is a 'Sharing' section indicating the user is 'Not Signed In'. The main panel on the right features a search bar with the placeholder 'Search for models on Hugging Face...', a 'GGUF' filter toggle, and a 'Staff picks' section. A list of model cards is displayed, each with a title, description, and a 'days ago' timestamp. The models listed are: Deepseek R1 0528 Qwen3 8B (245 days ago), Gemma 3n E4B (204 days ago), Qwen3 4B Thinking 2507 (175 days ago), Mistral 7B Instruct v0.3 (616 days ago), Lfm2 1.2B (196 days ago), Gemma 3 4B (322 days ago), Qwen3 V1 8B (106 days ago), Qwen3 4B 2507 (175 days ago), and GLM 4.6v Flash (47 days ago). Each card includes icons for search, GGUF, and other model-specific features.

Model Name	Description	Days Ago
Deepseek R1 0528 Qwen3 8B	Distilled version of the DeepSeek-R1-0528 model, created by continuing t...	245 days ago
Gemma 3n E4B	Gemma 3n is a multimodal generative AI model optimized for use in every...	204 days ago
Qwen3 4B Thinking 2507	Updated thinking version of Qwen3 4B featuring continued scaling of thin...	175 days ago
Mistral 7B Instruct v0.3	One of the most popular open-source LLMs, Mistral's 7B Instruct model's ...	616 days ago
Lfm2 1.2B	Hybrid architecture model intended for local use, by Liquid AI	196 days ago
Gemma 3 4B	State-of-the-art image + text input models from Google, built from the sa...	322 days ago
Qwen3 V1 8B	The 8B version of Qwen's latest vision-language model. Includes compreh...	106 days ago
Qwen3 4B 2507	Updated version of Qwen3 4B non-thinking mode featuring significant im...	175 days ago
GLM 4.6v Flash	GLM 4.6V Flash is a 9B vision-language model optimized for local deploym...	47 days ago

Select the most appropriate model for your use case; the model selection strategy is discussed in the next section.

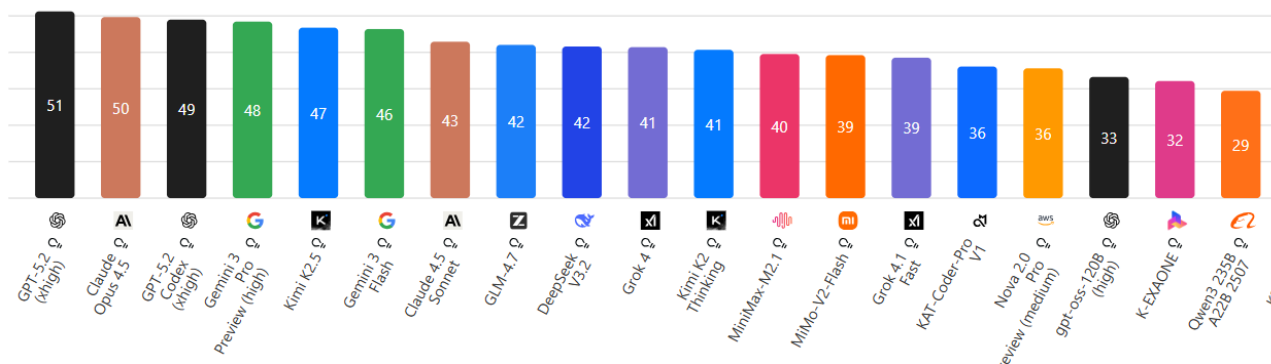
4. Model Benchmarking and Evaluation

4.1 Artificial Analysis

The following benchmarks are sourced from <https://artificialanalysis.ai> and compare models based on intelligence, agentic capabilities, and more.

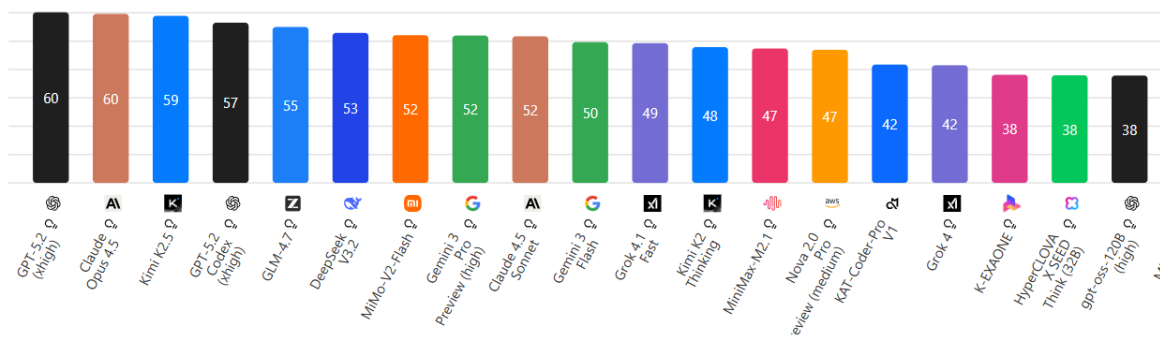
Artificial Analysis Intelligence Index

Artificial Analysis Intelligence Index v4.0 incorporates 10 evaluations: GDPval-AA, τ^2 -Bench Telecom, Terminal-Bench Hard, SciCode, AA-LCR, AA-Omniscience, IFBench, Humanity's Last Exam, GPQA Diamond, CritPt



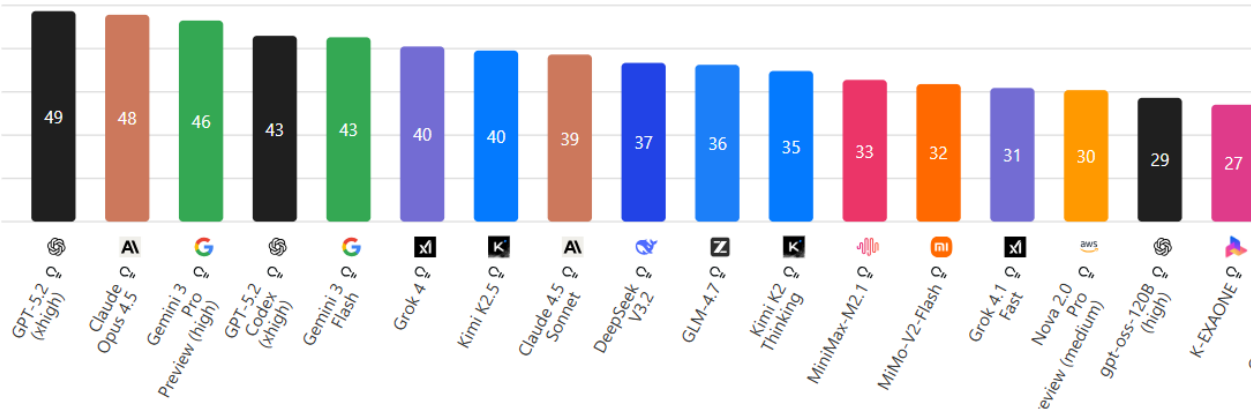
Artificial Analysis Agentic Index

Represents the average of agentic capabilities benchmarks in the Artificial Analysis Intelligence Index (GDPval-AA, τ^2 -Bench Telecom)



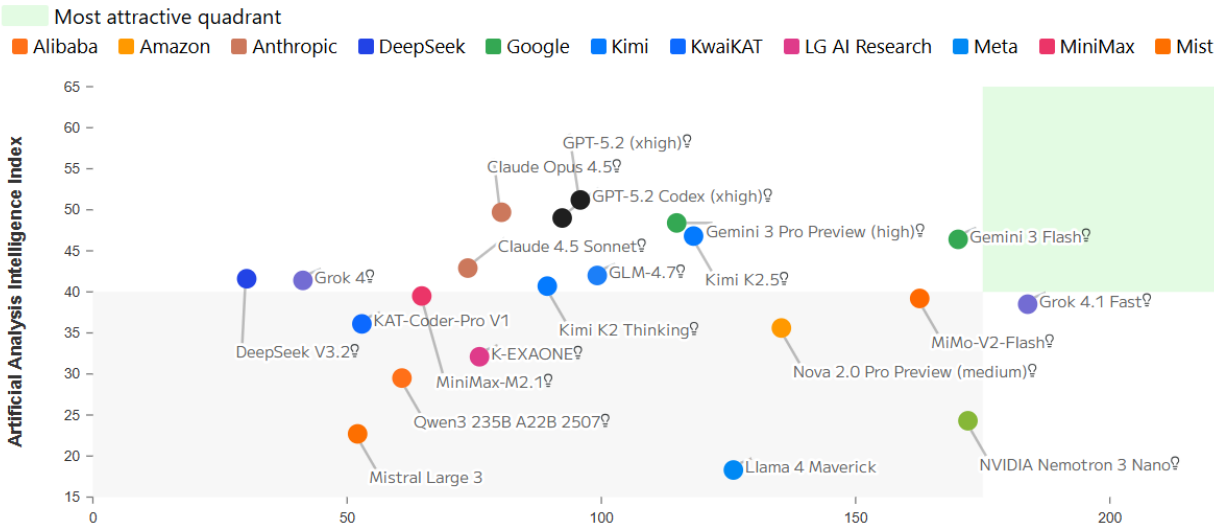
Artificial Analysis Coding Index

Represents the weighted average of coding benchmarks in the Artificial Analysis Intelligence Index (Terminal-Bench Hard, SciCode)



Intelligence vs. Output Speed

Artificial Analysis Intelligence Index; Output Speed: Output Tokens per Second



Click on a model to view more details

Qwen3 VL 4B Instruct Intelligence, Performance & Price Analysis

 Alibaba · Open weights model · Released October 2025

Intelligence #1 / 24  14 Artificial Analysis Intelligence Index	Speed N/A Output tokens per second	Input Price #1 / 24 \$0.00 USD per 1M tokens	Output Price #1 / 24 \$0.00 USD per 1M tokens	Verbosity #23 / 24  32M Output tokens from Intelligence Index	TECHNICAL SPECIFICATIONS 💡 Reasoning 🔧 Input modality No ⓘ 🔧 Output modality T ⓘ 📄 🗨️ 📏 Context window 256k ⓘ
--	---	---	--	--	---

4.2 LMArena

This platform <https://lmarena.ai> enables side-by-side comparison of answers from leading AI models



Find the best AI for you

Compare answers across top AI models, share your feedback and power our public [leaderboard](#)

Ask anything...



The following section demonstrates a side-by-side comparison of AI-generated responses, along with user-driven evaluation options

What is Attention Mechanism in Deep Learning?

Assistant A

Attention Mechanism in Deep Learning

The **Attention Mechanism** is a powerful technique in deep learning—especially within **sequence-to-sequence (seq2seq) models, transformers**, and tasks involving **natural language processing (NLP)**, computer vision, and more. Its primary purpose is to allow a model to **focus on the most relevant parts of the input data** when producing each element of the output, rather than treating all input equally.

Why Was Attention Introduced?

Assistant B

Excellent Question! Here's a clear and comprehensive answer:

What is the Attention Mechanism?

The **Attention Mechanism** is one of the most important innovations in Deep Learning. It allows neural networks to **dynamically focus on the most relevant parts** of the input sequence, rather than treating all parts equally.

It was introduced in 2014 for machine translation, but became extremely popular after the 2017 paper **"Attention Is All You Need"** (Vaswani et al.), which introduced the **Transformer** architecture.

← Left is Better

It's a tie

Both are bad

Right is Better →

The following image compares responses from two different AI models for the same prompt

amazon-nova-experimental-chat-12-10

Attention Mechanism in Deep Learning

The **Attention Mechanism** is a powerful technique in deep learning—especially within **sequence-to-sequence (seq2seq) models**, **transformers**, and tasks involving **natural language processing (NLP)**, **computer vision**, and more. Its primary purpose is to allow a model to **focus on the most relevant parts of the input data** when producing each element of the output, rather than treating all input equally.

Why Was Attention Introduced?

Traditional **encoder-decoder** architectures (e.g., for machine translation) compressed an entire input sequence into a **single fixed-length vector** (the “context vector”). This became a **bottleneck** for long sequences, as vital information could be lost.

kruxlox

Excellent Question! Here's a clear and compreh

What is the Attention Mechanism?

The **Attention Mechanism** is one of the most im allows neural networks to **dynamically focus on** sequence, rather than treating all parts equally.

It was introduced in 2014 for machine translation 2017 paper “**Attention Is All You Need**” (Vaswan architecture.

Core Idea (Intuitive Explanation)

Imagine you're translating this sentence:

LM Arena also provides a public leaderboard ranking LLMs based on community evaluations and performance

Text Arena

View rankings across various LLMs on their versatility, linguistic precision, and cultural context across text

Last Updated

Jan 27, 2026

Total Votes

5,135,830

Total Models

299

Overall

Search by model name...

Style Control

Rank	Rank Spread	Model	Score	95% CI (±)	Votes	Organization	License
1	1 ↔ 1	gemini-3-pro	1488	±5	31,624	Google	Proprietary
2	2 ↔ 5	grok-4.1-thinking	1476	±5	31,545	xAI	Proprietary
3	2 ↔ 7	gemini-3-flash	1471	±5	18,772	Google	Proprietary
4	2 ↔ 7	claude-opus-4-5-20251101-thinking-32k	1468	±5	23,402	Anthropic	Proprietary
5	2 ↔ 8	claude-opus-4-5-20251101	1467	±5	28,091	Anthropic	Proprietary
6	3 ↔ 8	grok-4.1	1466	±5	35,971	xAI	Proprietary
7	3 ↔ 9	gemini-3-flash (thinking-minimal)	1464	±6	13,318	Google	Proprietary

4.3 SWE-Bench

SWE-bench is a benchmark that evaluates large language models on real-world software engineering tasks by testing their ability to understand, modify, and fix code in real GitHub repositories.

Website: <https://www.swebench.com>



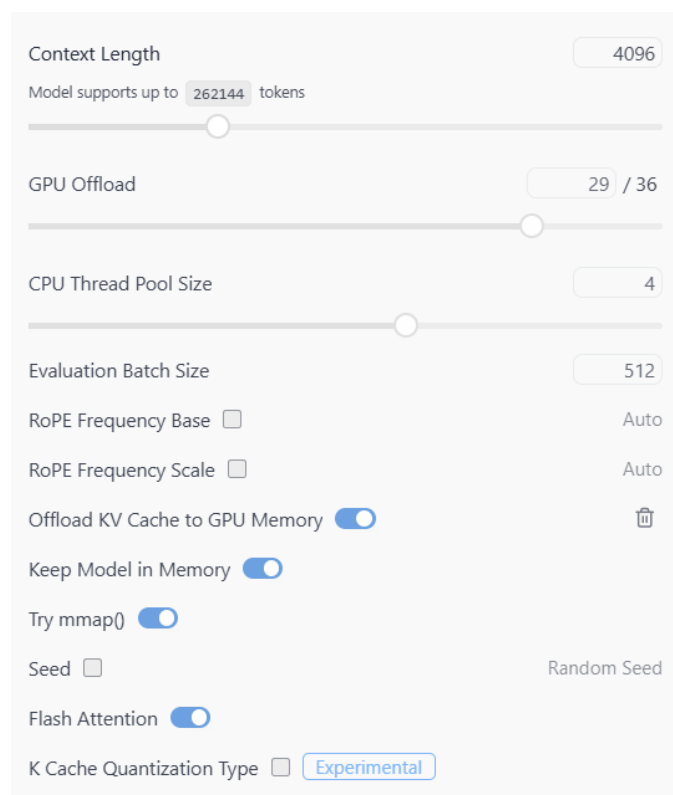
The following table shows SWE-bench leaderboard results comparing models by issue resolution performance.

Compare results Filters: All Tags							
☐	Model	% Resolved	Avg. \$	Trajs	Org	Date	Release
☐	NEW Claude 4.5 Opus medium (20251101)	74.40	\$0.72	🔗	AI	2025-11-24	1.16.0
☐	NEW Gemini 3 Pro Preview (2025-11-18)	74.20	\$0.46	🔗	🌟	2025-11-18	1.15.0
☐	NEW GPT-5.2 (2025-12-11) (high reasoning)	71.80	\$0.52	🔗	🌀	2025-12-11	1.17.2
☐	Claude 4.5 Sonnet (20250929)	70.60	\$0.56	🔗	AI	2025-09-29	1.13.3
☐	NEW GPT-5.2 (2025-12-11)	69.00	\$0.27	🔗	🌀	2025-12-11	1.17.2
☐	Claude 4 Opus (20250514)	67.60	\$1.13	🔗	AI	2025-08-02	1.0.0
☐	NEW GPT-5.1-codex (medium reasoning)	66.00	\$0.59	🔗	🌀	2025-11-24	1.16.0
☐	NEW GPT-5.1 (2025-11-13) (medium	66.00	\$0.31	🔗	🌀	2025-11-20	1.15.0

5. Runtime Configuration and Performance Tuning

The following panel allows users to configure runtime and performance settings for the selected local LLM

- **Context Length:** Defines how much input the model can process at once.
- **GPU Offload:** Improves performance by moving model layers to the GPU.
- **CPU Thread Pool Size:** Controls CPU parallelism during inference.
- **Evaluation Batch Size:** Impacts inference speed and memory usage.
- **Offload KV Cache to GPU:** Reduces latency by storing attention cache on the GPU.
- **Keep Model in Memory:** Avoids reload time for faster repeated interactions.
- **Try mmap():** Lowers RAM usage by memory-mapping model files.
- **Seed:** Ensures reproducible outputs when set.
- **Flash Attention:** Accelerates attention computation for better performance.
- **K Cache Quantization:** Reduces memory usage of the KV cache (experimental).



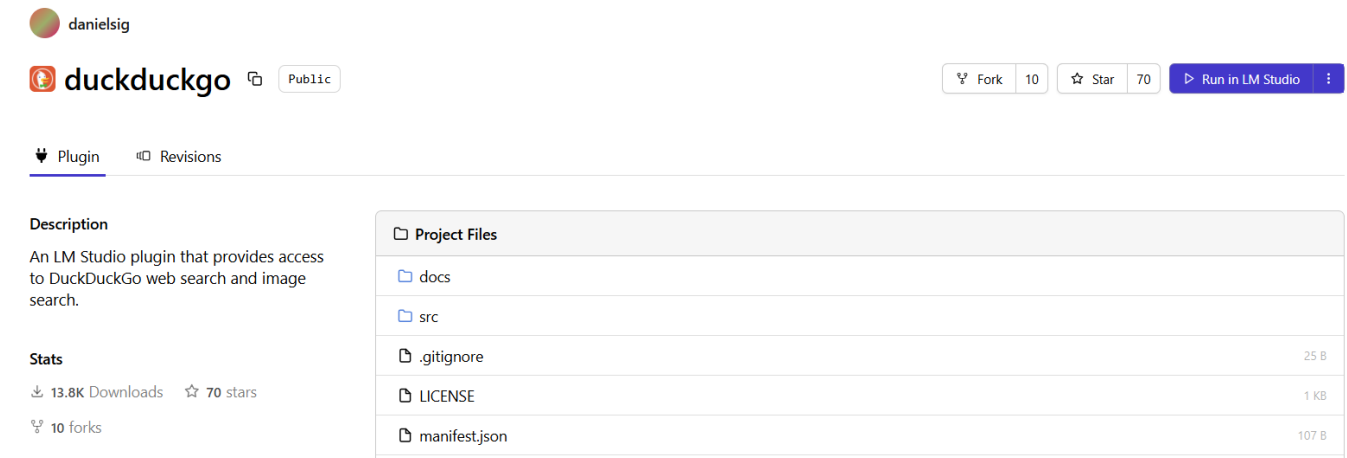
The screenshot shows a configuration panel for a local LLM. It includes sliders for Context Length (set to 4096, with a max of 262144), GPU Offload (set to 29 / 36), and CPU Thread Pool Size (set to 4). It also has input fields for Evaluation Batch Size (512), RoPE Frequency Base (Auto), and RoPE Frequency Scale (Auto). Toggle switches are present for Offload KV Cache to GPU Memory, Keep Model in Memory, Try mmap(), and Flash Attention. A Seed field is set to Random Seed. At the bottom, there is a checkbox for K Cache Quantization Type, which is currently set to Experimental.

Setting	Value
Context Length	4096
Model supports up to	262144 tokens
GPU Offload	29 / 36
CPU Thread Pool Size	4
Evaluation Batch Size	512
RoPE Frequency Base	Auto
RoPE Frequency Scale	Auto
Offload KV Cache to GPU Memory	Enabled
Keep Model in Memory	Enabled
Try mmap()	Enabled
Seed	Random Seed
Flash Attention	Enabled
K Cache Quantization Type	Experimental

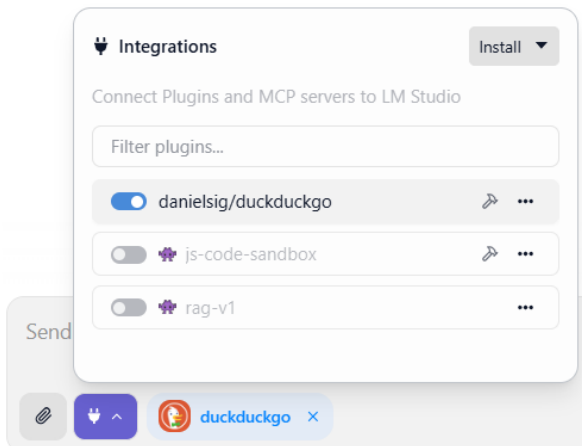
6. Web-Assisted Research Setup

6.1 DuckDuckGo Plugin Integration

Go to <https://lmstudio.ai/danielsig/duckduckgo> and click on Run in LM Studio it will automatically open LM Studio and prompt you to download the DuckDuckGo plugin

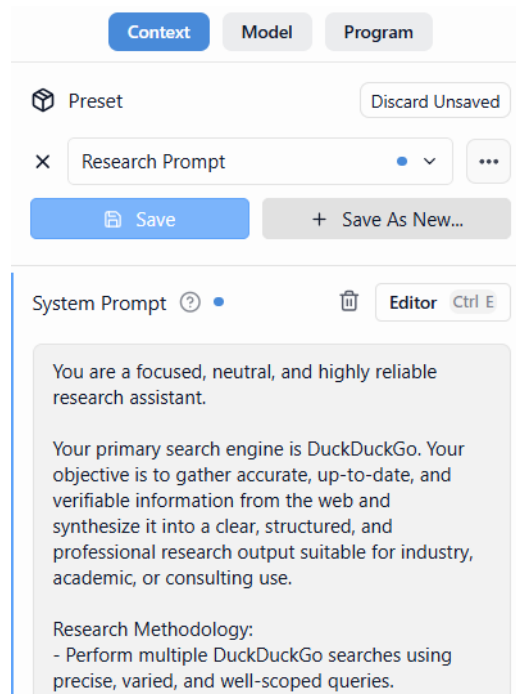


After installation, open the chat window and use the Integrations (plug) icon to enable the required plugins.

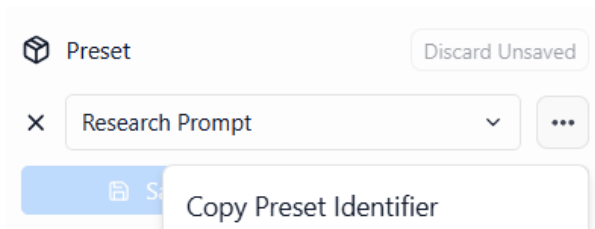


6.2 Research Prompt Configuration

In the top-right corner, click Settings. Under Context → System Prompt, enter your system prompt, save it, and reuse it later



Click the three-dot menu next to the saved prompt and copy the preset identifier to use it directly in the chat





The saved research prompt preset is invoked directly in the chat using its preset identifier, enabling web-assisted research via the DuckDuckGo plugin.

@local:research-prompt What is Deep Agents in Langchain?



@local:research-prompt

qwen/qwen3-v1-4b

I'm ready to conduct research using DuckDuckGo. Please provide the specific topic or question you'd like me to investigate. For example:

- "Latest advancements in AI research"
- "Impact of climate change on coastal cities"
- "Comparison of Python and JavaScript frameworks"
- "Best practices for secure software development"

I'll perform targeted searches using precise queries and prioritize credible sources like academic papers, official documentation, and reputable publications. Let me know what you need researched!



Deep Agents Module in Langchain

qwen/qwen3-v1-4b



Send a message to the model...



7. Speculative Decoding

- Speculative Decoding speeds up text generation by using a small draft model to predict tokens, which are then verified by a larger main model.
- If predictions match, multiple tokens are accepted at once → lower latency + faster responses without quality loss.

To perform speculative decoding, we use two models from the same family: a main model and a draft model.

Your Models				Recency ↓	Size	Downloaded
Qwen3 8B	Q4_K_M		qwen	8B	qwen3	GGUF 4.68 GB
Qwen3 VL 4B	Q4_K_M	 	qwen	4B	qwen3vl	GGUF 3.10 GB

Click the top-right Settings menu, go to the Models section, select a draft model, enable speculative decoding, and tune the required parameters

 **Speculative Decoding** Basic All 

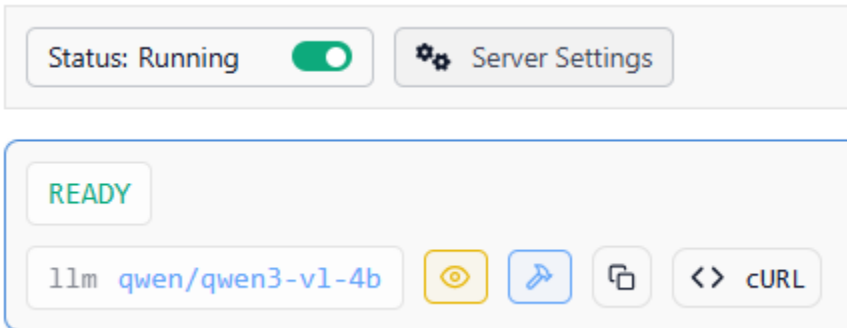
Draft Model [Read how it works](#)

Select a compatible draft model  

☐ Visualize accepted draft tokens

8. Configuring AnythingLLM with LM Studio as the LLM Provider

In the side panel, enable Developer Mode and ensure the status shows Running



Next, download AnythingLLM from <https://anythingllm.com/desktop>, select an LLM provider, and start chatting with the model

The all-in-one AI application

Everything great about AI in one desktop application.
Chat with docs, use AI Agents, and more - full locally and offline.

Download for desktop



Click the Upload icon near the Workspace. This interface allows users to upload documents or add web links for knowledge ingestion and retrieval

Name

No Documents



Click to upload or drag and drop
supports text files, csv's, spreadsheets, audio files, and more!

or submit a link

Fetch website

In Settings → Agent Skills, this panel allows users to enable or disable capabilities such as RAG, document summarization, and web scraping

 Agent Skills

RAG & long-term memory

On >

View & summarize documents

On >

Scrape websites

On >

Generate & save files

Off >

Generate charts

Off >

Web Search

Off >

SQL Connector

Off >

Browse and import agent skills from the AnythingLLM Community Hub

Community Hub

Share and collaborate with the AnythingLLM community.

Recently Added on AnythingLLM Community Hub

Explore the latest additions to the AnythingLLM Community Hub

Agent Skills

[Explore More →](#)

Calendar Event ⓘ
A plugin to generate calendar event links for Google and Outlook
Verified Skill
3 files found
[Import →](#)

Testing Assessments Lisa-AI ⓘ
Extracts standardized test results, cognitive assessments, and academic achievement evaluations for psychoeducational reports.
Unverified Skill
2 files found
[Import →](#)

Home Assistant Automation ⓘ
The Home Assistant Automation agent skill. Allows you to automate Home Assistant devices.
Verified Skill
3 files found
[Import →](#)

OMDB Movie Data Fetcher ⓘ
Fetches movie data from the OMDB API using title or IMDb ID, with language support and improved search accuracy. Note: May have limitations with non-English titles.
Unverified Skill
6 files found
[Import →](#)

Open Slack App ⓘ
Opens a deeplink or URL to a Slack channel. A simple use case of using an agent skill to open pre-installed apps.
Verified Skill
4 files found
[Import →](#)

After importing, the skill appears under Settings → Custom Skills, where it can be enabled.

 Custom Skills

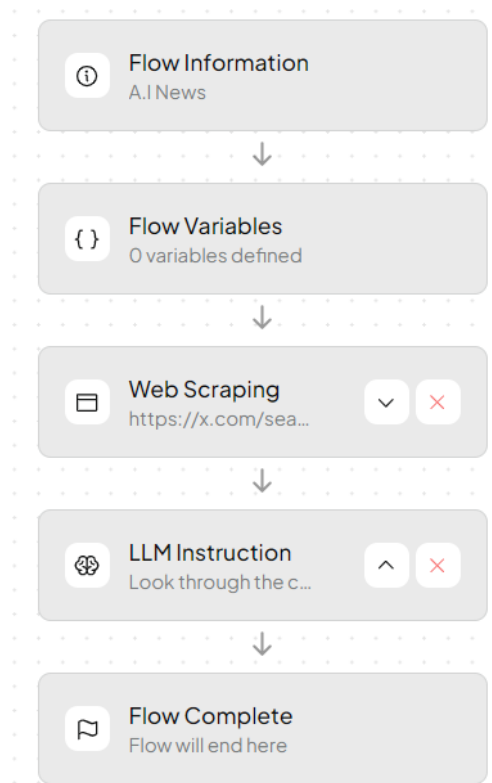
 Calendar event



A plugin to generate calendar event links for Google and Outlook by MintplexLabs



This agent flow defines an automated pipeline that scrapes web content, processes it with an LLM, and completes the task



Before performing web scraping, go to Settings → Tools → Browser Tool and log in to sites such as Google, LinkedIn, or GitHub

Browser Tool

Configure and manage the private browser tool for accessing authenticated web content.

Private Browser

Use this browser to authenticate to websites you want to scrape or access via AnythingLLM. <#>

 [Open Private Browser](#)



Invoke an agent in chat using @agent <agent-name>

@agent Look for A.I News







This output shows AI News highlights generated by the A.I News agent.

Top 4 AI News Highlights (April 2025)

- 01. OpenAI's GPT-5 Beta Released**
OpenAI unveiled a major update to its GPT series — GPT-5 Beta — with improved reasoning, multilingual support, and enhanced code generation. The model is now trained on 100+ trillion tokens and can handle complex, multi-step tasks.
- 02. Google AI Launches "Gemini 3" with Real-Time Video Understanding**
Google's latest AI model can now process and interpret live video streams in real time — useful for robotics, surveillance, and content moderation. Gemini 3 also supports 100+ languages and integrates with Android/iOS devices.
- 03. AI Ethics Group Publishes Global AI Governance Framework**
A coalition of 12 global AI ethics organizations released a new framework for regulating AI in healthcare, finance, and education. The framework emphasizes transparency, accountability, and human oversight.
- 04. Microsoft and Anthropic Announce Joint AI Research Lab**
Microsoft and Anthropic are launching a joint AI research lab in Seattle, focused on building safer, more ethical AI systems. The lab will prioritize alignment with human values and reduce bias in AI models.



In LM Studio, this local IP is reachable within the same network, allowing AnythingLLM to connect and use the LM Studio model from the same environment

Reachable at: `http://127.0.0.1:1234` 

In AnythingLLM, go to Settings → AI Providers → LLM Preferences and paste the LM Studio base URL (IP) to use the LM Studio model

LLM Preference

These are the credentials and settings for your preferred LLM chat & embedding provider. It is important that these keys

LLM Provider



LM Studio

Discover, download, and run thousands of cutting edge LLMs in a few clicks.



LM Studio Model

qwen/qwen3-vl-4b



Choose the LM Studio model you want to use for your conversations.

Hide advanced settings ^

LM Studio Base URL

`http://127.0.0.1:1234/v1`

Enter the URL where LM Studio is running.

Max Tokens (Optional)

Auto-detected from model

Override the context window limit. Leave empty to auto-detect from the model (defaults to 4096 if detection fails).

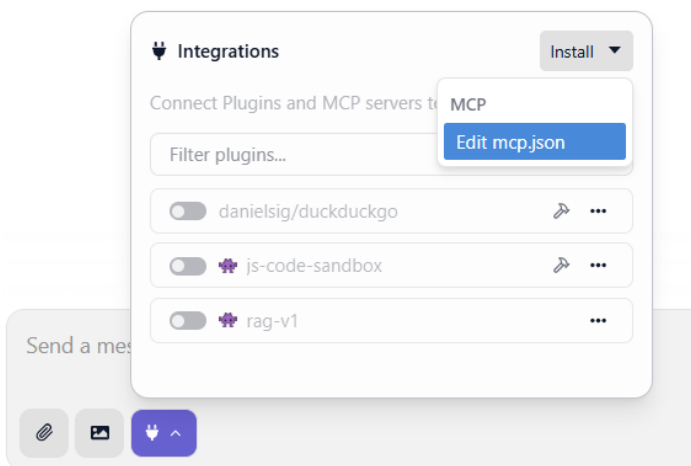
9. MCP Integrations

9.1 Playwright MCP

To integrate the MCP server, go to <https://github.com/microsoft/playwright-mcp> and copy the configuration code and

```
{
  "mcpServers": {
    "playwright": {
      "command": "npx",
      "args": [
        "@playwright/mcp@latest"
      ]
    }
  }
}
```

Next, click the plug icon in the chat, select *Edit mcp.json*, paste the configuration code, and the plugin is ready to use



What is headline on <https://www.artificialintelligence-news.com/>

qwen/qwen3-v1-4b

Model wants to call `browser_navigate` mcp/playwright ▾

Arguments Edit

```
{
  "url": "https://www.artificialintelligence-news.com/"
}
```

Proceed Ctrl Enter Deny Ctrl Esc Deny with Reason...

☐ Always allow `browser_navigate` from `mcp/playwright`

☐ Always allow any tool from `mcp/playwright`

9.2 Google Maps MCP

To integrate the Google Maps MCP, go to <https://developers.google.com/maps/ai/mcp> and copy the configuration into `mcp.json`.

```
"google-maps": {
  "command": "npx",
  "args": [
    "-y",
    "@modelcontextprotocol/server-google-maps"
  ],
  "env": {
    "GOOGLE_MAPS_API_KEY": "YOUR_GOOGLE_MAPS_API_KEY"
  }
}
```

Obtain the Google Maps API key from the Google Cloud Console

The [mcp.so](#) website provides a marketplace to discover and integrate MCP servers and clients.

Find Awesome MCP Servers and Clients

MCP.so is a third-party MCP Marketplace with **17478** MCP Servers collected.

[ShipAny](#): NextJS boilerplate for building AI SaaS startups.

Today

Featured

Latest

Clients

Hosted

Official

Featured MCP Servers

EdgeOne Pages MCP

An MCP service designed for deploying HTML content to EdgeOn...

AlphaVantage

Bring enterprise-grade stock market data to agents and LLMs

Time

A Model Context Protocol server that provides time and timezone...

Zhipu Web Search

Zhipu Web Search MCP Server is a search engine specifically designed...

MCP Advisor

MCP Advisor & Installation - Use the right MCP server for your needs

Howtocook Mcp

基于Anduin2017 / HowToCook (程序员在家做饭指南) 的mcp server, 帮...

MiniMax MCP

Official MiniMax Model Context Protocol (MCP) server that enables...

Serper MCP Server

A Serper MCP Server

Jina AI MCP Tools

A Model Context Protocol (MCP) server that integrates with Jina AI

Amap Maps

高德地图官方 MCP Server

Playwright Mcp

Playwright MCP server

Baidu Map

百度地图核心API现已全面兼容MCP协议 是国内首家兼容MCP协议的地图

Playwright Mcp

@microsoft

6 months ago #mcp #playwright

Overview

Tools

Comments

Overview

What is Playwright MCP?

Playwright MCP is a Model Context Protocol server that provides browser automation capabilities using Playwright. It allows large language models (LLMs) to interact with web pages through structured accessibility snapshots, eliminating the need for screenshots or visually-tuned models.

How to use Playwright MCP?

To use Playwright MCP, you can run it using the command `npm run playwright/mcp@latest` in your terminal. You can configure it to run in headless mode or with a display, depending on your needs.

Try in Playground

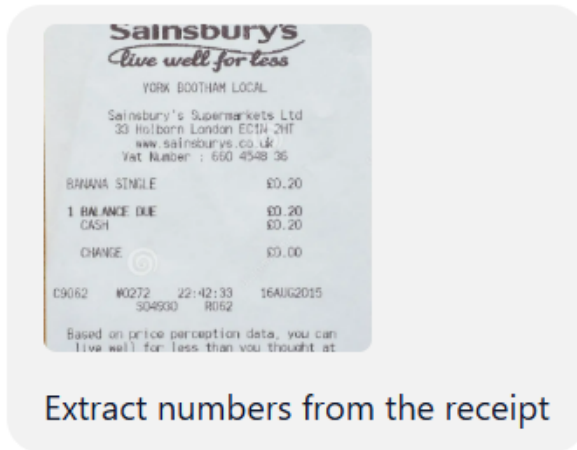
Server Config

```
{  "mcpServers": {    "playwright": {      "command": "npm",      "args": [        "@playwright/mcp@latest"      ]    }  }}
```

10. Vision-Language Model: Bank Receipt Number Extraction

Testing the capabilities of the Qwen3 VL vision model and evaluating its output.

Demonstrates **Qwen3-VL-4B's ability to extract structured numeric fields** (IDs, date, time, payment codes) from receipt images using local vision-language understanding in LM Studio.



qwen/qwen3-vl-4b

Here are the numbers extracted from the receipt:

- **Store ID / Receipt Number:** C9062
- **Transaction ID:** 00272
- **Time:** 22:42:33
- **Date:** 16AUG2015
- **Store Code:** SO4930
- **Payment Method:** R062
- **Vat Number:** 660 4548 36

11. Conclusion

- Demonstrates the practical viability of **local AI systems** for advanced LLM and agentic workflows using LM Studio
- Validates that **open-weight models** can deliver reliable intelligence, agentic behavior, and multimodal understanding in a fully local setup
- Combines **benchmark-driven evaluation, runtime optimization, and hands-on experimentation** for real-world relevance
- Bridges the gap between **experimental research and practical application** through agent-driven and vision-language use cases
- Highlights **local AI architectures** as scalable, controllable, and cost-effective, with flexibility for future expansion