Final Project

LLM Council Refactoring & Local Deployment

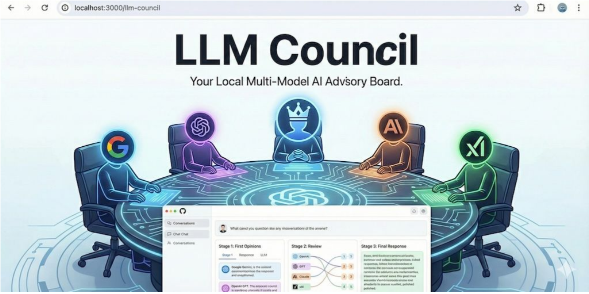
The LLM Council by Andrej Karpathy is a system where multiple LLMs independently answer a user query, anonymously review and rank each other's responses, and then a designated "Chairman" LLM synthesizes their outputs into a single final answer.

[1][Andrej karpathy "LLM Council" November 23, 2025]

Final Project

What is the LLM Council?

- Multiple LLMs answer the same user query
- Models review and rank each other’s answers anonymously
- A **Chairman LLM** synthesizes the final combined response
- Originally uses **OpenRouter** to access multiple models



[1][Andrej karpathy “LLM Council” November 23, 2025]

Final Project

Council Workflow

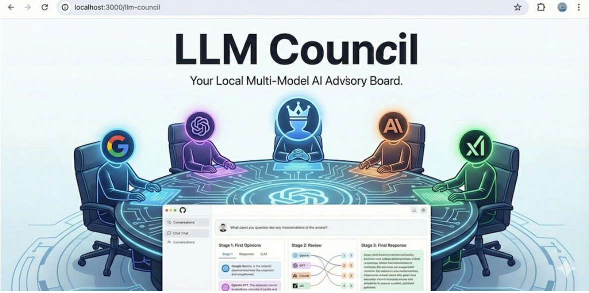
- **Stage 1 – First Opinions**
  - Each LLM answers the question independently
  - Responses shown in a tabbed interface
- **Stage 2 – Review**
  - Each LLM ranks the other models’ answers
  - Anonymized identities prevent bias
- **Stage 3 – Chairman Final Answer**
  - Chairman LLM aggregates all answers
  - Produces the final result

[1][Andrej karpathy “LLM Council” November 23, 2025]

Final Project

Project Goal

The goal is to refactor the project so that the different LLMs in the “LLM Council” run locally on your own machines, using Ollama instead of OpenRouter/OpenAI.



localhost:3000/llm-council

LLM Council

Your Local Multi-Model AI Advśory Board.

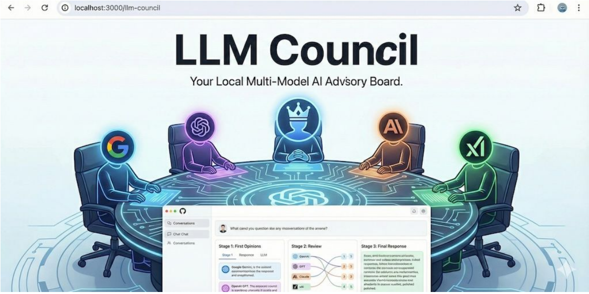
- Replace external LLM provider calls (**OpenRouter**) → with **local LLMs running via Ollama**
- Build a **distributed multi-LLM system** (LLM Council)
- Demonstrate a **working end-to-end prototype** in class

[1][Andrej karpathy “LLM Council” November 23, 2025]

Final Project

Team Requirements

- Groups of **maximum 5 students**
- Clone the GitHub repo by **Andrej Karpathy**
- Work collaboratively on refactoring and deployment
- Final evaluation will occur **during the last TD session**



[1][Andrej karpathy “LLM Council” November 23, 2025]

Final Project

Mandatory Requirements

- Replace **OpenRouter** with **Ollama** for all LLM calls
- Each team member must run **at least one model**
- Models may run:
  - on separate machines, or
  - on the same PC if resources allow
- Machines must communicate via **Ollama REST API**
- Chairman LLM must run on a **separate instance** (ideally separate machine)

[1][Andrej karpathy “LLM Council” November 23, 2025]

Final Project

Evaluation Criteria

- **Code quality** (refactoring, modularization, clean logic)
- **Functionality** (fully working council & chairman workflow)
- **Improvements** added to the original repo
- **Documentation** (README, setup guide, architecture, report)
- **Teamwork & clarity** during live demo

localhost:3000/llm-council

LLM Council

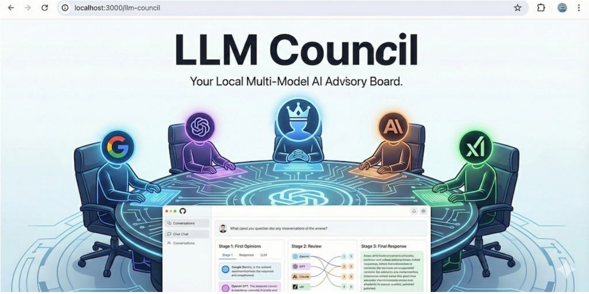
Your Local Multi-Model AI Advsory Board.

[1][Andrej karpathy “LLM Council” November 23, 2025]

Final Project

Final Notes

- Deadline will be announced later on **Moodle**
- Final demo will take place **in class**
- Make sure all machines are tested and connected before evaluation



LLM Council

Your Local Multi-Model AI Advisory Board.

[1][Andrej karpathy “LLM Council” November 23, 2025]

Final Project

Enhancement Ideas

- Backend Enhancement
  - Add Model Health & Status Checks
    - Automatic ping/heartbeat for each remote Ollama endpoint.
    - Display model load time, speed, availability.
  - API Rate & Token Cost Tracking
    - Even with local LLMs, track tokens in/out estimated, cost for cloud models
- Frontend Enhancement Ideas
  - Tab view can be improved with:
    - color-coded model responses
    - collapsible/expandable panels
    - side-by-side comparison mode
    - “diff view” highlighting differences between outputs
  - Model Performance Dashboard: latency per model, token throughput, consistency score, ranking results, etc
  - Dark Mode / Light Mode Simple UI enhancement but improves usability.

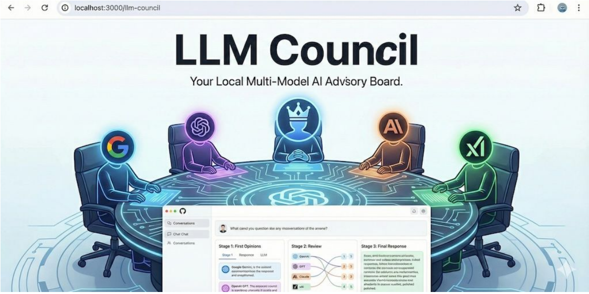
localhost:3000/llm-council
LLM Council
Your Local Multi-Model AI Advisory Board.

[1][Andrej karpathy “LLM Council” November 23, 2025]

Final Project

Ambitious Enhancements

- Drag-and-Drop Council Builder
  - Drag LLM icons into a “council room” UI
  - Assign Chairman role visually
- Plugins or Tool Use
  - Enable the council to call tools (calculator, web search, code executor)



localhost:3000/llm-council

LLM Council

Your Local Multi-Model AI Advisory Board.

[1][Andrej karpathy “LLM Council” November 23, 2025]