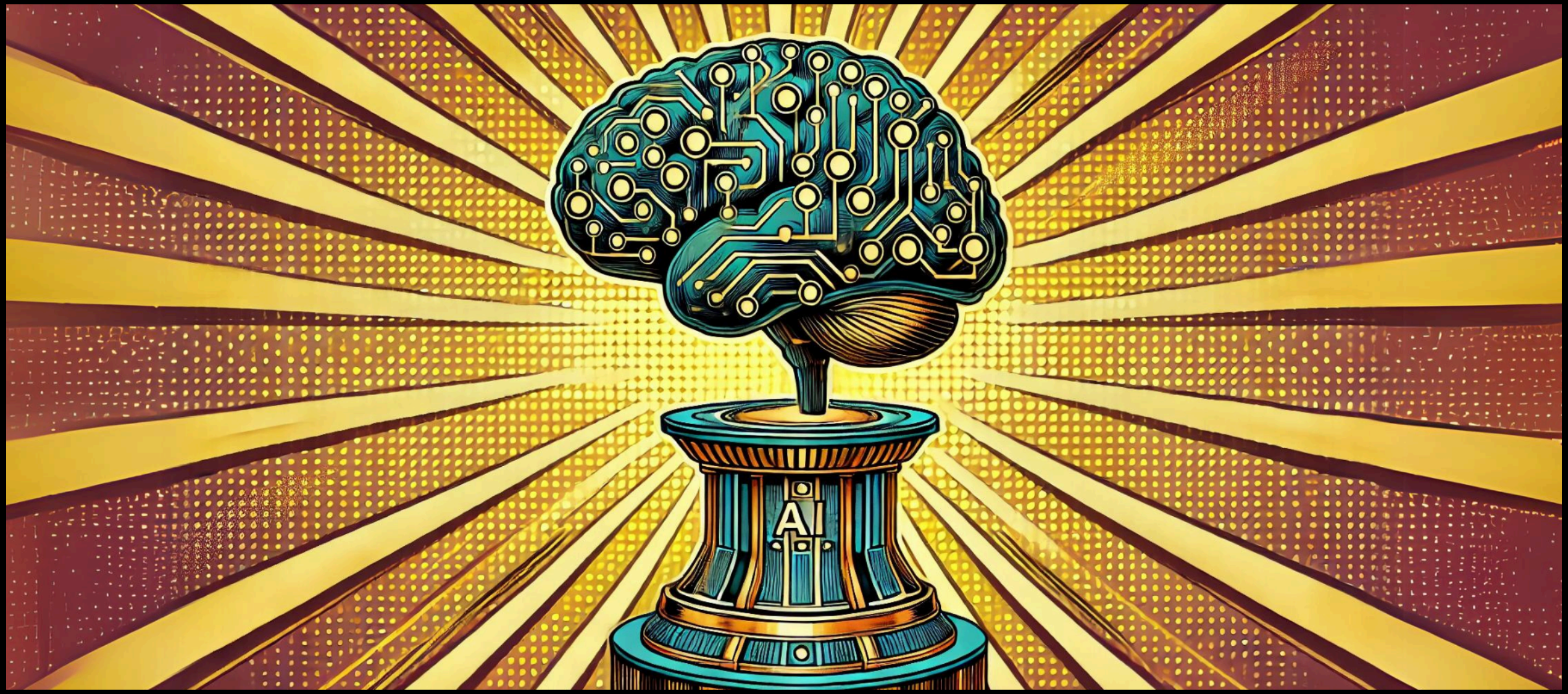
LLM Engineering

MASTER AI & LARGE LANGUAGE MODELS



WEEK 4 DAY 4

Can Open Source keep up?

What you can now do

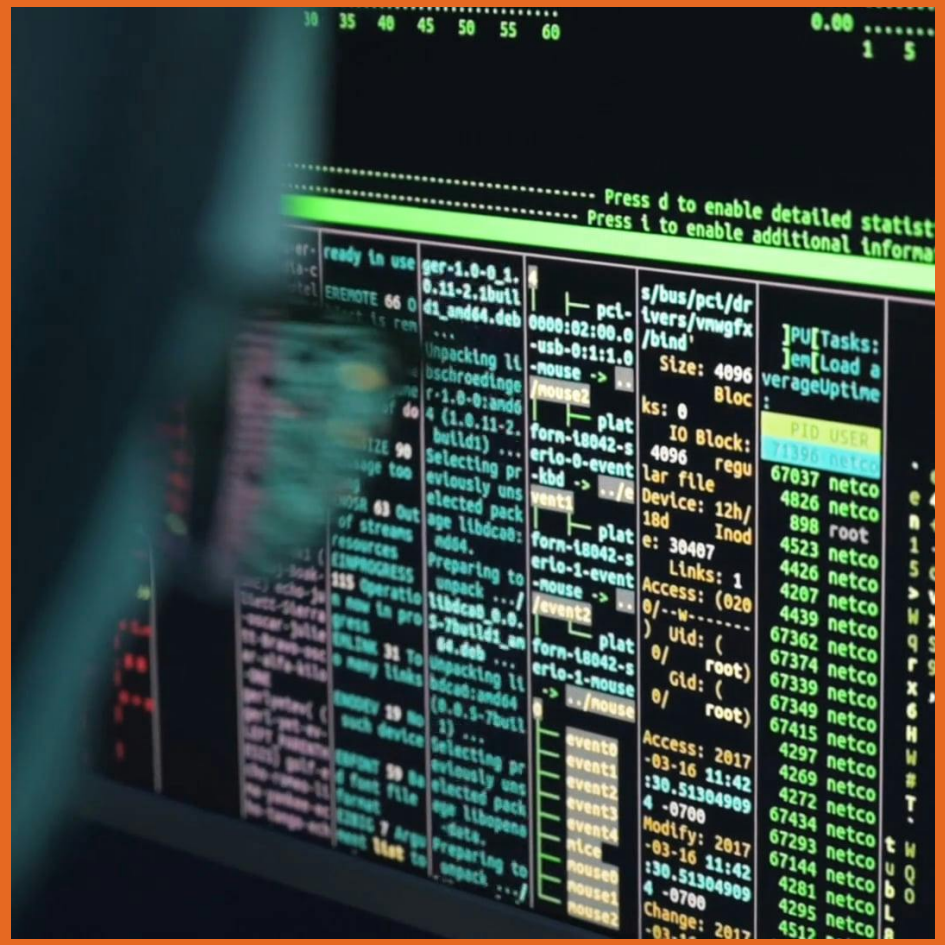
- Code with Frontier Models including AI Assistants with Tools, and with open-source models with HuggingFace transformers
- Confidently choose the right LLM for your project, backed by metrics
- Build solutions that use Frontier models to generate code

After today you'll be able to

- Assess open-source models for coding ability
- Use HuggingFace endpoints to deploy a model
- Build a solution that uses open-source LLMs to generate code

BUSINESS CHALLENGE

Reminder of the challenge



Build a product that converts Python code to C++ for performance

At the last session we used GPT-4o and Claude

Both wrote the same optimized C++ code for a simple program to estimate pi. It ran 100 times faster.

With the harder problem, GPT-4o optimized the code for a 40X speedup, but Claude rewrote the algorithm for a spectacular 60,000X gain!

BUSINESS CHALLENGE

Selecting the model using the Deep Code Leaderboard

T	Model	Win Rate	humaneval-python	java	javascript	cpp
🔶 EXT	OpenCodeInterpreter-DS-33B	55.83	75.23	54.8	69.06	64.47
🔶 EXT	Nxcode-CQ-7B-orpo	55.42	87.23	60.91	71.69	68.04
🔶	CodeQwen1.5-7B-Chat	55.08	87.2	61.04	70.31	67.85
🔶 EXT	CodeFuse-DeepSeek-33b	54.33	76.83	60.76	66.46	65.22
🔶 EXT	DeepSeek-Coder-33b-instruct	52	80.02	52.03	65.13	62.36
🔶 EXT	Artigenz-Coder-DS-6.7B	51.5	70.89	56.84	66.16	59.75
🔶 EXT	DeepSeek-Coder-7b-instruct	50.33	80.22	53.34	65.8	59.66
🔶 EXT	OpenCodeInterpreter-DS-6.7B	49.67	73.2	51.41	63.85	60.01
🔶	Phind-CodeLlama-34B-v2	49.12	71.95	54.06	65.34	59.59
🔶	Phind-CodeLlama-34B-v1	47.92	65.85	49.47	64.45	57.81
🔶	Phind-CodeLlama-34B-Python-v1	46.35	70.22	48.72	66.24	55.34
🟢	CodeQwen1.5-7B	45.08	50.79	42.15	50.07	48.35

BUSINESS CHALLENGE

Results:

```
Simple test

CodeQwen1.5-7B gave same C++ code:

Result: 3.141592658590
Execution Time: 0.073004417000 seconds
100X Speedup

Hard test

Unfortunately Qwen failed to reproduce the answer
```

PROGRESS REPORT

This time the Frontier Models win

What you can now do

- Code with Frontier Models including AI Assistants with Tools, and with open-source models with HuggingFace transformers
- Confidently choose the right LLM for your project, backed by metrics
- Use Frontier and open-source LLMs to generate code

Next time you'll be able to

- Compare performance of open-source and closed source models
- Describe different commercial use cases for code generation
- Build solutions that use code generation for diverse tasks