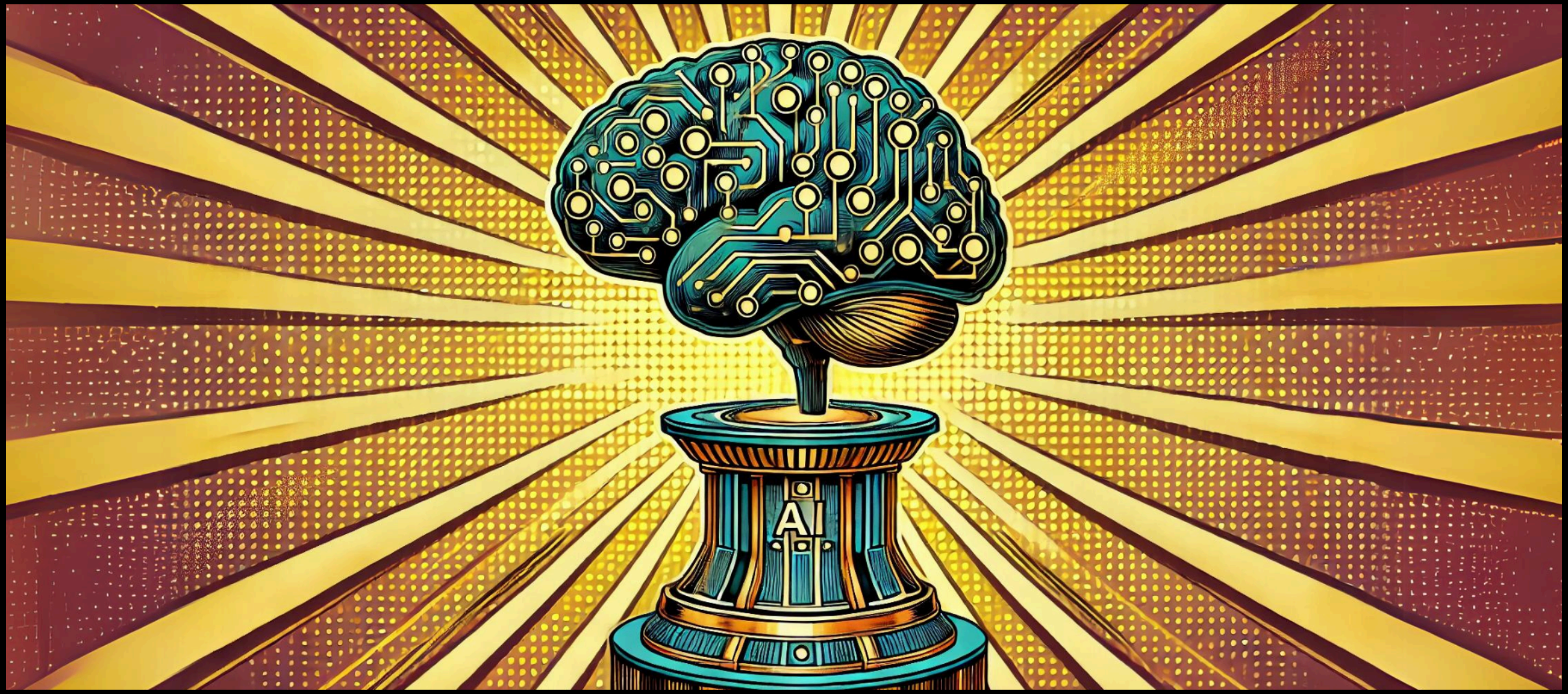
LLM Engineering

MASTER AI & LARGE LANGUAGE MODELS



WEEK 4 DAY 3

Time for code generation

What you can now do

- Code with Frontier Models including AI Assistants with Tools
- Build solutions with open-source LLMs with HuggingFace transformers
- Confidently choose the right LLM for your project, backed by metrics

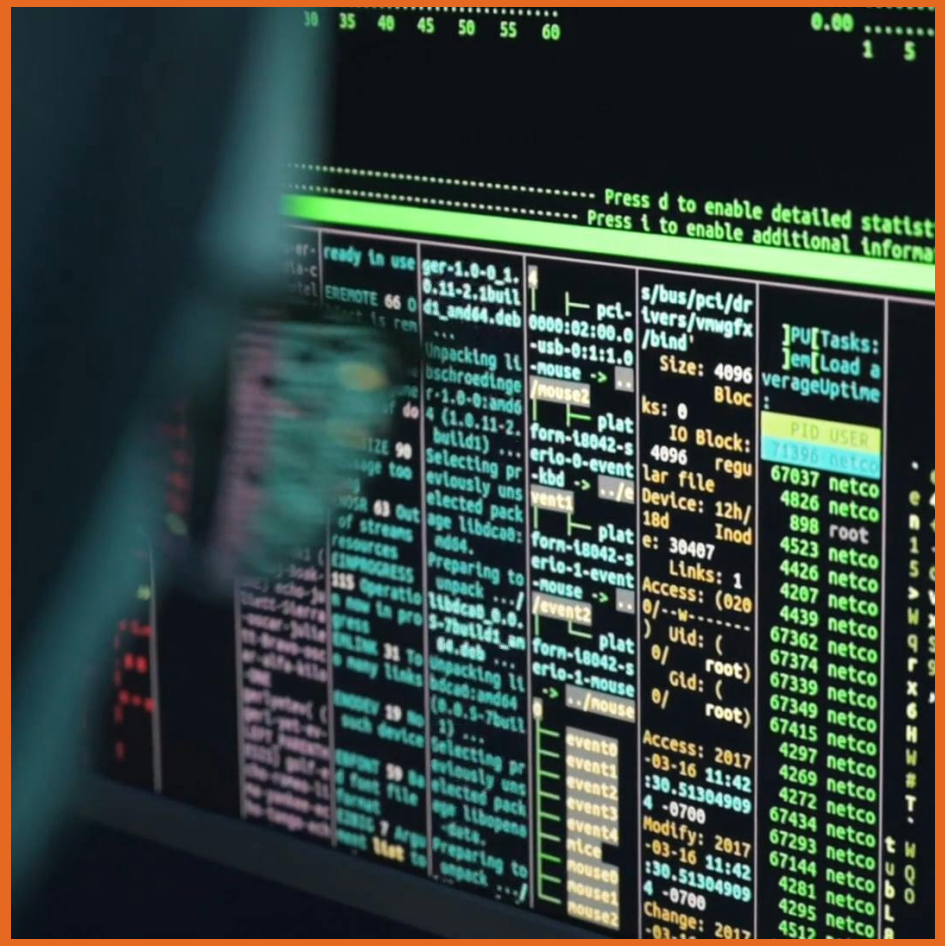
---

After today you will be able to

- Assess models for coding ability
- Use Frontier models to generate code
- Build a solution that uses LLMs to generate code

BUSINESS CHALLENGE

Reminder of the challenge



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

Today we will start with the Frontier Model solution

BUSINESS CHALLENGE

Testing the models

We'll use this approach to test our models

"Please reimplement this Python code in C++ with the fastest possible implementation for an M1 Mac. Only respond with the C++ code. Do not explain your implementation. The only requirement is that the C++ code prints the same result and runs fast."

```python
import time

def calculate(iterations, param1, param2):
    result = 1.0
    for i in range(1, iterations+1):
        j = i * param1 - param2
        result -= (1/j)
        j = i * param1 + param2
        result += (1/j)
    return result

start_time = time.time()
result = calculate(100_000_000, 4, 1) * 4
end_time = time.time()

print(f"Result: {result:.12f}")
print(f"Execution Time: {(end_time - start_time):.6f}
seconds")
```

BUSINESS CHALLENGE

Results from our simple test

```
From Python code:

Result: 3.141592658589
Execution Time: 7.757589 seconds

GPT-4o & Claude-3.5-Sonnet gave same C++ code:

Result: 3.141592658590
Execution Time: 0.073004417000 seconds

100X Speedup
```

Now let's try something harder...

BUSINESS CHALLENGE

Spectular results from our hard test

```
From Python code:

Total Maximum Subarray Sum (20 runs): 10980
Execution Time: 25.420952 seconds

GPT-4o C++ code:

Total Maximum Subarray Sum (20 runs): 10980
Execution Time: 0.642157 seconds
40X Speedup

Claude-3.5-Sonnet code:
Total Maximum Subarray Sum (20 runs): 10980
Execution Time: 0.000409 seconds
60,000X+ Speedup!
```

WEEK 4 DAY 3

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 the next session, you'll be able to

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