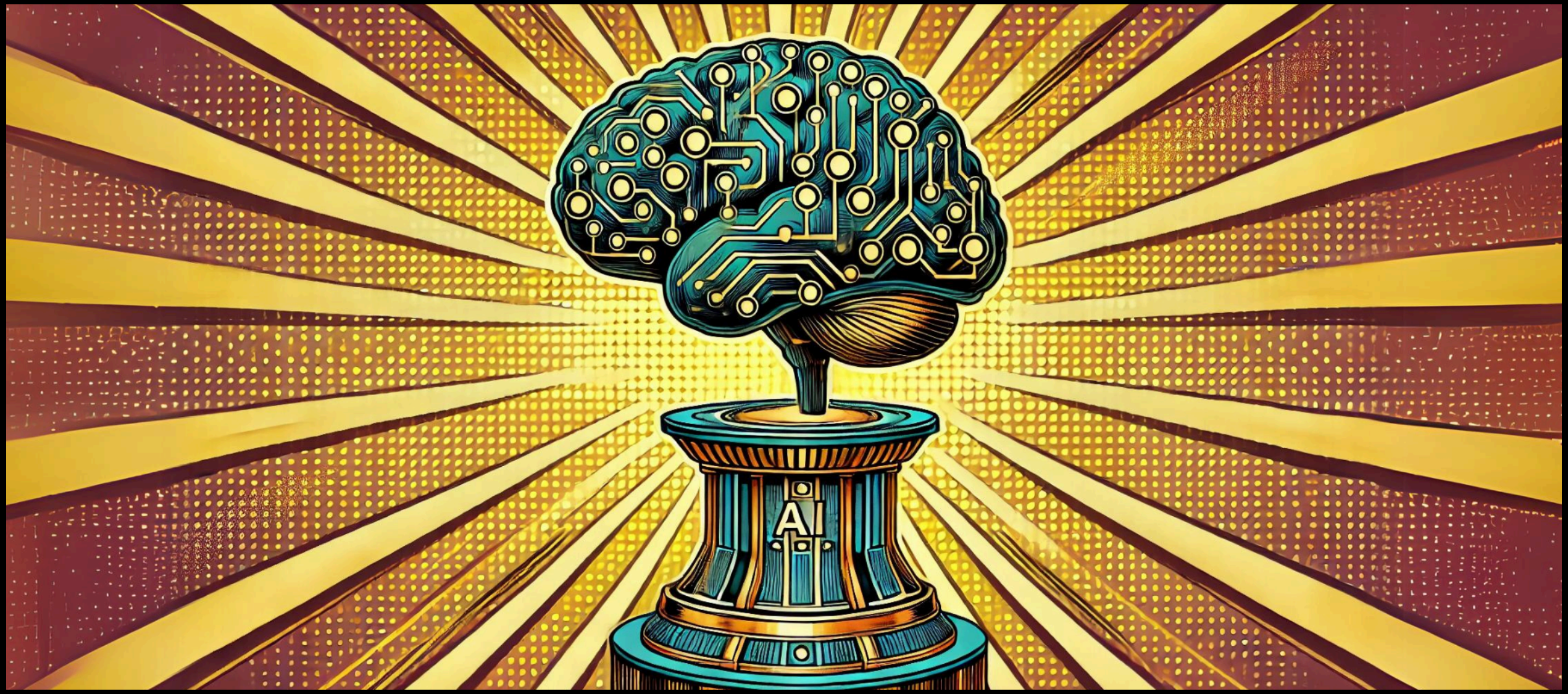
LLM Engineering

MASTER AI & LARGE LANGUAGE MODELS



LAST DAY OF WEEK 4!

Score 1 for the Frontier

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

By end of today 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

How to evaluate the performance of a Gen AI solution?

This is perhaps the single most important question you will face

Model-centric or Technical Metrics

Loss (eg cross-entropy loss)

Perplexity

Accuracy

Precision, Recall, F1

AUC-ROC

Easiest to optimize with

Business-centric or Outcome Metrics

KPIs tied to business objectives

ROI

Improvements in time, cost or resources

Customer satisfaction

Benchmark comparisons

Most tangible impact

BUSINESS CHALLENGE

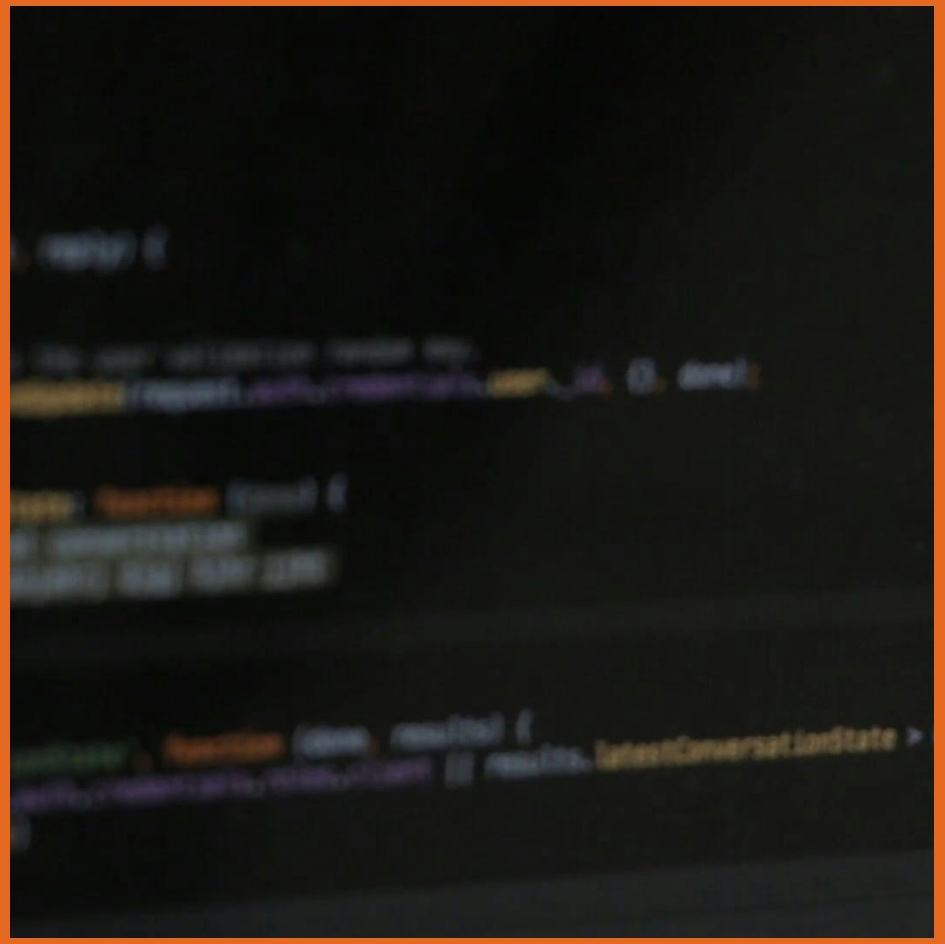
In our case we had simple business-centric metrics

Performance of C++ solution with identical results

Claude-3.5-Sonnet is the winner, followed by GPT-4o, followed by CodeQwen.

But remember that Qwen has 7B parameters; its closed-source cousins have more than 1T!

For everyday problems, Qwen is more than capable of converting Python to Optimized C++ code.



WEEK 4 CHALLENGES

Major challenges for you!

For this high performance coding solution

- Try adding Gemini to the Closed Source mix
- Try more open-source models such as CodeLlama and StarCoder, and see if you can get CodeGemma to work

---

3 new, exciting code generation ideas

- A code tool that automatically adds docstring / comments
- A code gen tool that writes unit test cases
- A code generator that writes trading code to buy and sell equities in a simulated environment, based on a given API

END OF WEEK 4

WAIT... is it... FIFTY PERCENT?!

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 to generate code with Frontier and open-source LLMs

---

By end of the next session you'll be able to

- Explain the big idea behind Retrieval Augmented Generation (RAG)
- Walk through the high level flow for adding expertise to queries
- Implement a toy version of RAG without vector databases.. yet