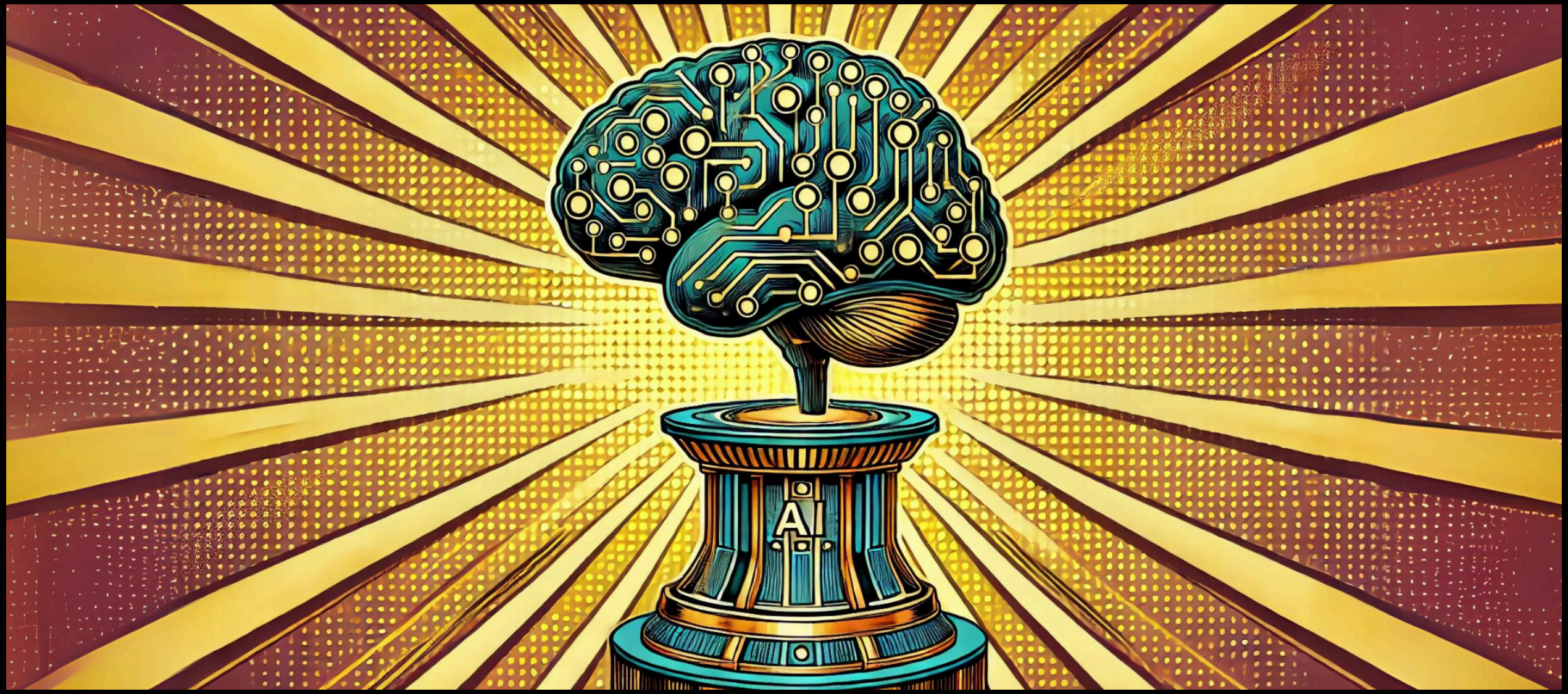
LLM Engineering

MASTER AI & LARGE LANGUAGE MODELS



WEEK 3 DAY 3

Sleeves rolled up

What you can now do

- Confidently code with Frontier Models
- Build a multi-modal AI Assistant with Tools
- Use HuggingFace pipelines for a wide variety of inference tasks

---

Today we get into the lower level HuggingFace API

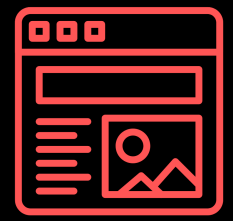
- Create tokenizers for models
- Translate between text and tokens
- Understand special tokens and chat templates

Introducing the Tokenizer

Maps between Text and Tokens for a particular model

Translates between Text and Tokens with encode() and decode() methods

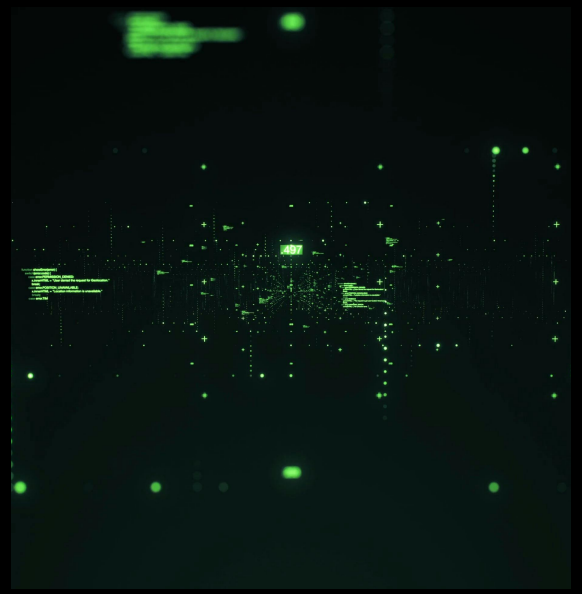
Contains a Vocab that can include special tokens to signal information to the LLM, like start of prompt



Can include a Chat Template that knows how to format a chat message for this model

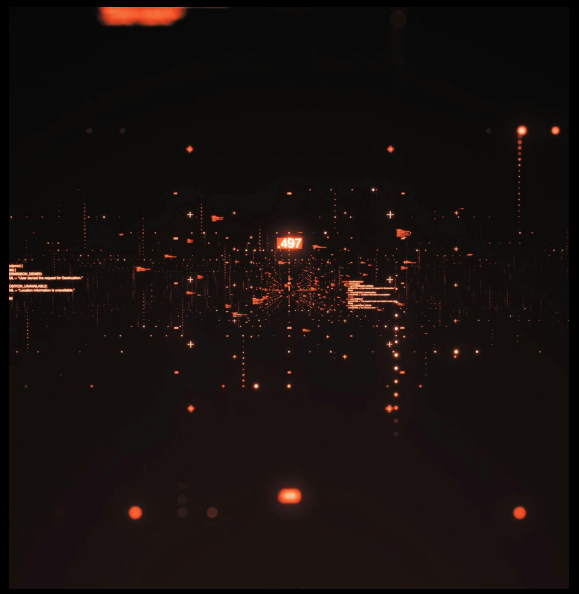
The Tokenizers for key models

We will experiment with tokenizers for a variety of open-source models



Llama 3.1

Meta led the way

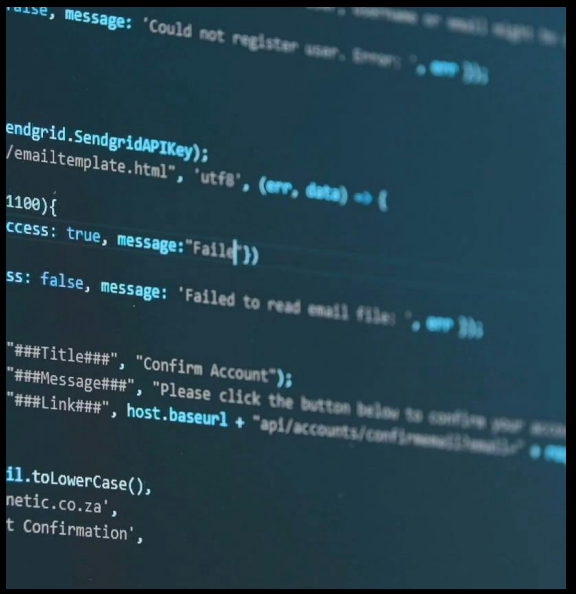


Phi 3

Microsoft's entrant

Qwen2

Leader from Alibaba Cloud



Starcoder2

Coding model

WEEK 3 DAY 3

Ready for transformers..

What you can now do

- Confidently code with Frontier Models
- Build a multi-modal AI Assistant with Tools
- Use HuggingFace pipelines and tokenizers

After next time you will be able to use the powerful APIs

- Work with HuggingFace lower level APIs
- Use HuggingFace models to generate text
- Compare the results across 5 open source models