LLM Talk Feb 16th

jacob

February 13, 2024

Who am I?

Jacob Windle

- Sr. Software Engineer at Instructure
- Sr. Opinion Holder at TriDev

What is an LLM?

- “BS Generator” ~ Stallman
- “The thing that does my homework” ~ Kids these days

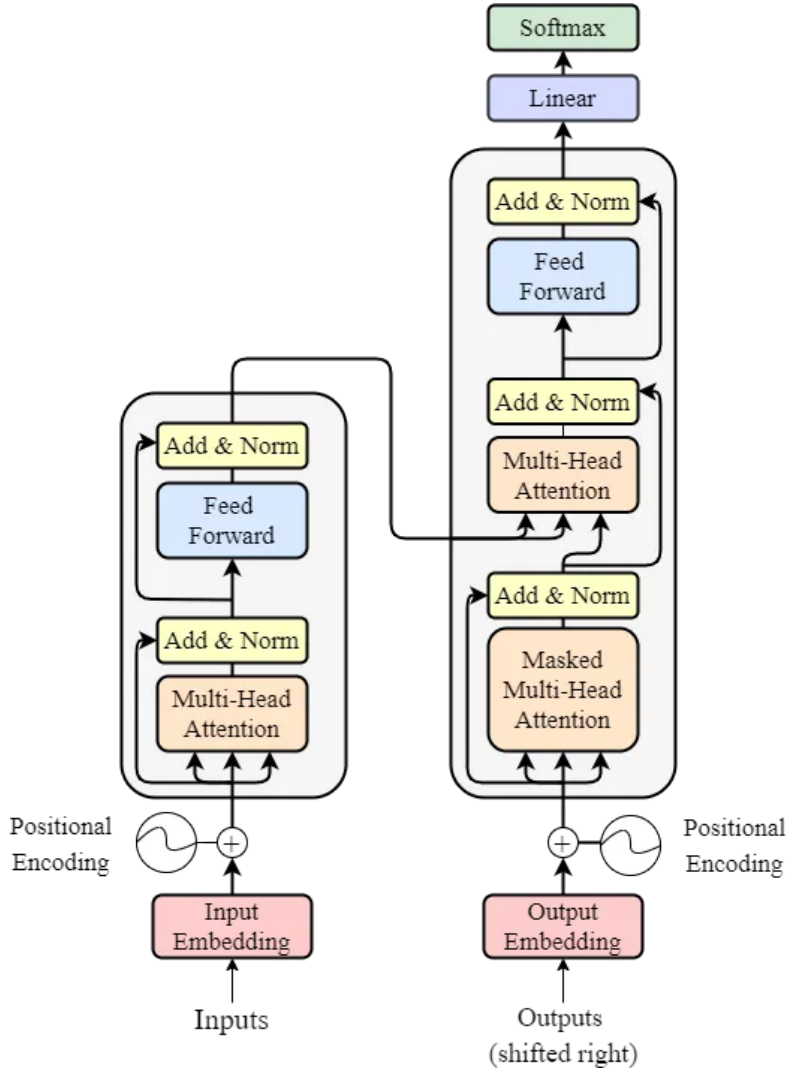
What is an LLM Actually?

- Stands for "Large Language Model"
- Fantastic at generating text (and images, and soon video).
- Fantastic at messy, unstructured data (text).
- Downright magical at times.

LLM Foundations

- Transformers: `https://arxiv.org/abs/1706.03762`
- Attention is All You Need paper.

Transformers in a Nutshell



Transformers

- Novel architecture by Google to string together deep nets.
- Enabled the rise of LLMs
- Based on simple yet complex “attention” concept.
- Attention looks for meaning between tokens, how does one relate to rest of text?
- Input and output come from embeddings vectors.

LLM demo

- OpenAI + Notebooks

What tasks are LLMs well suited for?

- Language Comprehension
- Text Classification
- Summarization
- Question answering
- Text generation
- Embeddings-based search

Do I need OpenAI?

- No, you can use plenty of open source LLMs to get the job done.
- Mistral
- Llama2
- NLLB (My work)

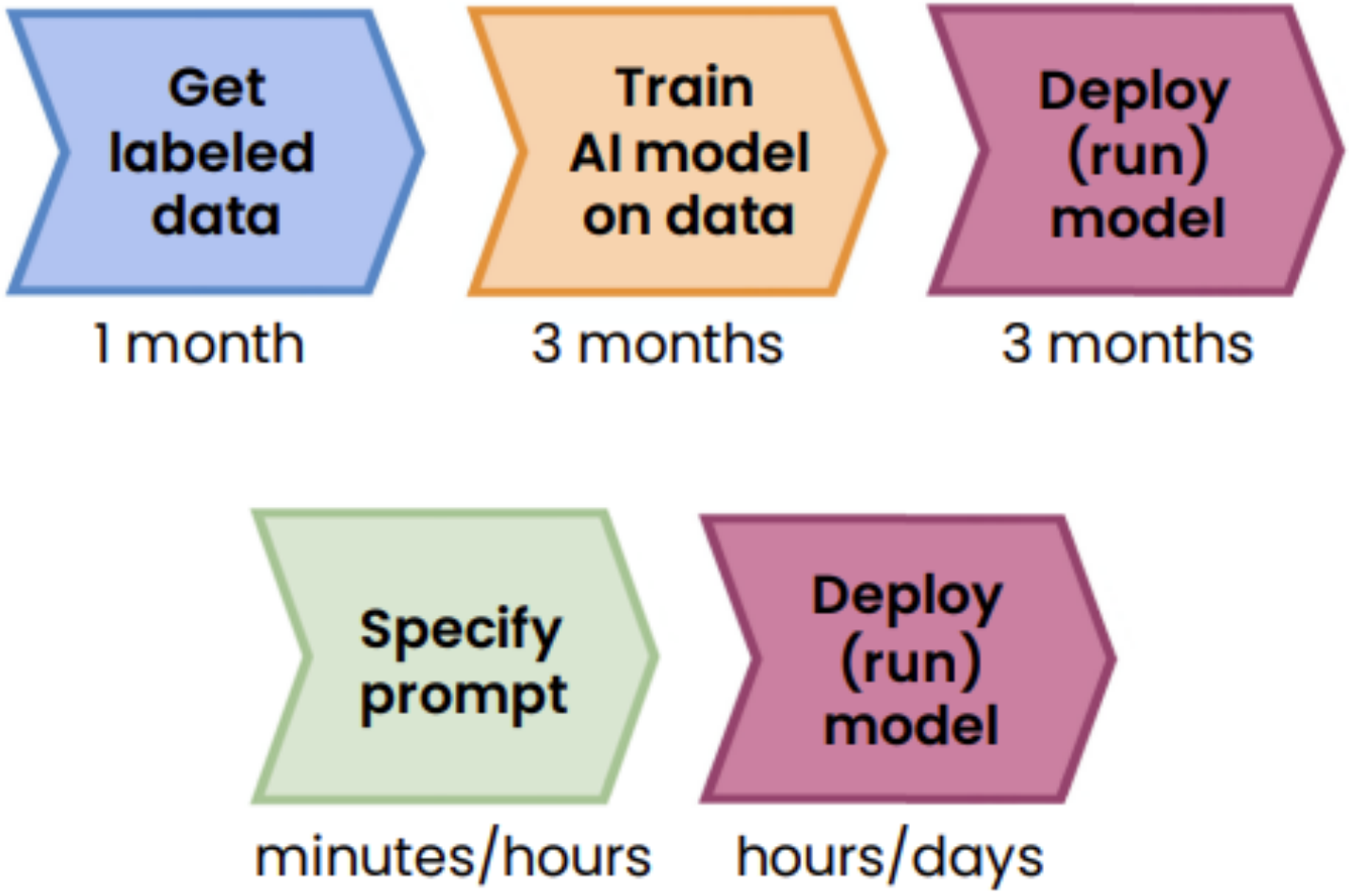
Getting Started without OpenAI

- use a local inference server like ollama (`https://ollama.com/`)
- test things out in chat interface
- use ollama inference server to develop applications.

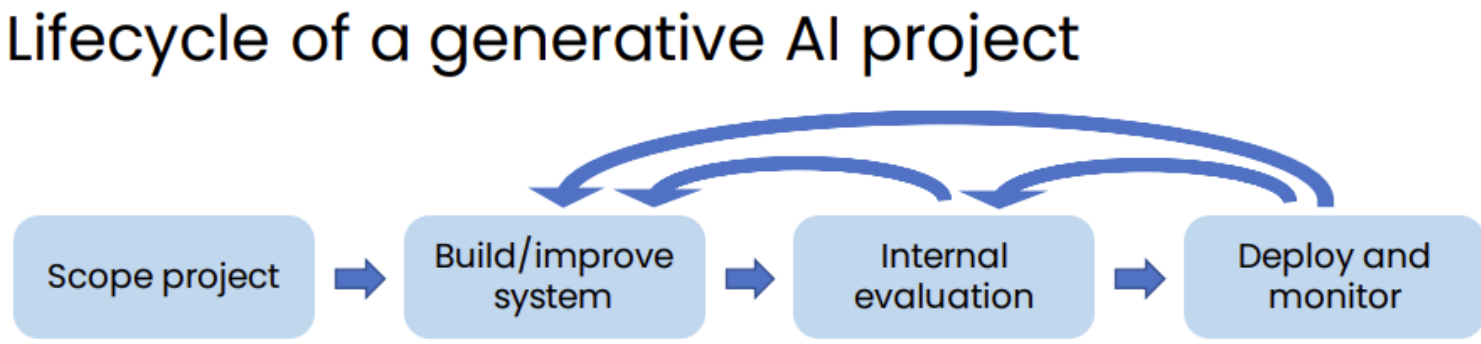
Ollama demo

- Demo Ollama

GenAI LifeCycle



GenAI Iterations



My Typical Prototype Cycle

- Have idea
- Prototype idea with Ollama or ChatGPT
- Start with single prompt, can I engineer it to get the results I want?
- Start adding examples, can I get good results with a few "shots"?

Past the Prototype Cycle

- Fine tune necessary? (probably not)
- Build my own LLM? (definitely not)

My current research

- Working with NLLB to run inference in browser.
- Have client-side translation for users of Canvas.