Doing It Ourselves: A (very) Brief Introduction to Locally Hosted Large Language Models and their Customization

John Fink

McMaster University

May 1st, 2024

What happened in 2017?

Attention Is All You Need

Ashish Vaswani[*]
Google Brain
avaswani@google.com

Noam Shazeer[*]
Google Brain
noam@google.com

Niki Parmar[*]
Google Research
nikip@google.com

Jakob Uszkoreit[*]
Google Research
usz@google.com

Llion Jones[*]
Google Research
llion@google.com

Aidan N. Gomez[*] [†]
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser[*]
Google Brain
lukaszkaiser@google.com

Illia Polosukhin[*] [‡]
illia.polosukhin@gmail.com

OpenAI and then... not so OpenAI

- 2018 - GPT-1 (117m parameters)
- 2019 - GPT-2 (1.5b parameters)
- (these two are both freely available, but ... not very good)
- 2020 - GPT-3 (175b parameters)
- 2023 - GPT-4 (??? parameters (probably trillions))
- (these two are... not freely available)

Enter... Meta?

- February 2023 - Meta/Facebook announces LLaMa, a series of models downloadable under restrictive access for researchers only.
- ...
- ...leaked within a week of being announced.

- Since then:
- LLaMa 2, 3 under more permissive licenses
- **Tons** of derivatives, relatives, etc (over 620k models on HuggingFace)
- Ranging from models on par with GPT-4 right down to ones that run on a cellphone or Raspberry Pi

- Which means, as libraries, with the equipment we have now or can get, we can potentially:
- **Run** stock models in house, and use low-compute technologies like RAG (Retrieval Augmented Generation) to return sensible, largely non-hallucinatory results.
- **Train** existing smaller models to create our own models with our own unique data, on commodity hardware.
- **Deploy** these models in-house, with less worry about uncontrollable costs and (lack of) privacy.

We have a real opportunity to do useful work in this area that respects our core values, like patron privacy and democratization of technology.

- Do it yourself with tools like these:
- Ollama (https://ollama.com/), for running models locally (smaller models will run on just about any modern laptop)
- WARC-GPT (Harvard Library, https://github.com/harvard-lil/warc-gpt), for applying RAG to WARC format files.