Teaching a Dog to Catalog: An Abbreviated History of Large Language Models and an Inquiry as to Whether They Can Replace Us

John Fink

McMaster University

October 24, 2023

John Fink
Digital Scholarship Librarian
McMaster University

A (series of) disclaimers:

I HAVE NO IDEA WHAT I'M DOING

Important things I **don't** address

- Copyright

Important things I **don't** address

- Copyright
- Pedagogical implications

Important things I **don't** address

- Copyright
- Pedagogical implications
- Ethics

What is *randomness*?

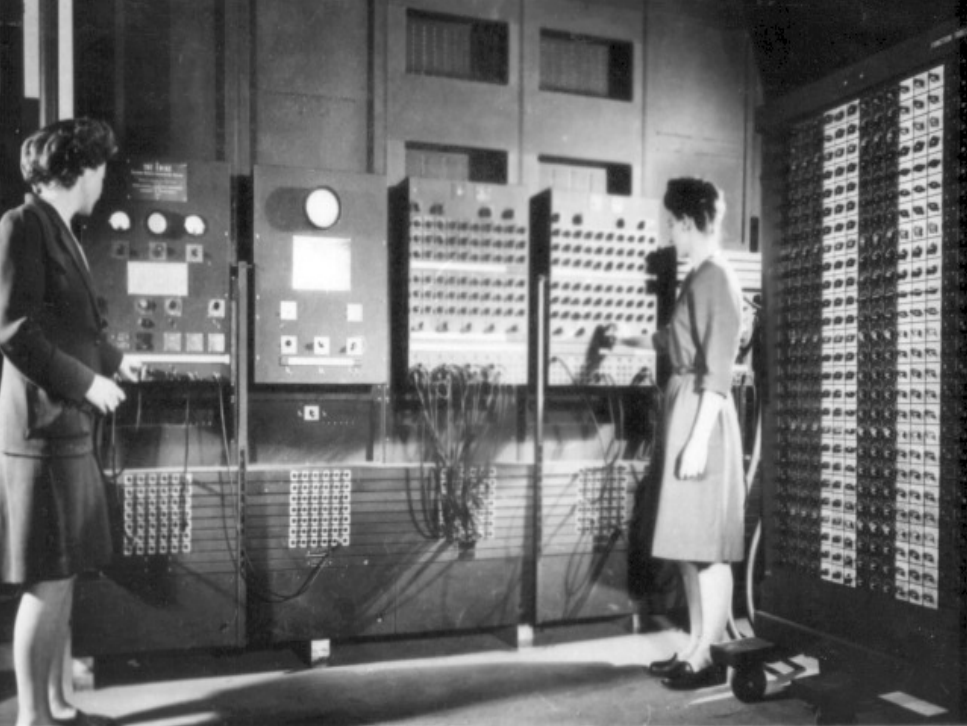A little conversation about the weather.

Change

Stormy RAIN FAIR Very Dry

29 30 31
28 27 26

T.P. WRIGHT

40 & 40E HOLLOWAY R. N

FALLS
FOR
WET or MORE WIND
S.W. & S.E. S.W.

RISES
FOR
DRY or LESS WIND
N.E. or N.W. N.E.

ANEROID BAROMETER

Why I say "Large Language Model" and not "AI"

So, about 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

Abstract

2017-now! Right now!

- 2017 - "Attention Is All You Need" paper

2017-now! Right now!

- 2017 - "Attention Is All You Need" paper
- 2018 - "Improving Language Understanding by Generative Pre-Training" paper

2017-now! Right now!

- 2017 - "Attention Is All You Need" paper
- 2018 - "Improving Language Understanding by Generative Pre-Training" paper
- 2020 - "Language Models are Few-Shot Learners" paper (GPT-3)

2017-now! Right now!

- 2017 - "Attention Is All You Need" paper
- 2018 - "Improving Language Understanding by Generative Pre-Training" paper
- 2020 - "Language Models are Few-Shot Learners" paper (GPT-3)
- 2022 - InstructGPT, and then ChatGPT

2017-now! Right now!

- 2017 - "Attention Is All You Need" paper
- 2018 - "Improving Language Understanding by Generative Pre-Training" paper
- 2020 - "Language Models are Few-Shot Learners" paper (GPT-3)
- 2022 - InstructGPT, and then ChatGPT
- 2023 - and then....

"Progress is now moving so swiftly that every few weeks the state-of-the-art is changing or models that previously required clusters to run now run on Raspberry PIs."
– https://github.com/brexhq/prompt-engineering

Important concepts for GPT and other models

- Context Window and Tokens

Important concepts for GPT and other models

- Context Window and Tokens
- Few-Shot / No-Shot

Important concepts for GPT and other models

- Context Window and Tokens
- Few-Shot / No-Shot
- Parameters

Important concepts for GPT and other models

- Context Window and Tokens
- Few-Shot / No-Shot
- Parameters
- Training

Important concepts for GPT and other models

- Context Window and Tokens
- Few-Shot / No-Shot
- Parameters
- Training
- The Prompt, aka "Programming for English Majors"

Important concepts for GPT and other models

- Context Window and Tokens
- Few-Shot / No-Shot
- Parameters
- Training
- The Prompt, aka "Programming for English Majors"
- And the Random Seed.

- Context Window is the "memory" of an LLM

- Context Window is the "memory" of an LLM
- And Tokens – words, roughly – fill up that "memory"

- Context Window is the "memory" of an LLM
- And Tokens – words, roughly – fill up that "memory"
- And the *response* also takes tokens.

Few-Shot / No-Shot

- *Few-Shot* – a few examples to "teach" an LLM, such as:

Few-Shot / No-Shot

- *Few-Shot* – a few examples to "teach" an LLM, such as:
- "I hate it when my phone battery dies." - negative

Few-Shot / No-Shot

- *Few-Shot* – a few examples to "teach" an LLM, such as:
- "I hate it when my phone battery dies." - negative
- "My day has been great!" - positive

Few-Shot / No-Shot

- *Few-Shot* – a few examples to "teach" an LLM, such as:
- "I hate it when my phone battery dies." - negative
- "My day has been great!" - positive
- "Here is an article." - neutral

Few-Shot / No-Shot

- *Few-Shot* – a few examples to "teach" an LLM, such as:
- "I hate it when my phone battery dies." - negative
- "My day has been great!" - positive
- "Here is an article." - neutral
- "This presentation is going fantastic!!!!" - positive(ly optimistic)

Few-Shot / No-Shot

- *Few-Shot* – a few examples to "teach" an LLM, such as:
- "I hate it when my phone battery dies." - negative
- "My day has been great!" - positive
- "Here is an article." - neutral
- "This presentation is going fantastic!!!!" - positive(ly optimistic)
- And *No-Shot* is exactly what you think it is.

Training

- Usually done on text corpuses

Training

- Usually done on text corpuses
- The Pile (825GiB), Github, ShareGPT, etc.

Training

- Usually done on text corpuses
- The Pile (825GiB), Github, ShareGPT, etc.
- And other terms like RLHF (Reinforcement Learning from Human Feedback)

Training

- Usually done on text corpuses
- The Pile (825GiB), Github, ShareGPT, etc.
- And other terms like RLHF (Reinforcement Learning from Human Feedback)
- The larger the model, the more resources it takes to train or re-train.

Parameters

- Roughly corresponds to how "Complex" or "Smart" a model is.

Parameters

- Roughly corresponds to how "Complex" or "Smart" a model is.
- (...very roughly)

Parameters

- Roughly corresponds to how "Complex" or "Smart" a model is.
- (...very roughly)
- But *definitely* correlates to resources needed to run the model.

Parameters

- Roughly corresponds to how "Complex" or "Smart" a model is.
- (...very roughly)
- But *definitely* correlates to resources needed to run the model.
- Which is why, say, GPT-4 requires this....

And you can run a 7B model on this....

llama.cpp – https://github.com/ggerganov/llama.cpp

We Have No Moat

In May 2023 a Google internal document was leaked to the public, titled "We Have No Moat, and Neither Does OpenAI". It's worth quoting some bits from it, because it's a doozy.

We Have No Moat

- "While our models still hold a slight edge in terms of quality, the gap is closing astonishingly quickly. Open-source models are faster, more customizable, more private, and pound-for-pound more capable. They are doing things with $100 and 13B params that we struggle with at $10M and 540B. And they are doing so in weeks, not months. This has profound implications for us."

We Have No Moat

- "While our models still hold a slight edge in terms of quality, the gap is closing astonishingly quickly. Open-source models are faster, more customizable, more private, and pound-for-pound more capable. They are doing things with $100 and 13B params that we struggle with at $10M and 540B. And they are doing so in weeks, not months. This has profound implications for us."

- "Indeed, in terms of engineer-hours, the pace of improvement from these models vastly outstrips what we can do with our largest variants, and the best are already largely indistinguishable from ChatGPT. Focusing on maintaining some of the largest models on the planet actually puts us at a disadvantage."

- ....Skynet????

THE FUTURE OF LARGE LANGUAGE MODELS

- ....Skynet????
- ....fully automated luxury communism???

THE FUTURE OF LARGE LANGUAGE MODELS

- ....Skynet????
- ....fully automated luxury communism???
- ....larger context windows?

deep breath....

Any questions?
jfink@mcmaster.ca
m https://glammr.us/@jbfink