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 (much)

- Copyright
- Pedagogical implications, including academic integrity
- Ethics
- Truth
- Beauty
- The state of the world prior to June 12, 2017.

These are VERY IMPORTANT concepts.

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

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 - GPT-4

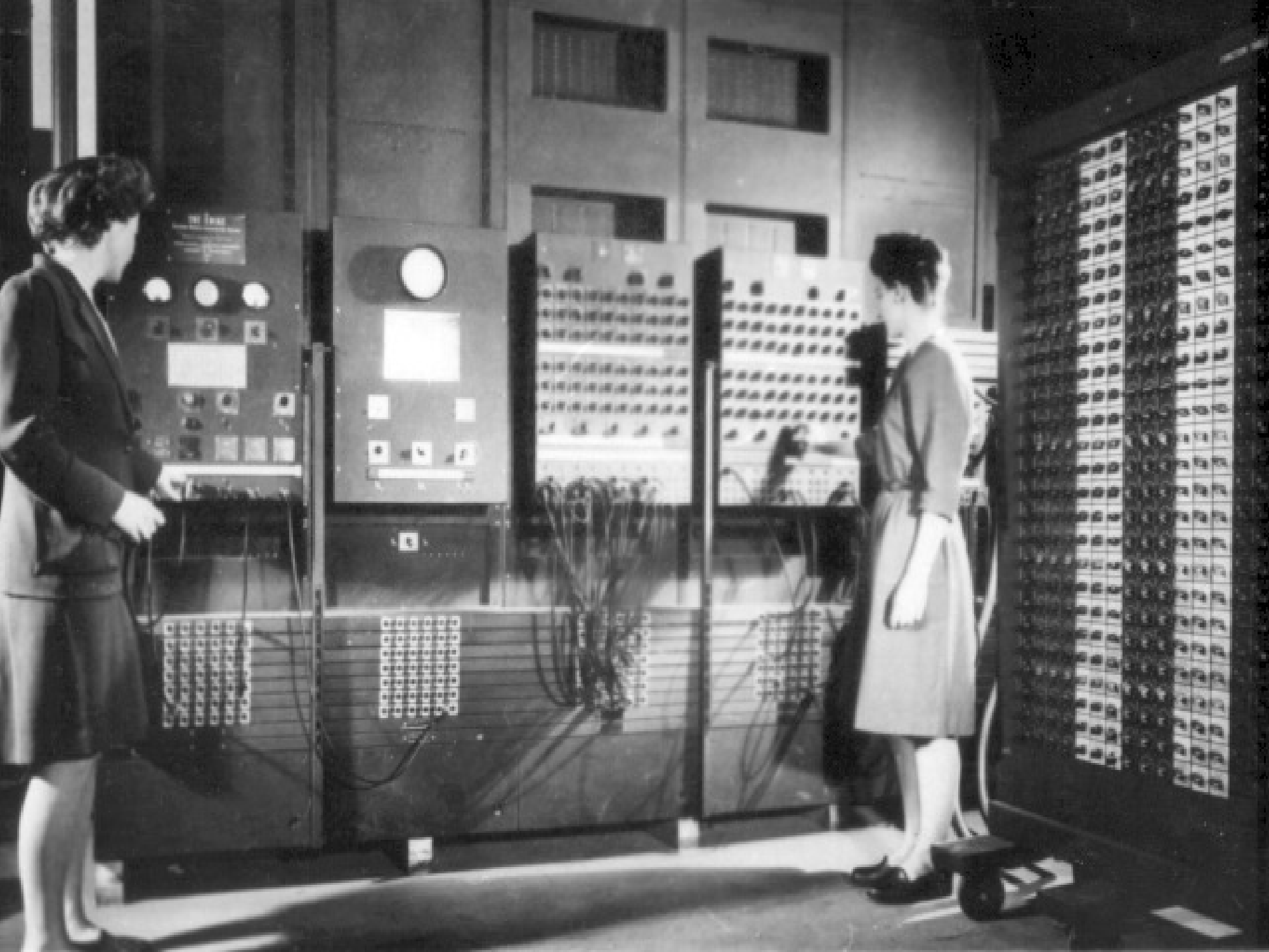
Meta's sort-of open source models

- February 2023 – Meta's LLaMa v1 released
- July 2023 – Meta's LLaMa v2 released

The immediate, post-2017 frantic present is a weird syncretism of Google and OpenAI, and to a somewhat lesser extent, Meta/Facebook, with other important minor players.

A little conversation about the weather.


Stormy
Rain
Change
Fair
Very Dry
28
29
30
31
27
26
T. P. WRIGHT
ANEROID BAROMETER



Right now, in late 2023, LLMs are sort of midway between weather modeling and...


Eat up
Martha.

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
- And Tokens – words, roughly – fill up that "memory"
- And the *response* also takes tokens.

More about tokens and the context window

- *Tokens* and *context windows* are one of the big limiting factors in nearly every present-day LLM
- But *it seems like every day* there is a new paper detailing some new method to get millions of tokens in context
- So who knows?

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
- The Pile (825GiB), Github, ShareGPT, etc.
- (cough) books3....
- 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.
- (...very roughly)
- But *definitely* correlates to resources needed to run the model.
- Which is why, say, GPT-4 (probably 1.7 trillion) requires this....

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

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

quantization of models

- llama.cpp runs on *quantized* models, which is roughly a compression technique that drastically cuts down the size, and more crucially, the processing power needed to run a model.
- So, Mistral's LLM – a popular open-source model – is roughly 15GB unquantized and pretty much requires dedicated GPU to run well, whereas with the most *extreme* quantization, it can get to about 3GB and can run on a reasonably modern smartphone.
- As running an LLM generally requires fitting the entire model in RAM or VRAM, this can be helpful.
- Quantization is like *lossy compression* – smaller the model, the more tradeoffs, although with less-extreme quantization the differences are fairly negligible.

The Prompt

Asking the right questions

```
== Running in interactive mode. ==
 - Press Ctrl+C to interject at any time.
 - Press Return to return control to LLaMa.
 - To return control without starting a new line, end your input with '/'.
 - If you want to submit another line, end your input with '\'.

Think of a good name for my cat.
A: Whiskers
```

```
== Running in interactive mode. ==
 - Press Ctrl+C to interject at any time.
 - Press Return to return control to LLaMa.
 - To return control without starting a new line, end your input with '/'.
 - If you want to submit another line, end your input with '\'.

A chat between a human and the world's foremost cat namer, Bob.  Bob has been na
ming cats for 15 years, and has won several prestigious awards for his efforts.

You: Hello, Bob.
### Bob: Oh, hello.  I sense you are someone who needs my services. Please descr
ibe your cat.
You: Hello Bob. I've just adopted an adorable orange tabby cat from the SPCA. He
's about a year old, he loves to eat sardines and watch TV, and he can be a litt
le mischevious. Can you think of five good names for him?
### Bob: Certainly! Here are five perfect names for your new feline friend:
1. Sardine Sammy - This name reflects his love for sardines and adds a playful t
ouch to it.
2. Telly Tom - A nod to his fondness for television.
3. Mischief Maxwell - It perfectly captures his mischevious personality.
4. Orange Ollie - An obvious reference to his beautiful orange coat.
5. Tabby Teddy - This name combines both his breed and fur pattern into a cute,
cuddly package.
```

teaching an old dog new... dog stuff

There are roughly *three* ways to get new information into a model.

- Train from scratch
- Consulting external sources (Retrieval Augmented Generation et al)
- Tuning on top of an existing model

Why even do this?

Why even do this?

- Running your own models on modest hardware is (or can be):
- slow
- (more) inaccurate
- annoying

Why even do this?

The obvious:

- Privacy
- Environmental Impact
- Cost

A *small* model tuned with high-quality information may be more effective than a huge, generic model

Possible applications for libraries

OK, so, what about libraries?

Possible applications for libraries

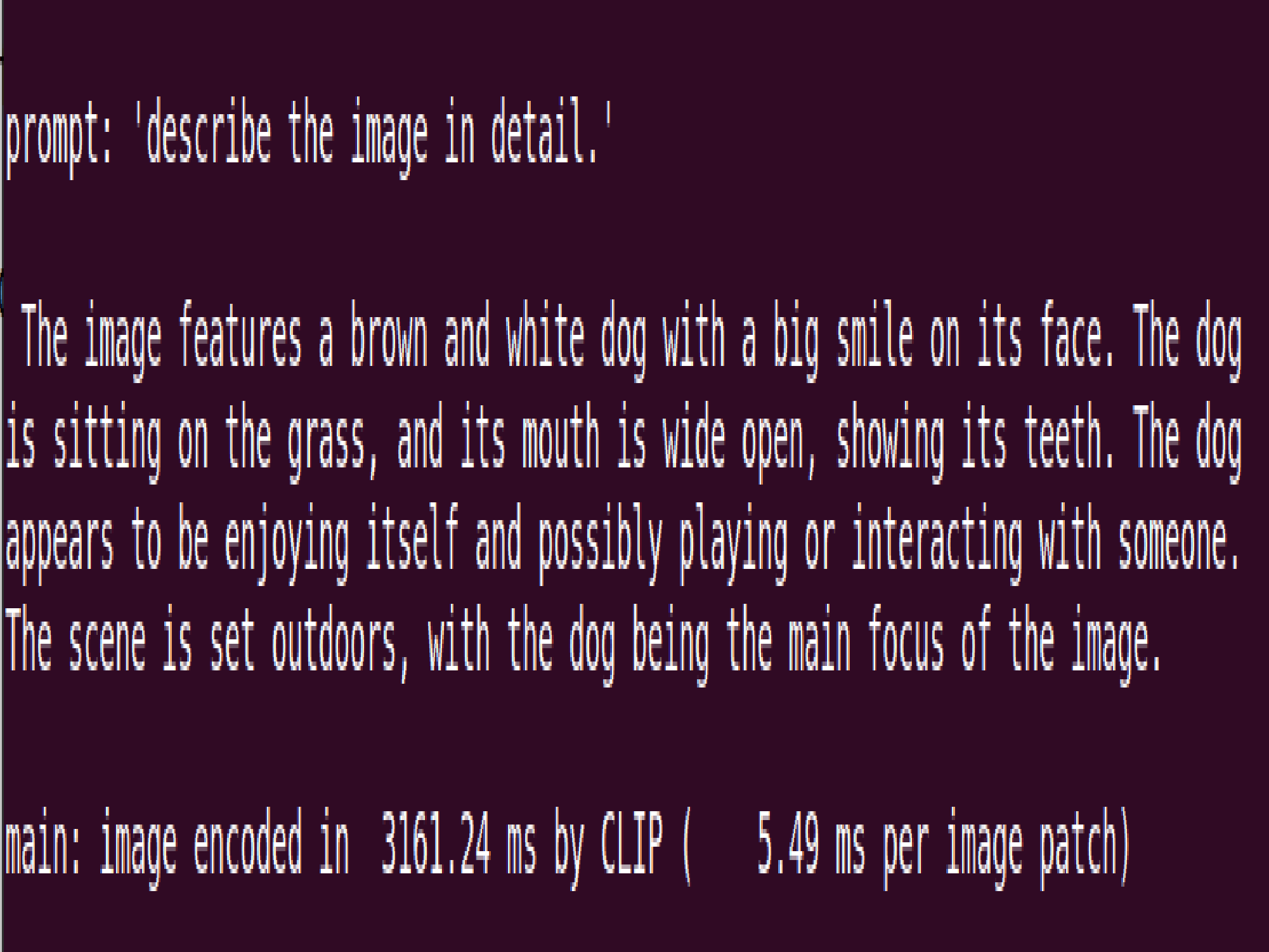
- Chatbot trained on site-specific reference transactions
- Pro: A possibly genuinely useful service, especially during off-hours
- Con: Privacy, hallucinations

Possible applications for libraries

- Programming-specific model tuned with examples of MARC records.
- Pro: Public data
- Con: Complex
- Con: Maybe only useful for edge cases
- Con: Will make catalogers mad if it works

Possible applications for libraries

- Multimodal model (trained on images + text) for image description of unique collections or OCR
- Pro: Public data, genuinely useful
- Con: Would still need editing/vetting due to (sometimes hilarious) inaccuracies


prompt: 'describe the image in detail.'
The image features a brown and white dog with a big smile on its face. The dog is sitting on the grass, and its mouth is wide open, showing its teeth. The dog appears to be enjoying itself and possibly playing or interacting with someone. The scene is set outdoors, with the dog being the main focus of the image.
main: image encoded in 3161.24 ms by CLIP (5.49 ms per image patch)

A COMPUTER

CAN NEVER BE HELD ACCOUNTABLE

THEREFORE A COMPUTER MUST NEVER

MAKE A MANAGEMENT DECISION

prompt: 'describe the image in detail.'

The image is a handwritten page from a book, likely a manual or a guide, discus
sing the importance of a computer in a business setting. The page is filled with
 text, emphasizing the need for a computer to make management decisions. The tex
t is written in a bold, capitalized style, making it easy to read. The page is d
ivided into sections, with each section focusing on a different aspect of the co
mputer's role in management.

main: image encoded in 2890.89 ms by CLIP (5.02 ms per image patch)

Any questions?

jfink@mcmaster.ca

https://glammr.us/@jbfink