Laptop-sized ML for Text, with Open Source

Nick Burch

Berlin Buzzwords 2023

Nick Burch

Director of Engineering

Nick Burch

@Gagravarr

@nick@social.earth.li

Code, Scripts, Slides

github.com/Gagravarr/BBuzz23-LaptopML

All code in slides is taken from here!

Our talk was going to be...

- LLMs - Large Language Models
- Why you can't have one
- Simpler vector-space language models
- Word2vec, GloVe, ELMo and BERT
- How they differ from traditional text relevancy like TF-IDF
- How to run them with Open Source libraries
- How to tweak them to get better relevance or speed or memory
- What you can do even though you don't have a LLM

But then...

Llamas happened

LLaMA

Large Language Model

from Facebook (they're still around!)

LLaMA

ai.facebook.com/blog/large-language-model-llama-meta-ai/

https://arxiv.org/abs/2302.13971v1

LLaMA changed my talk, and the ML-for-text world!

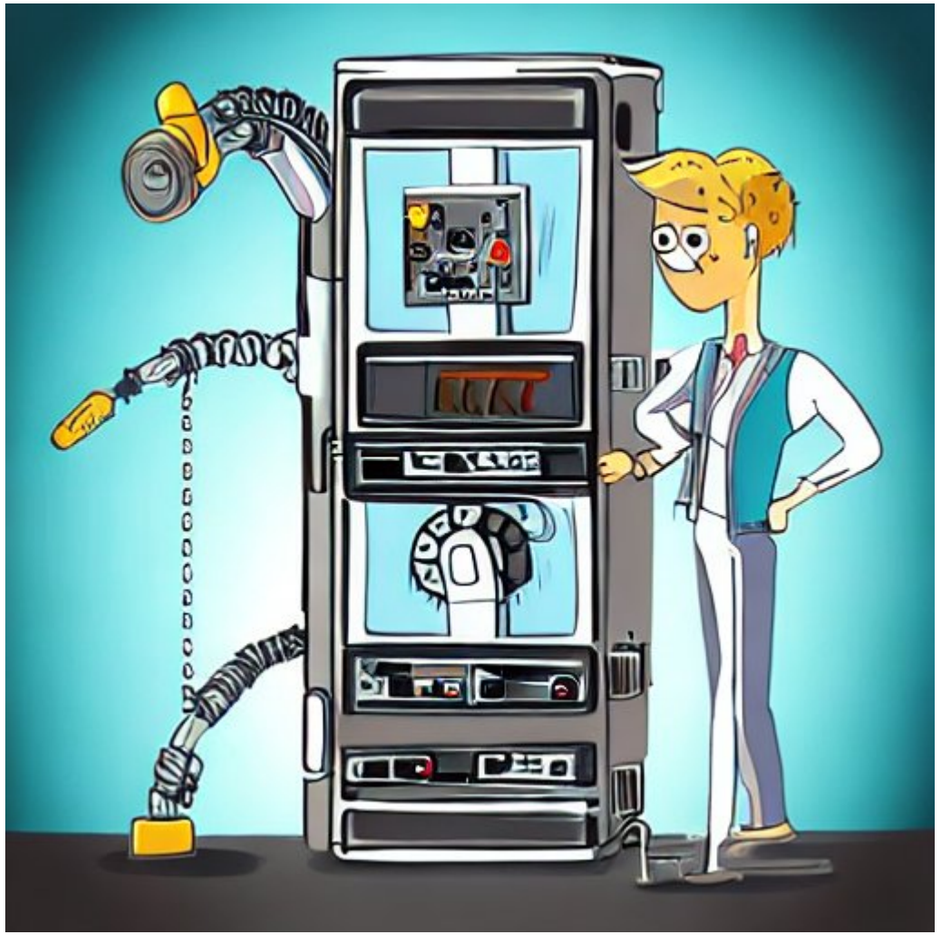
All images in this talk are from Stable Diffusion

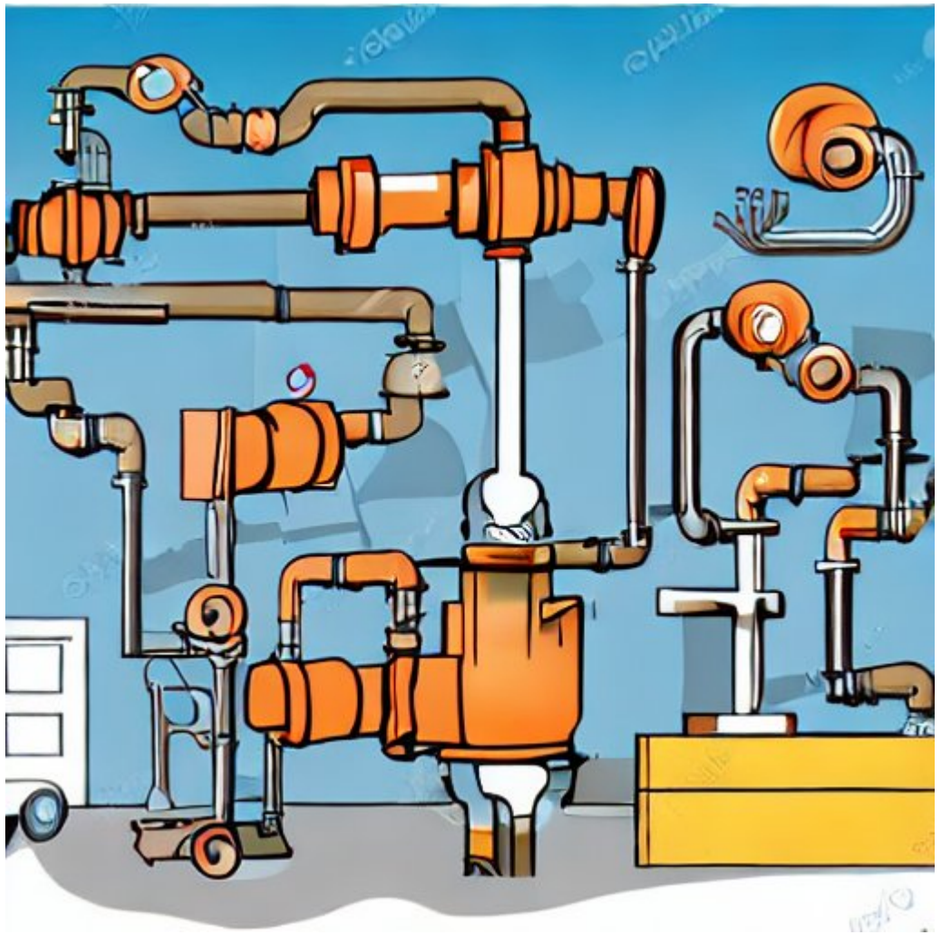
But this talk is all about text

So we'll skip over all of that crazy-fun area of AI/ML!

stablediffusionweb.com

So what is an LLM, and/or a LLaMA?





LLM - Large Language Model

LLaMA - Facebook's (slightly open) LLM

LLaMA - Slightly open?

Code is open, model is not

But more on that later!

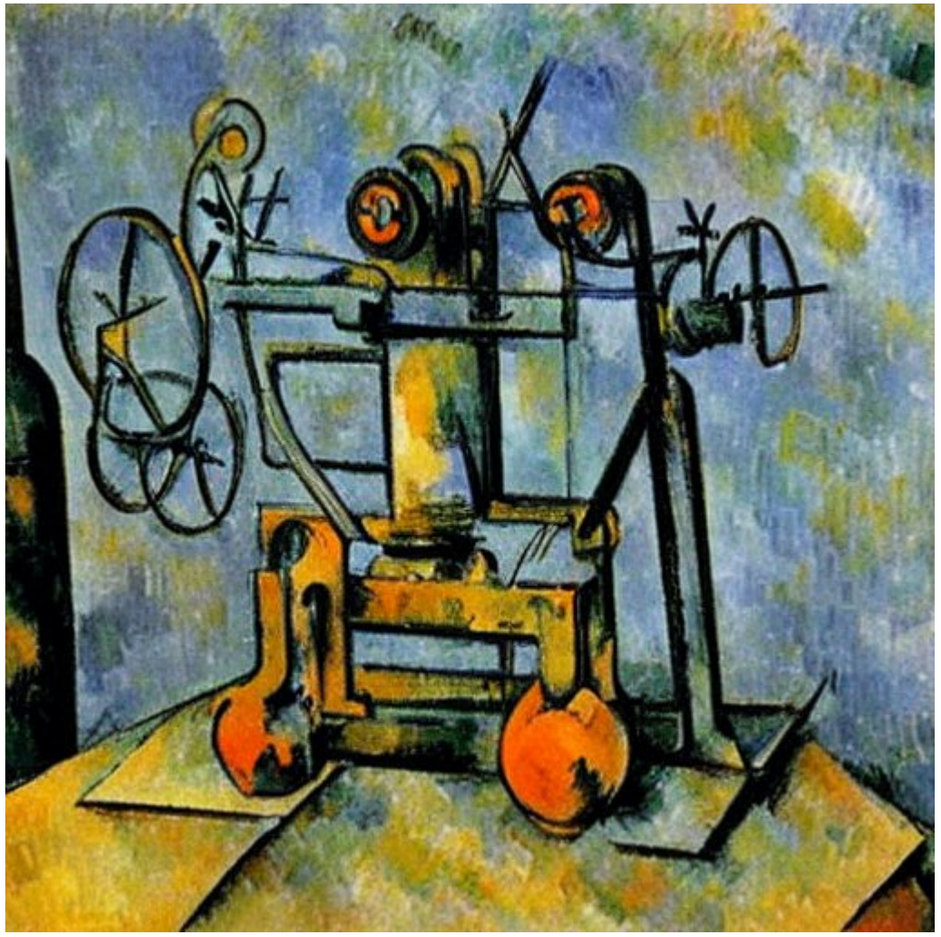
LLM - Large Language Model?

A text-focused neural network with billions of parameters, trained on large amounts of text.

LLMs are deep-learning models, typically general purpose, which excel at a wide range of text-related tasks.

Some level of encoding on syntax and sematics of human language, and some level of general knowledge / facts.

Bigger models (both number of parameters and training corpus) will tend to show more knowledge and better semantics.



Don't ignore the old stuff!

Everything you could do with BERT or ELMo, you can do with a LLM

- Embeddings? ✅
- Vector Search? ✅
- Similar Words? ✅
- Semantic Relationships? ✅

So look back through the Buzzwords 2022 and 2021 presentations

(They're all online on youtube)

If it mentioned BERT, you can try it with an LLM, and it'll (probably) be better!

But a LLM on a laptop?

Live demo time!

llama.cpp, Facebook's LLaMA 7B (7 billion parameters) model, and a few of your questions

Some ML + Information Retrieval terms to know

Tokenization

Breaking the input text into chunks, and possibly some simple transformations

eg *Hello there! Welcome to my talk.*

Could become

`hello there welcome to my talk`

Or… `hello there [!] welcome to my talk [end]`

Embeddings

Both IR systems like Lucene, and ML techniques, need to work on numeric representations.

Lucene has a *term dictionary*, eg `1=hello 142=there`

ML techniques have *embeddings*, a vector-space representation / projection, eg `[0, 0, 0.9, 0, 0.1, 0.8]`

Embeddings

If this is all new to you, or you're a bit hazy on the details...

What are embeddings

by *Vicki Boykis*

github.com/veekaybee/what_are_embeddings

What could the "old stuff" like BERT do?

Similar Words (semantically similar)

What other words are similar to X?

How similar are the words X and Y?

Similar tokens to: linux

- Cosine sim=0.849: unix
- Cosine sim=0.793: open-source
- Cosine sim=0.778: kernel

Similar tokens to: raise

- Cosine sim=0.890: raising
- Cosine sim=0.871: pay
- Cosine sim=0.848: benefit

Difference between raise and risen is 56

Difference between raise and above is 52

Difference between raise and below is 54

Difference between raise and shine is 30

Difference between raise and linux is 11

Word relationships

What is the equivalent of X to Y, for Z?

The analogy of X - Y for Z = ???

berlin - germany for paris = france

madrid - spain for lisbon = portugal

man - boy for woman = girl

Warning - it can be wrong, it can be sexist!

The analogy of X - Y for Z = ???

spain - madrid for portugal = spain

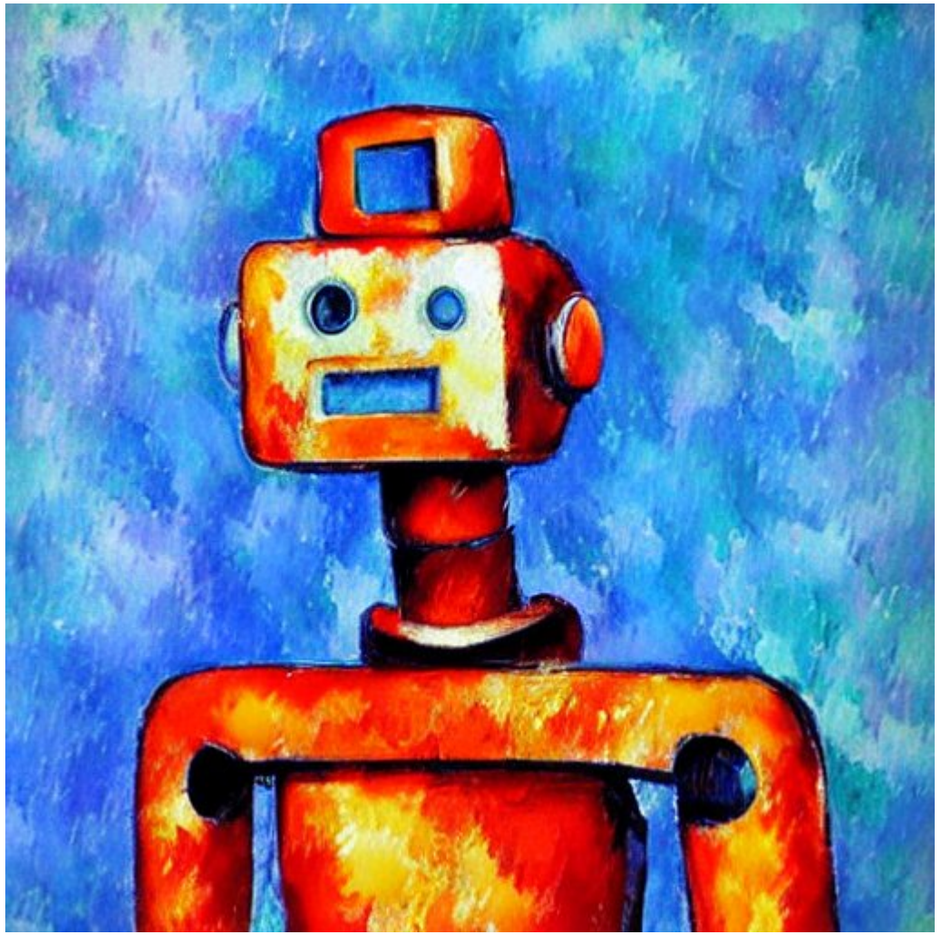
doctor - man for nurse = woman

Embeddings - sentence and/or document

Used in Vector Search (supported by Lucene, elastic etc)

- Transform the documents into the embedding vector space
- Transform the query into the embedding vector space
- Find documents similar to the query

Combination of embedding vector search and normal search can work well, see previous talks for details!



The LLM Revolution

2018 - OpenAI GPT-1

2019 - OpenAI GPT-2

2020 - OpenAI GPT-3

2021 - Github Co-Pilot

2022 March - OpenAI GPT-3.5

2022 November - OpenAI ChatGPT

2023 February - Microsoft Bing AI

2023 February - Google Bard

2023 February - Facebook LLaMA

2023 March - OpenAI GPT-4

2023 March - Microsoft 365 Co-pilot

2023 March - Stanford Alpaca

2023 April - Databricks Dolly

2023 May - MPT-7B

OpenAI is more than just ChatGPT

Whole bunch of different models, with different pricing

Accessed via their APIs, directly or via some wrapper

Include chat interfaces, but also embeddings (for search), summarisation, understanding etc

You're not running it directly on your laptop though!

Facebook's LLaMA

Source code for training and evaluating - GPLv3

Model - Non-commercial use, approval needed

- 7B parameters, 4096 dimensions, 32 layers, 13gb
- 13B parameters, 5120 dimensions, 40 layers, 25gb
- 33B parameters, 6656 dimensions, 60 layers, 61gb
- 65B parameters, 8192 dimensions, 80 layers, 122gb

Aims to be a generalist textual model, ready for further training

Stanford Alpaca

A fine-tuned version of LLaMA 7B, aimed at instruction following

Model - Non-commercial use, research only, needs LLaMA too

Used OpenAI's model to generate 52,000 example instructions, based on 175 human-written ones, then fine-tuned LLaMA with $100 of cloud computing.

New model behaves a lot like ChatGPT, in terms of being able to understand questions and give responses.

Methodology open, model even more closed than LLaMA

LLaMA vs Alpaca

What is the difference between a llama and an alpaca?

LLaMA - An alpaca is a smaller version of a llama.

Alpaca - The main differences between llama and alpacas are size, behavior, and fiber production. Llamas are typically larger than alpacas, with some varieties reaching over 15 hands (62 inches) in height. Behaviorally, llamas tend to be more independent while alpacas prefer the herd environment. Finally, llama's fleece is coarser and thicker compared to an alpaca’s finer fiber.

These both have the same source training data!

Databricks Dolly

Alpaca used OpenAI to generate the training prompts, so restricted by both LLaMA license and OpenAI's license

Databricks decided to crowd-source 15,000 questions+answers from their staff, suitable for model training + refining

Creative Commons Attribution-ShareAlike 3.0 - can be used in commercial applications!

`databricks-dolly-15k` is the name of the prompt/response pairs dataset

But you still need an open LLM to fine-tune, which LLaMA isn't...

github.com/databrickslabs/dolly/

Databricks Dolly 2.0

Applies `databricks-dolly-15k` to a EleutherAI pythia model, to give a LLM that can talk to people, without paying for API access or sharing data with third parties.

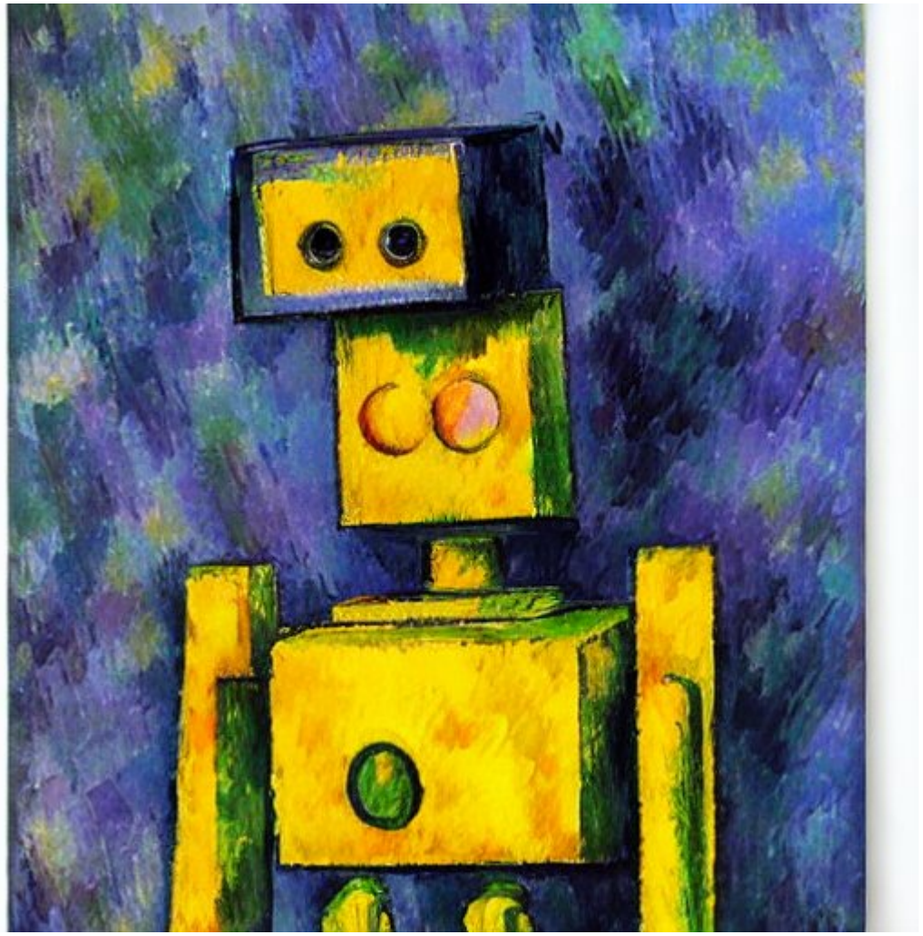
pythia isn't a state-of-the art model, more aimed at academic research on MLs, but is used in real life to guess missing characters in ancient writing, eg broken clay tablets where you can't read all the symbols

Despite that, works surprisingly well for Q&A, summarising, suggesting etc!

Creative Commons Attribution-ShareAlike 3.0 - can be used in commercial applications

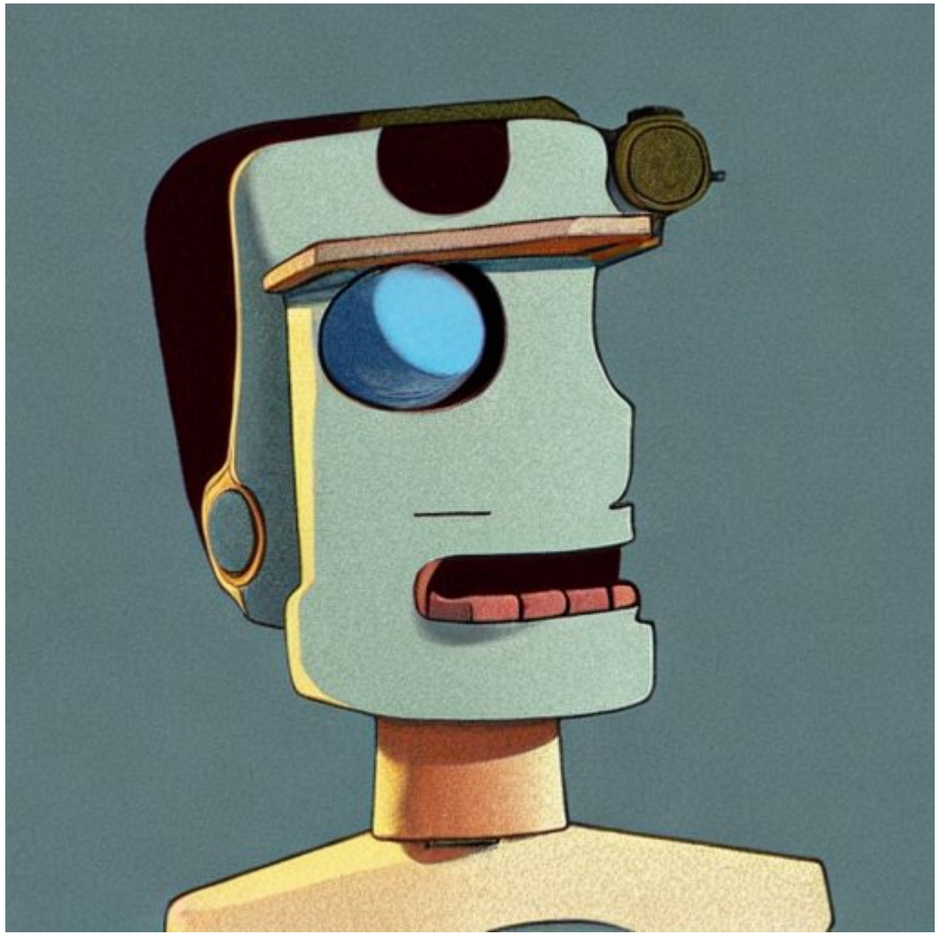
TODO More examples

- MPT
- Bloom
- QLoRA and/or Guanaco



Comparisons between LLMs later

(Or cheat and look in the github repo now!)



8 Things to Know about Large Language Models

by *Samuel R. Bowman*

cims.nyu.edu/~sbowman/eightthings.pdf

8 Things to Know about Large Language Models

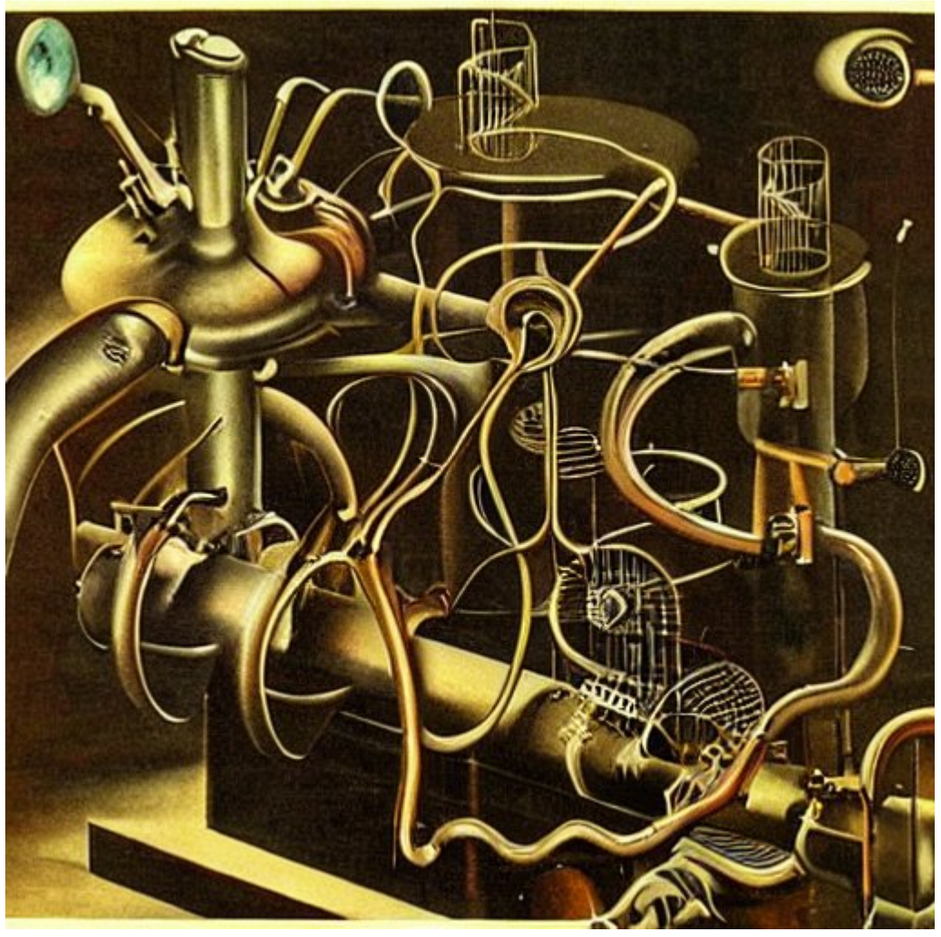
2 - Many important LLM behaviors emerge un-predictably as a byproduct of increasing investment.

4 - There are no reliable techniques for steering the behavior of LLMs.

5 - Experts are not yet able to interpret the inner workings of LLMs.

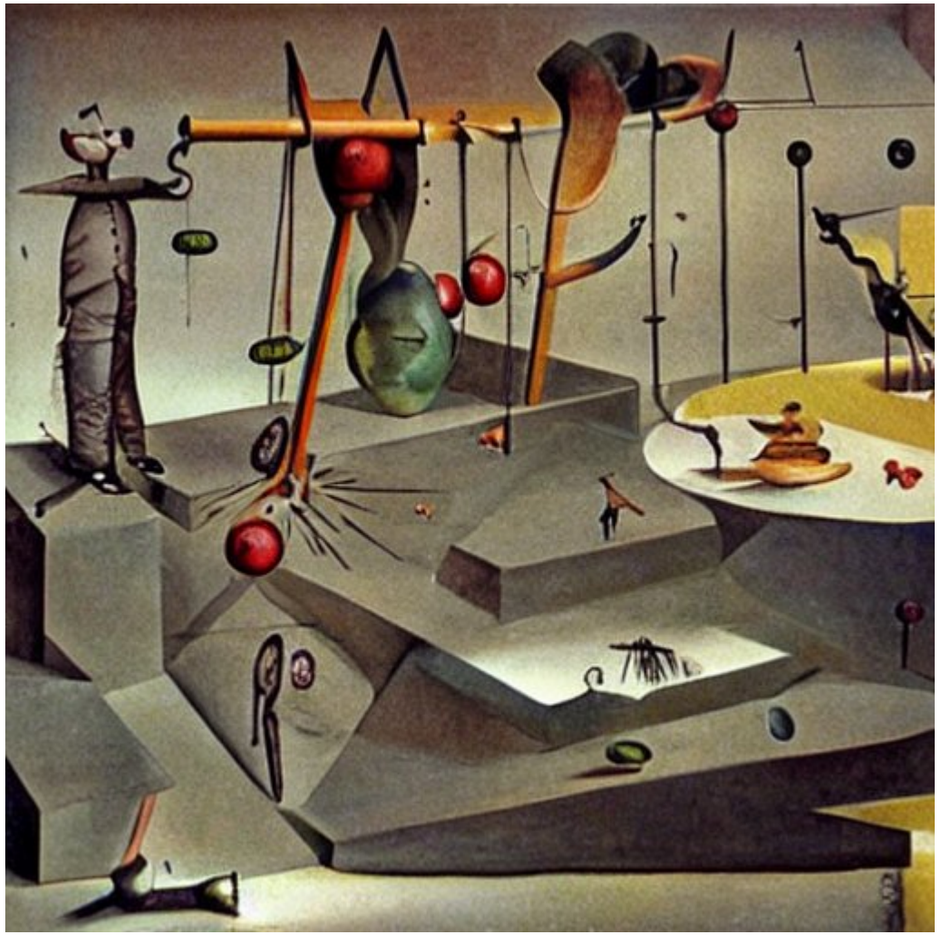
7 - LLMs need not express the values of their creators nor the values encoded in web text.

8 - Brief interactions with LLMs are often mis-leading.



TODO: Using a LLM for...

TODO: More using a LLM for...



TODO: Tokens and Windows

Context Matters to LLMs

Your previous prompts within a session will influcene answers

An example with LLaMA

Unprompted Session

Who is Nick Burch?

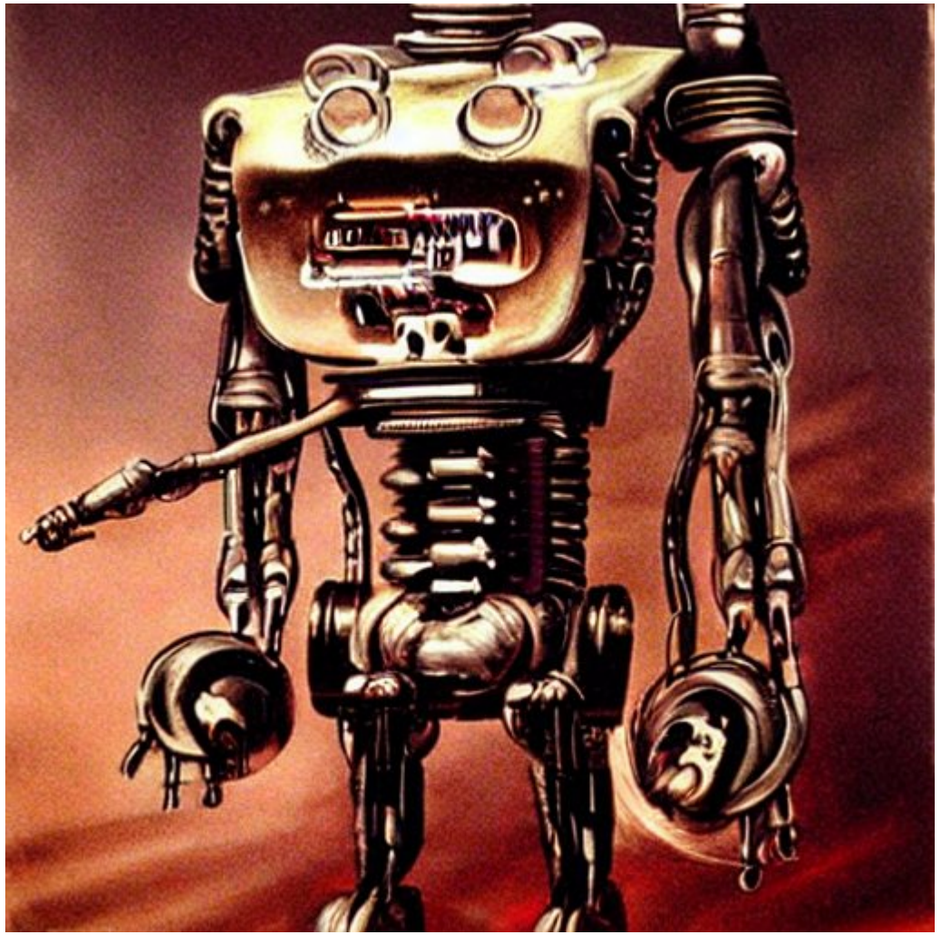
I'm sorry, but I don't know who Nick Burch is. Could you please be more specific?

Prompted Session

- **What is Berlin Buzzwords?**
- *Berlin Buzzwords is a technology conference held every year in Berlin.*

- **When was the first Berlin Buzzwords held?**
- *The first Berlin Buzzwords was held in 2004.*

- **Who is Nick Burch**
- *Nick Burch is a developer who started developing software when he was 14.*



TODO: Things that can go wrong

TODO: Getting it running on your laptop

TODO: Hugging Face

TODO: Improving your LLM

Code, Scripts, Slides

github.com/Gagravarr/BBuzz23-LaptopML

All code in slides is taken from here!