Start putting in frames to evaluate the future development of AI.

Frames/First principles

- To determine the impact of AI. How much of the current GDP can be produced by AI?
  - Understanding digital people (Karnovsky/Age of Em).
- That AI will have a 'transformative impact'. It's hard to deny that AI doesn't keep scaling but who says that it hits a point where it can duplicate us and/or go beyond us through its own feedback loops.
- Another tactic is to take the viewpoint of a smart person and deviate from that.

Dwarkesh w/ Carl Shulman

[[Carl Shulman]] [[Dwarkesk]]

I'd like to explore this notion that our brains are very efficient, so are other animals' brains but they evolved in a climate where there was optimasation along multiple dimensions and without too concerted a goal. With AI we're developing the algorithmic process in one direction, it doesn't have to worry about fat stores, muscle mass etc.

> really intensive massive brute force search and things like evolutionary algorithms can produce intelligence.

We spend more time in childhood, increasing compute, larger brain. Compared to a model and training.

A feedback loop of allocating resources to cognitive ability. The bigger your brain, the more compute, the smarter you are, the smarter everyone else is so they can teach each other. So the marginal benfit for optimising this dimension is much higher than any other dimension. (Without foresight here from evolution obviously).

He also highlights that the larger the group in pre-industrial society, the more technology propogated.

> . Humans and animals learn and adapt efficiently with relatively limited experiences compared to what AI models are exposed to. This 'undertraining' means that despite their efficient cognitive architectures, animals can't reach the kind of exhaustive training AI models undergo.

This comparison to evolutionary brain development is really interesting.

This niche of dimension expansion that we got is > In general, there are trade-offs where the extra fitness you get from a brain is not worth it and so creatures wind up mostly with small brains because

they can save that biological energy and that time to reproduce, for digestion and so on. Humans seem to have wound up in a self-reinforcing niche where we greatly increase the returns to having large brains. Language and technology are the obvious candidates. You have humans around you who know a lot of things and they can teach you. And compared to almost any other species we have vastly more instruction from parents and the society of the [unclear]. You're getting way more from your brain than you get per minute because you can learn a lot more useful skills and then you can provide the energy you need to feed that brain by hunting and gathering, by having fire that makes digestion easier.

Is democratising AI good in that it means powerful actors can't get a hold of it?

Prompt You goal is to read https://www.dwarkeshpatel.com/p/carl-shulman and provide intuition behind some of its core points. You will be in conversation with someone who's also interested in the topics discussed and has knowledge of explanations of things through evolutionary theory.

Each question asks you will frame your answer from what you've read.

got nothing good here

Intro to LLM's

https://www.youtube.com/watch?v=zjkBMFhNj_g&t=1069s

Think of LLMs as the kernel process of an LLM OS.

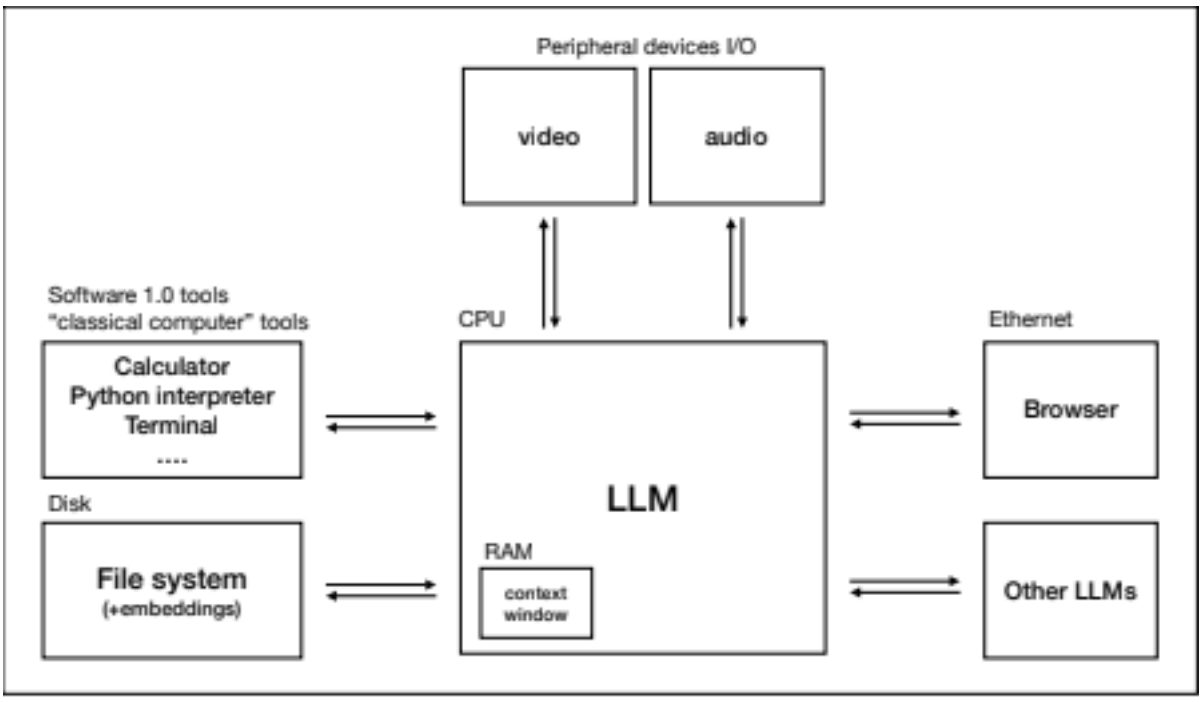

Figure 1: llm_os

- Pre-training is when a web crawler takes a large amount of text data and the parameters are calculate for the model through large amounts of computation with GPU's.
  - Compress 10TB of web text data
  - Get a cluster of~6000 GPUs.
  - Pay 2 million, wait 12 days, obtain base model.
- Once this done, fine-tuning involves more labelled instructions that specify how the model should behave. Collect high quality QA responses and train the model on that (assistance model).

Karpathy states that text prediction is a form of compression of the internet. If we think of a random sequence as the most un-compressed form of a string. A model can predict what comes next, therefore it reduces the information required to store that text.

- You can't use the base model directly. Not too sure why.
- You can do your own fine-tuning with the meta llama model, they'll give you the base model.
- Closed weight models are the highest rated currently (ELO wise).

Scaling Laws

Performance of these LLMs in terms of the *accuracy of the next word prediction task* is a well behaved function of two variables: The number of parameters and the amount of text we train on.

The accuracy is correlated with a lot of things we do care about.

Multi-modality.

- Tool use is a major aspect in how the models are becoming more capable (being able to use the browser, plot things Dall-E etc.).
- This is related to how we solve problems, we use calculators, search for information etc.

Future development

- LLM's currently function in a system 1 fashion. They can't dilate their pupils and put a tree of possibilities on front of their minds eye. What would it mean for them to have a system 2? We want them to think things through.
- Theres no simple reward function for self-learning of LLMs.

What are the limitations to the transformer architecture?

Neural Networks

http://neuralnetworksanddeeplearning.com/chap1.html

- If you try to write a program, algorithm, to recognise handwritten digits there are numerous caveats and fuzzy thinking that must be accounted for. > What seems easy when we do it ourselves suddenly becomes extremely difficult. Simple intuitions > about how we recognize shapes - "a 9 has a loop at the top, and a vertical stroke in the bottom > right" - turn out to be not so simple to express algorithmically. When you try to make such rules > precise, you quickly get lost in a morass of exceptions and caveats and special cases. It seems > hopeless.
- The goal of a neural network is to use examples to learn the rules and caveats this algorithm would have.

What is a neural network

- A *perceptron* takes a set of binary inputs and produces a single binary output.

- To compute its output, weights are used as a measure of importance of each input.
- The output is then determined if the *weighted sum* exceeds a threshold.
- Perceptrons should be thought of as simple decision makers weighing up evidence.

The output of the perceptron is

$$\text{output} = w \cdot x + b > 0$$

b is the threshold brought to the other side and the dot product $w \cdot x = \sum_j w_j x_j$.