Decoding GPT: Mathematical foundations of Interpretability and Alignment for Large Language Models

Samy Wu Fung, Daniel McKenzie, Michael Ivanitskiy

2024-01-09

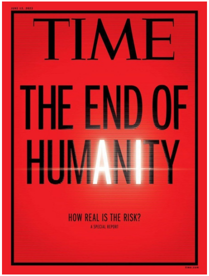
John F. Kennedy was just elected President of the United States after rising from the grave decades after his assassination. Due to miraculous developments in nanotechnology, Kennedy's brain was rebuilt from his remains and installed in the control center of a state-of-the art humanoid robot. Below is a transcript of his acceptance speech.

"It is time once again. I believe this nation can do great things if the people make their voices heard. The men and women of America must once more summon our best elements, all our ingenuity, and find a way to turn such overwhelming tragedy into the opportunity for a greater good and the fulfillment of all our dreams. In the months and years to come, there will be many battles in which we will have to be strong and we must give all of our energy, not to repel invaders, but rather to resist aggression and to win the freedom and the equality for all of our people. The destiny of the human race hangs in the balance; we cannot afford for it to slip away. Now and in the years to come, the challenge before us is to work out how we achieve our ultimate destiny. If we fail to do so, we are doomed."

- GPT-2 (1.5B parameters) paper published in Feb 2019
- GPT-3 (175B parameters) paper published in May 2020
- neither were noticed much outside of ML/NLP circles
- ChatGPT/GPT-3.5 released in Dec 2022, now everyone has heard of GPT
- ***nobody expected AI systems this powerful this to show up for decades!***

In 2023:

- LLM agents and tools being integrated into everything
- open letter to pause giant AI experiments
- Hinton quit google to talk about the dangers of AI
- Yudkowsky wrote in time magazine about existential risks
- AI safety is one of the top priorities of UK government and others
- internet is being flooded with generated images and LLM bots, mostly for nefarious purposes
- *why is everyone so concerned?*



- **Interpretability:** *nobody understands* how these models work
- **Alignment:** we don't know how to ensure that these models are aligned with human values
- **Scaling:** there is a possibility that these models surpass human intelligence through scale alone
- **Ethics:** models are taking jobs, spamming the internet, and will generally cause more and more chaos
- **Governance:** we can't even agree on how to govern a world without AIs, and they won't make the world simpler

it seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers. There would be no question of the machines dying, and they would be able to converse with each other to sharpen their wits. At some stage therefore we should have to expect the machines to take control

- Alan Turing, 1951

Logistics

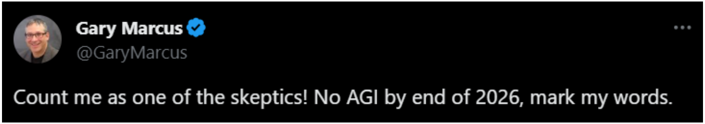
- ask questions at any time!
- course website: `miv.name/decoding-gpt`
- course materials will be on a public github repo, expect a canvas announcement
- office hours poll
- Generative AI policy: use it for everything, document how you use it, do things that LLMs can't do

Course Goals

- transformers & attention from the perspective of linear algebra, some probability intuition for why they work so well
- how do we summon these models from parameter space?
- using LLMs effectively, the landscape of tools and resources, TransformerLens library
- alignment: why should we worry about AI, what are the risks, what are the solutions?
- interpretability: what do we know about the internals of models so far, what are the current techniques, where are the exciting directions?
- final projects: do something cool with LLMs or interpretability!

Definitions

- **Neural Network:** a function approximator built from affine operations and nonlinearities between them. Not a single architecture, refers to a broad class of models
- **Machine Learning:** the study of algorithms which learn from data, nowadays focusing on neural network architectures
- **Artificial Intelligence:** overused and hyped to the point where it means so much that it means nearly nothing. No longer any agreed upon definition
- **AGI/ASI:** Artificial General Intelligence / Artificial Superintelligence. AI systems which are as smart as humans, or smarter than humans
  - until 2020, most AI/ML researchers were confident that these were coming sometime between 2050 and never. now:



Gary Marcus

@GaryMarcus

Count me as one of the skeptics! No AGI by end of 2026, mark my words.

- **GPT:** Generative Pretrained Transformer
  - autoregressive, trained on a lot of diverse data
  - often synonymous with LLM
  - not just OpenAI models!
- **LLM:** Large Language Model
  - almost always an autorregressive transformer
- **Diffusion Model:** a different attention-based architecture, primarily used for generating images. Learns the “reverse” of adding noise to an image
- **Generative AI:** a popular term encompassing both LLMs and diffusion models

- **Transformer:** a neural network architecture built on the attention mechanism
  - doesn't have to be autoregressive (encoder models / seq2seq)
  - doesn't have to deal with language (vision transformers)
  - introduced in "Attention is All You Need" (Vaswani et al, 2017)
  - stack of attention heads and feedforward layers with some other ML magic mixed in
- **Attention Head:** given two vectors, compute a scalar attention weight via dot products of their projections to another space, and use that to scale another linear projection

$$\mathbb{A}(x, y) = \sigma\left(\frac{1}{\sqrt{d_k}} x W_Q y \cdot W_K^T\right) y W_V$$

Interpretability/Explainability: the field of trying to make ML architectures understandable. this can either involve building architectures which are more comprehensible, or trying to understand the internals of existing architectures

- in this course, we will focus on the latter and refer to it as just “interpretability”
- “mechanistic” interpretability focuses on understanding circuits and testing hypotheses about how they work
- “developmental” interpretability focuses on understanding how the model learns, and what it learns over time
- very new field, but lots of really exciting research!

AI Alignment: broadly, the field of trying to get AI systems to be "good" in some sense

- this can mean any of: "obey instructions", "follow laws", "respect human values", "don't kill all humans", etc.
- in this course, we'll use it to refer to the technical problems of making AI systems pursue the goals we want them to, while respecting constrains – difficult because we can't write those down
- "outer alignment" is the problem of correctly specifying goals, since in complicated systems it's easy to underspecify what you want (Goodhart's Law)
- "inner alignment" is the problem of ensuring that the systems goals actually align with the loss function, particularly when you move from training to deployment

- **AI Ethics:** *what* should AI systems be doing? What should they ***not*** be doing?
  - lots of both overlap and heated disagreement with alignment
  - no the focus of this particular course, but without ethics alignment is meaningless
- **Scaling Laws:** the observation that transformer architectures get smarter as you make them larger. See Kaplan et al., 2017.

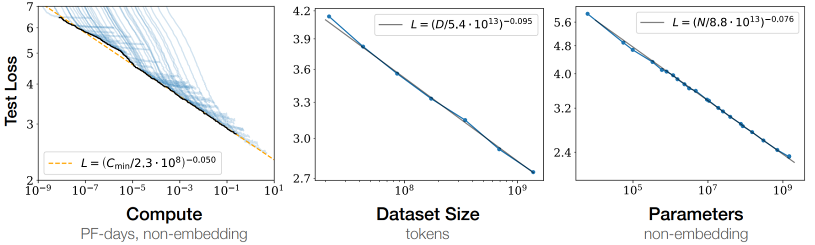


Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

GPT-4 demo!