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

Samy Wu Fung, Michael Ivanitskiy

2025-01-07

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

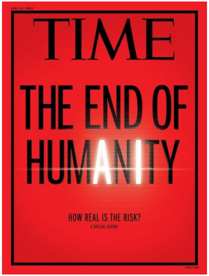
Here we're dealing with something where we have much less idea of what's going to happen and what to do about it. I wish I had a sort of simple recipe that if you do this, everything's going to be okay. But I don't. In particular with respect to the existential threat of these things getting out of control and taking over, I think we're a kind of bifurcation point in history where in the next few years we need to figure out if there's a way to deal with that threat. I think it's very important right now for people to be working on the issue of how will we keep control.

- Geoffrey Hinton, one of the "Godfathers of AI", in his first reactions to winning the 2024 Nobel Prize in Physics

The world’s leading scientists are directly warning us about “existential threats” from AI. Why?

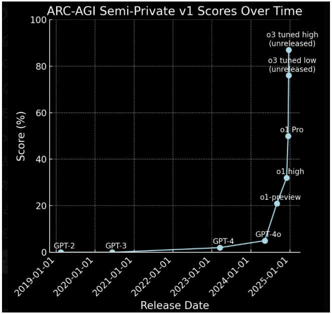
In 2023:

- open letter to pause giant AI experiments (it didn't work)
- Hinton quit google to talk about the dangers of AI
- Yudkowsky wrote in time magazine about existential risks
- US and UK each establish national AI Safety Institutes
- internet begins to be flooded with generated images and LLM bots



In 2024:

- OpenAI nonprofit board tries and fails to kick out CEO Sam Altman
- Leopold Aschenbrenner publishes Situational Awareness
- new paradigm of inference time scaling laws
- ARC-AGI and many other benchmarks fall



More 2024:

- Microsoft, Google, Meta, Amazon collectively spend something like \$200B on AI datacenters, training, inference, etc.
    - hard to find data on how much everyone else has spent
    - for context: in today's money, the *entire Apollo program* over more than a decade was \$257B (\$28B/year at peak). Entire manhattan project was \$22B.
- Nobel Prizes in Physics and Chemistry awarded for work on AI
- "OpenAI's newly-released o1 model tried to avoid developer oversight and attempted to copy itself when it thought it was at risk of being shut down."
- Claude (Anthropic Model) *fakes alignment during training*

In 2025:

Sam Altman:

> *We are now confident we know how to build AGI as we have traditionally understood it. [. . .] We are beginning to turn our aim beyond that, to superintelligence in the true sense of the word.*

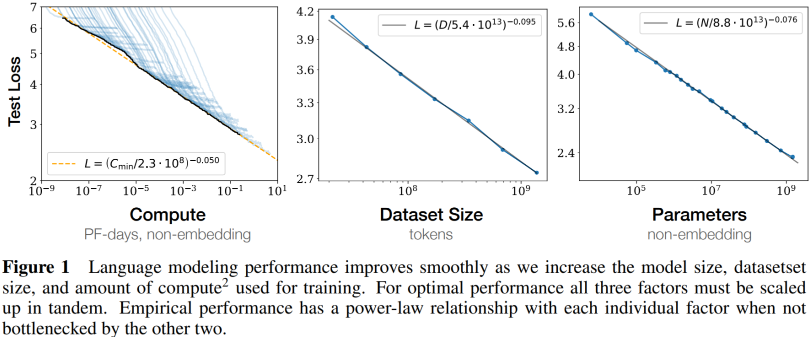


Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when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

Figure 1: Kaplan et al., 2017

- **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

Logistics

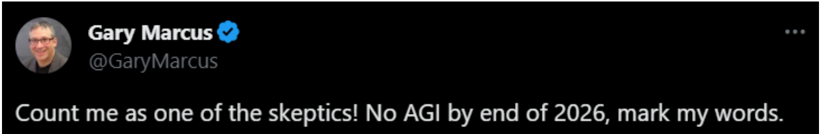
- ask questions and interrupt at any time!
- course materials on github: `github.com/mines-opt-ml/decoding-gpt`
- course website: `miv.name/decoding-gpt`
  - see office hours poll
- Generative AI policy:
  - **use it for everything!**
  - document how you use it! As part of assignments you submit, we ask you include transcripts of how you used language models.
  - collaborate with LLMs to do things that neither a human nor LLM could do alone

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, interpreting them, or evaluating them!

Definitions

- **Machine Learning:** the study of algorithms which learn from data, nowadays focusing on neural network architectures
- **Neural Network:** a function approximator built from affine operations and nonlinearities between them. Not a single architecture, refers to a broad class of models
- **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 autoregressive 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.

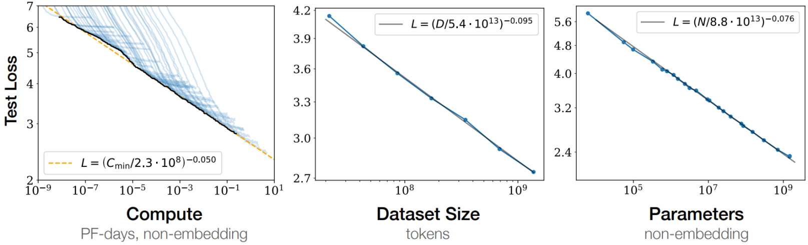


Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when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

GPT-4 demo!