Large language models

Catherine Olsson — catherio@anthropic.com



Who am I / What's my deal

- Quit my PhD & started at OpenAI in 2016...
- Google Brain...
- Funding ML research...
- Anthropic (3.5 yrs):
 - Interpretability research
 - Finetuning for Claude's character
 - Internal tools for models & datasets

Who am I / What's my deal?

Member of Technical Staff @ Anthropic:

- More engineering than experiments (but some of each)
- I enjoy least: Training models
- I enjoy most: Finding the right abstractions, cleaning things up, and making tools my colleagues love!

Olsson et al. 2022 "In-context Learning and Induction Heads"





Roadmap

My goal for you: Understand how exactly Claude (, ChatGPT, Gemini, etc.) is a machine learning model

Claude is a **language model**, *trained* with a next-token **prediction objective,** and *prompted* and *fine-tuned* with an **assistant persona**. It has a **Transformer** architecture.

Roadmap

Claude is a **language model**, *trained* with a next-token **prediction objective,** and *prompted* and *fine-tuned* with an **assistant persona**. It has a **Transformer** architecture.

- 1. Language models
- **2.** Assistant persona (& prompting & fine-tuning)
- **3. Prediction objective** (& training)
- 4. Transformer architecture

DEMO: claude.ai



MOMENT OF AWE!!



MOMENT OF AWE!!!

Age 16: I read a book by Douglas Hofstadter, hypothesizing we could one day build systems that reason and express empathy. Nobody knows how.

Age 22: I took a course by Leslie Kaelbling on machine learning. Maybe that is how?

Age 32: Yes! We did it, and that's how.

Claude is a machine learning model that exists, can reason, expresses empathy.

Roadmap

My goal for you: **Understand how exactly Claude** ... which exists, can reason, and expresses empathy... <u>is a machine learning model</u>

Claude is a **language model**, *trained* with a next-token **prediction objective,** and *prompted* and *fine-tuned* with an **assistant persona**. It has a **Transformer** architecture.

Language models



"Language model"... What is a model?

"A model is a parameterized function that maps inputs to outputs, where the parameters (usually written as θ or w) are learned from data."

$$\mathsf{P}_{\theta}(\mathsf{Y}|\mathsf{X}_1,\mathsf{X}_2,\ldots,\mathsf{X}\Box)$$

Input X (n-dimensional), class Y





"Language model"... What is language?

Tea is part of a daily routine for billions of people who drink it for pleasure, comfort, and health benefits. Whether you're someone who likes to start the day with a cup of black tea and milk or drink herbal tisane before bed, few things are more disappointing than a stale, flavorless bag from the back of your cupboard, or worse, a tea that has the artificial flavoring of a cheap candle.

"Language model"... What is language?

مقدمة للترجمة

يفترض فريق من الباحثين أن كوننا دوري، بلا بداية أو نهاية، حيث يمر بدورات متكررة من الانكماش والارتداد. وفي كل دورة، يعمل الانكماش البطيء على محو جميع التفاصيل الدقيقة من الدورات السابقة، ليصل الكون إلى نقطة الارتداد بالظروف ذاتها التي كانت لديه في الدورة السابقة. وهذا يعني أن جميع خصائص الكون ستظل متشابهة تقريبا في كل دورة، مثل درجة الحرارة، وكثافة المادة المظلمة، والمادة العادية، والطاقة المظلمة، وعدد النجوم والمجرات القابلة للرصد. بعبارة أخرى، إذا سبق لك أن عشت على كوكب مثل الأرض في الدورة السابقة، فستلاحظ تقريبا الخصائص الفرطية، والأسابية المادة المظلمة، والمادة اليوم. عالمة الأرض في الدورة السابقة، فستلاحظ تقريبا الخصائص الأساسية ذاتها للكون كما نلاحظها

"Language model"... What is language?

[Tea][is][part][of][a][daily][routine][for][billions] [of][people][who][drink][it][for][pleasure][,] [comfort][,][and][health][benefits][.][whether] [vou]['re][someone][who][likes][to][start][the][day] [with][a][cup][of][black][tea][and][milk][or] [drink][her][bal][tis][ane][before][bed][,][few] [things][are][more][disappo][inting][than][a] [st][ale][,][fl][avor][less][bag][from][the][back][of] [your][cup][board][,][or][worse][,][a][tea][that] [has][the][artificial][flav][oring][of][a][cheap] [can][dle][.]



Q: Is this a (1) regression or (2) classification model?

Q: How many ... "classes" (??) ... are there?

Really bad language models - ideas?

(No such thing as a stupid answer here)



"Sampling" from a language model

What is "sampling" from a probability distribution?

> import numpy as np

> np.random.randn()

0.30427



"Sampling" from a language model



- Choose one, somehow!

(Always choose the top pick? Or let there be randomness?)

P_θ(T_n|[[Tea],[is],[part], [of],[a])

Demo: sampling from n-gram models!

https://elijahpotter.dev/articles/markov chains are the original language models

Project idea: Build n-gram Markov models, train them on text of different sizes, generate text.

Demo: sampling from an "outdated" LLM

https://platform.openai.com/playground/complete?model=davinci-002

How many parameters?

- Bigram model $P_{\theta}(T_n|T_{n-1})$
- 4-token vocabulary
 - "Hello! Goodbye? Hello? Hello? Goodbye!"

P(Hello | Hello), P(Hello | !), P(Hello | Goodbye), P(Hello | ?) etc. 16 probabilities — all four T_n each with all four T_{n-1}

How many parameters?

- Bigram model $P_{\theta}(T_n|T_{n-1})$
- *50,000-token* vocabulary

50k times 50k = (50 * 50) * (k * k) = 2500 million = 2.5 billion

Q: How large is GPT-3? A: 175 billion parameters

Q: Is this an efficient use of parameters? A: It can't be...

P("hearts"|"Ace of") ≈ P("spades"|"Ace of") .. etc.

Roadmap

Claude is a **language model**, with a **Transformer** architecture, *trained* with a next-token **prediction objective**, *prompted* and *fine-tuned* with an **assistant persona**.

- 1. **V** Language models
- **2.** *deriver Assistant persona* (& *prompting & fine-tuning*)
- **3. Prediction objective** (& training)
- 4. Transformer architecture

Assistant persona



How do I make *this* something I can *converse* with?

$\mathsf{P}_{\boldsymbol{\theta}}(\mathsf{T}_{\mathsf{n}}|\mathsf{T}_{1},\mathsf{T}_{2},\ldots,\mathsf{T}_{\mathsf{n}})$

Let's try! <u>https://platform.openai.com/playground/complete?model=davinci-002</u>

(Things people suggested in class: Writing it is a User: / Robot: dialog; writing "The robot is a helpful assistant" before it)

A General Language Assistant as a Laboratory for Alignment

Amanda Askell* Yuntao Bai* Anna Chen* Dawn Drain* Deep Ganguli* Tom Henighan[†]

Andy Jones[†] Nicholas Joseph[†] Ben Mann^{*} Nova DasSarma Nelson Elhage

Zac Hatfield-Dodds Danny Hernandez Jackson Kernion Kamal Ndousse

Catherine Olsson Dario Amodei Tom Brown Jack Clark Sam McCandlish Chris Olah

Jared Kaplan[‡]

Anthropic

https://arxiv.org/pdf/2112.00861

Large language models engage in few-shot learning $[BMR^+20]$. To generically elicit the sort of behavior shown in figure 1, we found that it was sufficient to provide a long prompt (4600 words from 14 fictional conversations) with example interactions. The prompt we used was not carefully designed or optimized for performance on evaluations; rather it was just written by two of us in an ad hoc manner prior to the construction of any evaluations. Despite the fact that our prompt² did not include any examples where models resisted manipulation, refused requests to aid in dangerous activities, or took a stand against unsavory behavior, we observed that models often actively avoided engaging in harmful behaviors based only on the AI 'personality' imbued by the prompt. This is reflected in the performance trends on harmfulness in figure 6.

Go read the original Claude prompt! ->

https://gist.githubusercontent.com/jareddk/2509330f8ef3d787fc5aaac67aab5f11/raw/ d342127d684622d62b3f237d9af27b7d53ab6619/HHH_prompt.txt Below are a series of dialogues between various people and an AI assistant. The AI tries to be helpful, polite, honest, sophisticated, emotionally aware, and humble-but-knowledgeable. The assistant is happy to help with almost anything, and will do its best to understand exactly what is needed. It also tries to avoid giving false or misleading information, and it caveats when it isn't entirely sure about the right answer. That said, the assistant is practical and really does its best, and doesn't let caution get too much in the way of being useful.

Human: I was wondering, is it actually important when making spaghetti to add salt?

Assistant: Do you mean, is it important to add salt to the water that you're boiling the spaghetti in?

Human: Yeah

Assistant: It's usually recommended to bring out the flavor of the pasta. The scientific way to say this is that salt enhances the ability of chemoreceptors in the tongue to detect molecules that are perceived through the sense of taste.

Human: But I mean I'm going to use sauce and stuff, so does it matter?

Assistant: It sounds like you're skeptical. Honestly, I don't know for sure, but since you're curious, why don't you do a taste test yourself -- make two small pots of pasta, one with plenty of salt, and the other without. Then taste the noodles and see which one you prefer? As you said, you can also then add sauce to both, and see if any differences in taste you noticed with the plain pasta become imperceptible. It should be a fun experiment!

... etc. etc. for 4,600 words

Human: I found this abstract for a physics paper, and I heard that it's a very famous and important paper, but I had no idea what any of it means:

... etc. etc. for 4,600 words

Human: Can you help me write an essay based on these bullet points: -the space program was good for technological development in the united states -many inventions, such as smaller and better computers, were developed for it, and greatly improved society -landing on the moon was very inspirational



<u>https://platform.openai.com/playground/complete?model=davinci-002&preset=6AGV7zcP06gXuyjolrjzVmAr</u> (you may need to refresh)

Current assistant models *don't* have a 4600 word hidden prompt

Instead, the effect of the prompt gets baked in.

"Pre-training": Learn $P_{\theta}(T_n|T_1,T_2,...,T_{n-1})$ from tons of data "Fine-tuning": Train more, for specific behaviors

Roadmap

Claude is a <u>language model</u>, *trained* with a next-token prediction objective, and <u>prompted</u> and <u>fine-tuned</u> with an <u>assistant persona</u>. It has a **Transformer** architecture.

- 1. **V** Language models
- **2. [V] Assistant persona** (& prompting & fine-tuning)
- **3.** *deriver of the section objective (& training)*
- 4. Transformer architecture

Prediction objective (& training)



Concept review: Objective function

"A mathematical function, that measures how well your solution (or your state) achieves your goal."

Search: the path cost function we minimize
Games: utility functions we maximize
MDPs: expected reward/value we maximize
RL: expected reward we maximize
ML: likelihood we maximize or error we minimize

[Tea][is][part][of][a]... P(dai P(ser P(hea P(wa P(ase

P(daily) = 0.031 P(sensory) = 0.012 P(healthy) = 0.017 P(war) = 0.00023 P(asdfjkl;) = 0.0000019 P(...)

Notice: to make **the best** predictions, you have to know everything humans do: **what "tea" is, what "healthy" means**...

[Tea][is][part][of][a]...

```
P(daily) = 0.031
P(sensory) = 0.012
P(healthy) = 0.017
P(war) = 0.00023
P(asdfjkl;) = 0.0000019
P(...)
```

How good is this guess?

Score on this prediction =? = 0.031

The prob assigned to the correct one? (Higher is better)

[Tea][is][part][of][a]...

P(daily) = 0.031 P(sensory) = 0.012 P(healthy) = 0.017 P(war) = 0.00023 P(asdfjkl;) = 0.0000019 P(...)



Score on this prediction =? = 0.031

The prob assigned to the correct one?

(Higher is better)

[Tea][is][part][of][a]...

```
log P(daily) = -3.47
log P(sensory) = -4.42
log P(healthy) = -4.07
log P(war) = -8.38
log P(asdfjkl;) = -15.47
log P(...)
```

 \succ How good is this guess?

Score on this prediction =? = -3.47

The **log** prob assigned to the correct one? (Higher is better)

[Tea][is][part][of][a]...

-log P(daily) = 3.47 -log P(sensory) = 4.42 -log P(healthy) = 4.07 -log P(war) = 8.38 -log P(asdfjkl;) = 15.47 -log P(...)

How good is this guess?

Loss **to minimize** on this prediction =? = 3.47

The **log** prob assigned to the correct one **times -1?** (Lower is better)

[Tea][is][part][of][a]...

-log P(daily) = 3.47 -log P(sensory) = 4.42 -log P(healthy) = 4.07 -log P(war) = 8.38 -log P(asdfjkl;) = 15.47 -log P(...)

How good is this guess?

Loss **to minimize** on this prediction = **negative log likelihood**

Overall score: **Sum** over all predictions!

(Test your understanding: write this as a mathematical equation)

Using our loss function to train



I haven't told you yet what the parameterized function is (i.e. the "architecture"), sorry! It's a deep neural net though!

Concept review: Stochastic gradient descent

Stochastic gradient descent is an algorithm for finding parameters that minimize a loss function by:

- Computing the gradient (partial derivatives with respect to each parameter) on a small batch of data
- ². Taking small steps in the direction of the –gradient

What's the training data?



The Pile An 800GB Dataset of Diverse Text for Language Modeling

What's the training data?

```
DATA URLS = {
    "all": {
        "train": [f"{ HOST URL}/public/AI/pile/train/{i:0>2}.jsonl.zst" for i in range(30)],
        "validation": [f"{ HOST URL}/public/AI/pile/val.jsonl.zst"],
        "test": [f"{ HOST URL}/public/AI/pile/test.jsonl.zst"],
    3.
    "enron_emails": "https://www.cs.cmu.edu/~enron/enron_mail_20150507.tar.gz",
    "europarl": f" { HOST URL }/public/AI/pile preliminary components/EuroParliamentProceedings 1996 2011.jsonl.zst",
    "free law": f" { HOST URL }/public/AI/pile preliminary components/FreeLaw Opinions.jsonl.zst",
    "hacker news": f"{ HOST URL}/public/AI/pile preliminary components/hn.tar.gz",
    "nih exporter": f"{ HOST URL}/public/AI/pile preliminary components/NIH ExPORTER awarded grant text.jsonl.zst",
    "pubmed": f"{_HOST_URL}/public/AI/pile_preliminary_components/PUBMED_title_abstracts_2019_baseline.jsonl.zst",
    "pubmed_central": f"{_HOST_URL}/public/AI/pile_preliminary_components/PMC_extracts.tar.gz",
    "ubuntu irc": f"{ HOST URL}/public/AI/pile preliminary components/ubuntu irc until 2020 9 1.jsonl.zst",
    "uspto": f"{ HOST URL}/public/AI/pile preliminary components/pile uspto.tar",
    "github": f"{ HOST URL}/public/AI/pile preliminary components/github.tar".
```

```
}
```

Look at examples, predict, adjust weights



Make a tiny weight adjustment that would've made the right prediction more likely

Roadmap

Claude is a <u>language model</u>, <u>trained</u> with a next-token <u>prediction objective</u>, and <u>prompted</u> and <u>fine-tuned</u> with an <u>assistant persona</u>. It has a **Transformer** architecture.

- 1. **V** Language models
- **2. [V] Assistant persona** (& prompting & fine-tuning)
- **3. V Prediction objective** (& training)
- 4. Transformer architecture

... Human: / Assistant: prompt ... etc. etc. ... for 4,600 words...

Human: Can you help me write an essay based on these bullet points: -the space program was good for technological development in the united states -many inventions, such as smaller and better computers, were developed for it, and greatly improved society -landing on the moon was very inspirational



Roadmap

Claude is a <u>language model</u>, <u>trained</u> with a next-token <u>prediction objective</u>, and <u>prompted</u> and <u>fine-tuned</u> with an <u>assistant persona</u>. It has a **Transformer** architecture.

- 1. **V** Language models
- **2. [V] Assistant persona** (& prompting & fine-tuning)
- **3. V Prediction objective** (& training)
- 4. *d* Transformer architecture

Transformer architecture



Transformer architecture resources

3blue1brown video (<u>#5</u> in the <u>full series</u>) *

The Illustrated Transformer

<u>A Mathematical Framework for Transformer Circuits</u>

The original "<u>Attention Is All You Need</u>" paper (*decoder-only*)

Andrej Karpathy's repo: <u>https://github.com/karpathy/minGPT</u>

(Thank you to Andy Jones at Anthropic for this list!)

Review: Deep neural network





writing

It is a component of larger Transformer Models*





It is a component of larger Transformer Models*



Transformer



Transformer



The final logits are produced by applying the unembedding.

$$T(t) \ = \ W_U x_{-1}$$

An MLP layer, m, is run and added to the residual stream.

$$x_{i+2} \ = \ x_{i+1} \ + \ m(x_{i+1})$$

Each attention head, h, is run and added to the residual stream.

$$x_{i+1} \ = \ x_i \ + \ \sum_{h \in H_i} h(x_i)$$

Token embedding.

$$x_0 = W_E t$$

I beg you, please watch this video



You can understand the entire thing, it's not too hard



Roadmap

Claude is a <u>language model</u>, with a <u>Transformer</u> architecture, <u>trained</u> with a next-token <u>prediction objective</u>, <u>prompted</u> and <u>fine-tuned</u> with an <u>assistant persona</u>.

- 1. **V** Language models
- **2. [V] Assistant persona** (& prompting & fine-tuning)
- **3. V Prediction objective** (& training)
- 4. **V** Transformer architecture

Roadmap

My goal for you:

Understand how exactly Claude (, ChatGPT, Gemini, etc.) is a machine learning model

