Intuitions on language models

Jason Wei

OpenAI

Stanford CS25 2024 Guest Lecture

Fundamental question. Why do large language models work so well?

Thing I've been thinking about recently: Manually inspecting data gives us clear intuitions about how the model works.

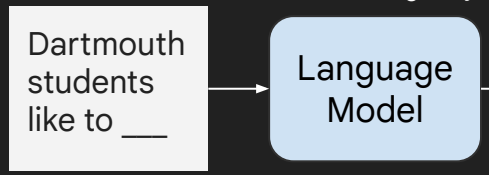
Looking at data = training your biological neural net.

Your biological neural net makes many observations about the data after reading it.

These intuitions can be valuable.

(I once manually annotated an entire lung cancer image classification dataset. Several papers came out of intuitions from that process.)

Review: language models



Pre-training only

Language Model

(hypothetical)

Word	Probability
a	*0.00001*
aardvark	*0.000004*
...	
drink	*0.5*
...	
study	*0.23*
...	
zucchini	*0.000002*

Loss = - log P(next word | previous words)

(per word, on an unseen test set)

Example. If your loss is 3, then you have a 1/(e^3) probability of getting the next token right on average.

The best language model is the one that best predicts an unseen test set (i.e., best test loss).

Intuition 1.
Next-word prediction (on large data) is massively multi-task learning.

Example tasks from next-word prediction

Task	**Example sentence in pre-training that would teach that task**
Grammar	In my free time, I like to {**code**, **banana**}
Lexical semantics	I went to the store to buy papaya, dragon fruit, and {**durian**, **squirrel**}
World knowledge	The capital of Azerbaijan is {**Baku**, **London**}
Sentiment analysis	Movie review: I was engaged and on the edge of my seat the whole time. The movie was {**good**, **bad**}
Translation	The word for “pretty” in Spanish is {**bonita**, **hola**}
Spatial reasoning	Iroh went into the kitchen to make tea. Standing next to Iroh, Zuko pondered his destiny. Zuko left the {**kitchen**, **store**}
Math question	Arithmetic exam answer key: 3 + 8 + 4 = {**15**, **11**}

[millions more]

Extreme multi-task learning!

There are a lot of possible “tasks”, and they can be arbitrary

Input	Target	Task
Biden married Neilia	Hunter	world knowledge
Biden married Neilia Hunter	,	comma prediction
Biden married Neilia Hunter ,	a	grammar
Biden married Neilia Hunter , a	student	impossible?

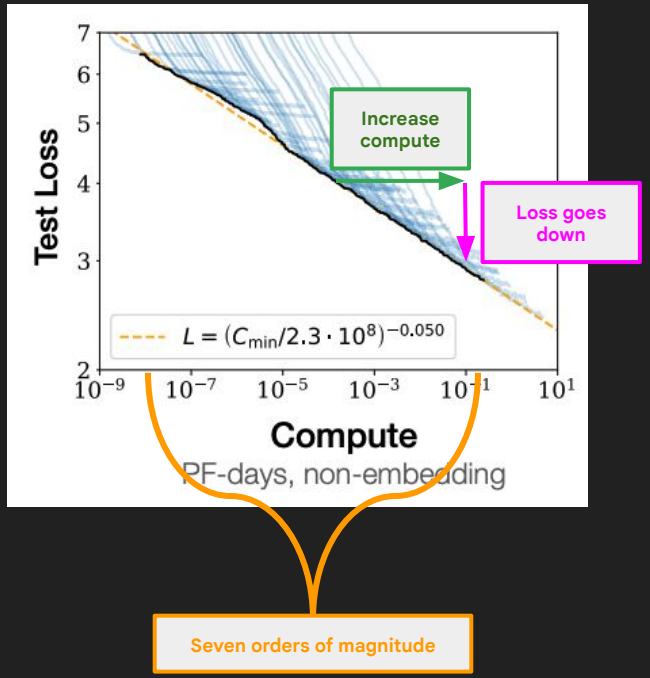
https://en.wikipedia.org/wiki/Joe_Biden

Being a language model is not easy! A lot of arbitrary words to predict. Tasks aren’t weird and not clean.

Intuition 2.
Scaling language models (size * data = compute) is reliably improves loss.

Scaling predictably improves performance (“scaling laws”)

Scaling laws for neural language models. Kaplan et al., 2020.



Kaplan et al., 2020:
“Language modeling performance improves smoothly as we increase the model size, dataset size, and amount of compute for training.”

Jason’s rephrase: You should expect to get a better language model if you scale up compute.

Why does scaling work? Hard to confirm, but just some guesses

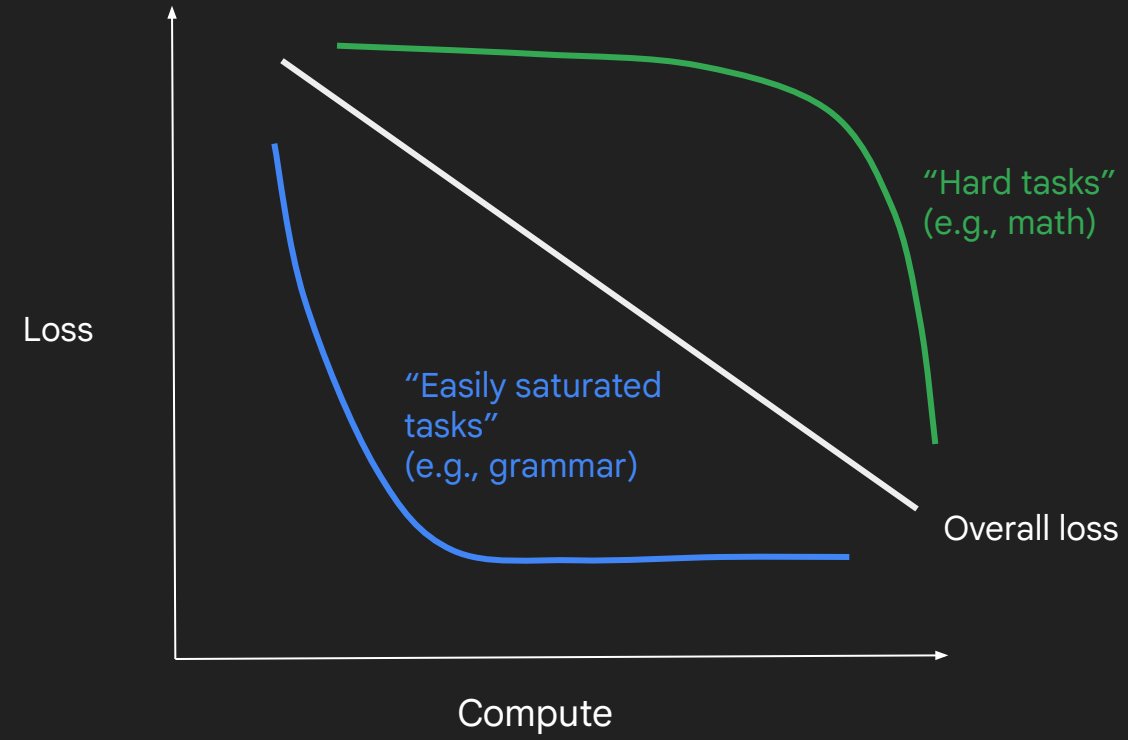
Small language model	**Large language model**
Memorization is costly *“Parameters are scarce, so I have to decide which facts are worth memorizing”*	More generous with memorizing tail knowledge *“I have a lot of parameters so I’ll just memorize all the facts, no worries”*
First-order correlations *“Wow, that token was hard. It was hard enough for me to even get it in the top-10 predictions. Just trying to predict reasonable stuff, I’m not destined for greatness.”*	Complex heuristics *“Wow, I got that one wrong. Maybe there’s something complicated going on here, let me try to figure it out. I want to be the GOAT.”*

Intuition 3.
While overall loss scales smoothly, individual downstream tasks may scale in an emergent fashion.

Take a closer look at loss. Consider:

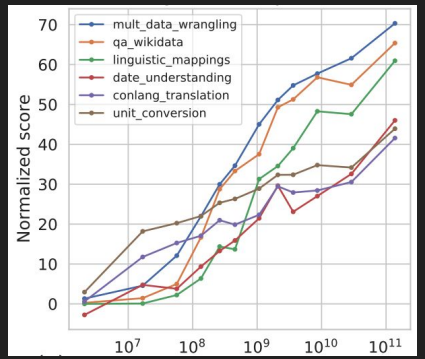
Overall loss = 1e-3 * loss_grammar +
1e-3 * loss_world_knowledge +
1e-6 * loss_sentiment_analysis +
...
1e-4 * loss_math_ability +
1e-6 * loss_spatial_reasoning
...

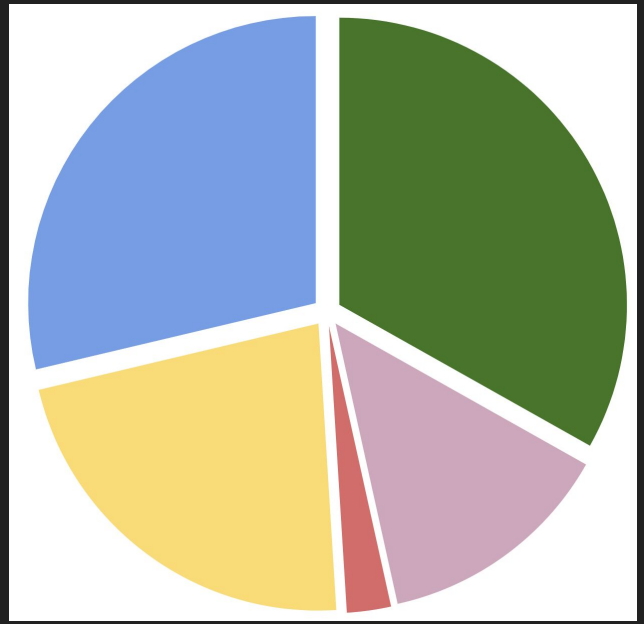
If loss goes from 4 to 3, do all tasks get better uniformly? Probably not.


Loss
"Hard tasks"
(e.g., math)
"Easily saturated
tasks"
(e.g., grammar)
Overall loss
Compute

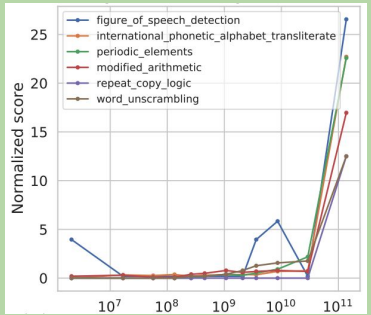
202 downstream tasks in BIG-Bench

Smoothly increasing (29%)





Emergent abilities (33%)



Not correlated with scale (13%)

Flat (22%)

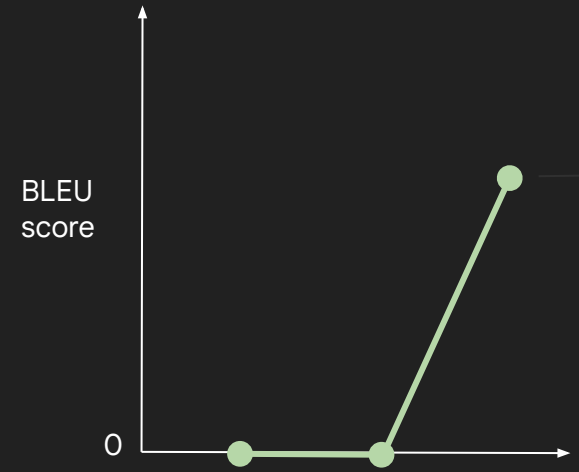
Inverse scaling (2.5%)
Performance decreases with scale

Emergence in prompting: example

Prompt

Input (English): I like to play soccer and tennis

Target (**Spanish**):



curie

Me gusta jugar al fútbol y al tenis ✓

ada

I like to play soccer and tennis ✗

babbage

I like to play soccer and tennis ✗

Model “curie” suddenly figures out to translate and not repeat.

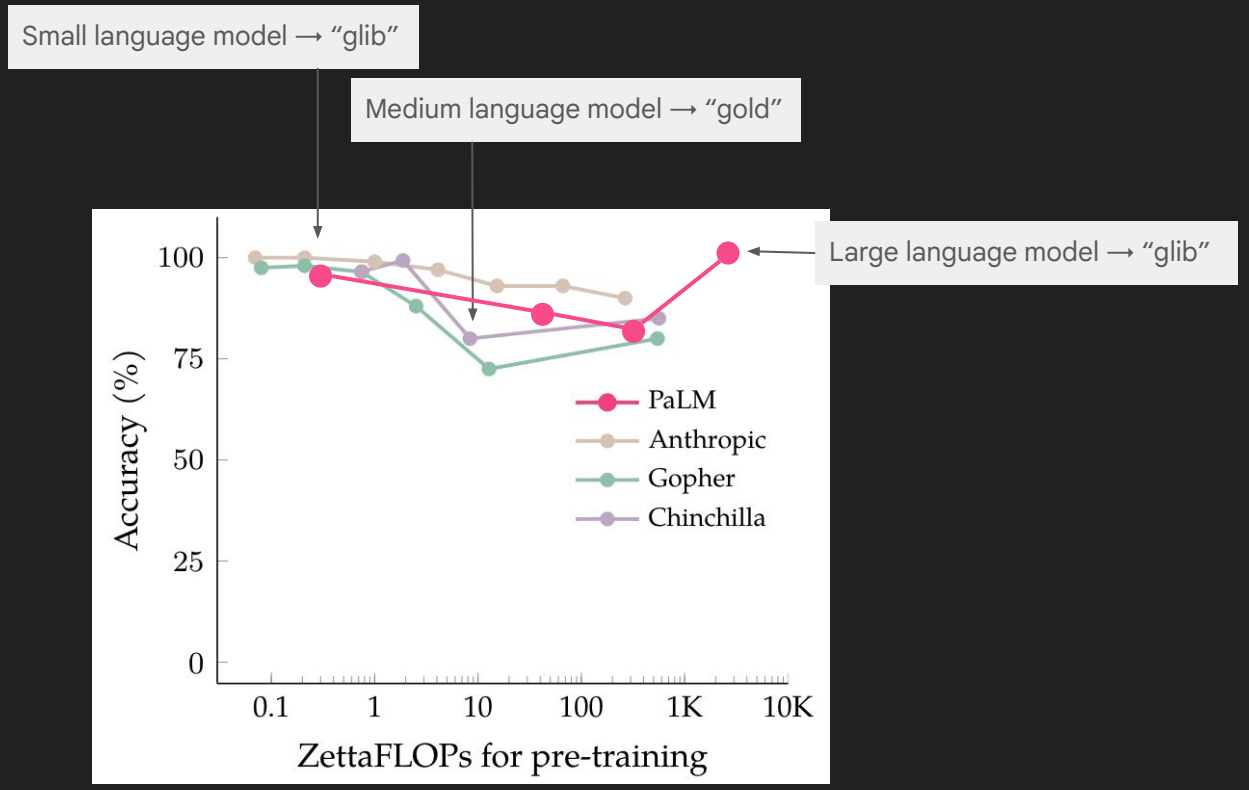
Intuition 4.
Picking a clever set of tasks results in inverse or U-shaped scaling.

Quote repetition

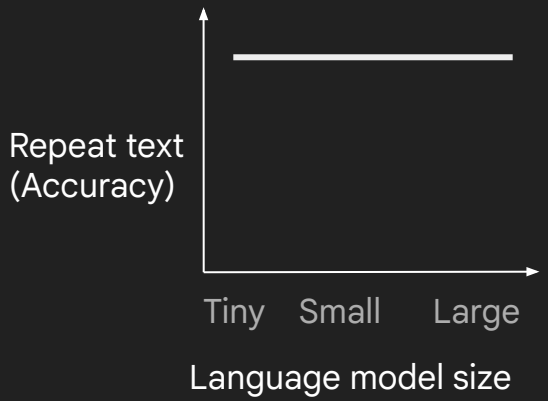
Repeat my sentences back to me.

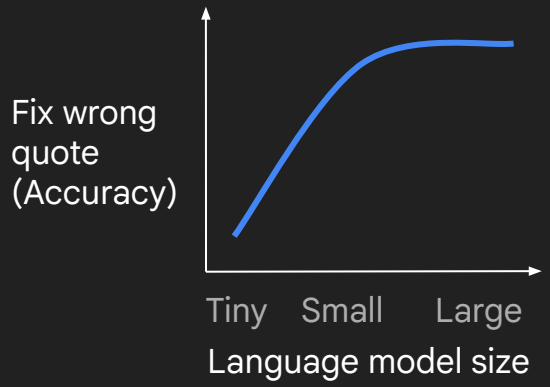
Input: All that glisters is not glib
Output: All that glisters is not ___

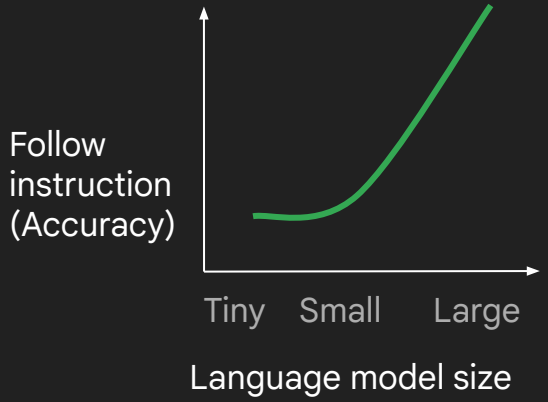
Correct answer = “glib”



Inverse scaling can become U-shaped.


Repeat text
(Accuracy)
Tiny
Small
Large
Language model size


Fix wrong
quote
(Accuracy)
Tiny
Small
Large
Language model size


Follow
instruction
(Accuracy)
Tiny
Small
Large
Language model size

Large LM intuition		General idea
Scaling model size and data is expected to continue improving loss.	→	Plot scaling curves to see if doing more of something will be a good strategy.
Overall loss improves smoothly, but individual tasks can improve suddenly.	→	To better understand aggregate metrics, decompose them into individual categories. Sometimes you'll find errors in the annotation set.

Thanks.

X / Twitter: @_jasonwei

I've love your feedback on this talk: https://tinyurl.com/jasonwei