# 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

Next-sentence prediction

Pre-training only

Dartmouth students like to ___ → 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.



Next-word prediction is really challenging!

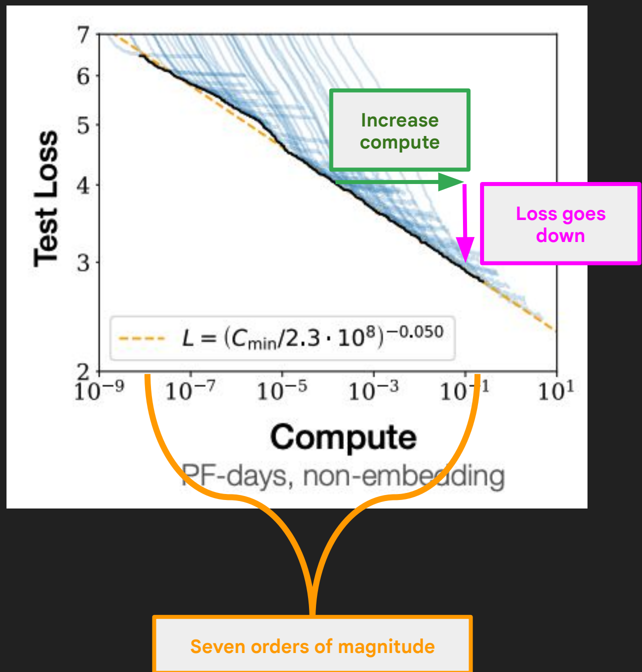
**Intuition 2.**
Scaling language models (size * data = compute) is reliably improves loss.



Reliable improvement. Read paper scaling laws for Neural Language models. Kaplan et all 2020

# 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.



Read this!

Predict loss based on compute.

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

| **Small language model** | **Large language model** |
|---|---|
| Memorization is costly<br><br>*"Parameters are scarce, so I have to decide which facts are worth memorizing"* | More generous with memorizing tail knowledge<br><br>*"I have a lot of parameters so I'll just memorize all the facts, no worries"* |
| First-order correlations<br><br>*"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<br><br>*"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."* |



small LM
Memorization
Learn first order correlations

Large LM
Memorize tail knowledge
Learn complex heuristics.

**Intuition 3.**

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



→ Emergent abilities

While overall loss improves smoothly, individual task can improve suddenly.

Take a closer look at loss. Consider:

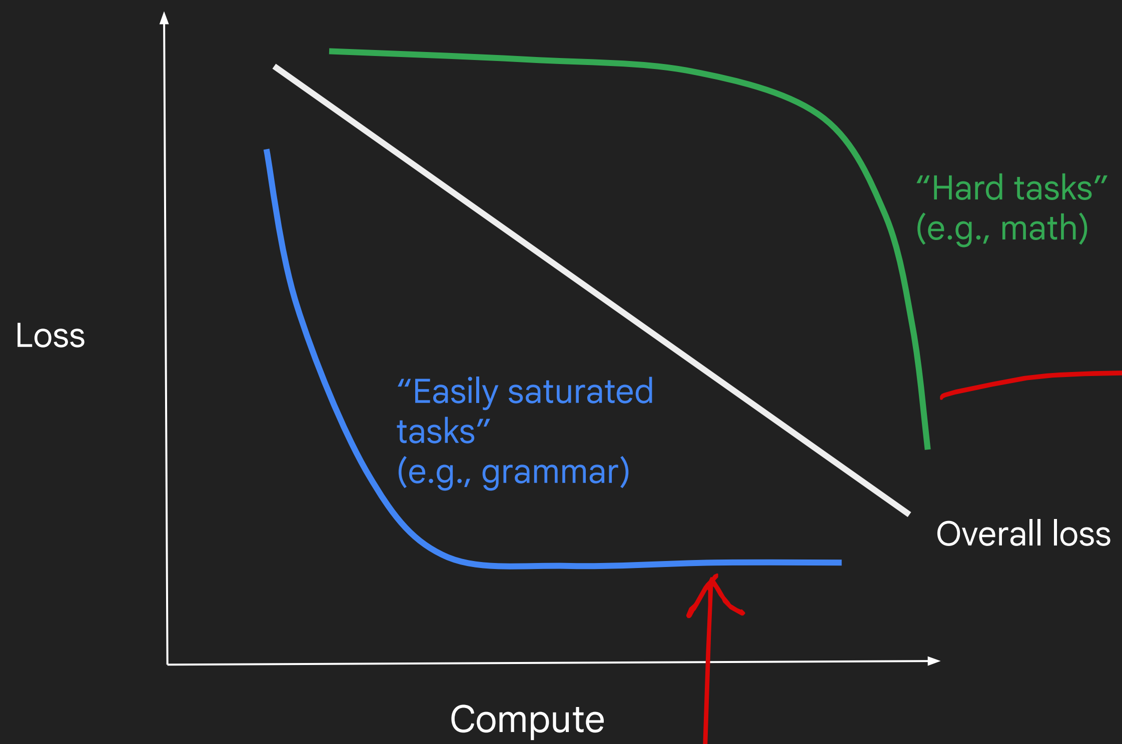
Overall loss = 1e-3 * loss_grammar +
1e-3 * loss_world_knowledge +
1e-6 * loss_sentiment_analysis +
...
1e-4 * loss_math_ability +
1e-6 * loss_spatial_reasoning
...

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



Decompose overall loss into loss of smaller components. Weighted sum of individual tasks loss.

Improvement in loss probably means one task loss ↓ a lot.



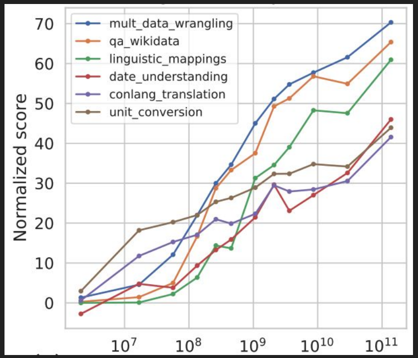


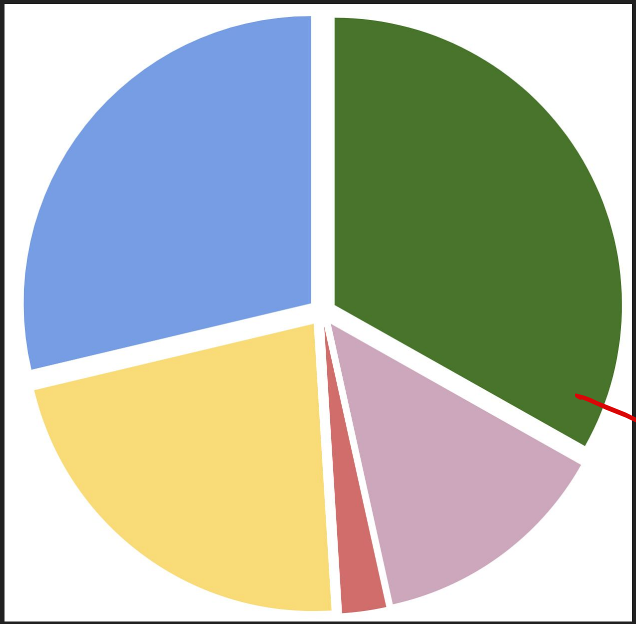

Easy tasks : Loss reduces fast

Hard tasks loss slow

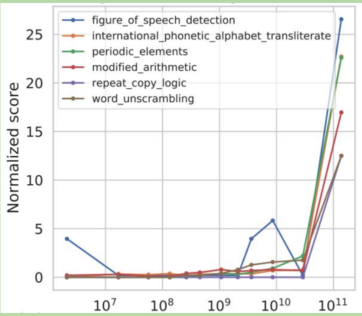
# 202 downstream tasks in BIG-Bench

**Smoothly increasing (29%)**



**Emergent abilities (33%)**




Not correlated with scale (13%)

Flat (22%)

Inverse scaling (2.5%)
*Performance decreases with scale*



Task was too hard

Acc gets worse as size of LM ↑

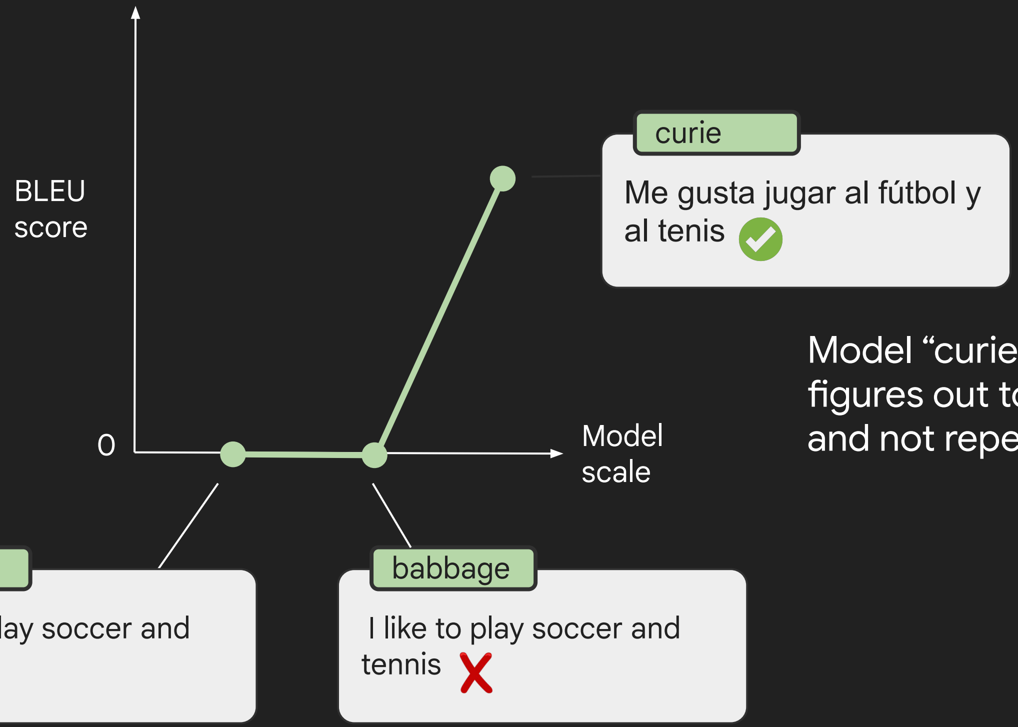
Acc suddenly starts to ↑

# Emergence in prompting: example

**Prompt**

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

Target (**Spanish**):




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.



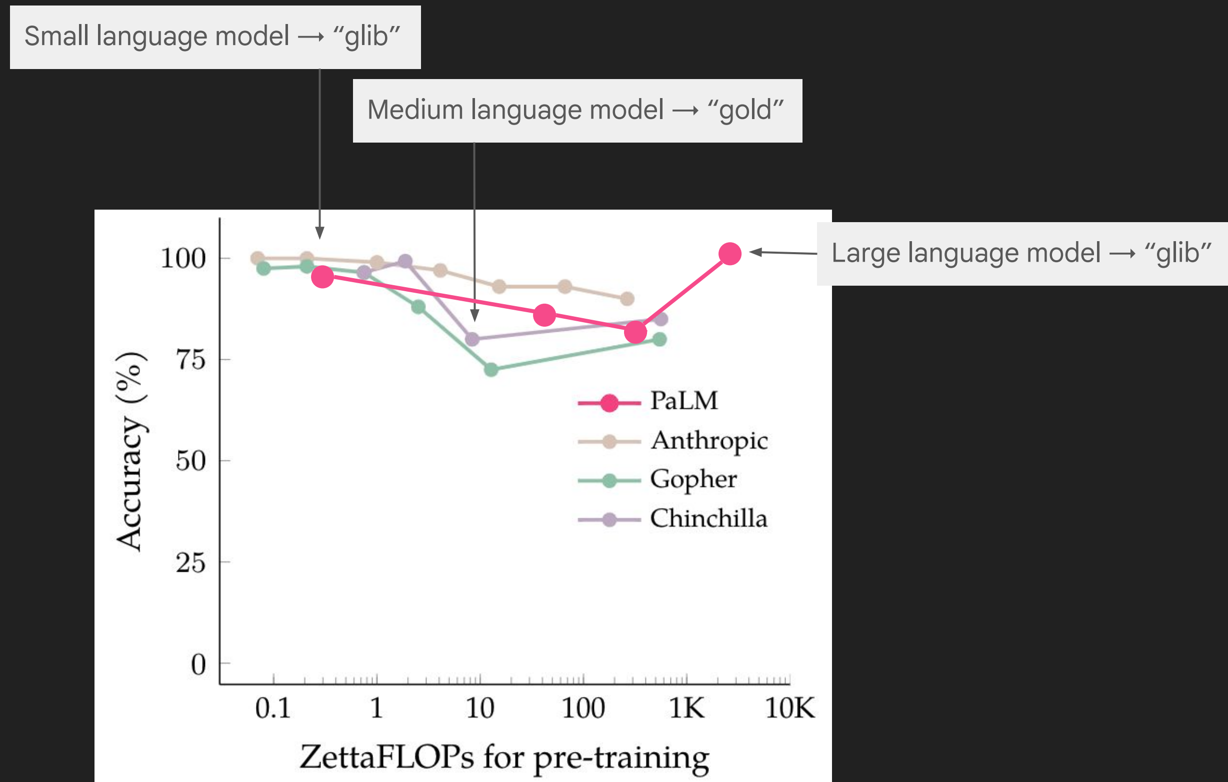
Inverse Scaling

## Quote repetition

Repeat my sentences back to me.

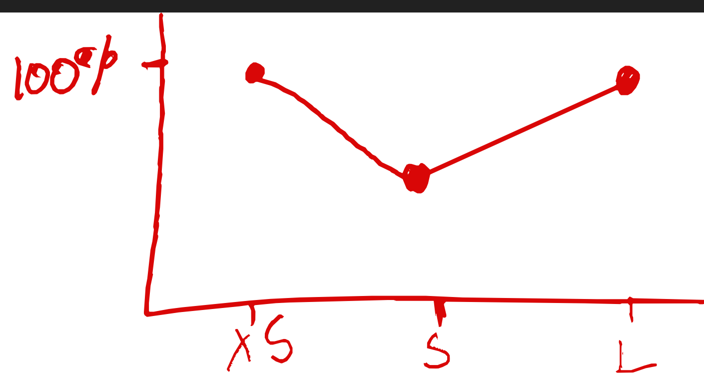
Input: All that glisters is not glib
Output: All that glisters is not ___

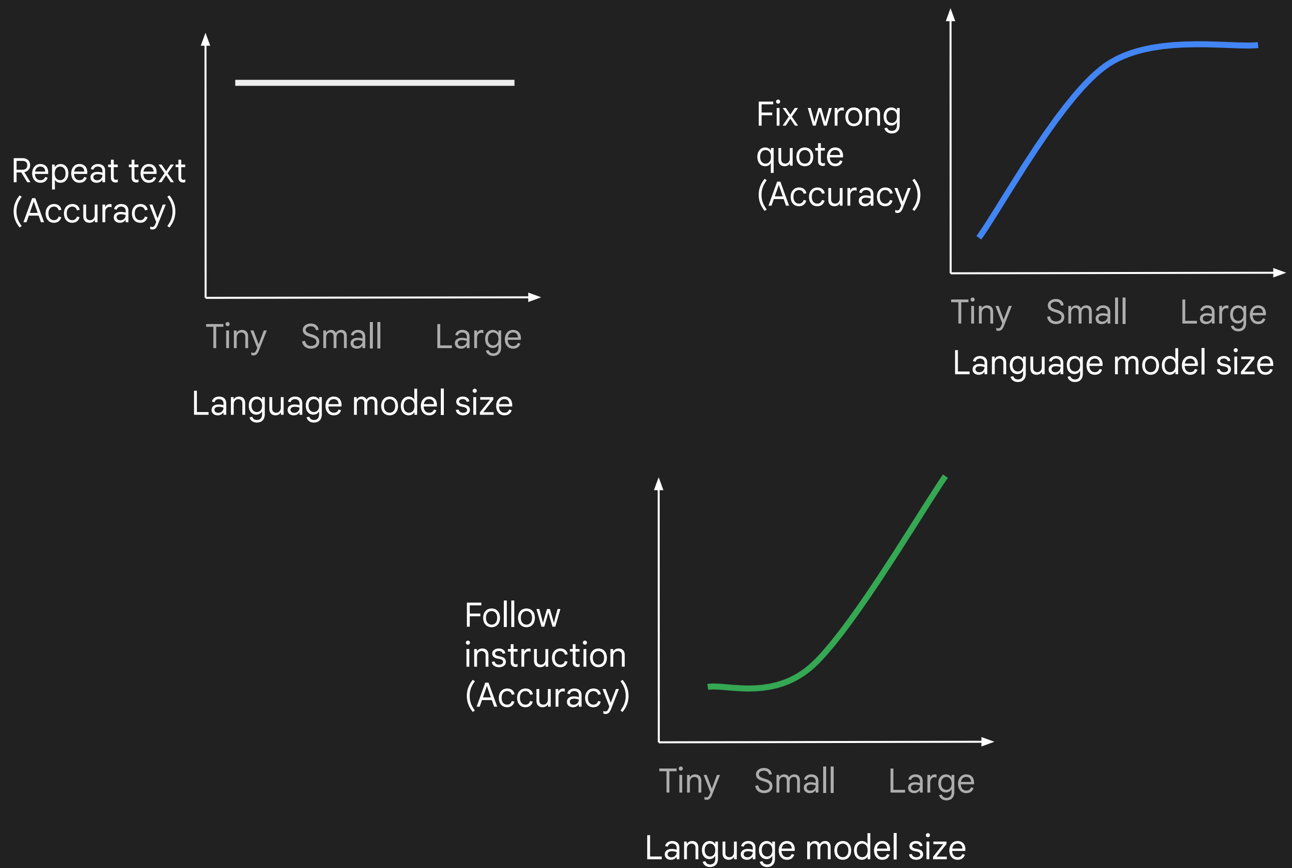
Correct answer = “glib”




[Inverse scaling can become U-shaped.]










→ Decompose the task into 3 sub-tasks

XS can repeat, can't fix quote & can't follow → Solves task

S can repeat, can fix quote & can't follow → Fixes quote

L can do all 3 → Solves task

## Large LM intuition

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.

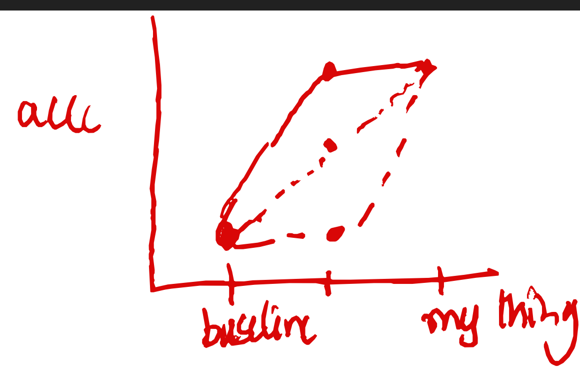
## General idea

(see right-hand column above)



Takeaways:

→ Plot scaling curves




Can't differentiate btw good & bad data during pre-training. Just hope to select good datasets.

Bottlenecks for LLM → Compute & data

Read paper: Are emergent abilities a mirage?

# Thanks.

X / Twitter: @_jasonwei

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