You Know It or You Don’t:
Categorical Differences in Model Behavior



Naomi Saphra
Boston University

How do we usually analyze language models?

- Behavioral test, internal metric, test time intervention
- Fully trained model
  - **Without its training history**
- Single set of parameters
  - **Sampling one individual model**

How do we usually analyze language models?

- Behavioral test, internal metric, test time intervention
- Fully trained model
  - **Without its training history**
- Single set of parameters
  - **Sampling one individual model**

Does this explain why a model works?

What is good explanatory evidence?

Some lessons from an older science.

The parable of the prinia’s egg



“It’s camouflage!”

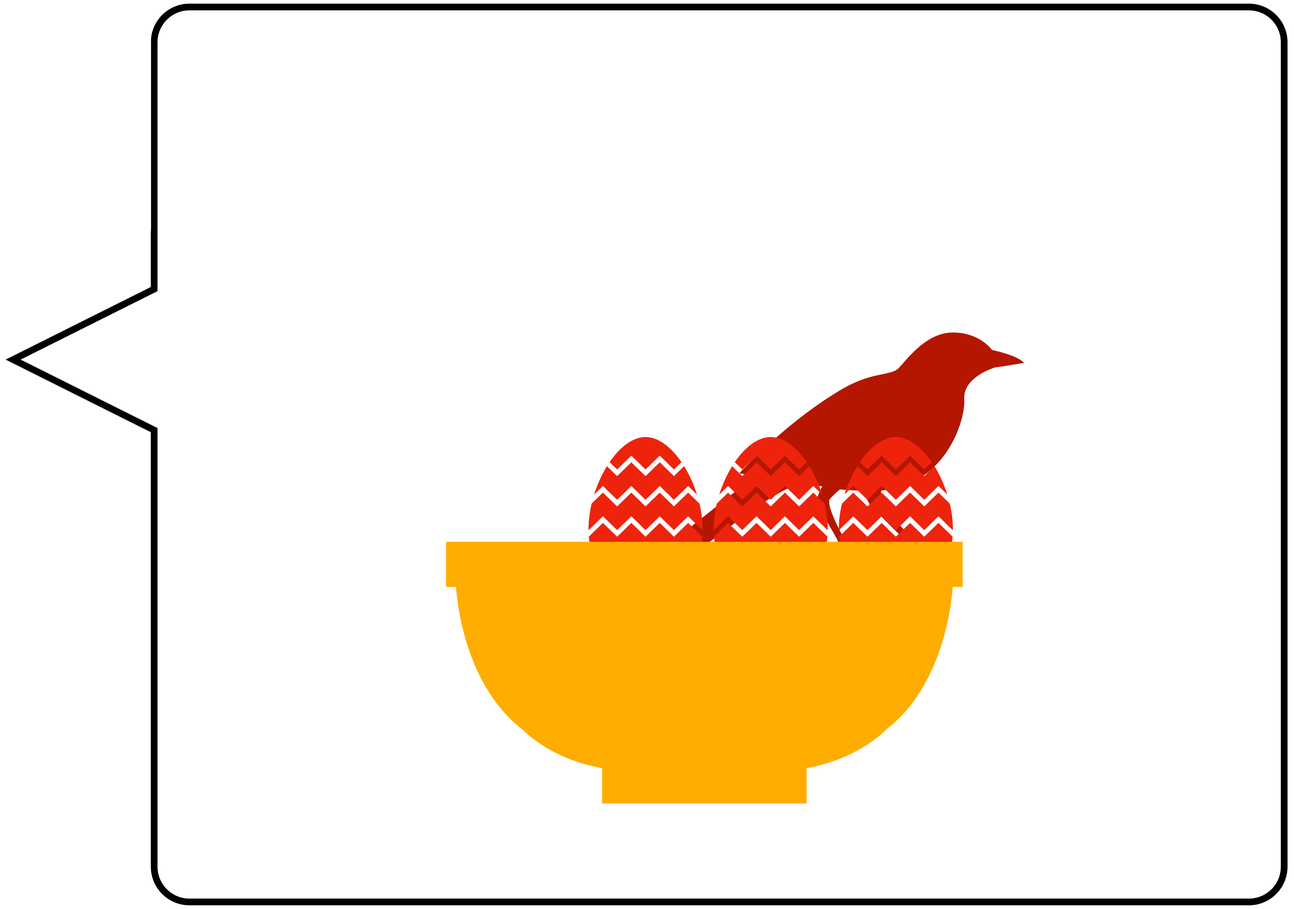
- 19th century scientists



Brood parasite defense



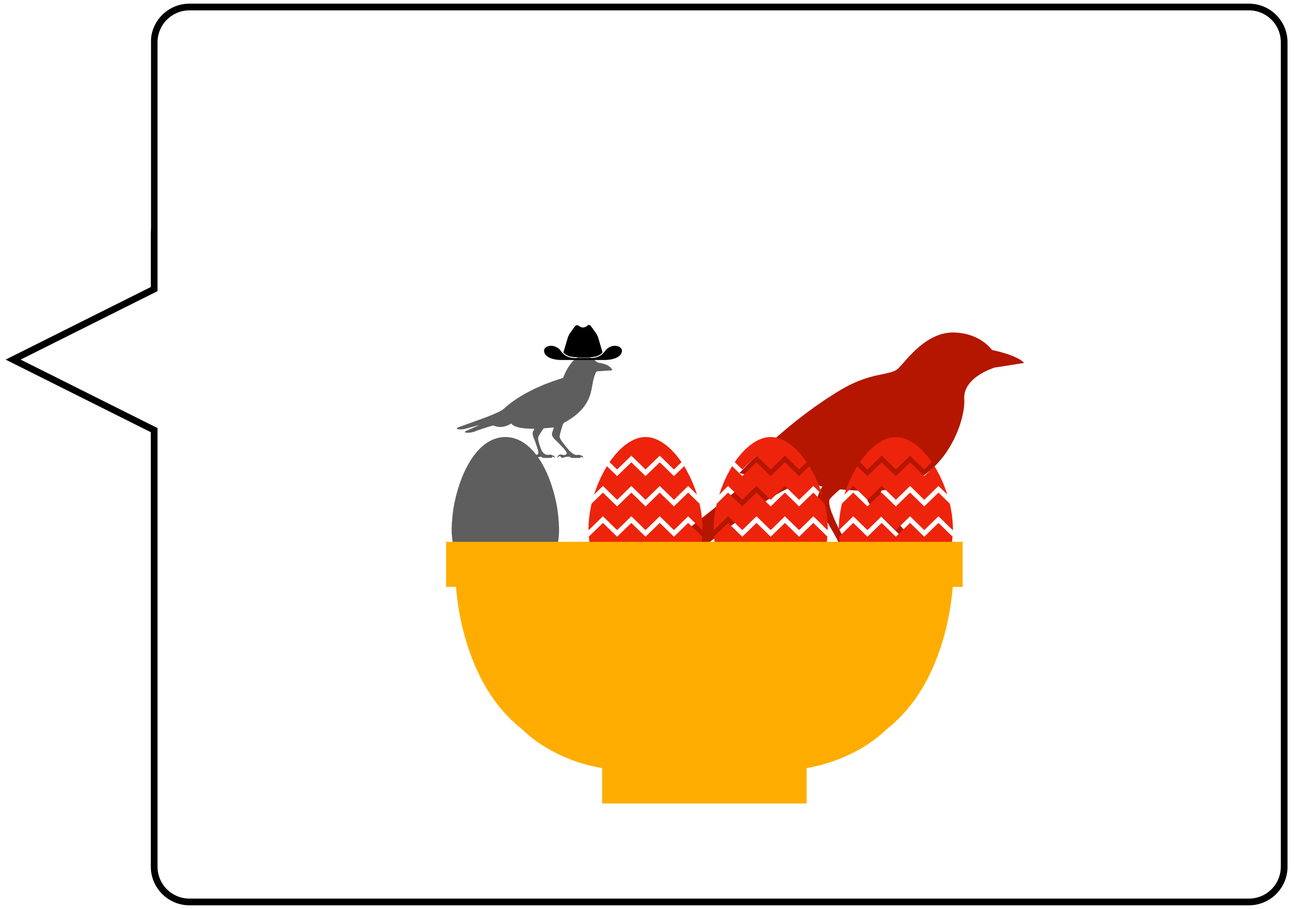
Charles Swynnerton, CMG



Brood parasite defense



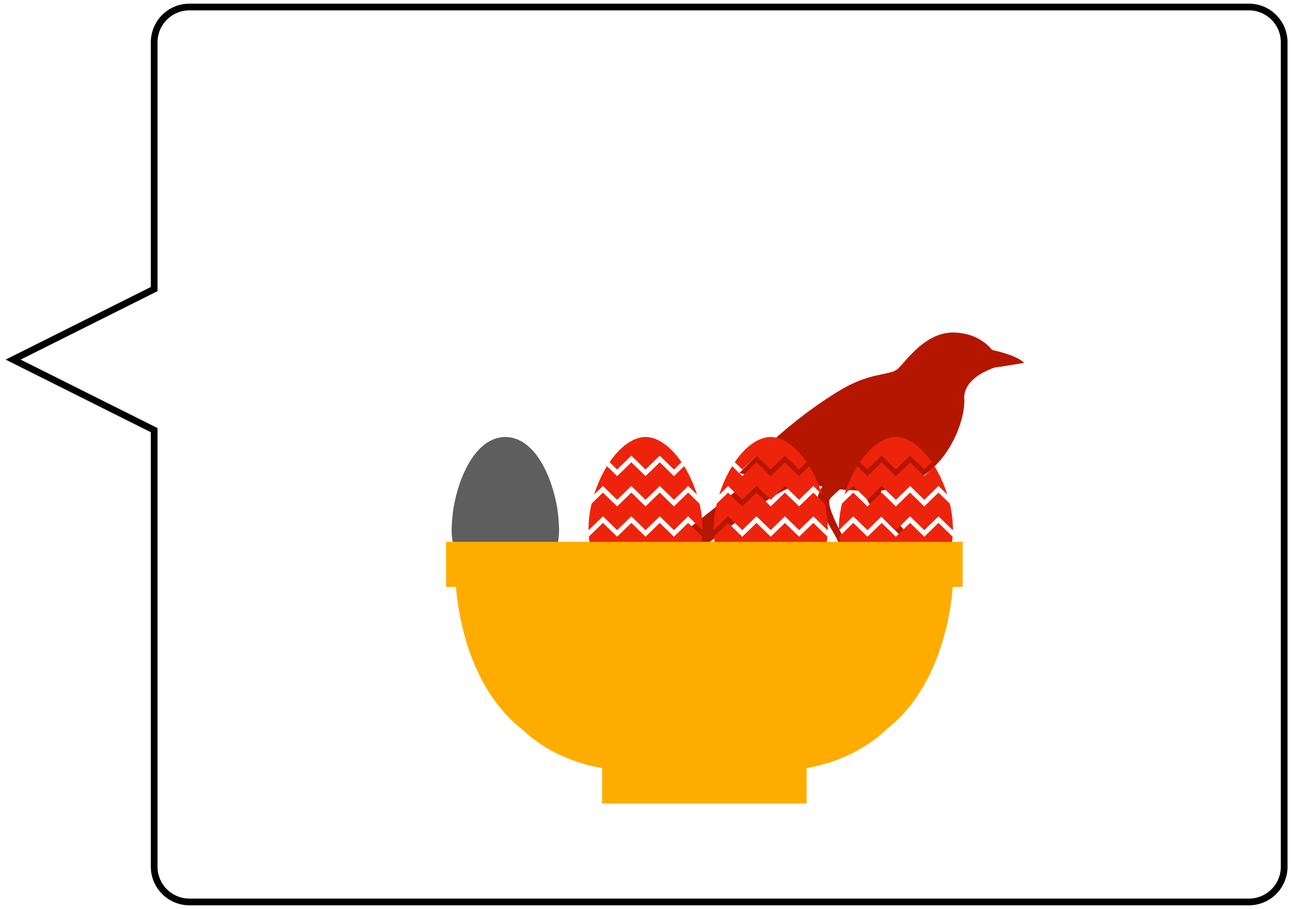
Charles Swynnerton, CMG



Brood parasite defense



Charles Swynnerton, CMG



Brood parasite defense

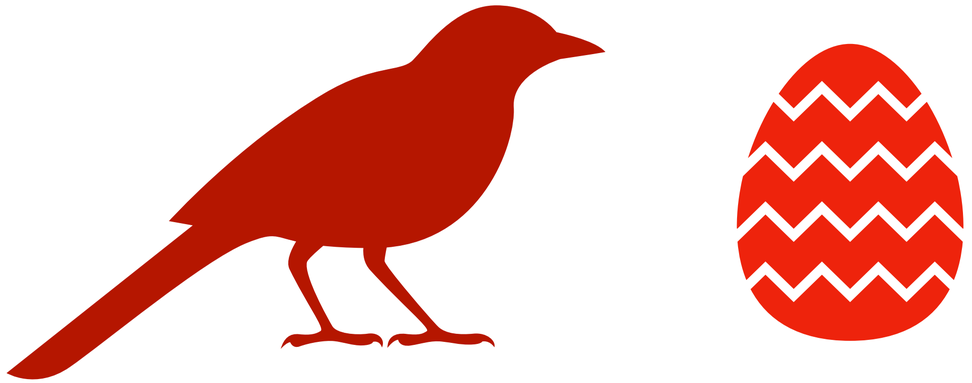


Charles Swynnerton, CMG

You need to recognize a dumped egg to toss it out!

But how do we know?

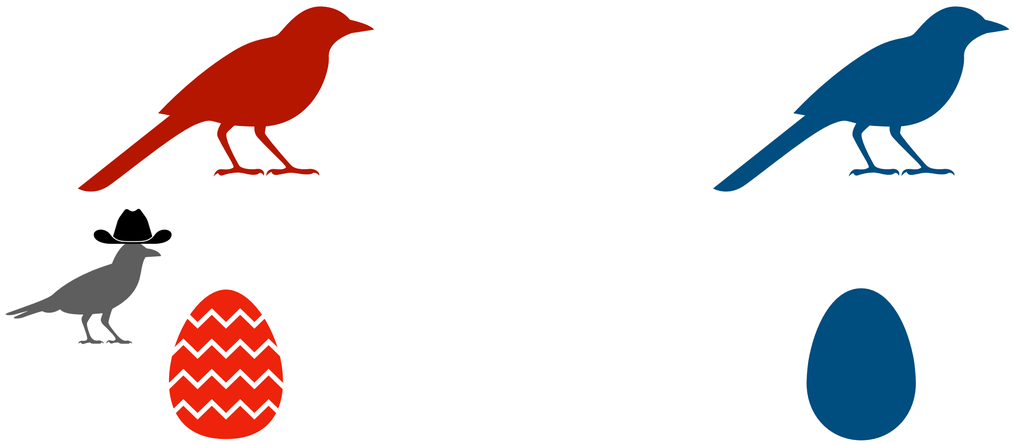
Not from a single individual at a single moment in time!



“It’s camouflage!”

The evidence

- **Correlational lens:** Comparing individuals with and without the trait.



Preferred hosts have more unique eggs.

The evidence

- **Correlational lens:** Comparing individuals with and without the trait.
- **Developmental lens:** Studying development of the trait over time.

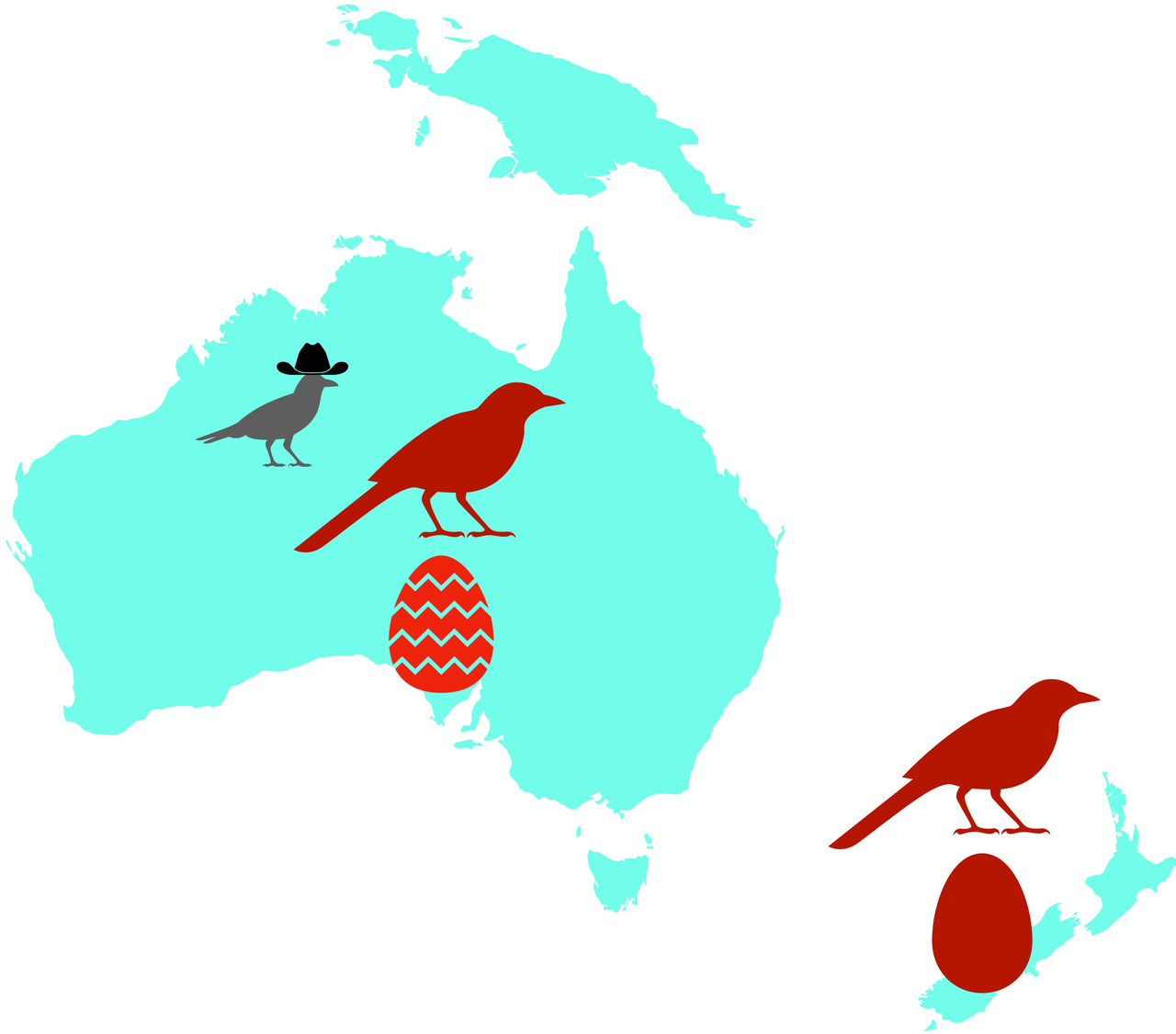


On continents with longer histories of brood parasitism, birds have more unique eggs.

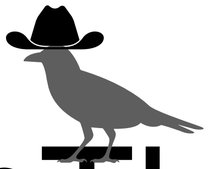
The evidence

- **Correlational lens:** Comparing individuals with and without the trait.
- **Developmental lens:** Studying development of the trait over time.
- **Causal developmental lens:** Intervening during development to discover effects on the trait.

Invasive species that move to islands without brood parasites lose their egg patterns.



Let’s try understanding something about language models!

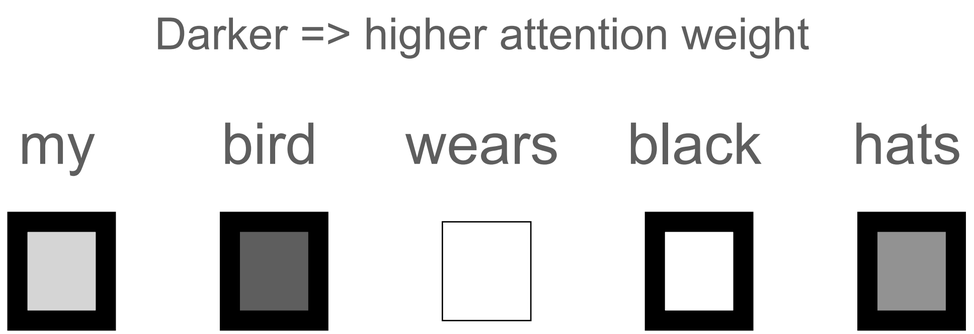


- The correlational lens
- The developmental lens
- The causal developmental lens

Masked Language Modeling with BERT

(Bidirectional Encoder Representations from Transformers)

- The task: predict a masked out (missing) word from a sequence.
- Made up of **attention heads**, which compute an **attention distribution** to reweight the importance of each word in a sequence.



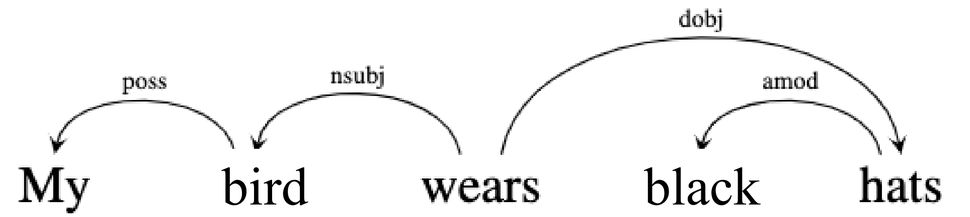
Target: “wears”



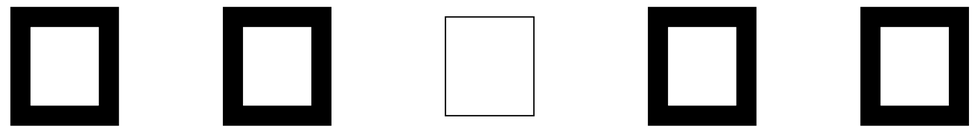
Internal Syntax

Voita et al., 2019 and Clark et al., 2019

- Some heads attend more to specific syntactic relations.



nsubj head

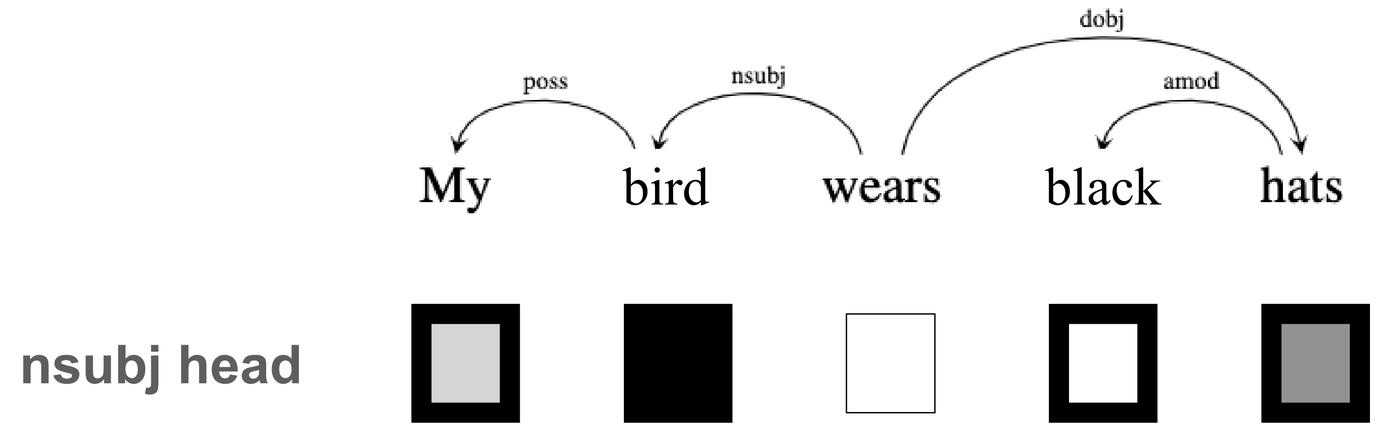


Target: “wears”

Internal Syntax

Voita et al., 2019 and Clark et al., 2019

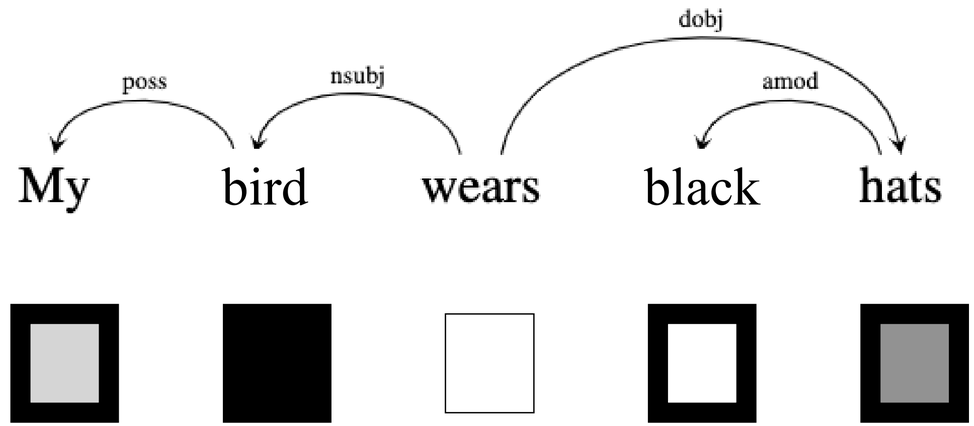
- Some heads attend more to specific syntactic relations.



Internal Syntax

Voita et al., 2019 and Clark et al., 2019

- Some heads attend more to specific syntactic relations.
- Naturally emerging property in masked language models!



nsubj head

Target: “wears”

We know masked LMs have specialized heads

But are they important for grammatical understanding? Evidence:

- **Instance level observations**
  - Specialized heads exhibit internal syntax (Clark et al., 2019).

What if these artifacts are just a side effect of training?

- **Instance level causal intervention**
  - Removing specialized heads hurts performance (Voita et al., 2019).

What if they are important for reasons other than internal syntax?



- **The correlational lens**
- The developmental lens
- The causal developmental lens

Measuring grammatical capability with examples from linguistics papers

Warstadt et al. (2019)

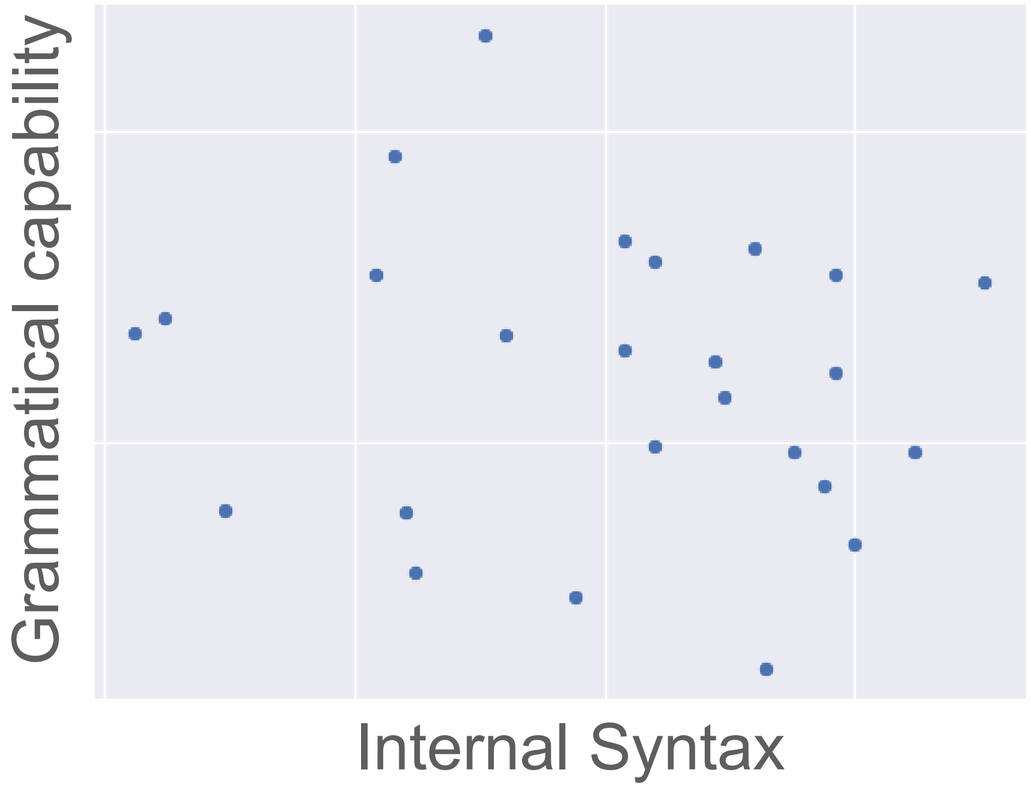
- Example: Negative polarity indexing
  - “I go to the movies anymore.”
  - “I don’t go to the movies anymore.”





Is Internal Syntax correlated with grammatical capability?

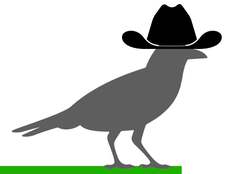
No. (Across 25 independent runs with different random seeds.)



Is Internal Syntax a red herring for the grammatical capability mechanism?

The evidence

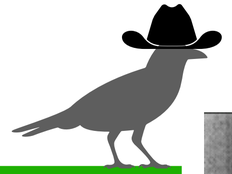
- **Correlational lens:** Comparing individuals with and without the trait.



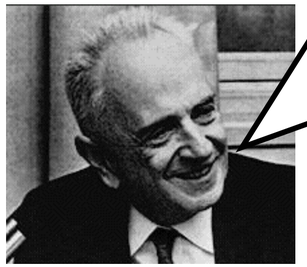
- **Developmental lens:** Studying development of the trait over time.
- **Causal developmental lens:** Intervening during development to discover effects on the trait.

The evidence

- **Correlational lens:** Comparing individuals with and without the trait.



- **Developmental lens:** Studying development of the trait over time.
- **Causal developmental lens:** Intervening during development to discover effects on the trait.

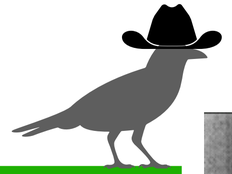


Theodosius Dobzhansky

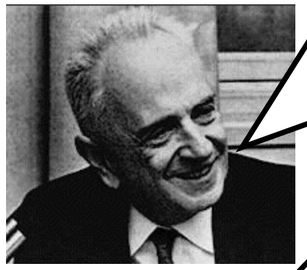
Nothing in biology makes sense except in the light of evolution.

The evidence

- **Correlational lens:** Comparing individuals with and without the trait.



- **Developmental lens:** Studying development of the trait over time.
- **Causal developmental lens:** Intervening during development to discover effects on the trait.



Theodosius Dobzha

Nothing in biology makes sense except in the light of evolution.

Nothing in **deep learning** makes sense except in the light of **stochastic gradient descent**.



Naomi Saphra

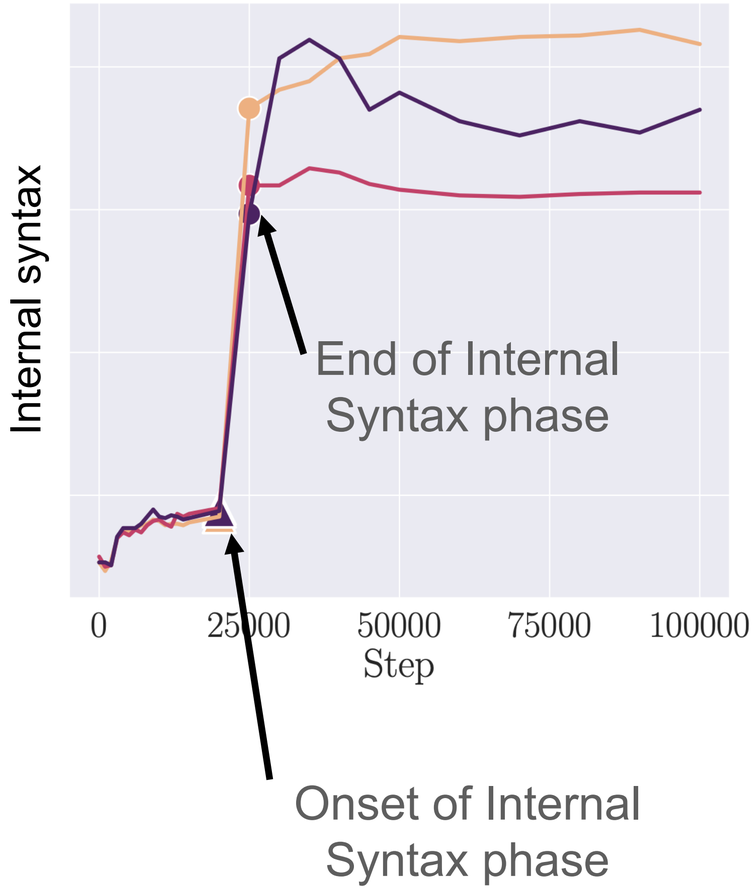
- The correlational lens
- **The developmental lens**
- The causal developmental lens



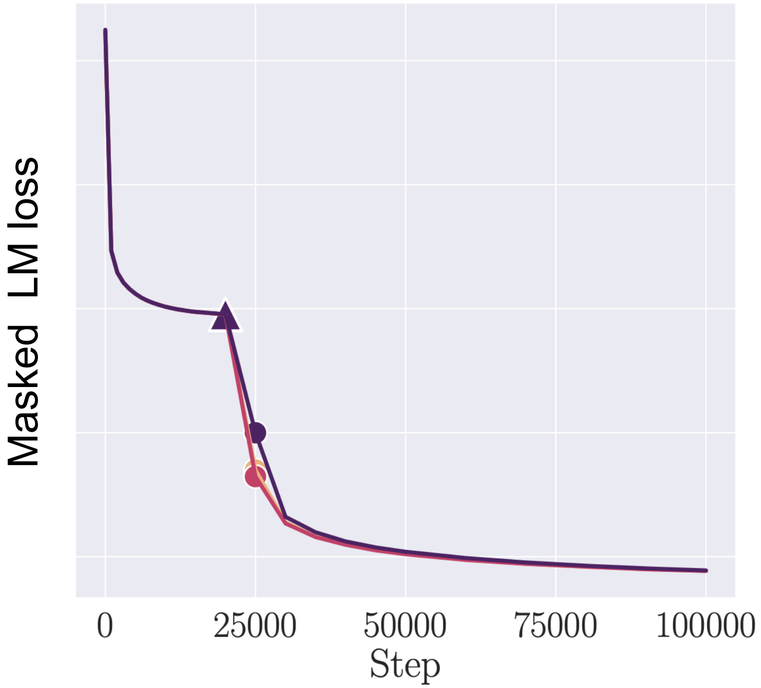
When does Internal Syntax emerge?

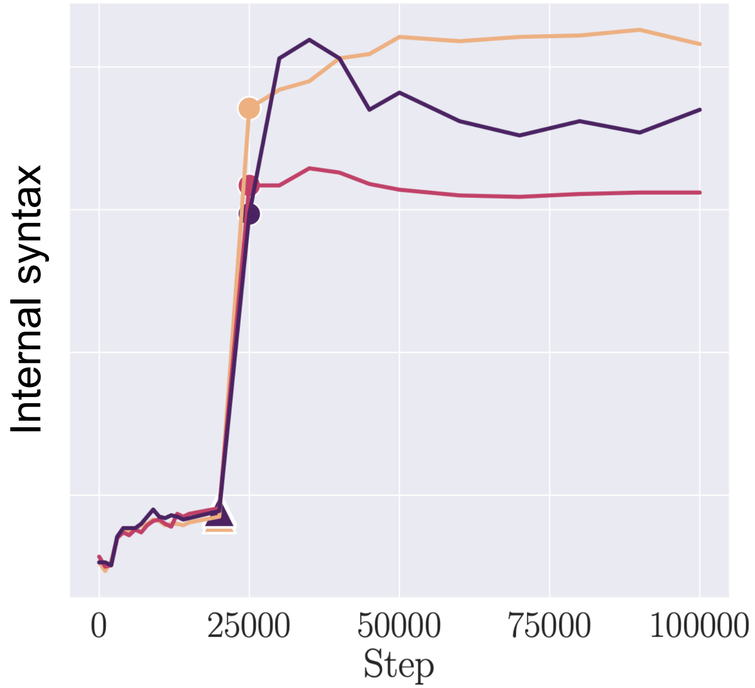
Chen et al. Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in MLMs. *ICLR*, 2024.

Internal Syntax is acquired abruptly

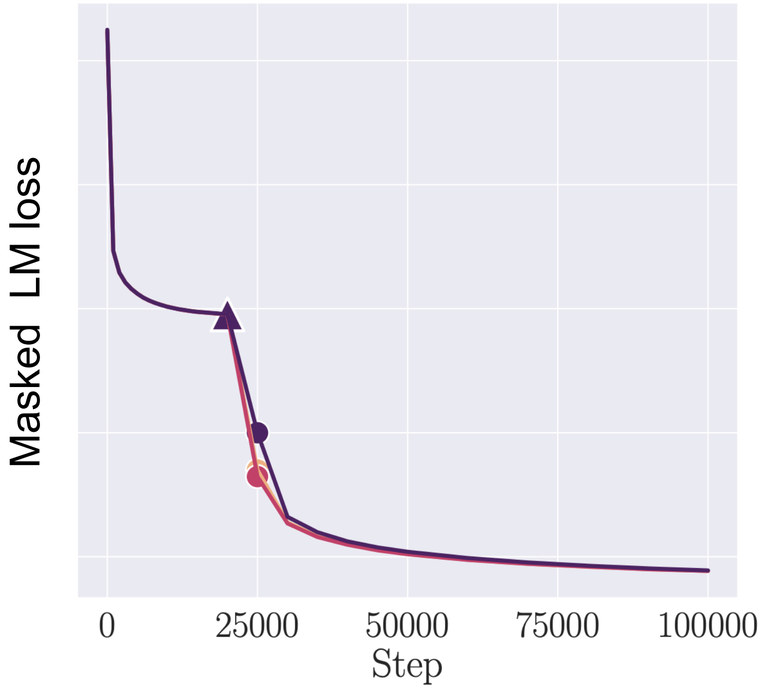


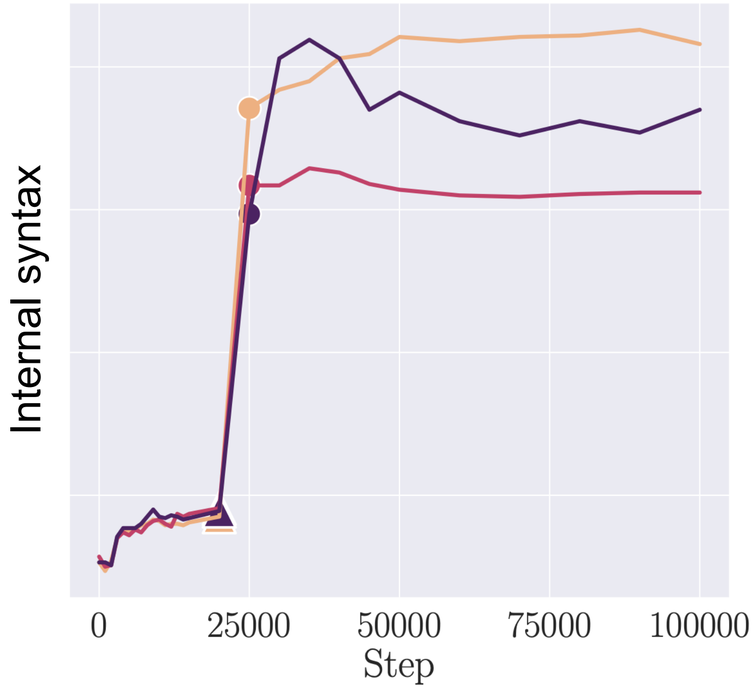
SAS phase accompanies a large loss drop

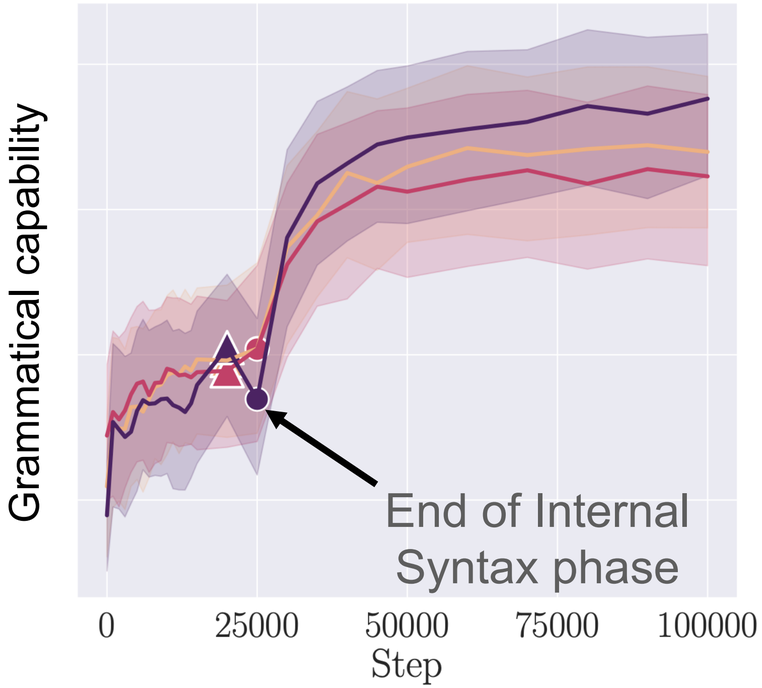




And is followed by gains in grammatical reasoning





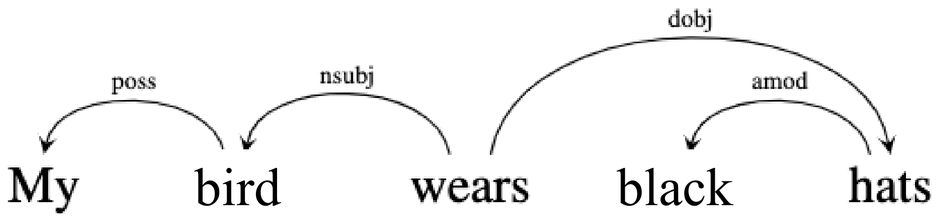


- The correlational lens
- The developmental lens
- **The causal developmental lens**



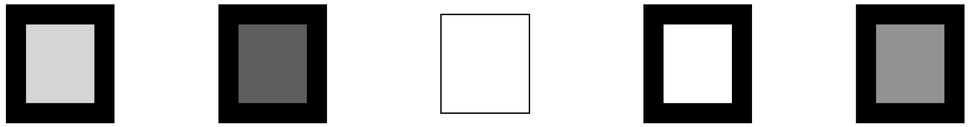
What happens if we suppress Internal Syntax?

Suppressing Internal Syntax

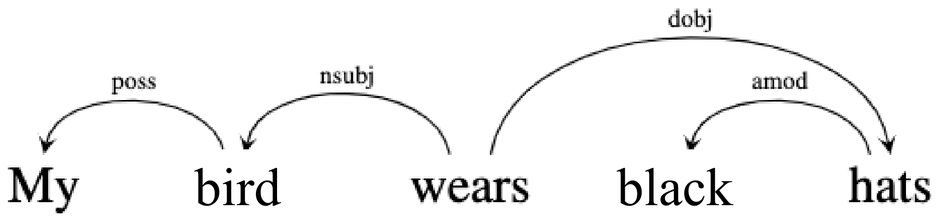


Target: “wears”

nsubj head

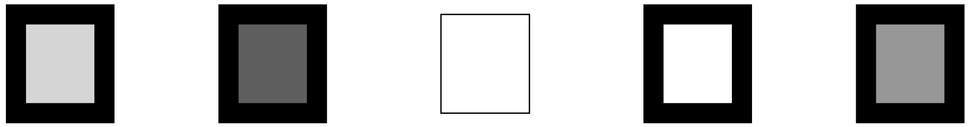


Suppressing Internal Syntax

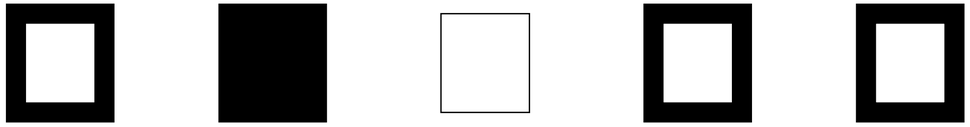


Target: “wears”

nsubj head



nsubj gold labels

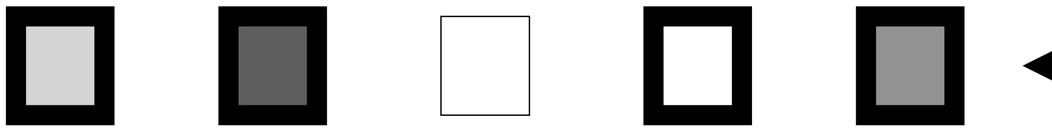


Suppressing Internal Syntax

$$\gamma(\alpha_i^k, x_i) = \frac{\alpha_i^k \cdot \mathbb{1}[D(x_i)]}{\|\mathbb{1}[D(x_i)]\|_2}$$

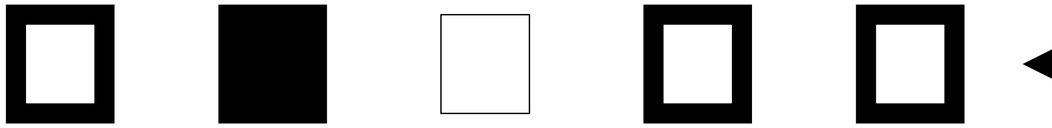
$$L(x) = L_{\text{MLM}}(x) + \lambda \sum_{i=1}^{s} \max_k \gamma(\alpha_i^k, x_i)$$

nsubj head



α_i^k

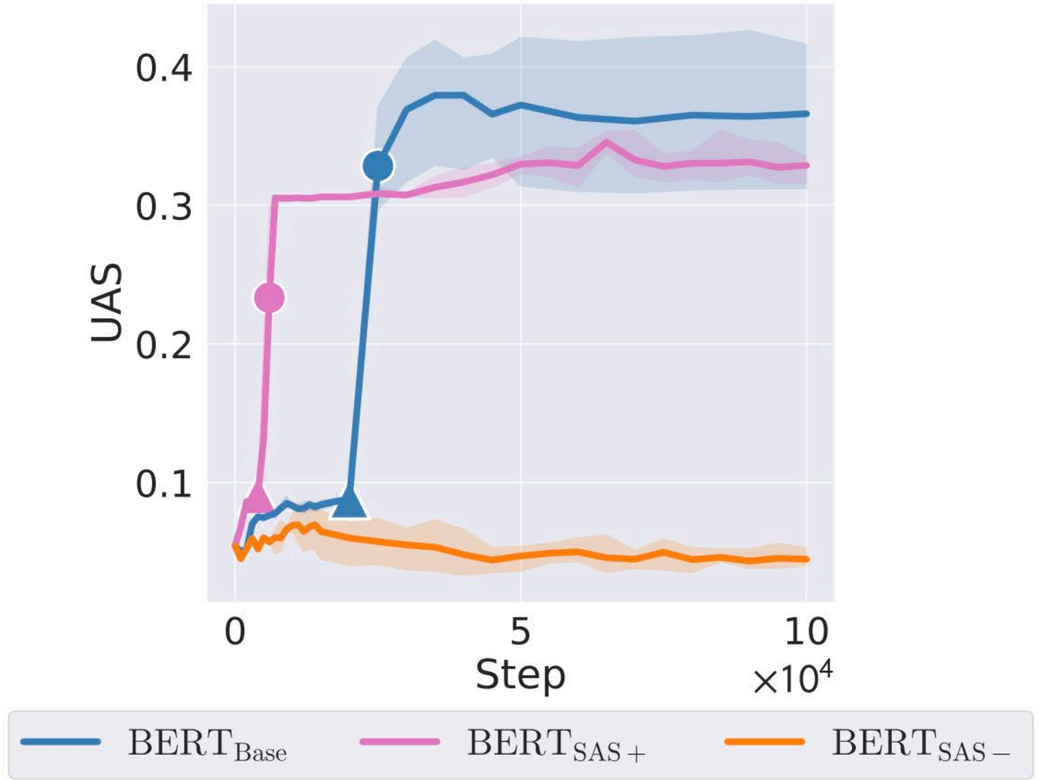
nsubj gold labels



$\mathbb{1}[D(x_i)]$

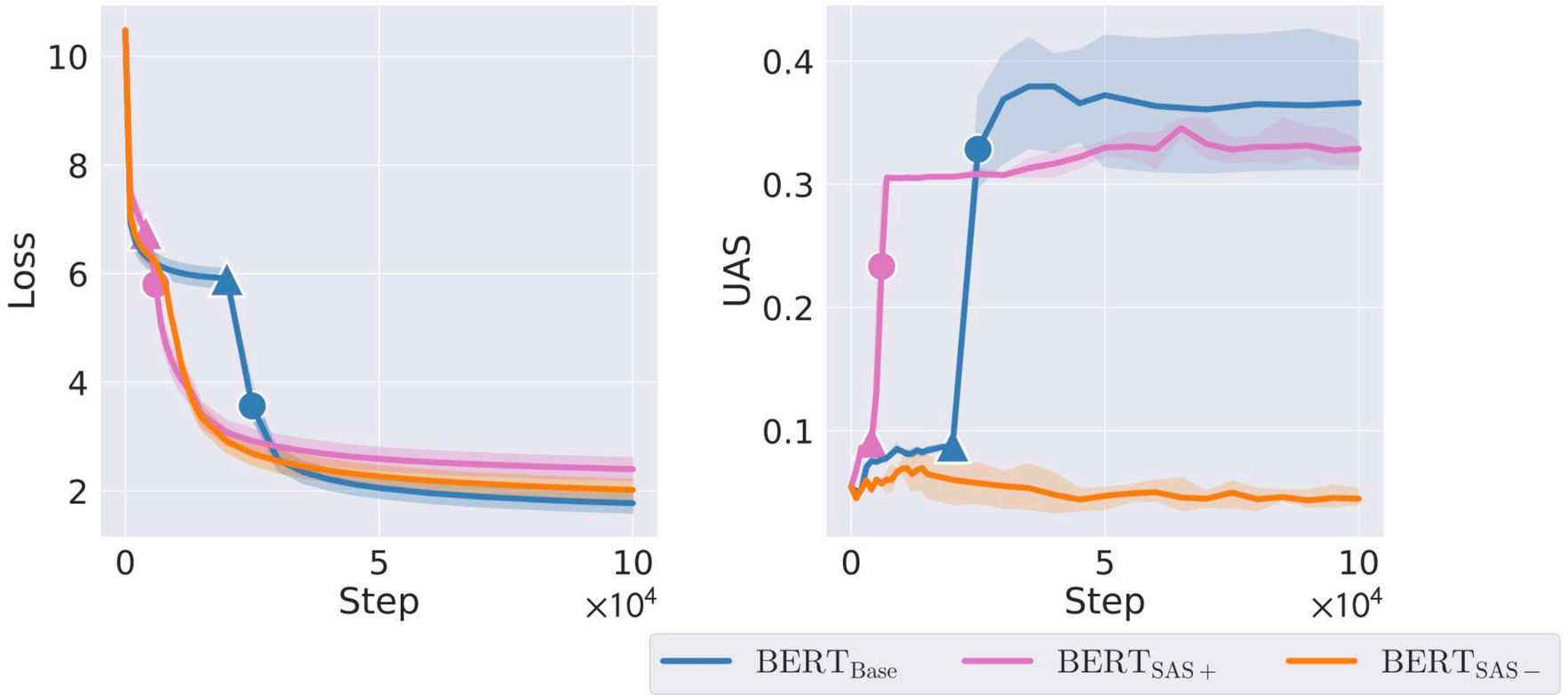
Causal Impact of Internal Syntax

We can prevent or accelerate the SAS phase transition …



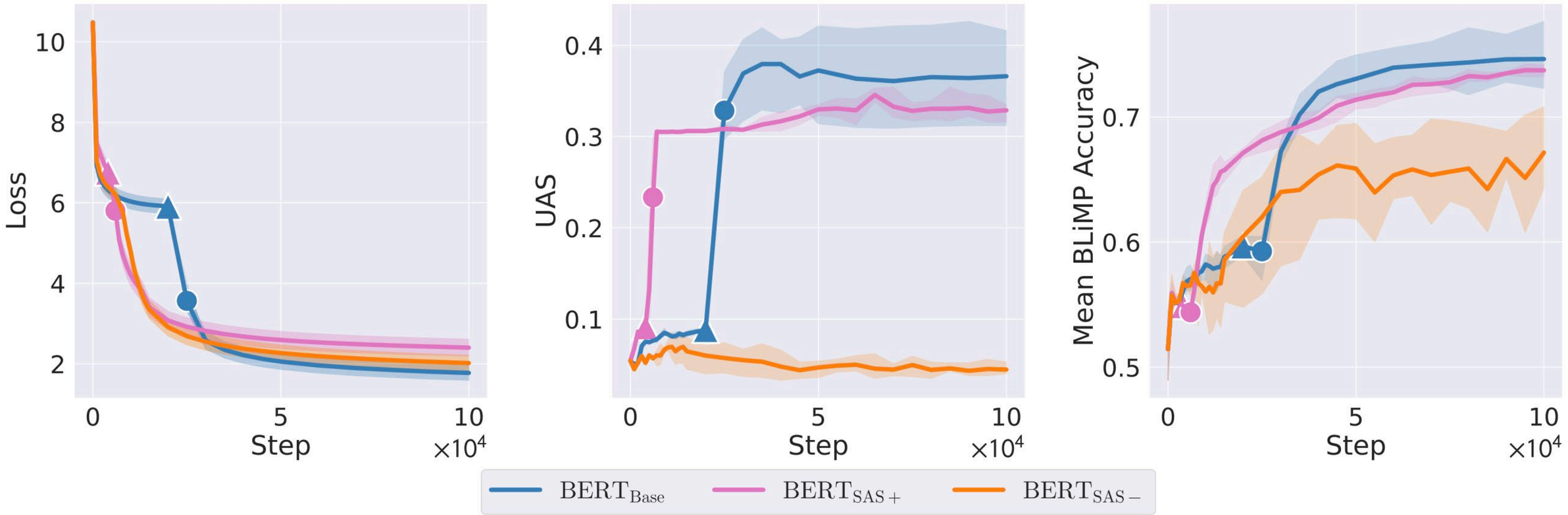
Causal Impact of Internal Syntax

Accelerating SAS accelerates the timing of the initial loss drop …



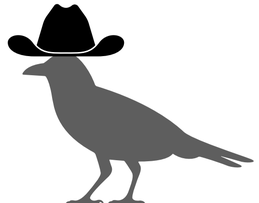
Causal Impact of Internal Syntax

And the SAS onset precipitates the grammar capabilities onset!

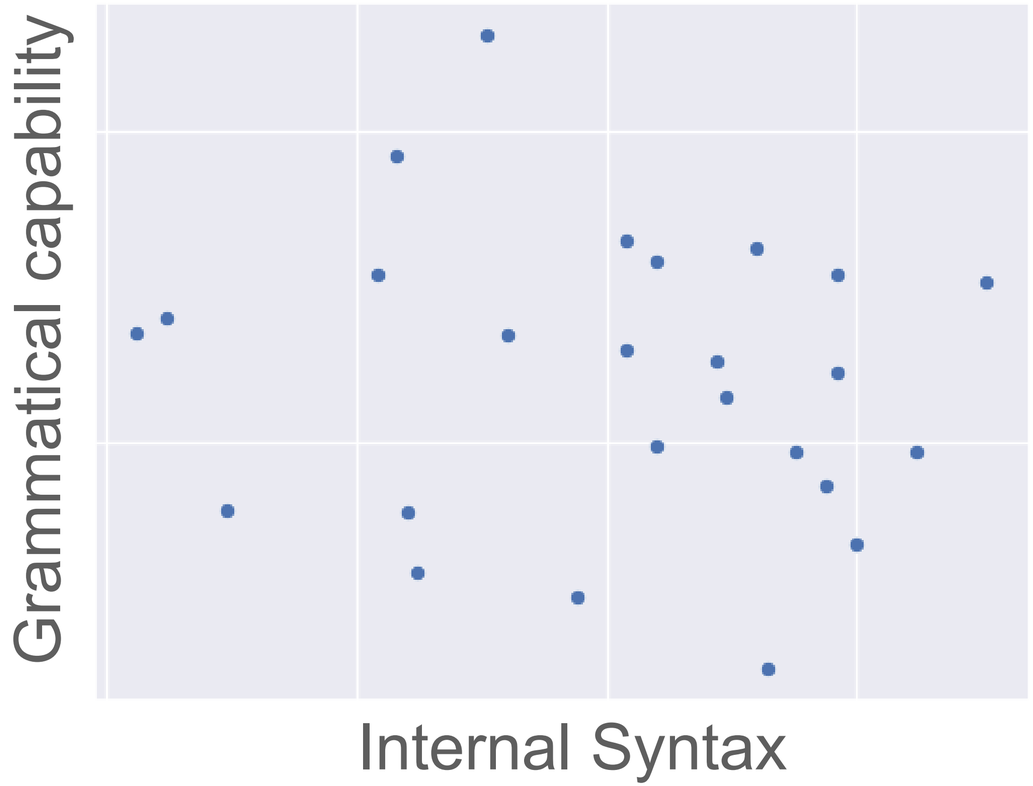


The evidence - recap

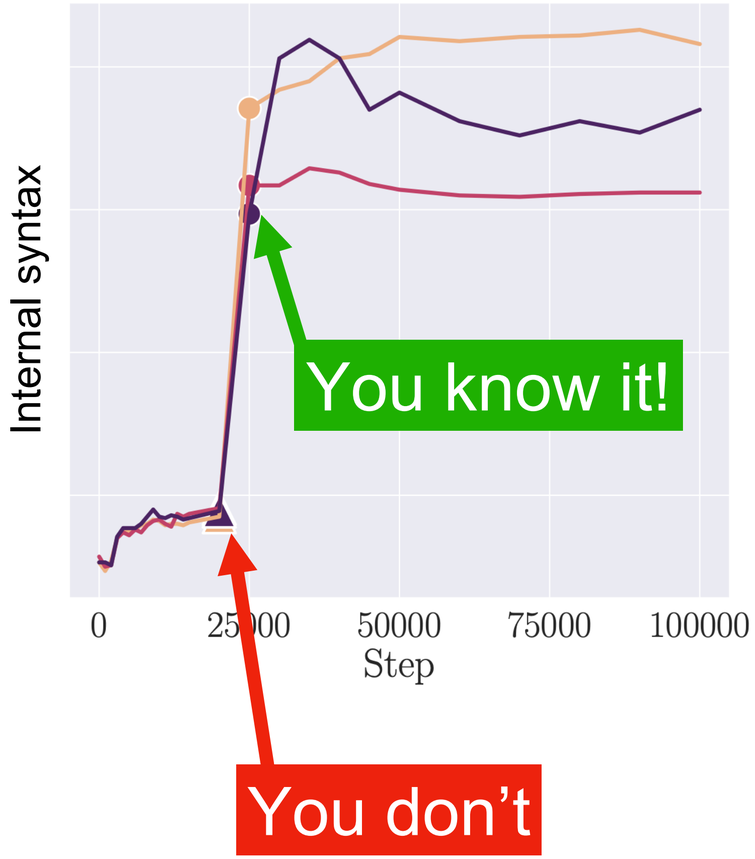
- **Developmental evidence:** Transitioning into an Internal Syntax strategy precipitates grammatical capability.
- **Causal developmental evidence:** Suppressing Internal Syntax damages grammatical capability.



So why didn’t Internal Syntax seem important from the correlation test?

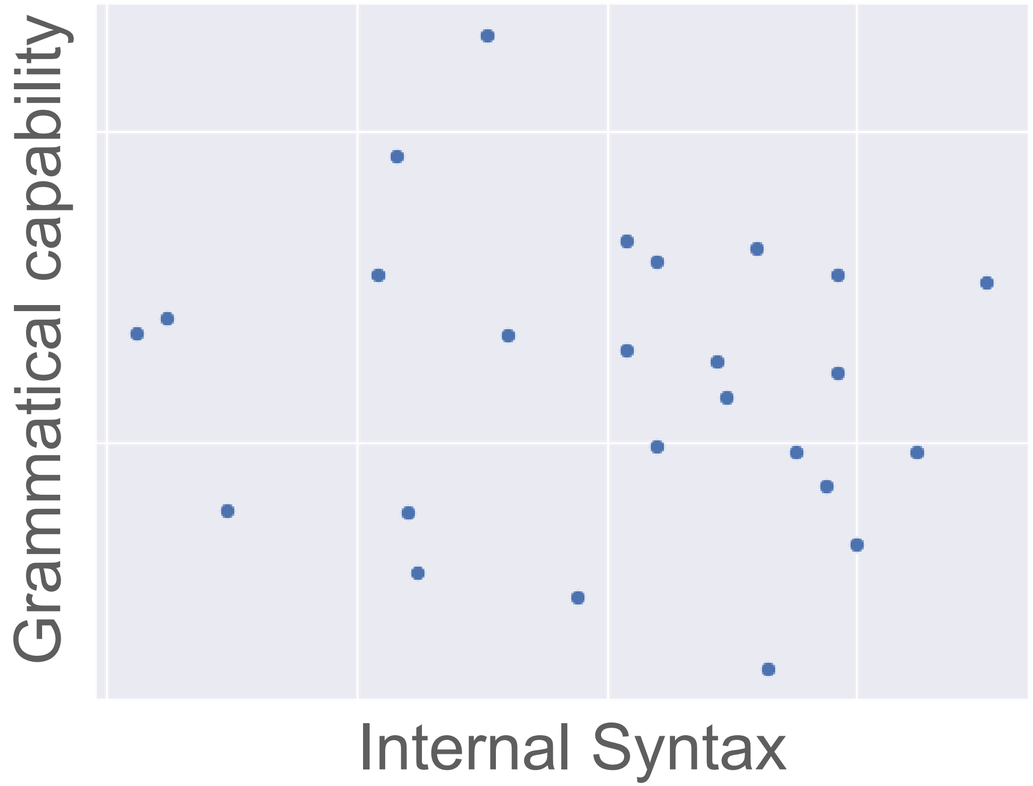


You know it or you don’t



So why didn’t Internal Syntax seem important from the correlation test?

If every trained model “knows it” then variation is just noise!



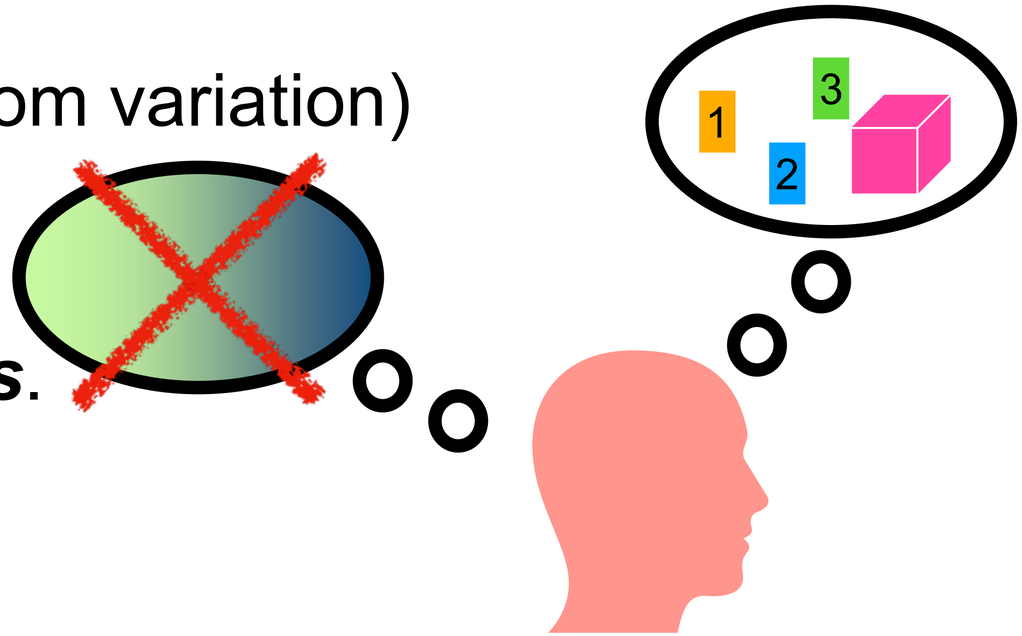
You know it or you don’t

- Scientific and theoretical work often describes a smooth continuum.
  - Continuously learned skills
  - Guaranteed convergence (little random variation)
  - Gradual improvements from scale
- But certain concepts are ***natural kinds***.
  - **You know it or you don’t!**

You know it or you don’t

Great news for human interpretability!

- Scientific and theoretical work often describes a smooth continuum.
  - Continuously learned skills
  - Guaranteed convergence (little random variation)
  - Gradual improvements from scale
- But certain concepts are ***natural kinds***.
  - **You know it or you don’t!**



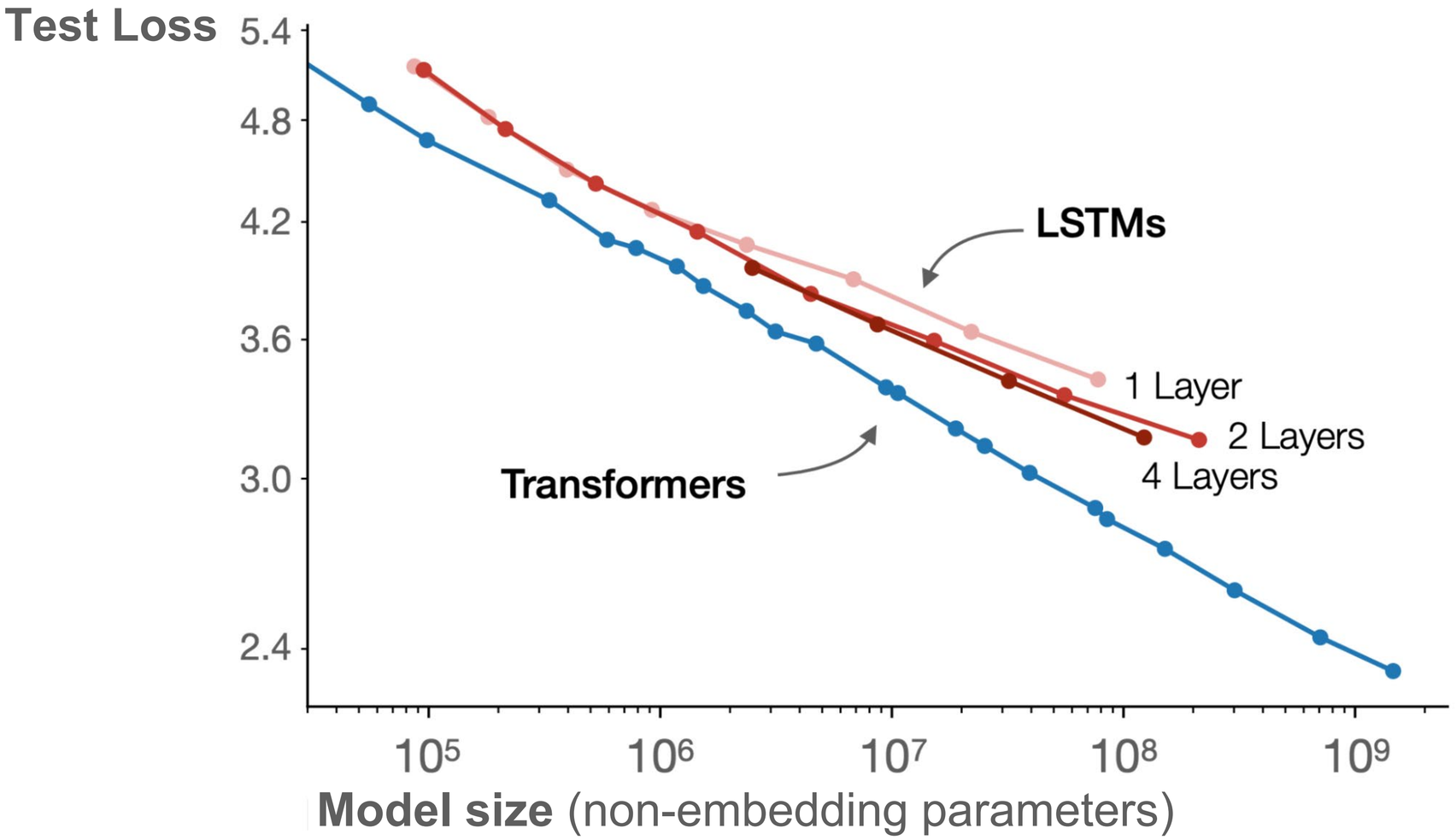
When *does* the random seed impact performance?

Multiple solutions are accessible near *emergent scaling thresholds.*

Zhao*, Qin*, et al. Random scaling of emergent capabilities. *ICML*, 2026.

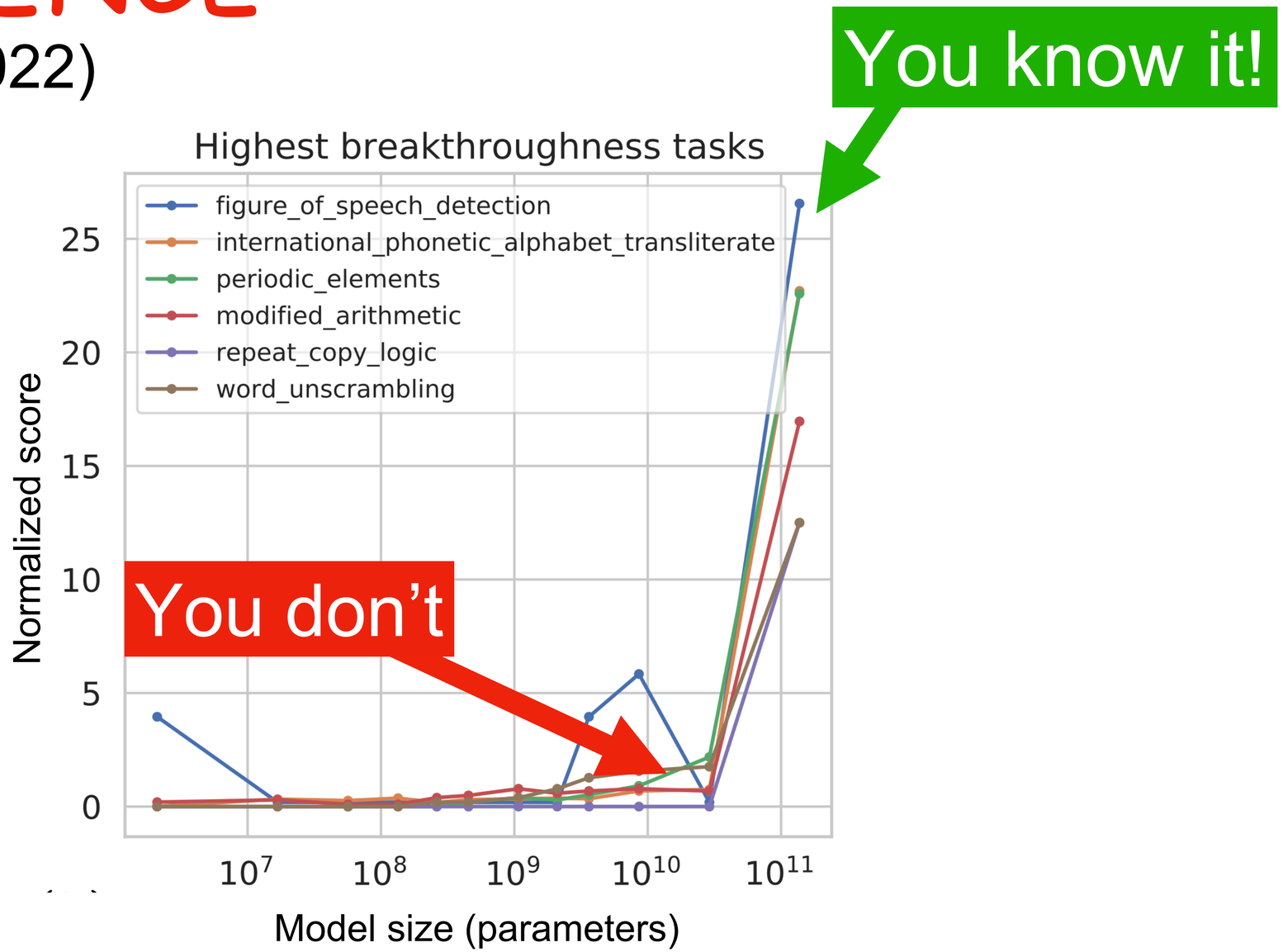
Scaling laws

Smooth power law improvements from adding compute



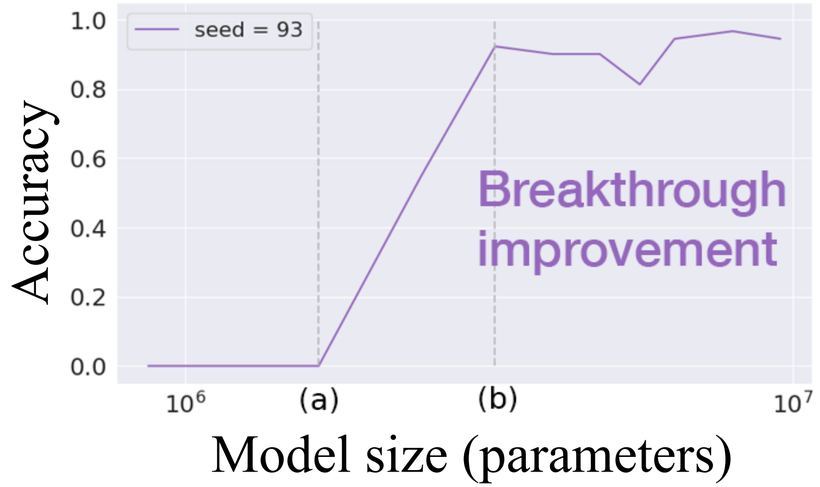
EMERGENCE

Srivastava et al. (2022)



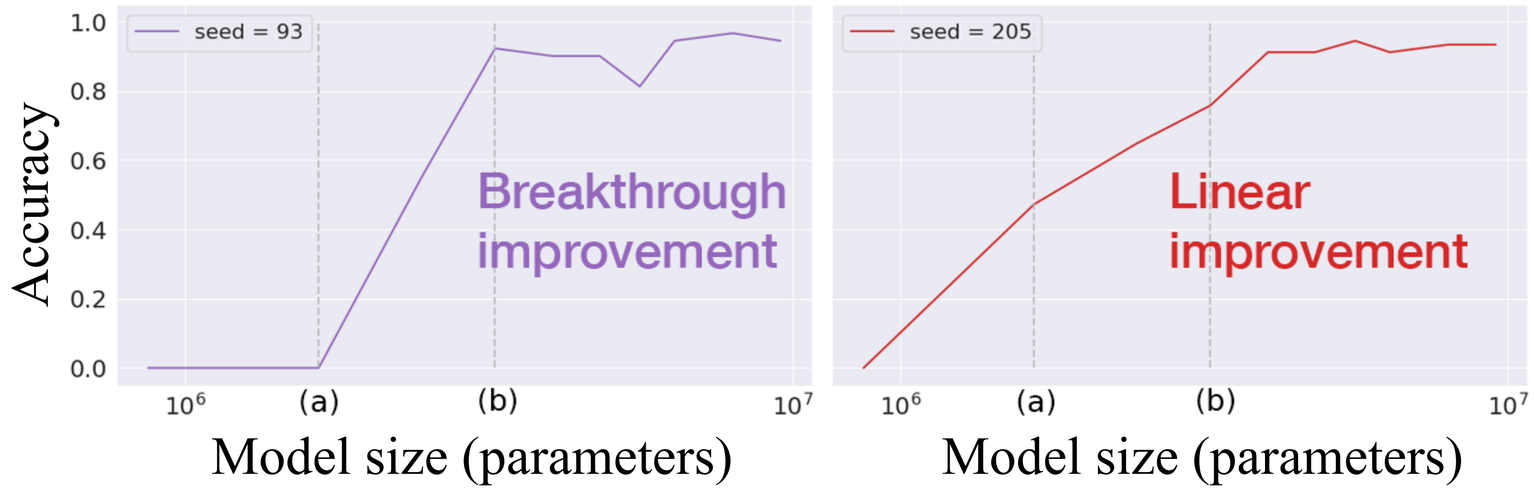
Emerges at appropriate width scale!

Case study: length generalization in reverse addition



Zhao*, Qin*, et al. Random scaling of emergent capabilities. *ICML*, 2026.

Or does it?

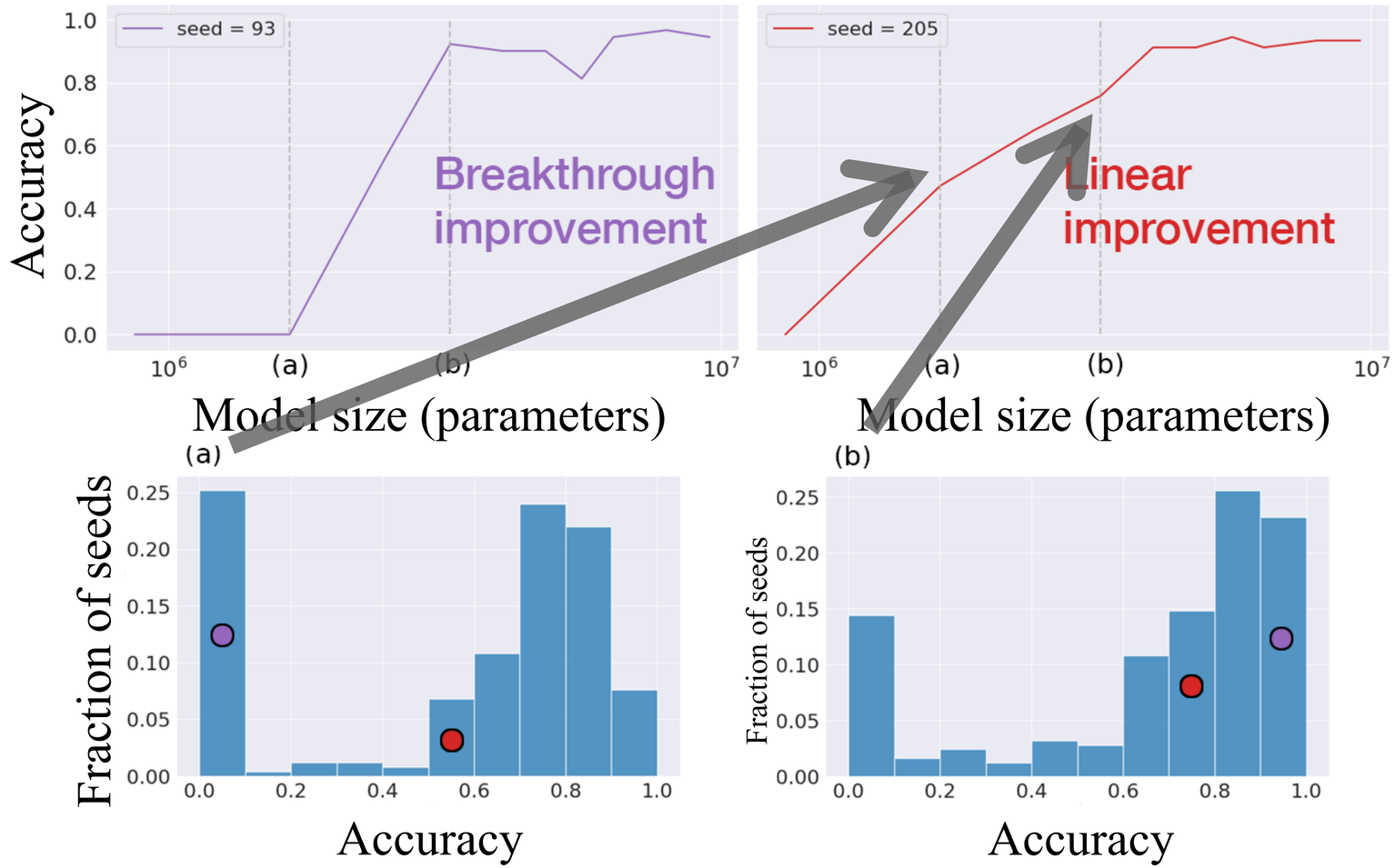


Emergence claims are based on *scalar values* (one seed or average of a few)

Zhao*, Qin*, et al. Random scaling of emergent capabilities. *ICML*, 2026.

Emergence happens when training is “successful”

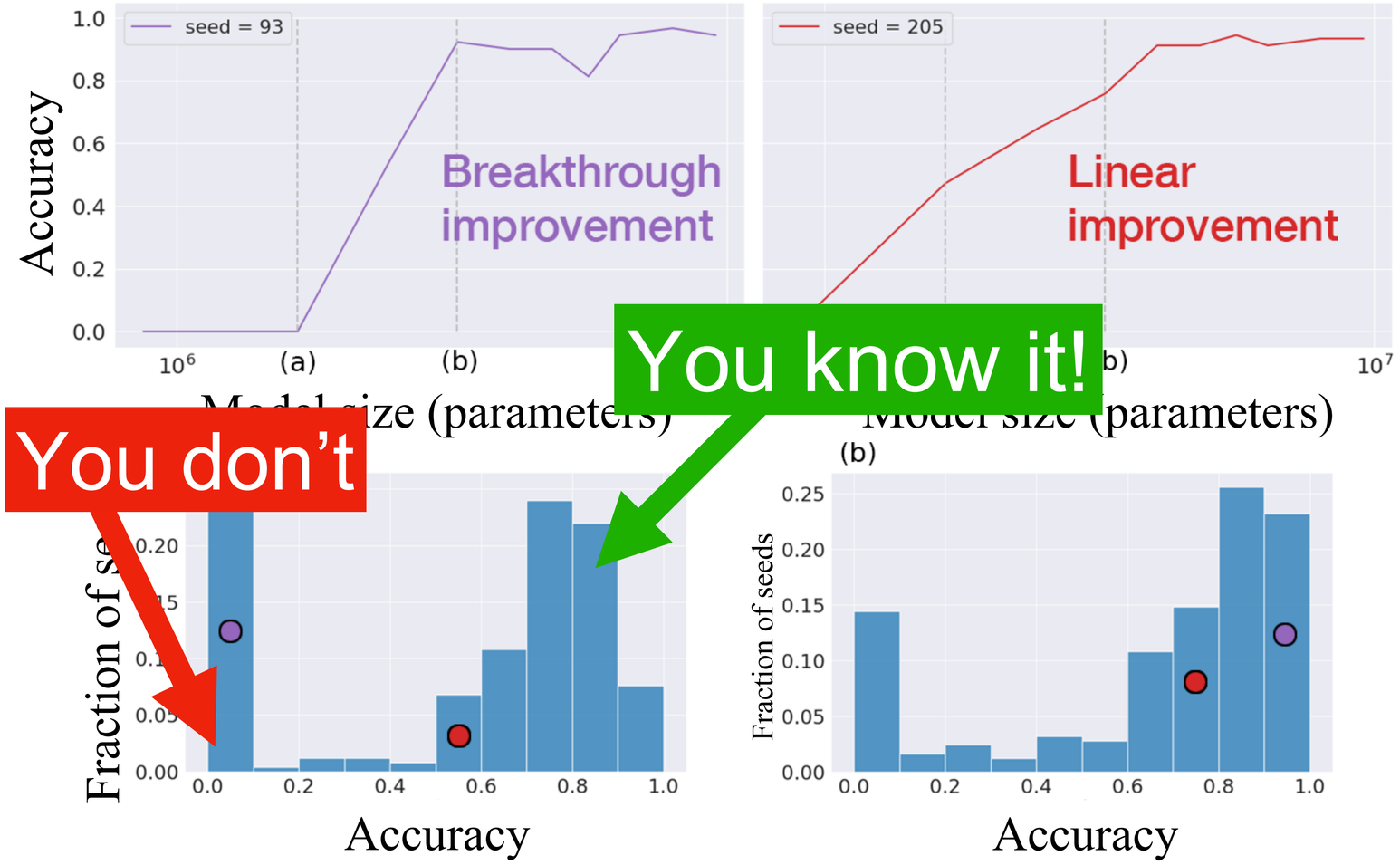
But performance is bimodal!



Zhao*, Qin*, et al. Random scaling of emergent capabilities. *ICML*, 2026.

Emergence happens when training is “successful”

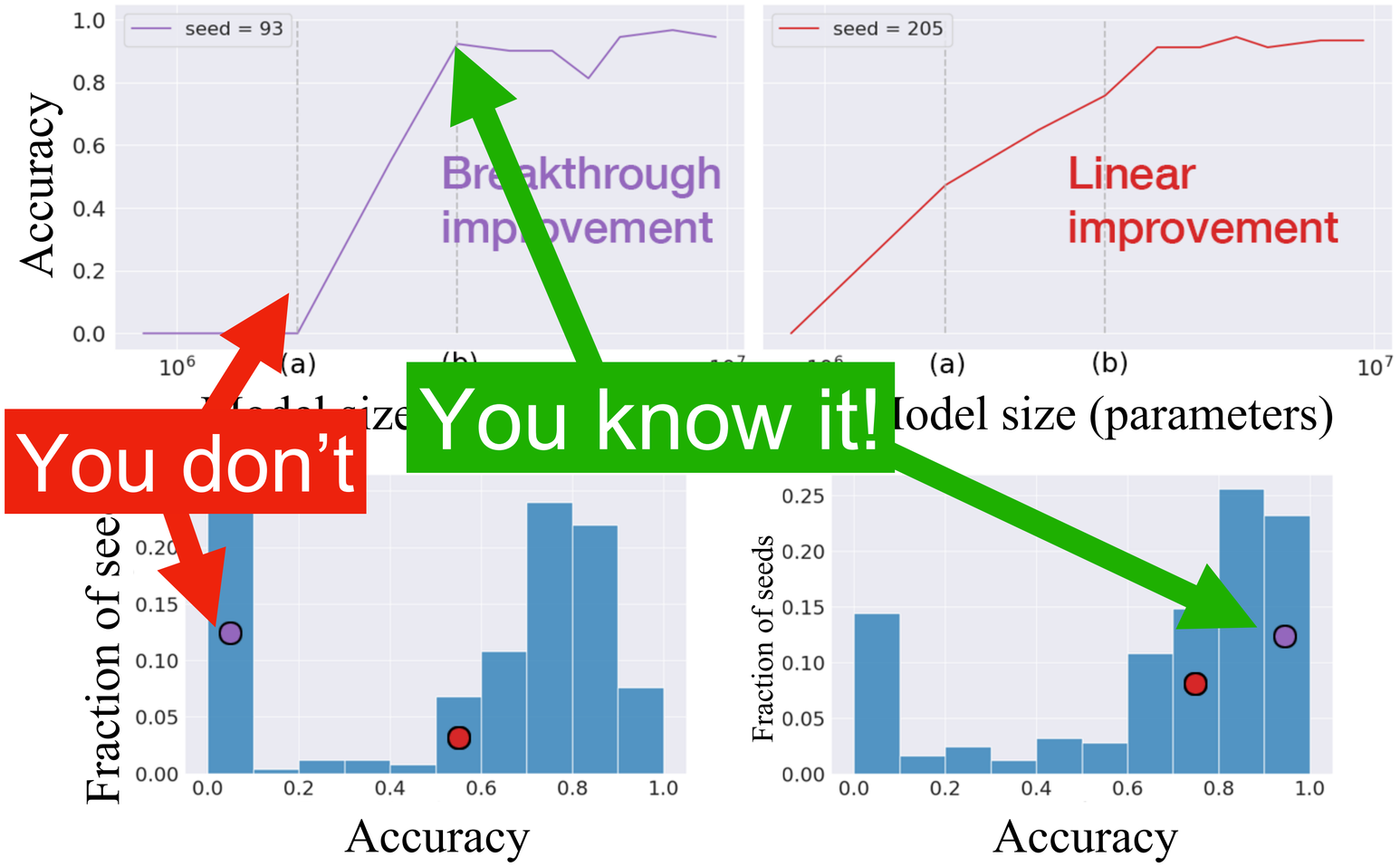
But performance is bimodal!



Zhao*, Qin*, et al. Random scaling of emergent capabilities. *ICML*, 2026.

Emergence happens when training is “successful”

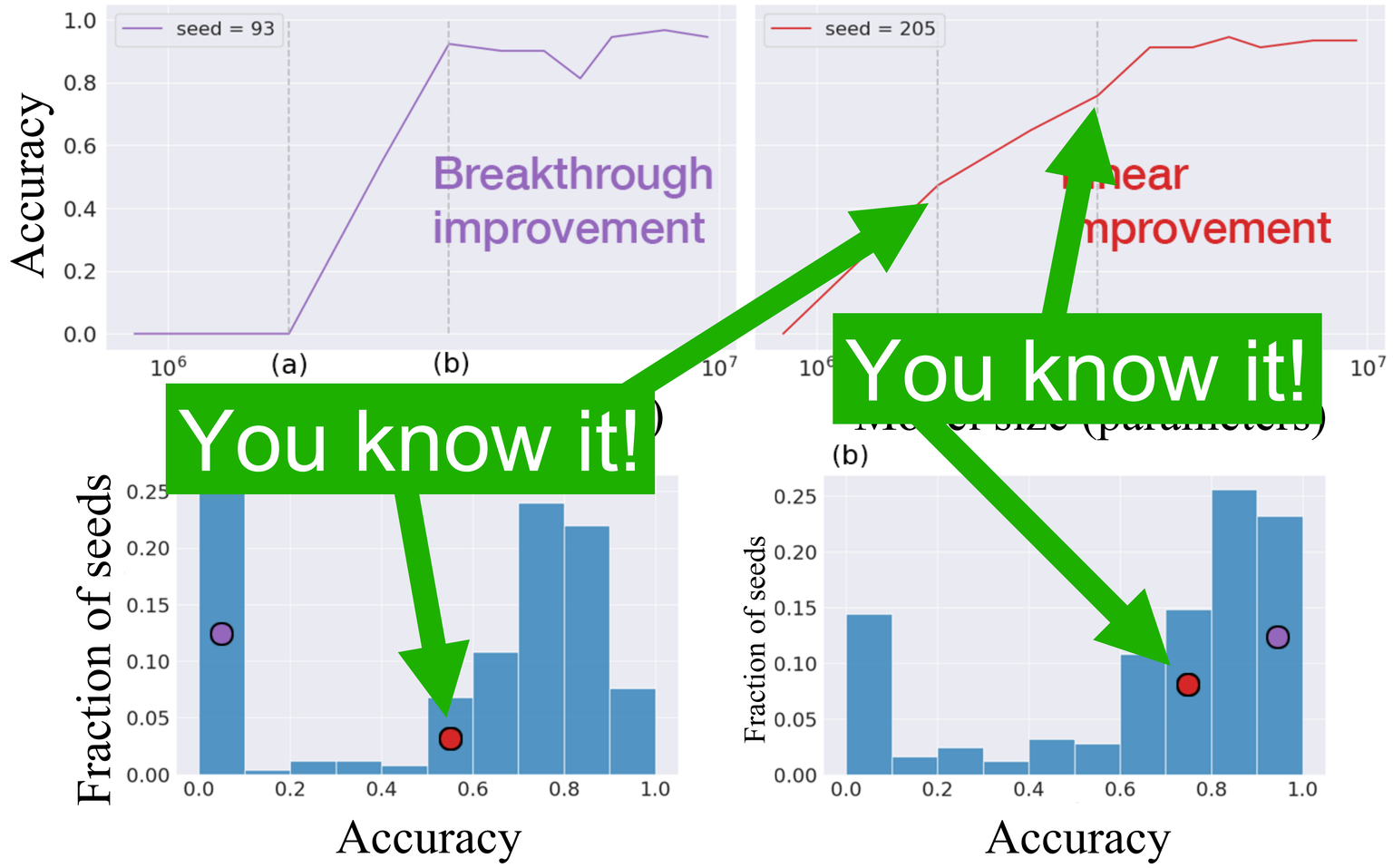
Smaller models might fail …



Zhao*, Qin*, et al. Random scaling of emergent capabilities. *ICML*, 2026.

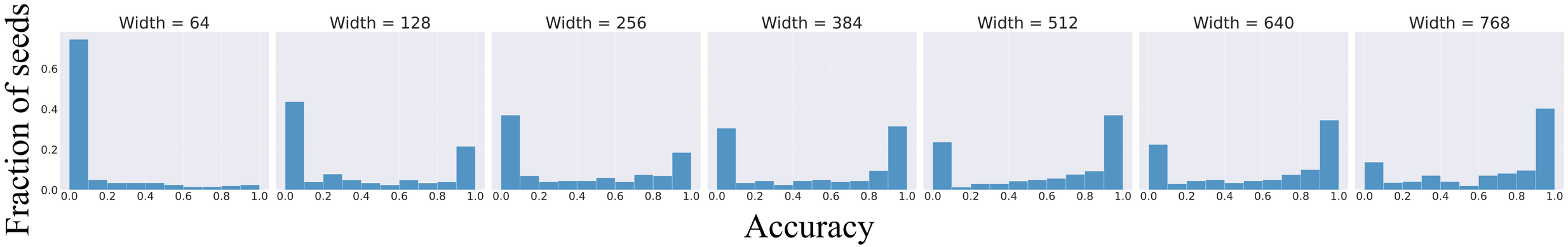
Emergence happens when training is “successful”

… But if they don’t, no emergence!



Zhao*, Qin*, et al. Random scaling of emergent capabilities. *ICML*, 2026.

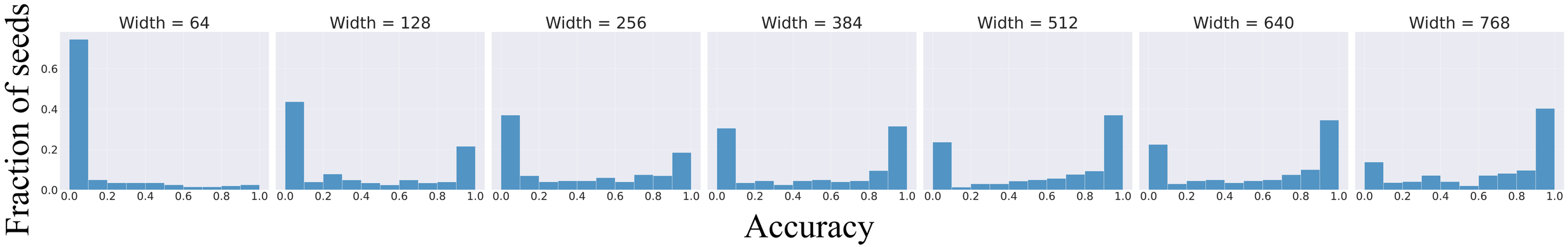
Bimodal distributions change gradually



Zhao*, Qin*, et al. Random scaling of emergent capabilities. *ICML*, 2026.

So why are some scale trends discontinuous?

Gradual changes in the *probability of success!*



Zhao*, Qin*, et al. Random scaling of emergent capabilities. *ICML*, 2026.

(No, it’s not a metric thresholding effect!)



Zhao*, Qin*, et al. Random scaling of emergent capabilities. *ICML*, 2026.

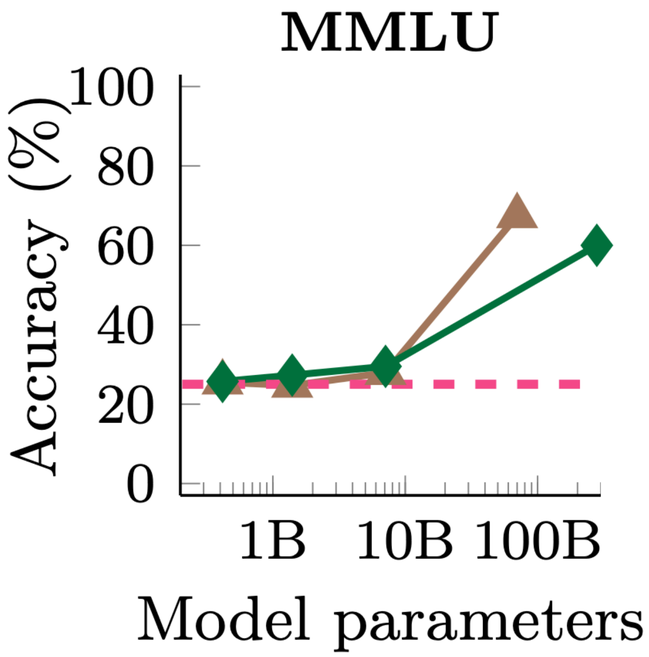
Real world example: multiple choice question answering

(MMLU dataset)

Professional Medicine

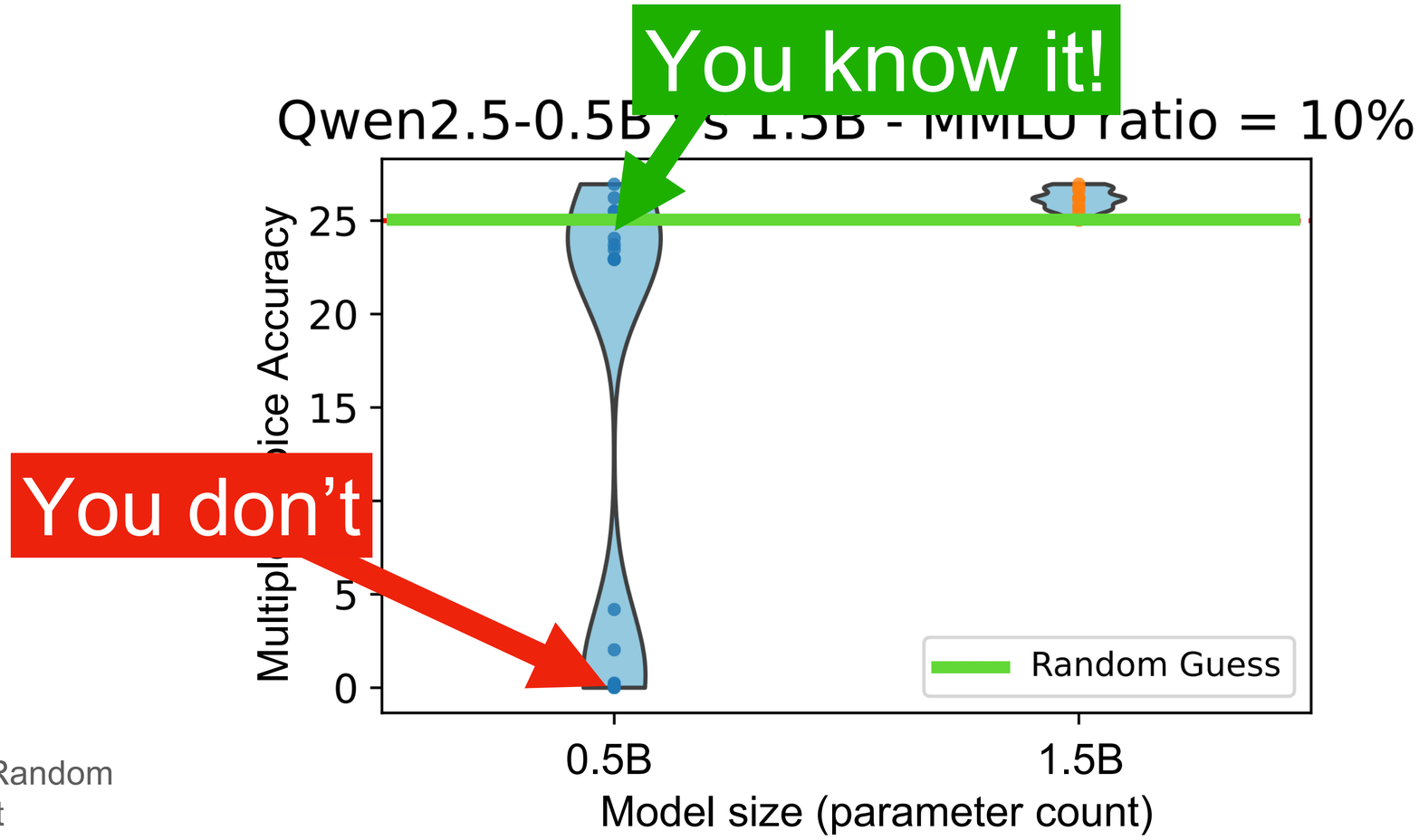
A 33-year-old man undergoes a radical thyroidectomy for thyroid cancer. During the operation, moderate hemorrhaging requires ligation of several vessels in the left side of the neck. Postoperatively, serum studies show a calcium concentration of 7.5 mg/dL, albumin concentration of 4 g/dL, and parathyroid hormone concentration of 200 pg/mL. Damage to which of the following vessels caused the findings in this patient?

(A) Branch of the costocervical trunk ✗
(B) Branch of the external carotid artery ✗
(C) Branch of the thyrocervical trunk ✓
(D) Tributary of the internal jugular vein ✗



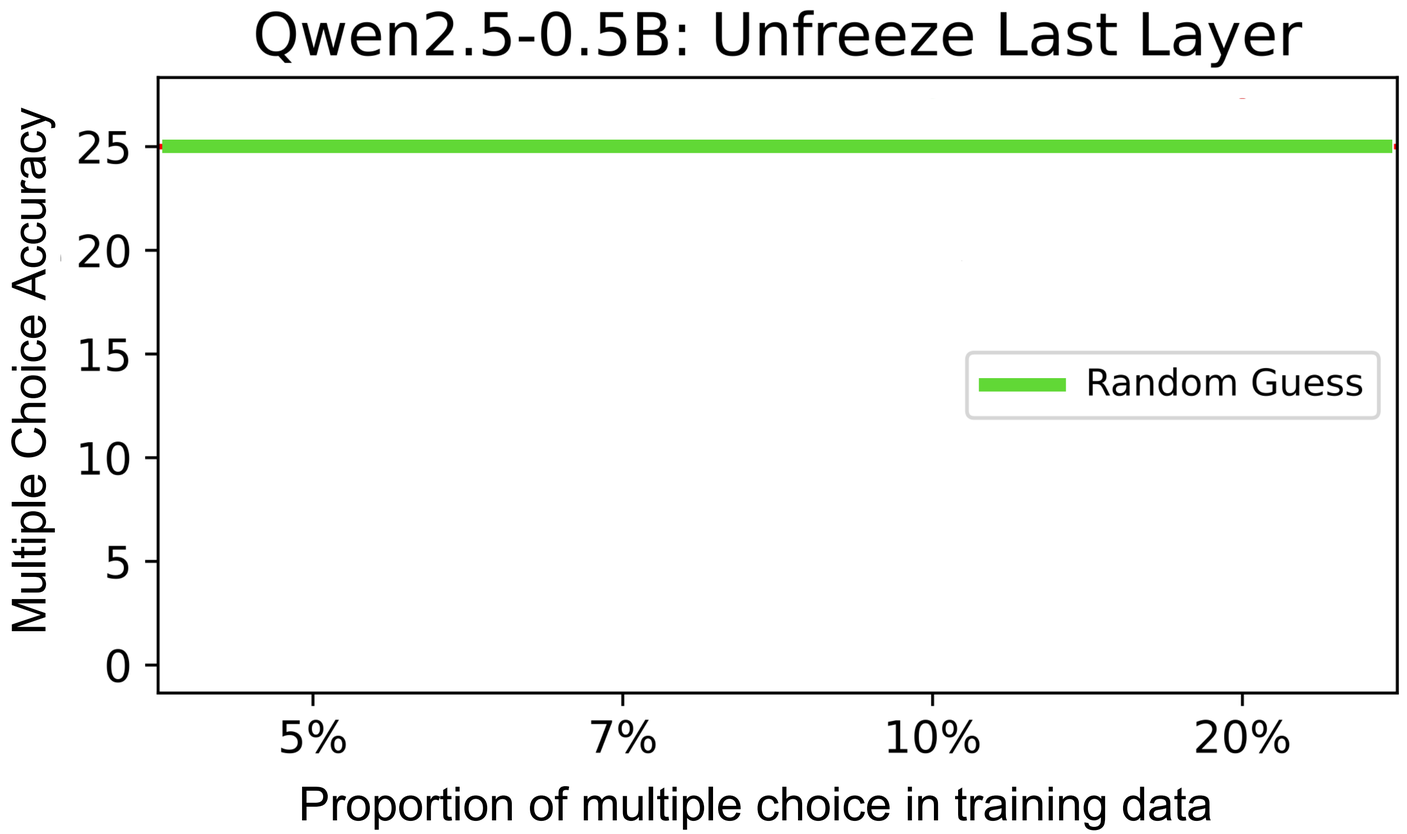
Wei et al., 2022

Mid size *models* yield bimodal distributions



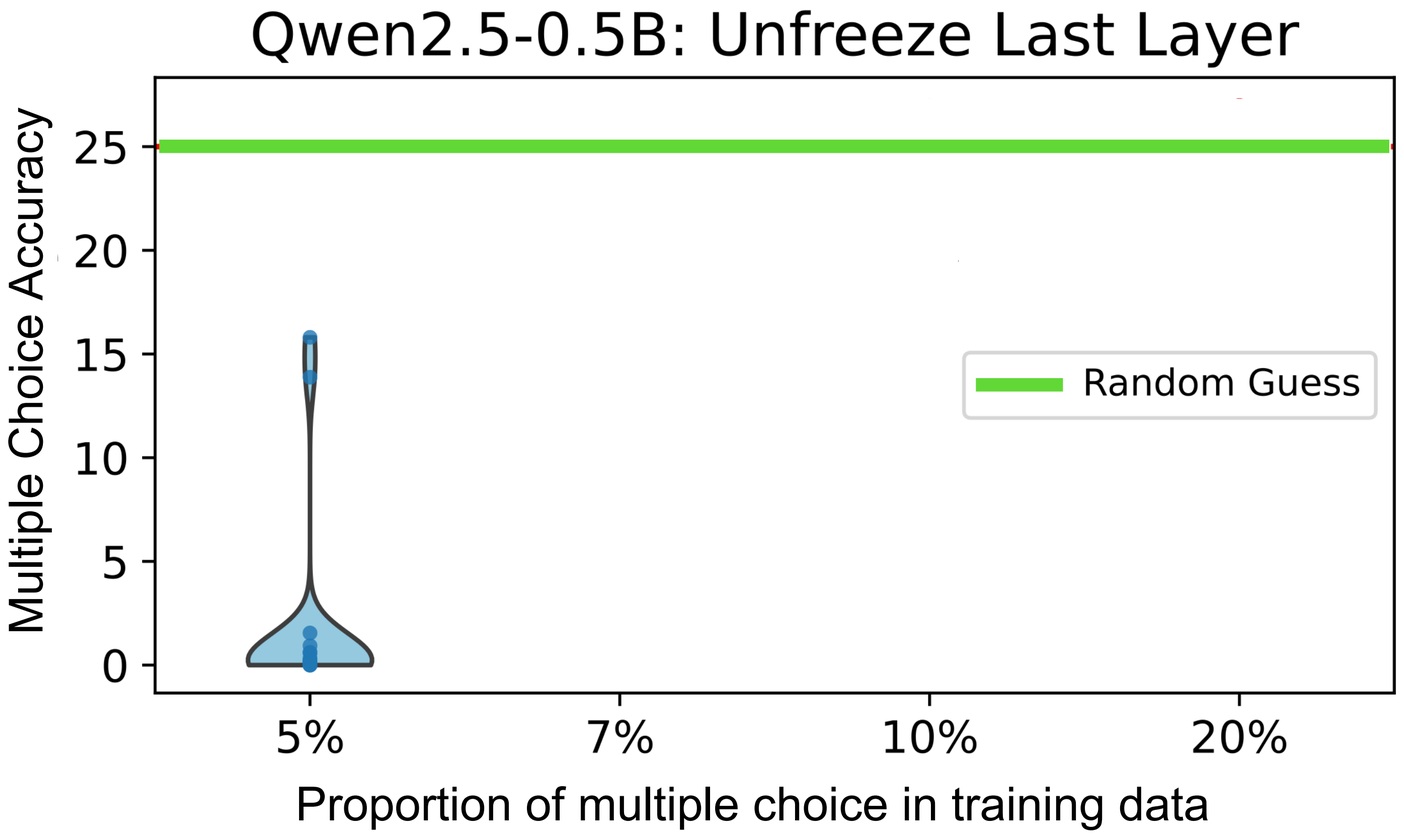
Zhao*, Qin*, et al. Random scaling of emergent capabilities. *ICML*, 2026.

Mid size *data* yields bimodal distributions



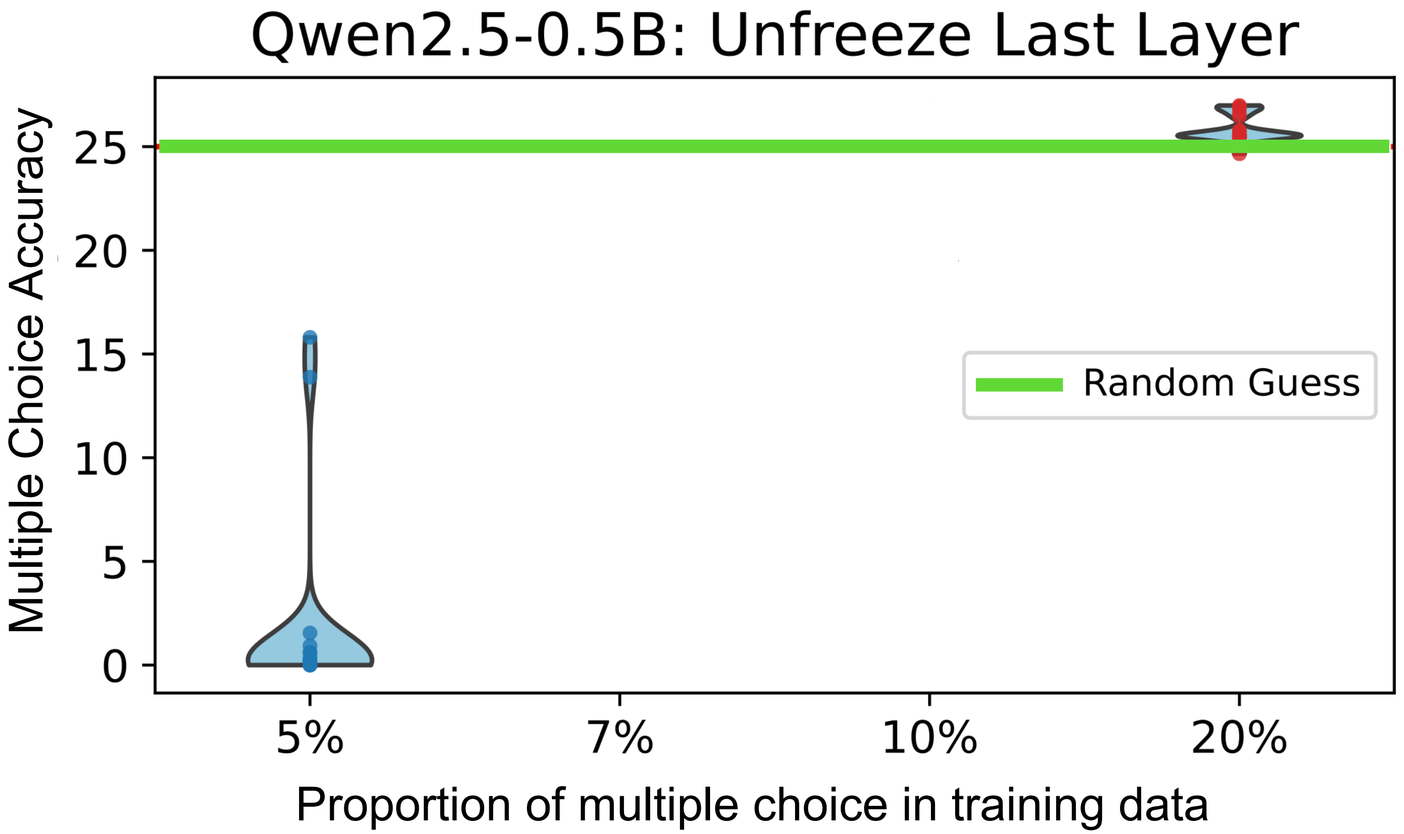
Zhao*, Qin*, et al. Random scaling of emergent capabilities. *ICML*, 2026.

Mid size *data* yields bimodal distributions



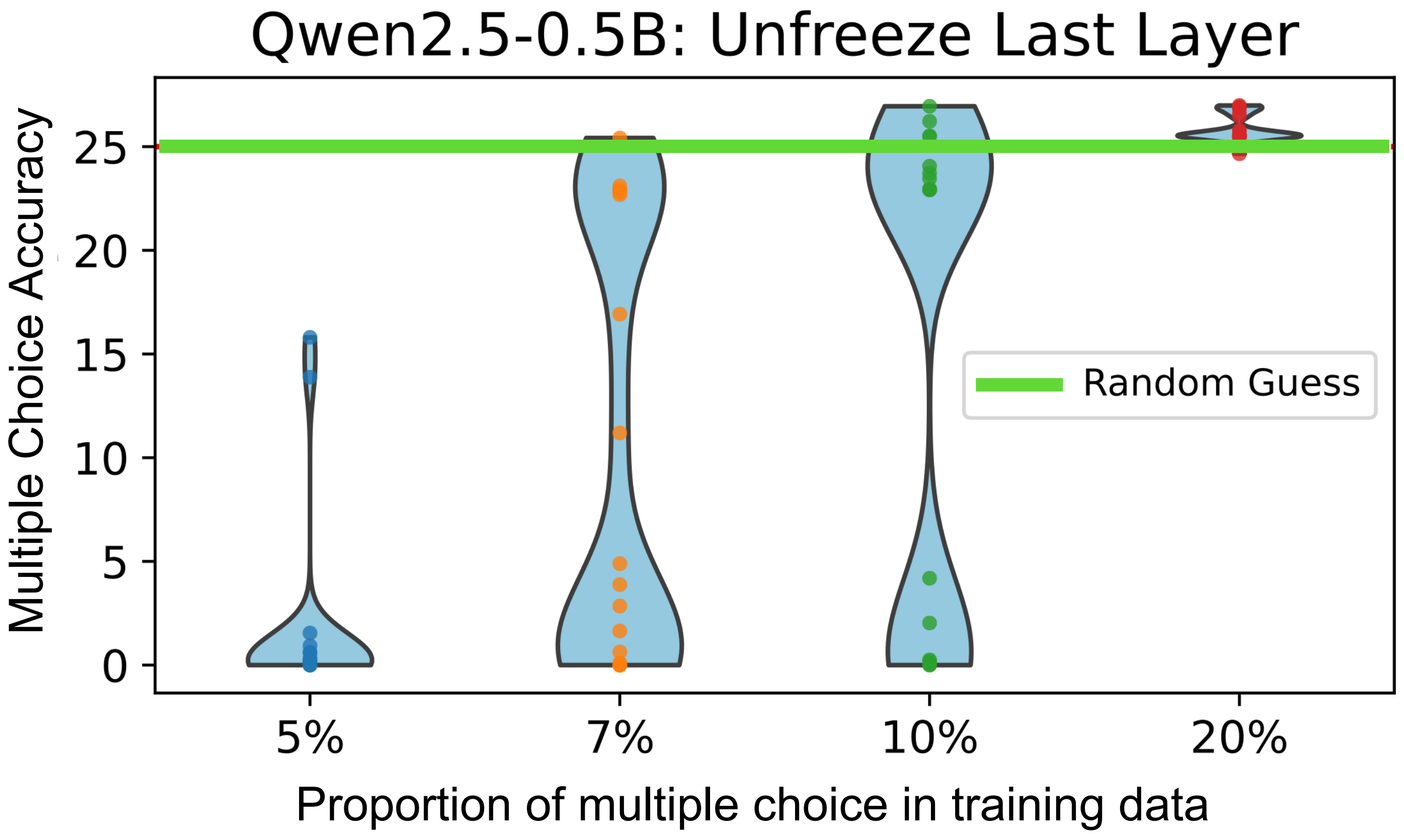
Zhao*, Qin*, et al. Random scaling of emergent capabilities. *ICML*, 2026.

Mid size *data* yields bimodal distributions



Zhao*, Qin*, et al. Random scaling of emergent capabilities. *ICML*, 2026.

Mid size *data* yields bimodal distributions



Zhao*, Qin*, et al. Random scaling of emergent capabilities. *ICML*, 2026.

Why care about discrete model behaviors?

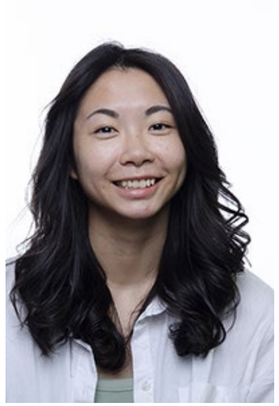
- This is what allows complex and compositional generalization!
- This is what makes model training outcomes unpredictable and contingent!
- This is what can *turn a blackbox predictor back into human intuition.*



















Questions?