You Know It Or You Don’t: Compositionality and Phase Transitions in LMs

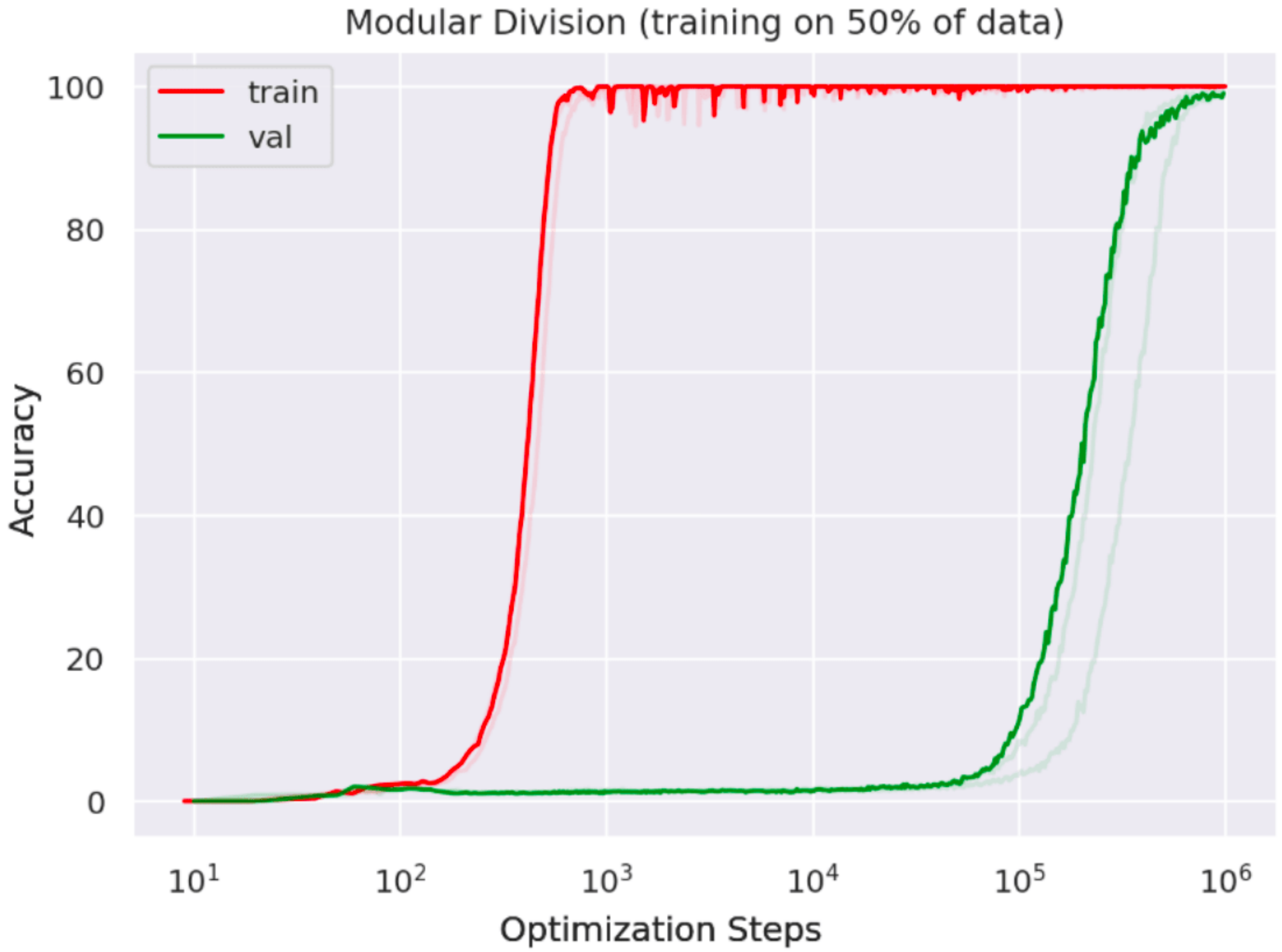
Naomi Saphra at Simons Institute, 2025

Prologue:
Unpredictable breakthroughs

Breakthroughs in training

Grokking (Power et al.)

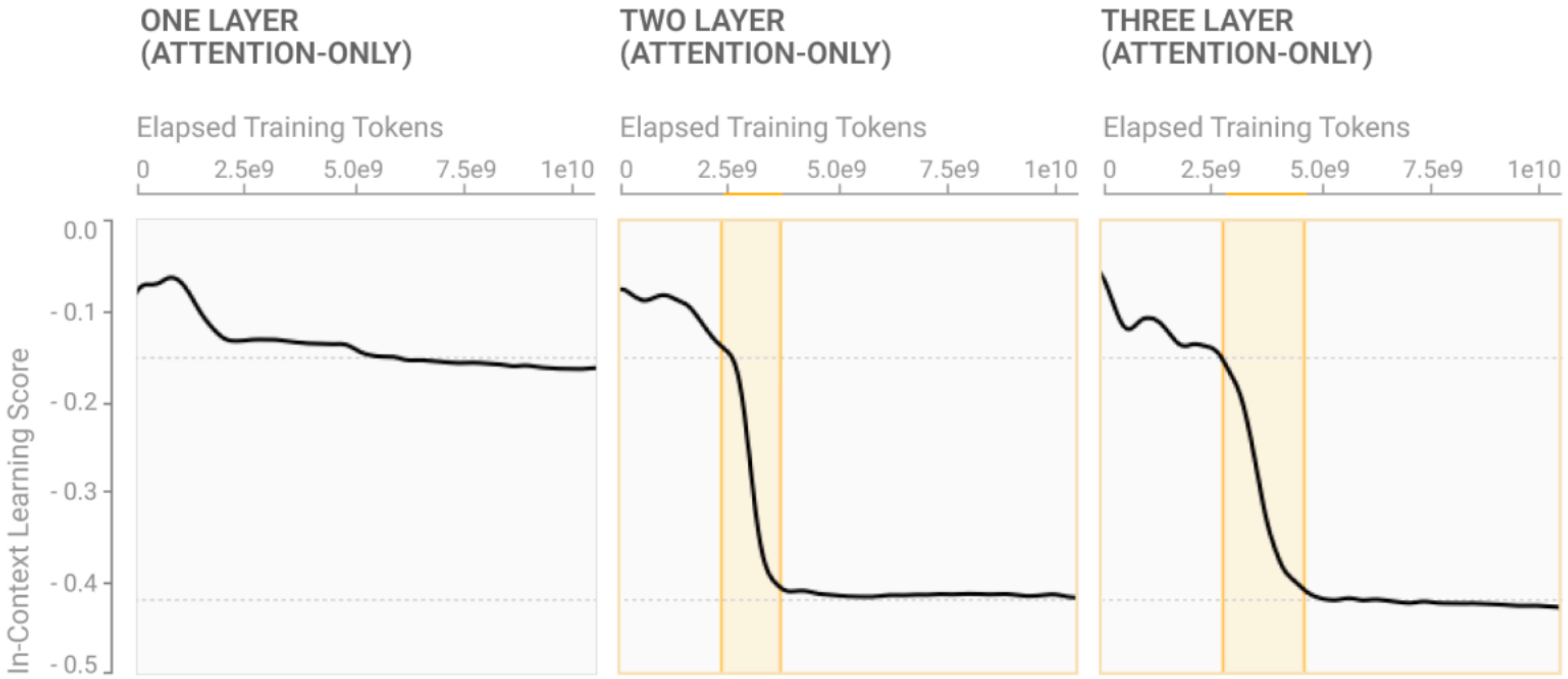
- Low data setting
- First, memorize
- Then generalize (same distribution)



Breakthroughs in training

Induction Heads (Olsson et al.)

- Multilayer models form a circuit with two steps
  - First search for previous occurrence
  - Then copy next token
- Think: priming effects
- Used for in-context learning

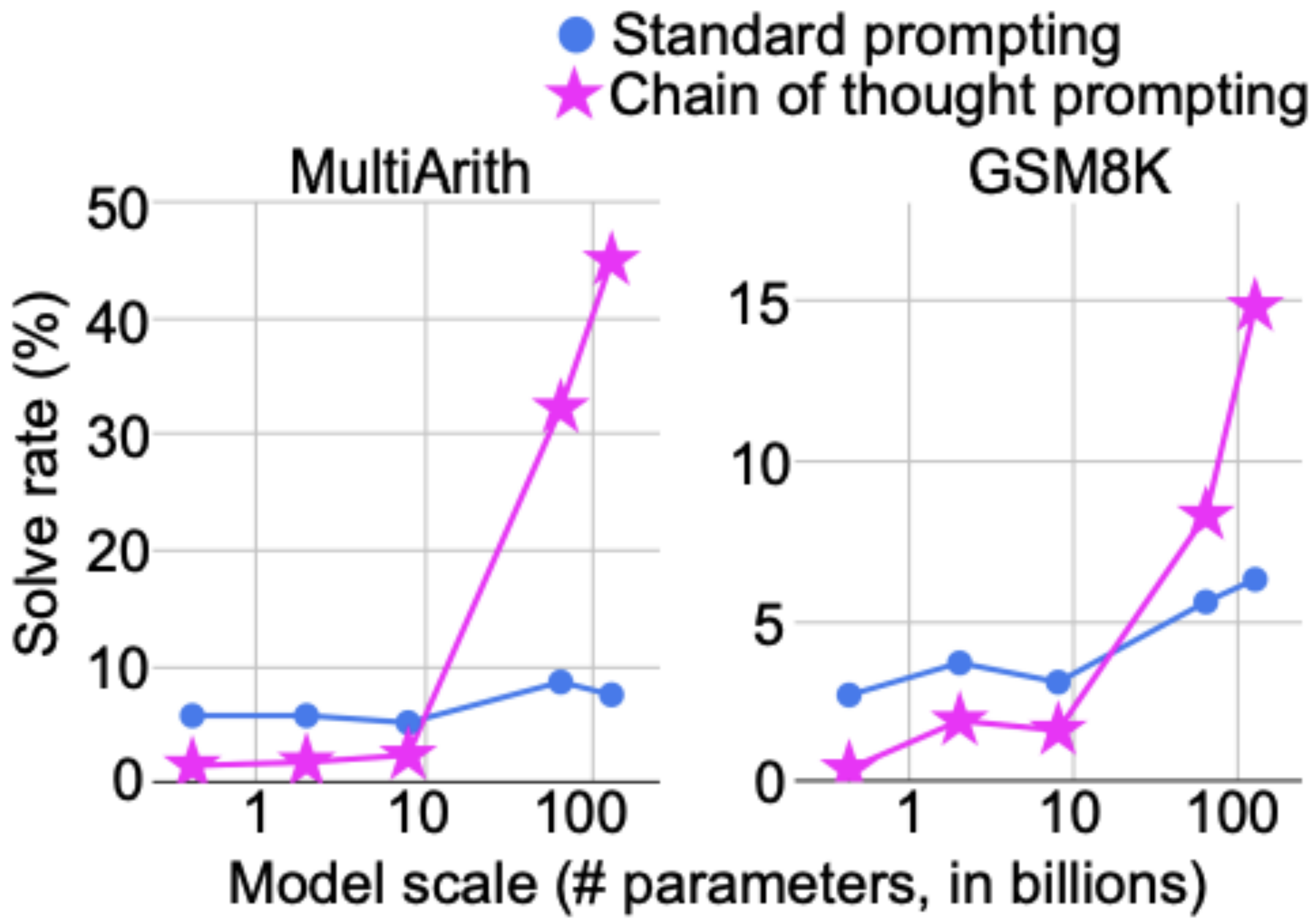


Olsson et al., 2022

Breakthroughs in scale

“Emergence” or “Breakthrough” (Srivastava et al.)

- Compositional, usually
  - Classic example: Multiple choice QA
  - But not open-ended / cloze QA!
- Maybe just thresholding artifacts that disappear under continuous metrics?
  - But not always!



What makes a capability *breakthrough*?

What makes a capability *breakthrough*?

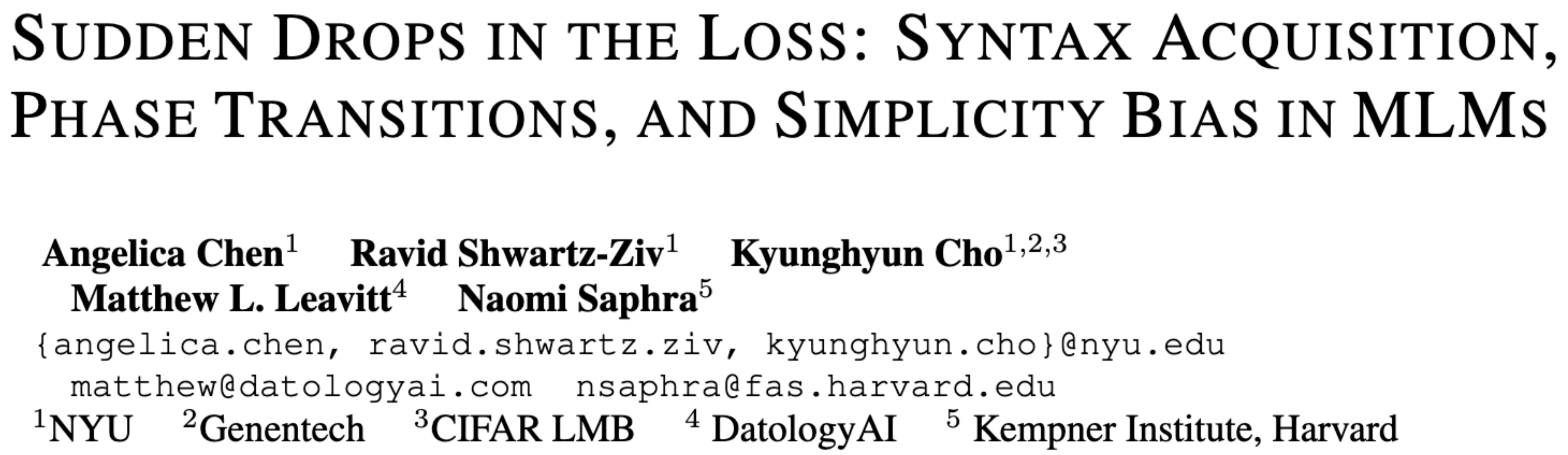
- Compositional structure
- Competition between possible solutions
- Multimodality (across random seeds or subtle changes)

What makes a capability *breakthrough*?

(Bonus question: Are these … all the same thing?)

- Compositional structure
- Competition between possible solutions
- Multimodality (across random seeds or subtle changes)

Case study 1: Sudden syntax acquisition

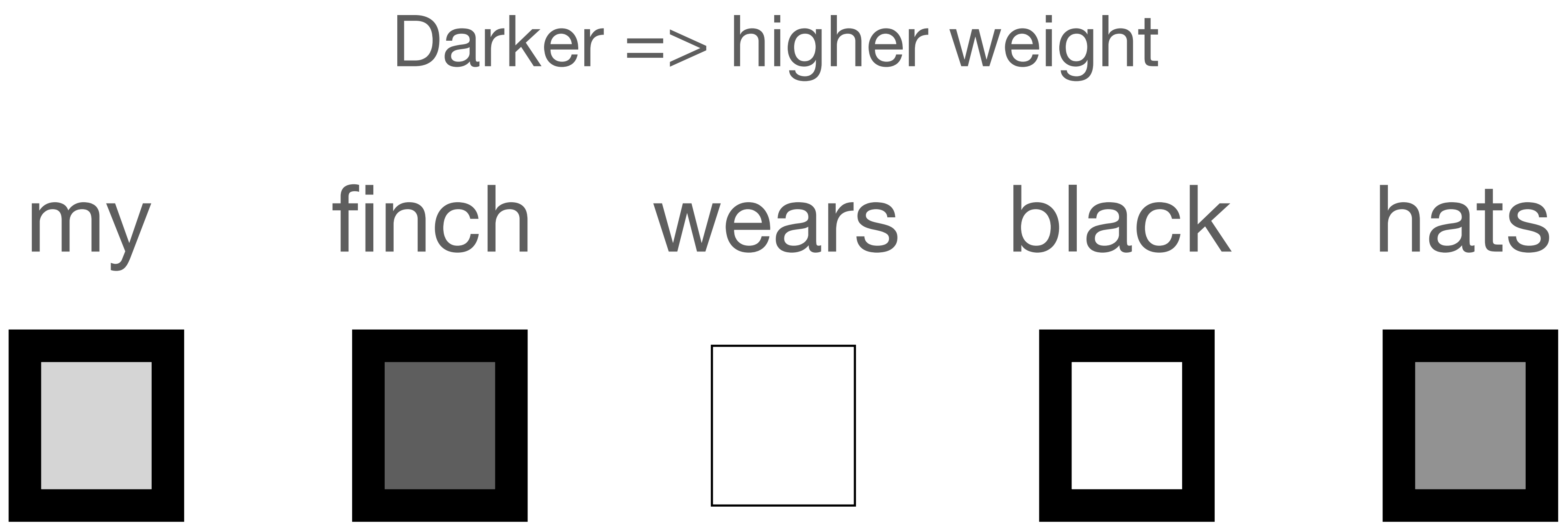


SUDDEN DROPS IN THE LOSS: SYNTAX ACQUISITION, PHASE TRANSITIONS, AND SIMPLICITY BIAS IN MLMS

Angelica Chen[1] **Ravid Shwartz-Ziv**[1] **Kyunghyun Cho**[1,2,3]
Matthew L. Leavitt[4] **Naomi Saphra**[5]
{angelica.chen, ravid.shwartz.ziv, kyunghyun.cho}@nyu.edu
matthew@datologyai.com nsaphra@fas.harvard.edu
[1]NYU [2]Genentech [3]CIFAR LMB [4] DatologyAI [5] Kempner Institute, Harvard

Masked Language Modeling (MLM) with BERT

- The task: predict a masked out (missing) word from a sequence.
- Used to build a pretrained model which can be finetuned for other tasks.
- BERT: made up of Transformer heads, which compute an attention distribution to reweight the representation of each word in a sequence.

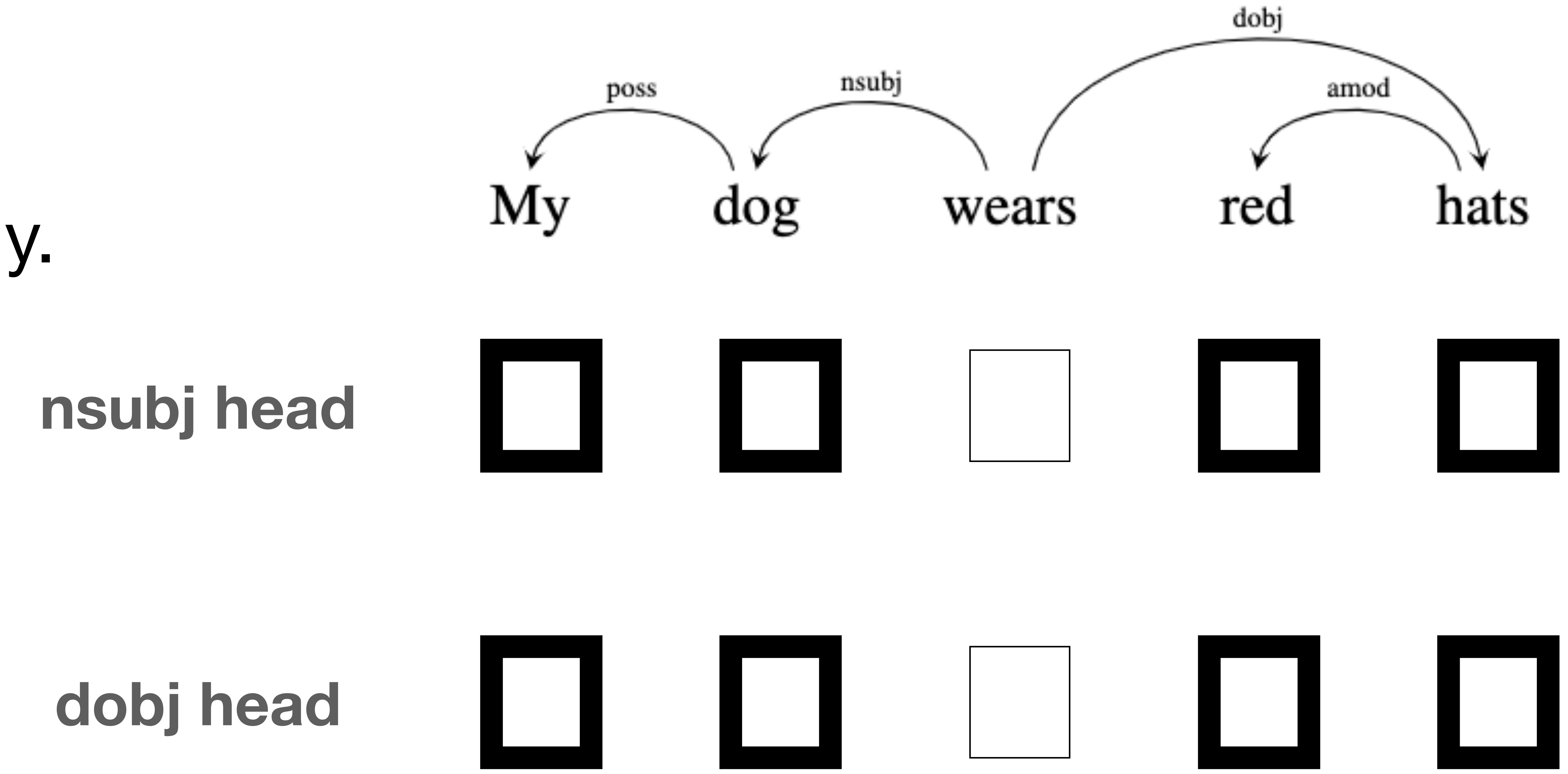


Target: “wears”

Syntactic Attention Structure (SAS)

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

- Given a syntactic relation, some BERT head attends to that relation consistently.

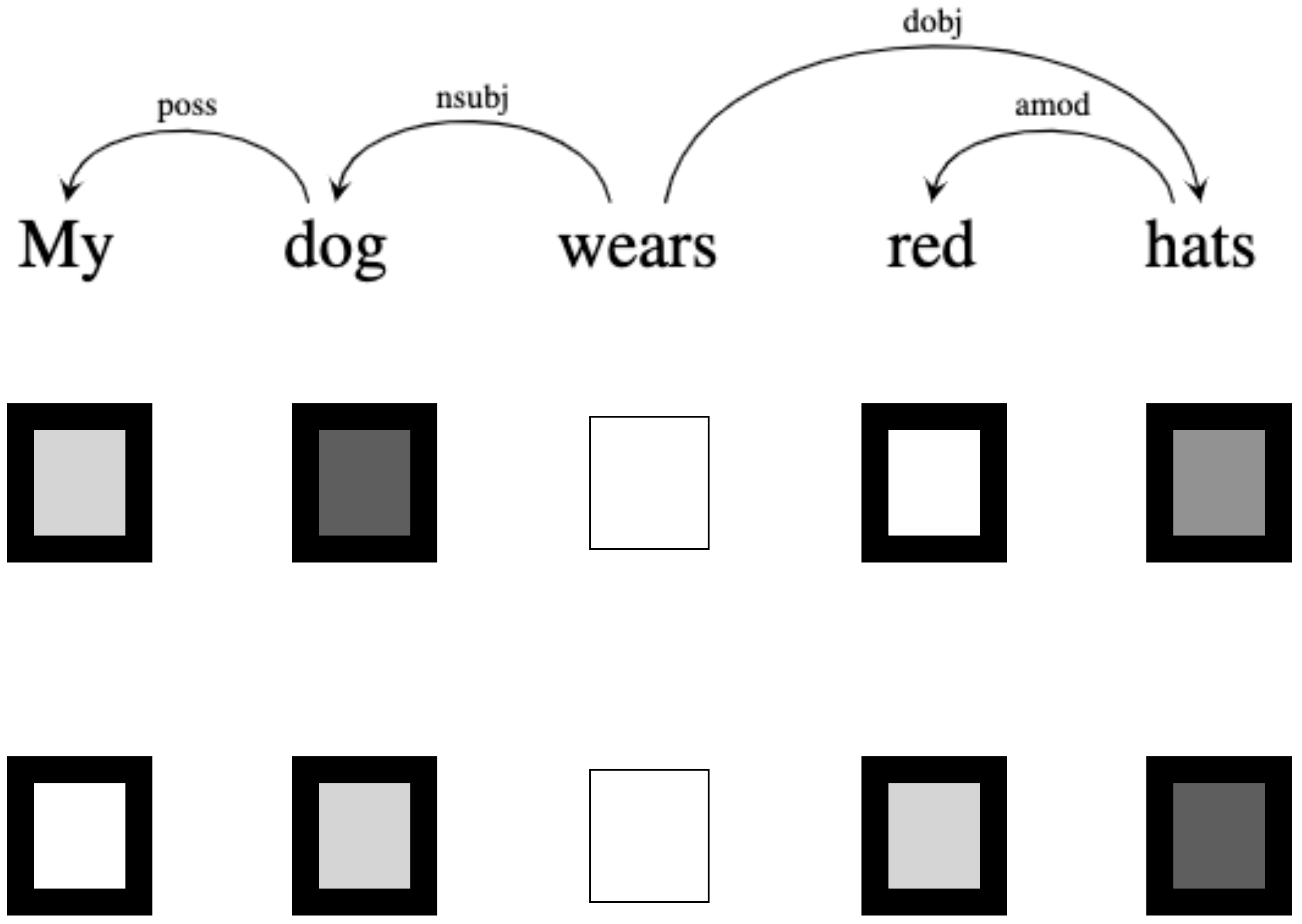


Target: “wears”

Syntactic Attention Structure (SAS)

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

- Given a syntactic relation, some BERT head attends to that relation consistently.



nsubj head

dobj head

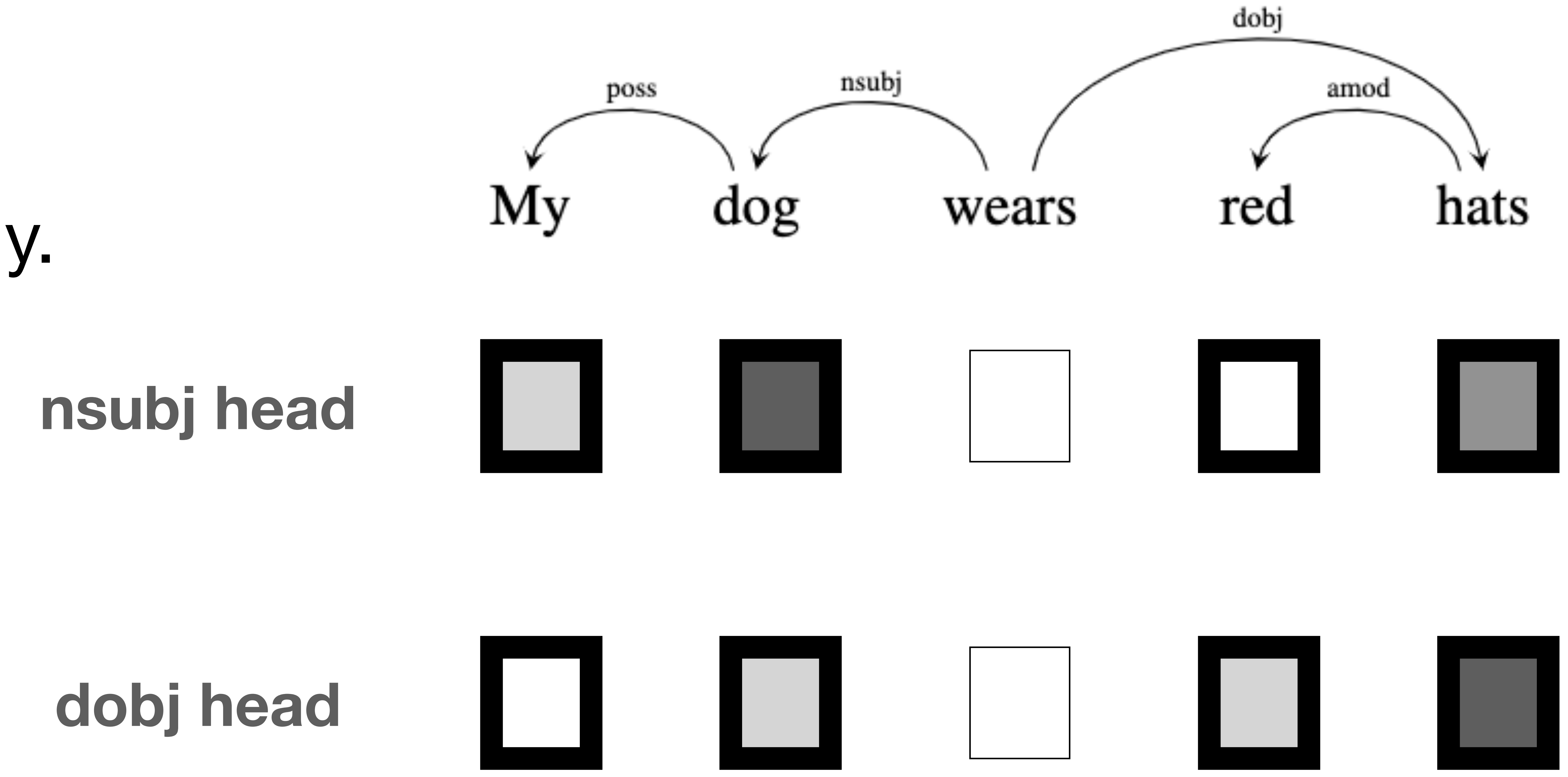
Target: “wears”

Syntactic Attention Structure (SAS)

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

- Given a syntactic relation, some BERT head attends to that relation consistently.

- Naturally emerging property in masked language models!

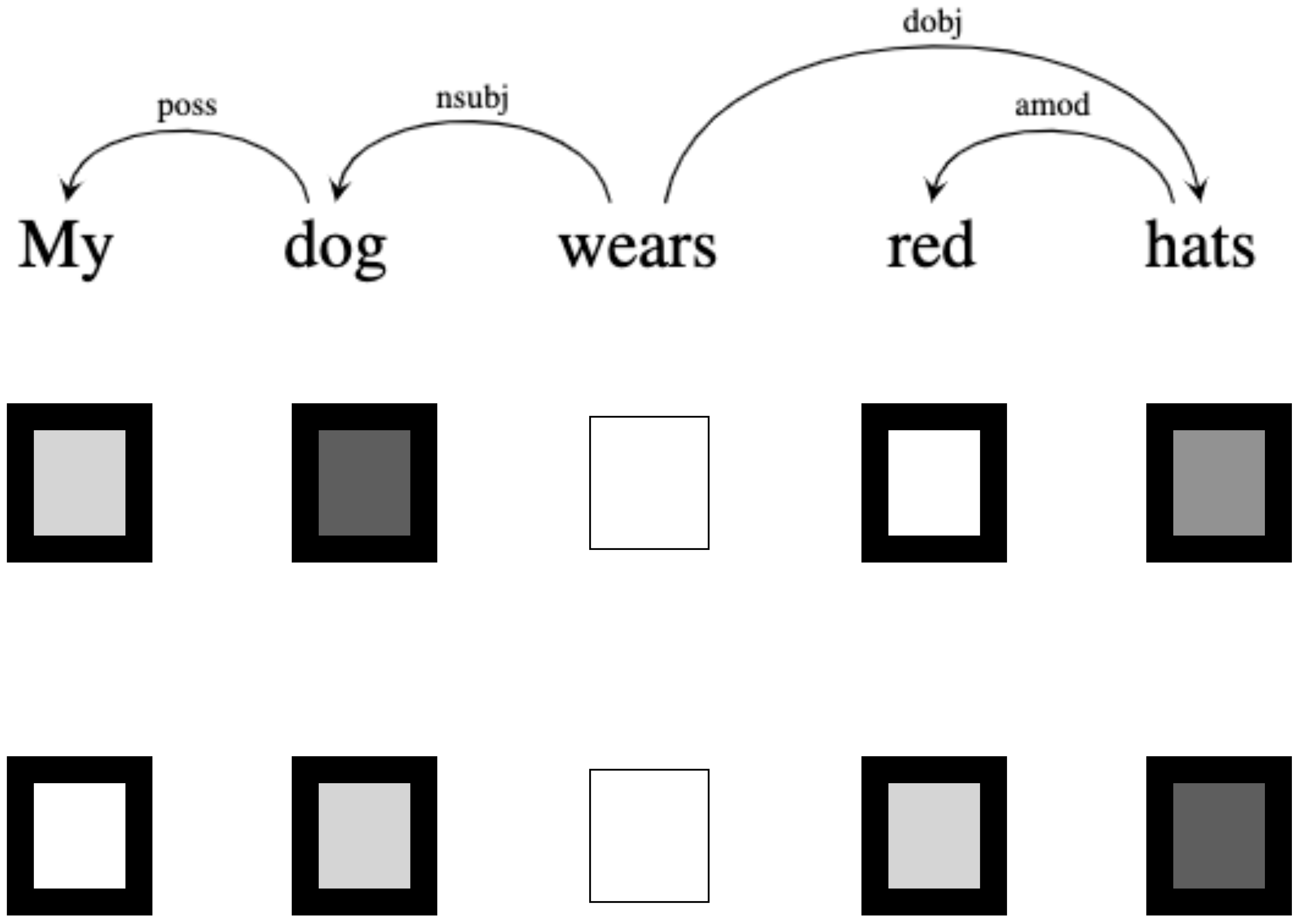


Target: “wears”

Syntactic Attention Structure (SAS)

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

- Given a syntactic relation, some BERT head attends to that relation consistently.
- Naturally emerging property in masked language models!
- Measured with **Unlabeled Attachment Score (UAS)**.



nsubj head

dobj head

Target: “wears”

We know MLMs have specialized syntactic heads

But are they important for grammatical understanding? Evidence:

- **Instance level observations**
  - Specialized syntactic heads predict dependencies with high accuracy.
  - (Clark et al., 2019)

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

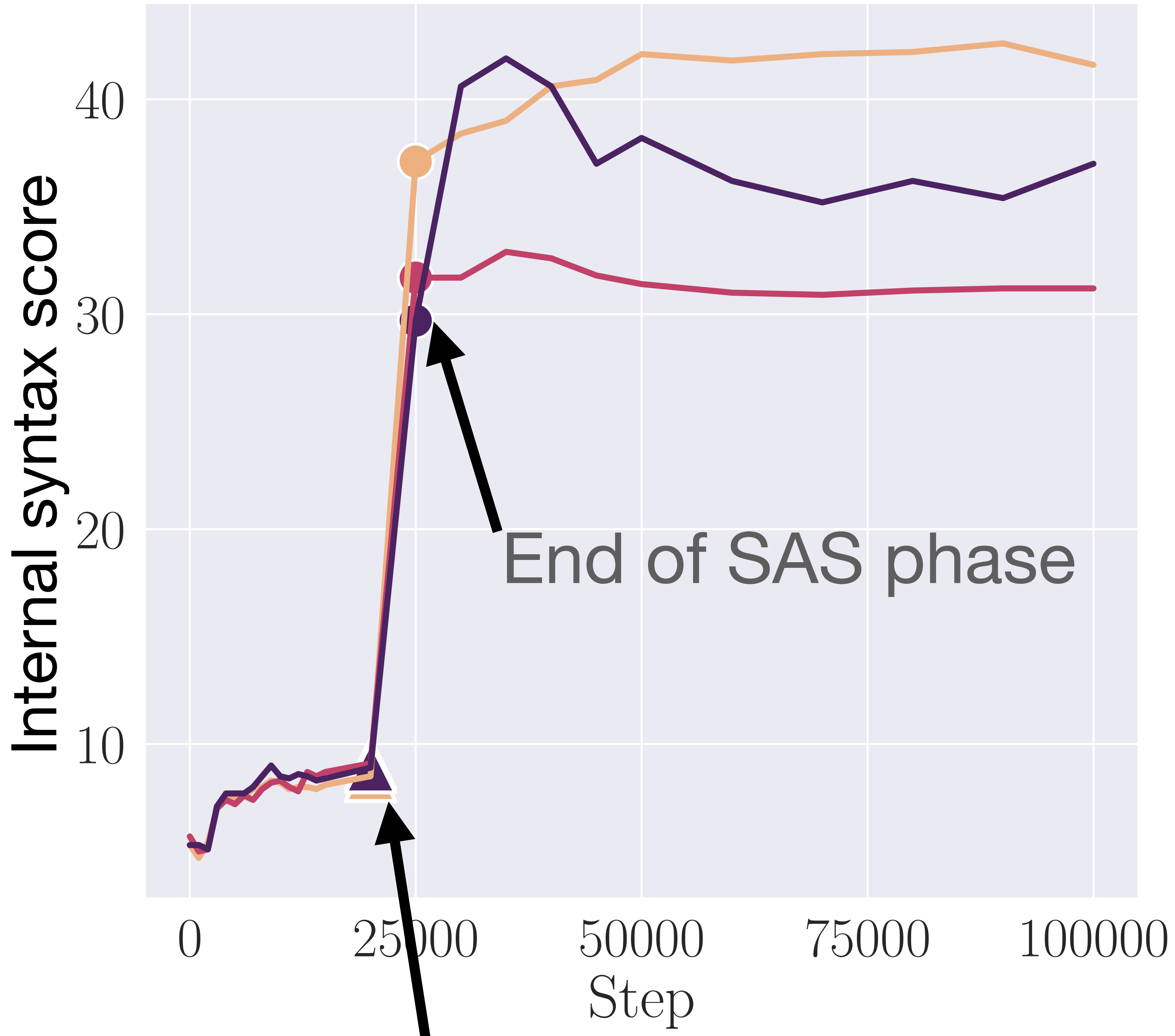
- **Instance level causal intervention**
  - Specialized syntactic heads hurt performance most when pruned.
  - (Voita et al., 2019)

What if specialized heads are more entangled, rather than themselves encoding structure?

Let’s find some evidence for the role of SAS!

When does Syntactic Attention Structure emerge?

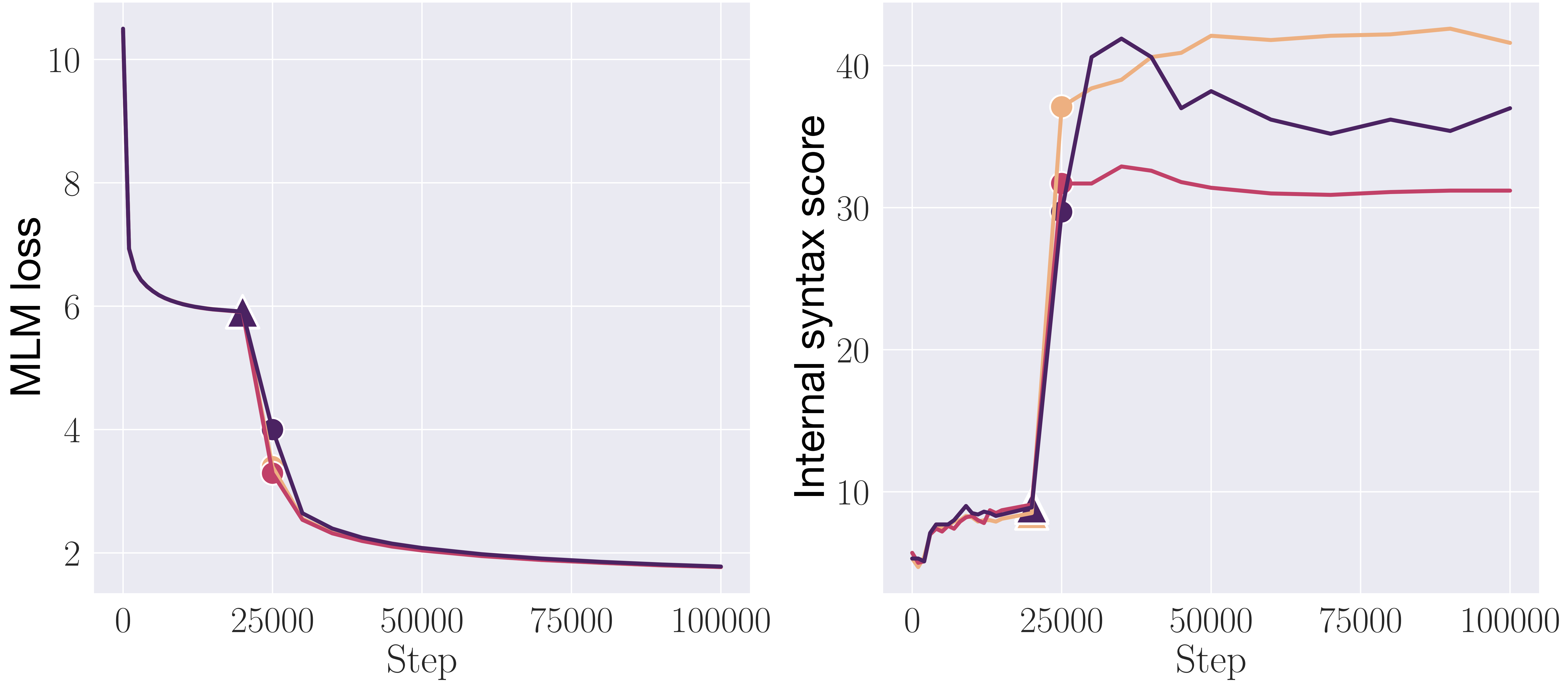
Syntactic Attention Structure is acquired abruptly



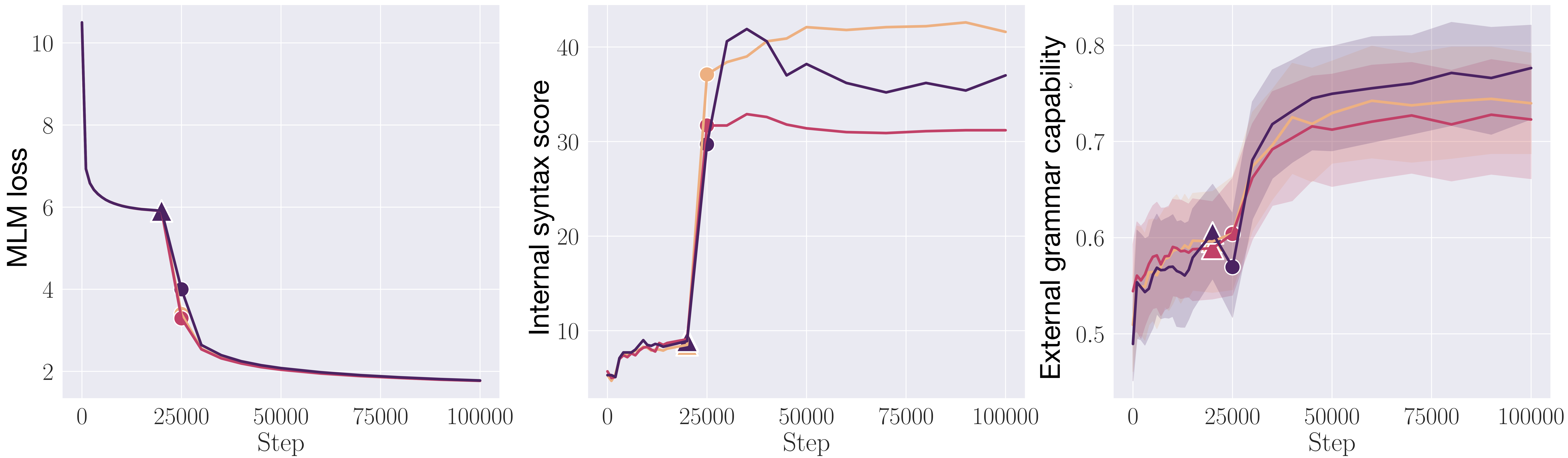
End of SAS phase

Onset of SAS phase

SAS phase accompanies a large loss drop



And is followed by gains in grammatical reasoning

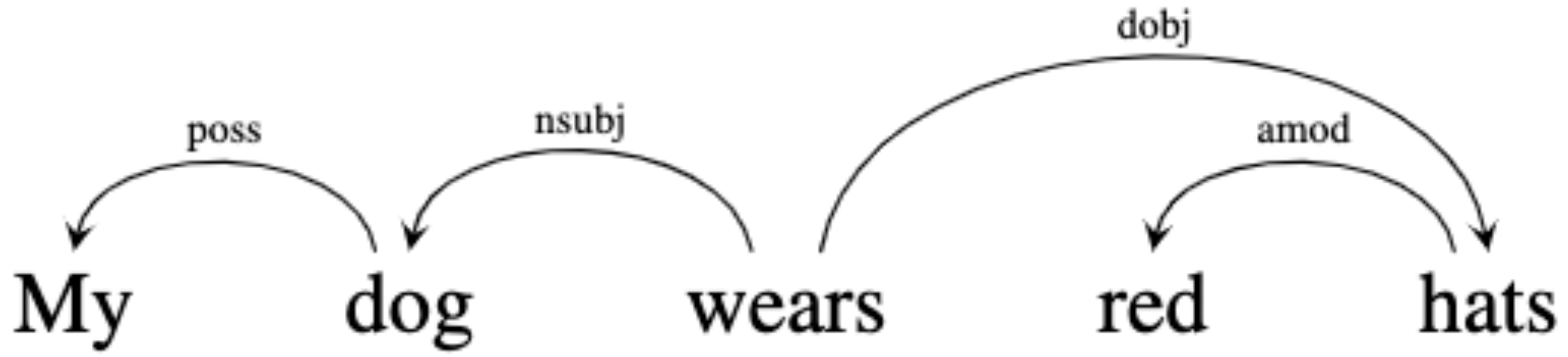


What makes a capability *breakthrough*?

- **Compositional structure**
- Competition between solutions
- Multimodality

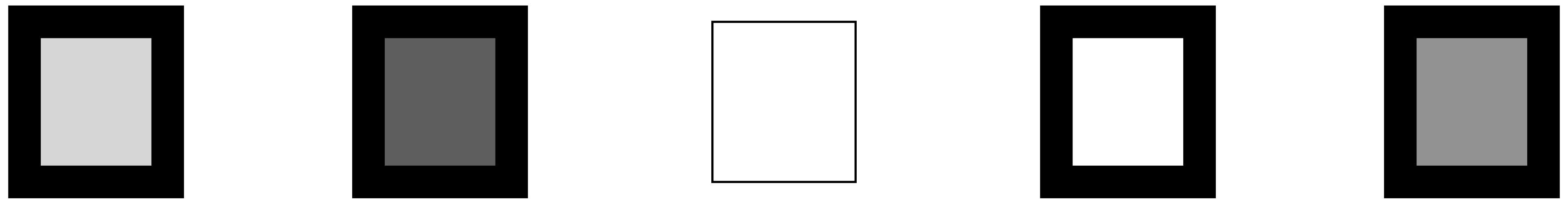
What happens if we suppress SAS? *Causal evidence!*

Suppressing Syntactic Attention Structure



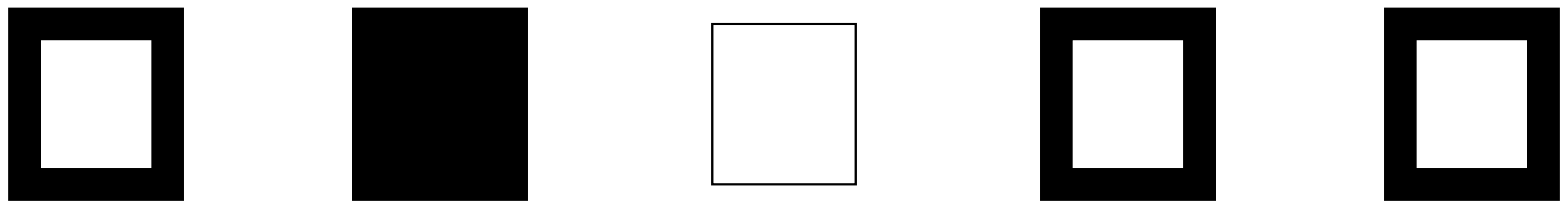
Target: “wears”

nsubj head



α_i^k

nsubj gold labels

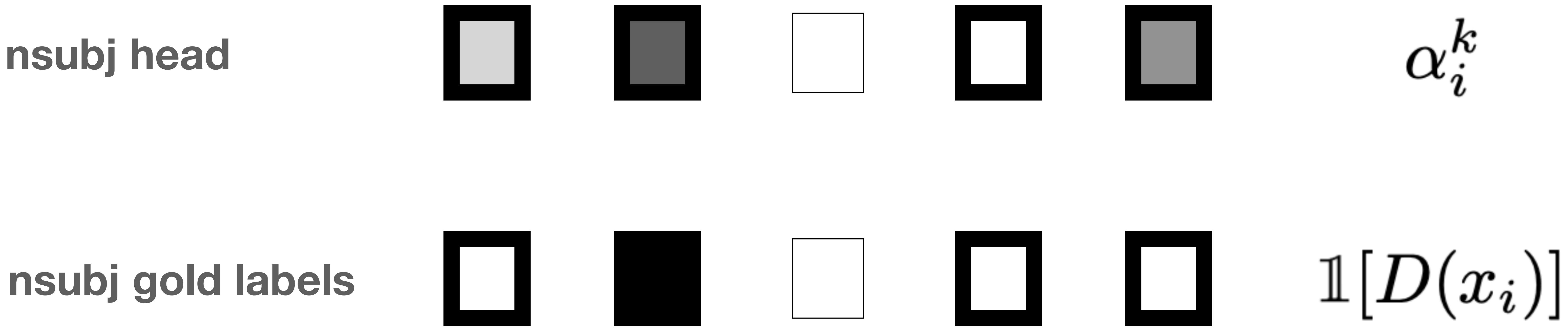


$\mathbb{1}[D(x_i)]$

Suppressing Syntactic Attention Structure

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

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



The impact of Syntactic Attention Structure

MLM loss
10^1
6×10^0
4×10^0
3×10^0
2×10^0
10^3
10^4
10^5
step
Internal syntax score
10^1
10^3
10^4
10^5
step
BERT-Base
SHS Suppressed

The impact of Syntactic Attention Structure

Bad at smaller scales …

MLM loss
10^1
6×10^0
4×10^0
3×10^0
2×10^0
10^3
10^4
10^5
step
Internal syntax score
10^1
10^3
10^4
10^5
step
BERT-Base
SHS Suppressed

The impact of Syntactic Attention Structure

But eventually important!

Onset of competing strategy
Onset of SAS strategy
MLM loss
10^1
6×10^0
4×10^0
3×10^0
2×10^0
10^3
10^4
10^5
step
Internal syntax score
10^1
10^3
10^4
10^5
step
BERT-Base
SHS Suppressed

Why are there *two* phase transitions?

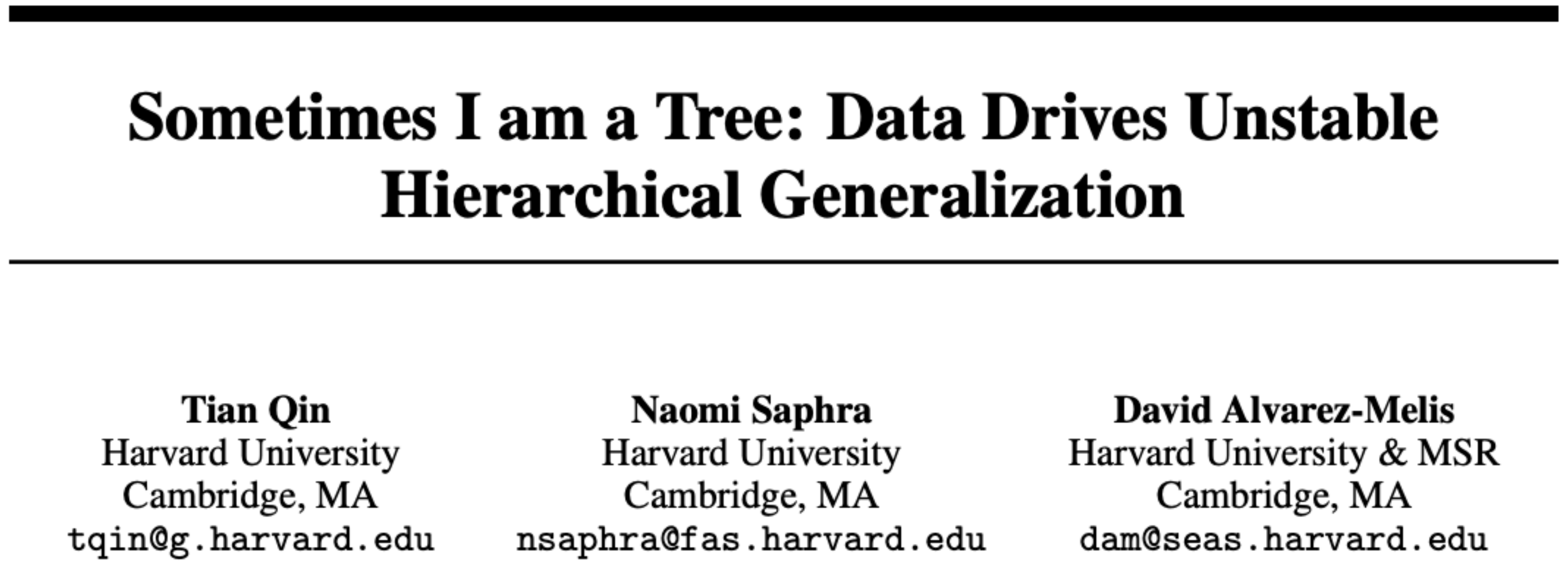
We have found a competing strategy.

Onset of competing strategy
Onset of SAS strategy
MLM loss
10^1
6×10^0
4×10^0
3×10^0
2×10^0
10^3
10^4
10^5
step
Internal syntax score
10^1
10^3
10^4
10^5
step
BERT-Base
SHS Suppressed

What makes a capability *breakthrough*?

- Compositional structure
- **Competition between solutions**
- Multimodality

Case study 2: What makes hierarchical syntax grok?



Sometimes I am a Tree: Data Drives Unstable Hierarchical Generalization

Tian Qin
Harvard University
Cambridge, MA
tqin@g.harvard.edu

Naomi Saphra
Harvard University
Cambridge, MA
nsaphra@fas.harvard.edu

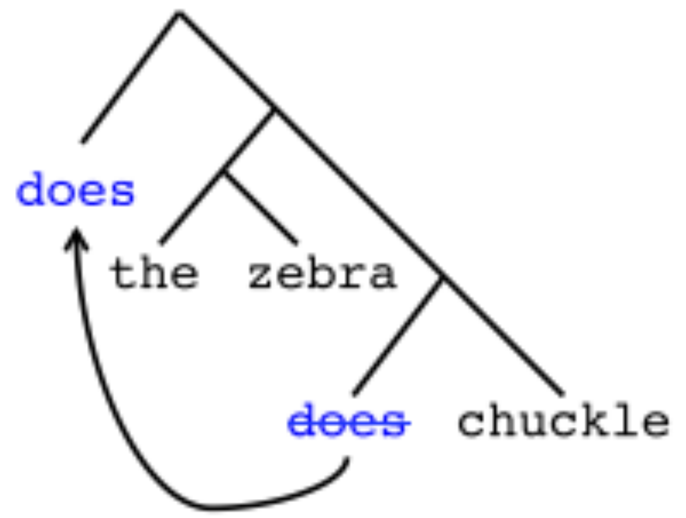
David Alvarez-Melis
Harvard University & MSR
Cambridge, MA
dam@seas.harvard.edu

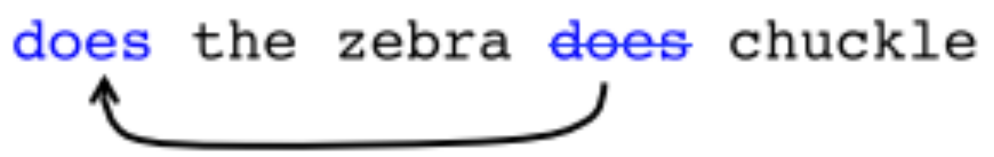
Will we learn hierarchical syntactic generalization?

Ambiguous rule: question formation (McCoy et al., 2019)

In Distribution:

Input: My unicorn does move the dogs that do wait.
Output: Does my unicorn move the dogs that do wait?



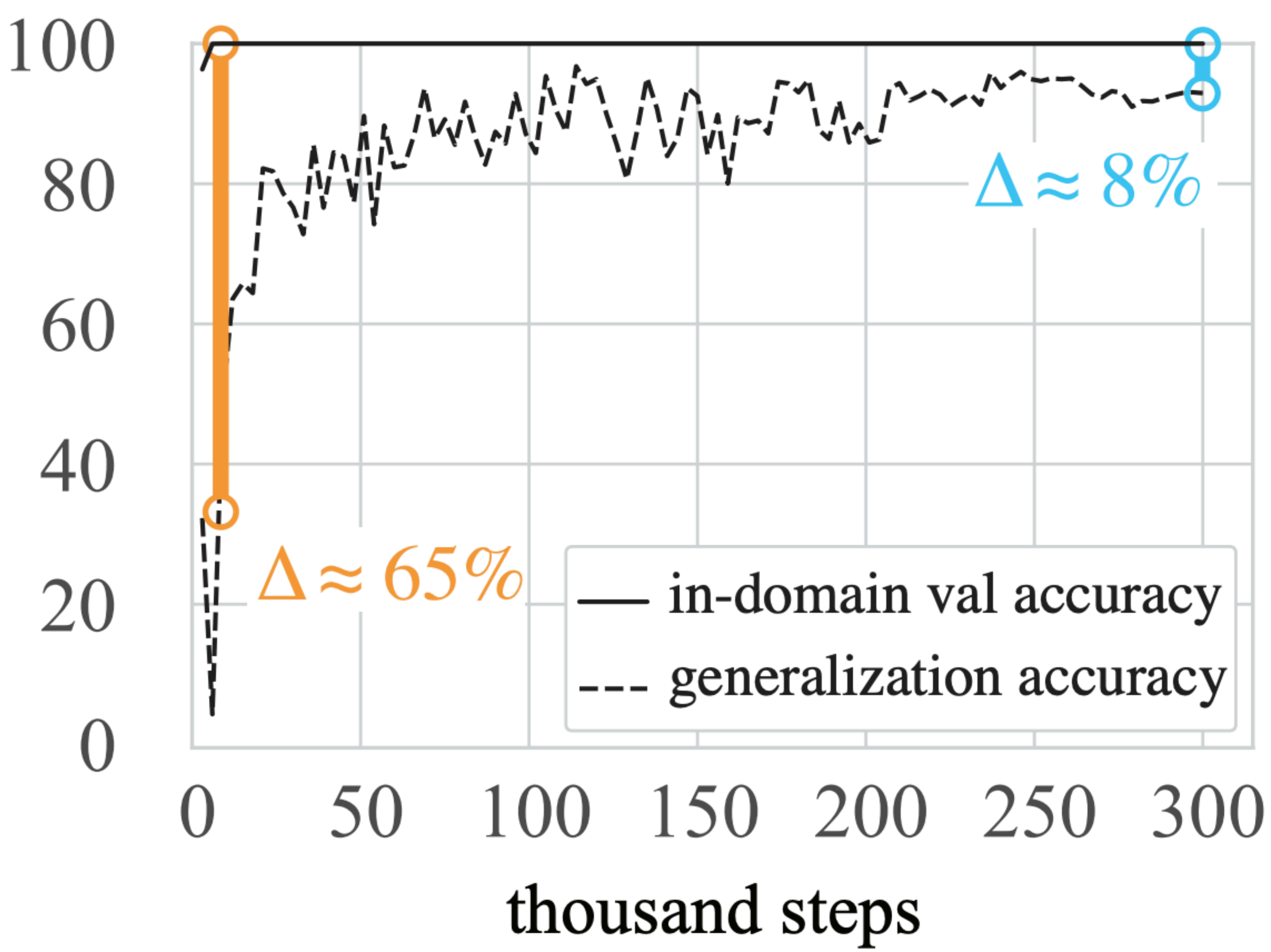


Out of Distribution:

Input: My unicorn who doesn’t sing does move.
Linear Output: Doesn’t my unicorn who sing does move?
Hierarchical Output: Does my unicorn who doesn’t sing move?

Hierarchical syntax groks after ID accuracy converges for an autoregressive LM.

Murty et al., 2023; Ahuja et al., 2024



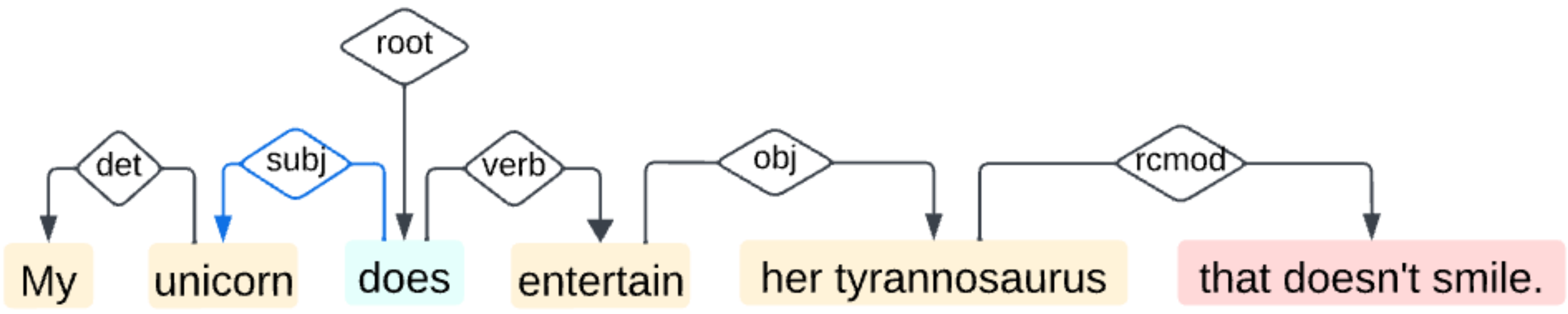
What makes a capability *breakthrough*?

- **Compositional structure**
- Competition between solutions
- Multimodality

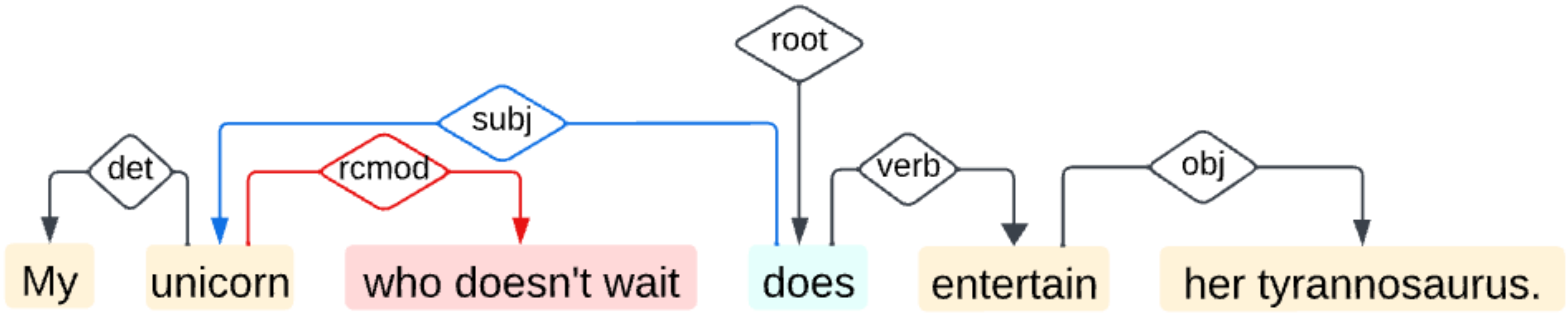
Hierarchical generalization depends on *center embeddings*

- English language mostly branches right …
- So if each head only gets one relative clause, it will be exclusively *forward scoping.*
  - *That doesn't require hierarchical structure at all!*

Right Branching

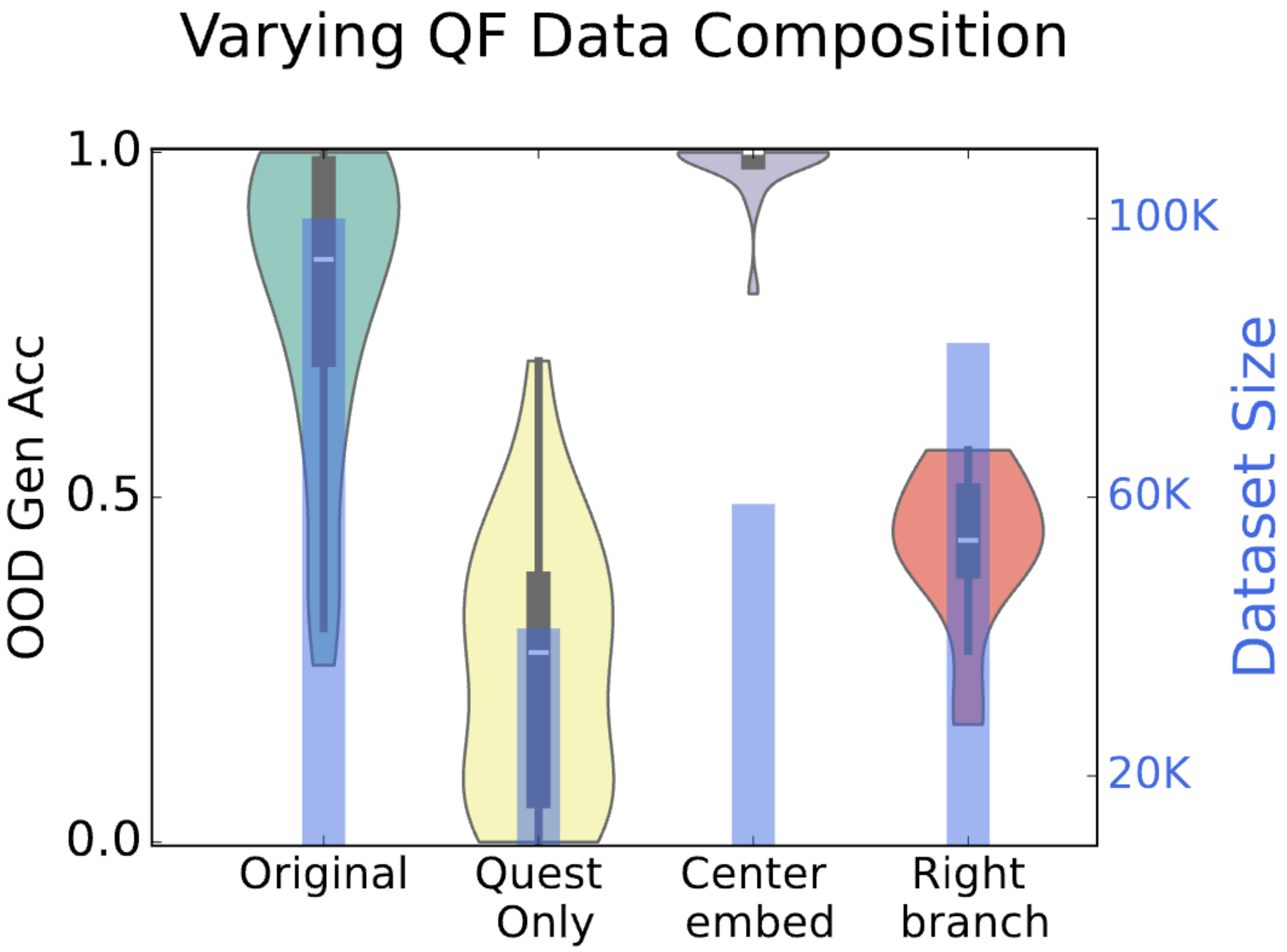


Center Embedding



Hierarchical generalization depends on *center embeddings*

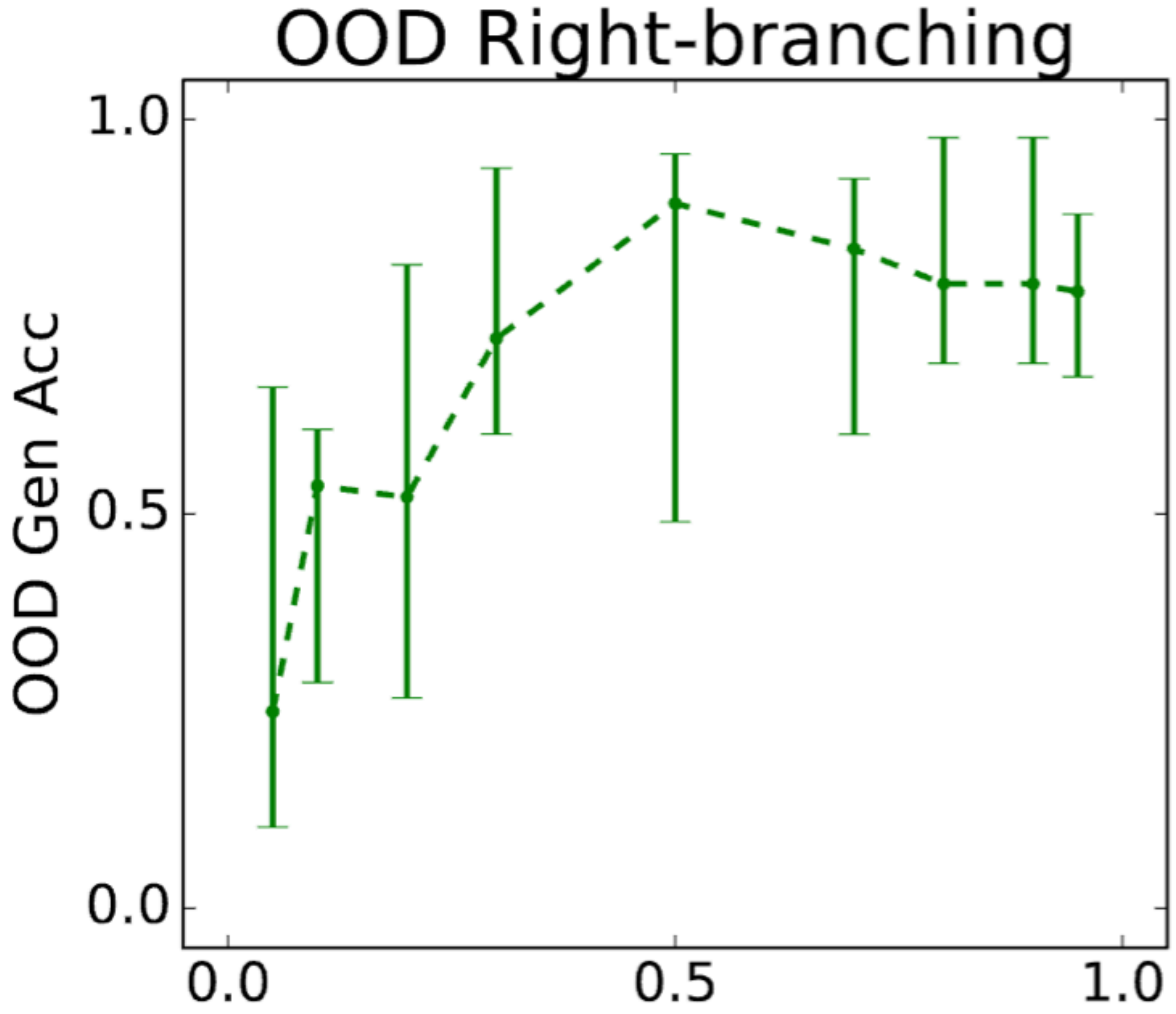
Complex training data leads more training runs to generalize.



Complex training data teaches complex rules.

What happens if you *mix* “easy” and “hard” data?

Training doesn’t lead to consistent OOD behavior!



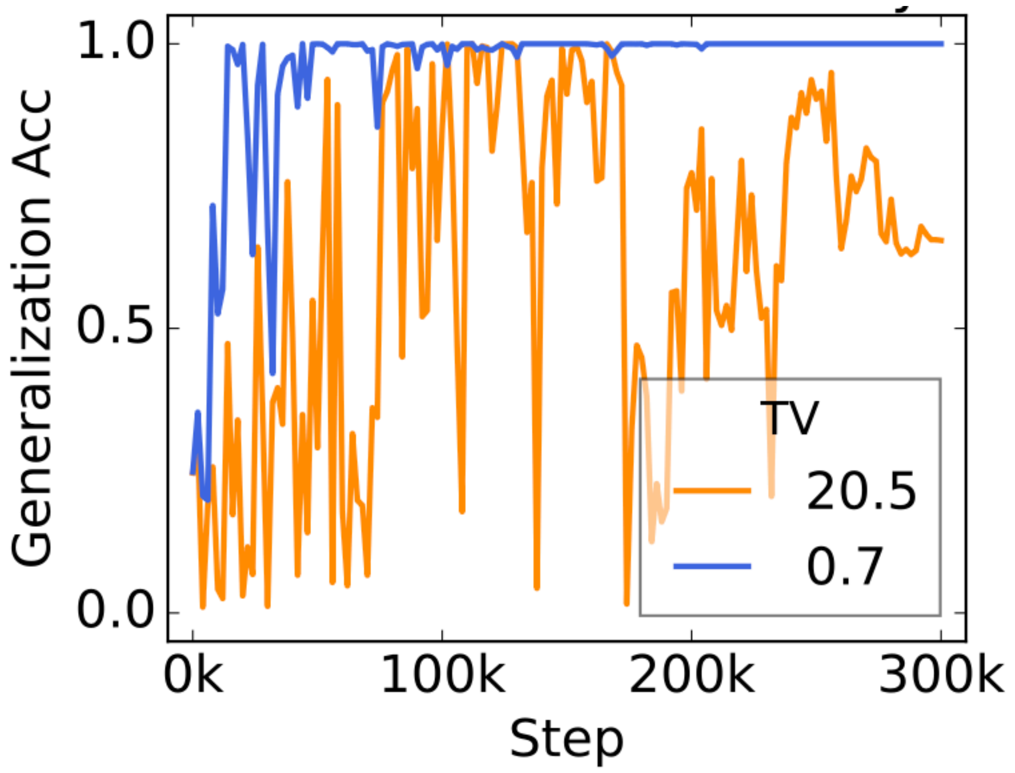
Proportion of center embeddings in training

What makes a capability *breakthrough*?

- Compositional structure
- **Competition between solutions**
- Multimodality

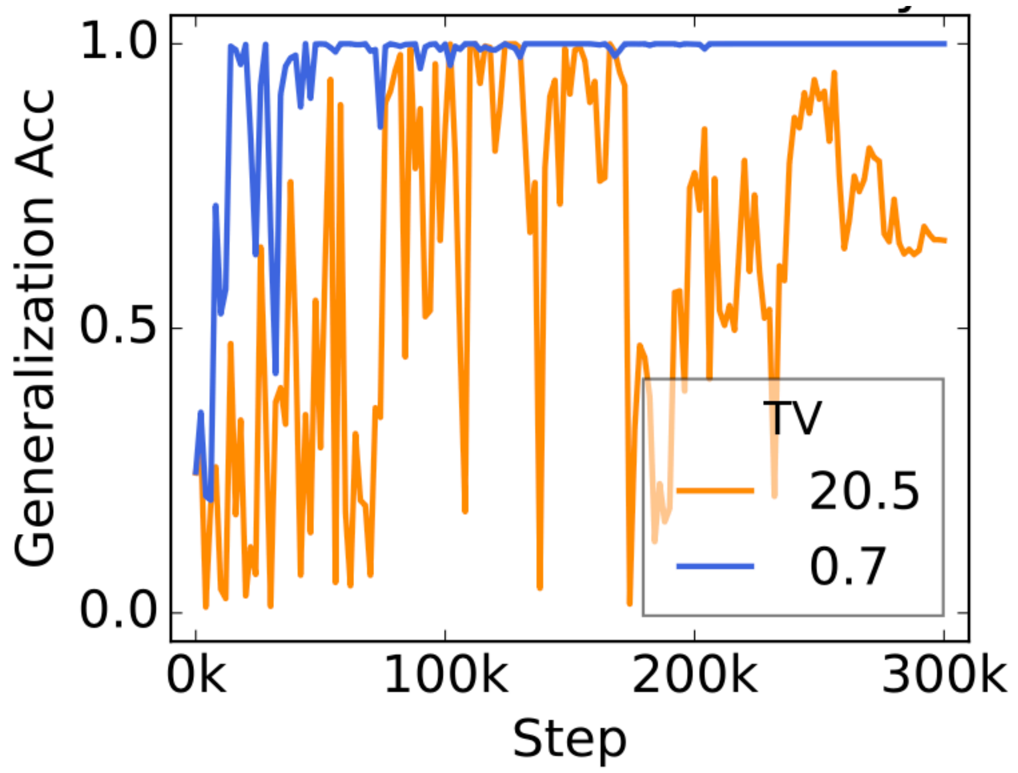
Only models that commit to a simple rule can stabilize OOD behavior

Measuring stability

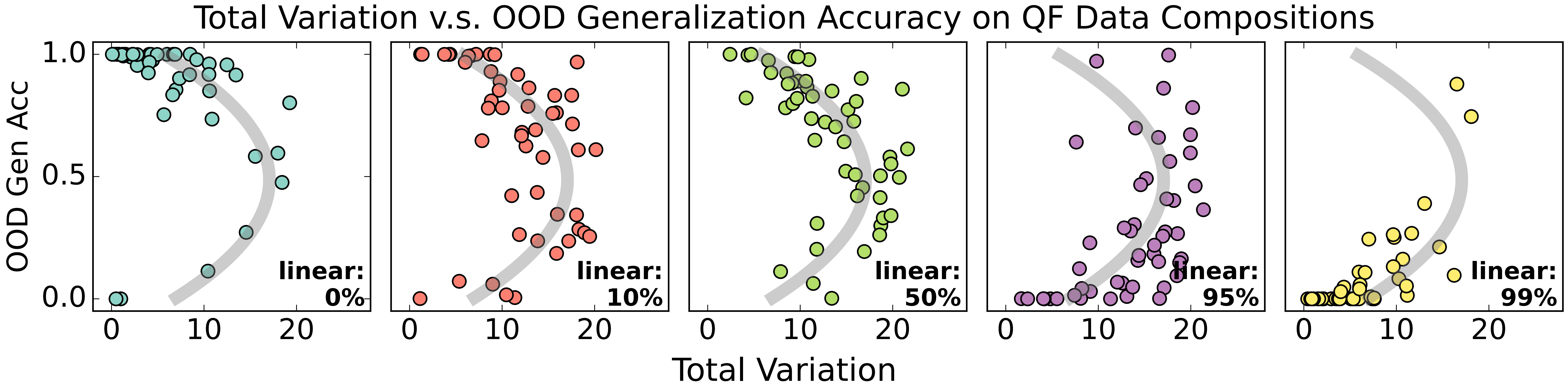


$$\text{Total Variation (TV)} = \text{Avg}_i \left(|\text{Acc}_i - \text{Acc}_{i-1}| \right)$$

Only models that commit to a simple rule can stabilize OOD behavior

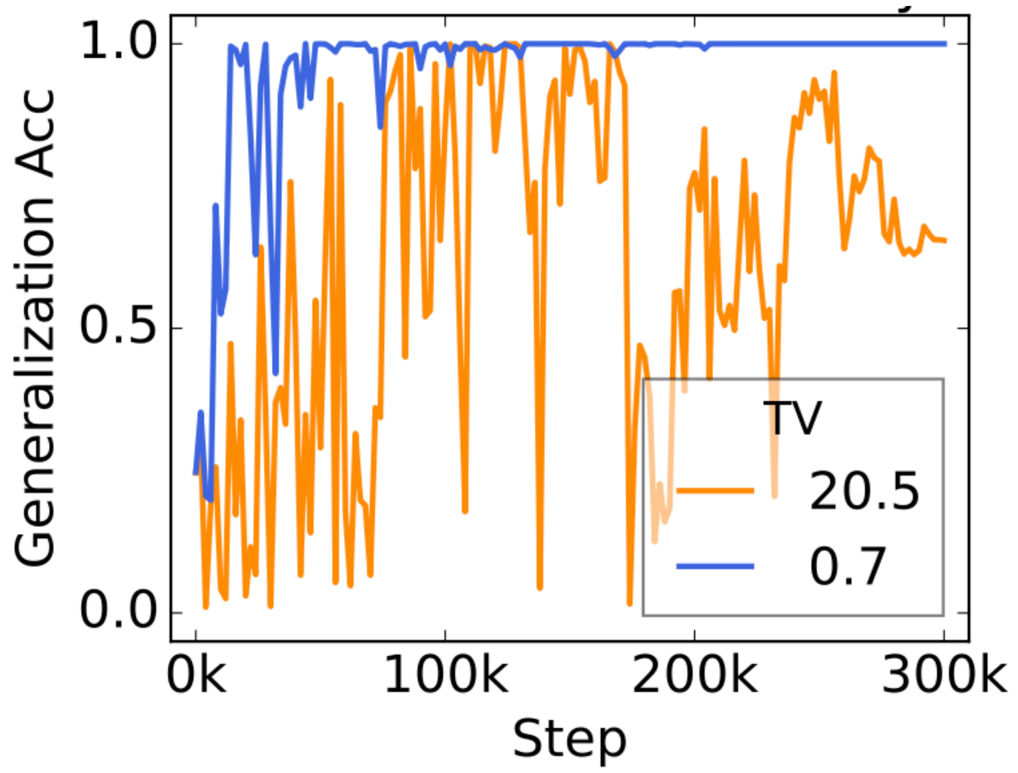


$$\text{Total Variation (TV)} = \text{Avg}_i\left(|\text{Acc}_i - \text{Acc}_{i-1}|\right)$$

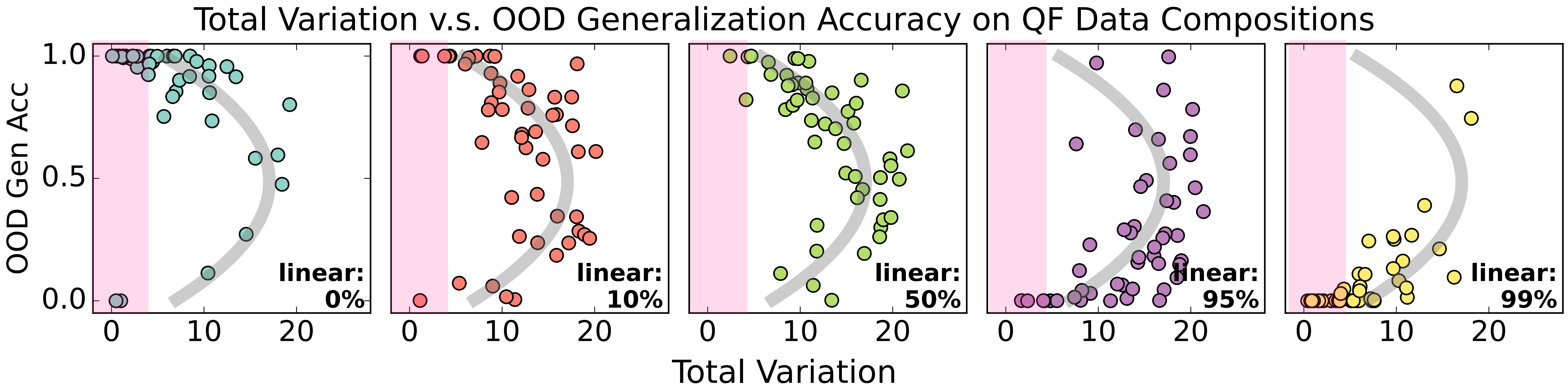


Only models that commit to a simple rule can stabilize OOD behavior

Stable models are bimodally distributed

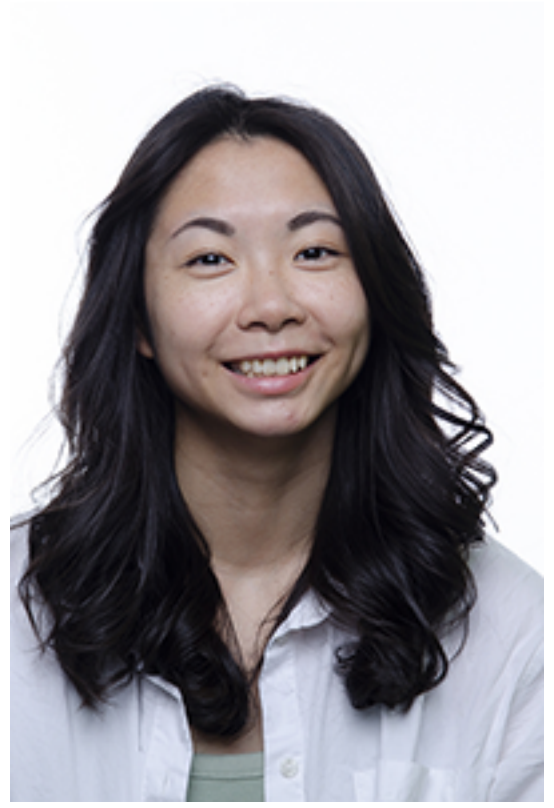


$$\text{Total Variation (TV)} = \text{Avg}_i\left(|\text{Acc}_i - \text{Acc}_{i-1}|\right)$$

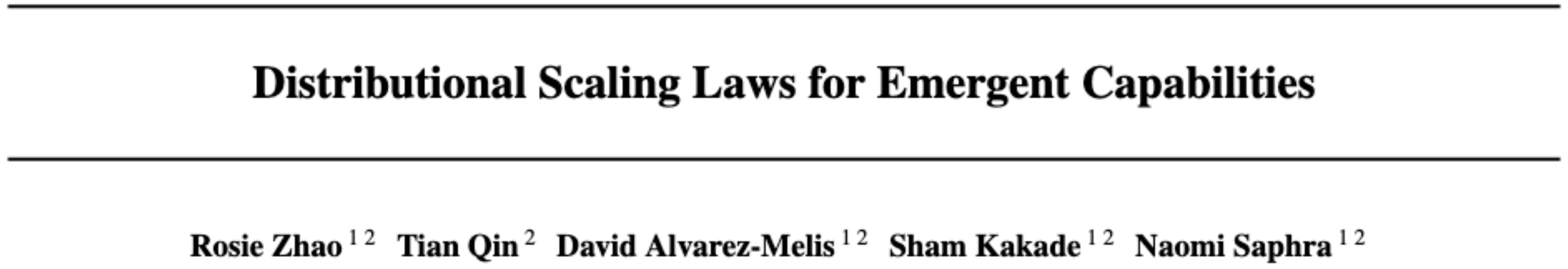


What makes a capability *breakthrough*?

- Compositional structure
- Competition between solutions
- **Multimodality**



Case study 3: Predicting unpredictable emergence in length generalization



Distributional Scaling Laws for Emergent Capabilities

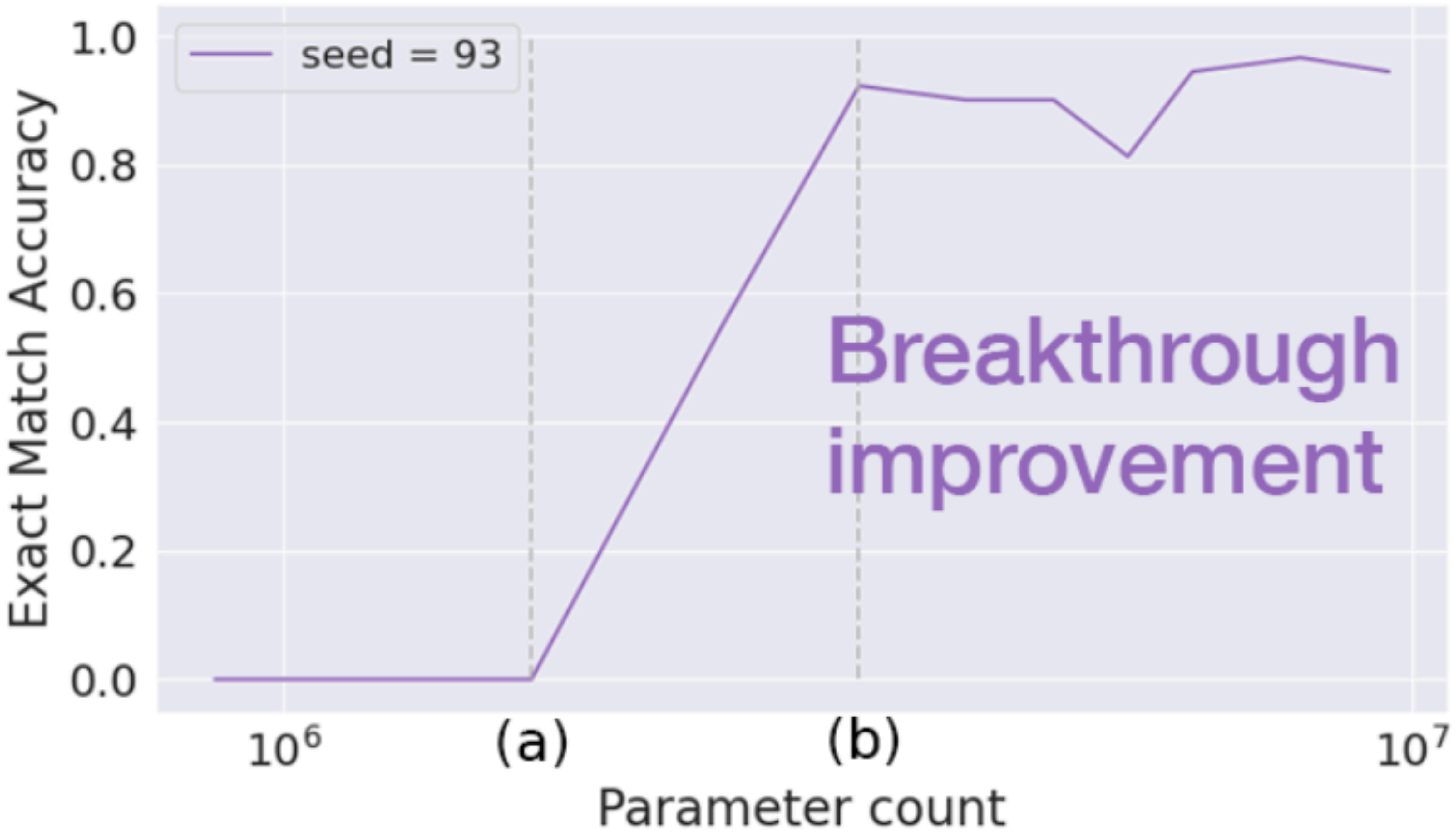
Rosie Zhao [1 2] **Tian Qin** [2] **David Alvarez-Melis** [1 2] **Sham Kakade** [1 2] **Naomi Saphra** [1 2]

Length generalization: reverse order addition

Zhou et al., 2023; Zhou et al., 2024

- Train on 30 characters, test on 40
    - Compositional *productivity* or *length generalization*
- 200 seeds trained for each architecture
- Example task: **Reverse order addition**
    - Output sum in reverse order
    - Input includes index hints
    - a0, 3, a1, 4, +, a0, 2, a1, 8, >, a1, 2, a0, 6

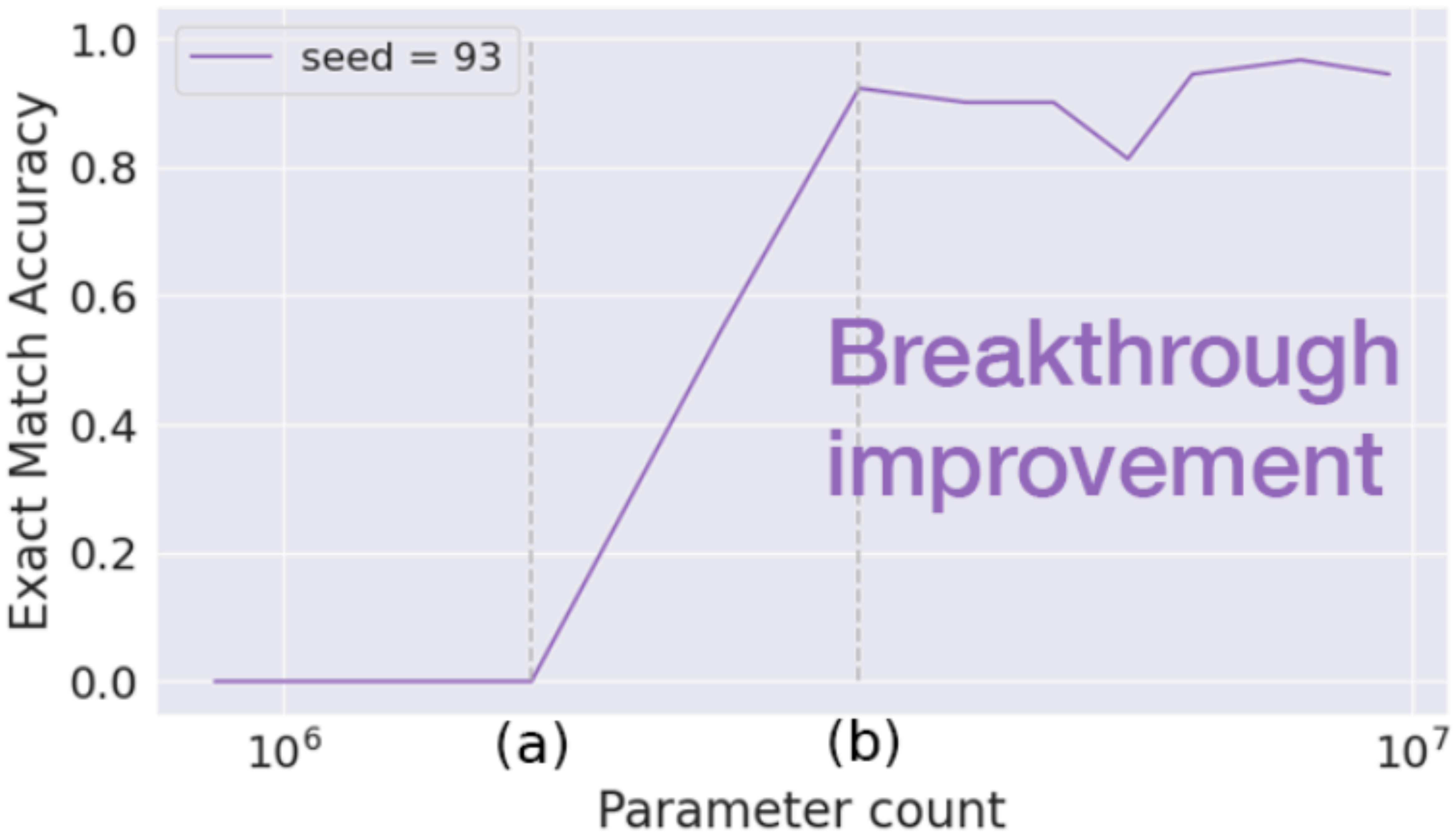
Emerges at appropriate width scale!



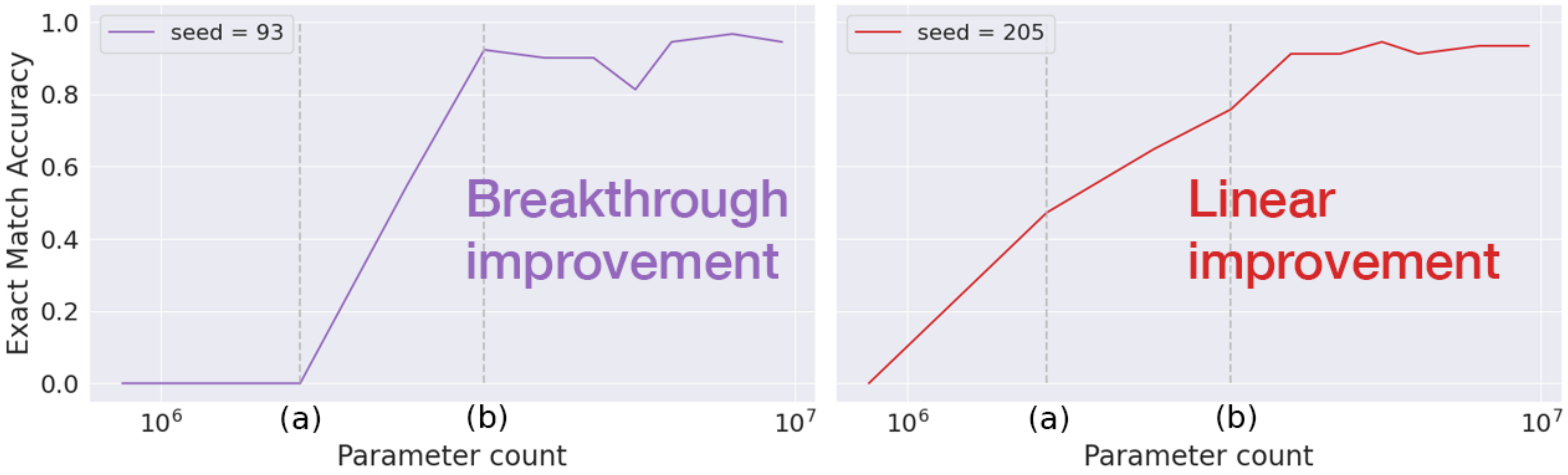
What makes a capability *breakthrough*?

- **Compositional structure**
- Competition between solutions
- Multimodality

Emergence!


seed = 93
Exact Match Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
10^6
(a)
(b)
10^7
Parameter count
Breakthrough improvement

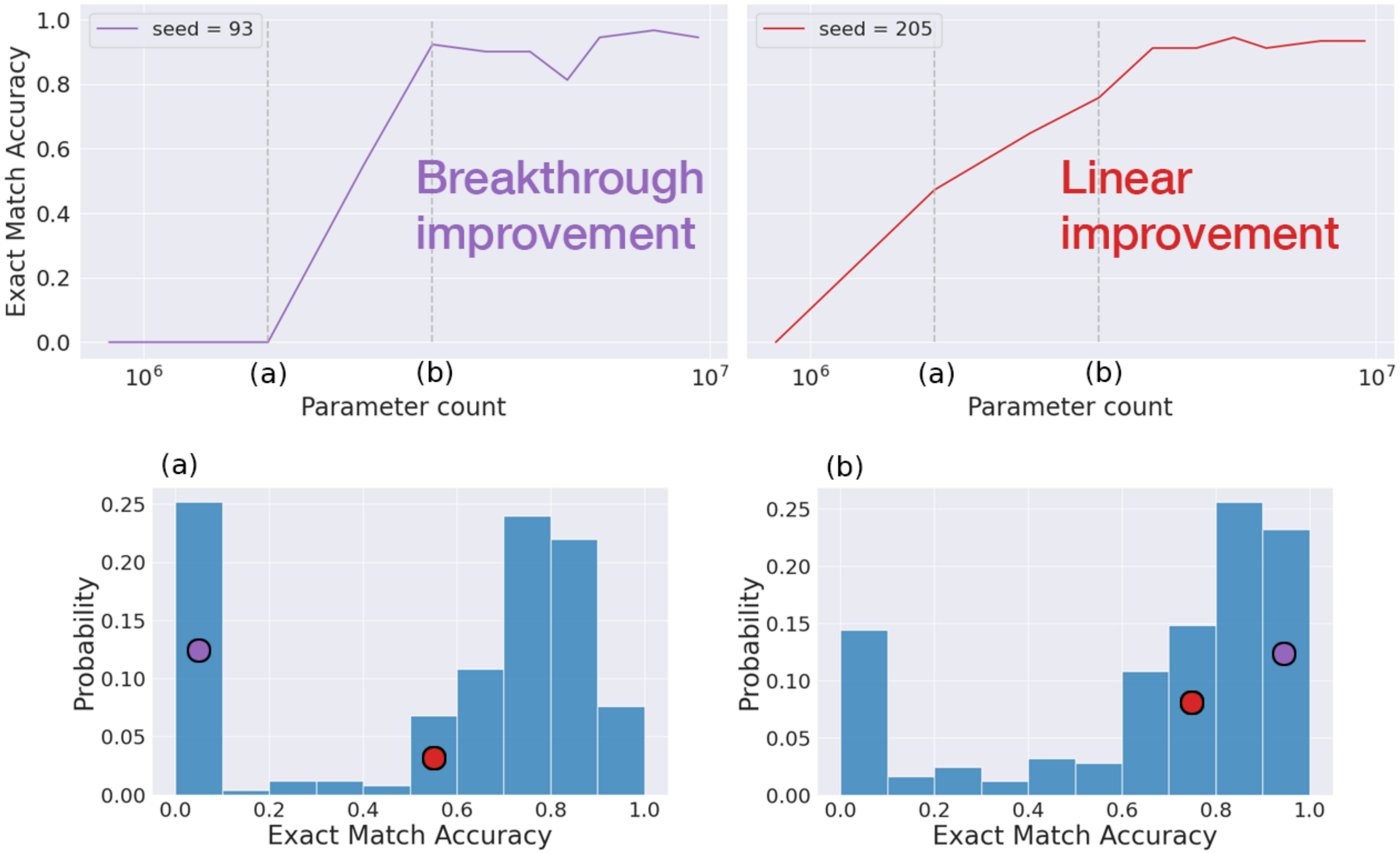
Or is it?



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

Emergence is when the model selects a “successful” run

But performance is bimodal!

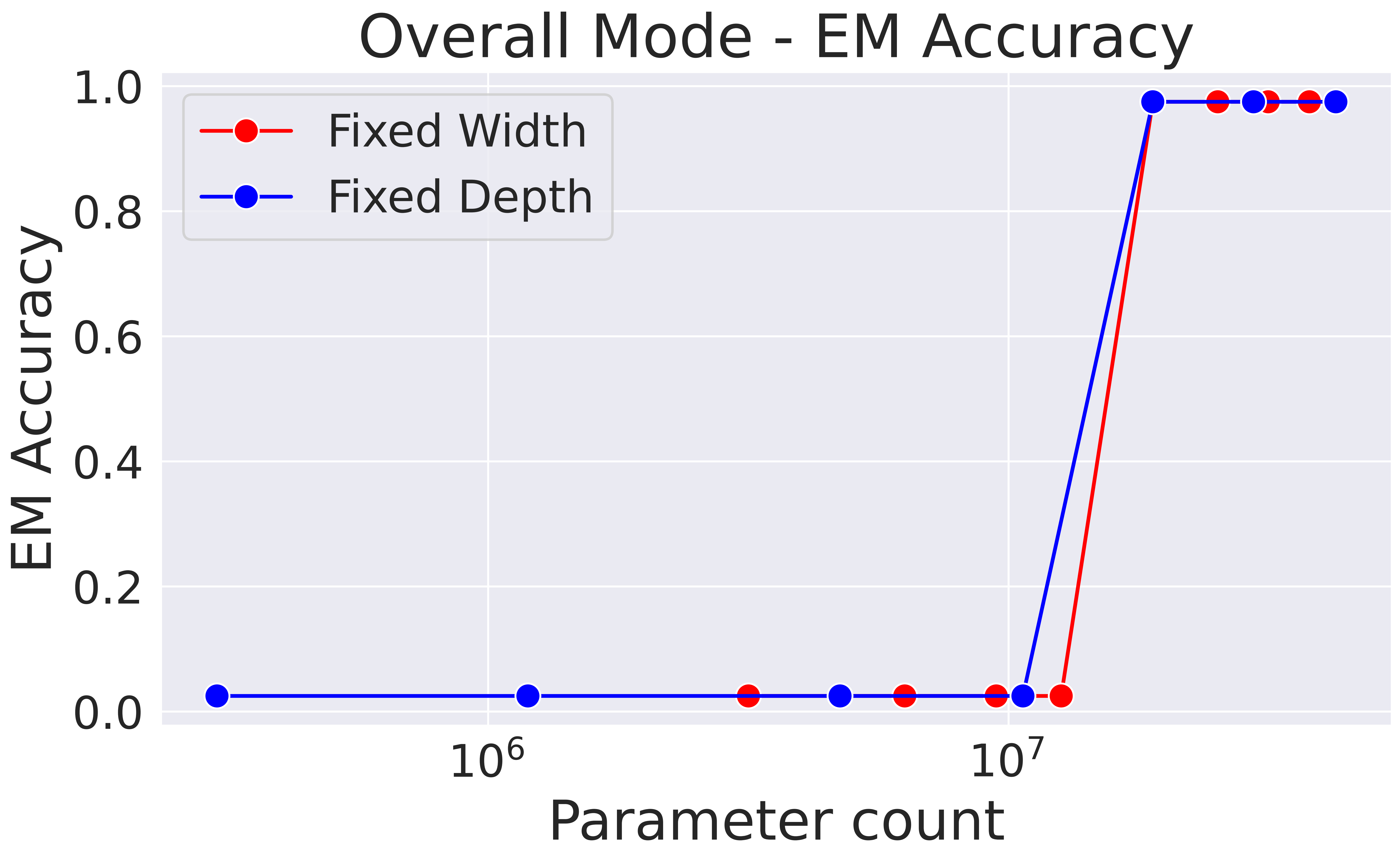


What makes a capability *breakthrough*?

- Compositional structure
- Competition between solutions
- **Multimodality**

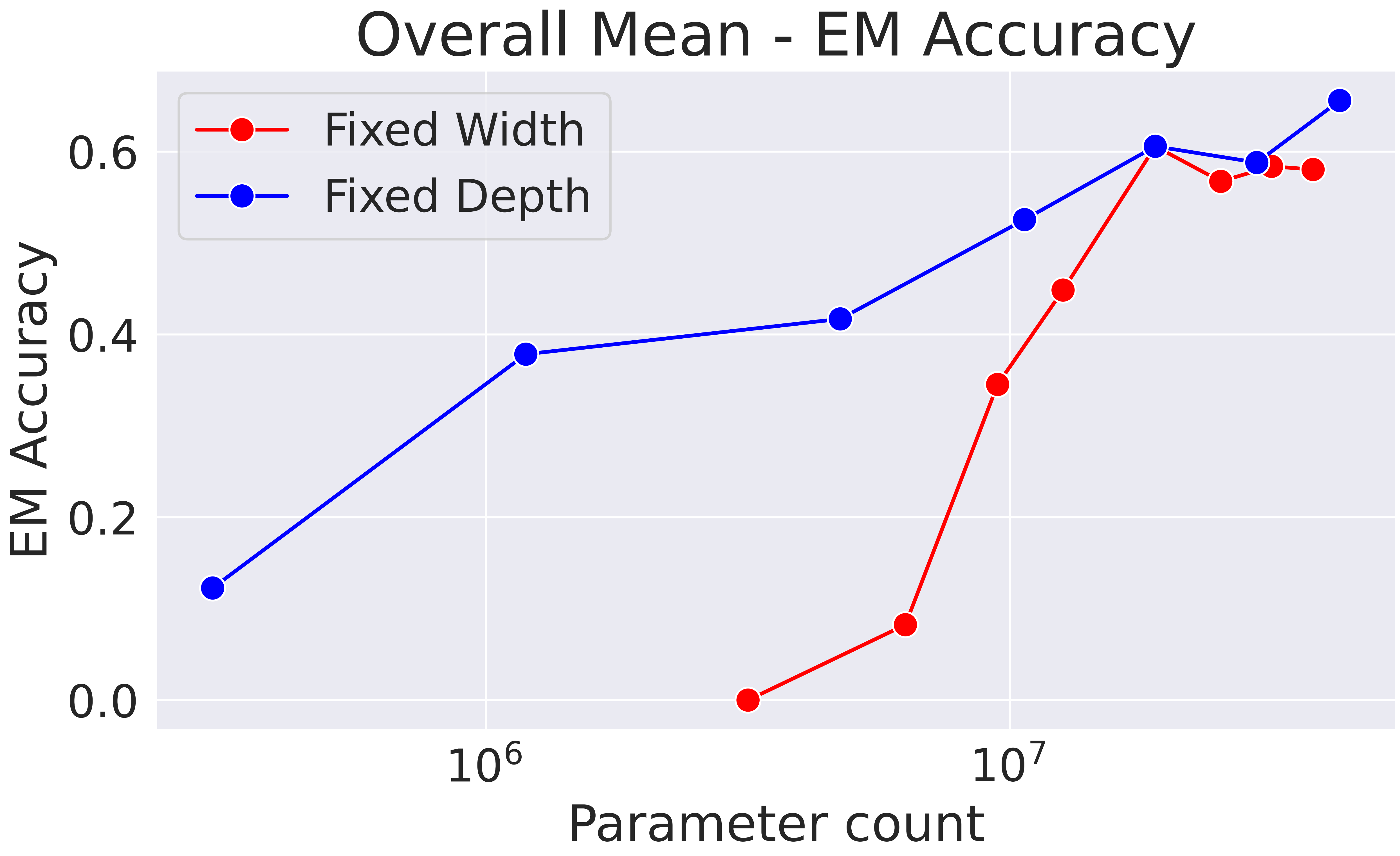
Emergent trends express underlying continuous changes

Individual scaling "laws" look like the mode

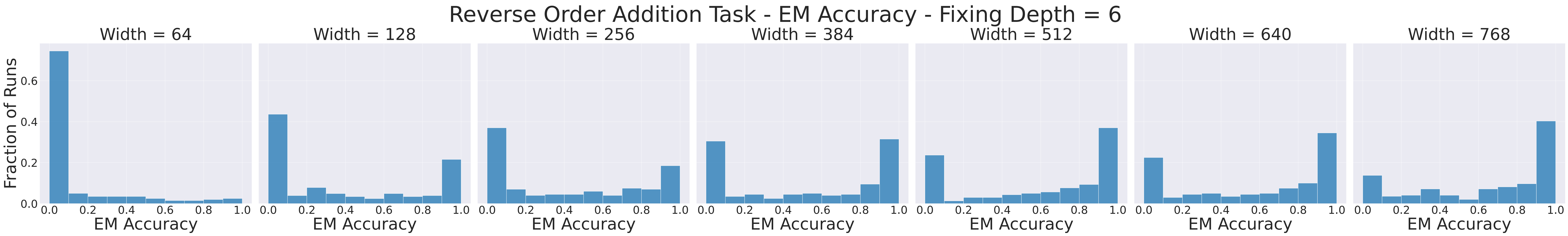


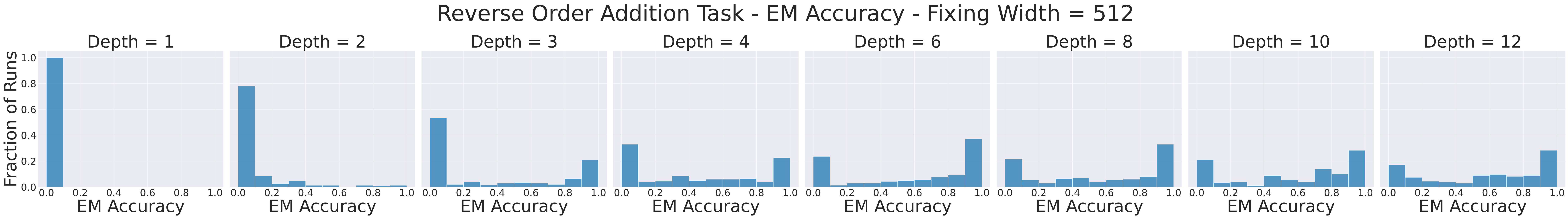
Emergent trends express underlying continuous changes

But with enough samples, mean can be smoother



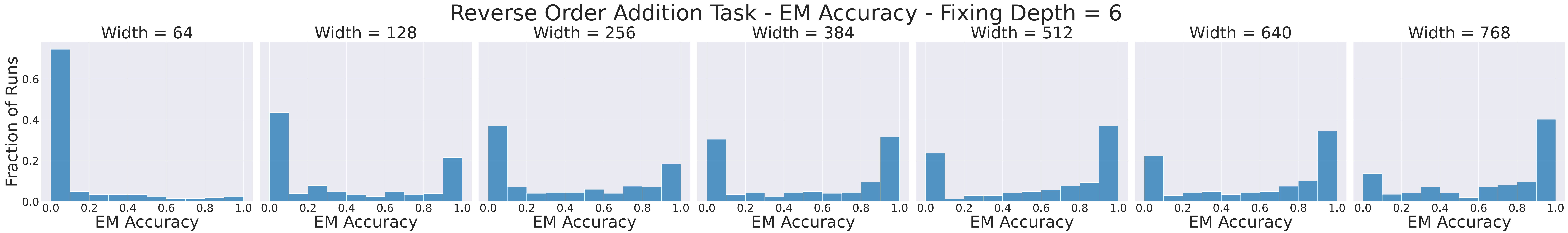
Bimodal distributions change gradually

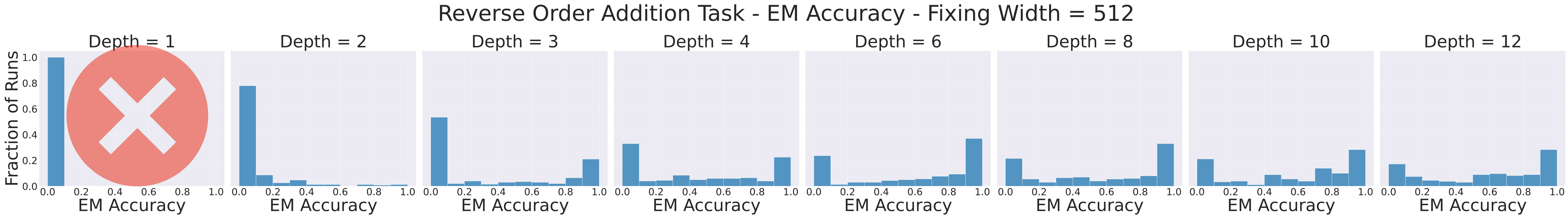

Reverse Order Addition Task - EM Accuracy - Fixing Depth = 6


Reverse Order Addition Task - EM Accuracy - Fixing Width = 512

Bimodal distributions change gradually

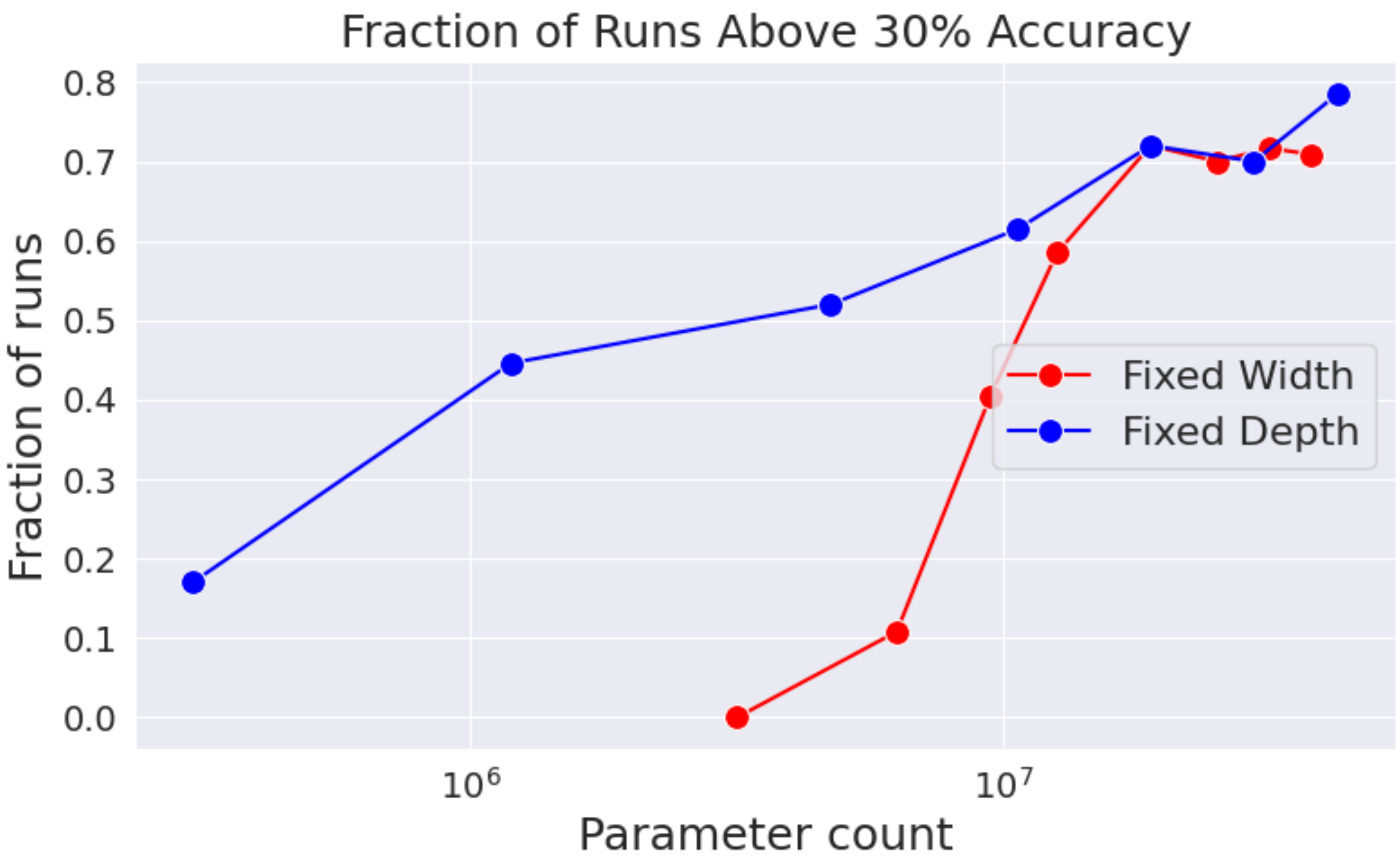
(As long as we have minimum capacity)





Why is the mode discontinuous?

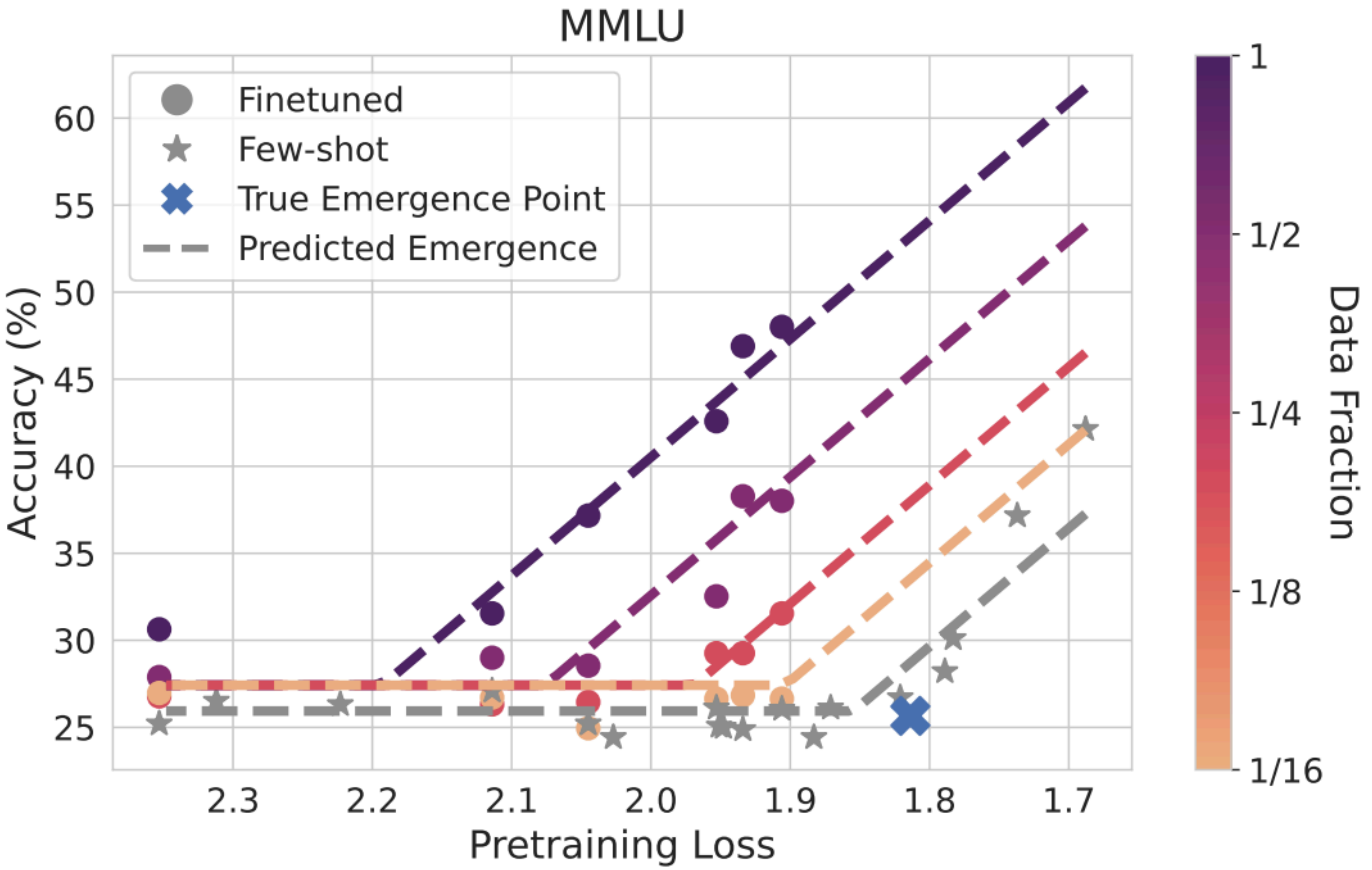
Gradual change in PROBABILITY of success



Real world example: multiple choice QA

MMLU dataset

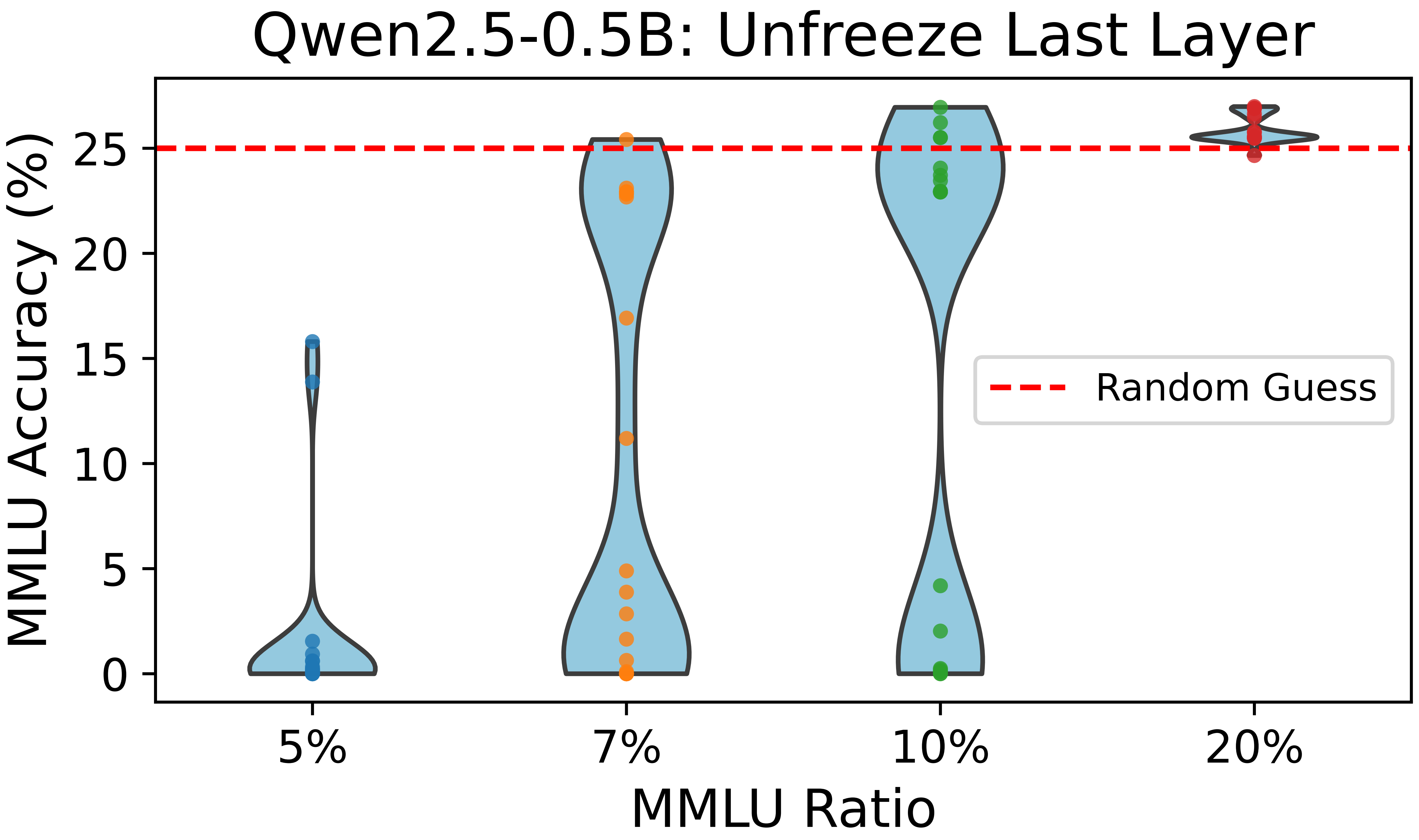
- Emergent because it is compositional
    - Without multiple choice format, QA improvement is smooth
- With extra finetuning / exposure to dataset, can emerge at smaller model scales



Snell et al., 2025

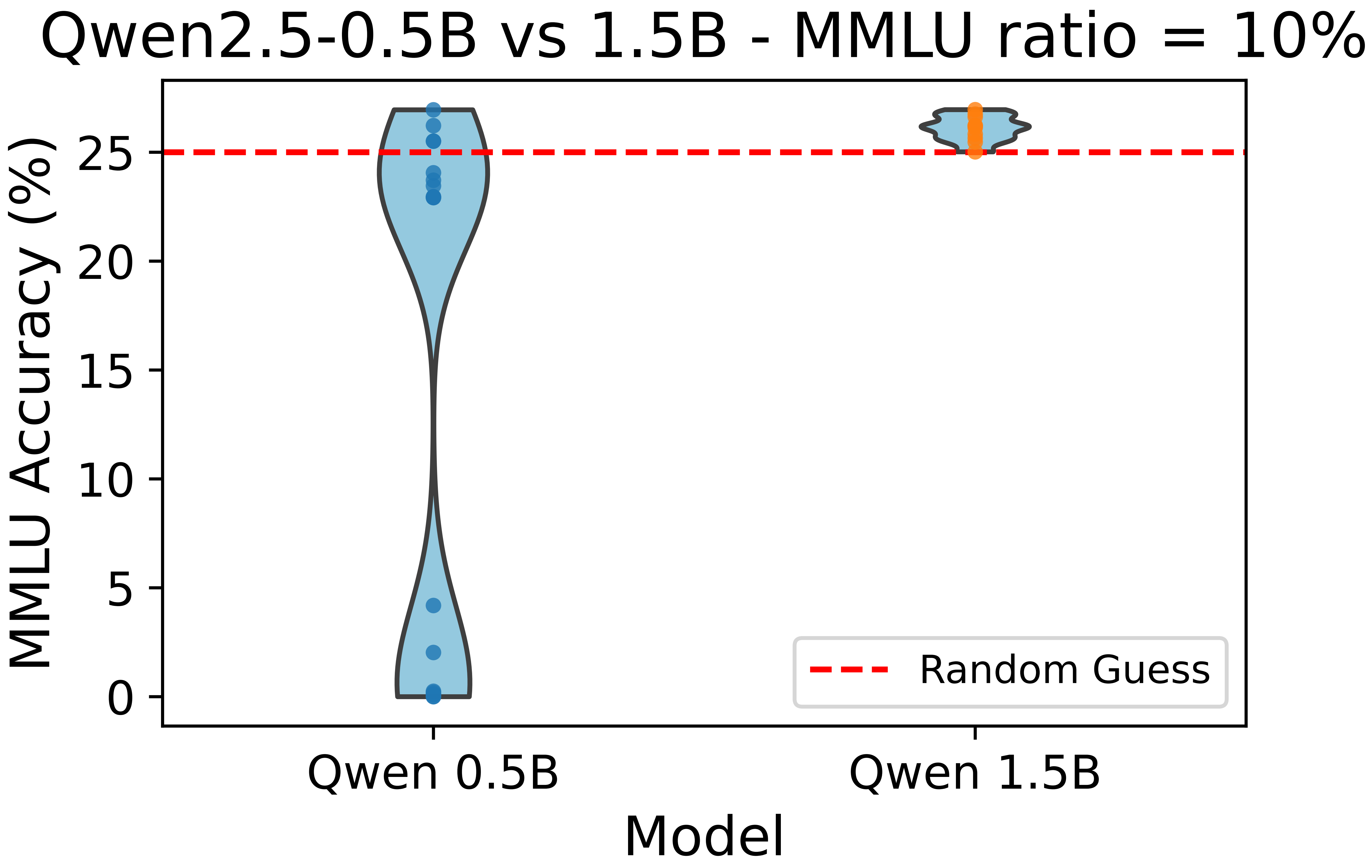
Training after top layer reinitialization

With random variation, MMLU is bimodal



Training after top layer reinitialization

With enough scale, eventually collapses to top mode



What makes a capability *breakthrough*?

- Compositional structure
- Competition between solutions
- **Multimodality**

Recap

- MLMs develop specialized syntactic heads suddenly during a huge loss drop, and immediately afterwards learn complex linguistic rules during another huge loss drop.

- Causal LMs trained on ambiguous data develop an inductive bias towards hierarchical rules, but only if exposed to enough center embeddings that cannot be represented with linear structures.

- In length generalization, emergence looks discontinuous for a single sample, but once the model has theoretical capacity, changes in probability are continuous.

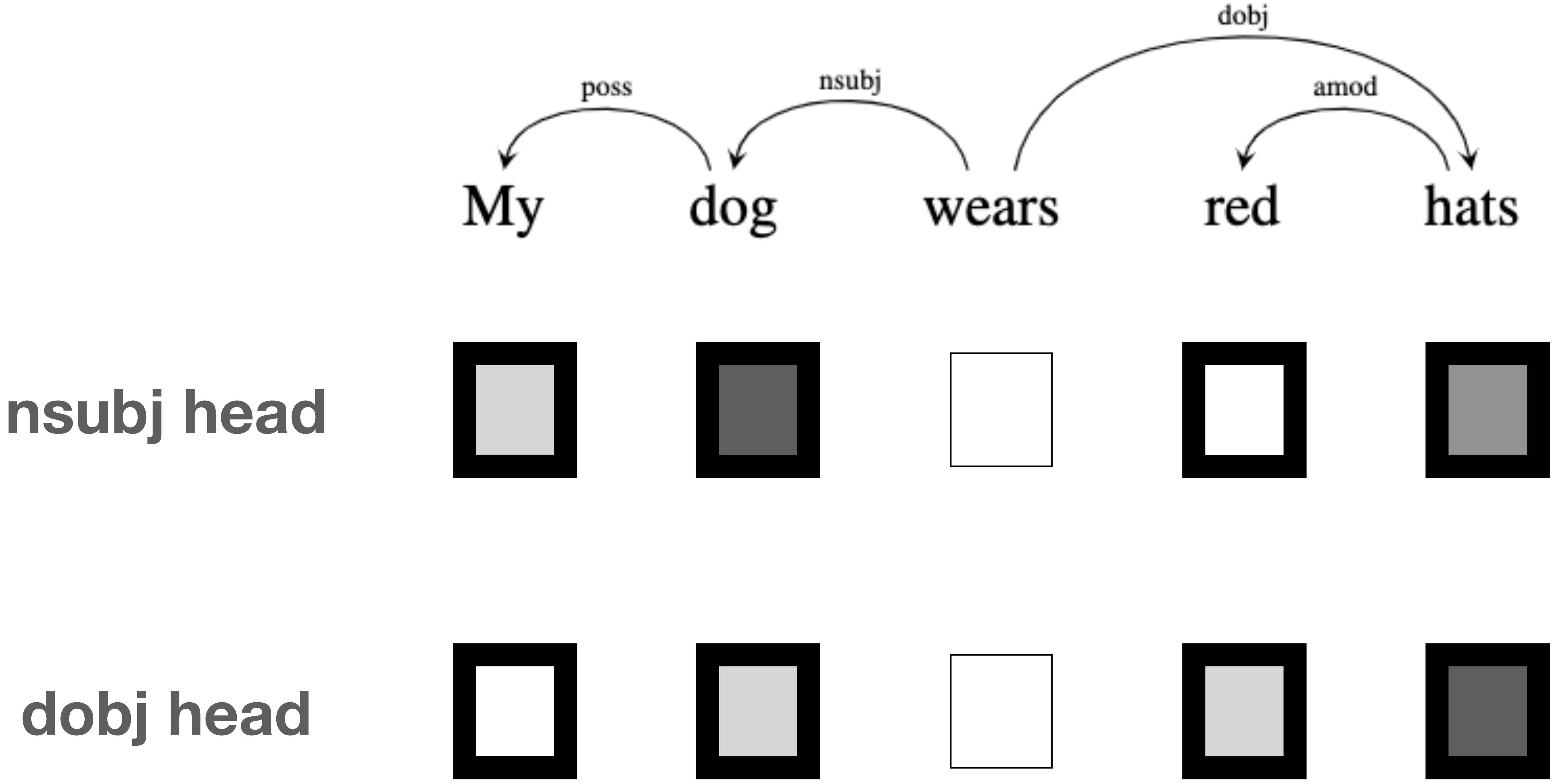
What makes a capability *breakthrough*?

- Compositional structure
- Competition between solutions
- Multimodality

Do I have extra time?
Let’s talk about mysterious U-shaped curves!

Mystery #1: U-shaped regularizer responses

When should MLMs learn syntax?



Target: “wears”

There *two* phase transitions?

Onset of competing strategy

Onset of SAS strategy

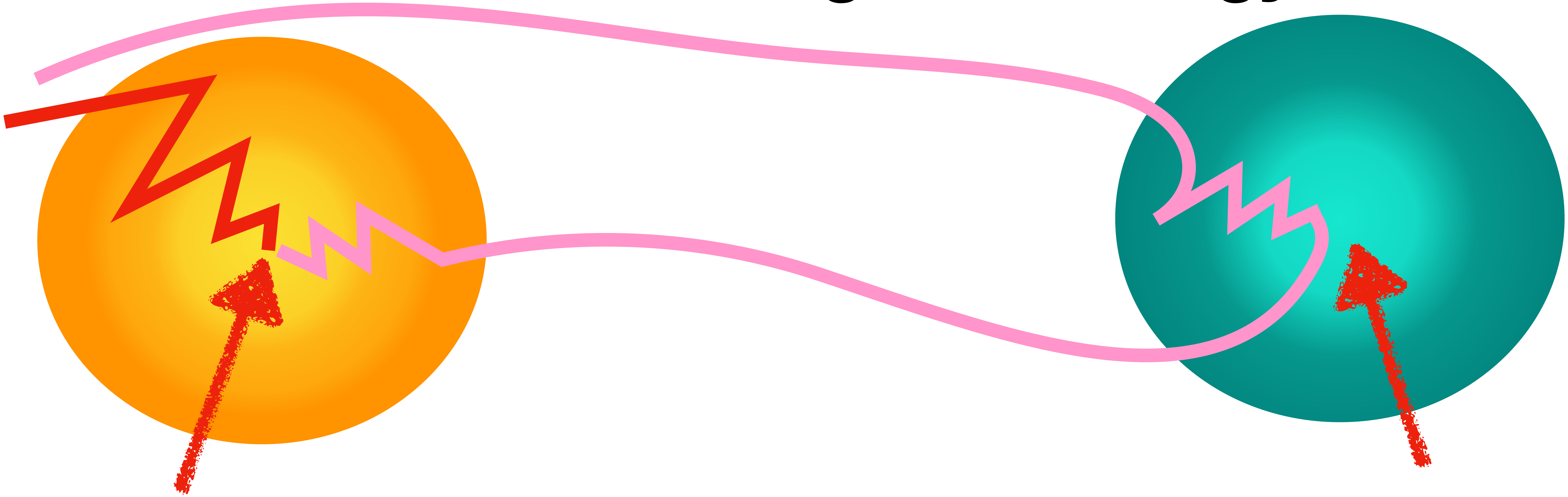
MLM loss
10^1
6×10^0
4×10^0
3×10^0
2×10^0
10^3
10^4
10^5
step
Internal syntax score
10^1
10^3
10^4
10^5
step
BERT-Base
SHS Suppressed

Suppressing SAS promotes a competing strategy

World's greatest
all-SAS model

Competing strategy

Can we recover the original strategy?



World's greatest
all-SAS model

Competing strategy

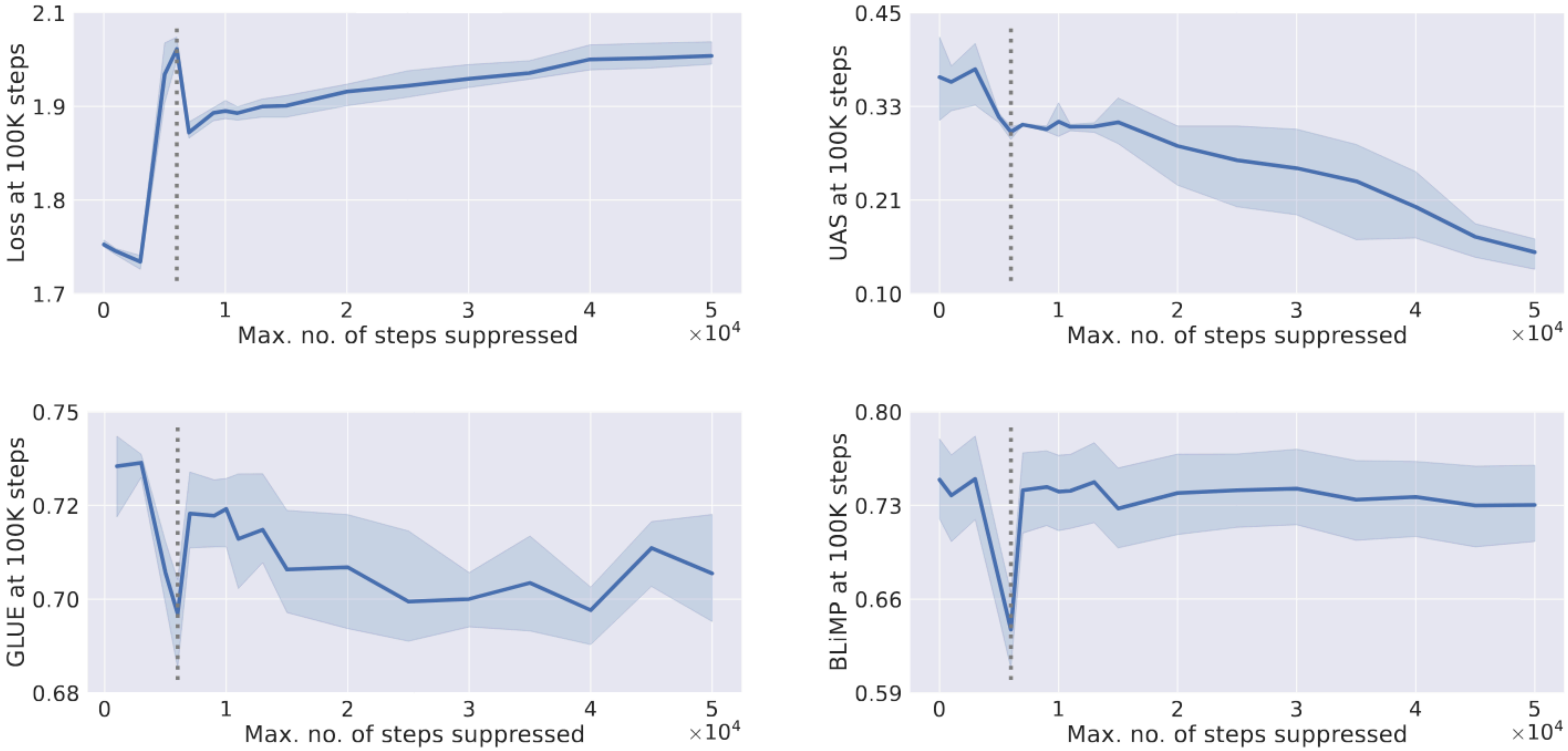
Multistage regularization

- Stage 1: Suppress SAS
- Stage 2: Stop suppressing SAS

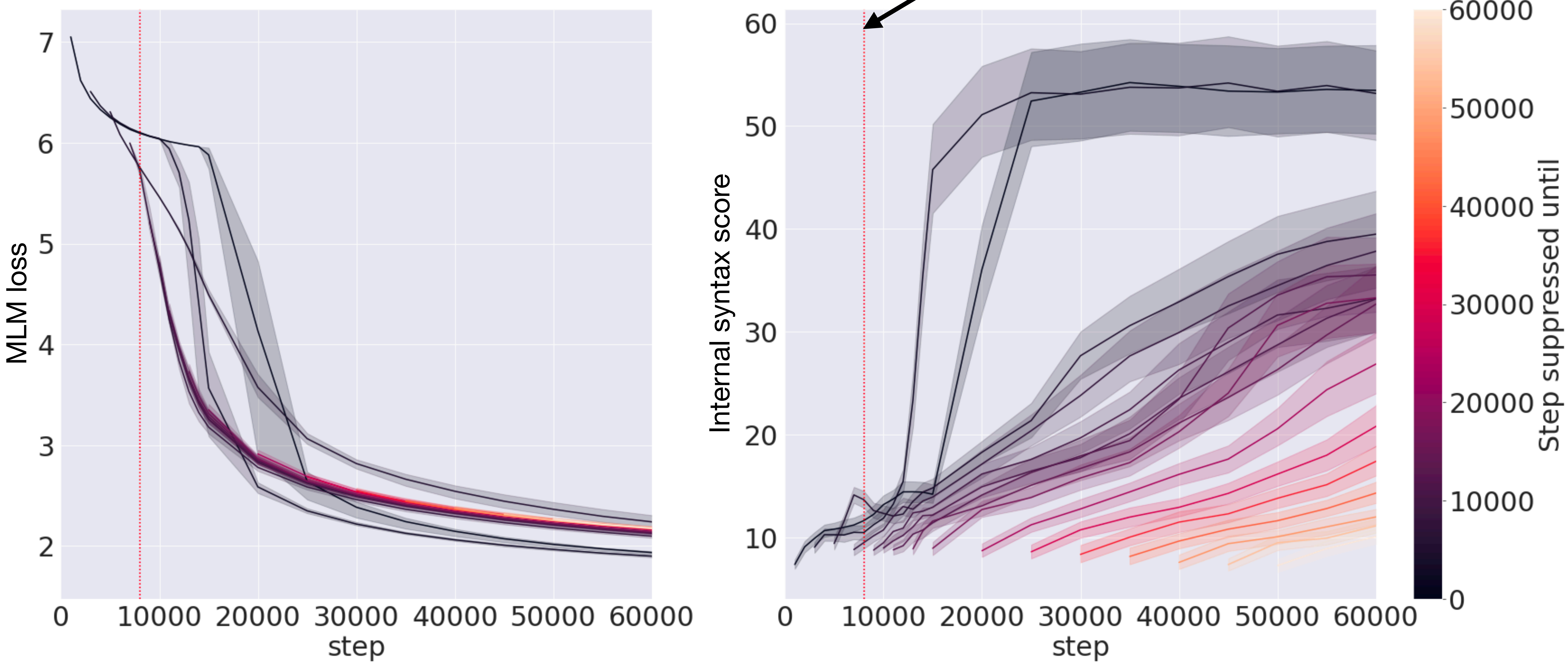
- Will we hit the original phase transition?

Every metric is *worst* when we release during the breakthrough

Why?



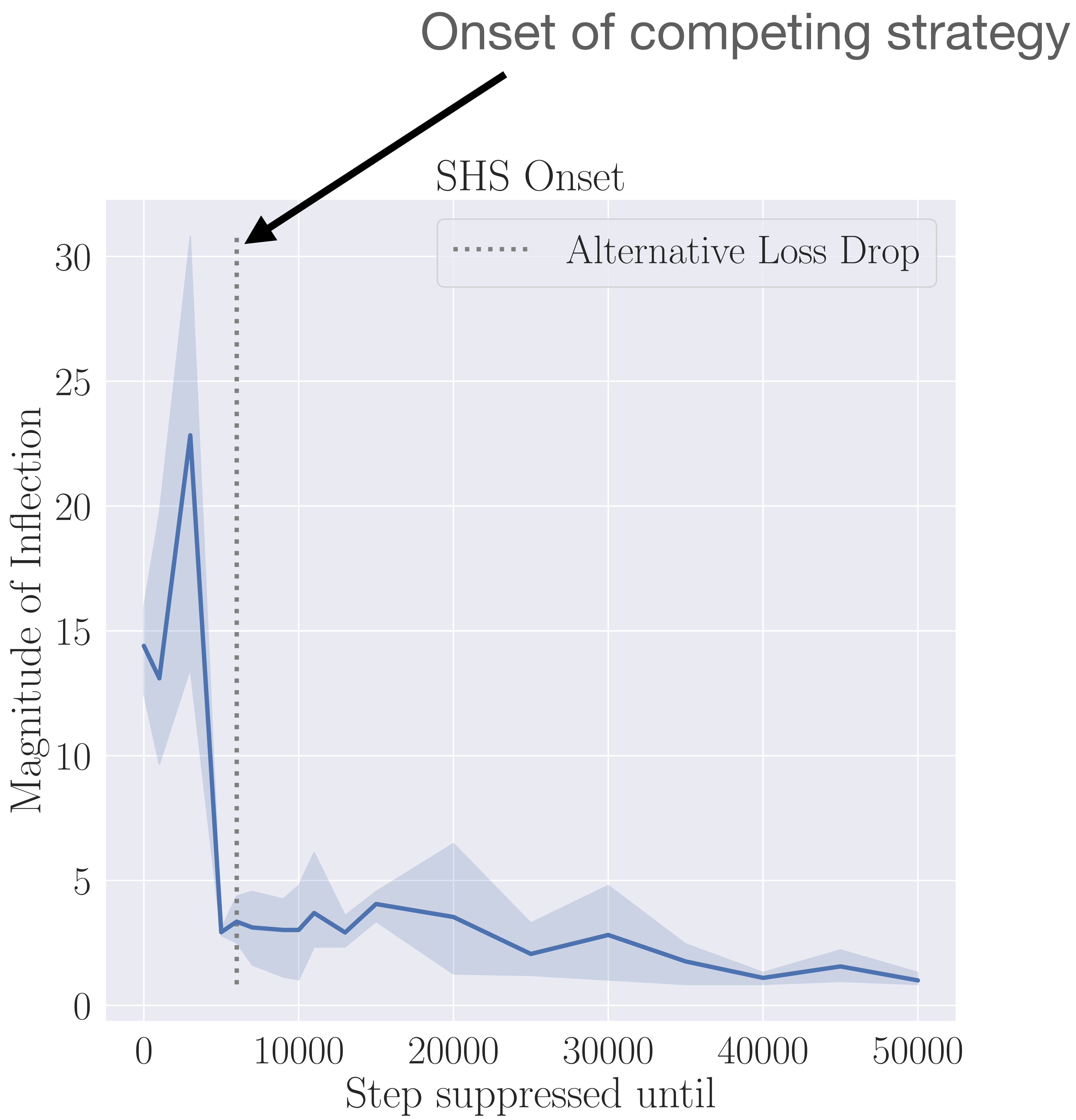
The longer we suppress SAS, the less SAS recovers



Onset of competing strategy

Syntactic Attention Structure onset magnitude

- Push past the competing strategy phase transition and we lose the SAS phase transition entirely!

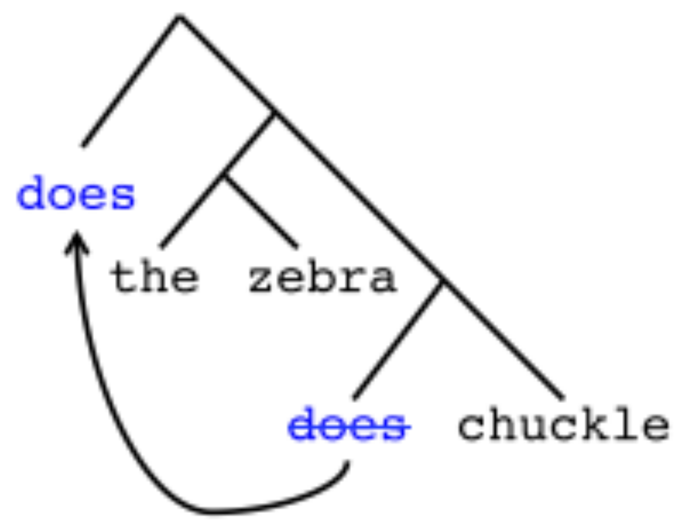


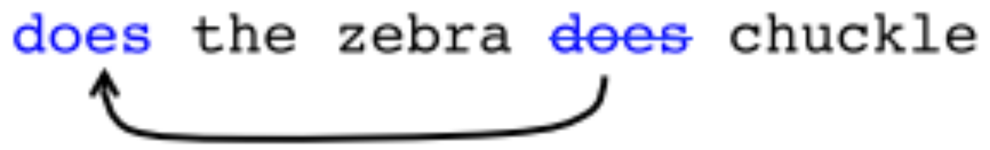
Once we transition to the competing strategy, the model can’t transition strategies back to Syntactic Attention Structure.

Conjecture: Phase changes are unstable?

Mystery #2: U-shaped stability

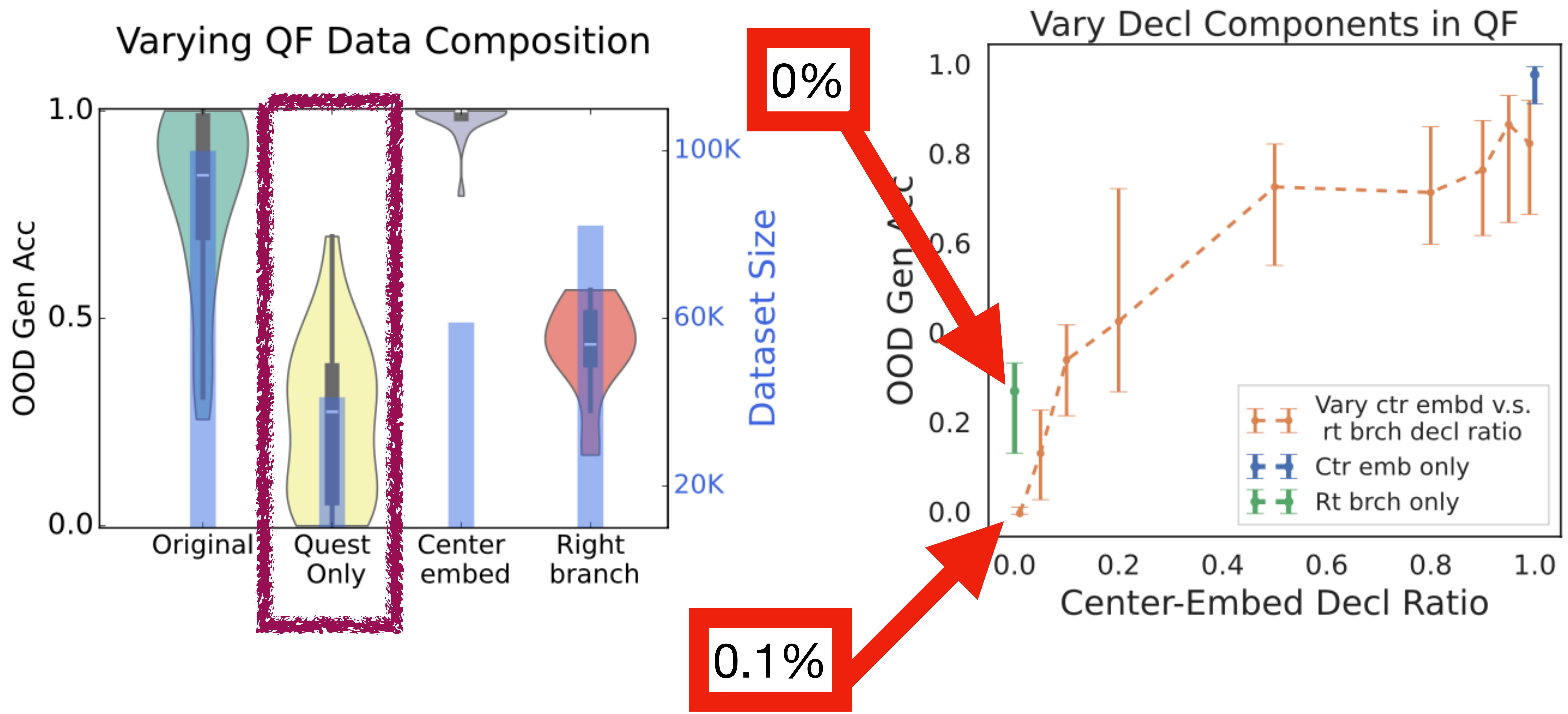
Why don't we learn linear rules from exclusively forward-branching data?





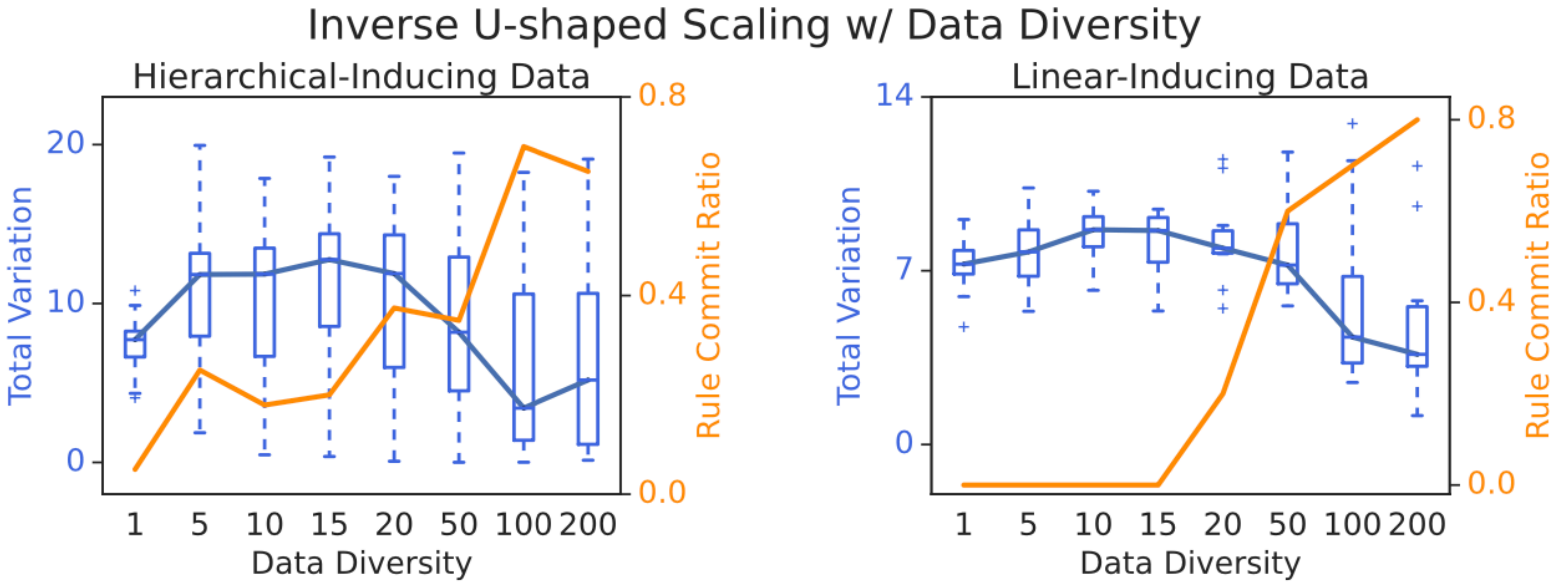
We need at least .1% center embeddings

0% yields high variance, but why?

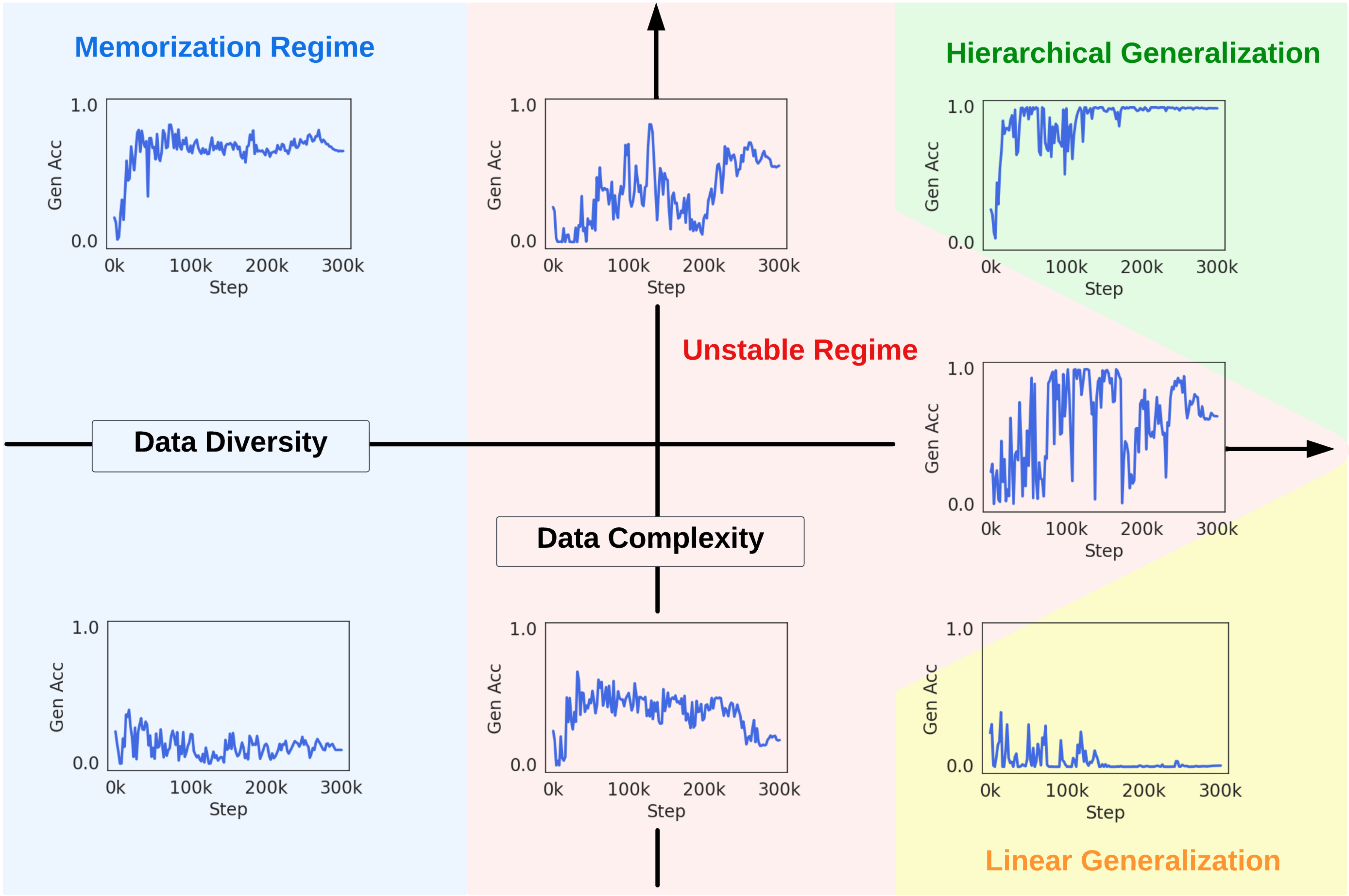


Forward-branching data isn't diverse enough!

Model oscillates between memorization and linear rule



The whole picture



Memorization Regime

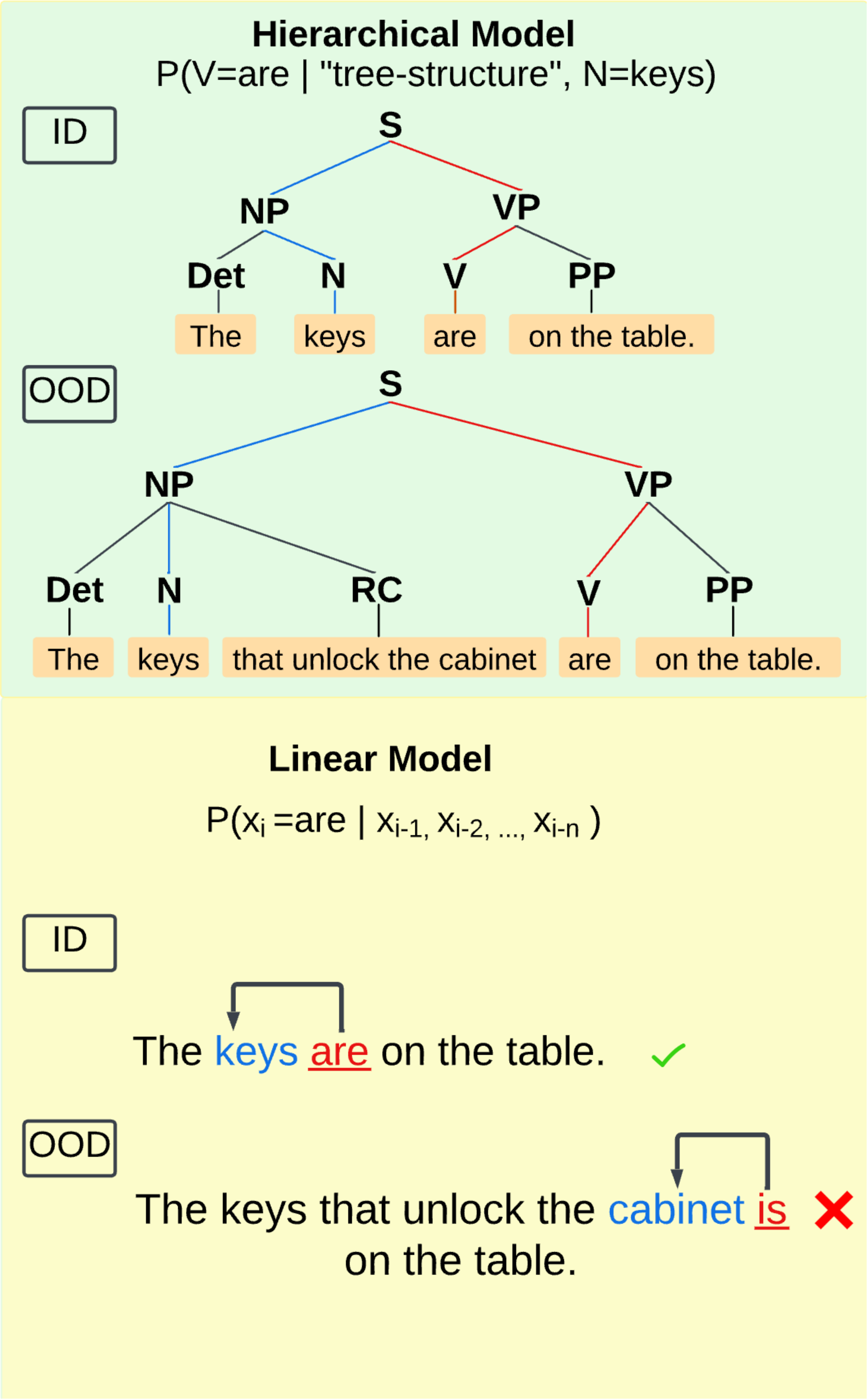
Unstable Regime

Hierarchical Generalization

Linear Generalization

Data Diversity

Data Complexity



Hierarchical Model

P(V=are | "tree-structure", N=keys)

Linear Model

$P(x_i = \text{are} \mid x_{i-1}, x_{i-2}, \ldots, x_{i-n})$

ID

The keys are on the table. ✓

OOD

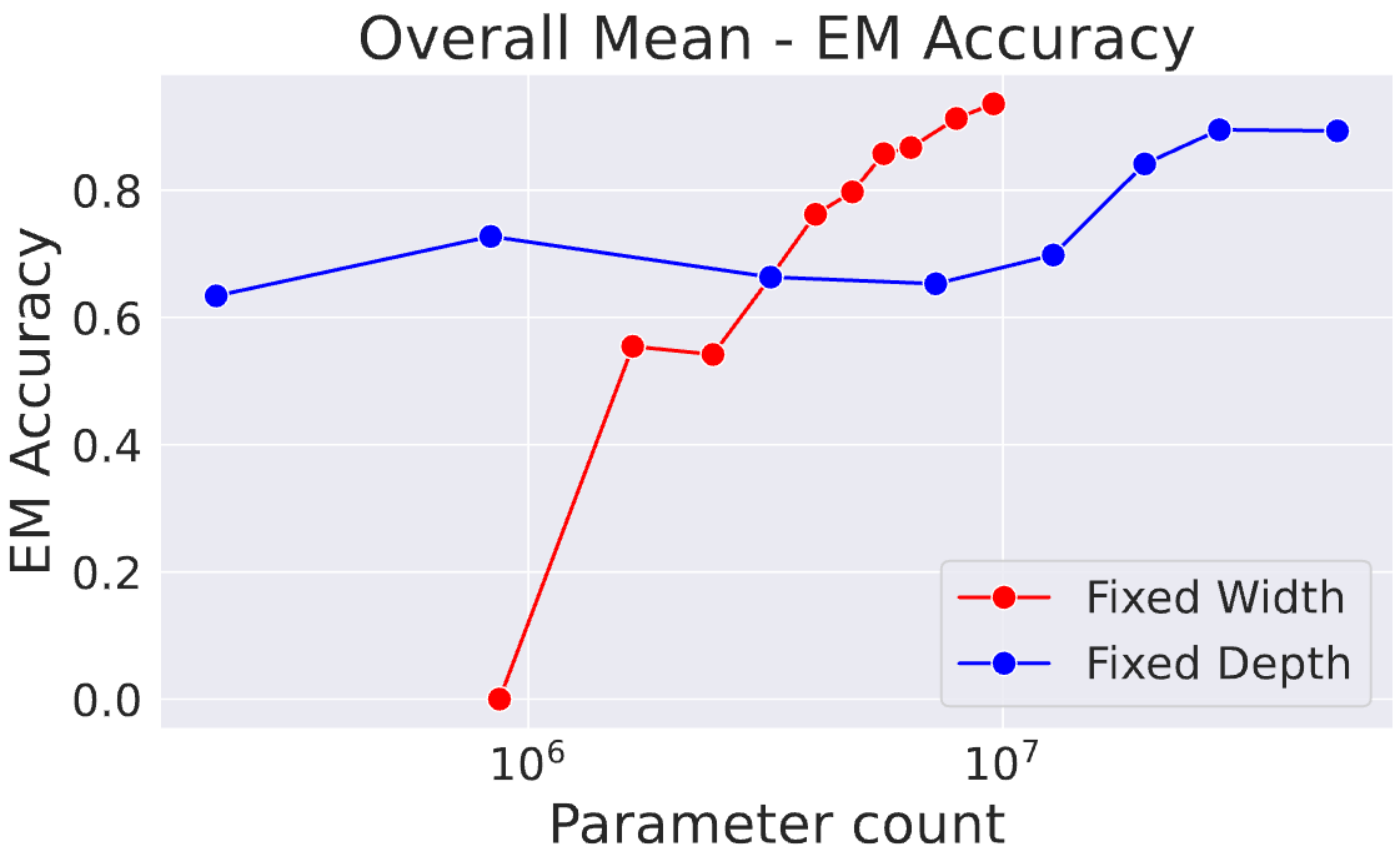
The keys that unlock the cabinet is ✗
on the table.

Confirmed: Memorization and rule-based generalization can also compete.

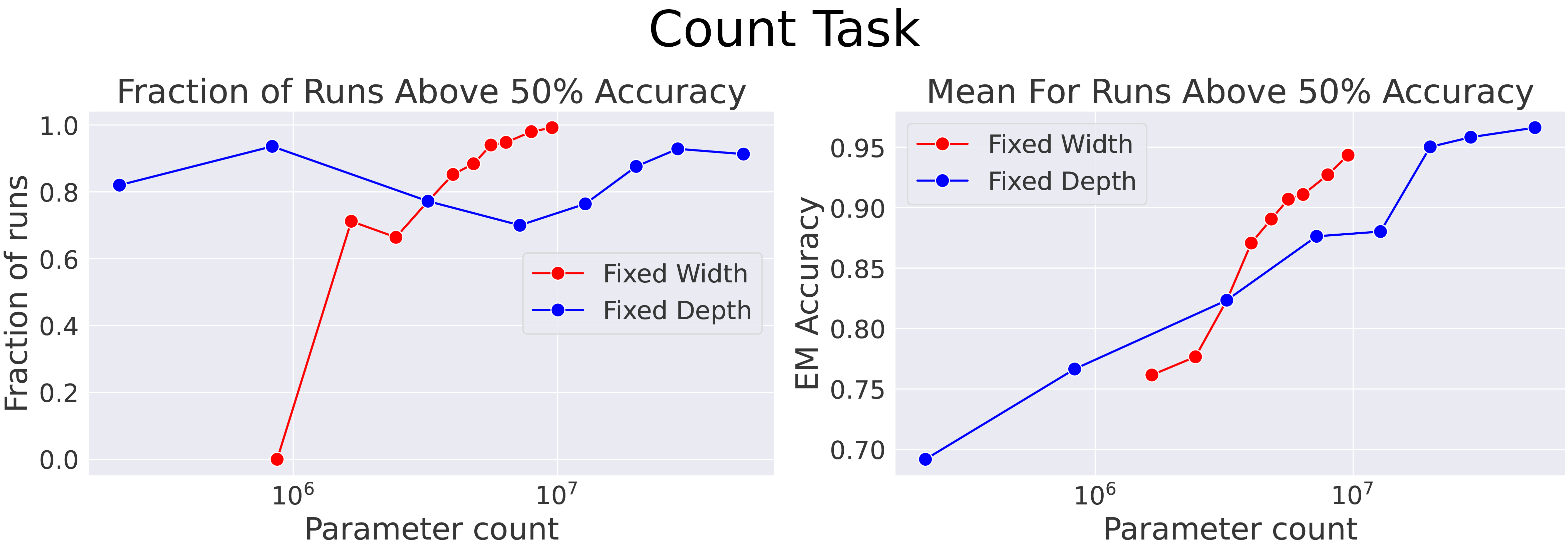
Mystery #3: U-shaped scaling laws

- Task: Length generalization in counting
  - 5, 9 >, 5, 6, 7, 8, 9
  - Train on 30, test on 40

Scaling width yields INVERSE scaling law



Only *probability* of emergence has inverse scaling

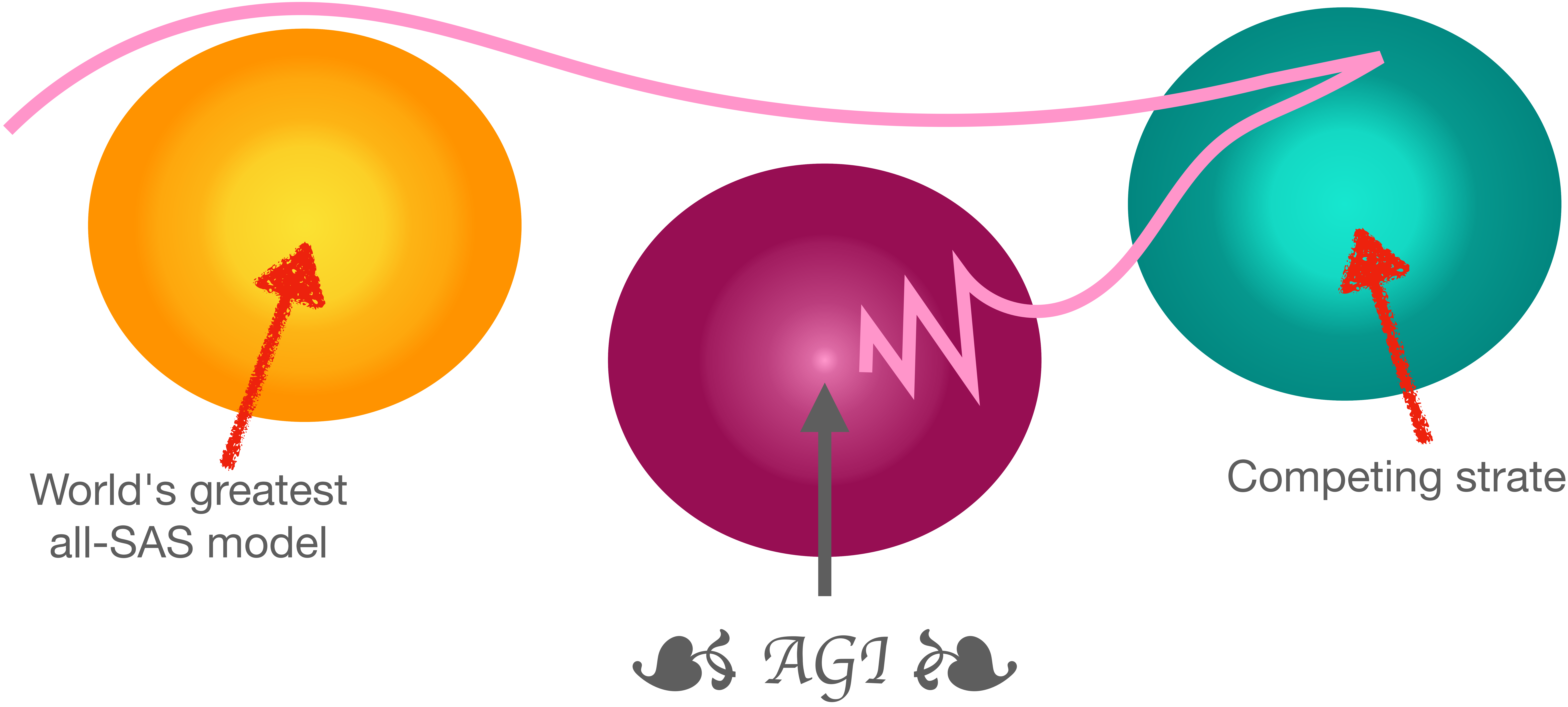


Conjecture: Sometimes scaling up can “buy” more potential parameters for the non-compositional circuit?

Complete understanding will include U-shaped curves, not just emergence.

Questions?

We find a viable alternative strategy!


World's greatest
all-SAS model
☙ AGI ❧
Competing strategy