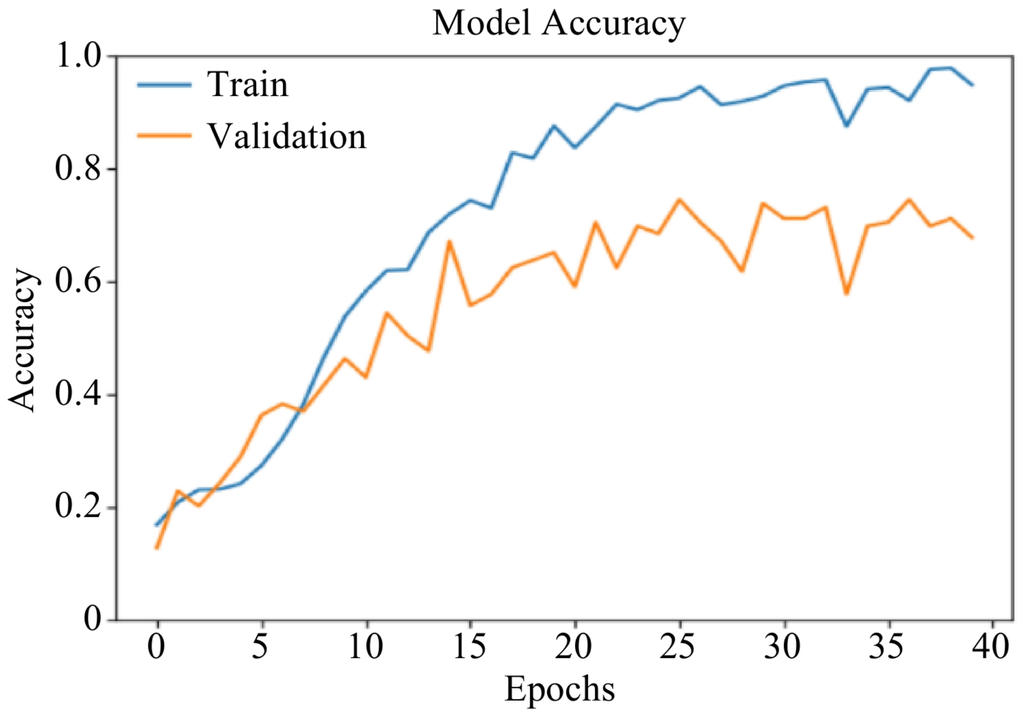
REFORM 2/26: Grokking

Charlotte Peale, Silvia Casacuberta

What is “grokking”?

https://towardsdatascience.com/i-used-to-hate-overfitting-but-now-im-grokking-it-e6e1dcfbdfd8/

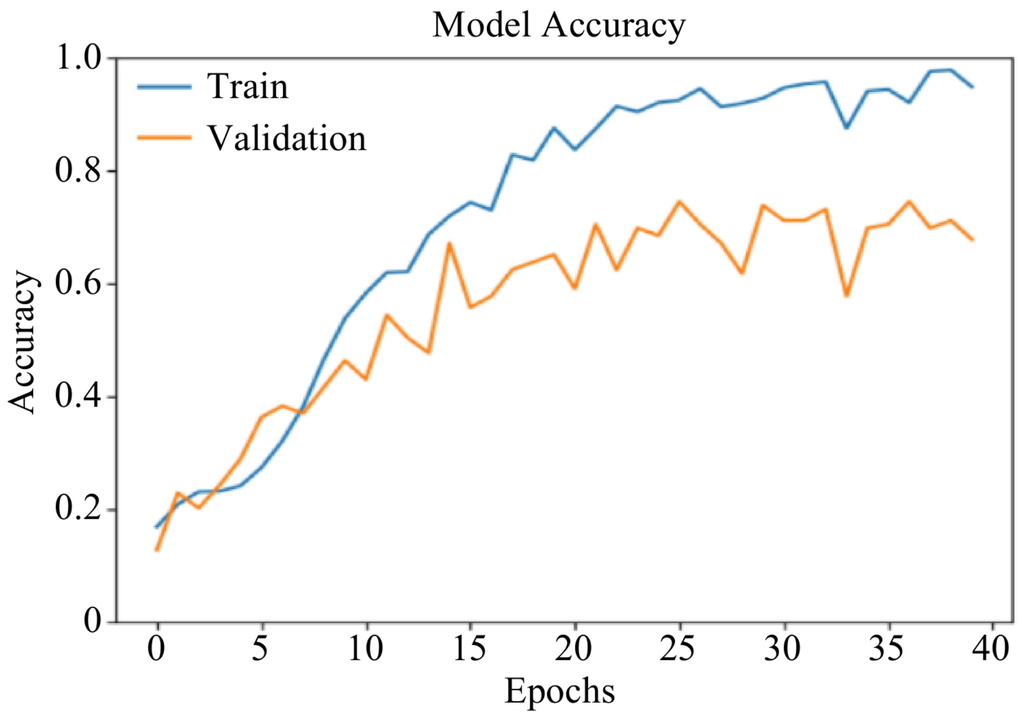
What is “grokking”?



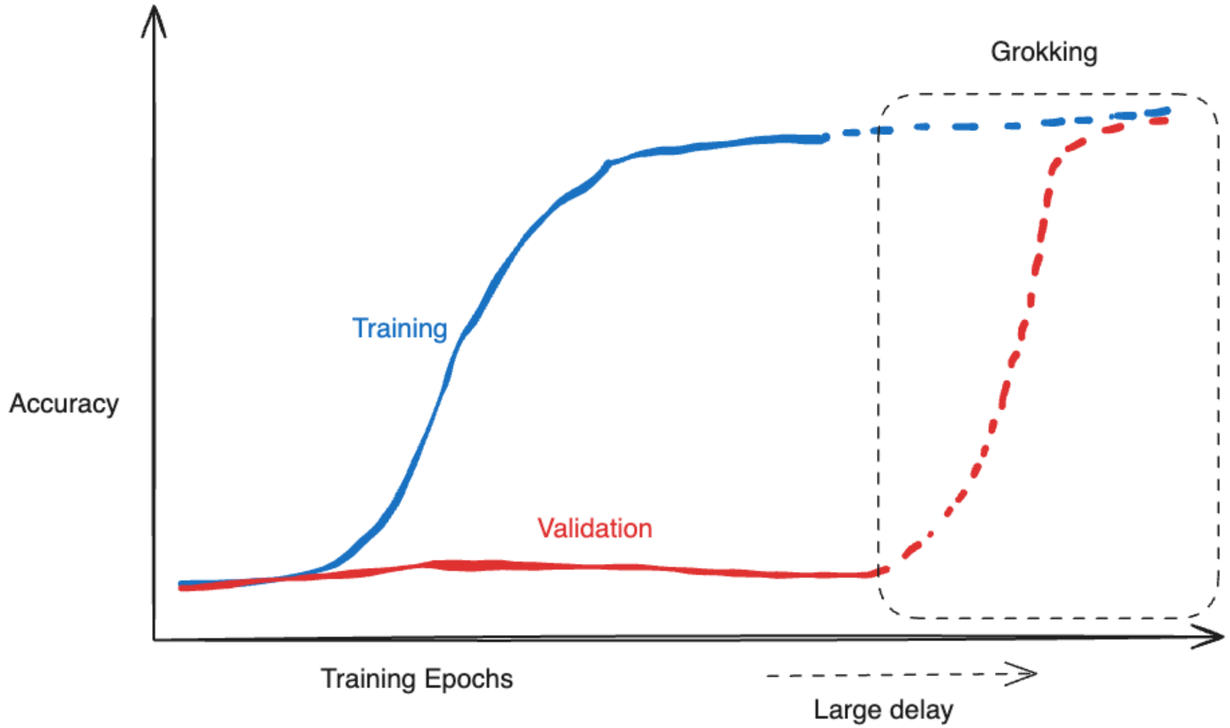
Typical Training Dynamics

https://towardsdatascience.com/i-used-to-hate-overfitting-but-now-im-grokking-it-e6e1dcfbdfd8/

What is “grokking”?



Typical Training Dynamics

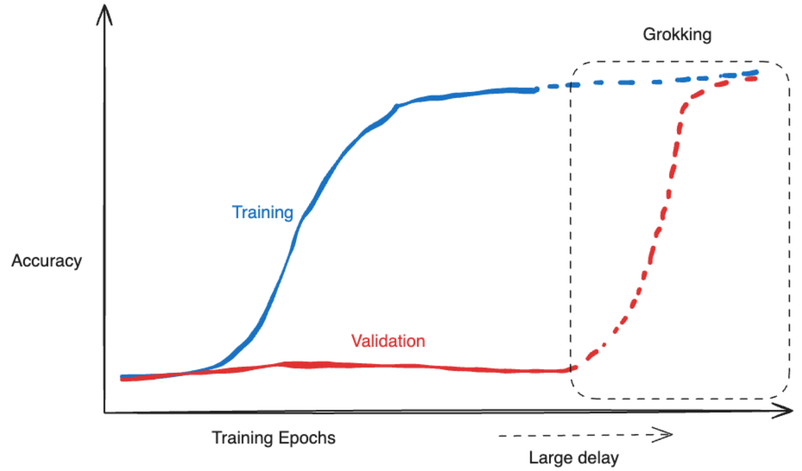


Grokking

https://towardsdatascience.com/i-used-to-hate-overfitting-but-now-im-grokking-it-e6e1dcfbdfd8/

What is “grokking”?

- Observed in transformers
- Structured/algorithmic learning tasks



Task-of-interest for today: **modular addition**

$$a, b \in \{0,1,...,P-1\} \longrightarrow (a+b) \mod P$$

https://towardsdatascience.com/i-used-to-hate-overfitting-but-now-im-grokking-it-e6e1dcfbdfd8/

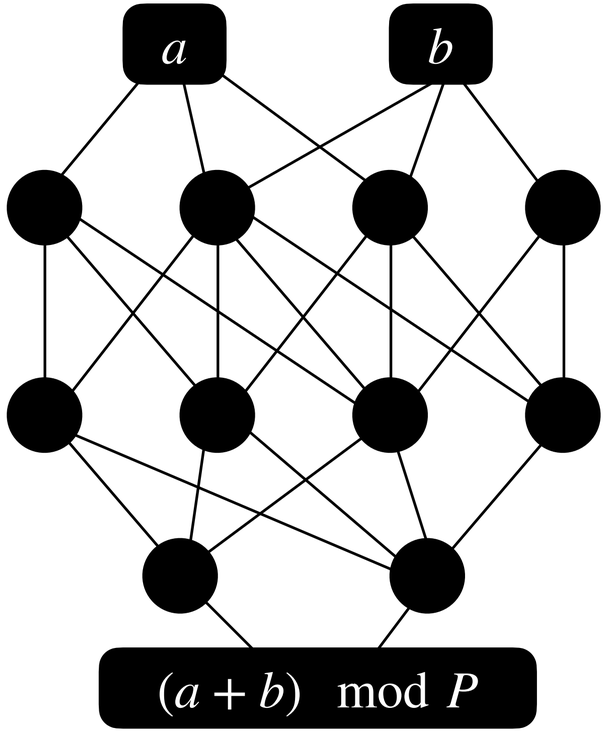
Why does grokking happen?

- **Current hypothesis:** competition between memorization and generalization.
  Explaining grokking through circuit efficiency, Varma et al. 2023

a, b	f(a,b)
1, 2	3
45, 34	79
23, 6	29
17, 21	38
51, 10	61

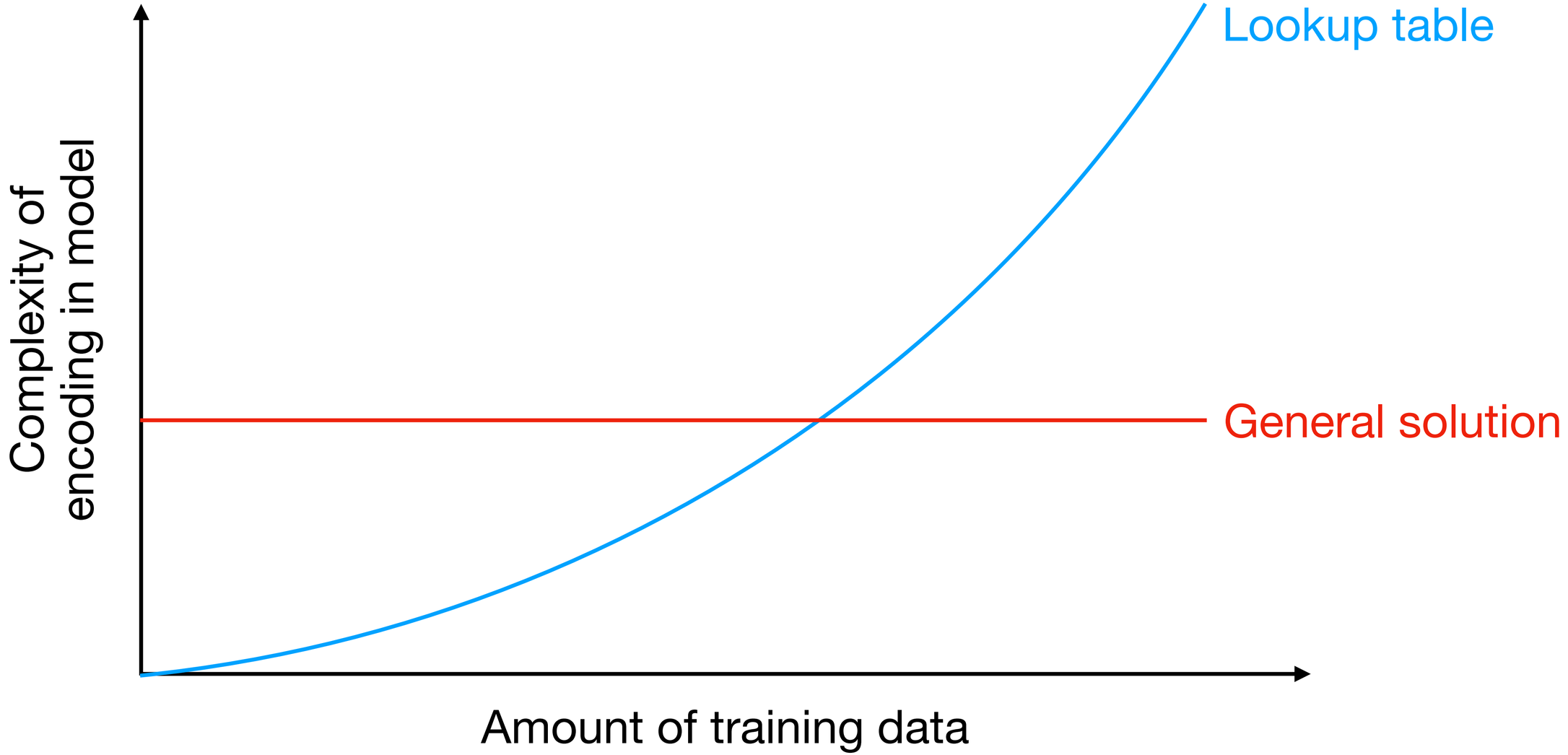
lookup table for training data

vs

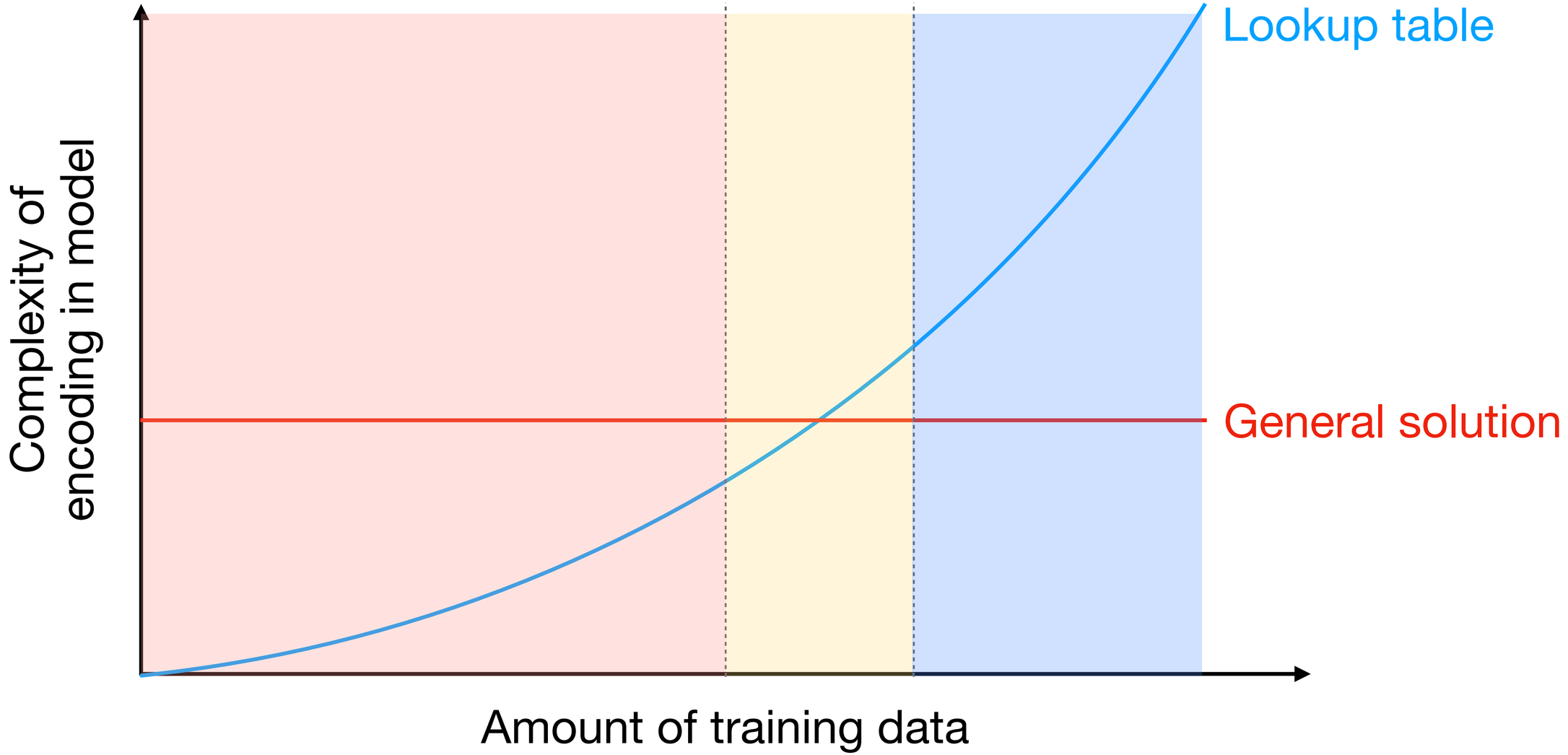


General circuit for addition

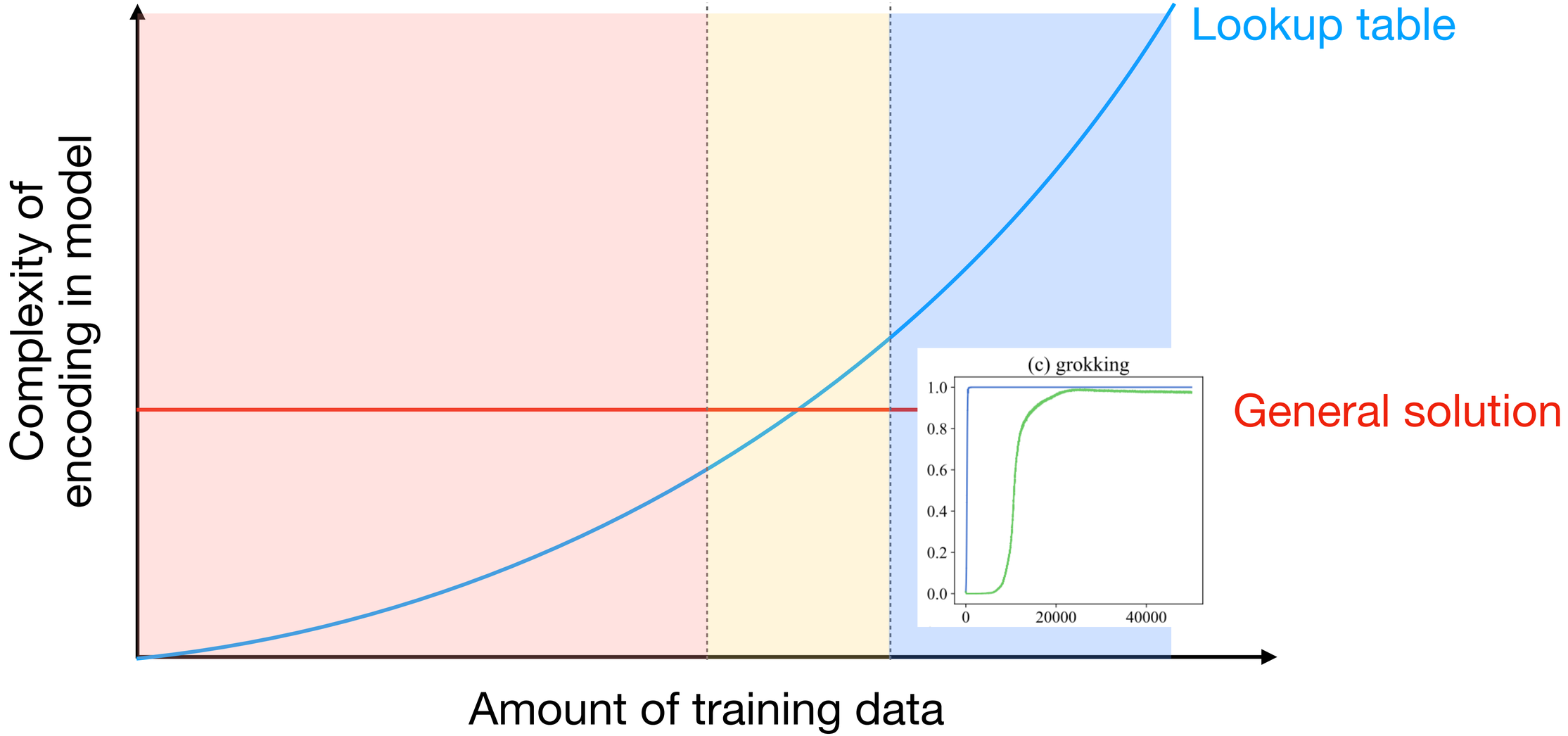
Memorization vs Generation Complexity



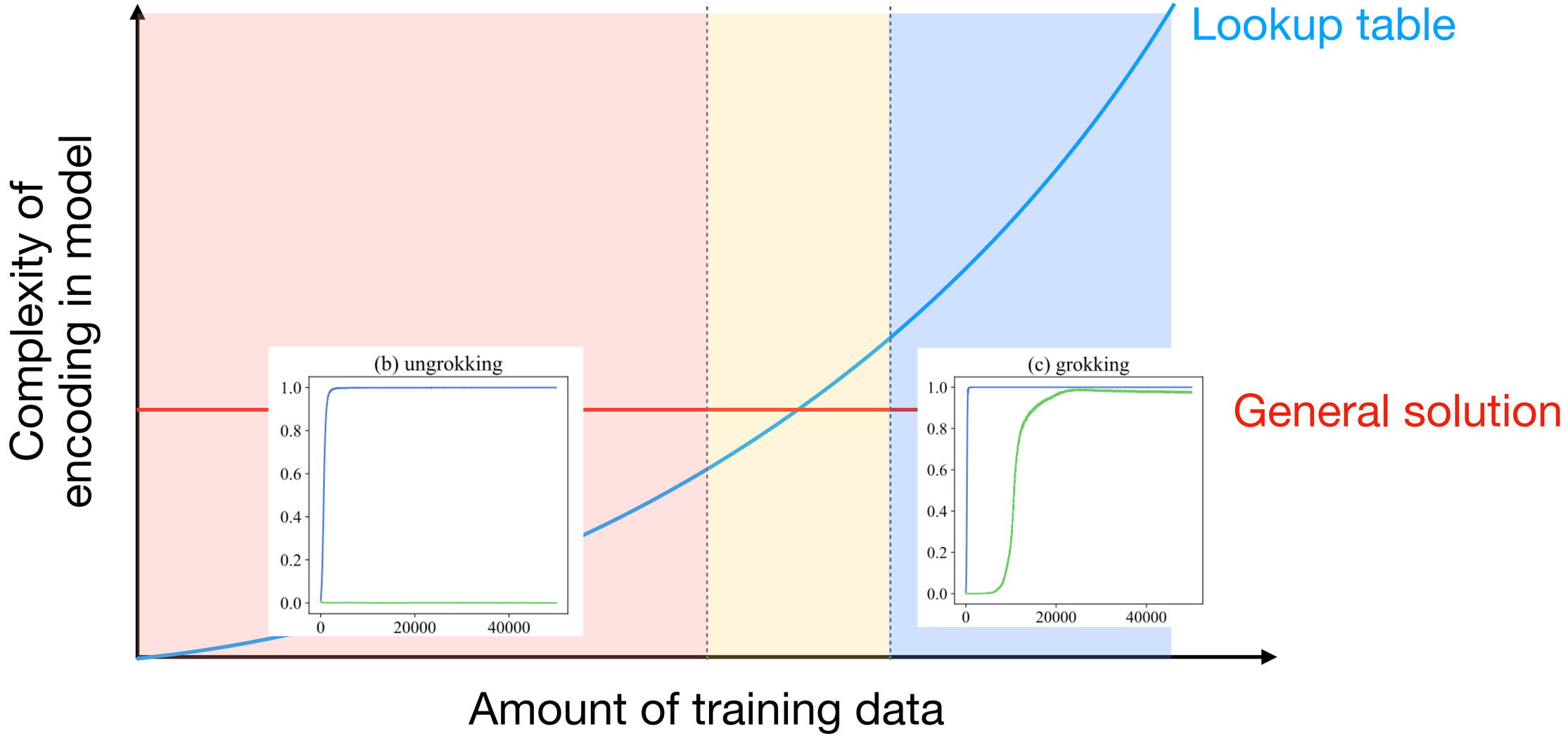
Memorization vs Generation Complexity



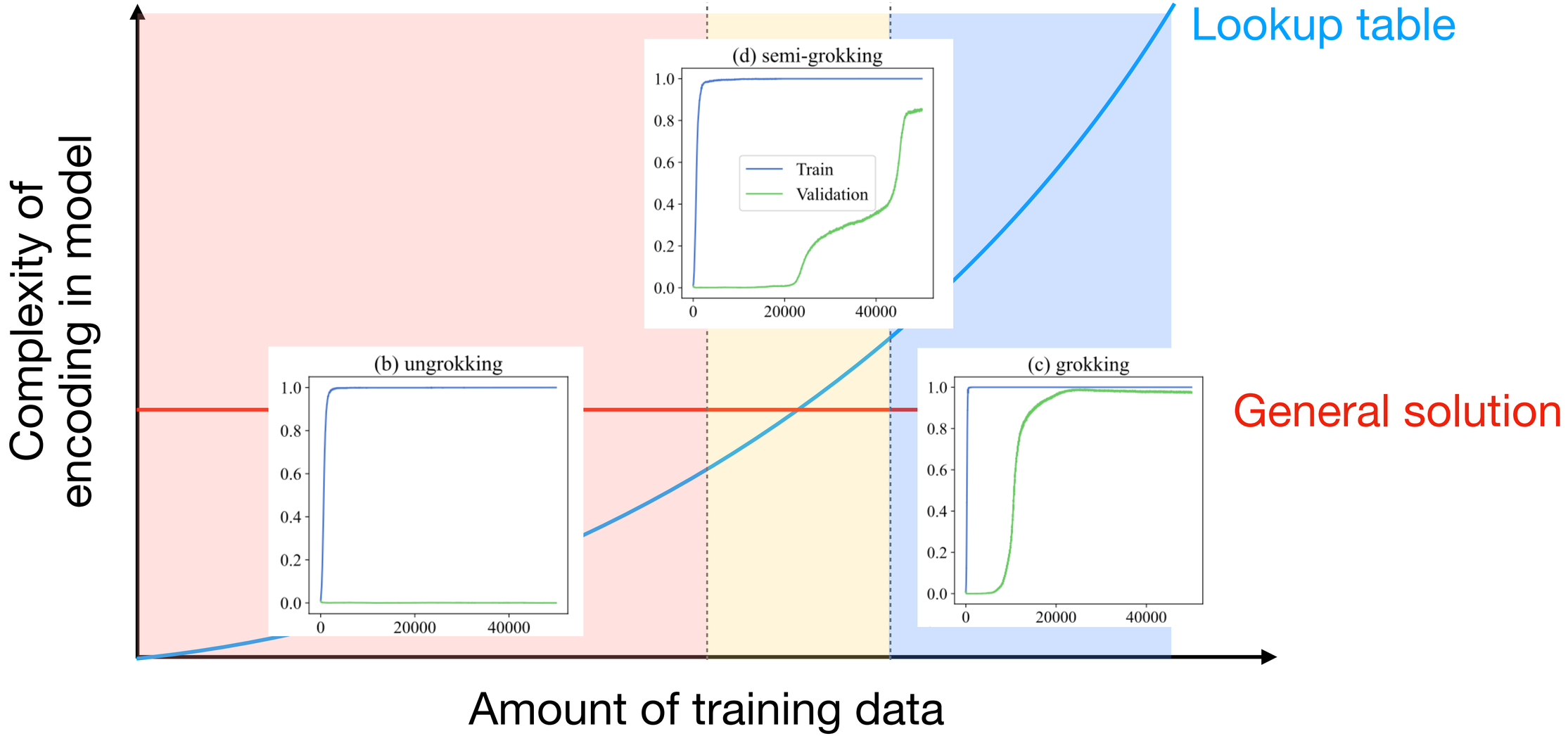
Memorization vs Generation Complexity



Memorization vs Generation Complexity



Memorization vs Generation Complexity



Unified View of Grokking, Double Descent and Emergent Abilities: A Perspective from Circuits Competition

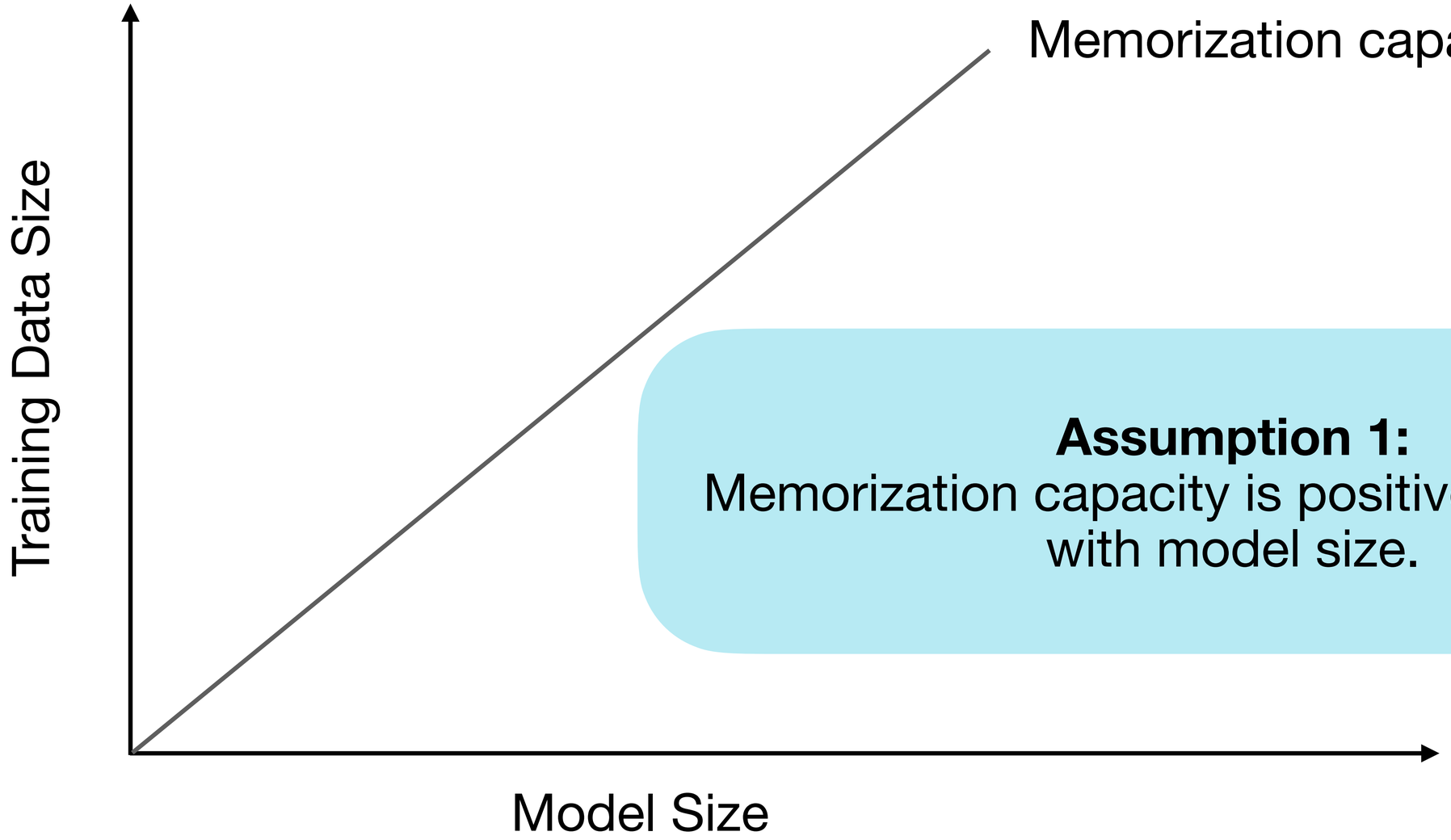
Yufei Huang [1] **Shengding Hu** [1] **Xu Han** [1] **Zhiyuan Liu** [1] **Maosong Sun**[† 1]

1. Map out landscape of training dynamics varying train set size *and* model size.

2. Use landscape to give an explanation for double descent.

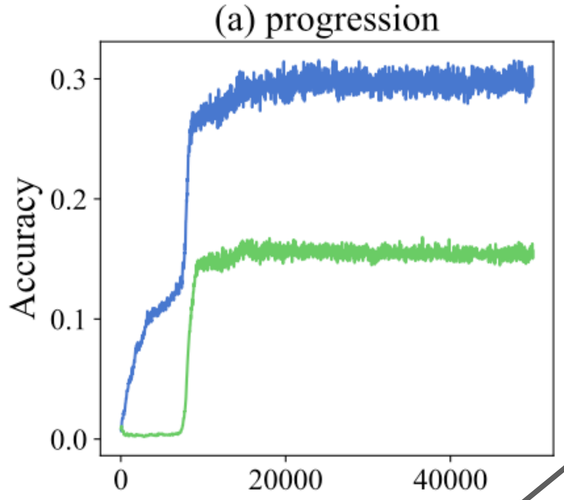
3. [Look at how emergent abilities relate to grokking]

Memorization and Model Size



Assumption 1:
Memorization capacity is positively correlated with model size.

Memorization and Model Size



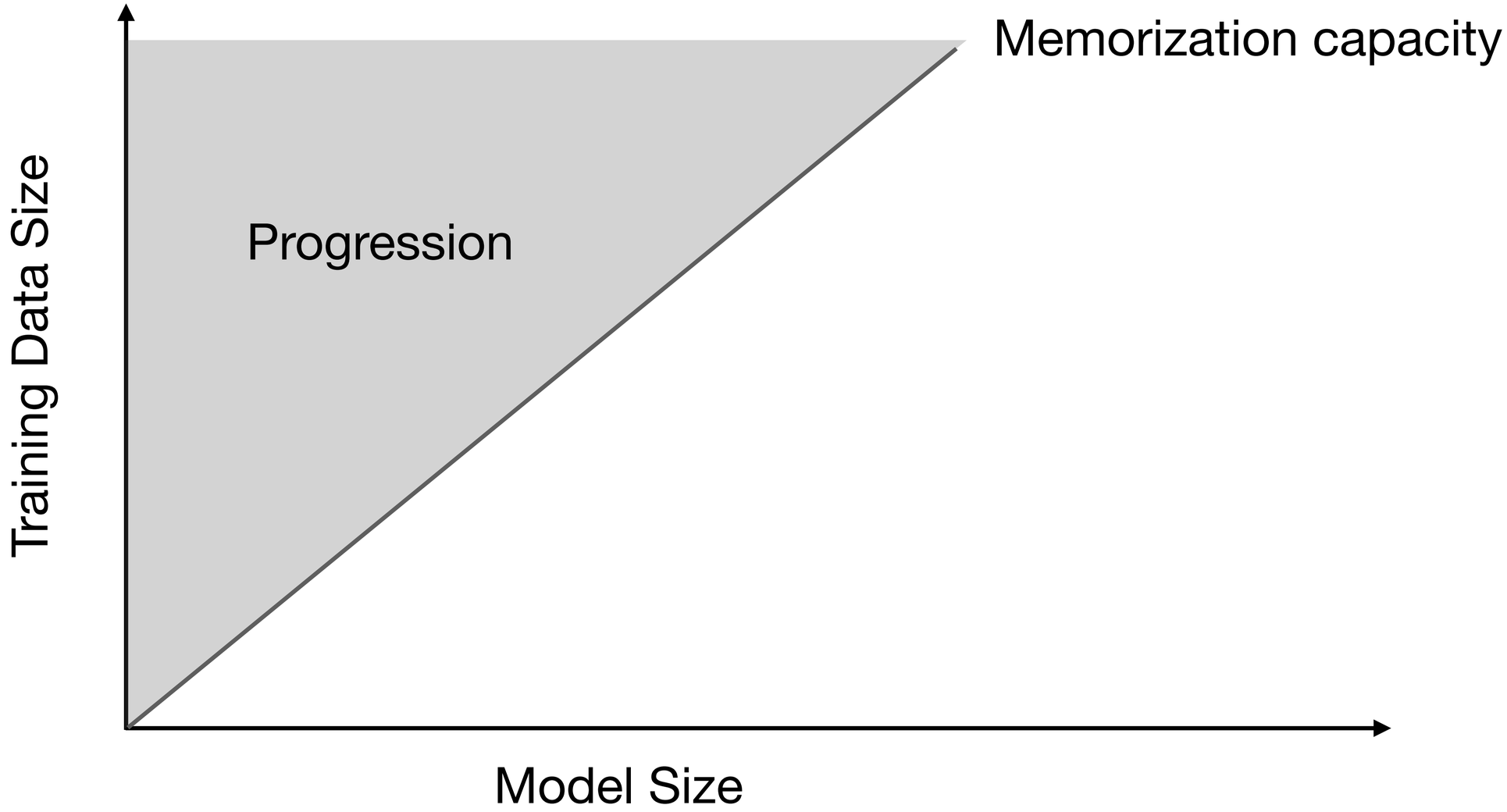
Memorization capacity

Training Data Size

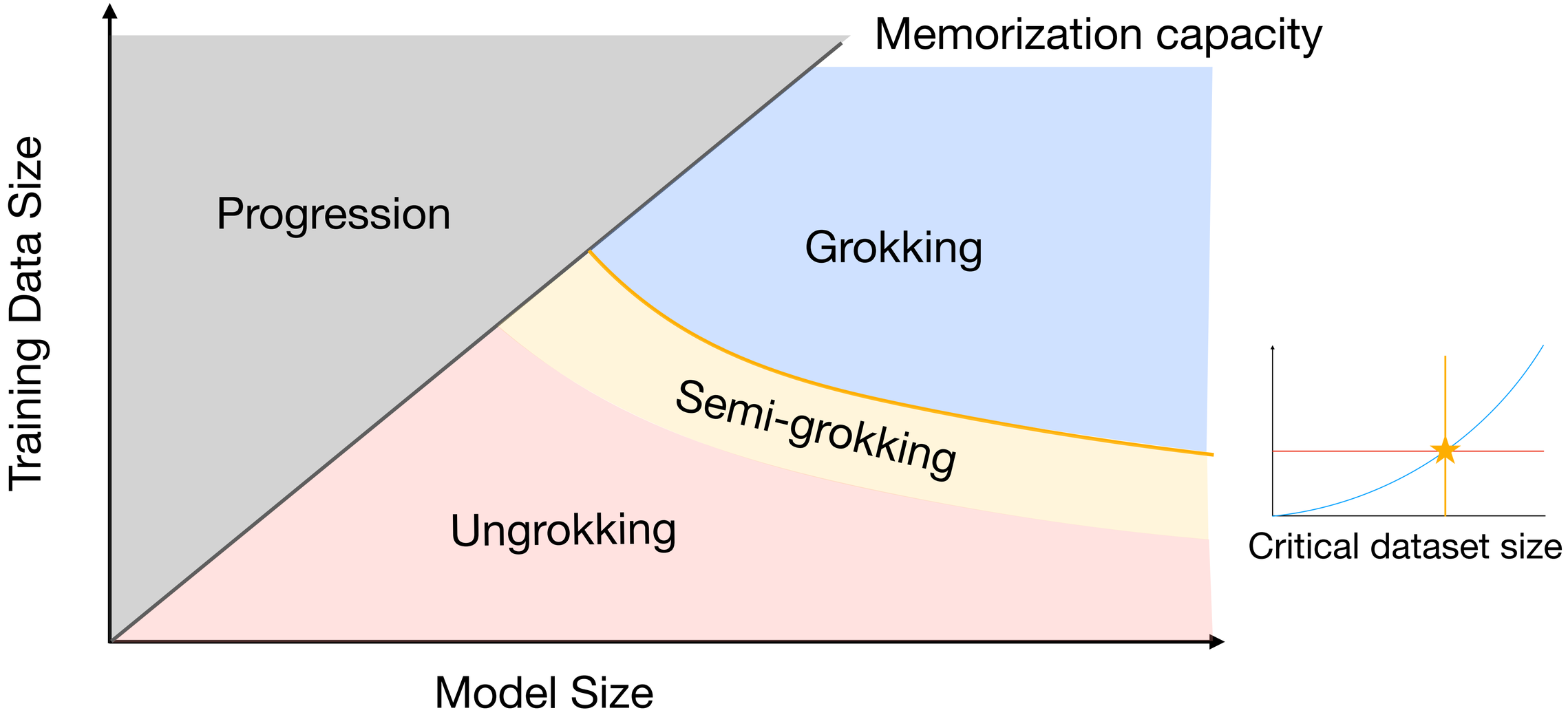
Model Size

Assumption 1:
Memorization capacity is positively correlated with model size.

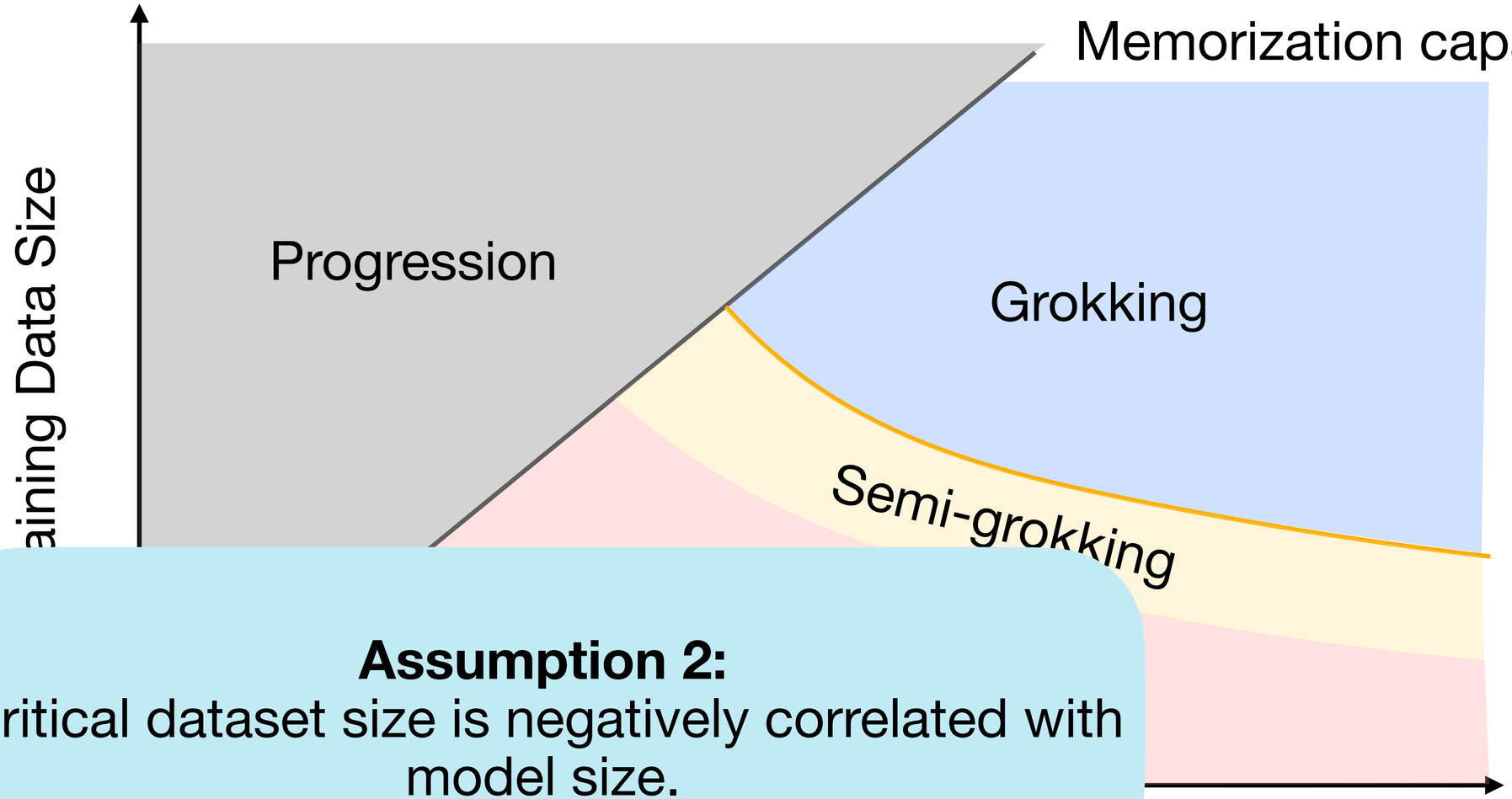
Grokking and Critical Dataset Size

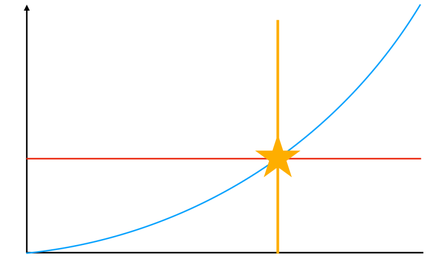


Grokking and Critical Dataset Size



Grokking and Critical Dataset Size

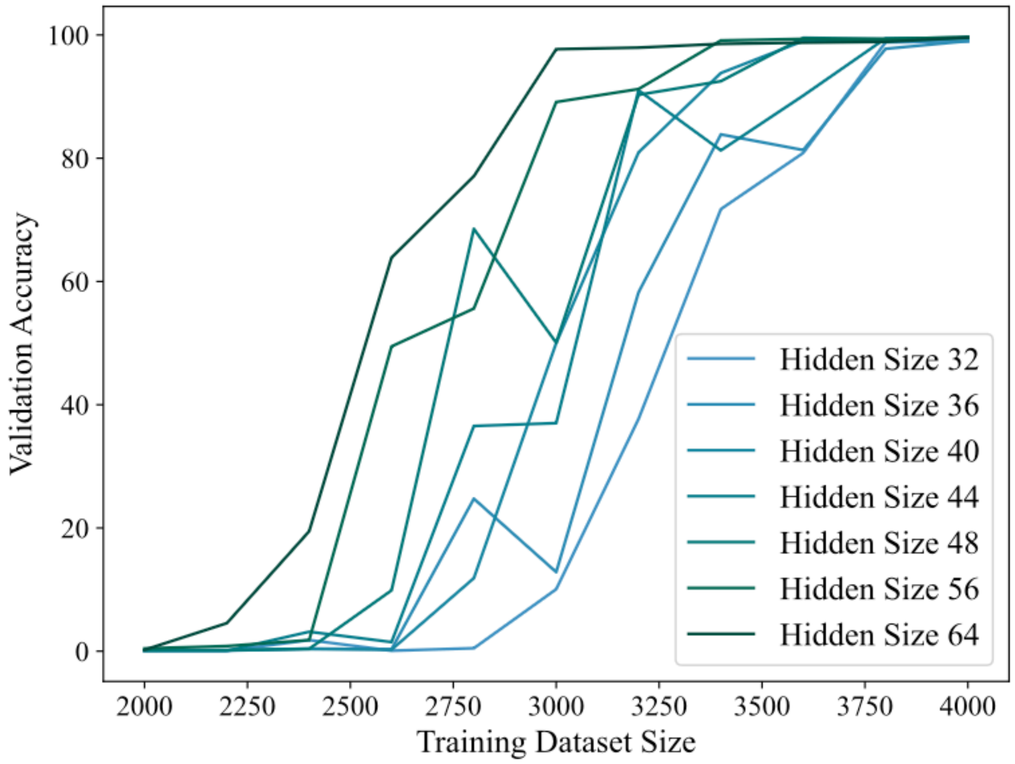




Critical dataset size

Assumption 2:
Critical dataset size is negatively correlated with model size.

Critical Dataset Size and Model Size



Assumption 2:
Critical dataset size is negatively correlated with model size.

Empirical Justification: Larger models achieve perfect validation accuracy with smaller train datasets.

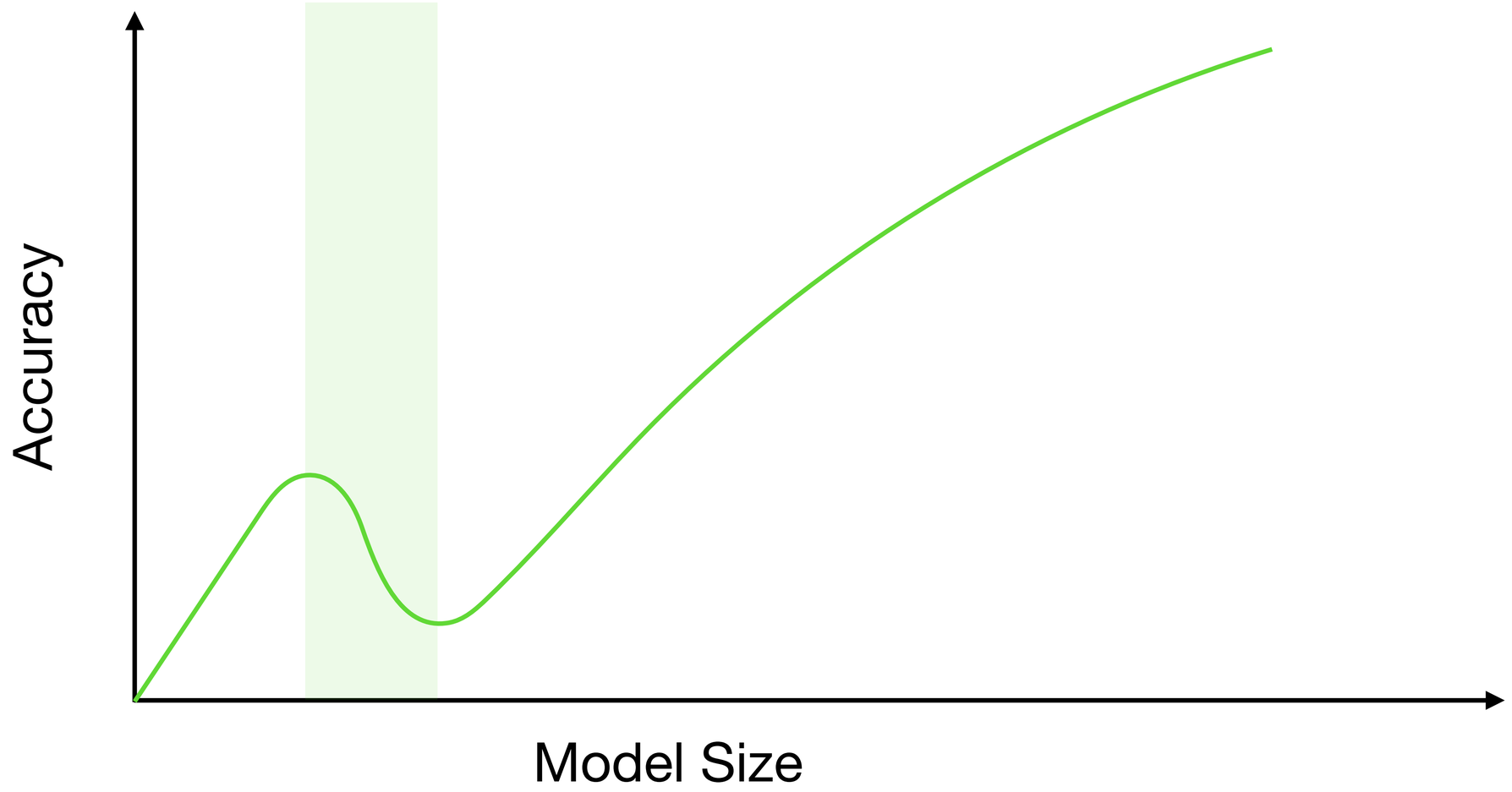
Unified View of Grokking, Double Descent and Emergent Abilities: A Perspective from Circuits Competition

Yufei Huang [1] Shengding Hu [1] Xu Han [1] Zhiyuan Liu [1] Maosong Sun† [1]

1. Map out landscape of training dynamics varying train set size *and* model size.
2. Use landscape to give an explanation for **double descent**.
3. [Look at how emergent abilities relate to grokking]

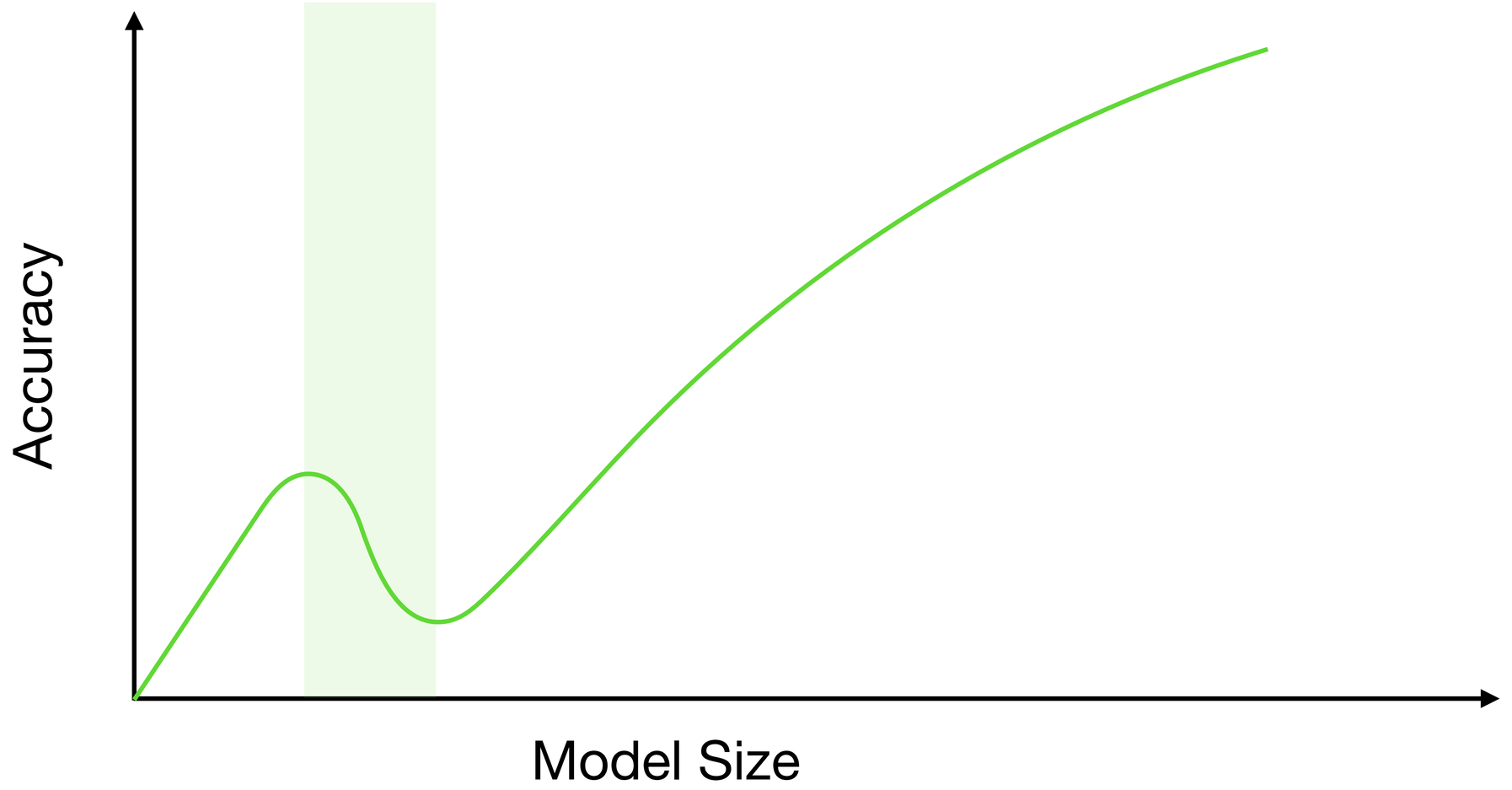
Refresher: Double Descent

- Temporary drop in accuracy as model size increases.



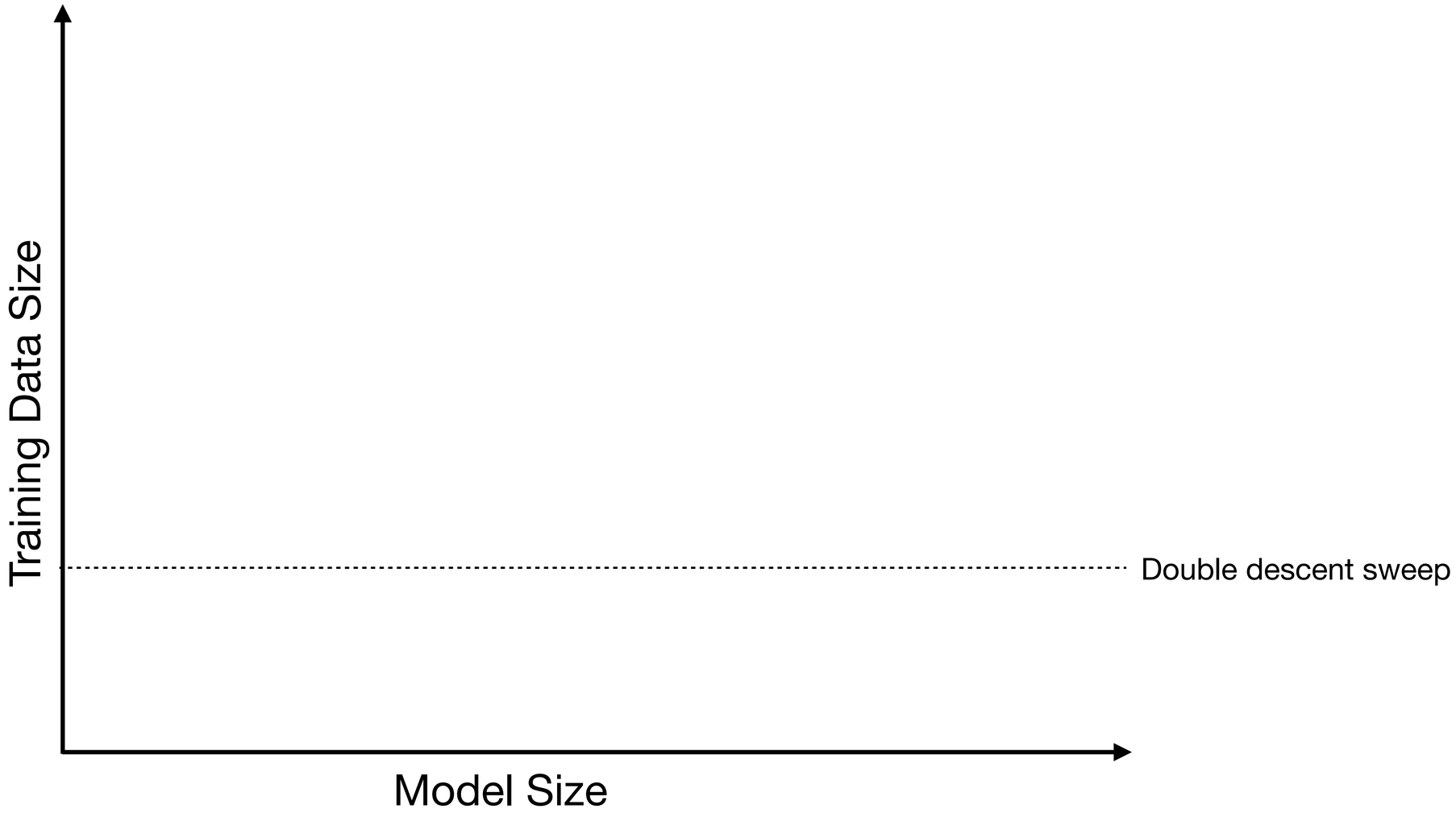
Double Descent and Grokking Landscape

- Double descent plots a fixed train set size while varying model size.



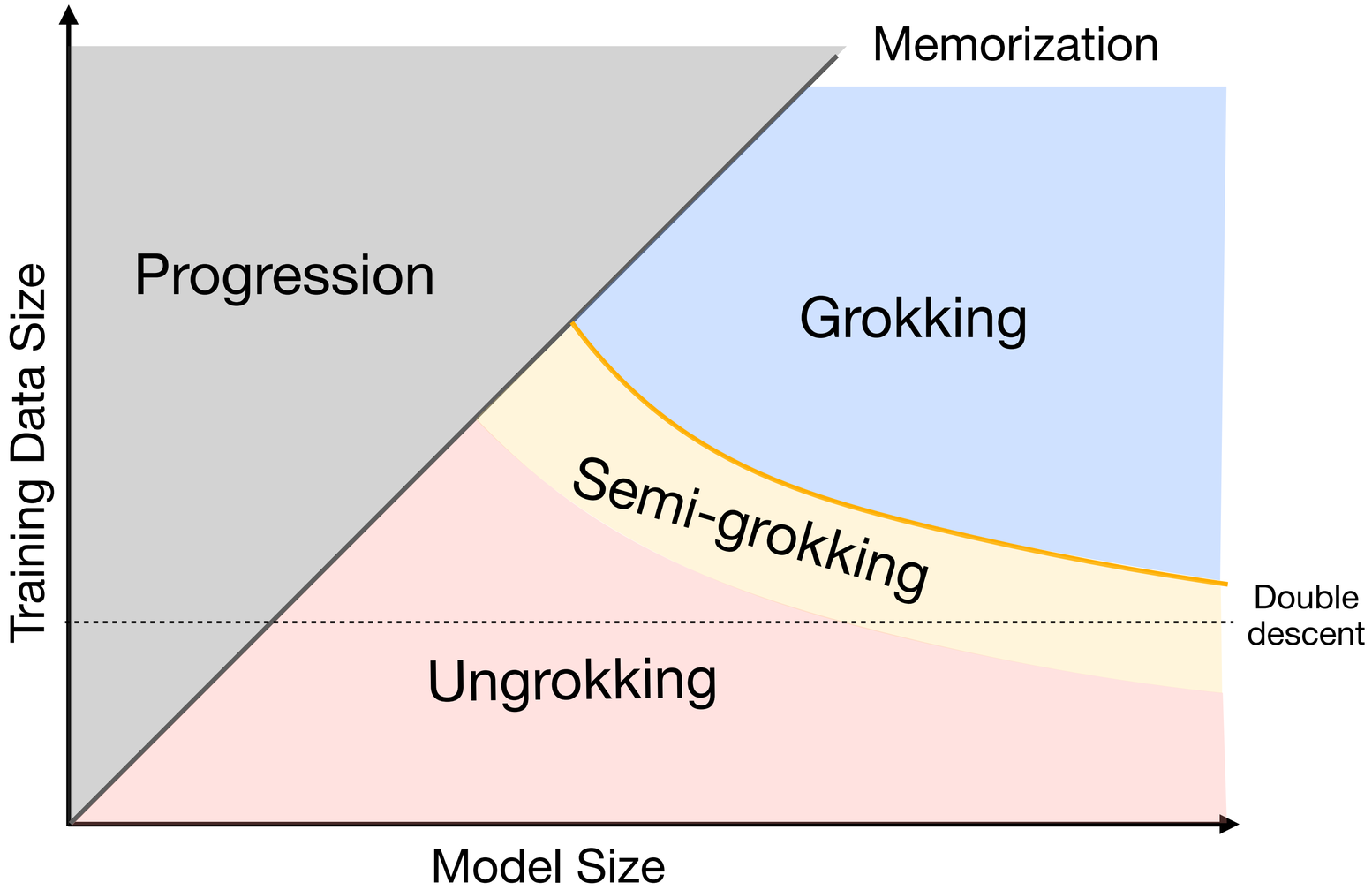
Double Descent and Grokking Landscape

- Double descent plots a fixed train set size while varying model size.



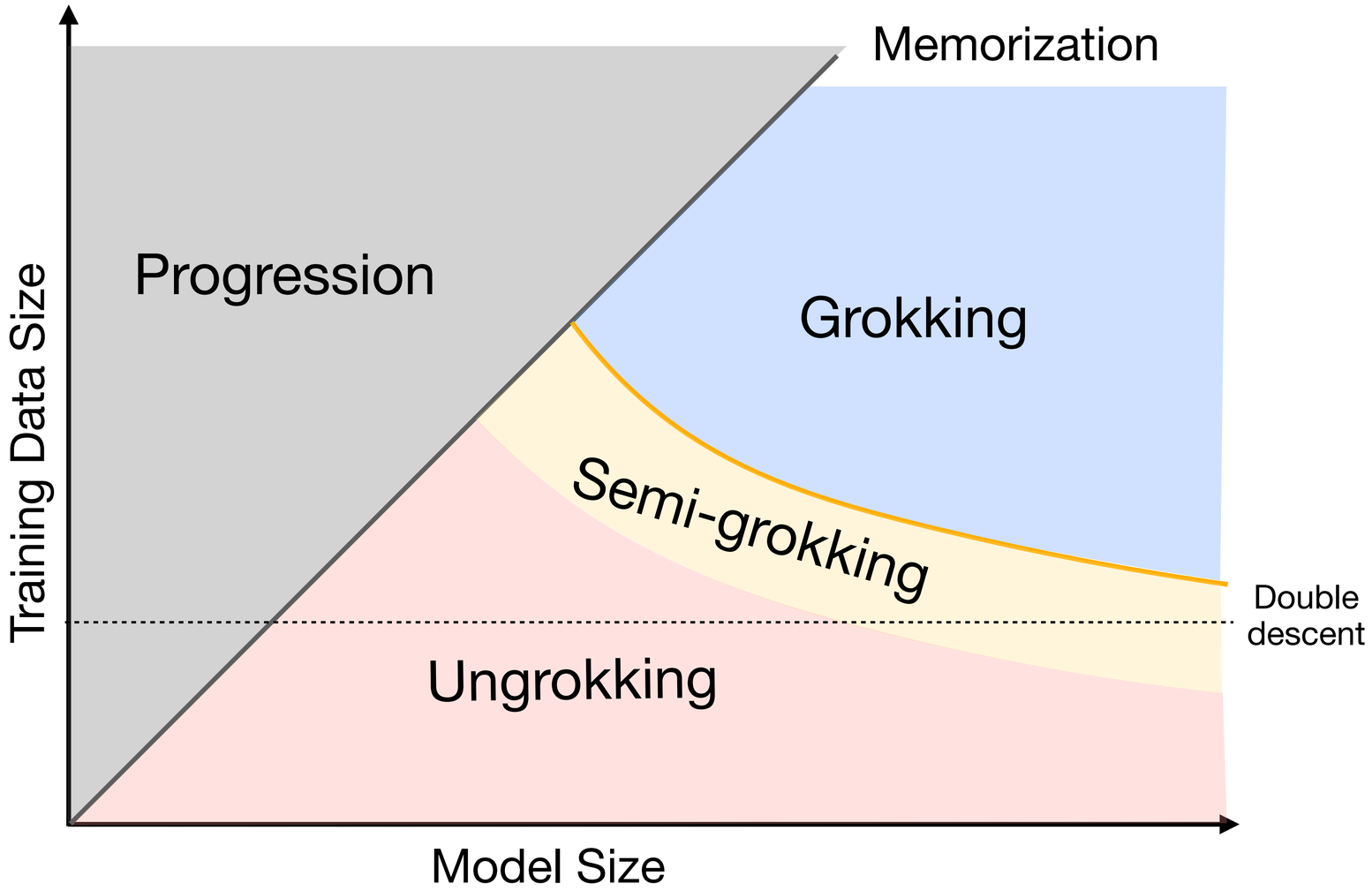
Double Descent and Grokking Landscape

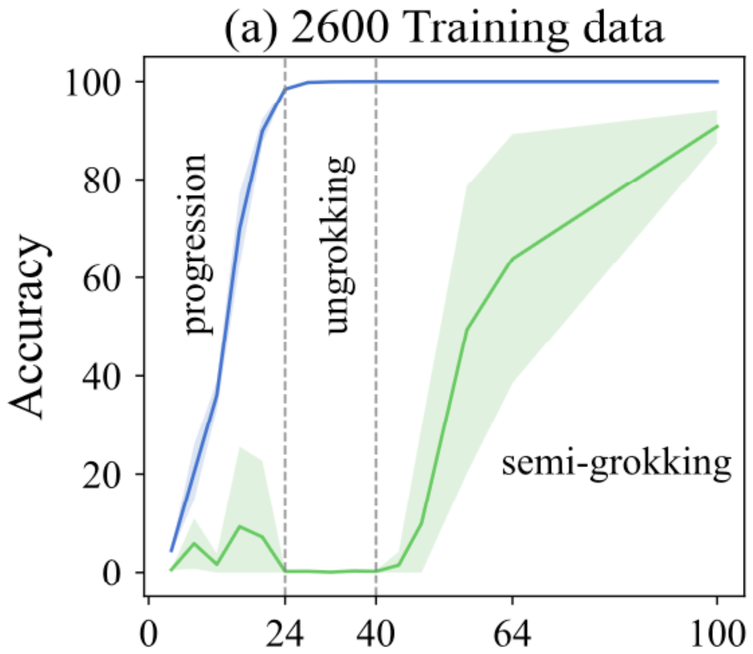
- Observation: Model size sweeps pass through multiple regions.



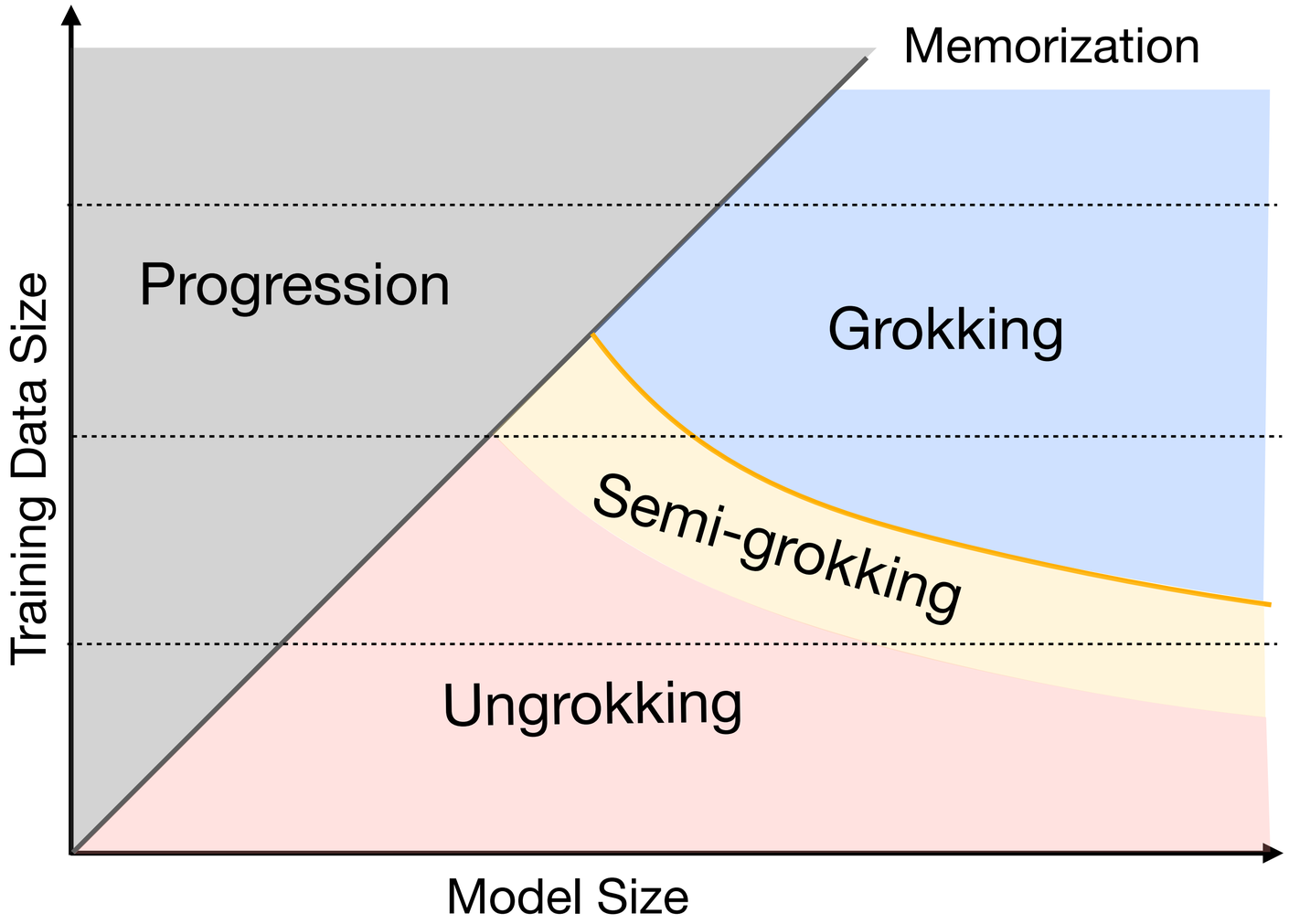
Double Descent and Grokking Landscape

- Observation: Model size sweeps pass through multiple regions.

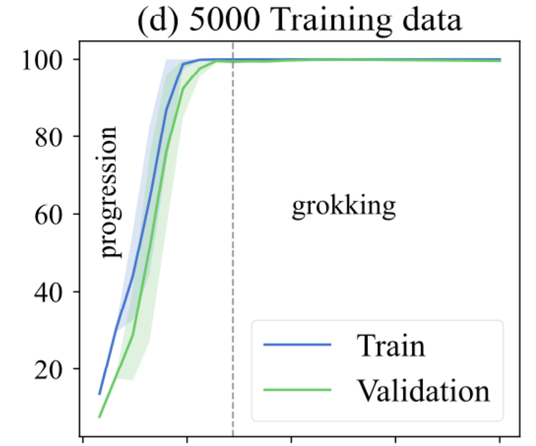


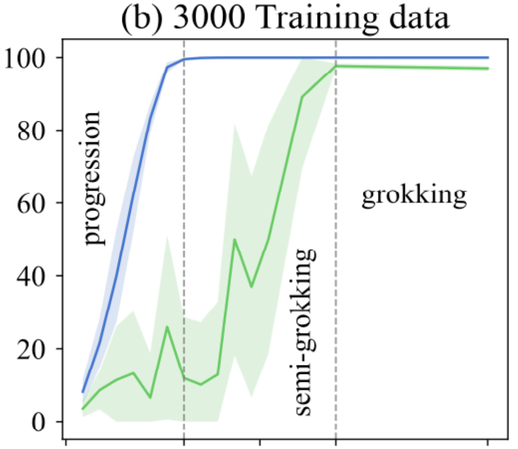


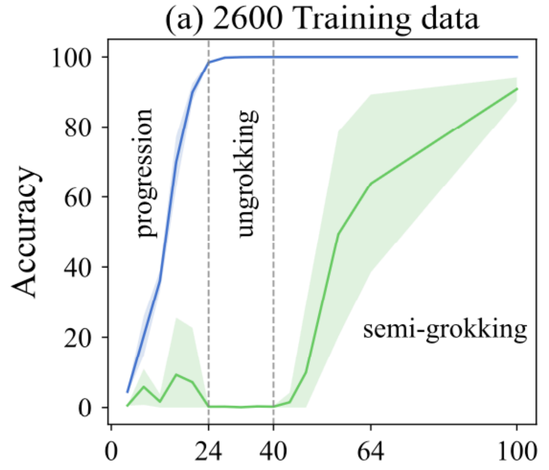
Sweeping Training Data



Double descent







Unified View of Grokking, Double Descent and Emergent Abilities: A Perspective from Circuits Competition

Yufei Huang [1] Shengding Hu [1] Xu Han [1] Zhiyuan Liu [1] Maosong Sun† [1]

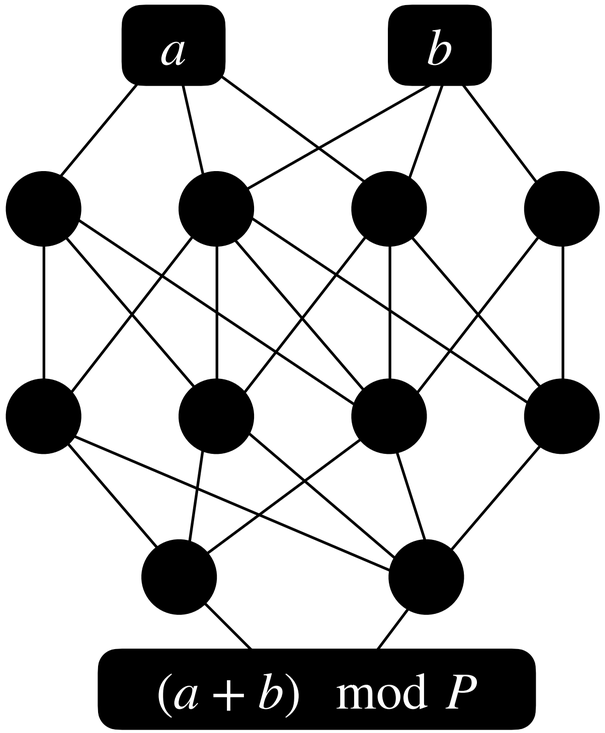
1. Map out landscape of training dynamics varying train set size *and* model size.

2. Use landscape to give an explanation for **double descent**.

3. [Look at how emergent abilities relate to grokking]

Up next from Silvia

What does this general circuit actually look like?



General circuit for addition

Unified View of Grokking, Double Descent and Emergent Abilities: A Perspective from Circuits Competition

Yufei Huang [1] Shengding Hu [1] Xu Han [1] Zhiyuan Liu [1] Maosong Sun† [1]

1. Map out landscape of training dynamics varying train set size *and* model size.

2. Use landscape to give an explanation for **double descent**.

3. How emergent abilities relate to grokking.

Multi-Task Learning

- Data is a mixture of different tasks (e.g. LLM pretraining)
- Emergent task abilities — smoothly decreasing overall loss, but sudden increases in individual task success.

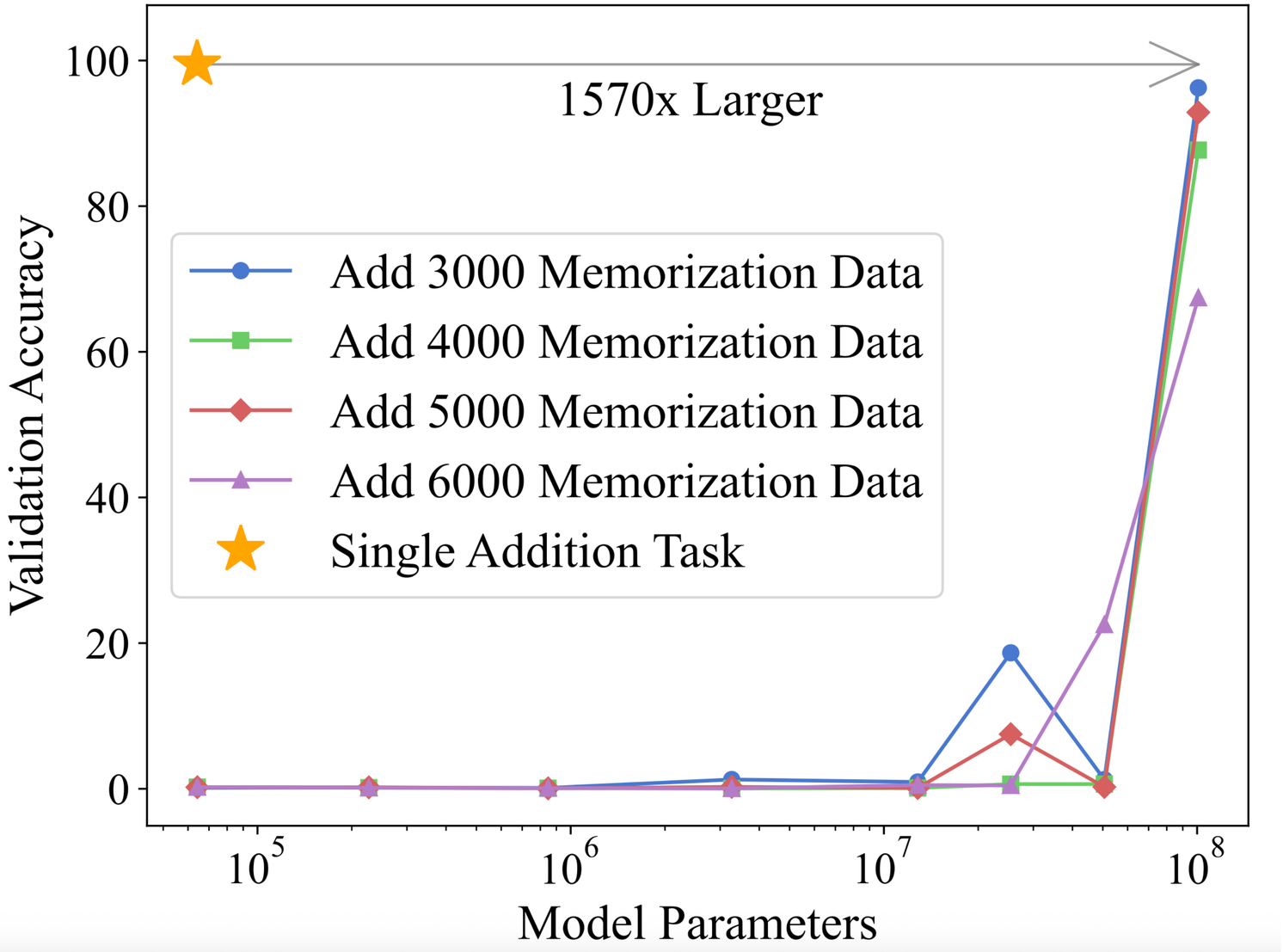
How does learning multiple tasks affect grokking behavior?

Multi-Task Learning

- Data is a mixture of different tasks (e.g. LLM pretraining)
- Emergent task abilities — smoothly decreasing overall loss, but sudden increases in individual task success.
- Setup:
  - Same modular addition task (can generalize)
  - Additional task with random labels (memorization only)
  - Train on various ratios of data, varying model size.

Multi-Task Learning

- Addition task + memorization task.
- Memorization data hinders addition task success
  - but amount of data doesn't seem to matter much
- *plot best of 3 runs



Multi-Task Learning

- Addition task + memorization task.
- Memorization data hinders addition task success
- Purposefully **separating** how the model treats addition vs memorization data helps somewhat.