The emergence of sparse attention

Impact of data distribution and benefits of repetition




Nicolas Zucchet | Francesco D'Angelo | Andrew Lampinen | Stephanie Chan

ETH zürich | EPFL | Google DeepMind

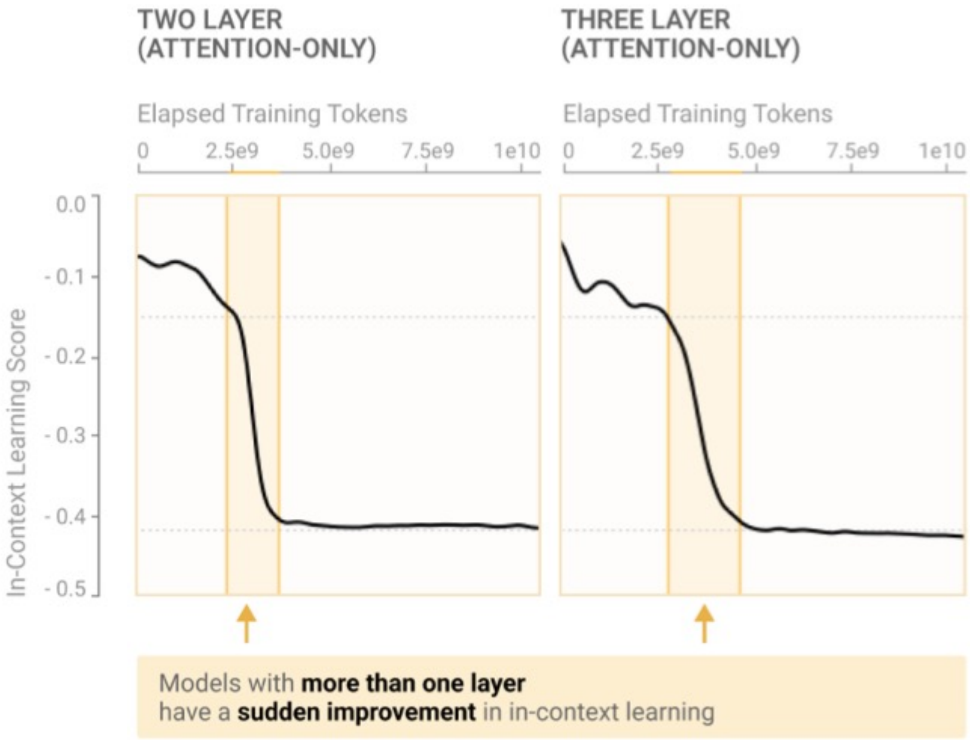
We propose a simple framework to think about **how long** it takes for **Transformers** to **learn** certain abilities: **sparse attention**.

We explain why:

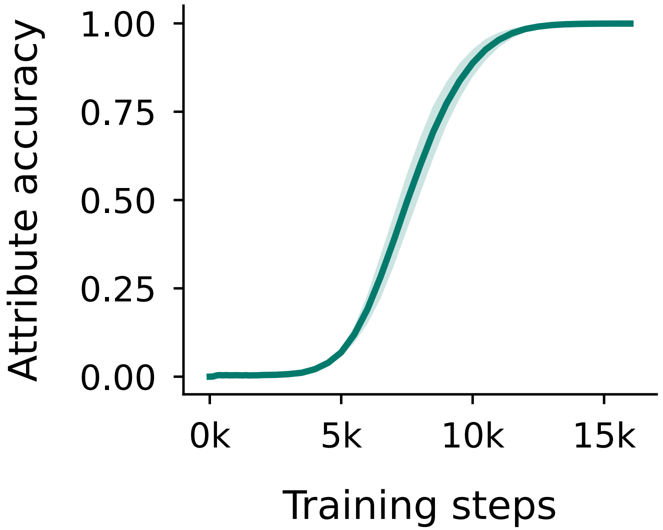
1. Sparse attention **emerges**.
2. Learning time increases as **sequence length increases** and as data gets more **diverse**.
3. **Repetition** can speed up learning.

Motivation

What are the **mechanisms** underlying **emergence** during learning?



Olsson et al. 2022

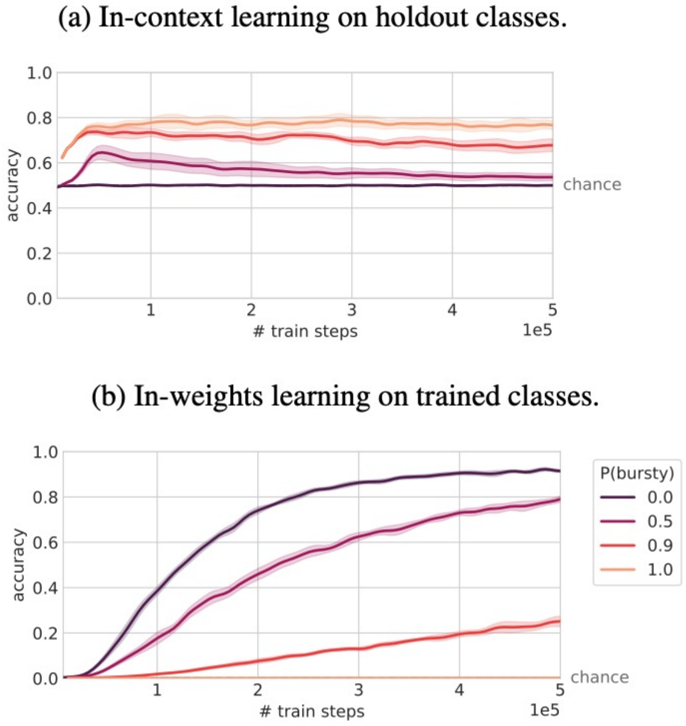


Zucchet et al. 2025

Motivation

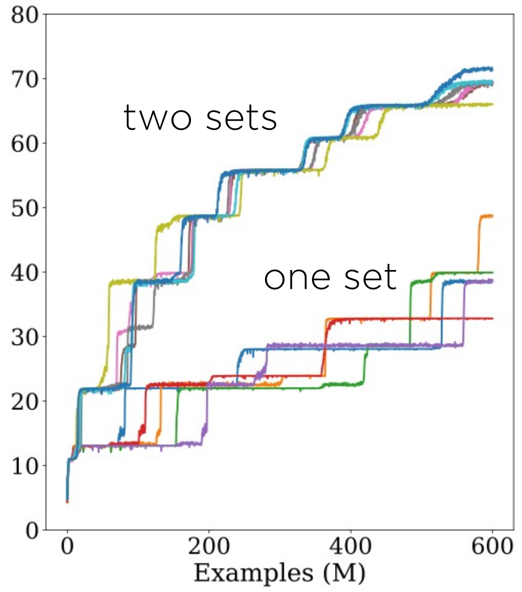
How does **data** influence emergence? Why is **repetition** useful?

In-context learning



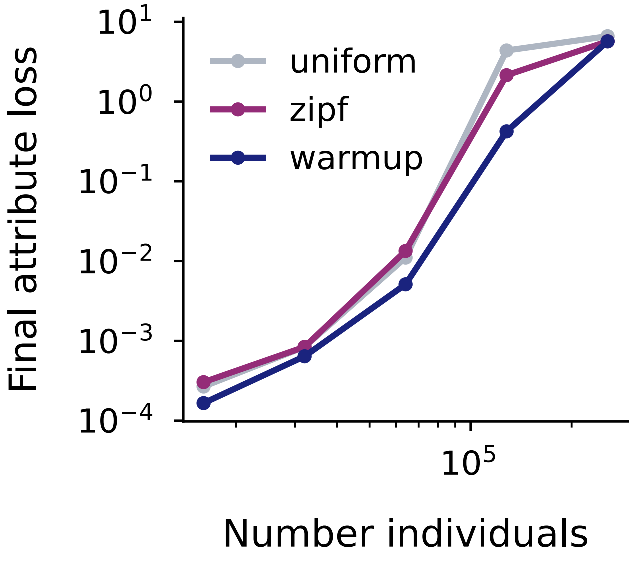
Chan et al. 2022

Computing GCD



Charton & Kempe 2024

Factual recall



Zucchet et al. 2025

Why sparse attention?

Empirical intuition

Many phase transitions **coincide** with the **development of sparse attention** layers

Theoretical intuition

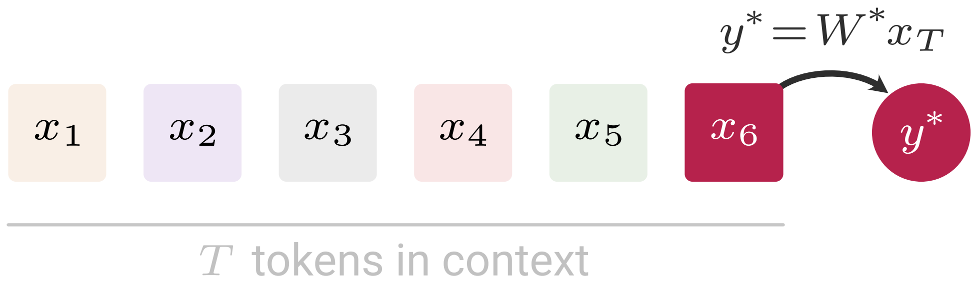
Attention is typically uniform at initialization, so the **information** flowing between two tokens **decreases as context length increases**

A theoretically tractable toy model

What is a sparse attention mechanism doing?

- **Filtering relevant information** out of "noise"
- **Transformation of this information** into desired answer (e.g. an associative memory)

Task. Single-location linear regression



x, y dimension d

Model. Simplified Transformer

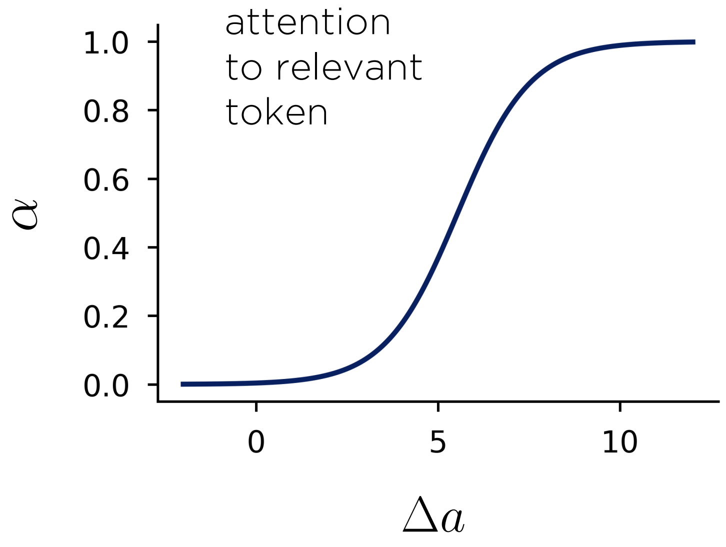
$$y = W \sum_{t=1}^{T} \mathrm{softmax}(a)_t \, x_t$$

Learning dynamics

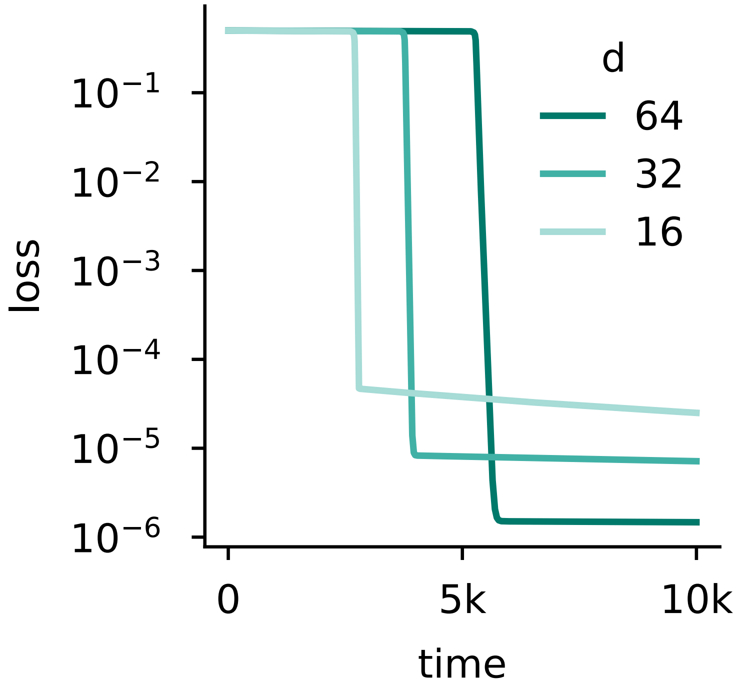
Under reasonable assumptions, we can reduce the learning dynamics to **two variables**

Δa — **logit difference** between relevant and non-relevant tokens

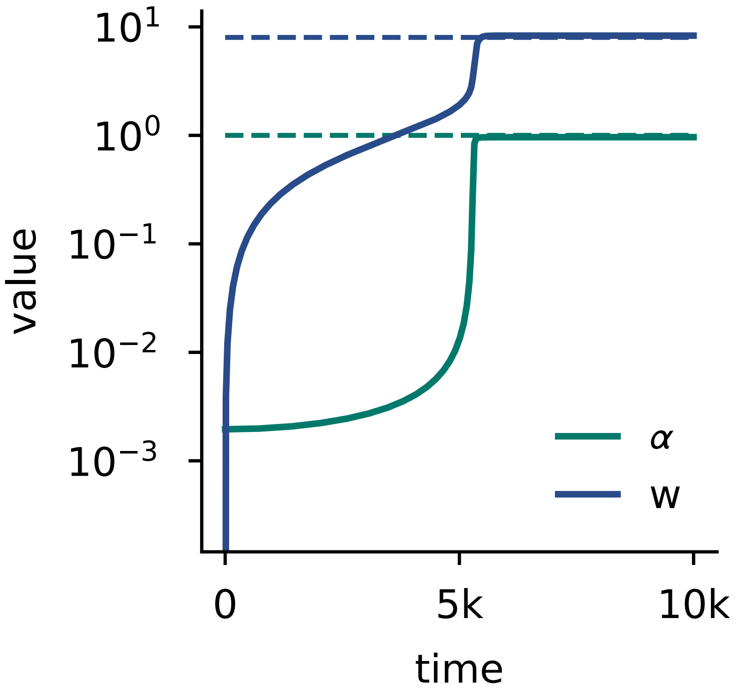
w — projection of **W** on **W***



Learning dynamics



✅ Exhibits sharp phase transitions



w learns before attention focuses on the relevant token

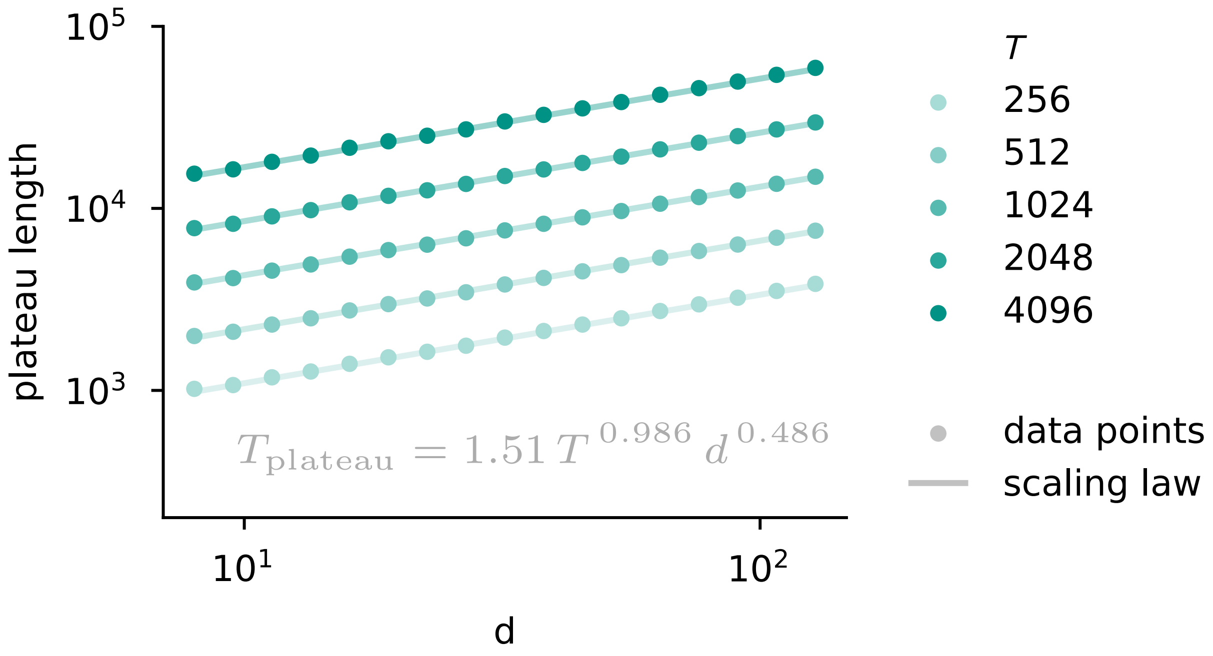
Initial learning dynamics

Linearized dynamics at initialization

$$\begin{pmatrix} \dot{w} \\ \Delta a \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{dT}} \\ 0 \end{pmatrix} + \begin{pmatrix} 0 & \frac{1}{\sqrt{dT}} \\ \frac{1}{\sqrt{dT}} & 0 \end{pmatrix} \begin{pmatrix} w \\ \Delta a \end{pmatrix}$$

Escape time (time to decrease loss by ε)

$$T_\varepsilon = \frac{\sqrt{dT}}{2} \ln\left(\varepsilon\sqrt{dT}\right) \sim \sqrt{dT}$$



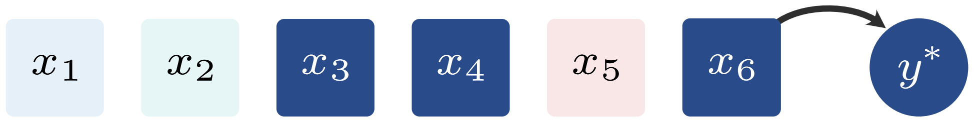
Almost perfect **empirical fit**!

Learning time **increases** when:

- Attention gets **sparser**
- **Less signal** to learn the feedforward mapping

Introducing repetition

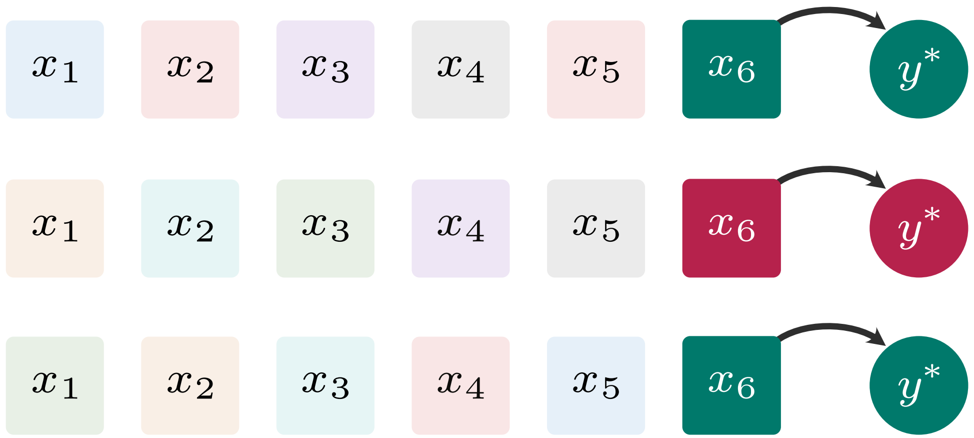
In-context repetition



the relevant token x_T appears B times

Example.
In a Harry Potter chapter, `[Harry Potter]` appears multiple time within the context

Cross-sample repetition

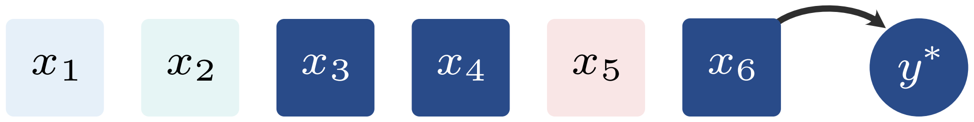


the same x_T appears with probability p

Example.
In Harry Potter books, `[Harry Potter]` appears more often than `[Sirius Black]`

Understanding in-context repetition

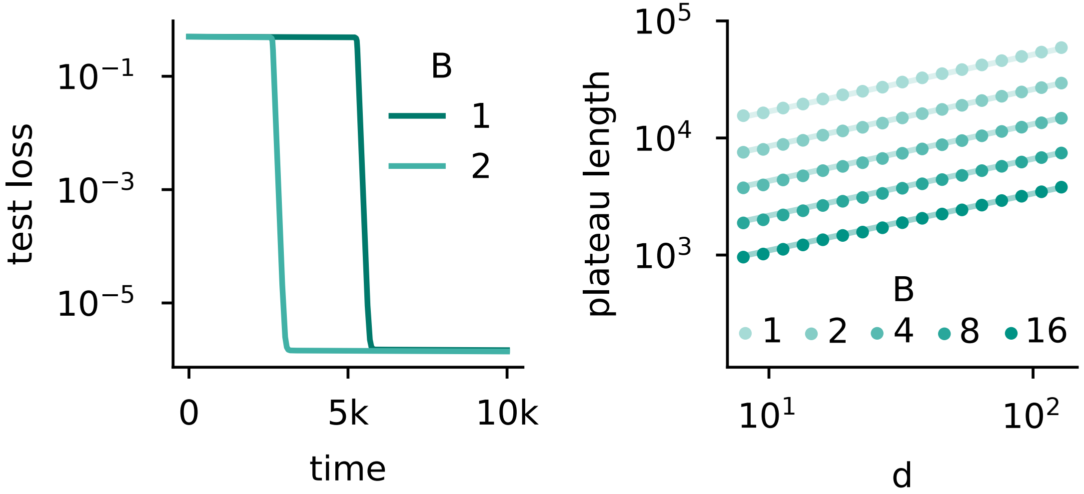
In-context repetition



the relevant token x_T appears B times

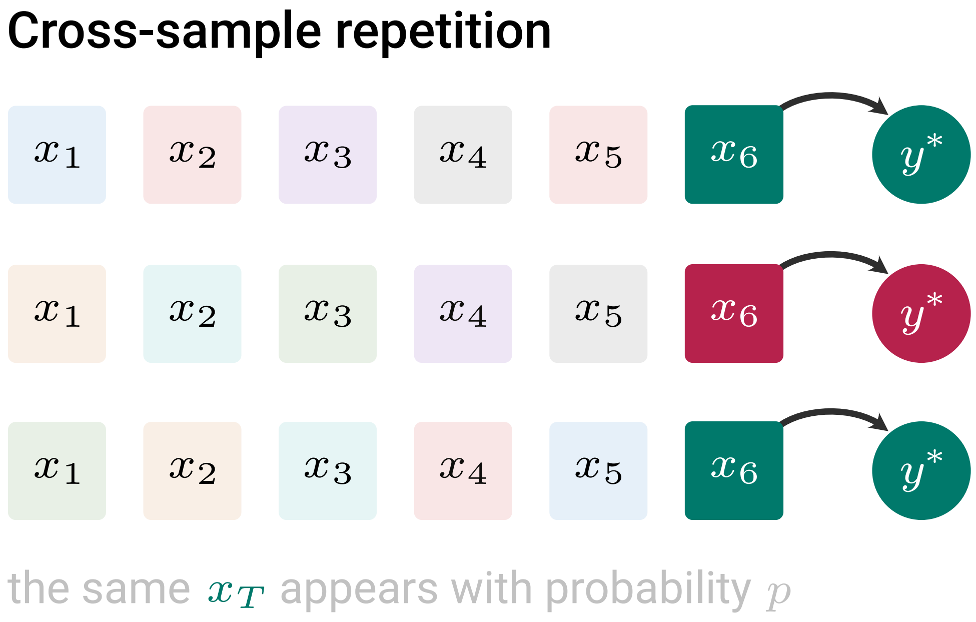
Reduces the sparsity of the target attention, so speeds up emergence

In-context repetition



$$T_{\text{plateau}} = 1.51\, d^{0.49} \left(\frac{T}{B}\right)^{0.99}$$

Understanding cross-sample repetition



Cross-sample repetition speeds up emergence

W learns faster on the repeated dimension

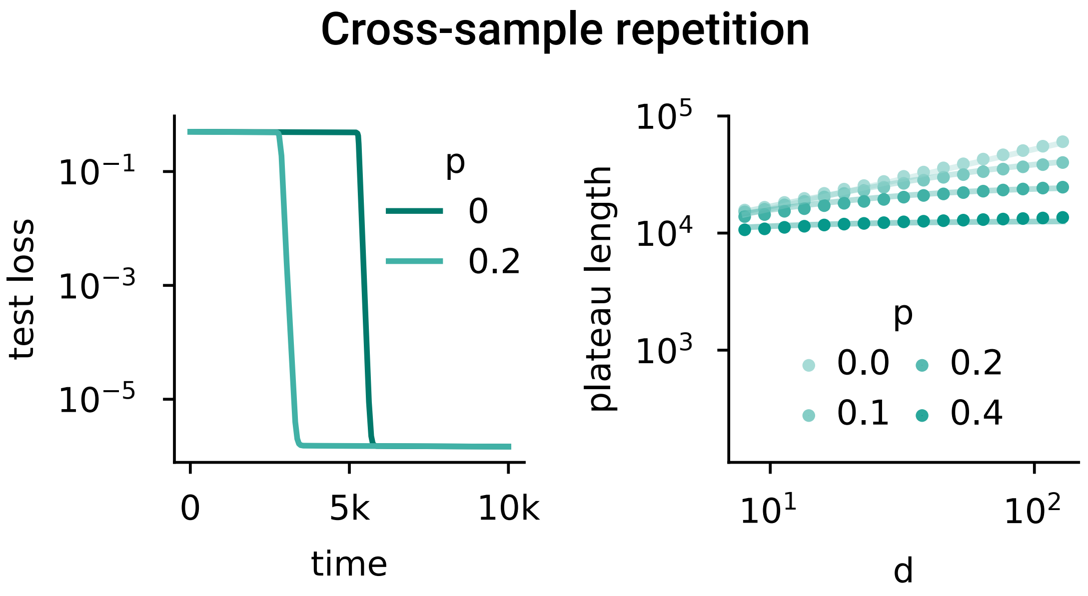
↓

Attention learns faster overall because W provides **better teaching signal** on the repeated data

↓

This speeds up the learning of W on non-repeated data, and thus learning overall

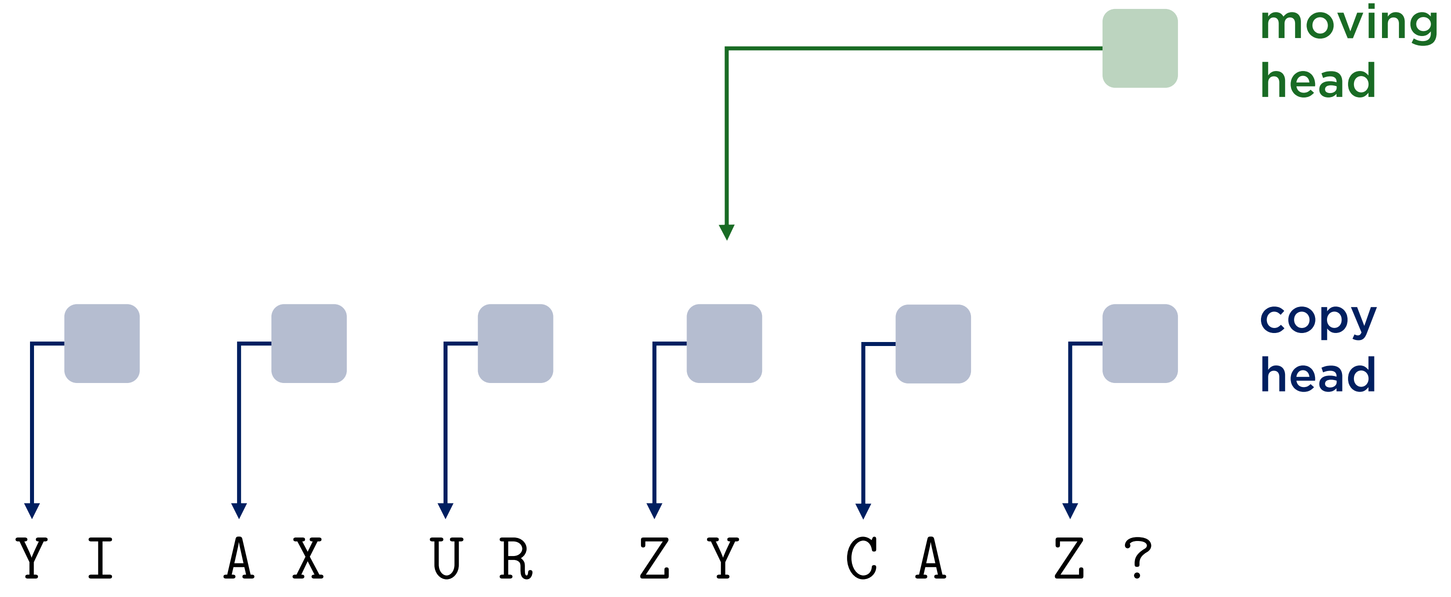
Understanding cross-sample repetition



$$T_{\text{plateau}} = 2.15 \left(\frac{\sqrt{d}T}{\sqrt{p^2 d + (1-p)^2}} \right)^{1.02}$$

Main factor can be **derived theoretically** (similar analysis but with three variables)

Validation on an in-context associative recall task

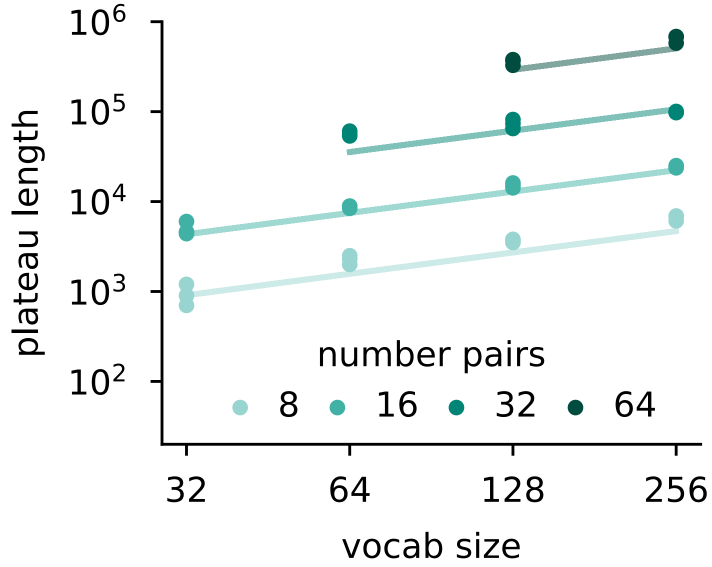


N_{tokens}	vocab size (~ d)
N_{pairs}	number of pairs in context ($(\mathrm{T}-1)/2$)

Combination of **two sparse attention layers**: we should be able to say something about it!

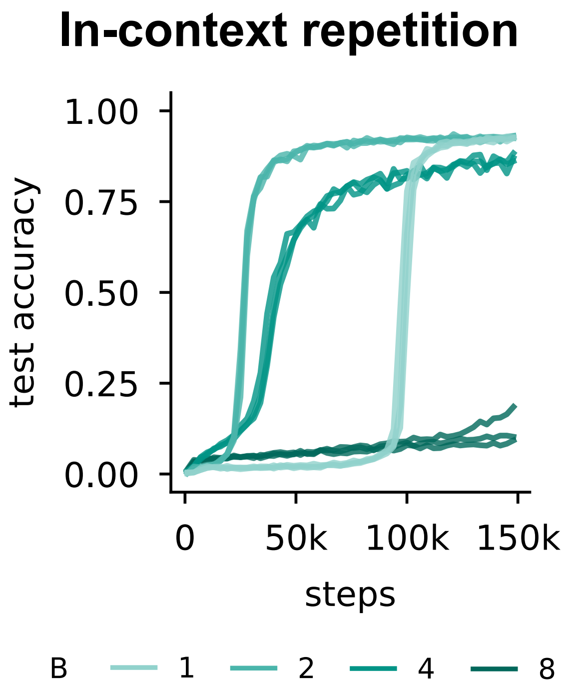
Olsson et al. 2022

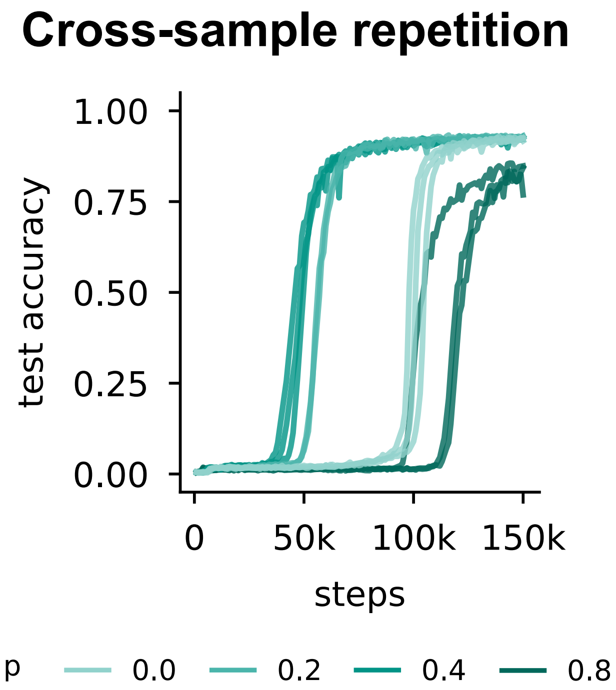
Validation on the in-context associative recall task



$$T_{\text{plateau}} = 0.55\, N_{\text{tokens}}^{0.79}\, N_{\text{pairs}}^{2.25}$$

✅ Same (qualitative) behavior as in the toy task





Increasing in-context repetition is more efficient (cf. power law)

Repetition **speeds up** training, but leads to **overfitting**

Dynamics are messy, hard to get a clean power law

We propose a simple framework to think about **how long** it takes for **Transformers** to **learn** certain abilities: **sparse attention**.

We explain why:

1. Sparse attention **emerges**.
2. Learning time increases as **sequence length increases** and as data gets more **diverse**.
3. **Repetition** can speed up learning.

Paper