STAT 992: Science of Large Language Models

Lecture 3: Out-of-distribution generalization, induction heads

Spring 2026

Yiqiao Zhong

Mechanistic interpretability (MI)

- Anthropic (and early OpenAI) pioneered MI—reverse engineering of neural networks and microscopic understanding
- As LLMs grow larger in scale and complexity, MI becomes difficult
- Yet, many fundamental mechanism and viewpoints remain relevant

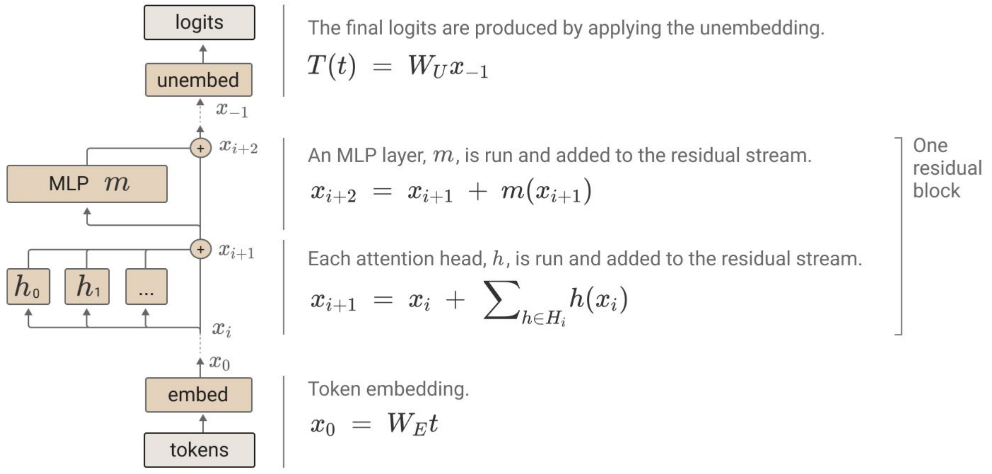
	2017–2019	2020	2021	2022–2023	2024–Present
Key idea	**Analyzing Transformer**	**The Zoom In**	**The Circuit Era**	**Superposition & SAEs**	**Scaling & Automation**
Core Concept	Initial "Probing" (e.g., checking if BERT knows grammar) and Attention Head visualization.	OpenAI's papers begin treating individual neurons as specialized features.	Anthropic's "Mathematical Framework" introduces Induction Heads and "In-Context Learning" as a mechanism.	Focus shifts to **Polysemanticity** (neurons doing multiple things) and using **Sparse Autoencoders (SAEs)** to untangle them.	Automated interpretability (using LLMs to explain LLMs) and mapping features in frontier models like Claude 3.

Table edited by Gemini-3

Mechanistic interpretability (MI)

- A intuitive of viewpoint of transformers (useful but not always accurate)

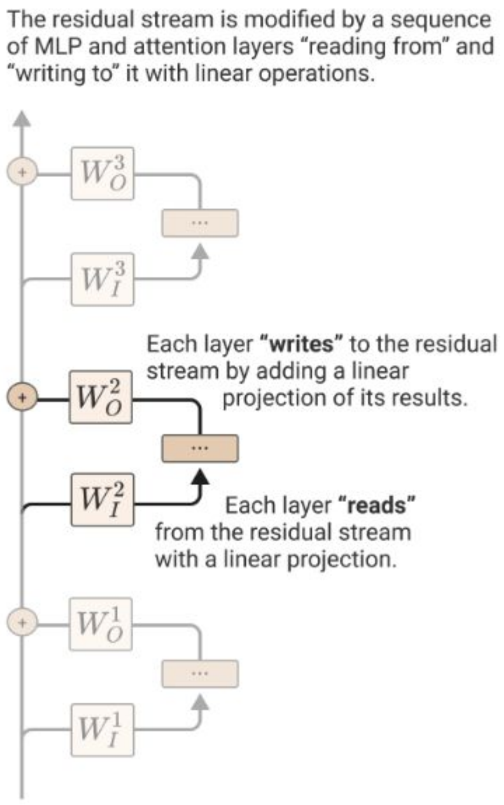
- **Self-attention** (SA) and **MLP** enrich representations by adding to the **residual stream** (identity map from residual connection).
- MLP stores [static knowledge]() as it applies nonlinear transformation token by token
- SA implements dynamic algorithm as it computes interaction between tokens



[Anthropic, A Mathematical Framework for Transformer Circuits, 2021]()

Mechanistic interpretability (MI)

- “**Circuits**” heuristics
- SA and MLP “**read**” (accept input embeddings) from residual stream, process vectors, and “**write**” (return vectors as outputs) to residual stream.
- Attention head as **pattern detector**: activates for one or several patterns in a prompt
- Attention matrix: for given a prompt, how the current token interacts with another token
- Idealized interpretation: logical / algebraic operations in the vector space

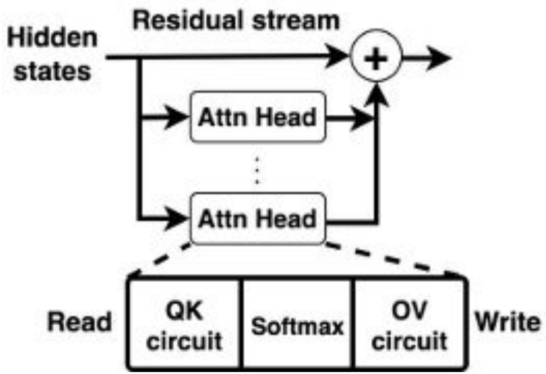


Anthropic, A Mathematical Framework for Transformer Circuits, 2021

Mechanistic interpretability (MI)

- Input or hidden states $\boldsymbol{X} \in \mathbb{R}^{T \times d}$, T is seq length, d is embed dim

- How is this plausible?
    - In theory, transformers can express algo
    - In exploratory work, modified transformers are trained and binarized into programs

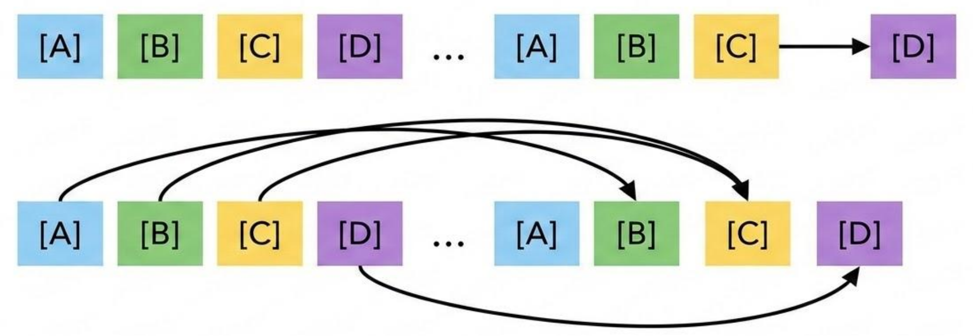


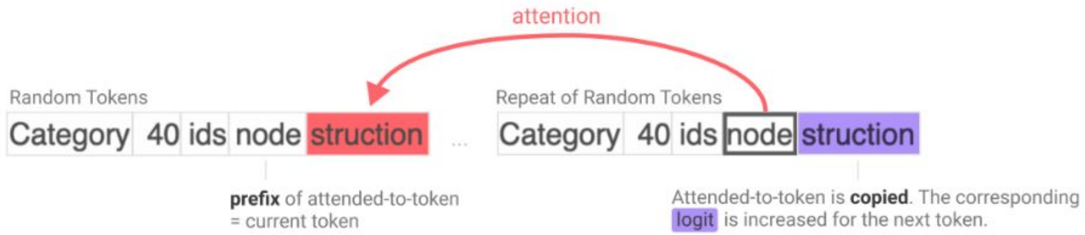
$$\mathrm{MSA}(\boldsymbol{X};\boldsymbol{W}) := \underbrace{\boldsymbol{X}}_{\substack{\text{residual stream stores} \\ \text{info from previous layer}}} + \sum_{j=1}^{H} \overbrace{\mathrm{Softmax}\underbrace{\left(\boldsymbol{X}\boldsymbol{W}_{\mathrm{QK},j}\boldsymbol{X}^{\top}\right)}_{\substack{\text{QK circuit reads and} \\ \text{matches info from stream}}}}^{\text{attention matrix}} \underbrace{\boldsymbol{X}\boldsymbol{W}_{\mathrm{OV},j}}_{\substack{\text{OV circuit writes and} \\ \text{adds info to stream}}}$$

Induction head: a basic building block underlying emergence and ICL

Copying in context

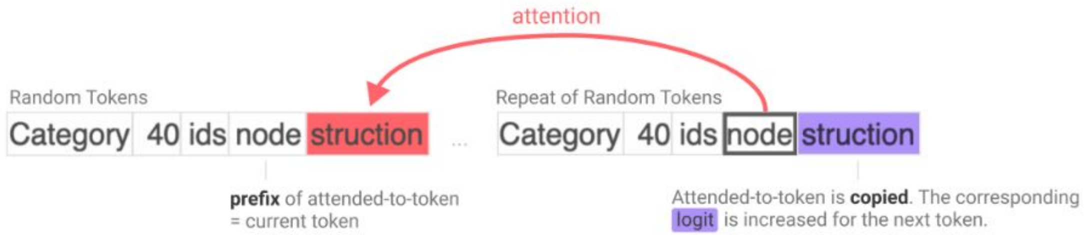
Suppose that is a pattern—consecutive tokens [A], [B], [C], [D] in the sequence–to be completed





Copying in context

- How would a classical statistical model learn to copy?
  - Estimate the joint probability distribution of p([A], [B], [C], [D])
  - Modeling [A], [B], [C], [D] as a (hidden) Markov chain
- General-purpose statistical models can't generalize beyond training data
  - Different token distributions
  - Different pattern length
- In transformers, composition of two self-attention heads solves copying:
  - First head: **previous-token head** (attending to previous token)
  - Second head: **induction head** (attention to to-be-copied token)



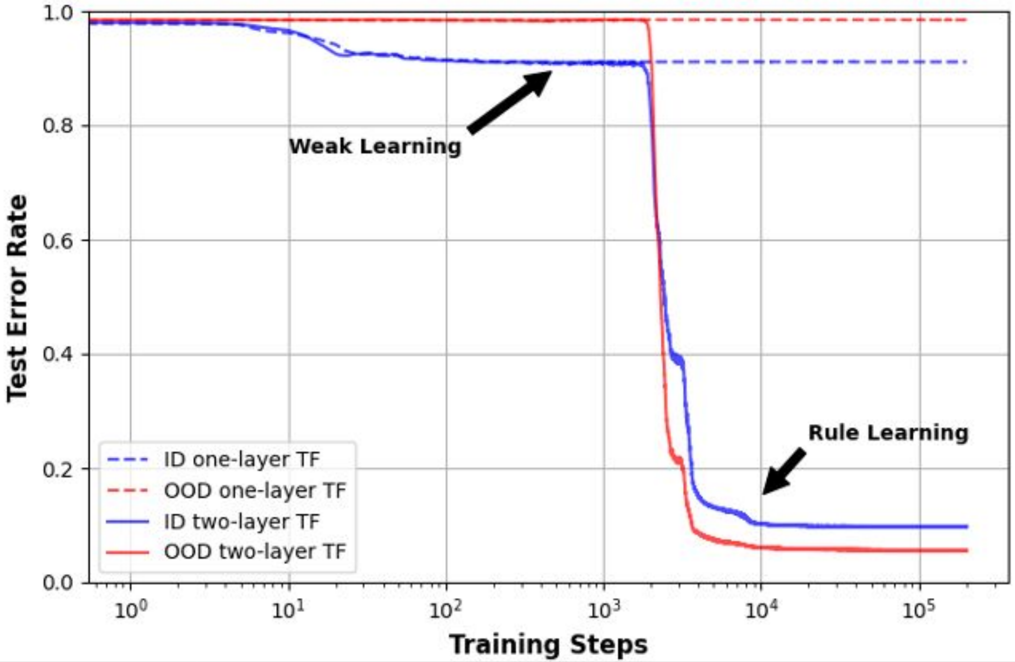
Copying in context: simple synthetic experiment

- Training data
  - Vocabulary size 64, sequence len 64, draw i.i.d. tokens from a power law distribution to form "noisy background" in a prompt
  - Sample segment len $L \in \{10, 11, \dots, 19\}$ uniformly, and then sample a segment $\boldsymbol{s}^{\#}$ of len L
  - Place two copies of $\boldsymbol{s}^{\#}$ at random non-overlapping locations in the prompts. Prompt format $(*, \boldsymbol{s}^{\#}, *, \boldsymbol{s}^{\#}, *)$

- OOD Test data
  - Change token distribution to uniform
  - Change L to 25

- Model: 2-layer transformer without MLPs

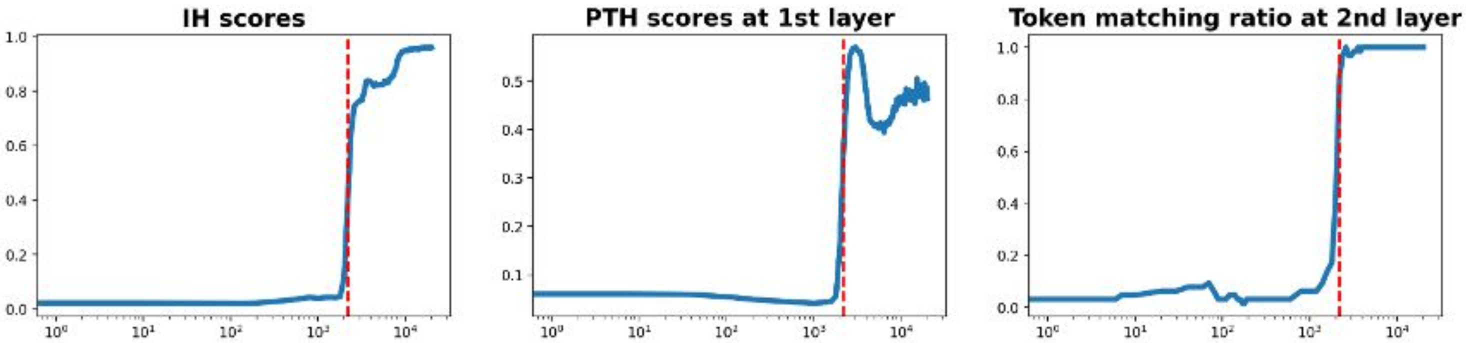
Copying in context: simple synthetic experiment

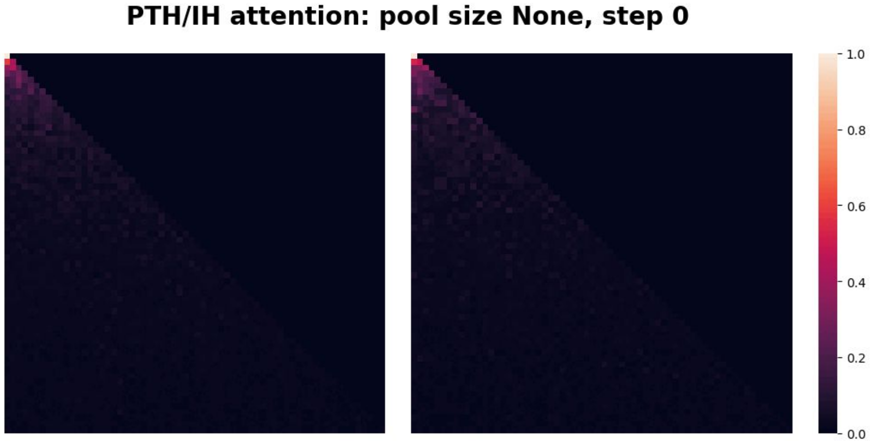


- **Weak learning phase**: rely on simple statistics of ID data and fail to generalize OOD
- **Rule-learning phase**: two-layer TF learns the rule of copying from ID data

Out-of-distribution generalization via composition: A lens through induction heads in Transformers, 2024

Copying in context: Induction head mechanism





1st attention head
focuses on position info

PTH attends to
previous token

[A], [B], [C] · · · [A], [B]

PTH shifts
embedding

Induction head: training on corpus and emergence of ICL

- Training small transformers on natural language data

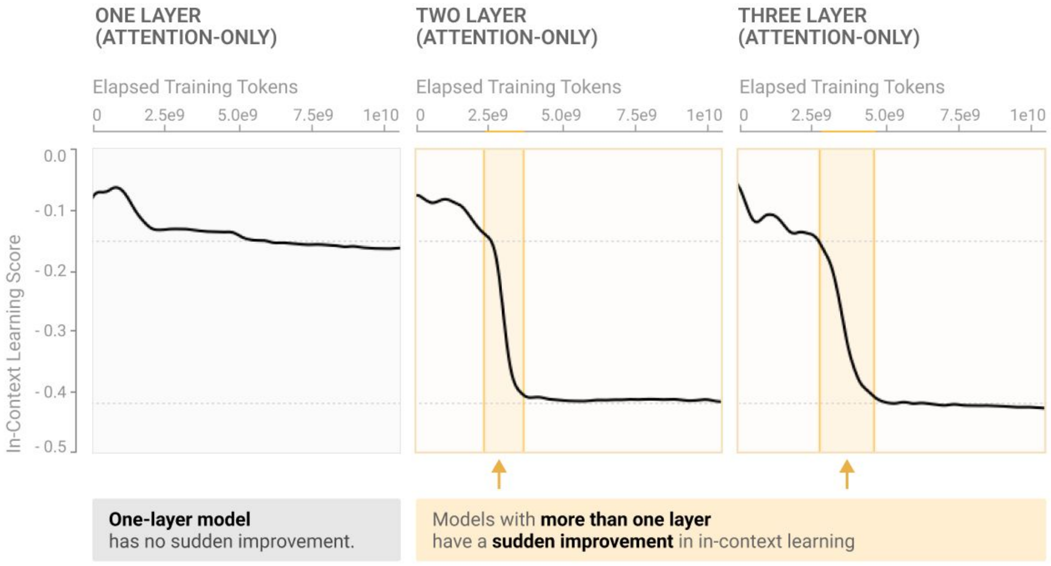
ICL score: $\ell_{500}(t) - \ell_{50}(t)$

➔ Recall that the autoregressive loss is $\mathcal{L}(t) = \sum_{k=1}^{L} \ell_k(t)$.

➔ On average, it is cross entropy between language and model prediction.

➔ Intuitively, a longer context helps prediction (conditioning reduces entropy)

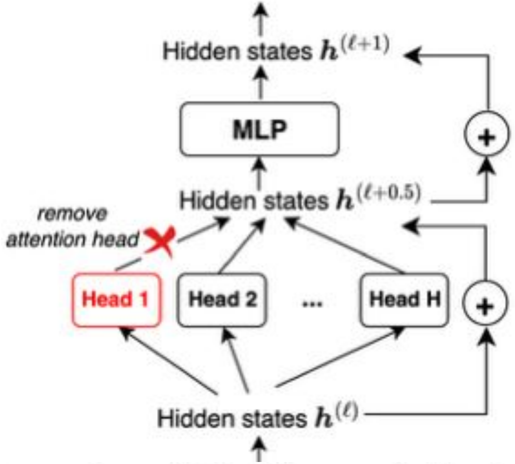
ICL scores reflects how much better longer context helps prediction

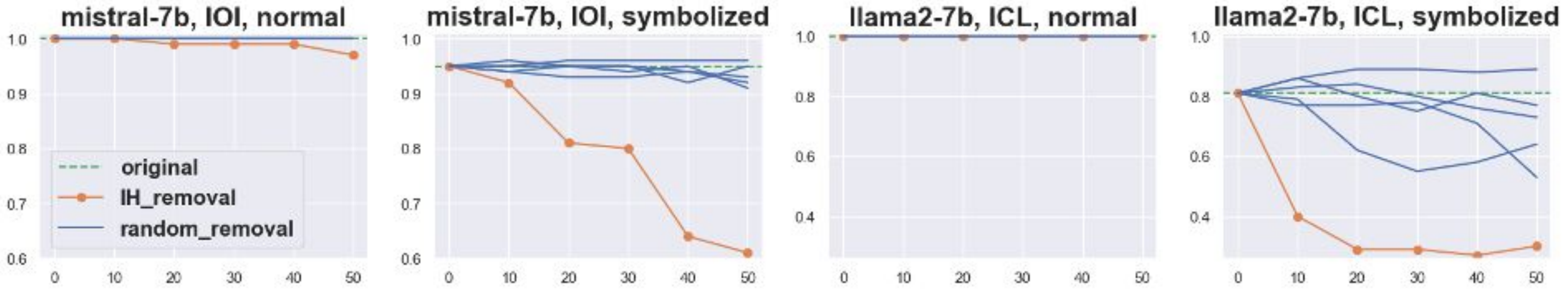


Anthropic, In-context learning and induction heads, 2022

Intervention experiment from pretrained LLMs

- Many attention heads in LLMs (even GPT2-small has 12*12 heads)
- Ranking heads and screen top ~50 as induction heads
- Evaluating models with normal prompts (ID) vs unnatural / abstract prompts (OOD)



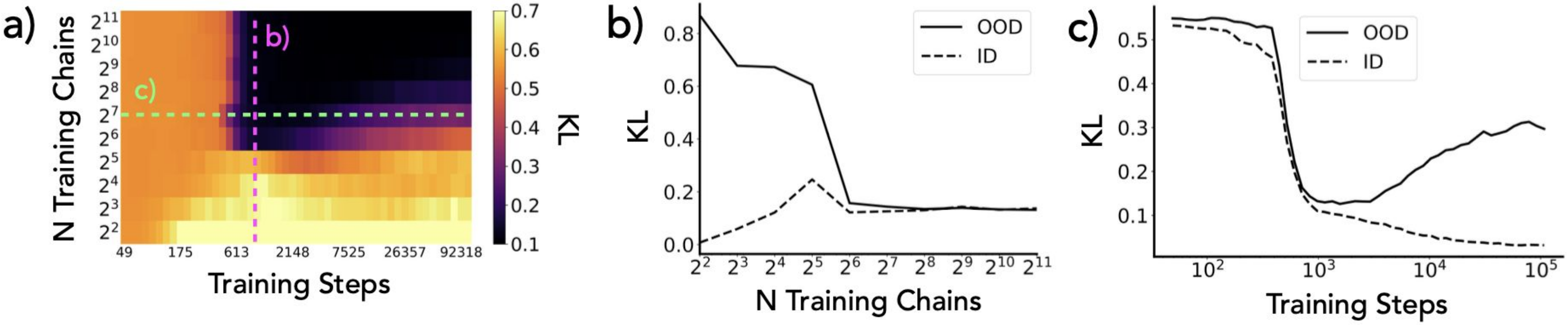


[Out-of-distribution generalization via composition: A lens through induction heads in Transformers](), 2024

Beyond copying: induction head learns Markov chain in context

Finite mixtures of Markov chains

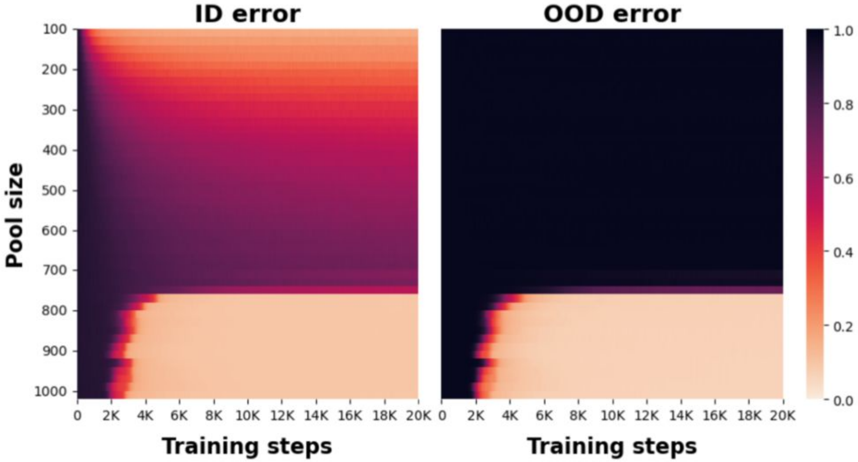
- Each input sequence [A], [B], [C], [D], … is a Markov Chain (MC)
- The model trained on different MCs (e.g., different transition matrices)
- Can it generalize on new MCs? (OOD generalization)
- Copying is a special case of MC, as transition is deterministic

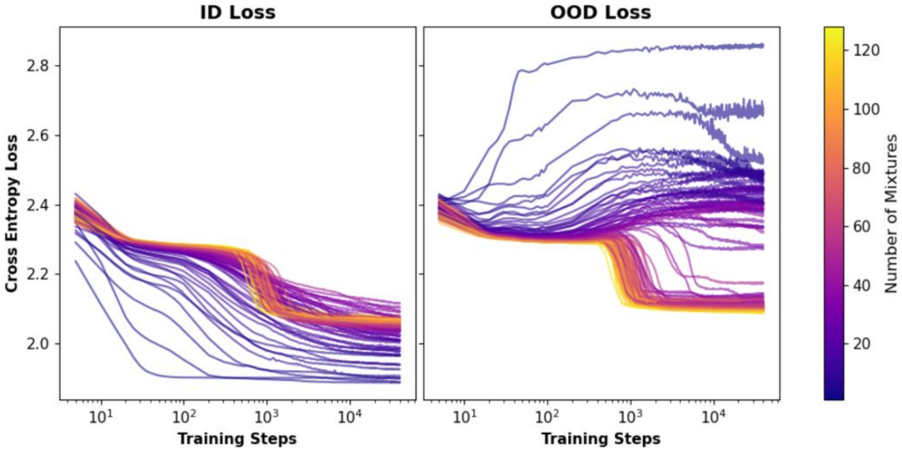


Competition Dynamics Shape Algorithmic Phases of In-Context Learning, 2024

Phase transitions in data diversity and training steps

- Left: finite patterns for copying task
- Right: finite transition kernels for learning MCs





Open problems & research ideas

- Conclusion: induction heads are critical to ICL and OOD generalization
  - Copying patterns from context
  - Inferring from new Markov chains
- How are phase changes developed in training?
- How do models represent algorithms beyond induction heads?
- What are other mechanisms of OOD generalization
- What is the role of training data?
  - Diversity of patterns / tasks