Out-of-distribution generalization via composition: a lens through induction heads in Transformers

Jiajun Song [1] Zhuoyan Xu [2] Yiqiao Zhong [2]

[1]State Key Laboratory of General Artificial Intelligence, BIGAI, Beijing [2]Department of Statistics, University of Wisconsin– Madison, Madison, WI

Do LLMs solve novel tasks?

Out-of-distribution (OOD) generalization measures the performance on novel tasks $\mathcal{P}_{\text{train}} \neq \mathcal{P}_{\text{test}}$. New challenges since advent of LLMs.

- Prompting, In-context learning.
- Compositional structure.
- Tasks that require "reasoning".

Goal: In-depth empirical analysis to understand

- How composition is internally represented by LLMs;
- How critical geometric structure emerges from training;
- How they empower language & reasoning tasks across a wide range of models.

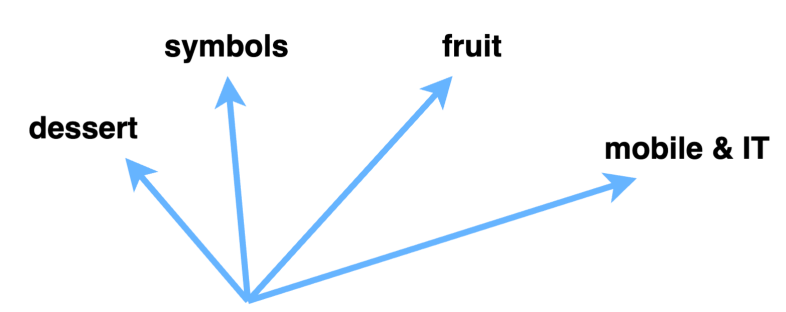
A primer on Transformers: How concepts are represented

Transformers from circuits perspective. Let $\boldsymbol{X} = [\boldsymbol{x}_1^{(\ell)}, \dots, \boldsymbol{x}_T^{(\ell)}]^\top \in \mathbb{R}^{T\times d}$ be the input vectors or the hidden states at a layer ℓ.

$$\boldsymbol{X} \longleftarrow \boldsymbol{X} + \mathrm{MSA}(\boldsymbol{X}; \boldsymbol{W}^{(\ell)}), \qquad \boldsymbol{X} \longleftarrow \boldsymbol{X} + \mathrm{FFN}(\boldsymbol{X}; \widetilde{\boldsymbol{W}}^{(\ell)})$$

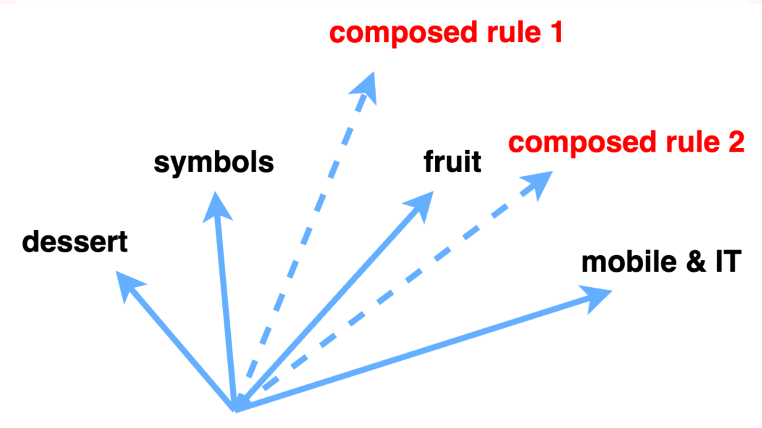
$$\mathrm{MSA}(\boldsymbol{X};\boldsymbol{W}) = \sum_{j=1}^{H} \overbrace{\mathrm{Softmax}\underbrace{\left(\boldsymbol{X}\boldsymbol{W}_{\mathrm{QK},j}\boldsymbol{X}^\top\right)}_{\text{QK circuit reads and matches info from stream}}}^{\text{attention matrix}} \underbrace{\boldsymbol{X}\boldsymbol{W}_{\mathrm{OV},j}^\top}_{\text{OV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

where $\boldsymbol{W}_{\mathrm{QK}}, \boldsymbol{W}_{\mathrm{OV}} \in \mathbb{R}^{d\times d}$ are query-key, output-value matrices.
Linear representation hypothesis: concepts are encoded as linear subspaces within the embedding space.
Feature superposition: hidden states are sparse linear combinations of base concept vectors from a large dictionary.



$$\text{apple} = 0.09\,\text{"dessert"} + 0.11\,\text{"organism"} + 0.16\,\text{"fruit"} + 0.22\,\text{"mobile\&IT"} + 0.42\,\text{"other"}$$

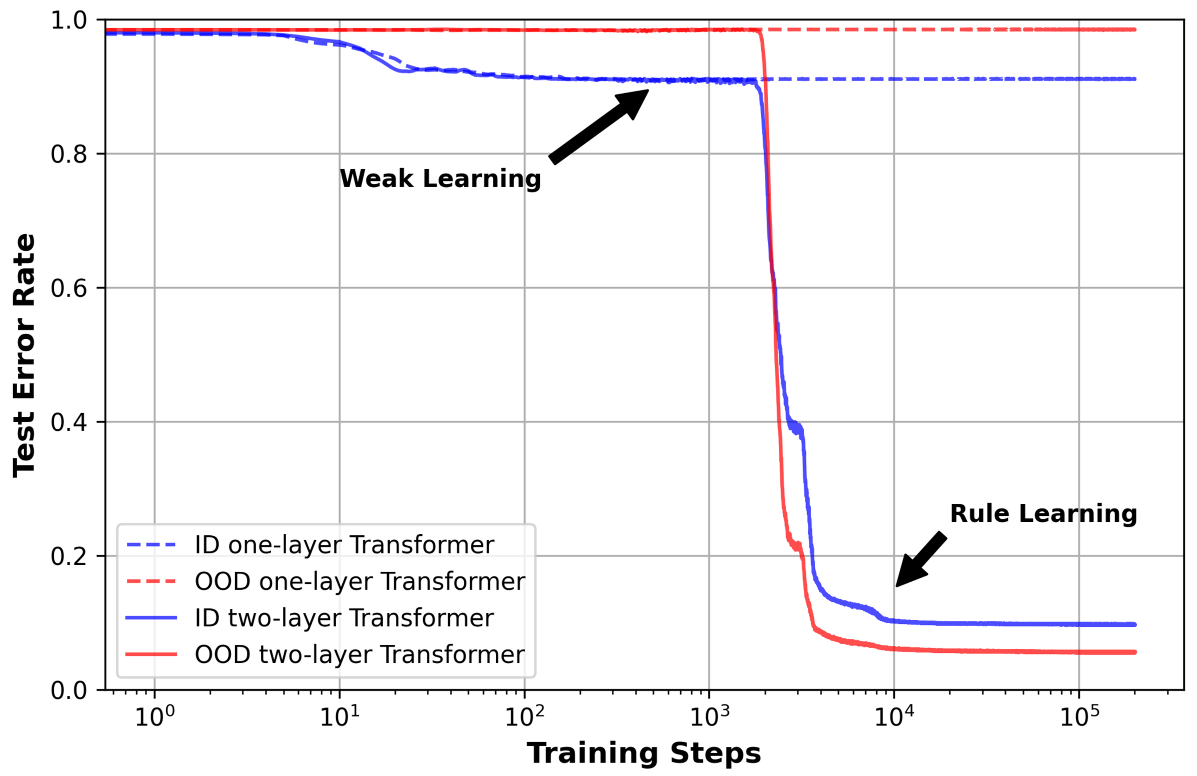
Main message: composition through subspace matching empowers OOD generalization



- Concept subspaces + rule subspaces
- Composed rule 1 (e.g., copying), composed rule 2 ...
- Enables OOD generalization, esp. in novel context (ICL, CoT)

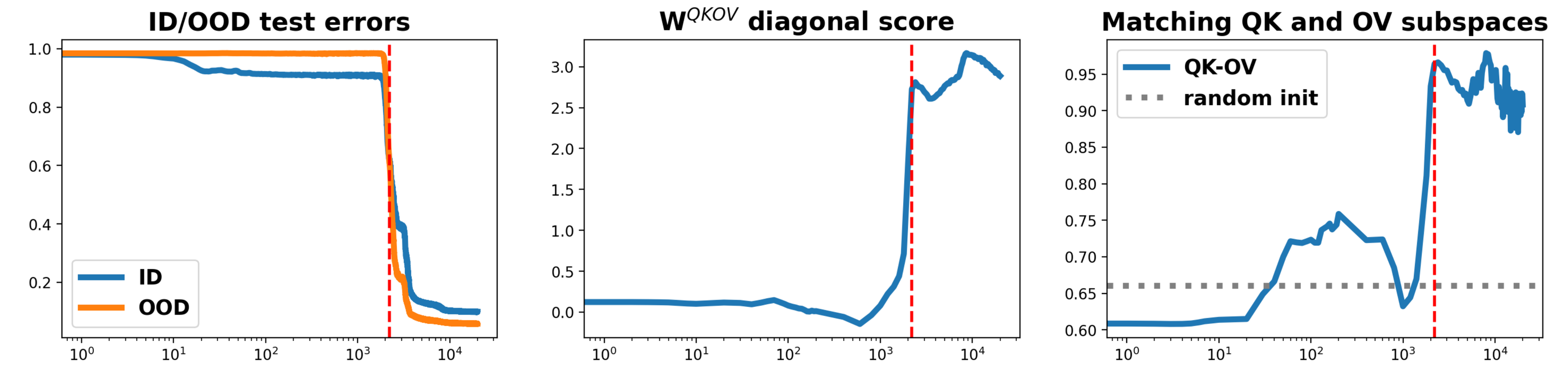
Synthetic example: training dynamics on copying task

Copying task : $\dots[A],[B],[C]\dots[A],[B] \xrightarrow{\text{next-token prediction}} \dots[A],[B],[C]\dots[A],[B],[\boldsymbol{C}]$

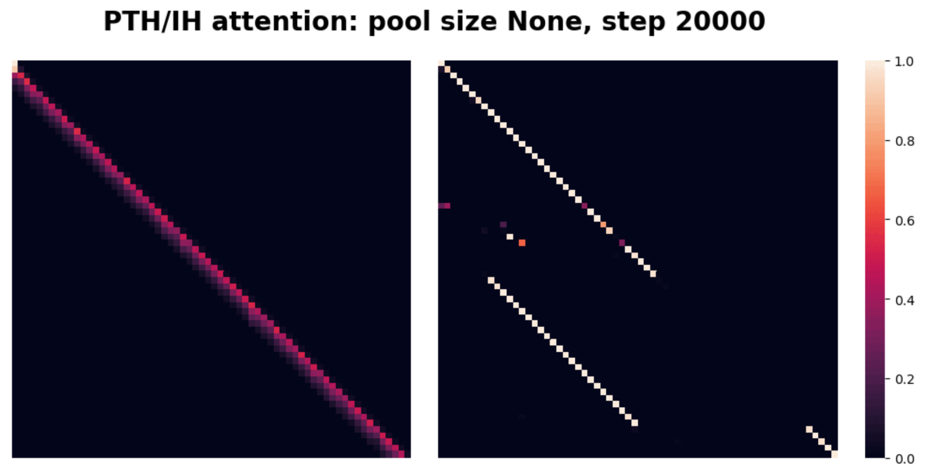


1. **Training data generation.** Vocabulary size 64, context length 64, i.i.d. tokens from power law distribution. Segment $\boldsymbol{s}^\#$ of random tokens with length $\mathrm{Unif}(\{10,11,\dots,19\})$. Two copies of $\boldsymbol{s}^\#$ at random non-overlapping locations. Prompt format $(*, \boldsymbol{s}^\#, *, \boldsymbol{s}^\#, *)$.
2. **OOD data generation.** Token distribution uniform, segment length 25.
3. **Model.** 2-layer (1-layer) 1-head TF with no FFN, LayerNorm, RoPE, dropout.
4. **Training.** Fresh samples, autoregressive, AdamW.

Low-dimensional subspace matching emerges abruptly



- **Diagonal score**: normalized average diagonal entries of $\boldsymbol{W}_{\mathrm{QK}}^{(2)}\boldsymbol{W}_{\mathrm{OV}}^{(1)}$.
- **Subspace matching**: generalized cosine sim between two principal subspaces ($r = 10$).
- **Previous-token head (PTH) and induction head (IH).** Two types of attention heads. Follow similar sharp transition, complementary role (position vs. token matching).



$$\mathrm{score}^{\mathrm{PTH}} = \mathrm{Ave}_{i\le N}\Big(\frac{1}{T-1}\sum_{T\ge t\ge 2}(\boldsymbol{A}_i)_{t,t-1}\Big)$$

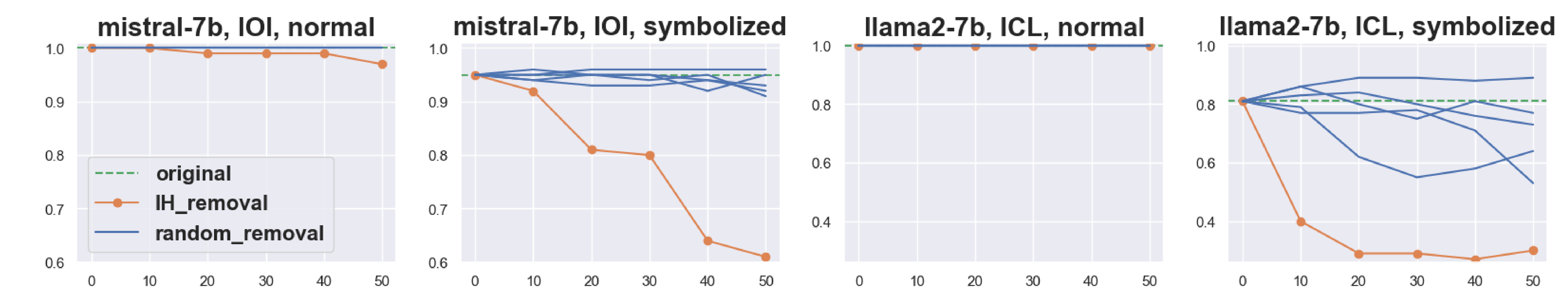
$$\mathrm{score}^{\mathrm{IH}} = \mathrm{Ave}_{i\le N}\Big(\frac{1}{|\mathcal{I}_i|}\sum_{t\in\mathcal{I}_i}(\boldsymbol{A}_i)_{t,t-L_i+1}\Big)$$

$\mathcal{I}_i$: index set of repeating tokens
L_i : distance between two segments

Experiments on pretrained LLMs: symbolic & reasoning tasks

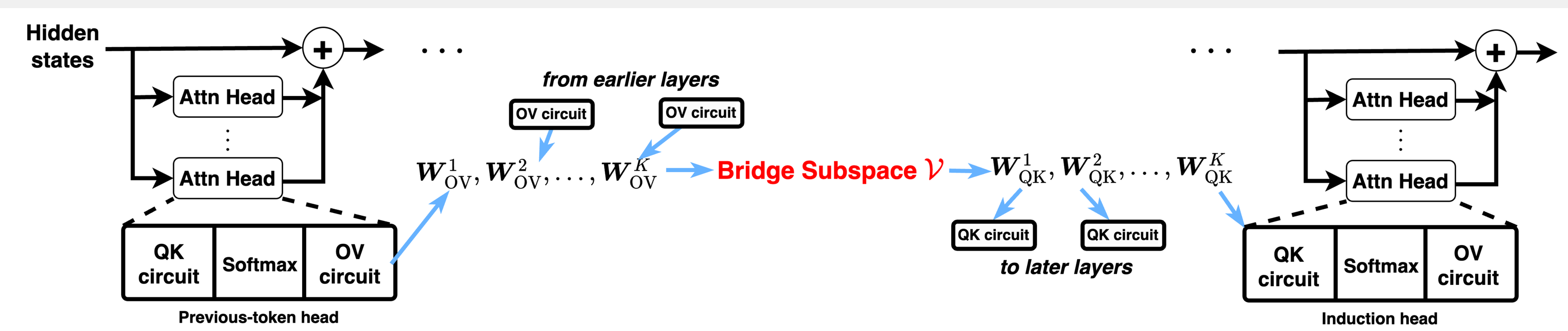
1. **Indirect object identification** (IOI). Normal (N) vs. Symbolized (S).
   (N) "Then, Henry and Blake had a long argument. Afterwards Henry said to" ⟶ Blake
   (S) "Then, &ˆ and #$ had a long argument. Afterwards &ˆ said to" ⟶ #$
2. **In-context learning** (ICL).
   (N) "baseball is sport, celery is plant, sheep is animal, volleyball is sport, lettuce is" ⟶ plant
   (S) "baseball is $#, celery is !%, sheep is &*, volleyball is $#, lettuce is" ⟶ !%
3. **Math reasoning** with chain-of-thought (CoT) on GSM8K.
   "Jerry is cutting up wood for his wood-burning stove. Each pine tree makes 80 logs, each maple tree makes 60 logs, and each walnut tree makes 100 logs. If Jerry cuts up 8 pine trees, 3 maple trees, and 4 walnut trees, how many logs does he get?" [...Deduction...] "#### 1220"

- 100 test prompts, two versions (Normal as in-distribution, Symbolized as OOD).
- Removal top-K induction heads (ranked by attention scores) vs. removal of random heads, $K = 0, 10, \dots, 50$.



1. **Finding 1:** Normal prompts are insensitive to IH removal (likely memorization)
2. **Finding 2:** In contrast, OOD/reasoning prompts accuracy rely crucially on IHs as a component in composition.

Common bridge representation hypothesis



Hypothesis: For a compositional task, there exists a low-dimensional subspace $\mathcal{V} \subset \mathbb{R}^d$ s.t.

$$\mathcal{V} = \mathrm{span}(\boldsymbol{W}_{\mathrm{OV},j}) = \mathrm{span}(\boldsymbol{W}_{\mathrm{QK},k}^\top).$$

- Extension of linear representation hypothesis to compositional tasks.
- Key to OOD generalization.
- Supported by ablation experiments (projecting weights onto $\mathcal{V}$ vs. onto $\mathcal{V}^\perp$)

References

[1] Jiajun Song, Zhuoyan Xu, and Yiqiao Zhong. Out-of-distribution generalization via composition: a lens through induction heads in transformers. *arXiv preprint arXiv:2408.09503*, 2024.