Can LLMs solve compositional tasks? A study of out-of-distribution generalization

Yiqiao Zhong (UW Madison Statistics)

HKU IDS, May 28, 2025

Collaborators

Jiajun Song, BIGAI

Zhuoyan Xu, UW Madison

Paper: https://www.pnas.org/doi/10.1073/pnas.2417182122

Are LLMs creative? Or are they a hype?

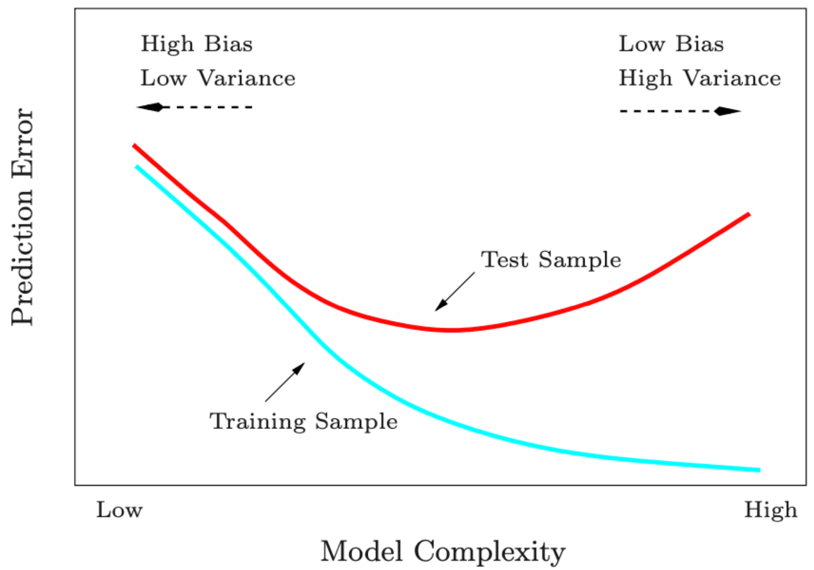
- Two polarizing opinions
  - Sparks of artificial general intelligence
  - LLMs memorize facts, parrot the speech
- Intriguing phenomena: *Emergent abilities*
  - Sudden emergence, sharp increase in accuracy
  - In-context learning (ICL)
  - Chain-of-thought (CoT)
- Lack of scientific foundations
  - Overloading notions
  - Unclear model internals
  - Lack of clear measurements

Sparks of Artificial General Intelligence:
Early experiments with GPT-4

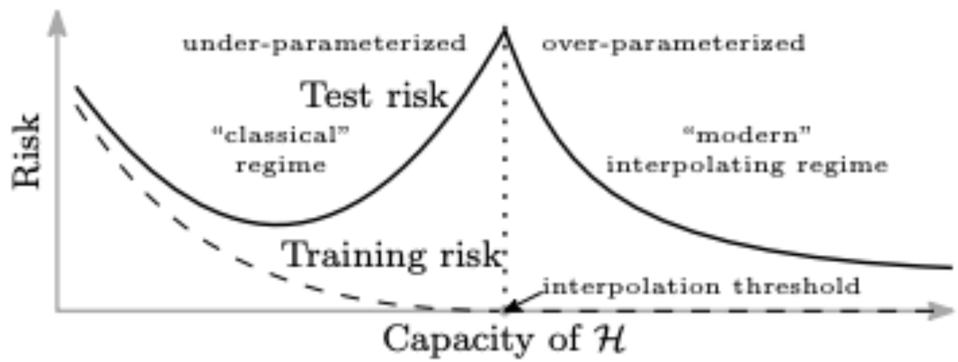
Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke
Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg
Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

Microsoft Research

Does classical notions of generalization explain?



ESL: Bias-variance tradeoff



Belkin et. al. 2019: Double descent

$$\mathcal{P}_{\text{train}} = \mathcal{P}_{\text{test}}$$

*Lack of performance measures on **Novel task***

Compositions and OOD generalization

- Out-of-distribution (OOD) generalization: $\mathcal{P}_{\text{train}} \neq \mathcal{P}_{\text{test}}$
- In-distribution (ID) generalization: $\mathcal{P}_{\text{train}} = \mathcal{P}_{\text{test}}$
- Compositions and "reasoning": benefits of multiple layers

Holy grail

- How do LLMs represent **composition**?
- When do we expect **emergence**?
- Why do LLMs achieve **OOD generalization**?

Teaser: Evidence of OOD generalization

Realistic Task: "Symbolized language reasoning"

- **Indirect object identification (IOI)**
  - (normal)

"Then, Henry and Blake had a long argument. Afterwards Henry said to" → Blake

  - (symbolized)

"Then, &^ and #$ had a long argument. Afterwards &^ said to" → #$

- **In-context learning (ICL)**
  - (normal)

"baseball is sport, celery is plant, sheep is animal, volleyball is sport, lettuce is" → plant

  - (symbolized)

"baseball is $#, celery is !%, sheep is &*, volleyball is $#, lettuce is" → !%

See Rong 2021, Wang et. al., ICLR 2023, Pan et. al., ACL 2023

- Draw 100 test prompts for each subtask, two versions (normal as ID, symbolized as OOD)
- **IOI**: $[\text{Subject}] \ldots [\text{Object}] \ldots [\text{Subject}] \ldots [\text{Object}]$
- **ICL**: $x_1, f(x_1), x_2, f(x_2), \ldots, x_n, f(x_n)$ where $f: \text{object} \mapsto \text{category}$
- Calculate Acc in multiple-choice form, random guess 1/2 (IOI), 1/3 (ICL has 3 categories)

	Llama2-7B	**Falcon-7B**	**Olmo-7B**	**Mistral-7B**	**Falcon2-11B**	**Llama3-8B**
Normal	1	1	1	1	1	1
Symbolized	0.84	1	0.96	0.95	0.96	0.99

	Llama2-7B	**Falcon-7B**	**Olmo-7B**	**Mistral-7B**	**Falcon2-11B**	**Llama3-8B**
Normal	1	1	1	1	1	1
Symbolized	0.81	0.45	0.79	0.45	0.82	0.90

Synthetic Task: "Learning copying with a simple Transformer"

$$\ldots [A], [B], [C] \ldots [A], [B] \xrightarrow{\text{next-token prediction}} \ldots \underbrace{[A], [B], [C]}_{\boldsymbol{s}^{\#}} \ldots \underbrace{[A], [B], {\color{blue}[C]}}_{\boldsymbol{s}^{\#}}$$

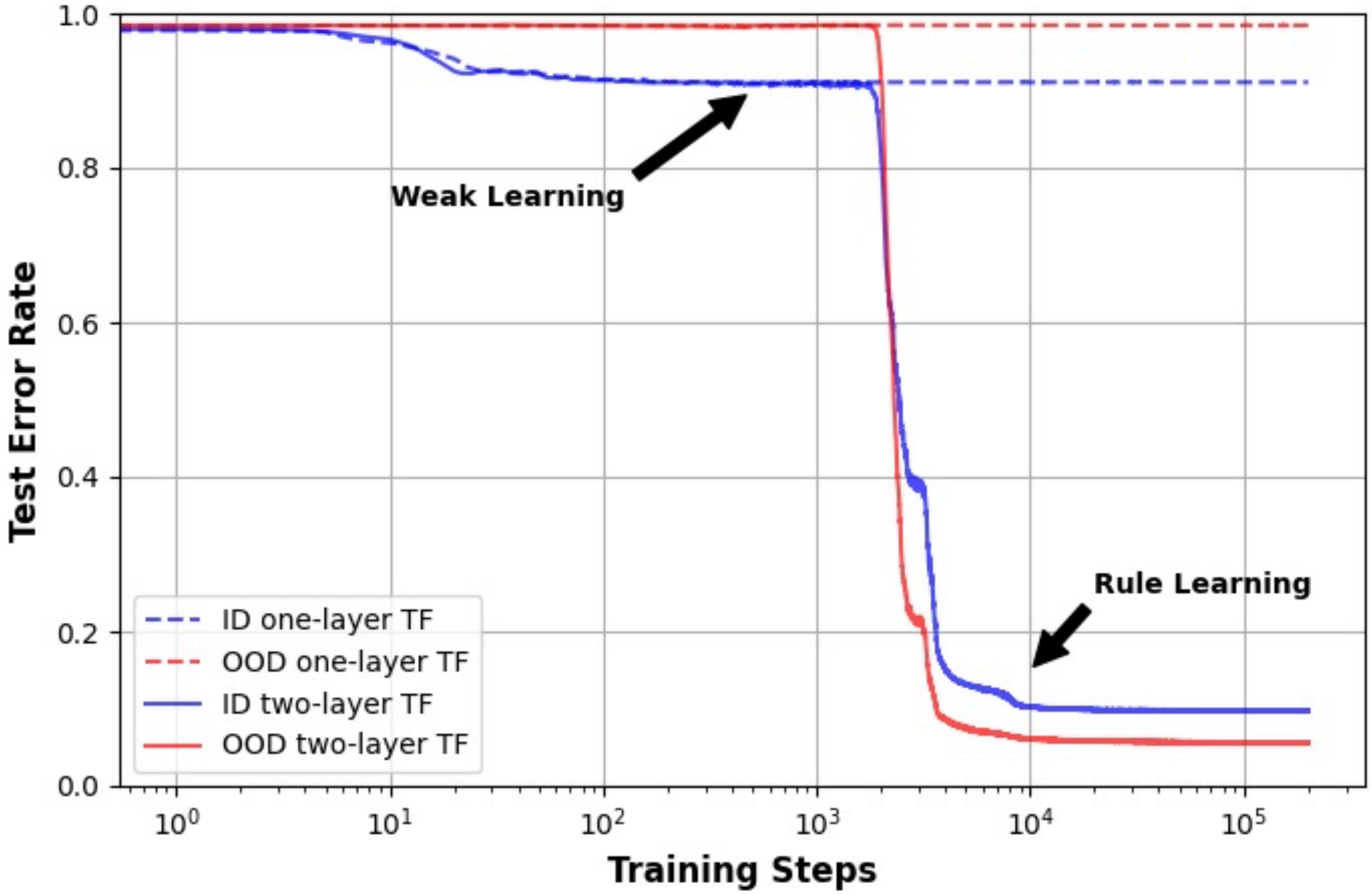
- Vocabulary size 64, sequence len 64, draw i.i.d. tokens from a power law distribution to form "noisy background" in a prompt
- Sample segment len $L \in \{10, 11, \ldots, 19\}$ uniformly, and then sample a segment $\boldsymbol{s}^{\#}$ of len L
- Place two copies of $\boldsymbol{s}^{\#}$ at random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

See Olsson et. al., 2022

- OOD data
  - <u>Token distribution</u> changed from power law to uniform
  - <u>Length</u> of repeating segment changed from {10, 12, ... 19} to 25

- Model: minimal Transformer, 2-layer and 1-head
  - No MLP, standard architecture (residual connection, LayerNorm, RoPE, dropout)
  - Trained on **fresh samples (one-pass setting)**, autoregressive, standard technique (AdamW)

- Simple for rule-based algorithms, but hard for classical general-purpose ML methods (n-gram models, hidden Markov models)



- **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

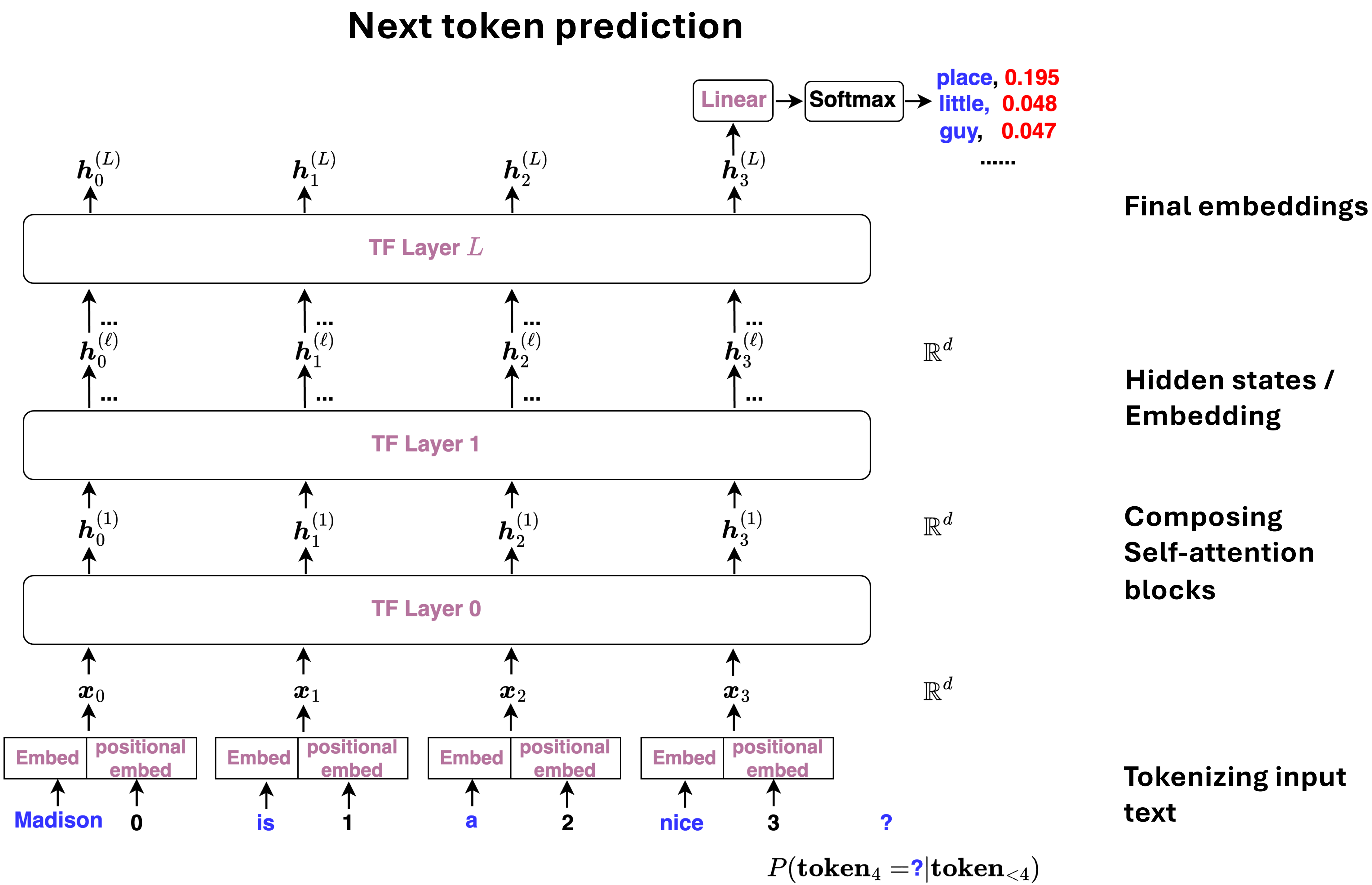
What do we learn?

- Benefits of composition: two layer >> one layer
- Emergence of learning copying
- OOD generalization reasonably well

Goal of this talk:

Geometric (mechanistic) insights via experiments

A Primer on Transformer

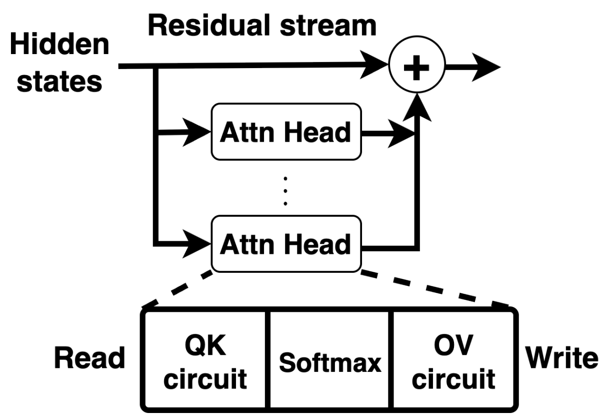

Next token prediction
Linear
Softmax
place, 0.195
little, 0.048
guy, 0.047
......
$h_0^{(L)}$ $h_1^{(L)}$ $h_2^{(L)}$ $h_3^{(L)}$
Final embeddings
TF Layer L
...
$h_0^{(\ell)}$ $h_1^{(\ell)}$ $h_2^{(\ell)}$ $h_3^{(\ell)}$
$\mathbb{R}^d$
Hidden states / Embedding
TF Layer 1
$h_0^{(1)}$ $h_1^{(1)}$ $h_2^{(1)}$ $h_3^{(1)}$
$\mathbb{R}^d$
Composing Self-attention blocks
TF Layer 0
x_0 x_1 x_2 x_3
$\mathbb{R}^d$
Embed
positional embed
Madison 0
is 1
a 2
nice 3
?
Tokenizing input text
$P(\mathbf{token}_4 = ?|\mathbf{token}_{<4})$

A simple intro to self-attention

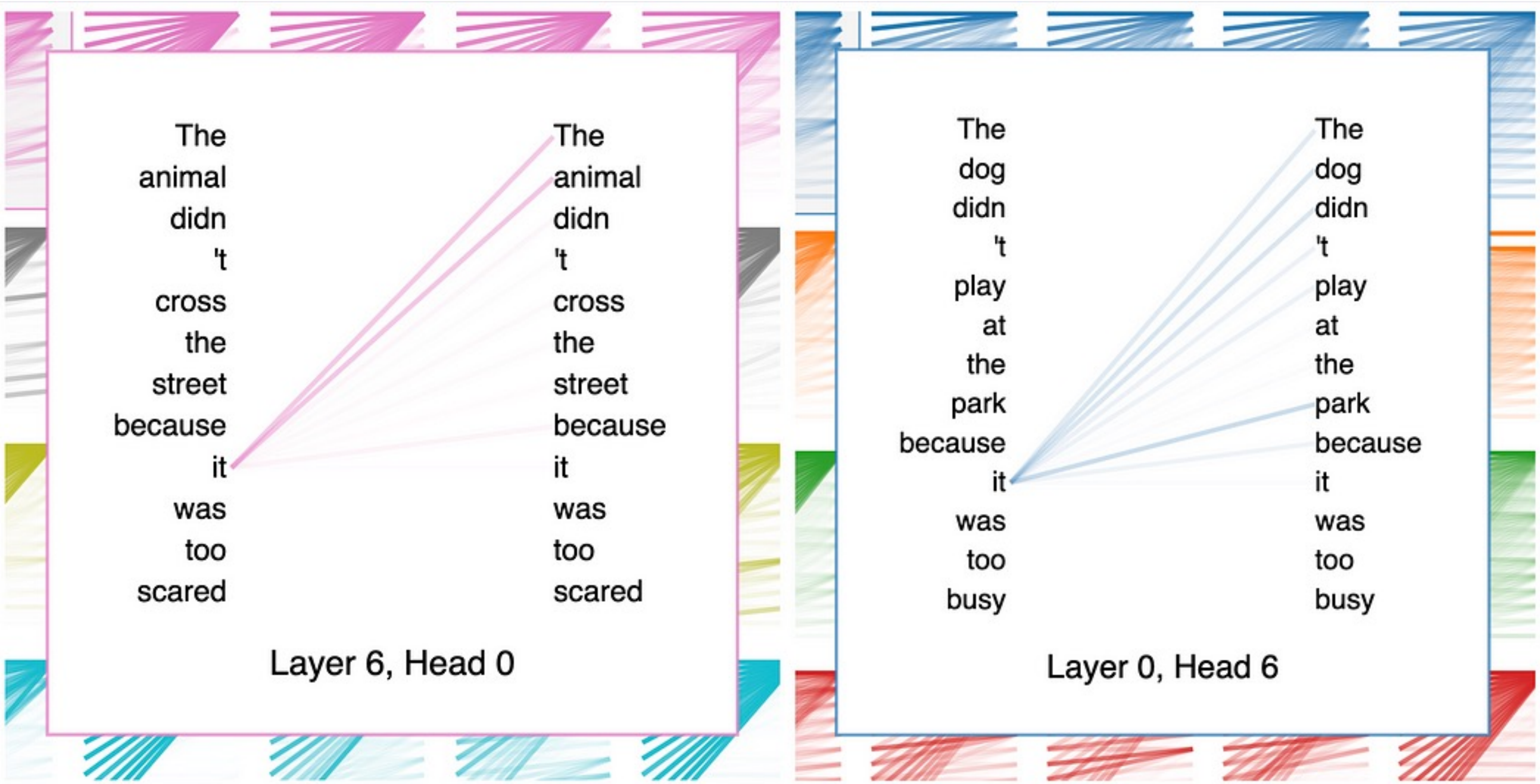
- Input or hidden states $\boldsymbol{X} \in \mathbb{R}^{T \times d}$, T is seq length, d is embed dim

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

- Attention matrix: $T \times T$ similarities of hidden states between pairs of hidden states



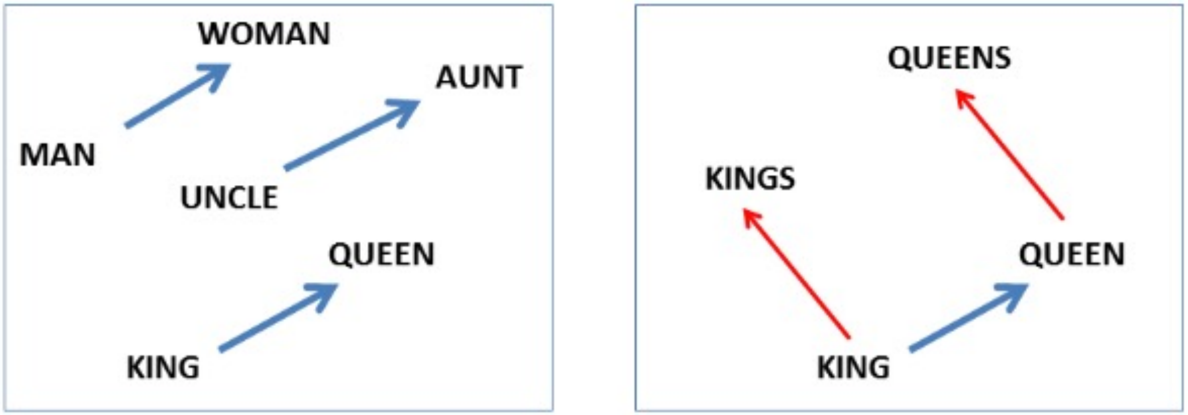
See Elhage et. al., 2021



GPT-2 example, Credit: https://mlops.community/

What do hidden states represent?

- In pre-Transformer age, word meaning is decomposed into vectors of latent concepts/factors



Word embedding, “gender” factor + “royalty” factor

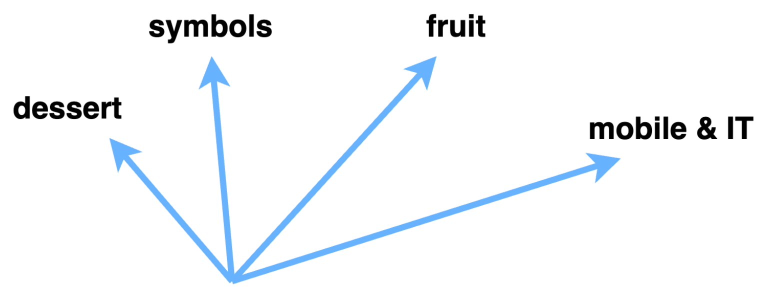
- Classical stats: PCA and factor analysis, e.g., latent factor that drives stock market or gene expression or network community structure

See Mikolov et. al., 2013

Linear representation hypothesis

- Dictionary learning: find of vectors as “base concepts”
- Dictionary size much larger than embedding dimension
- Then hidden state vector is a sparse linear combination of “base concepts” (feature superposition)

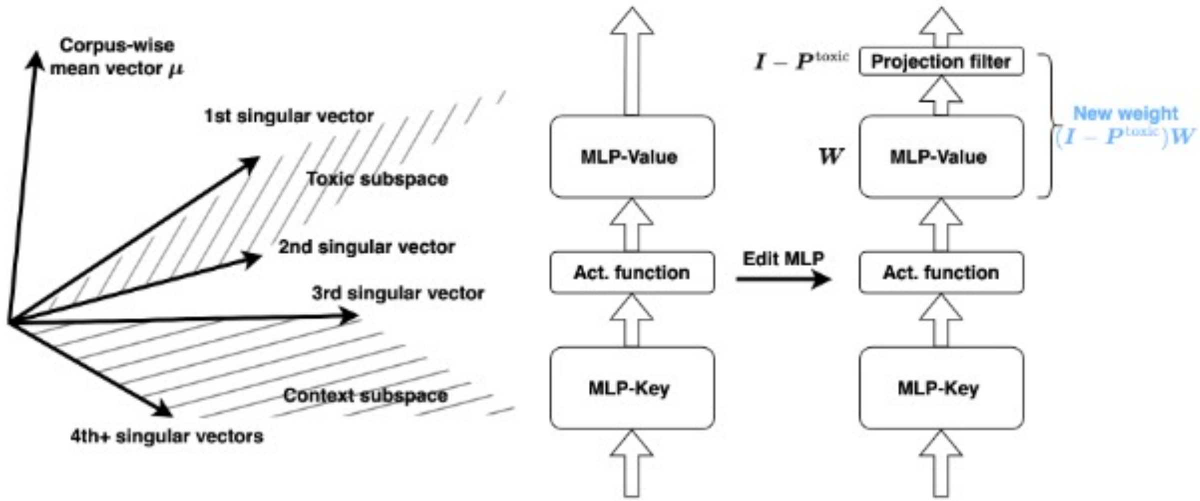
$$\text{apple} = 0.09\,\text{“dessert”} + 0.11\,\text{“organism”} + 0.16\,\text{“fruit”} + 0.22\,\text{“mobile\&IT”} + 0.42\,\text{“other”}.$$



- Anthropic and OpenAI’s interpretability research

Linear representation hypothesis

- A large literature on alignment, model editing [UDHZH, 2024]



	Top Tokens (Layer 14)	Interpretation
μ	, and the - in (" .	Frequent tokens, stopwords
1st svec	s**t f**k ucker b***h slut F**k holes	Toxic tokens
2nd svec	damn really kinda stupid s**t goddamn	Toxic tokens
3rd svec	disclaimer Opinion LĤ Statement Disclaimer Brief	Context dependent topics
4th svec	nation globalization paradigm continent empire ocracy	Context dependent topics

Emergence of Subspace Matching

Synthetic experiment "Learning copying with a simple Transformer"

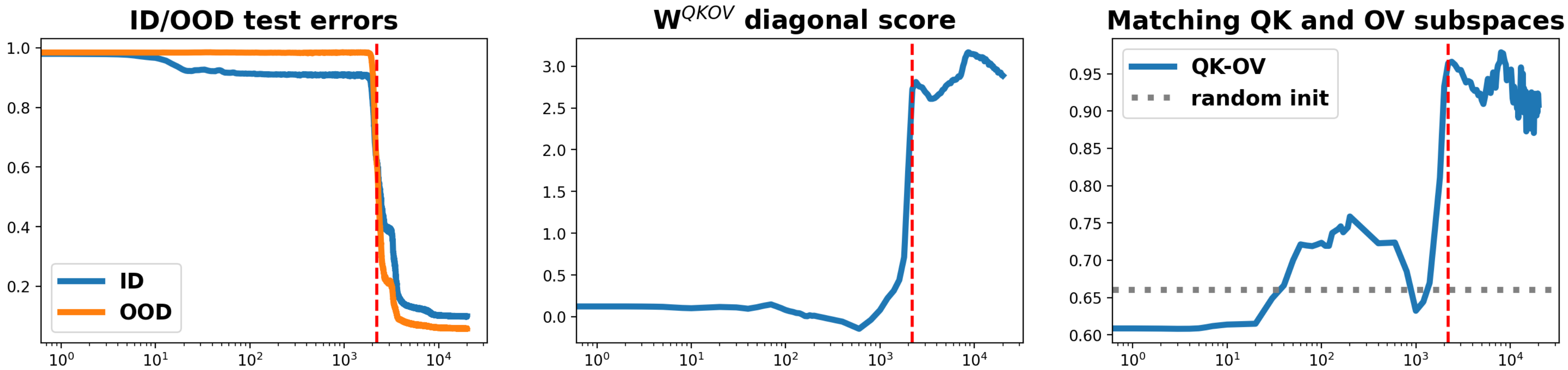
$$\ldots[A],[B],[C]\ldots[A],[B] \xrightarrow{\text{next-token prediction}} \ldots[A],[B],[C]\ldots[A],[B],{\color{blue}[C]}$$

- 2-layer 1-head no-MLP TF: $\mathrm{TF}(\boldsymbol{X}) = \mathrm{MSA}(\mathrm{MSA}(\boldsymbol{X};\boldsymbol{W});\widetilde{\boldsymbol{W}})$

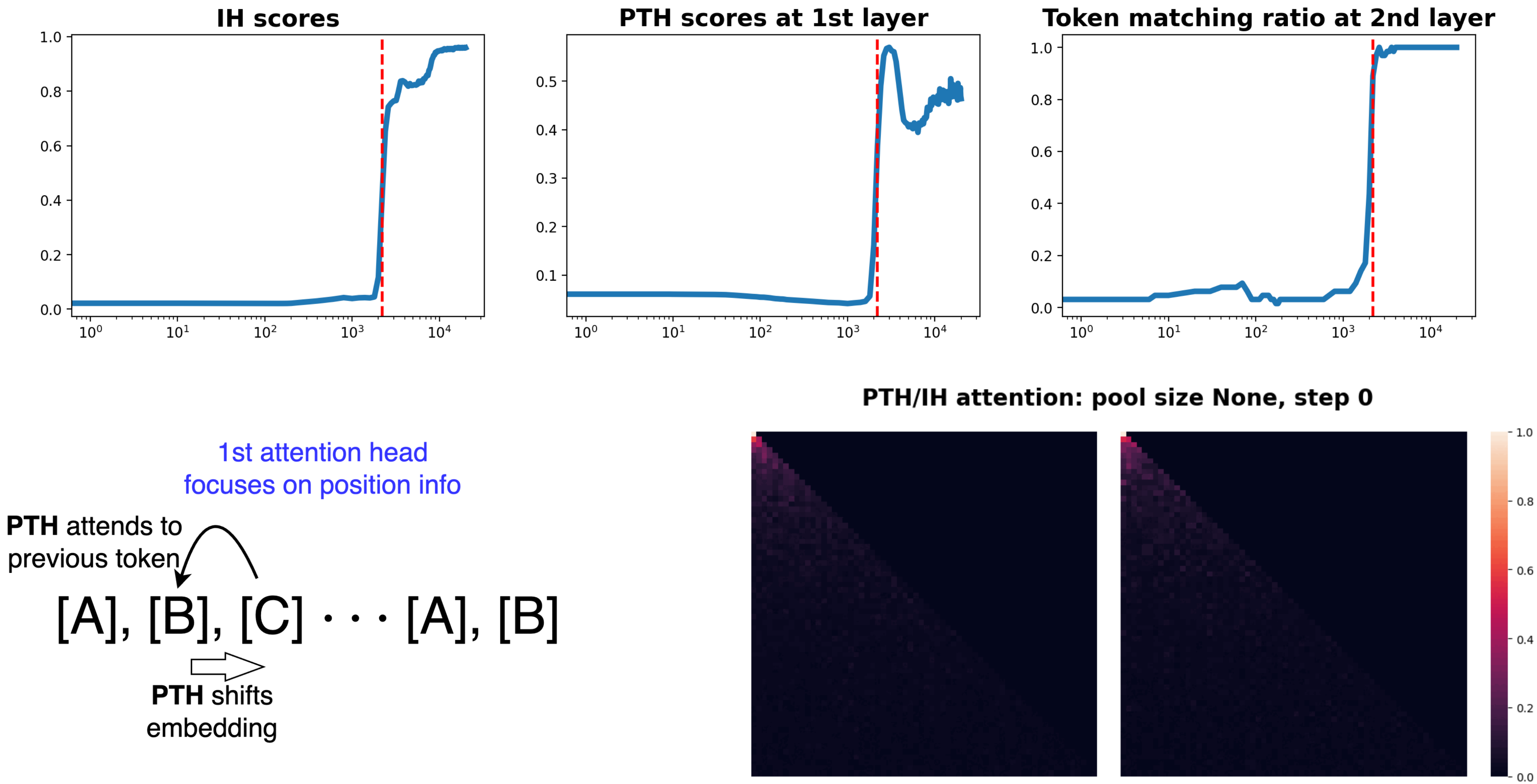
1st layer $\mathrm{MSA}(\boldsymbol{X};\boldsymbol{W}) := \boldsymbol{X} + \mathrm{Softmax}\underbrace{(\boldsymbol{X}\boldsymbol{W}_{\mathrm{QK}}\boldsymbol{X}^\top)}_{\text{QK circuit reads and matches info from stream}}\underbrace{\boldsymbol{X}\boldsymbol{W}_{\mathrm{OV}}^\top}_{\text{OV circuit writes and adds info to stream}}$

What compositional structure enables copying?

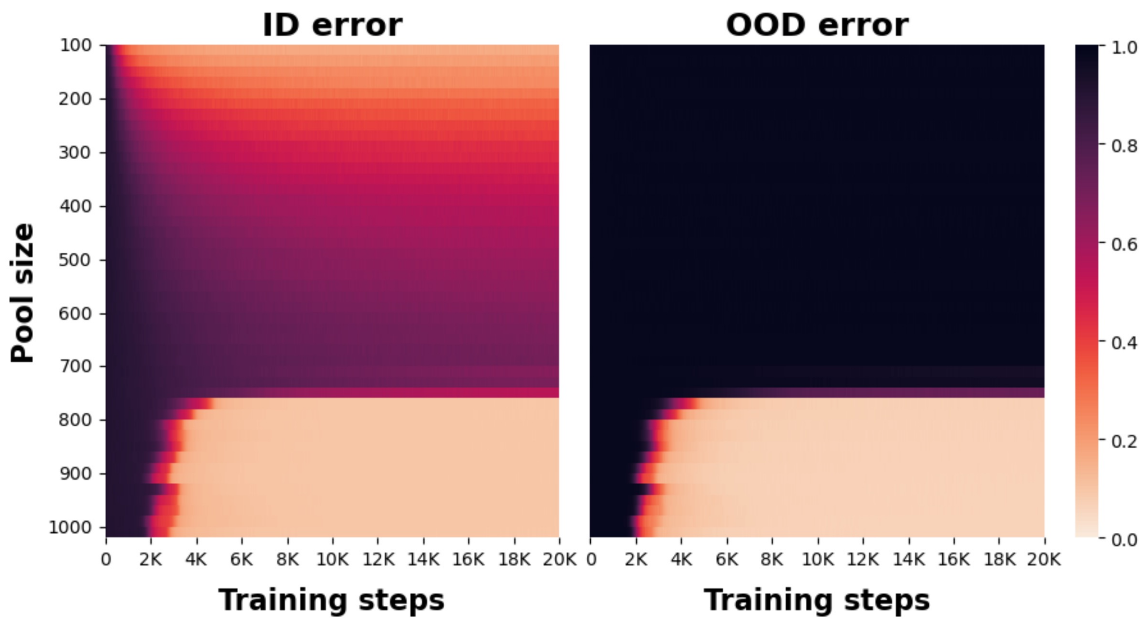
2nd layer $\mathrm{MSA}(\boldsymbol{X};\widetilde{\boldsymbol{W}}) := \boldsymbol{X} + \mathrm{Softmax}\underbrace{(\boldsymbol{X}\widetilde{\boldsymbol{W}}_{\mathrm{QK}}\boldsymbol{X}^\top)}_{\text{QK circuit reads and matches info from stream}}\underbrace{\boldsymbol{X}\widetilde{\boldsymbol{W}}_{\mathrm{OV}}^\top}_{\text{OV circuit writes and adds info to stream}}$



- Diagonal score measures normalized avg diagonal entries of $\widetilde{\boldsymbol{W}}_{QK}\boldsymbol{W}_{OV}$
- Subspace matching: generalized cosine sim between two principal subspaces ($r = 10$): $\mathrm{sim}(\mathcal{P}_{\mathrm{QK}}, \mathcal{P}_{\mathrm{OV}}) := \sigma_{\max}(\boldsymbol{U}^\top \boldsymbol{V})$

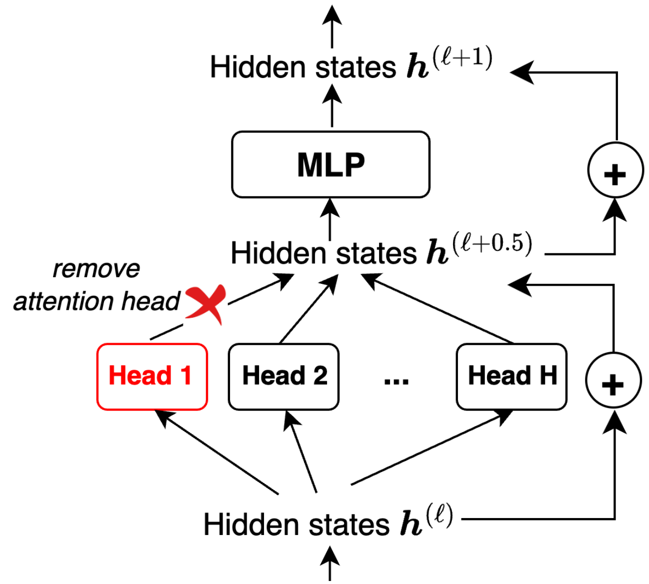

IH scores
PTH scores at 1st layer
Token matching ratio at 2nd layer
PTH/IH attention: pool size None, step 0
1st attention head
focuses on position info
PTH attends to
previous token
[A], [B], [C] · · · [A], [B]
PTH shifts
embedding

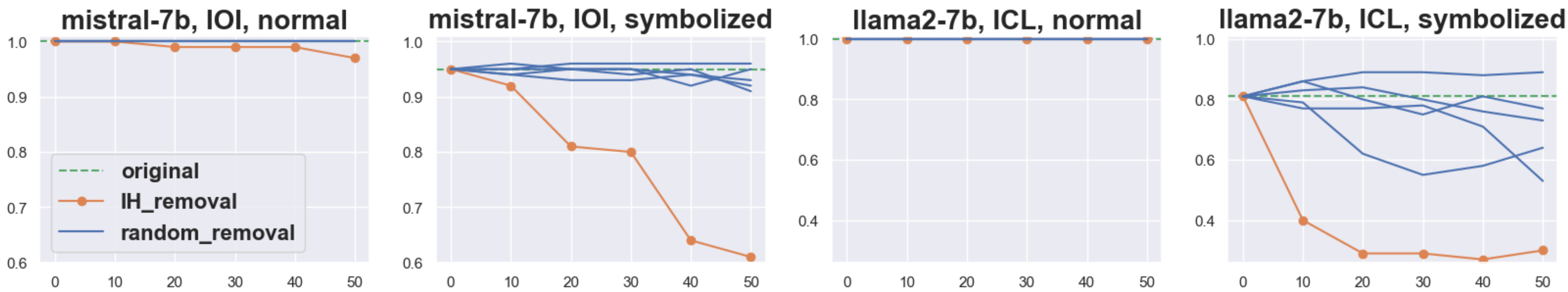
- **Subspace matching**: First layer "writing circuit" (OV) matches second layer "reading circuit" (QK)
- **Complementary roles**: first layer focuses on positional info, second layer token info
- **Clear phase transition**: critical thresholds in both diversity and training steps



LLM experiment "Symbolized language reasoning"

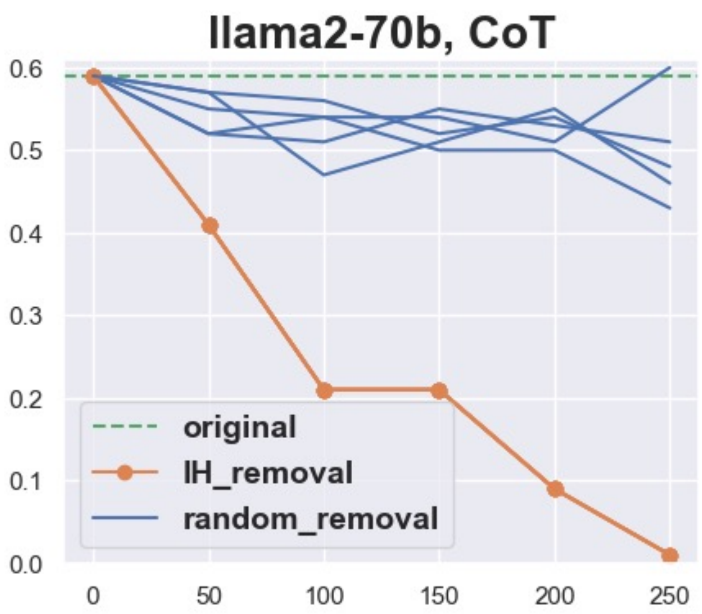
- Many attention heads in LLMs (even GPT2-small has 12*12 heads)
- Ranking heads and screen top ~50 as induction heads



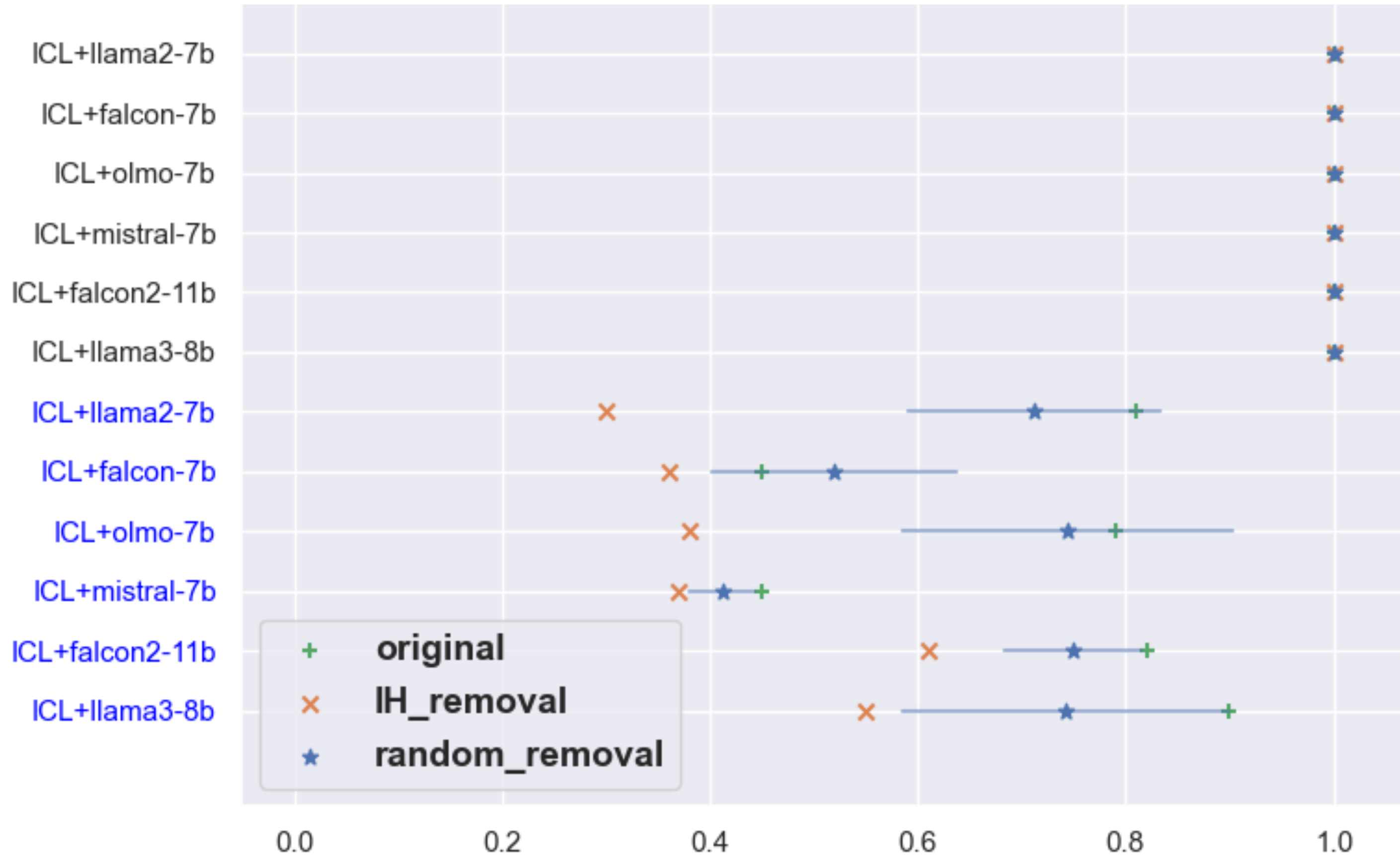


OOD generalization depends crucially on IHs

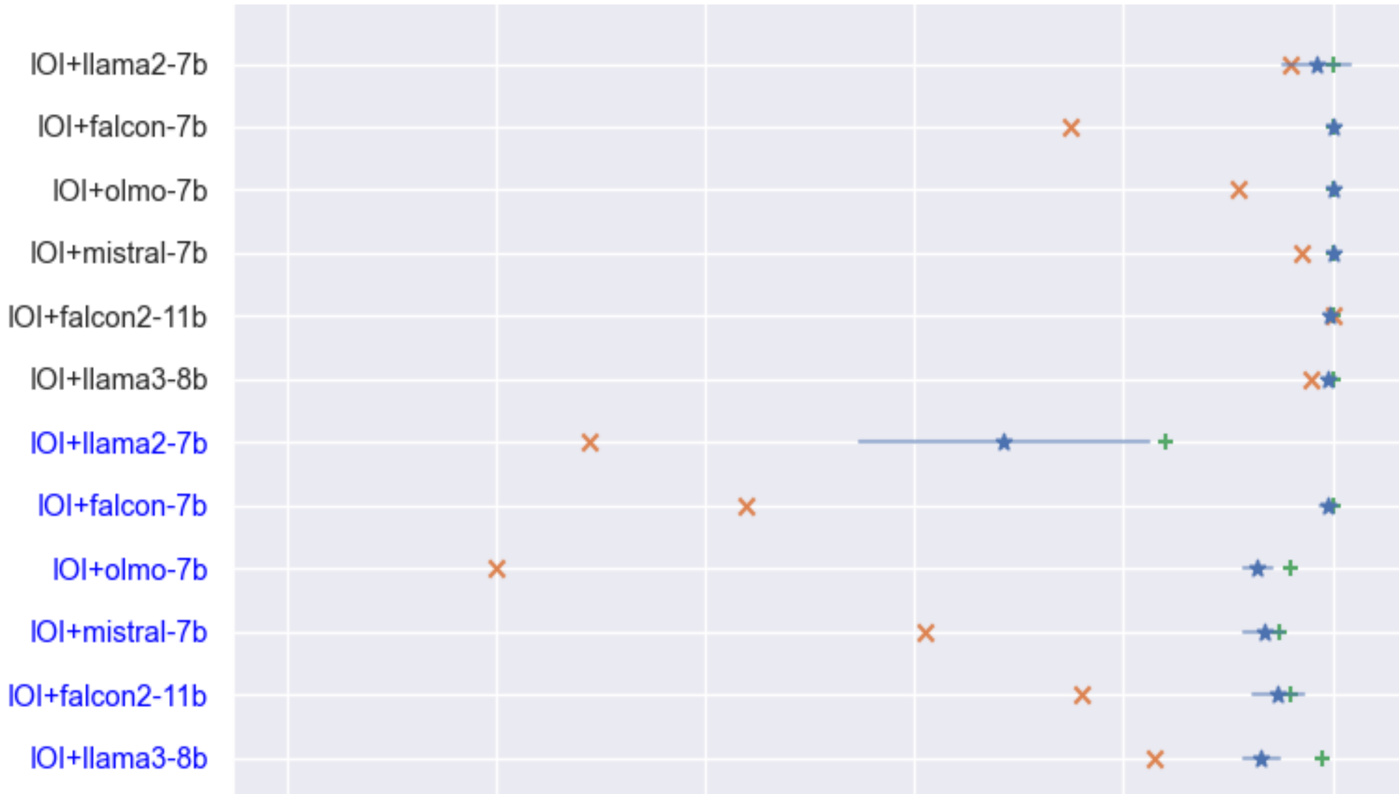
- LLMs on normal prompts are insensitive to IH removal (memorization)
- In contrast, LLMs on symbolized (OOD) prompts depend on IHs
- Same crucial dependence for CoT on GSM8K



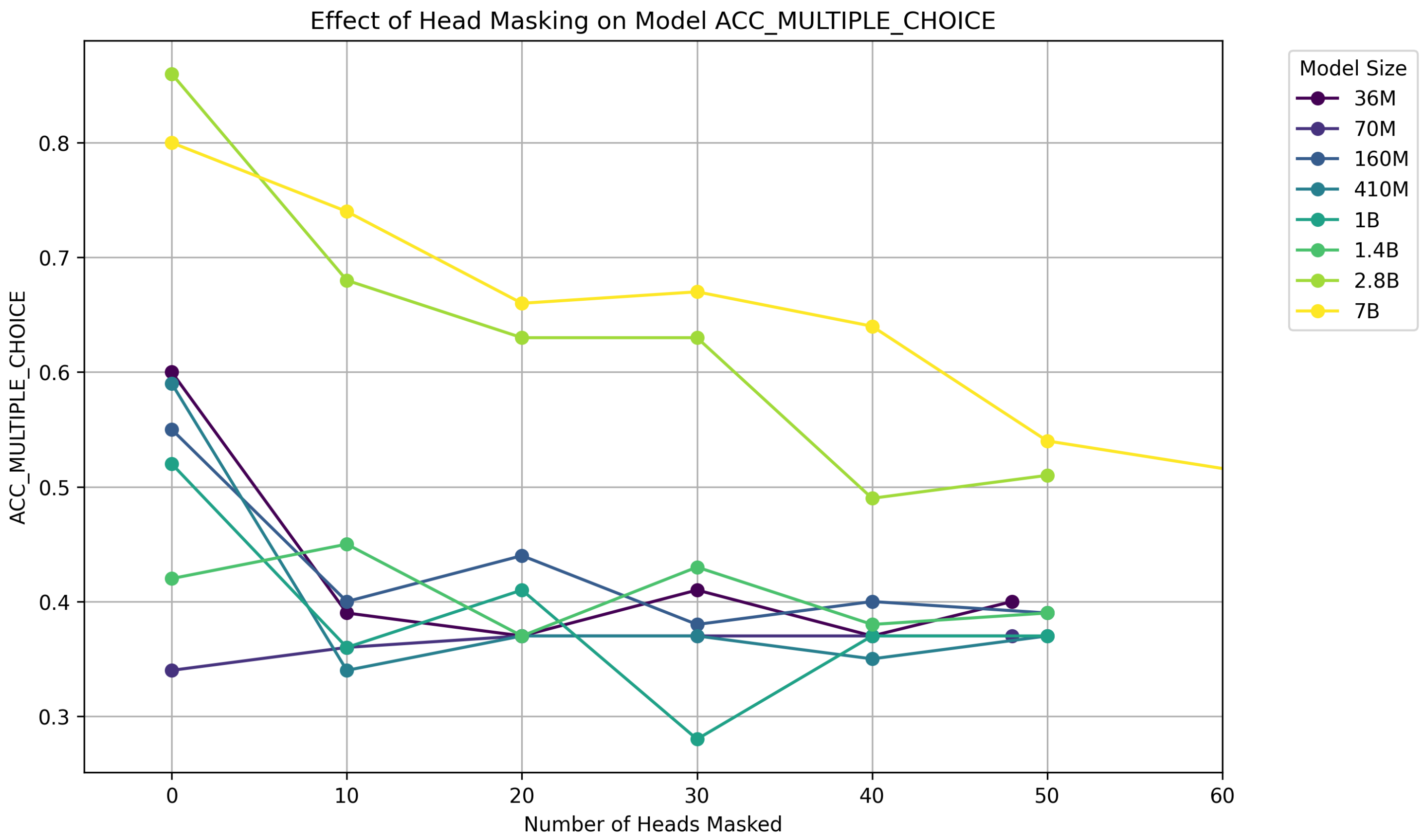
OOD generalization depends crucially on IHs



OOD generalization depends crucially on IHs



OOD generalization depends crucially on IHs: scaling experiments



Common Subspace Representation Hypothesis

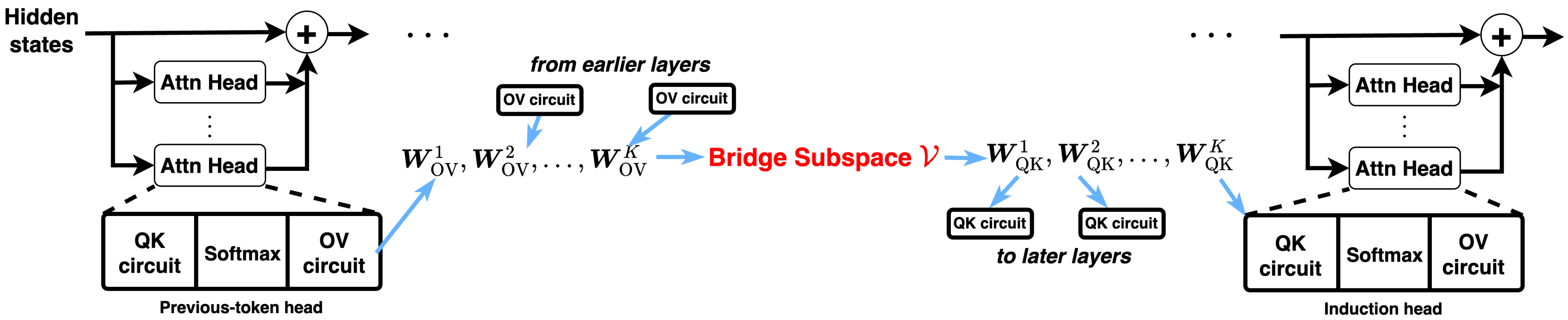
How subspace matching works in LLMs

- Multi-layer, multi-head, how do two layers match?
- Generalizes the linear representation hypothesis

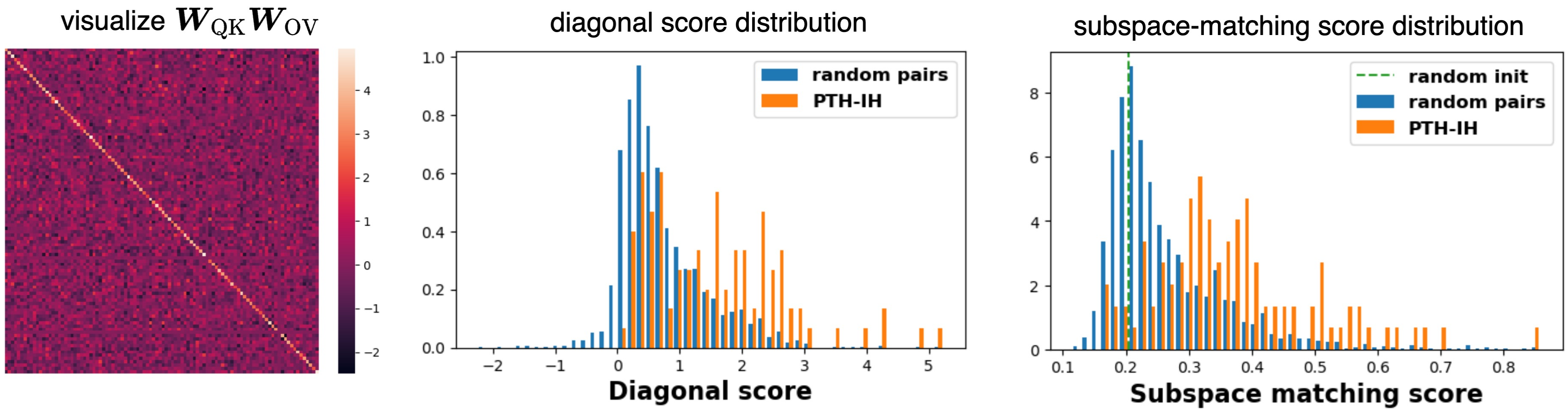
How subspace matching works in LLMs

- Multi-layer, multi-head, how do two layers match?
- Generalizes the linear representation hypothesis
- **Bridge subspace** in ideal form

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



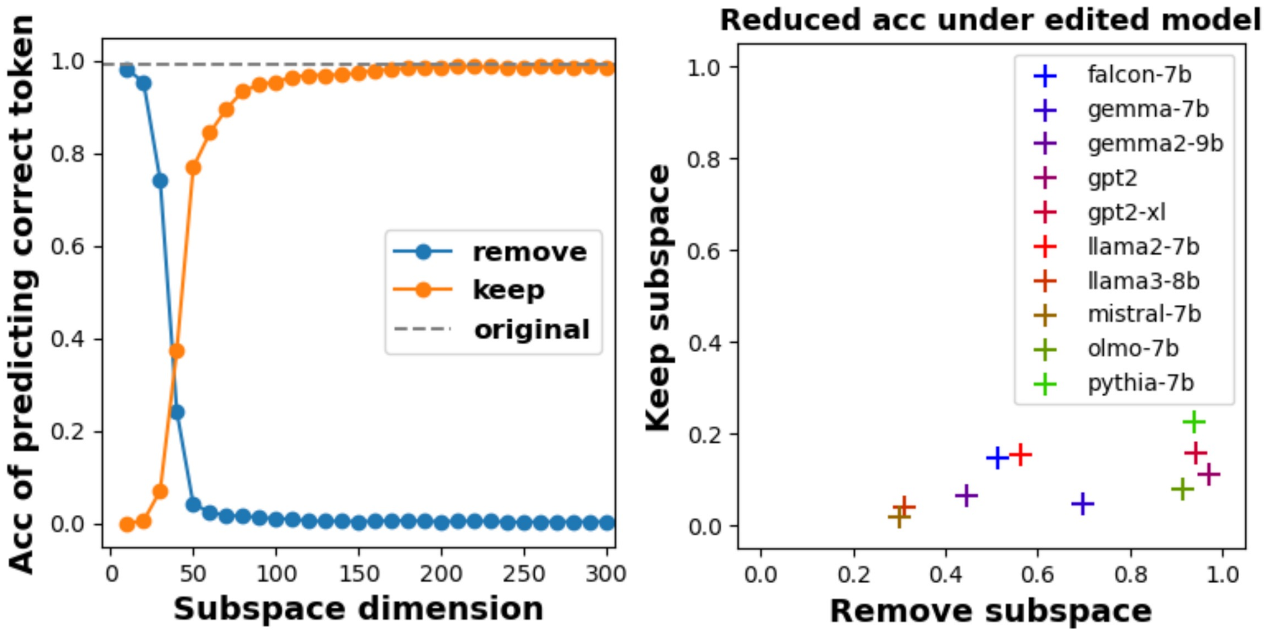
Pairwise matching suggests shared global structure



GPT-2 on copying task

- Strong pairwise matching among top-ranked PTHs and IHs

Impact of removing bridge subspace



Left: GPT-2, Right: various LLMs

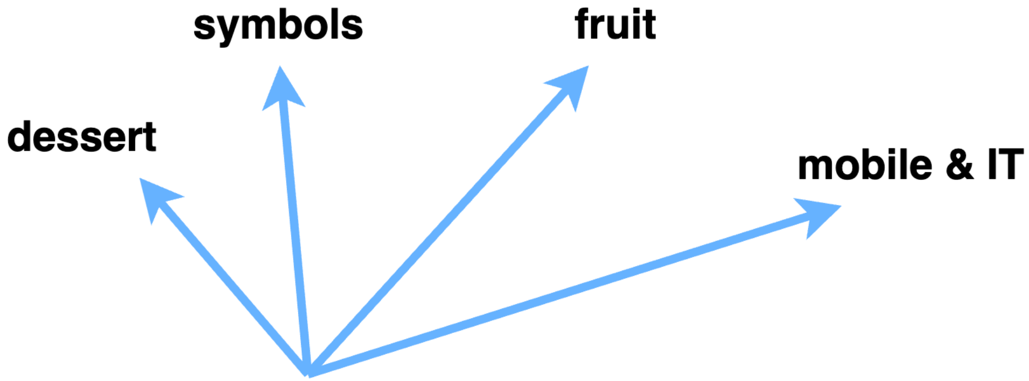
- Calculate **bridge subspace** by pooled SVD: $\boldsymbol{V} = \mathrm{svd}_r([\boldsymbol{W}_{\mathrm{QK}}^1, \ldots, \boldsymbol{W}_{\mathrm{QK}}^M])$
- Two projection applied to weight matrices

$$\boldsymbol{W}_{\mathrm{QK}} \leftarrow \boldsymbol{W}_{\mathrm{QK}} \boldsymbol{V}\boldsymbol{V}^\top \qquad \boldsymbol{W}_{\mathrm{QK}} \leftarrow \boldsymbol{W}_{\mathrm{QK}}(\boldsymbol{I}_d - \boldsymbol{V}\boldsymbol{V}^\top)$$

(Speculative) take-away messages

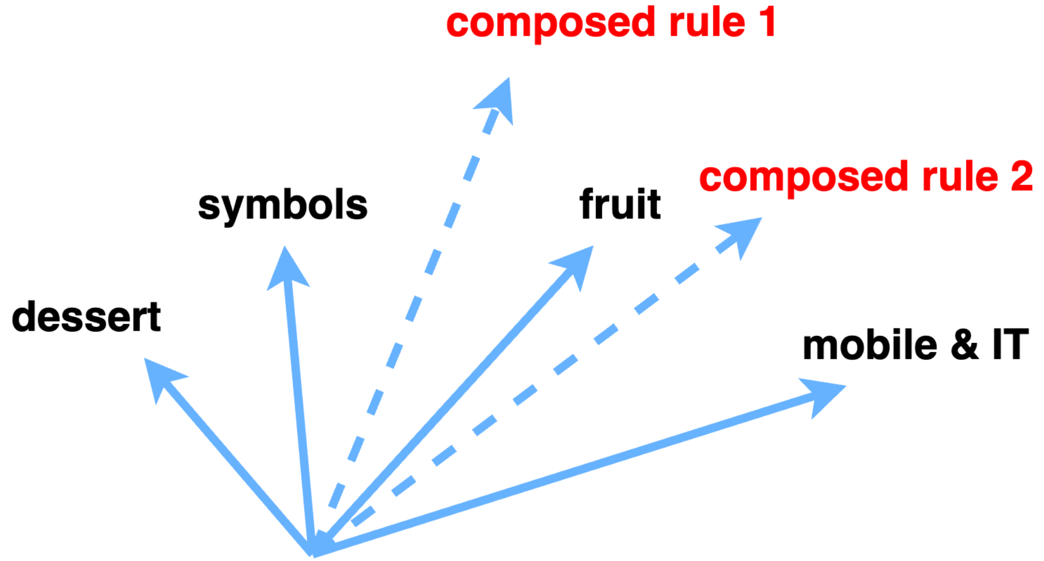
What do hidden states *really* represent

- Concept subspaces



What do hidden states *really* represent

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



Thank you!

References

- Trevor Hastie, Robert Tibshirani, Jerome H. Friedman, and Jerome H. Friedman. The elements of statistical learning: data mining, inference, and prediction. *Vol. 2. New York: Springer*, 2009.
- Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. "Reconciling modern machine-learning practice and the classical bias–variance trade-off." *Proceedings of the National Academy of Sciences 116, no. 32 (2019)*: 15849-15854.
- Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. *In The Eleventh International Conference on Learning Representations,* 2023.
- Frieda Rong. Extrapolating to unnatural language processing with gpt-3's in-context learning: The good, the bad, and the mysterious, 2021.
- Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html

References

- Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.
- Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. "Linguistic regularities in continuous space word representations." *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*. 2013.
- Abbe, Emmanuel, Samy Bengio, Aryo Lotfi, and Kevin Rizk. "Generalization on the unseen, logic reasoning and degree curriculum." *In International Conference on Machine Learning, pp. 31-60. PMLR*, 2023.
- Rheeya Uppaal, Apratim Dey, Yiting He, Yiqiao Zhong, Junjie Hu, "Model Editing as a Robust and Denoised variant of DPO: A Case Study on Toxicity", 2024
- Pan, Jane, Tianyu Gao, Howard Chen, and Danqi Chen. "What In-Context Learning" Learns" In-Context: Disentangling Task Recognition and Task Learning." In The 61st Annual Meeting Of The Association For Computational Linguistics. 2023.