Transformer Circuits: Mathematical Framework and In-context Learning

Hwai-Liang Tung & Yu Huang

Wharton Statistics & Data Science

March 25, 2025

- Elhage, et al., "*A Mathematical Framework for Transformer Circuits*", Transformer Circuits Thread, 2021.
- Olsson, et al., "*In-context Learning and Induction Heads*", Transformer Circuits Thread, 2022.

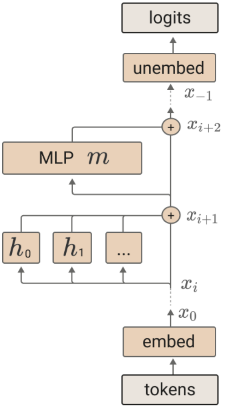
Outline

Mathematical Framework

Induction head & In-context learning

Outline

Review of Transformers



The final logits are produced by applying the unembedding.

$$T(t) = W_U x_{-1}$$

An MLP layer, m, is run and added to the residual stream.

$$x_{i+2} = x_{i+1} + m(x_{i+1})$$

Each attention head, h, is run and added to the residual stream.

$$x_{i+1} = x_i + \sum_{h \in H_i} h(x_i)$$

One residual block

Token embedding.

$$x_0 = W_E t$$

Review of Attention Heads

- We can describe applying attention as

$$
\begin{aligned}
h(x) &= (Id \otimes W_O) \cdot (A \otimes Id) \cdot (Id \otimes W_V) \cdot x \\
&= (A \otimes W_O W_V) \cdot x
\end{aligned}
$$

  - W_V computes the value vector for each token
  - W_O projects the result vector for each token
  - $A = \text{softmax}(x^T W_Q^T W^K x)$
- Both W_Q, W_K and W_O, W_V always operate together so we may let $W_{OV} = W_O W_V$, $W_{QK} = W_Q^T W_K$
- $(A^{h_2} \otimes W_{OV}^{h_2}) \cdot (A^{h_1} \otimes W_{OV}^{h_1}) = (A^{h_2} A^{h_1}) \otimes (W_{OV}^{h_2} W_{OV}^{h_1})$

Zero-Layer Transformers

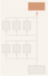
- A zero-layer transformer embeds an input token and unembeds it to produce logits predicting the next token
- Can represent this as $T = W_U W_E$
  - W_E is token embedding matrix
  - W_U is token unembedding matrix
- Optimal behavior of $W_U W_E$ is to approximate bigram log-likelihood

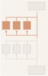
Overview of Composition of Attention Heads

- Attention heads read in a subspace of the residual stream and writes to another subspace
- Once we have two or more layers we have composition of attention heads
- W_Q, W_K, W_V read in subspaces affected by a previous head and perform Q-composition, K-composition, V-composition respectively
- Q-composition and K-composition affect attention pattern
- V-composition affects what information is moved when attending to a certain position

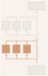
Path Expansion of Logits

$$T = \mathrm{Id} \otimes W_U \cdot \left(Id + \sum_{h \in H_2} A^h \otimes W_{OV}^h\right) \cdot \left(Id + \sum_{h \in H_1} A^h \otimes W_{OV}^h\right) \cdot \mathrm{Id} \otimes W_E$$





The second **attention layer** has multiple attention heads which add into the residual stream

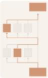


The first **attention layer** has multiple attention heads which add into the residual stream

$$= \mathrm{Id} \otimes W_U W_E + \sum_{h \in H_1 \cup H_2} A^h \otimes (W_U W_{OV}^h W_E) + \sum_{h_2 \in H_2} \sum_{h_1 \in H_1} (A^{h_2} A^{h_1}) \otimes (W_U W_{OV}^{h_2} W_{OV}^{h_1} W_E)$$

"Direct path" term contributes to bigram statistics.

The **individual attention head** terms describe the effects of individual attention heads in linking input tokens to logits, similar to those we saw in the one layer model.



The **virtual attention head** terms correspond to V-composition of attention heads. They function a lot like individual attention heads, with their own attention patterns (the compositon of the heads patterns) and own OV matrix.

QK-Circuits and OV-Circuits

- Recall $A^h = \text{softmax}^*(t^T \cdot W_E^T W_{QK}^h W_E \cdot t)$ where softmax* denoted the softmax with autoregressive maxing
- Two key matrices present in an attention head
- We can call $W_E^T W_{QK}^h W_E$ the query-key or QK-circuit
- The QK-circuit provides the attention score for every query and key token
- We can call $W_U W_{OV}^h W_E$ the output-value or OV-circuit
- The OV-circuit describes how a given token will affect the output logits if attended to and are involved in copying behavior

Path Expansion of Attention Scores QK-Circuit

- Recall $A^h = \text{softmax}^*(t^T \cdot W_E^T W_{QK}^h W_E \cdot t)$
- Can take a closer look at the operations of A^h
- For the first layer QK-circuit we have

$$C_{QK}^{h \in H_1} = x_0^T W_{QK}^h x_0 = W_E^T W_{QK}^h W_E$$

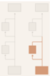
- For the second layer QK-circuit we have

$$C_{QK}^{h \in H_2} = x_1^T W_{QK}^h x_1$$

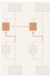
where x_1 is the residual stream after the first layer attention heads

Path Expansion of Attention Scores QK Circuit

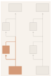
$$C_{QK}^{h\in H_2} = \left(\mathrm{Id}\otimes\mathrm{Id}\otimes W_E^T + \sum_{h_q\in H_1} A^{h_q}\otimes\mathrm{Id}\otimes(W_{OV}^{h_q}W_E)^T\right) \cdot \mathrm{Id}\otimes\mathrm{Id}\otimes W_{QK}^{h} \cdot \left(\mathrm{Id}\otimes\mathrm{Id}\otimes W_E + \sum_{h_k\in H_1} \mathrm{Id}\otimes A^{h_k}\otimes W_{OV}^{h_k}W_E\right)$$



The **"query side" residual stream** at the start of the second layer contains both the layer 1 direct path and layer 1 attention heads. All terms are of the form $\ldots\otimes\mathrm{Id}\otimes\ldots$ because they don't move key information.

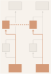


W_{QK} of the second layer head combines both sides into attention scores.

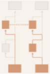


The **"key side" residual stream** at the start of the second layer contains both the layer 1 direct path and attention heads. All terms are of the form $\mathrm{Id}\otimes\ldots$ because they don't move query information.

$$= \mathrm{Id}\otimes\mathrm{Id}\otimes\left(W_E^T W_{QK}^h W_E\right) + \sum_{h_q\in H_1} A^{h_q}\otimes\mathrm{Id}\otimes\left(W_E^T {W_{OV}^{h_q}}^T W_{QK}^h W_E\right)$$

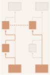


The no composition term. Both first layer follows the direct path on both the key and query side.

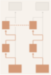


These terms correspond to pure **Q-composition**. A previous attention head is used to generate the query side, but the key side is the first layer direct path.

$$+ \sum_{h_k\in H_1} \mathrm{Id}\otimes A^{h_k}\otimes\left(W_E^T W_{QK}^h W_{OV}^{h_k} W_E\right) + \sum_{h_q\in H_1}\sum_{h_k\in H_1} A^{h_q}\otimes A^{h_k}\otimes\left(W_E^T {W_{OV}^{h_q}}^T W_{QK}^h W_{OV}^{h_k} W_E\right)$$



These terms correspond to pure **K-composition**. A previous attention head is used to generate part of the key, but the query side is the first layer direct path.

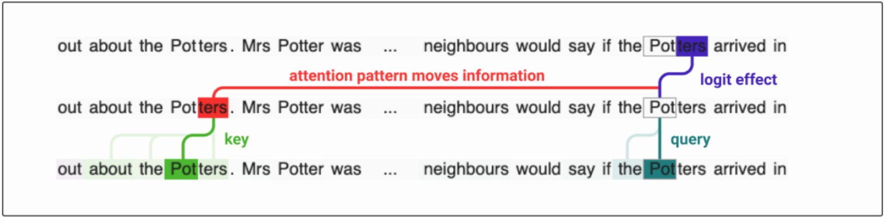


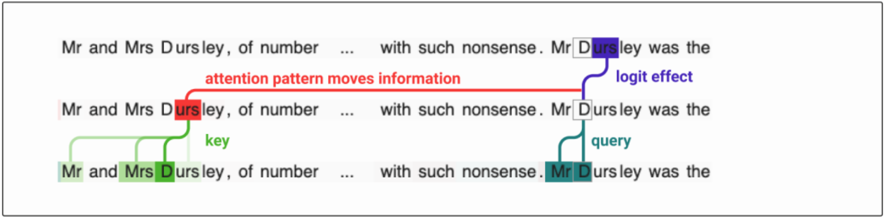
These terms are interactions between both **Q-composition** and **K-composition**. A previous attention head is used to generate the query and key sides.

Induction Heads

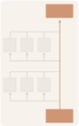
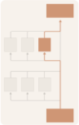
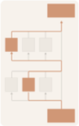
- Starting from two layers a transformer can progress beyond copying: [b]...[a]$\rightarrow$[b]
- Induction heads search for previous examples of the present token and allow copying the next token from the previous example: [a][b]...[a]$\rightarrow$[b]
- Suggested that induction heads are composed of two parts: a "copying" OV circuit matrix and "same matching" QK circuit matrix

Induction Heads





Term Importance Analysis

Type	Example	Equation	Marginal Loss Reduction
direct path order 0		$W_U W_E$	**- 1.8 nats** relative to uniform predictions -1.8 nats/term (- 1.8 nats / 1 term)
individual attention head order 1		$A^h \otimes (W_U W_{OV}^h W_E)$	**- 5.2 nats** relative to only using direct path -0.2 nats/term (5.2 nats / 24 terms)
virtual attention head order 2		$(A^{h_2} A^{h_1}) \otimes (W_U W_{OV}^{h_2} W_{OV}^{h_1} W_E)$	**- 0.3 nats** relative to only using above -0.002 nats/term (0.3 nats / 144 terms)

Term Importance Analysis

- Virtual attention heads have little impact but this may change with more layers
- This composition of attention heads may be able to implement the attending to the start of a clause or sentence
- The number of virtual attention heads grows faster than the number of individual heads

Outline

Remarkable emergent ability for LLMs: In-context learning (ICL)

Remarkable emergent ability for LLMs: In-context learning (ICL)

For pre-trained LLMs:

Remarkable emergent ability for LLMs: In-context learning (ICL)

For pre-trained LLMs:

- At inference time, model receives in-contexting examples from certain task.

Remarkable emergent ability for LLMs: In-context learning (ICL)

For pre-trained LLMs:

- At inference time, model receives in-contexting examples from certain task.
- Given a new query input, model can return corresponding output **without further fine-tuning.**

Remarkable emergent ability for LLMs: In-context learning (ICL)

For pre-trained LLMs:

- At inference time, model receives in-contexting examples from certain task.
- Given a new query input, model can return corresponding output **without further fine-tuning.**

Circulation revenue has increased by 5% in Finland. // Positive

Panostaja did not disclose the purchase price. // Neutral

Paying off the national debt will be extremely painful. // Negative

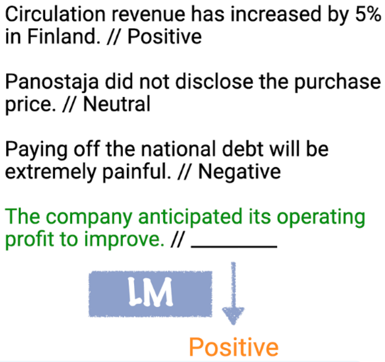
The company anticipated its operating profit to improve. // ________

Positive

Remarkable emergent ability for LLMs: In-context learning (ICL)

For pre-trained LLMs:

- At inference time, model receives in-contexting examples from certain task.
- Given a new query input, model can return corresponding output **without further fine-tuning.**



General framing: decreasing loss at increasing token indices

Remarkable emergent ability for LLMs: In-context learning (ICL)

For pre-trained LLMs:

- At inference time, model receives in-contexting examples from certain task.
- Given a new query input, model can return corresponding output **without further fine-tuning.**

Circulation revenue has increased by 5% in Finland. // Positive

Panostaja did not disclose the purchase price. // Neutral

Paying off the national debt will be extremely painful. // Negative

The company anticipated its operating profit to improve. // ________

Positive

ICL score: the loss of the 500th token in the context minus the average loss of the 50th token in the context

What is an induction head?

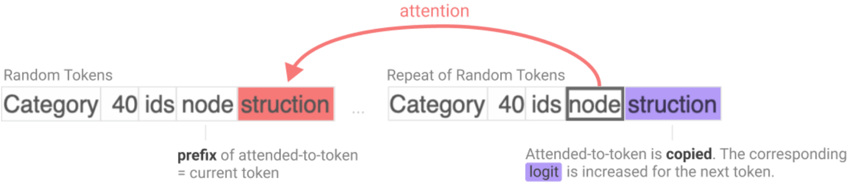
Two Key Properties

- **Prefix matching:** attends back to previous tokens that were followed by the current and/or recent tokens.
- **Copying:** The head's output increases the logit corresponding to the attended-to token.

What is an induction head?

Two Key Properties

- **Prefix matching:** attends back to previous tokens that were followed by the current and/or recent tokens.
- **Copying:** The head's output increases the logit corresponding to the attended-to token.



Does **induction head** provide the primary mechanism for the majority of ICL for transformers in general?

SUMMARY OF EVIDENCE FOR SUB-CLAIMS (STRONGEST ARGUMENT FOR EACH)

	Small Attention-Only	Small with MLPs	Large Models
Contributes Some	Strong, Causal	Strong, Causal	Medium, Correlational & Mechanistic
Contributes Majority	Strong, Causal	Medium, Causal	Medium, Correlational

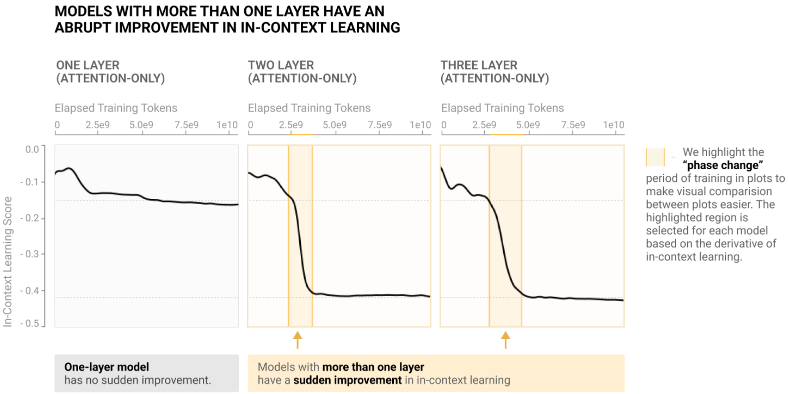
Argument 1: Macroscopic co-occurrence

Transformer language models undergo a **"phase change"** early in training, during which i). induction heads form and ii). simultaneously ICL improves dramatically.

STRENGTH OF ARGUMENT FOR SUB-CLAIMS

	Small Attention-Only	Small with MLPs	Large Models
Contributes Some	Medium, Correlational	Medium, Correlational	Medium, Correlational
Contributes Majority	Medium, Correlational	Medium, Correlational	Medium, Correlational

ICL improves dramatically

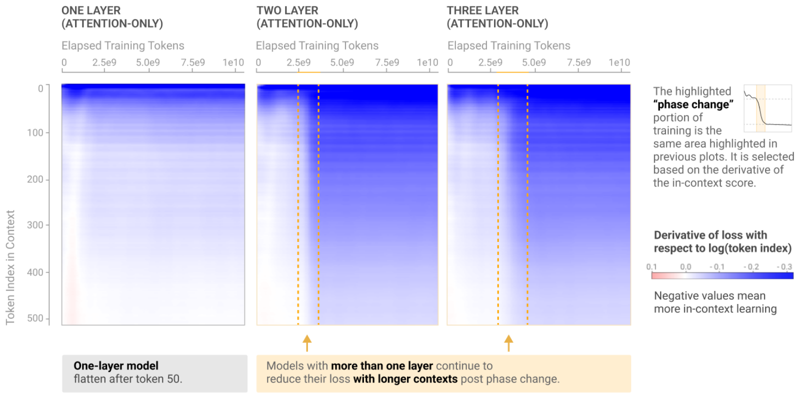


An **artifact** of the choice to define ICL in terms of the difference between the 500th and 50th tokens?

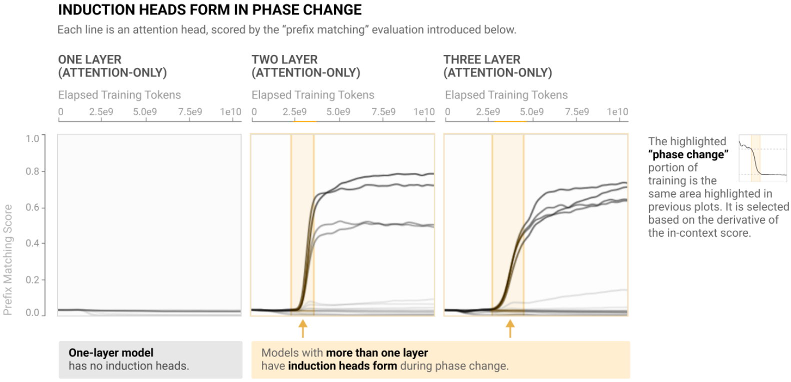
ICL improves dramatically

DERIVATIVE OF LOSS WITH RESPECT TO LOG TOKEN INDEX

The rate at which loss decreases with increasing token index can be thought of as something like “in-context learning per token”. This appears to be most naturally measured with respect to the log number of tokens.

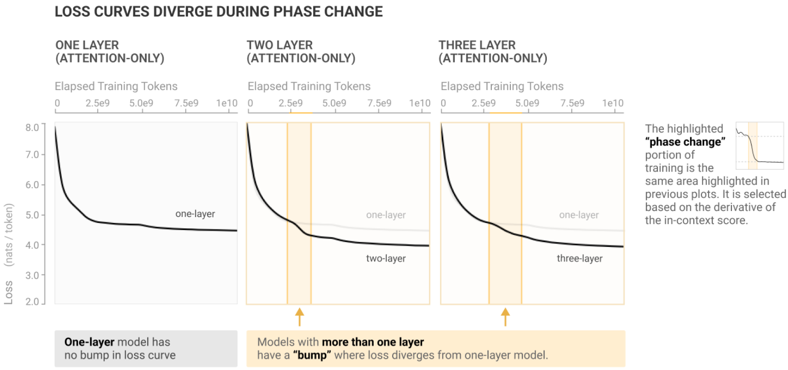


Induction heads form



Suggests some **connection** between induction heads and ICL

The window is a key turning point in training



Behaviors **visible** on the loss curve, unlike subtle ones like emergent arithmetic, signal **broad and significant** changes in model behavior.

The window is a key turning point in training

- Capacity for ICL sharply improves
- Induction heads form
- Loss undergoes a small "bump"

Assessing the Evidence

- Low time resolution on the analysis over training for **larger** models.
- The observed co-occurrence could stem from **other mechanisms** rather than a direct causal link

Argument 2: Macroscopic co-perturbation

When we adjust the transformer architecture to influence induction head formation, the ICL improvement shifts accordingly.

STRENGTH OF ARGUMENT FOR SUB-CLAIMS

	Small Attention-Only	Small with MLPs	Large Models
Contributes Some	Medium, Interventional	Medium, Interventional	Weak, Interventional
Contributes Majority	Medium, Interventional	Medium, Interventional	Weak, Interventional

“Smeared key” architecture

- **Key observation:** the phase change and the corresponding improvement in in-context learning only occurs in transformers with more than one layer.
  - induction heads require a composition of attention heads

"Smeared key" architecture

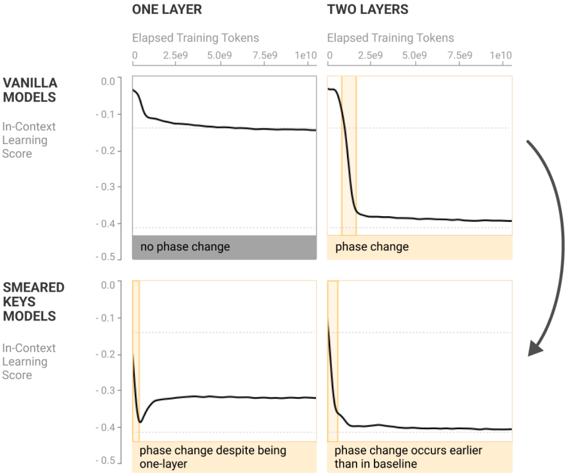
- **Key observation:** the phase change and the corresponding improvement in in-context learning only occurs in transformers with more than one layer.
  - induction heads require a composition of attention heads
- **Predictable minimum** architectural requirements

“Smeared key” architecture

- **Key observation:** the phase change and the corresponding improvement in in-context learning only occurs in transformers with more than one layer.
  - induction heads require a composition of attention heads
- **Predictable minimum** architectural requirements
- **“smeared key” architecture:** for each head h, introducing a trainable α^h with $\sigma(\alpha^h) \in [0,1]$

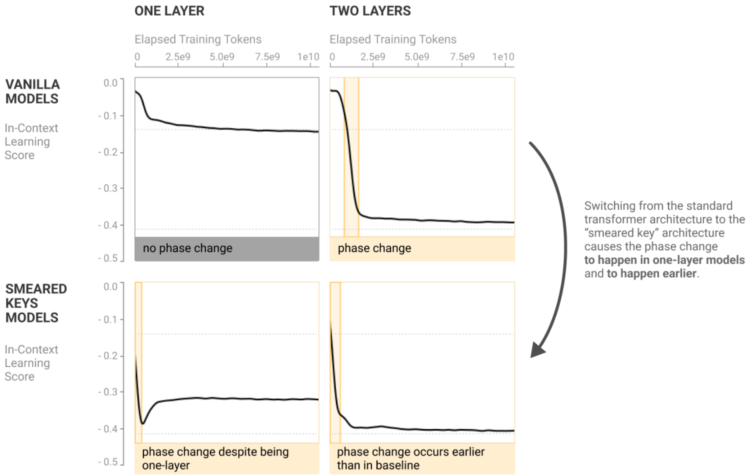
$$k_j^h = \sigma\left(\alpha^h\right) k_j^h + \left(1 - \sigma\left(\alpha^h\right)\right) k_{j-1}^h$$

"Smeared key" architecture



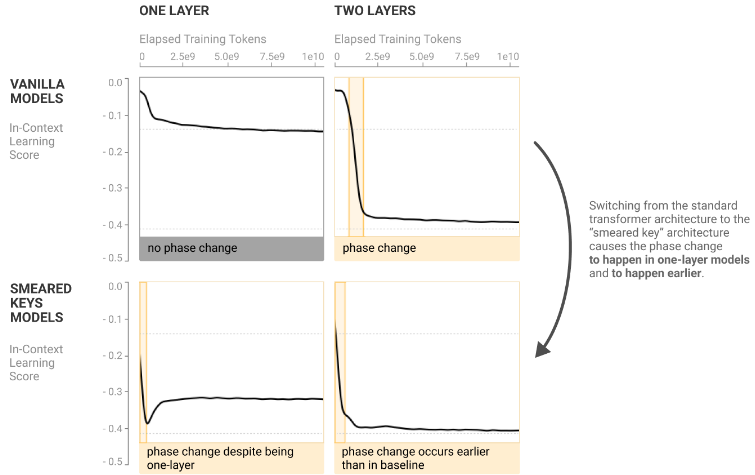
Switching from the standard transformer architecture to the "smeared key" architecture causes the phase change **to happen in one-layer models** and **to happen earlier**.

“Smeared key” architecture



ICL does **indeed form** for one-layer models (when it didn’t before), and it forms **earlier** for two-layer and larger models.

“Smeared key” architecture



induction heads are the **minimal** mechanism for greatly increased ICL, but may **not the whole story** for larger models

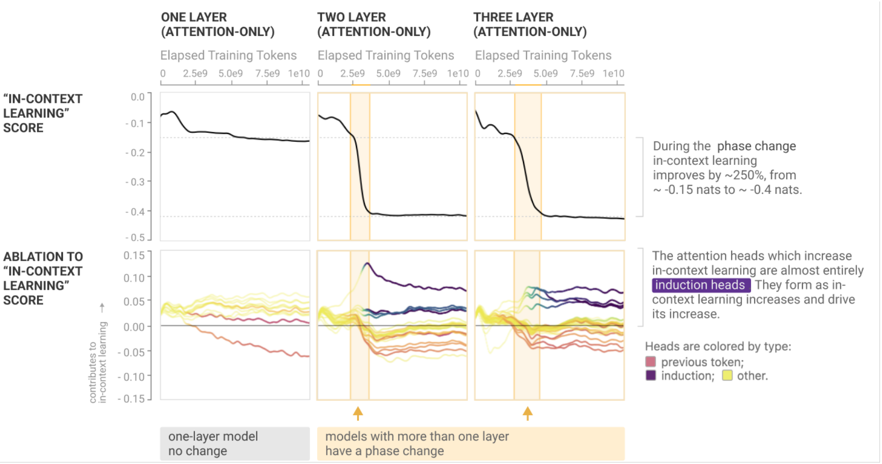
Argument 3: Direct ablation

When we directly ablate induction heads in small models at test-time, the ICL score drops dramatically.

STRENGTH OF ARGUMENT FOR SUB-CLAIMS

	Small Attention-Only	Small with MLPs	Large Models
Contributes Some	Strong, Causal	Strong, Causal	
Contributes Majority	Strong, Causal	Medium, Causal	

Ablations: knocking out induction heads



strongest evidence: almost all the ICL in small attention-only models appears to come from these induction heads

No ablations for full-scale models

Can we further infer that induction heads are the primary mechanism?

- In **attention-only** models, ICL arises solely from attention heads, making their contribution clearer.
- In **MLP** models, interactions between MLP and attention layers could also drive ICL.
- Ablations only measure **marginal effects**, may obscure **individual** head contributions

Argument 4: induction head generality

Despite being defined narrowly as copying random sequences, induction heads can implement surprisingly abstract types of ICL.

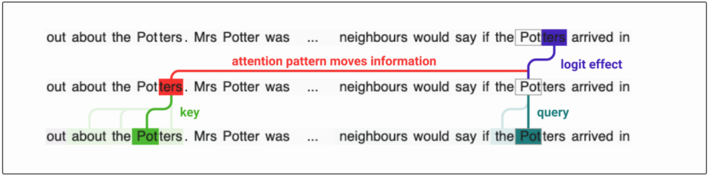
STRENGTH OF ARGUMENT FOR SUB-CLAIMS

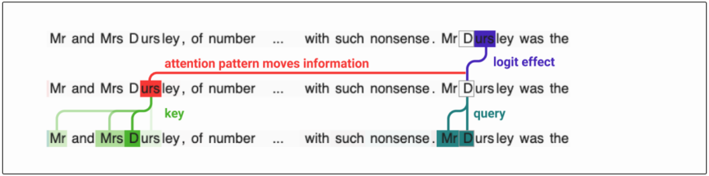
	Small Attention-Only	Small with MLPs	Large Models
Contributes Some			Plausibility
Contributes Majority			Plausibility

- **Induction heads:** empirically copy arbitrary token sequences using a "prefix matching" attention pattern.
- **Goal:** find heads that meet this definition but also perform **more interesting and sophisticated** behaviors

Head	Layer Depth	Copying score (?)	Prefix matching score (?)
Literal copying head	21 / 40	0.89	0.75
Translation head	7 / 40	0.20	0.85
Pattern-matching head	8 / 40	0.69	0.94

Recap: attention & logit effect





Behavior 2: Translation

Induction head from Layer 7 of our 40-layer model, showcasing **translation** between English, French, and German

where the head is attending to predict the *next* token.

<EOT>EN: This is the largest temple that I've ever seen.
FR: C'est le plus grand temple que j'ai jamais vu.
DE: Das ist der größte Tempel, den ich je gesehen habe.

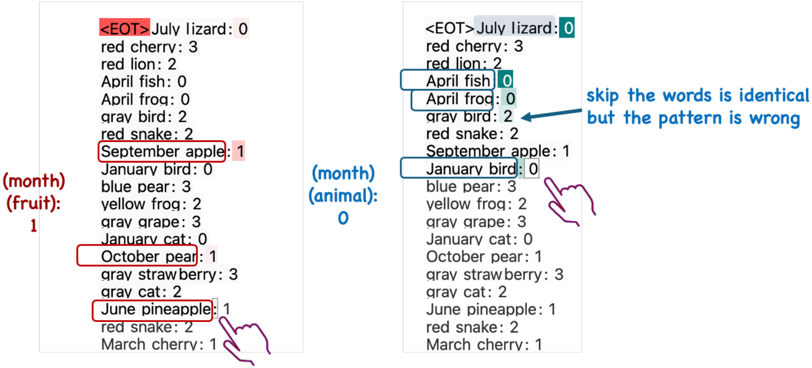
the earlier tokens that contributed to the prediction of the *current* token.

<EOT>EN: This is the largest temple that I've ever seen.
FR: C'est le plus grand temple que j'ai jamais vu.
DE: Das ist der größte Tempel, den ich je gesehen habe.

Behavior 3: Pattern matching

Induction head found at layer 26 of 40-layer model does more complex **pattern matching:**

(month) (animal): 0; (month) (fruit): 1; (color) (animal): 2; (color) (fruit): 3



Why do the same heads that inductively copy random text also exhibit these other behaviors?

Copying:	`[A] [B] ... [A] → [B]`
Spiritually similar:	`[A*] [B*] ... [A] → [B]`

Why do the same heads that inductively copy random text also exhibit these other behaviors?

Copying:	**[A] [B] … [A] → [B]**
Spiritually similar:	**[A*] [B*] … [A] → [B]**

- the first behavior is a special case of the second

Why do the same heads that inductively copy random text also exhibit these other behaviors?

Copying:	`[A] [B] … [A] → [B]`
Spiritually similar:	`[A*] [B*] … [A] → [B]`

- the first behavior is a special case of the second
- induction heads copy literally when **isolated** in the residual stream but perform abstract tasks when processing earlier layers' outputs.

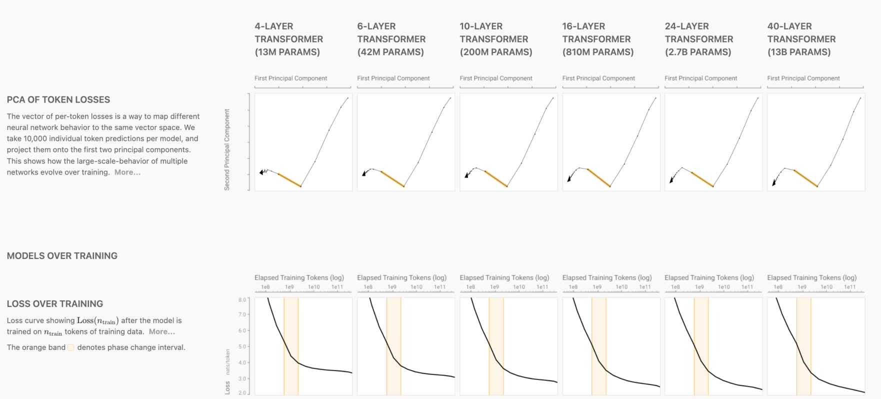
Argument 6: Continuity from small to large models

Extrapolation from small models suggests induction heads are responsible for the majority of in-context learning in large models.

STRENGTH OF ARGUMENT FOR SUB-CLAIMS

	Small Attention-Only	Small with MLPs	Large Models
Contributes Some			Analogy
Contributes Majority			Analogy

Full-scale Transformers



Full-scale Transformers

LOSS OVER TRAINING, BROKEN DOWN BY CONTEXT INDEX

Heatmap of $\mathrm{Loss}(n_{\mathrm{train}},\ i_{\mathrm{ctx}})$, the average loss of i_{ctx} token in context after n_{train} elapsed tokens of training. More...

5.5 5.0 4.5 4.0

DERIVATIVE OF LOSS WRT CONTEXT INDEX

Heatmap of $d\mathrm{Loss}(n_{\mathrm{train}},\ i_{\mathrm{ctx}})\ /\ d\ln(i_{\mathrm{ctx}})$. The log vertical partial derivative of the above graph. More...

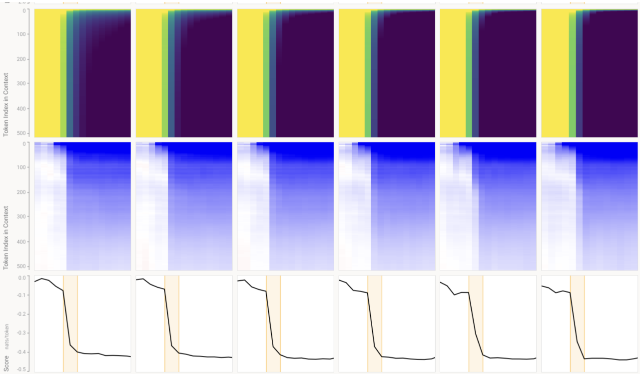
loss decreases with longer context

0.3 0.2 0.1 0.0 -0.1 -0.2 -0.3

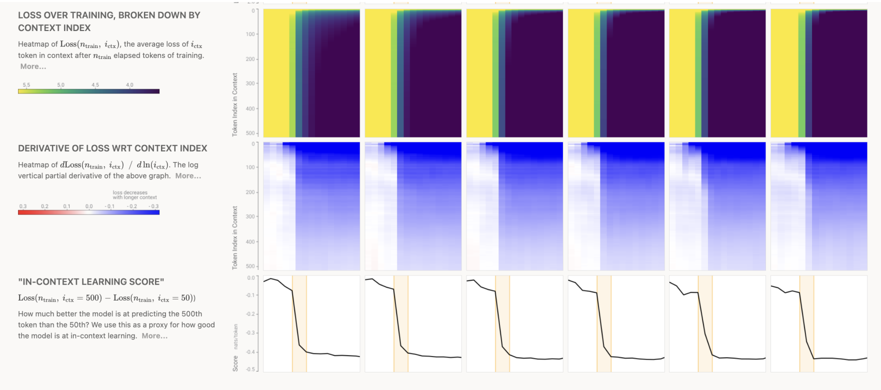
"IN-CONTEXT LEARNING SCORE"

$\mathrm{Loss}(n_{\mathrm{train}},\ i_{\mathrm{ctx}} = 500) - \mathrm{Loss}(n_{\mathrm{train}},\ i_{\mathrm{ctx}} = 50))$

How much better the model is at predicting the 500th token than the 50th? We use this as a proxy for how good the model is at in-context learning. More...

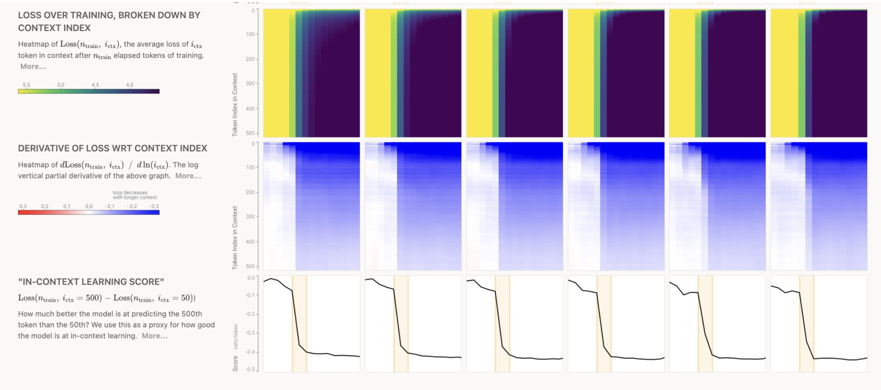


Full-scale Transformers



- **Same** phase change across models.

Full-scale Transformers



- **Same** phase change across models.
- Larger models may develop **other composition mechanisms** during phase change, enhancing ICL.

Concluding remarks

- Argument 1: Macroscopic co-occurence
- Argument 2: Macroscopic co-perturbation
- Argument 3: Direct ablation
- Argument 4: Specific examples of induction head generality
- Argument 5: Mechanistic plausibility of induction head generality (not included)
- Argument 6: Continuity from small to large models

Concluding remarks

- Argument 1: Macroscopic co-occurence
- Argument 2: Macroscopic co-perturbation
- Argument 3: Direct ablation
- Argument 4: Specific examples of induction head generality
- Argument 5: Mechanistic plausibility of induction head generality (not included)
- Argument 6: Continuity from small to large models

Strong mechanistic evidence that induction heads drive ICL in **small** models; Correlational evidence for **larger models with MLPs**

Thanks for your attention!