

In-context Learning and Induction Heads

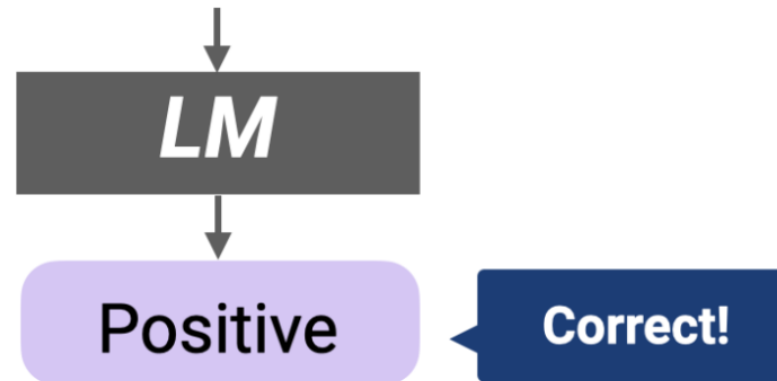
Olsson et al., Anthropic
2022



What is In-Context Learning?

What is In-Context Learning?

Circulation revenue has increased by 5% in Finland.	\n	Positive
Panostaja did not disclose the purchase price.	\n	Neutral
Paying off the national debt will be extremely painful.	\n	Negative
The company anticipated its operating profit to improve.	\n	_____



Zero shot learning

Classify the following review:

The product works great! //

Few shot learning

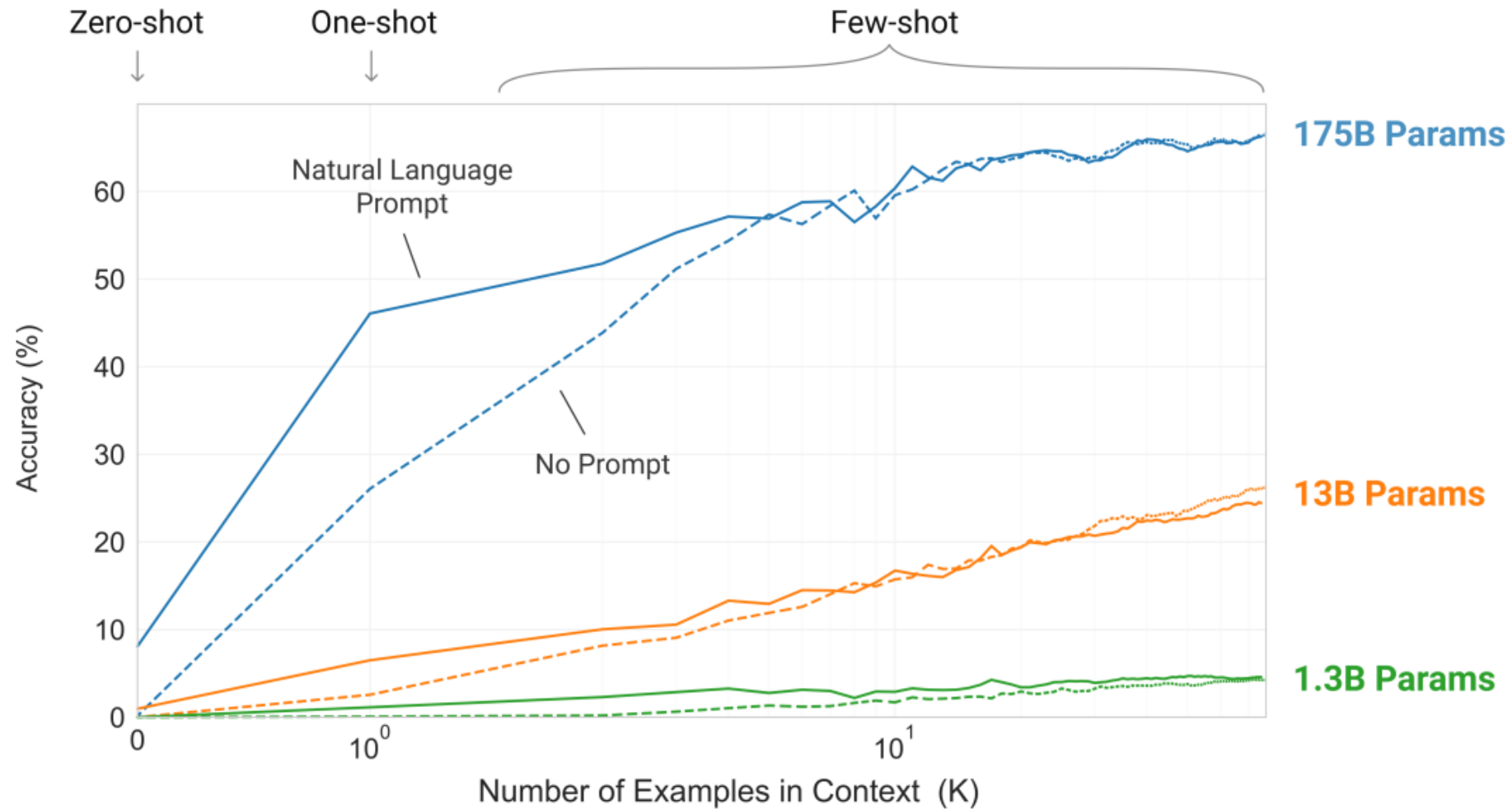
Classify the following review:

This product is the best! // Positive
The product doesn't work // Negative
The product is alright I guess // Neutral

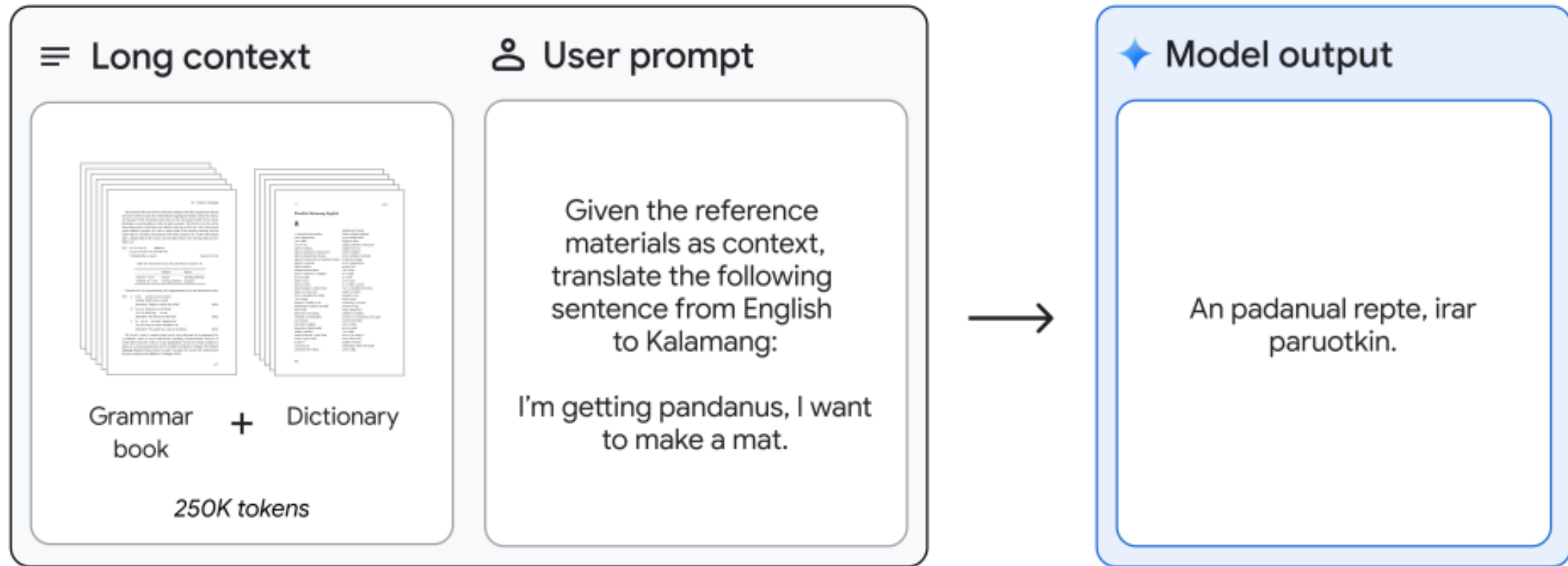
The product works great! //

In-Context Learning in Large Language Models

Performance of ICL

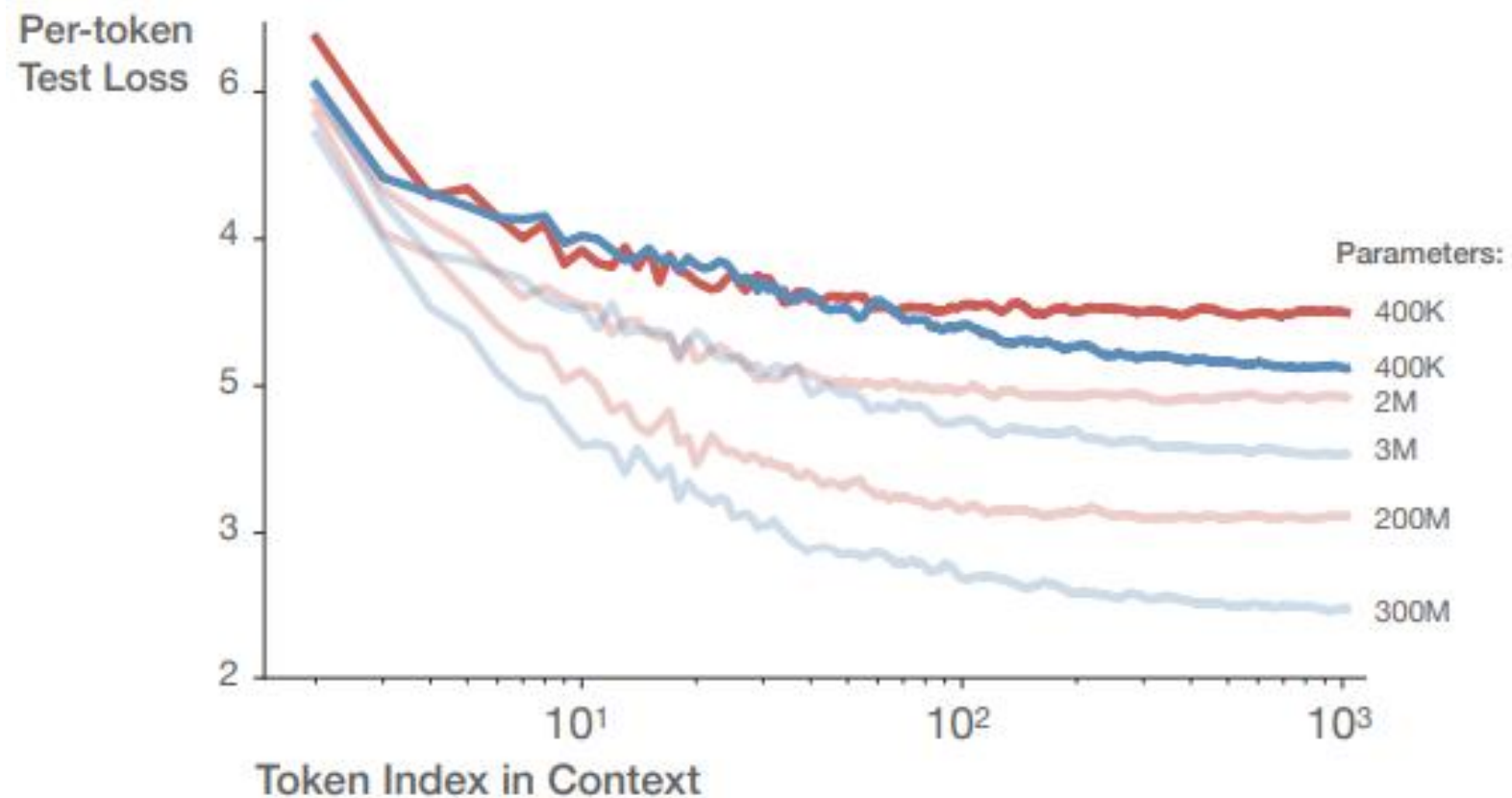


New tasks using ICL – Learning a new language



Comparison with LSTM

LSTM plateaus after <100 tokens
Transformers improves through the whole context



CoT Prompting

Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. 

Chain-of-Thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

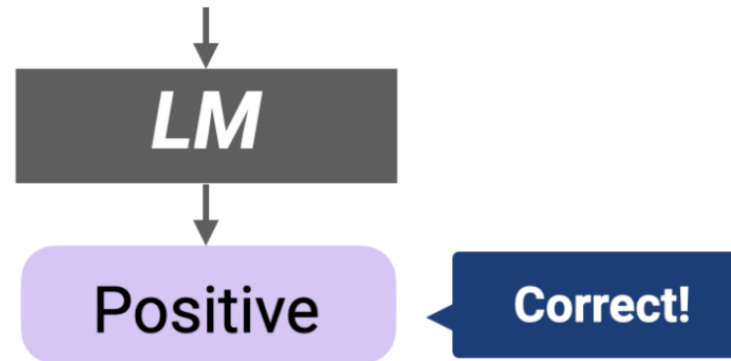
Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. 

Reasons behind ICL – Bayesian Framework

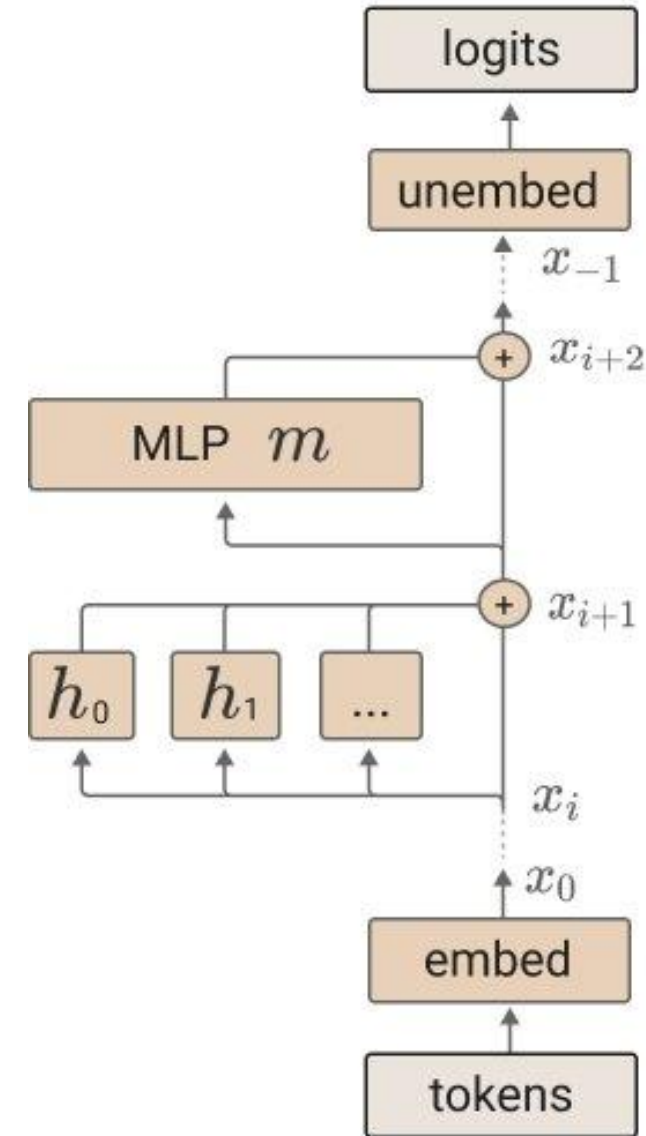
Circulation revenue has increased by 5% in Finland.	\n	Neutral
Panostaja did not disclose the purchase price.	\n	Negative
Paying off the national debt will be extremely painful.	\n	Positive
The company anticipated its operating profit to improve.	\n	_____



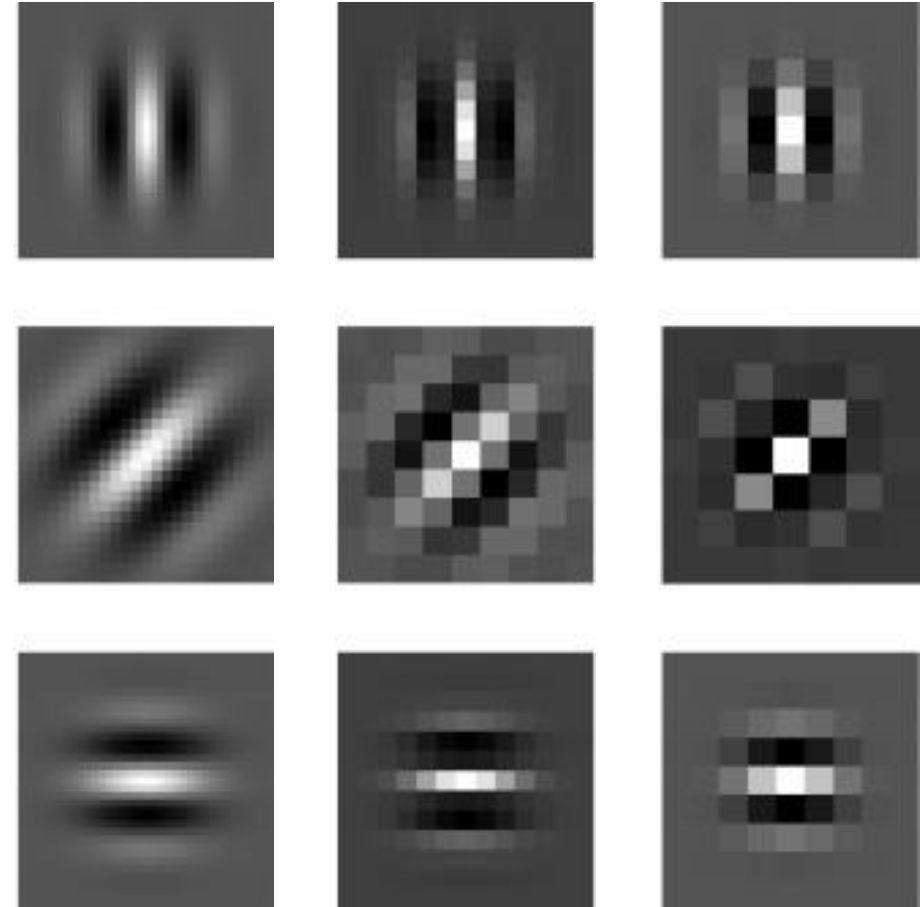
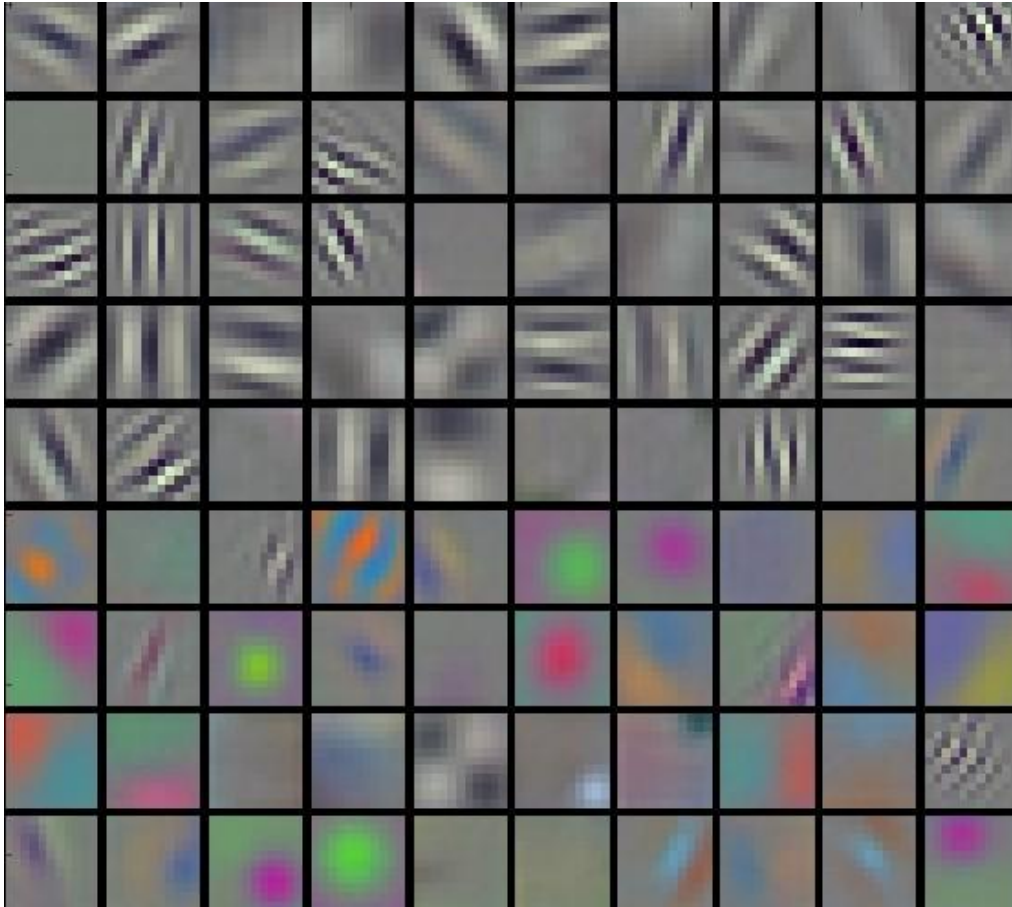
Mechanistic Interpretability

What is Mechanistic Interpretability?

It is the study of how neural networks compute their outputs, by reverse-engineering their internal structures into components that are understandable to humans.



Why toy models? – Gabor filters



Claim of the paper

Induction heads constitute the mechanism underlying the majority of all in-context learning in large transformer models.

What are induction heads?

[A][B] ... [A] → [B]

What are induction heads?

$[D][urs][ley] \dots [D][urs] \longrightarrow [ley]$

Visualizing induction heads

Beginning of Harry Potter

<EOT> Mr. and Mrs. Dursley, of number four, Privet Drive, were proud to say that they were perfectly normal, thank you very much. They were the last people you'd expect to be involved in anything strange or mysterious, because they just didn't hold with such nonsense. Mr. Dursley was the director of a firm called Grunnings, which made drills. He was a big, beefy man with hardly any neck, although he did have a very large mustache. Mrs. Dursley was thin and blonde and had nearly twice the usual amount of neck, which came in very useful as she spent so much of her time craning over garden fences, spying on the neighbors. The Dursleys had a small son called Dudley and in their opinion there was no finer boy anywhere. The Dursleys had everything they wanted, but they also had a secret, and their greatest fear was that somebody would discover it.

Mr

Visualizing induction heads

Beginning of Harry Potter

<EOT> Mr. and Mrs. Dursley, of number four, Privet Drive, were proud to say that they were perfectly normal, thank you very much. They were the last people you'd expect to be involved in anything strange or mysterious, because they just didn't hold with such nonsense. Mr. Dursley was the director of a firm called Grunnings, which made drills. He was a big, beefy man with hardly any neck, although he did have a very large mustache. Mrs. Dursley was thin and blonde and had nearly twice the usual amount of neck, which came in very useful as she spent so much of her time craning over garden fences, spying on the neighbors. The Dursleys had a small son called Dudley and in their opinion there was no finer boy anywhere. The Dursleys had everything they wanted, but they also had a secret, and their greatest fear was that somebody would discover it.

Mr D

Visualizing induction heads

Beginning of Harry Potter

<EOT> Mr. and Mrs. Dursley, of number four, Privet Drive, were proud to say that they were perfectly normal, thank you very much. They were the last people you'd expect to be involved in anything strange or mysterious, because they just didn't hold with such nonsense. Mr. Dursley was the director of a firm called Grunnings, which made drills. He was a big, beefy man with hardly any neck, although he did have a very large mustache. Mrs. Dursley was thin and blonde and had nearly twice the usual amount of neck, which came in very useful as she spent so much of her time craning over garden fences, spying on the neighbors. The Dursleys had a small son called Dudley and in their opinion there was no finer boy anywhere. The Dursleys had everything they wanted, but they also had a secret, and their greatest fear was that somebody would discover it.

Mr Durs

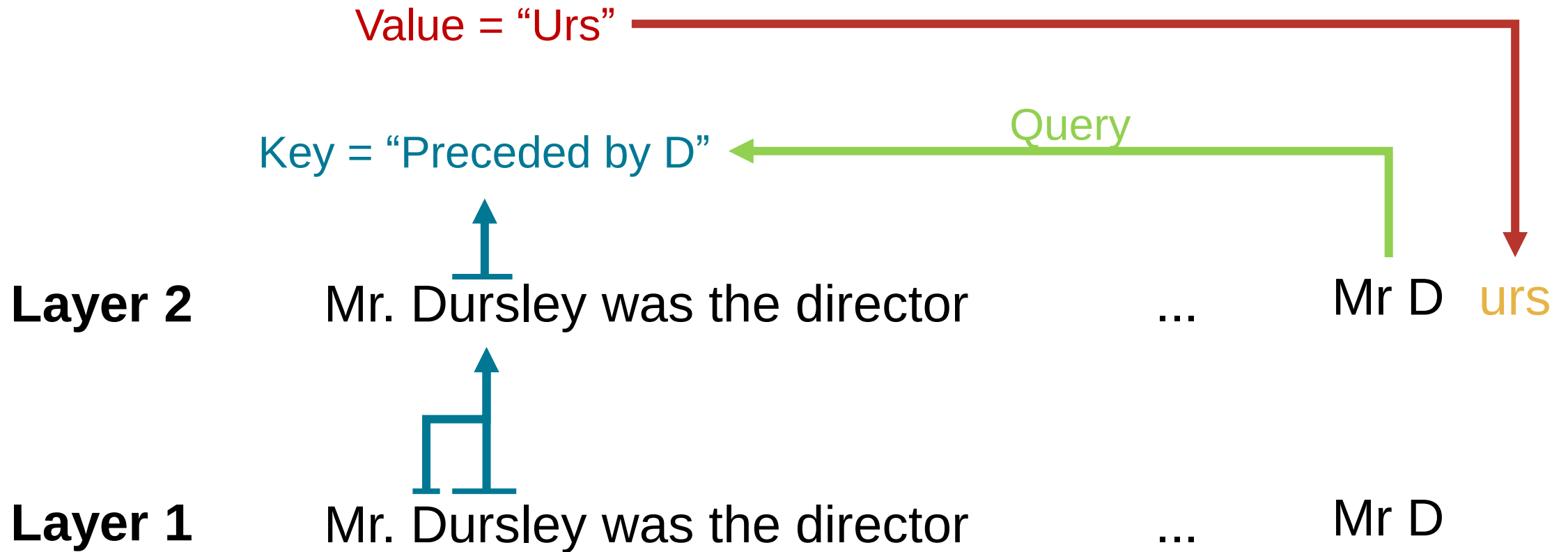
Visualizing induction heads

Beginning of Harry Potter

<EOT> Mr. and Mrs. Dursley, of number four, Privet Drive, were proud to say that they were perfectly normal, thank you very much. They were the last people you'd expect to be involved in anything strange or mysterious, because they just didn't hold with such nonsense. Mr. Dursley was the director of a firm called Grunnings, which made drills. He was a big, beefy man with hardly any neck, although he did have a very large mustache. Mrs. Dursley was thin and blonde and had nearly twice the usual amount of neck, which came in very useful as she spent so much of her time craning over garden fences, spying on the neighbors. The Dursleys had a small son called Dudley and in their opinion there was no finer boy anywhere. The Dursleys had everything they wanted, but they also had a secret, and their greatest fear was that somebody would discover it.

Mr Dursley

How are they implemented?



Abstract induction heads

$[A^*][B^*] \dots [A] \rightarrow [B]$

where $[A] \approx [A^*]$ and $[B] \approx [B^*]$

Generalization of induction heads : Translation

<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.

Generalization of induction heads : Pattern Matching

Category 1	Category 2	Output
Month	Animal	0
Month	Fruit	1
Color	Animal	2
Color	Fruit	3

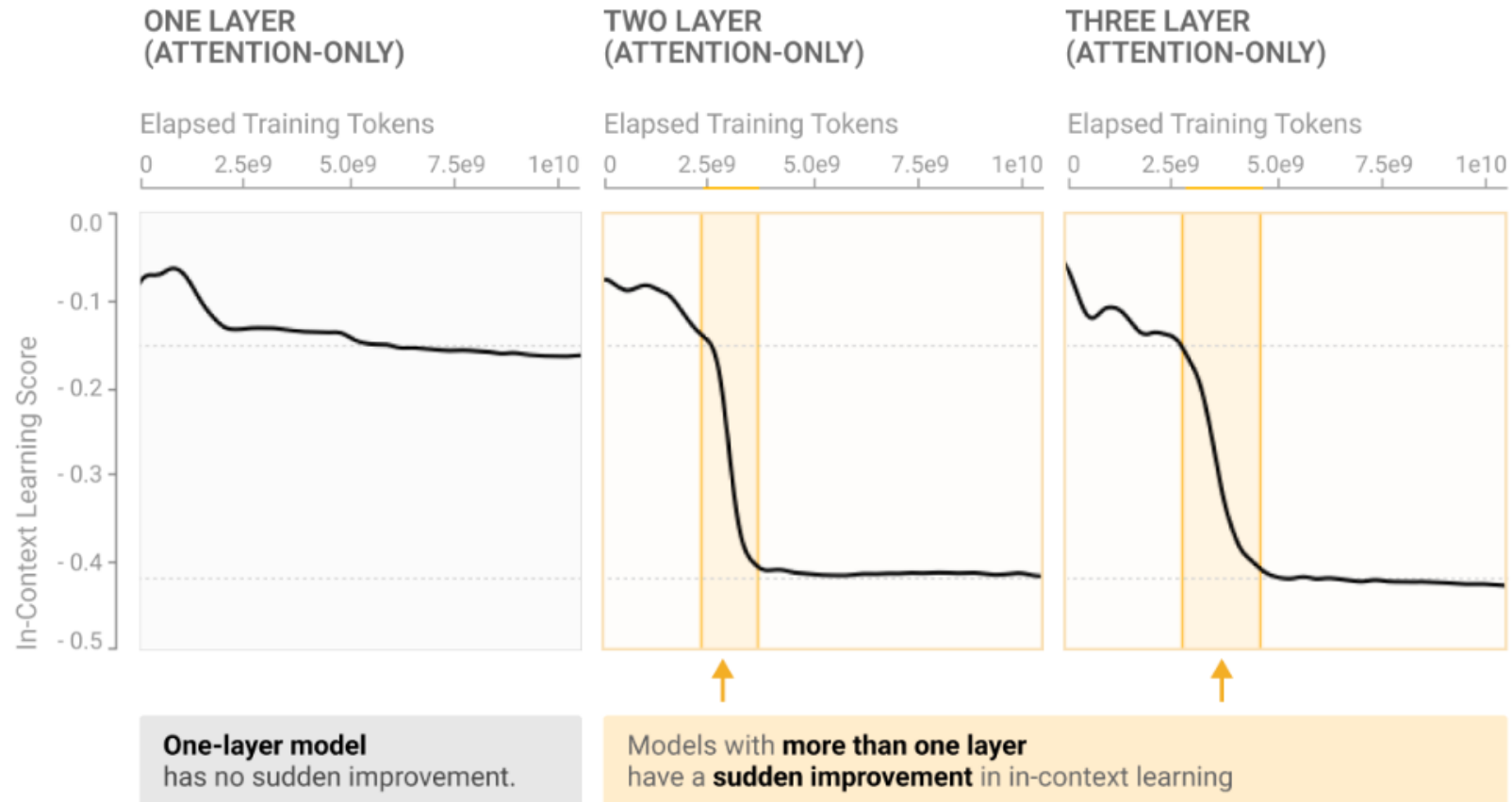
PROMPT

July Lizard: 0
Red Cherry: 3
Blue Lion: 2
September Pineapple: 1
Gray Cat: 2
Green Banana: 3
Red Snake: 2
March Cherry: 1
...
June Banana: 1

Phase change

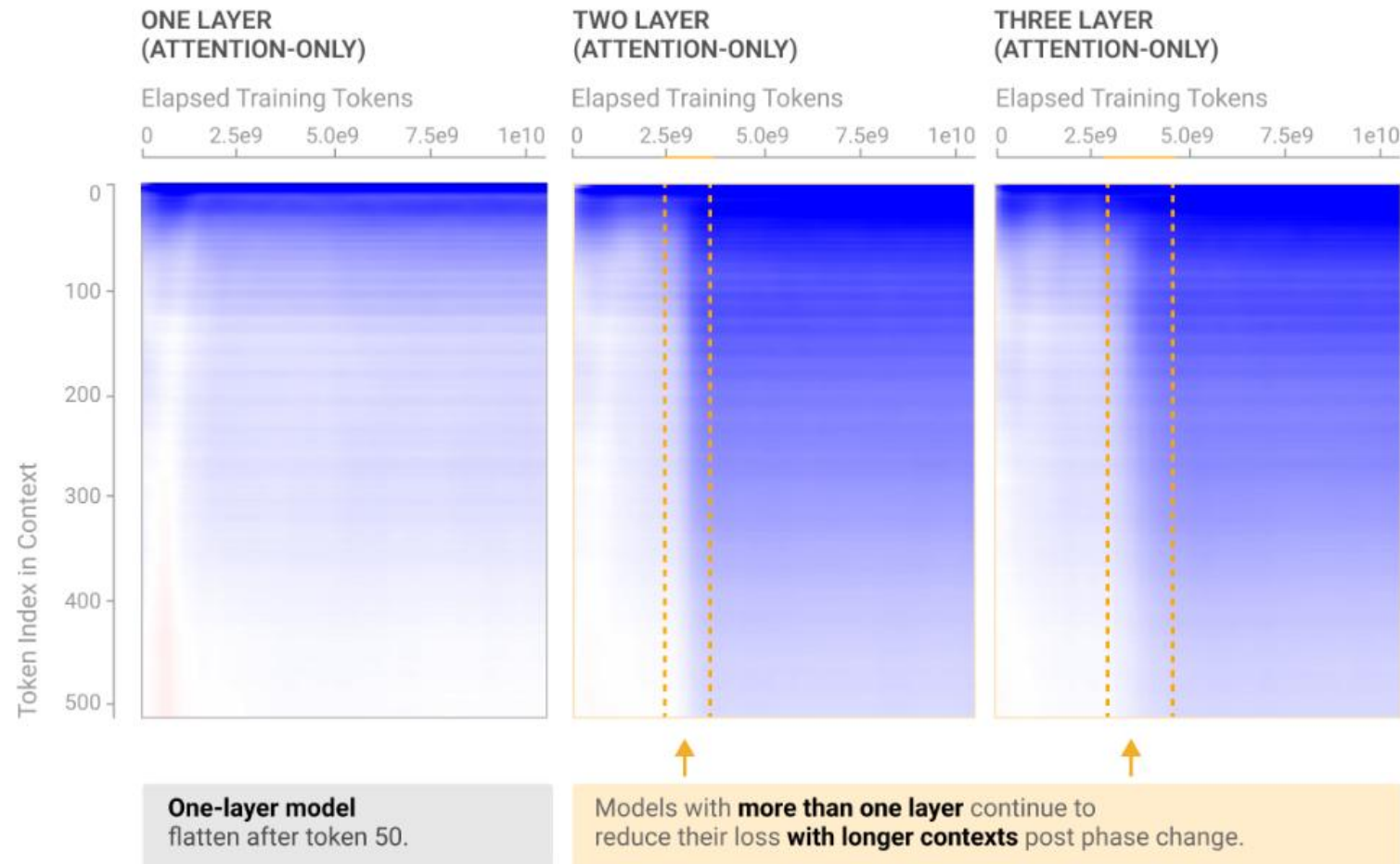
“Transformer language models undergo a “phase change” during training, during which induction heads form and simultaneously in-context learning improves dramatically.”

Behavior of the ICL score over training



In-context learning score = average loss(500th token – 50th token)

Behavior of the derivative of the loss over training



$$\frac{d \text{ loss}}{d \log(\text{context index})}$$

What changes in the phase transition?

Mr and Mrs Dursley, of number four, Privet Drive, were proud to say that they were perfectly normal, thank you very much. They were the last people you'd expect to be involved in anything strange or mysterious, because they just didn't hold with such nonsense. Mr Dursley was the director of a firm called Grunnings, which made drills. He was a big, beefy man with hardly any neck, although he did have a very large moustache. Mrs Dursley was thin and blonde and had nearly twice the usual amount of neck, which came in very useful as she spent so much of her time craning over garden fences, spying on the neighbours. The Dursleys had a small son called Dudley and in their opinion there was no finer boy anywhere. The Dursleys had everything they wanted, but they also had a secret, and their greatest fear was that somebody would discover it. They didn't think they could bear it if anyone found out about the Potters. Mrs Potter was Mrs Dursley's sister, but they hadn't met for several years; in fact, Mrs Dursley pretended she didn't have a sister, because her sister and her good-for-nothing husband were as unDursleyish as it was possible to be. The Dursleys shuddered to think what the neighbours would say if the Potters arrived in the street. The Dursleys knew that the Potters had a small son, too, but they had never even seen him. This boy was another good reason for keeping the Potters away; they didn't want Dudley mixing with a child like that.

Early Prediction

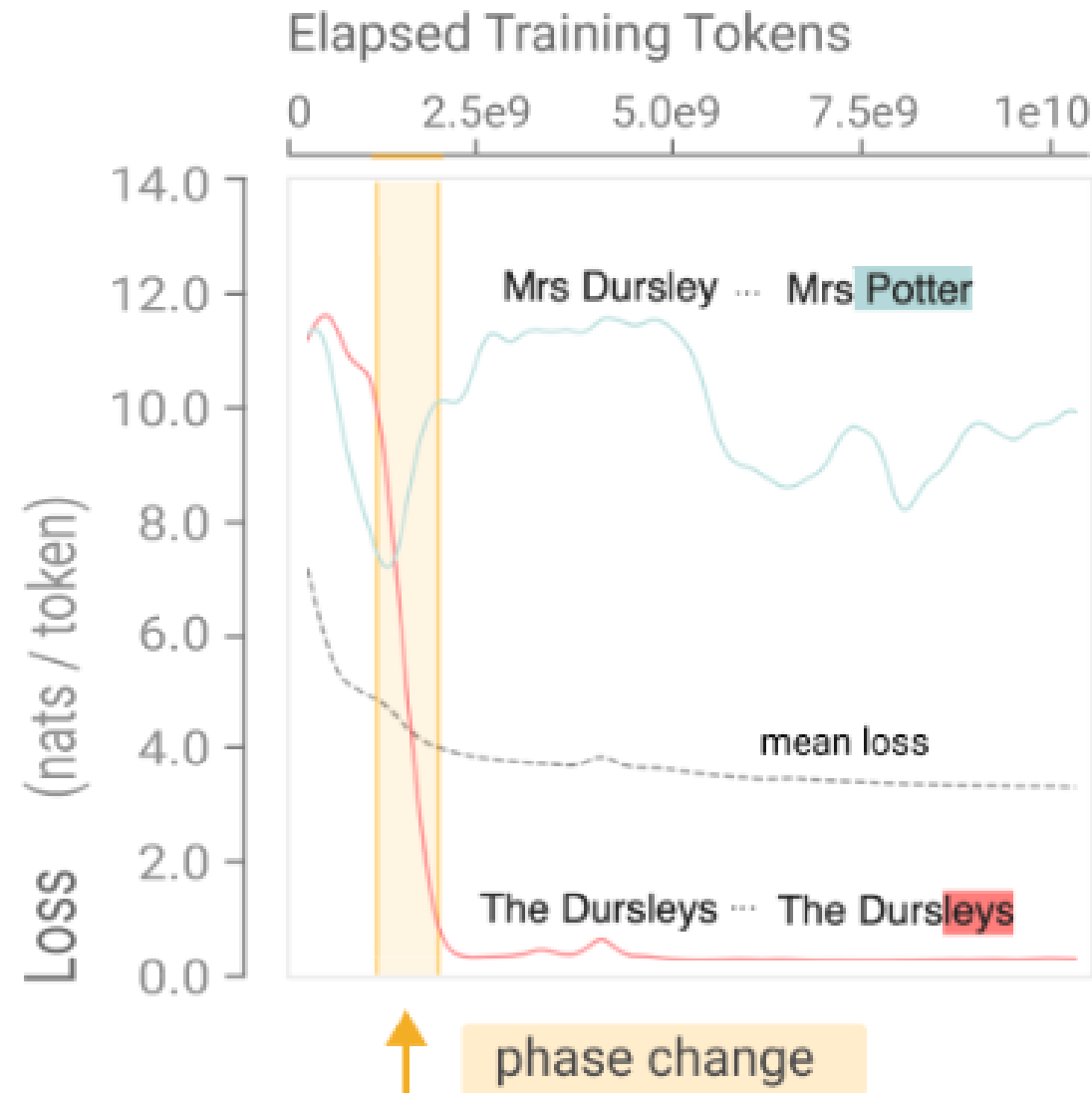
Mrs Dursley

Late Prediction

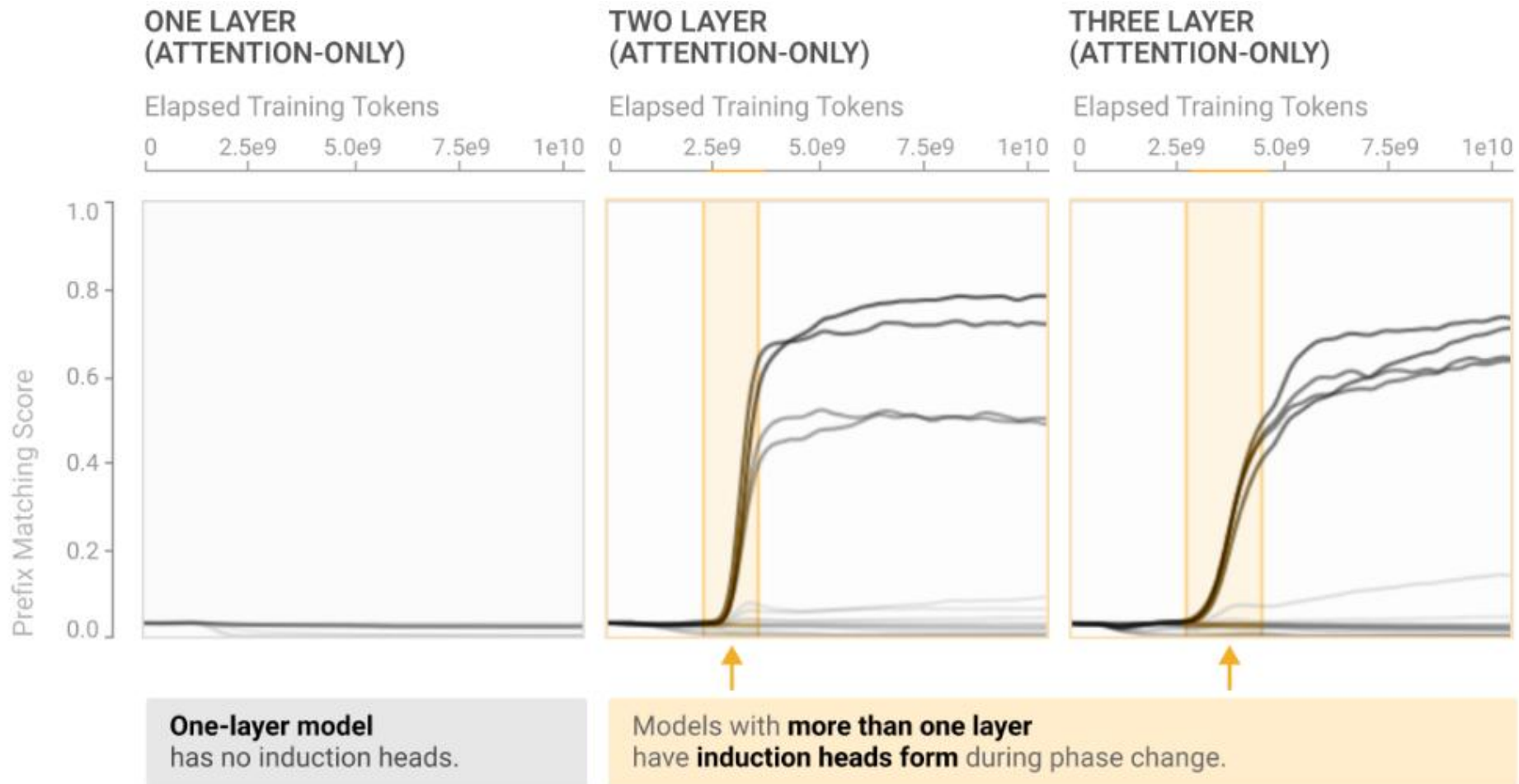
Mrs Dursley

Mrs Potter

What changes in the phase transition?

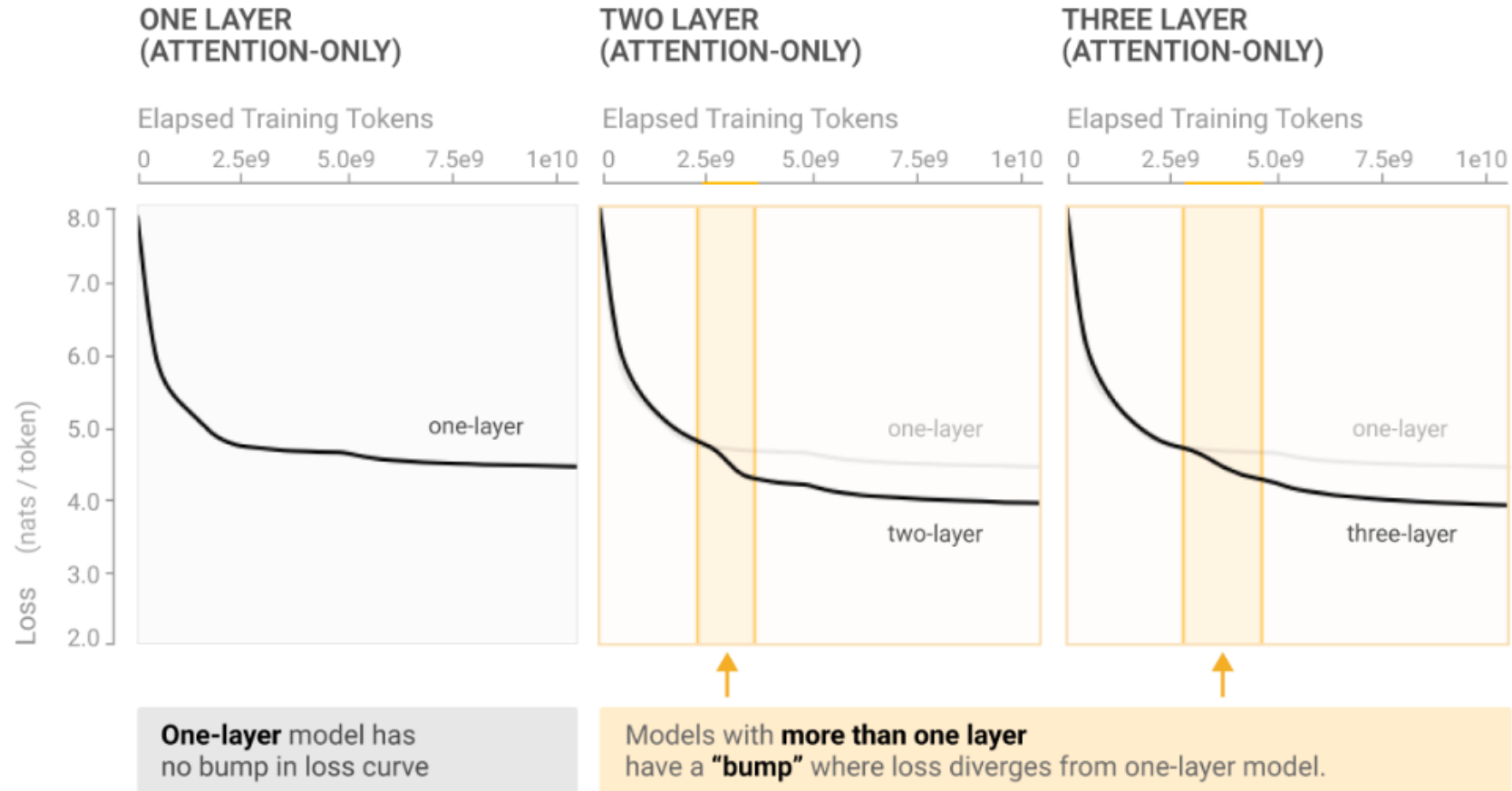


Behavior of induction heads over training



Prefix matching score

Behavior of the loss over training



Loss of one random token per example, averaged over all examples

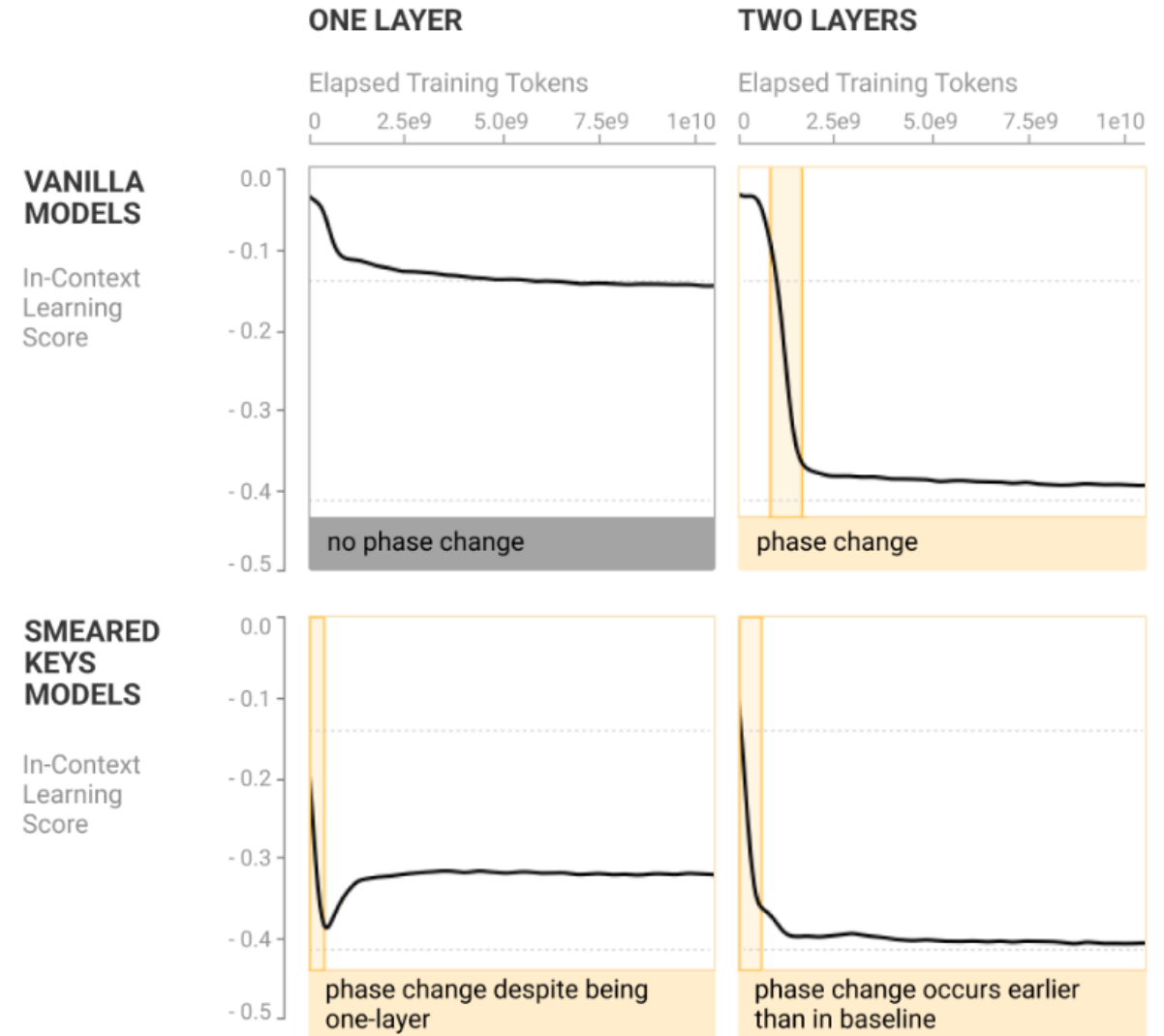
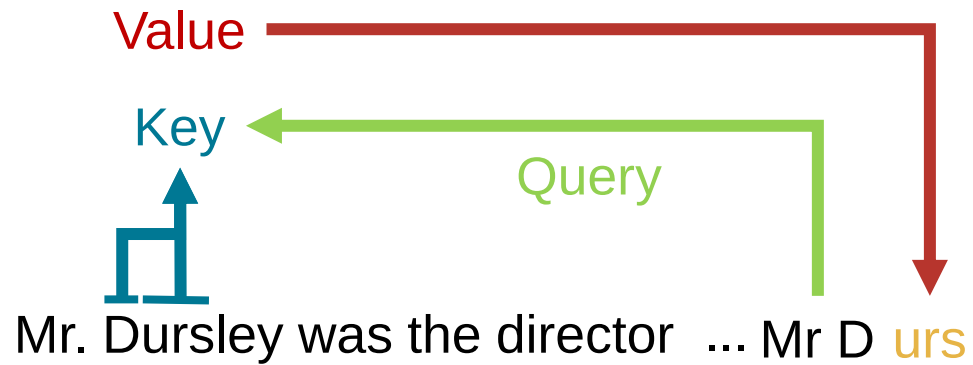
The evidence so far...

In a small window of training we have that:

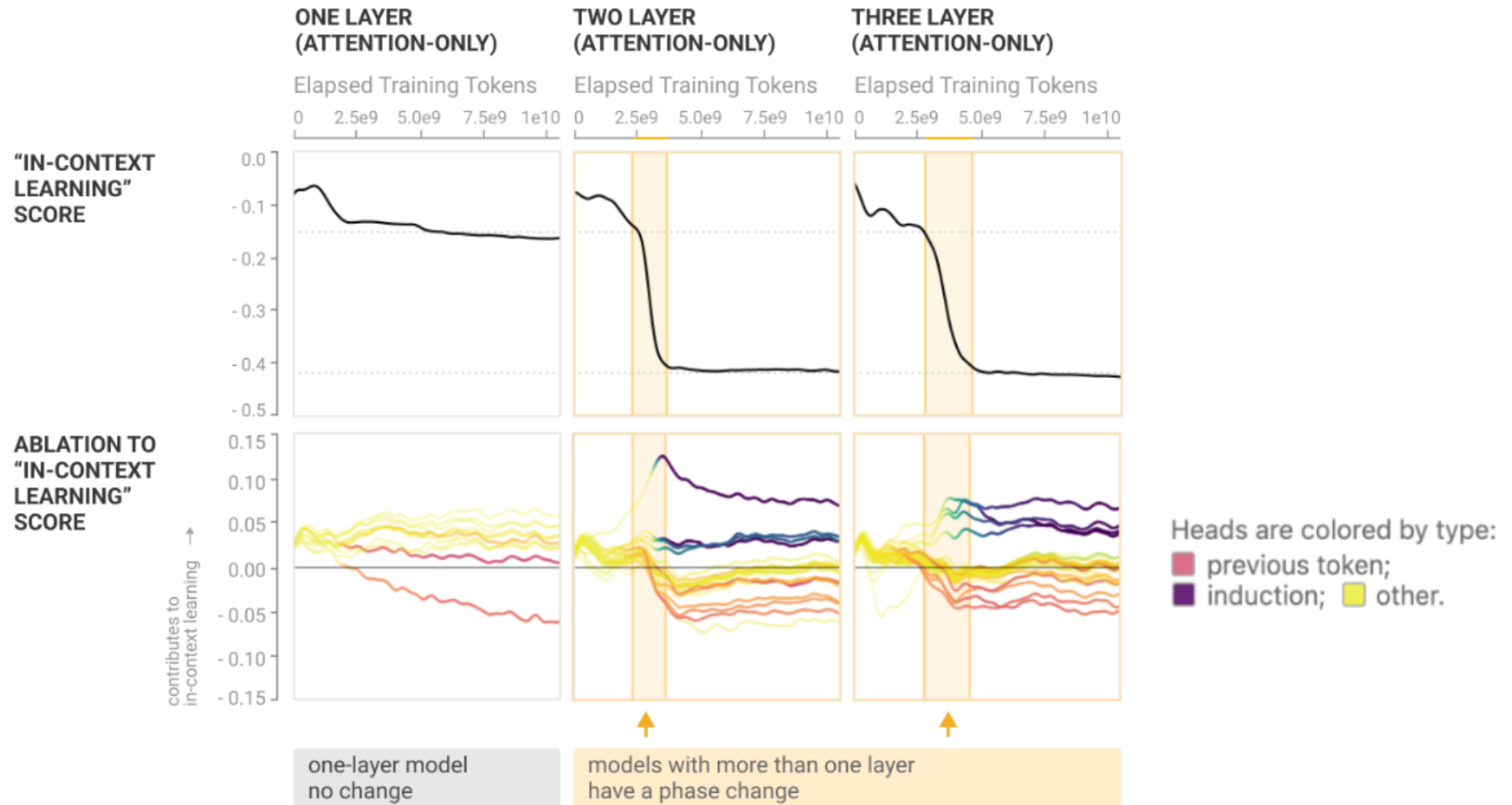
1. In-context learning abilities sharply improves
2. Induction heads form abruptly
3. The loss undergoes a small but visible bump

Smeared architecture

The structural modification of attention:



Ablation study



Limitations of the paper

The paper provides some evidence that induction heads are related to in-context learning in large language models, and shows that this effect is visible also in the loss curve.

However:

1. Evidence is causal and very strong only for small and simple models
2. Evidence is correlational and indirect for larger models
3. For large models, they only had 15 checkpoints of the model during training, making the co-occurrence argument weaker.

Thank you!