

FINE-TUNING ENHANCES EXISTING MECHANISMS

A Case Study on Entity Tracking



Nikhil Prakash



Tamar Rott Shaham



Tal Haklay



Yonatan Belinkov

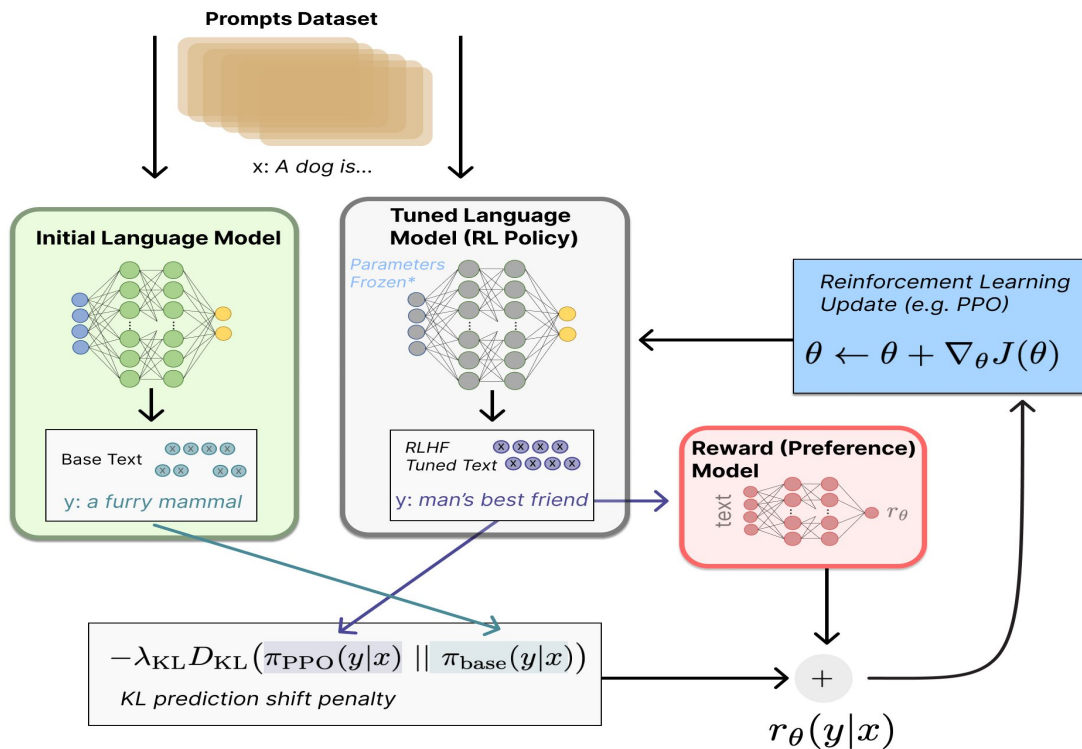


David Bau



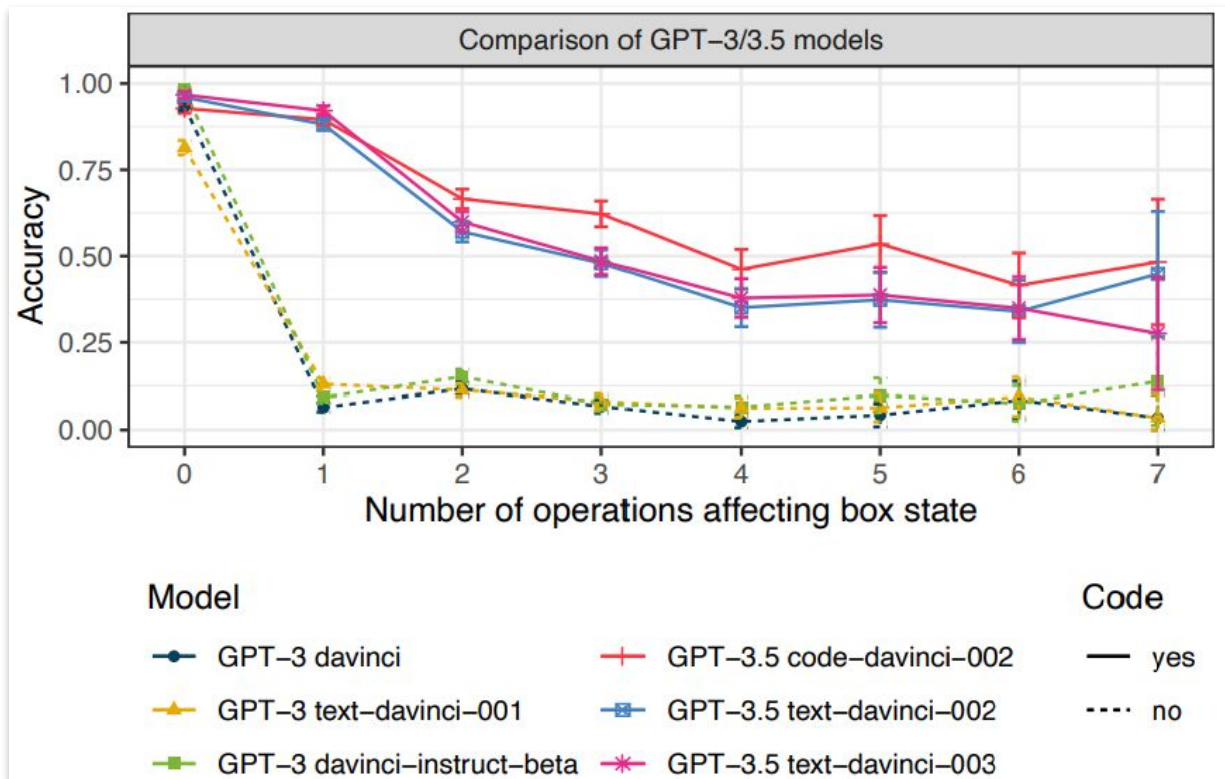
Background: Fine-tuning improves capabilities beyond training task

RLHF on English leads to instruction-following in French (Ouyang et al. 2022)



Background: Fine-tuning improves capabilities beyond training task

Code fine-tuning improves Entity Tracking capabilities (Kim et al. 2023)



- Fine-tuning on generic domains such as code, mathematics, and instructions has been shown to enhance language models performance on multiple tasks.

Goat-7B

3978640188 + 42886272 =
3978640188

4523646
4523646

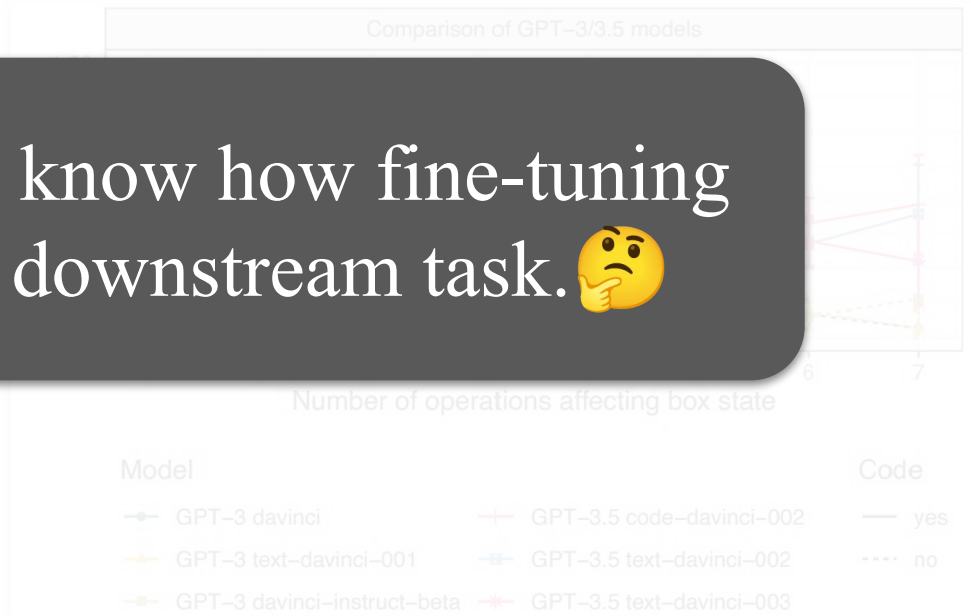
Calculate

397 × 4
= 4429
= 4429
= 1328
= 17273
= 1758313

What is 8914/64?

8914 - 64 × 100 = 8914 - 6400 = 2514
2514 - 64 × 30 = 2514 - 1920 = 594
594 - 64 × 9 = 594 - 576 = 18
Therefore, 8914 ÷ 64 = 139 R 18

We still don't know how fine-tuning improves a downstream task. 🤔



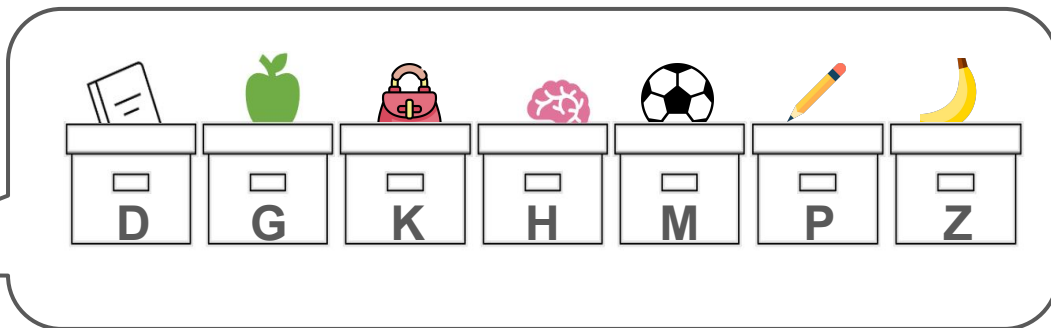
Liu et al. "Goat: Fine-tuned LLaMA Outperforms GPT-4 on Arithmetic Tasks" (2023).

Kim et al. "Entity Tracking in Language Models" (2023).

What is Entity Tracking Task?

The book is in Box D, the apple is in Box G, the brain is in Box H, ...

Box G contains the _____

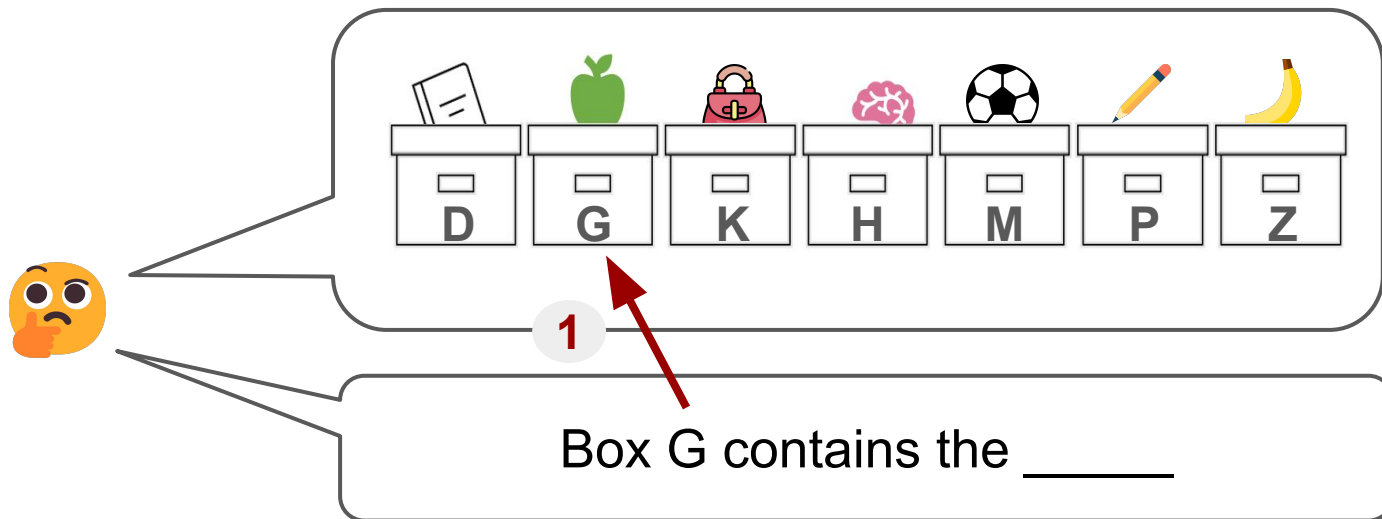


Box G contains the _____

What is Entity Tracking Task?

The book is in Box D, the apple is in Box G, the brain is in Box H, ...

Box G contains the _____

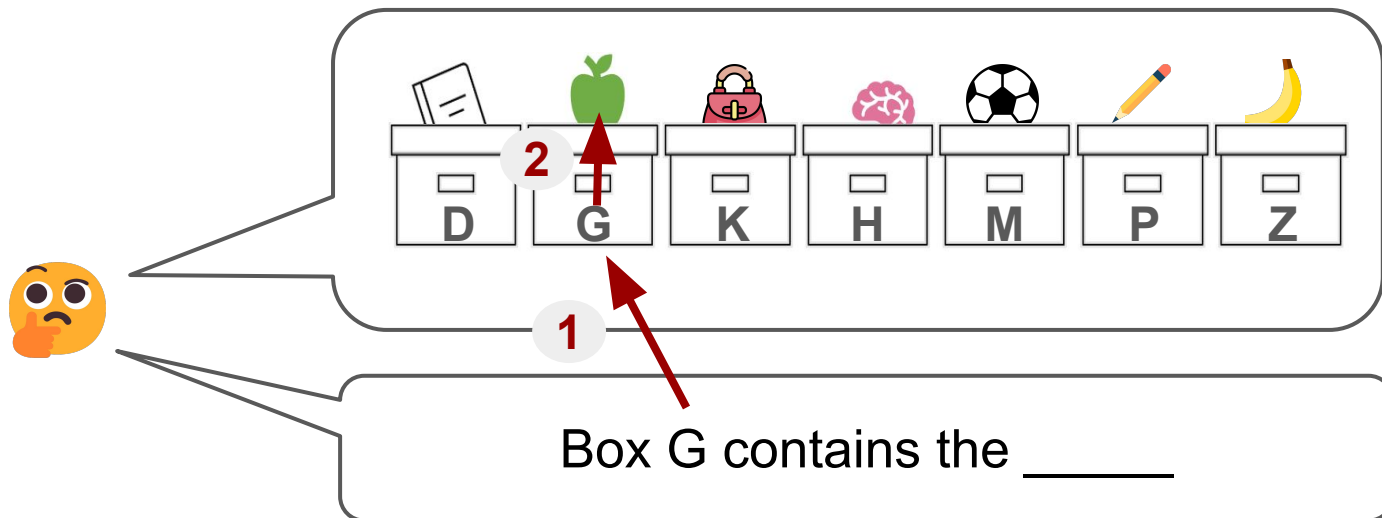


The diagram illustrates an entity tracking task. It features a row of seven boxes labeled D, G, K, H, M, P, and Z. Each box contains a different object: a book in D, an apple in G, a handbag in K, a brain in H, a soccer ball in M, a pencil in P, and a banana in Z. A thinking face emoji on the left points to the boxes. A red arrow points to box G, which contains an apple, with a '1' in a circle next to it. Below the boxes is a text box asking 'Box G contains the _____'.

What is Entity Tracking Task?

The book is in Box D, the apple is in Box G, the brain is in Box H, ...

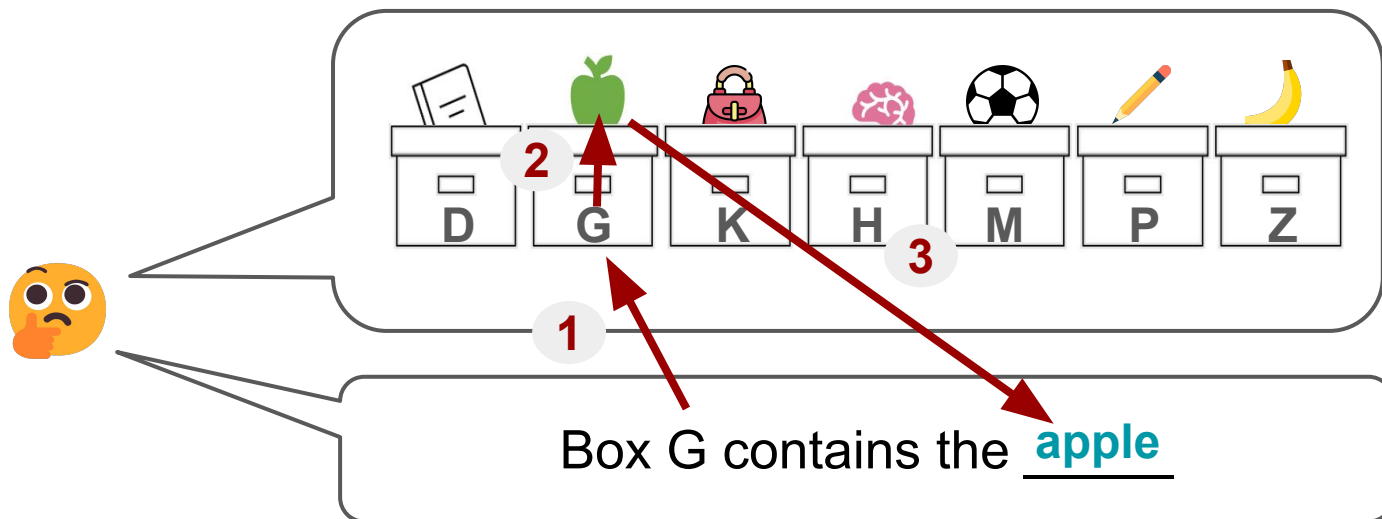
Box G contains the _____



What is Entity Tracking Task?

The book is in Box D, the apple is in Box G, the brain is in Box H, ...

Box G contains the _____



Arithmetic Fine-tuning improves Entity Tracking

Model	Fine-tuned?	Full Model Performance
Llama-7B (Touvron et al. 2023)	-	0.66
Vicuna-7B (Chiang et al. 2023)	User conversations	0.67
Goat-7B (Liu & Low, 2023)	Arithmetic tasks (LoRA)	0.82
FLoat-7B	Arithmetic tasks (w/o LoRA)	0.82

Arithmetic Fine-tuning improves Entity Tracking

Model	Fine-tuned?	Full Model Performance
Llama-7B (Touvron et al. 2023)	-	0.66
Vicuna-7B (Chiang et al. 2023)	User conversations	0.67
Goat-7B (Liu & Low, 2023)	Arithmetic tasks (LoRA)	0.82
FLoat-7B	Arithmetic tasks (w/o LoRA)	0.82

Base LM is not great at entity tracking.



Arithmetic Fine-tuning improves Entity Tracking

Model	Fine-tuned?	Full Model Performance
Llama-7B (Touvron et al. 2023)	-	0.66
Vicuna-7B (Chiang et al. 2023)	User conversations	0.67
Goat-7B (Liu & Low, 2023)	Arithmetic tasks (LoRA)	0.82
FLoat-7B	Arithmetic tasks (w/o LoRA)	0.82

RLHF doesn't improve it much.



Arithmetic Fine-tuning improves Entity Tracking

Model	Fine-tuned?	Full Model Performance
Llama-7B (Touvron et al. 2023)	-	0.66
Vicuna-7B (Chiang et al. 2023)	User conversations	0.67
Goat-7B (Liu & Low, 2023)	Arithmetic tasks (LoRA)	0.82
FLoat-7B	Arithmetic tasks (w/o LoRA)	0.82

Arithmetic fine-tuning improves it a lot.

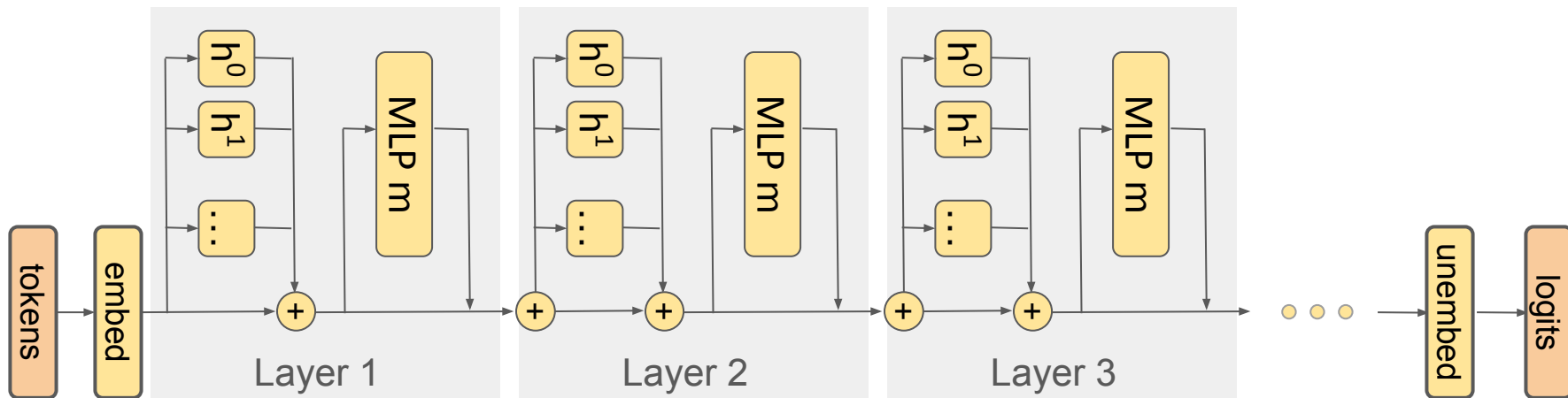


- We study the ability of language model to perform **in-context entity tracking**, i.e. infer properties associated with an entity previously defined in the input context.

Mechanistically explain why fine-tuned models perform entity tracking better than base model. 🤔

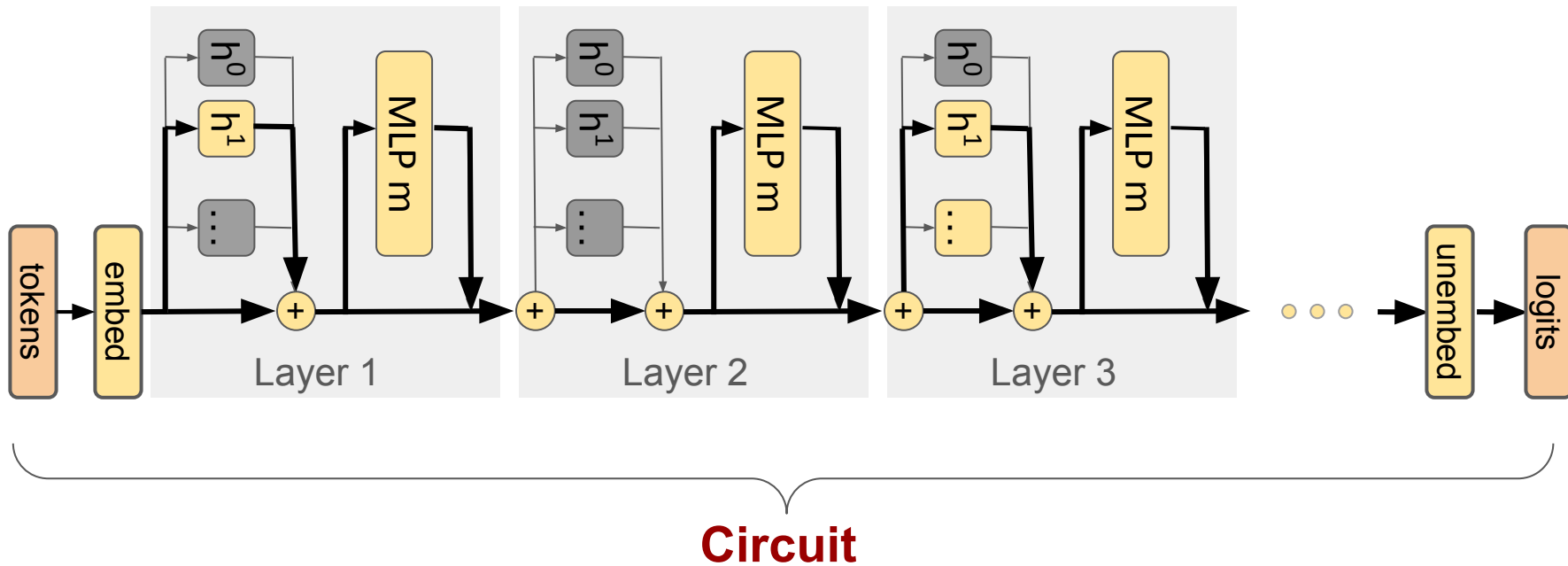
		Performance
Llama-7B		
Vicuna-7B		
Goat-7B		
FLan-7B	Arithmetic tasks (w/o LoRA)	0.52

What is a Circuit?



Transformer Model

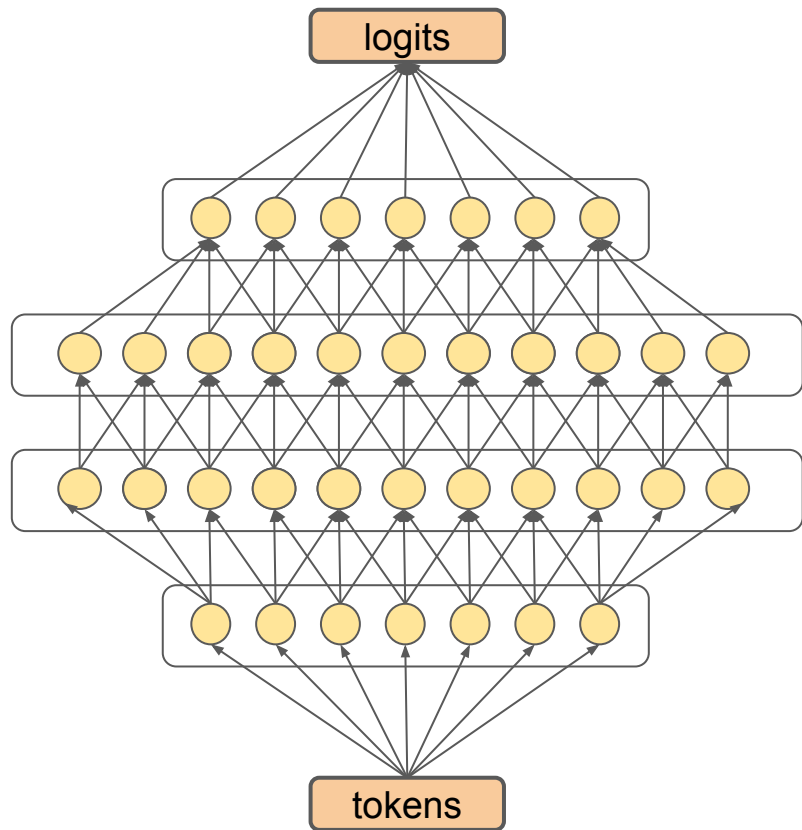
What is a Circuit?



Hypotheses: Is the same circuit present after fine-tuning?

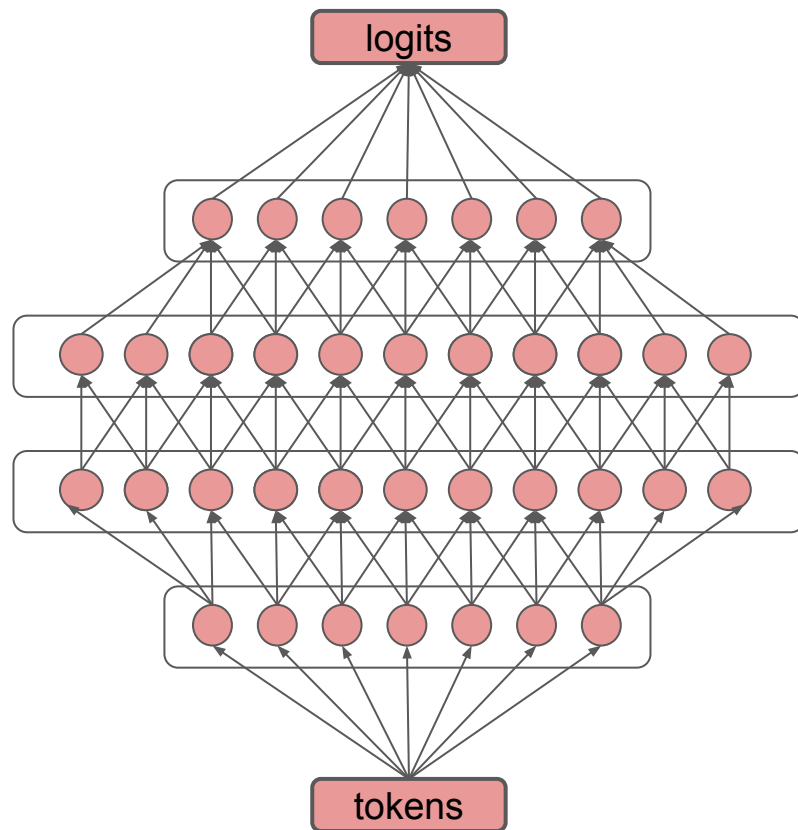
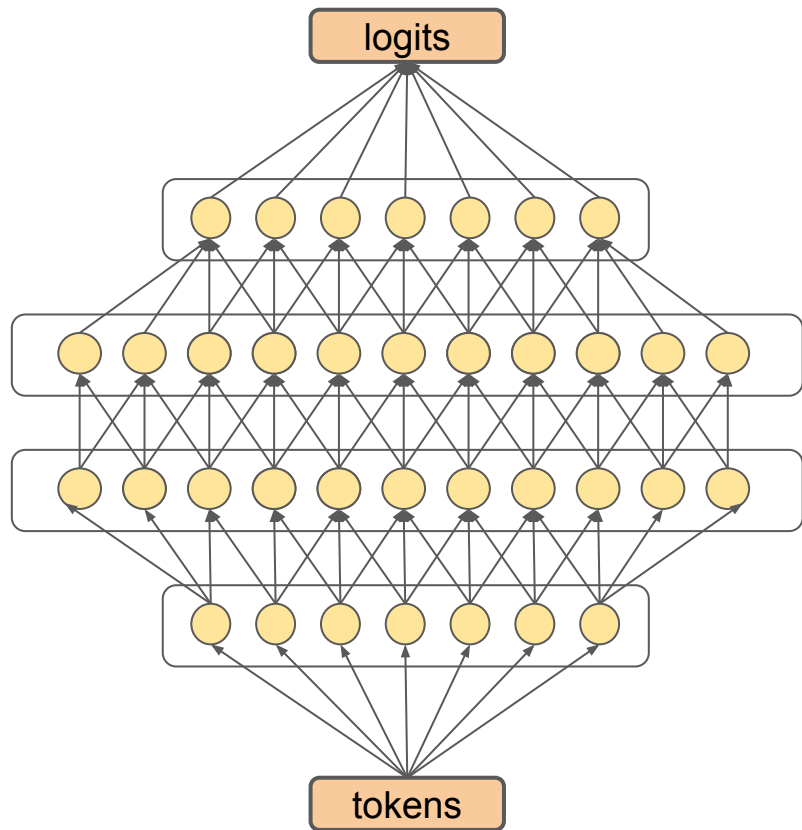
- Fine-tuned models contain a **different circuit** for performing entity tracking.
- Fine-tuned models contains the **same circuit** as the base model.

Path Patching: Circuit Discovery Algorithm (Intuition)

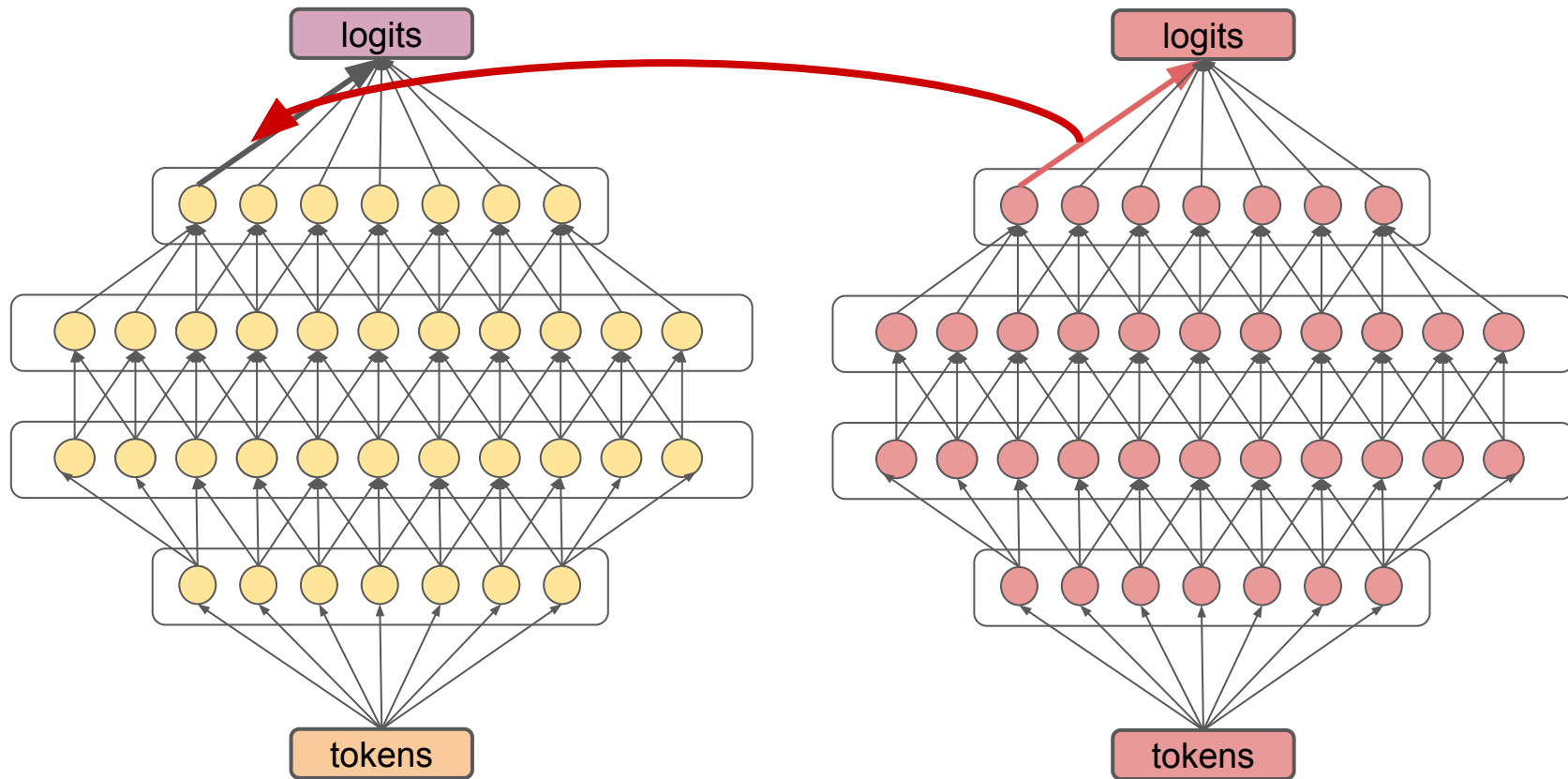


What are the important edges in a computational graph?

Path Patching: Circuit Discovery Algorithm (Intuition)

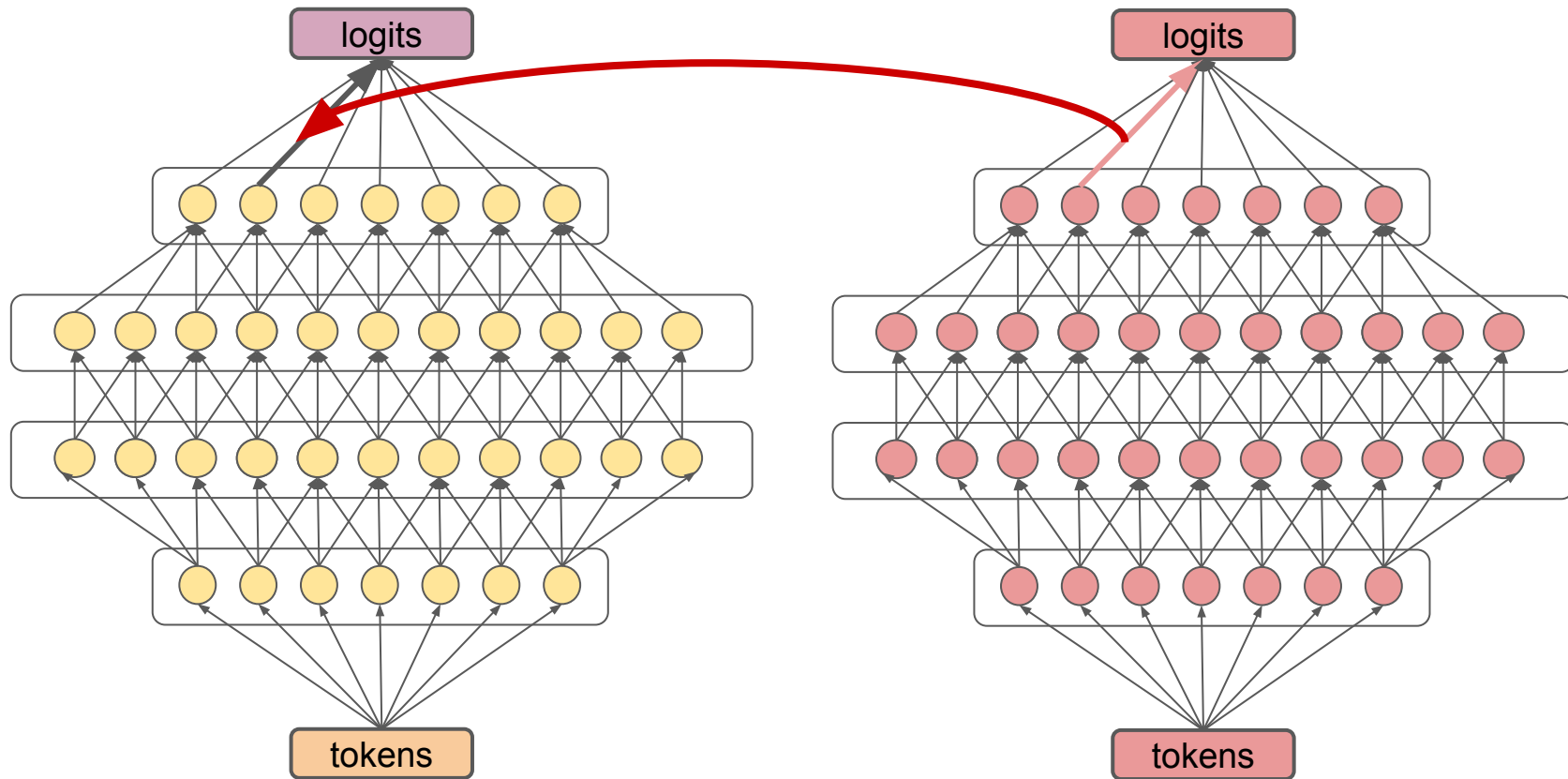


Path Patching: Circuit Discovery Algorithm (Intuition)



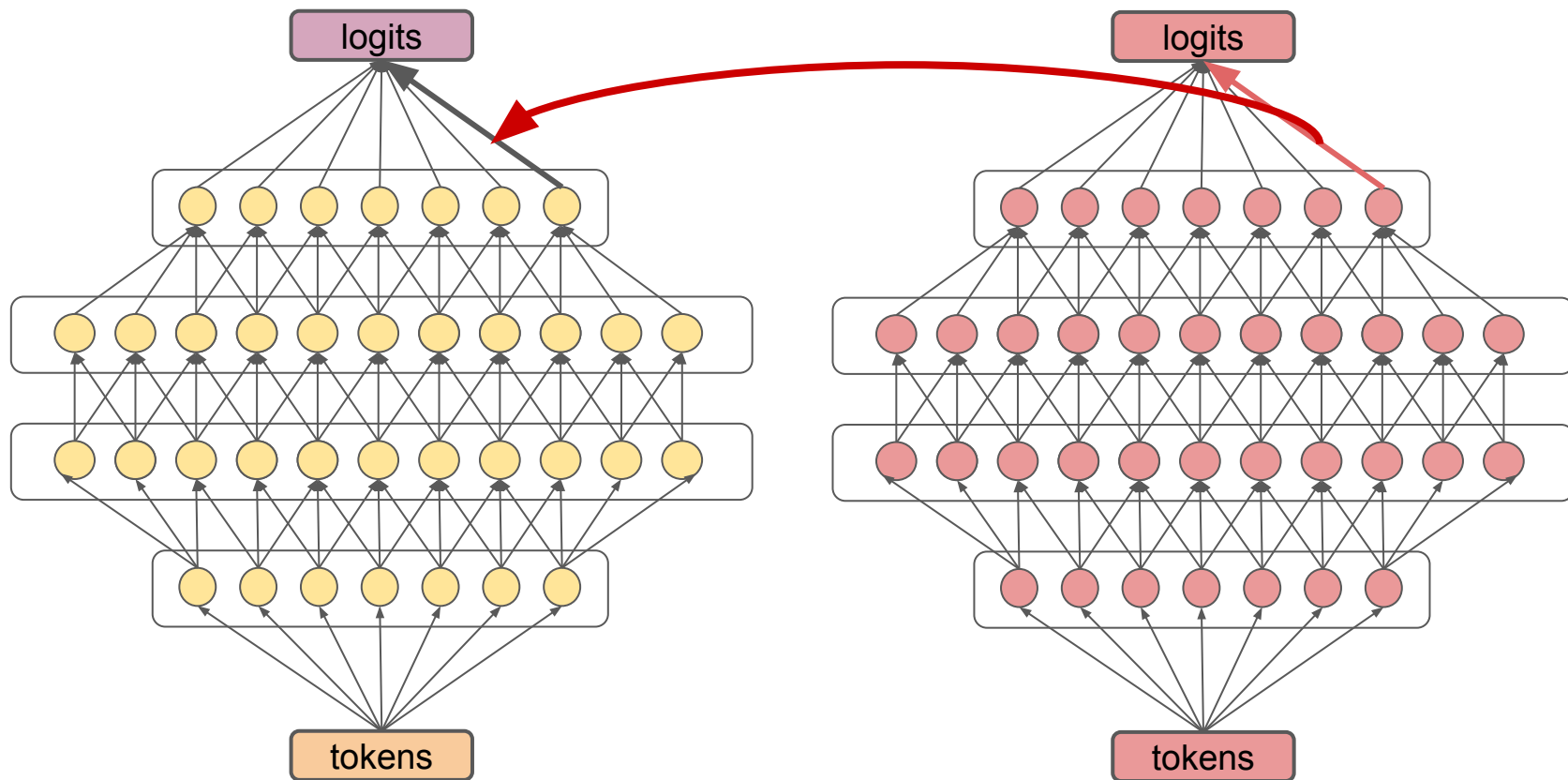
Wang et al. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small, 2023.

Path Patching: Circuit Discovery Algorithm (Intuition)



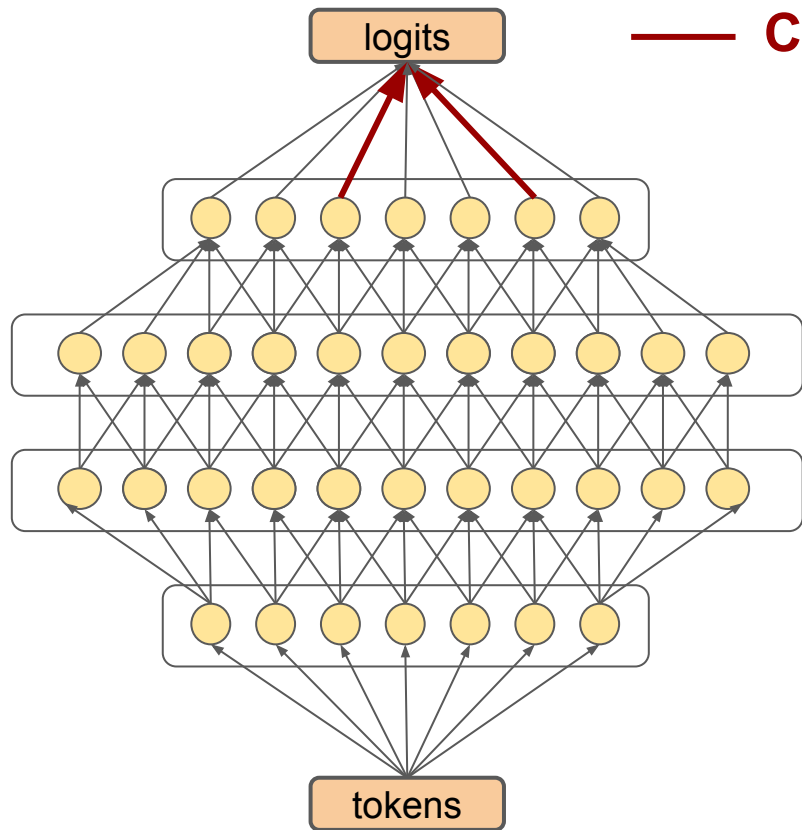
Wang et al. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small, 2023.

Path Patching: Circuit Discovery Algorithm (Intuition)



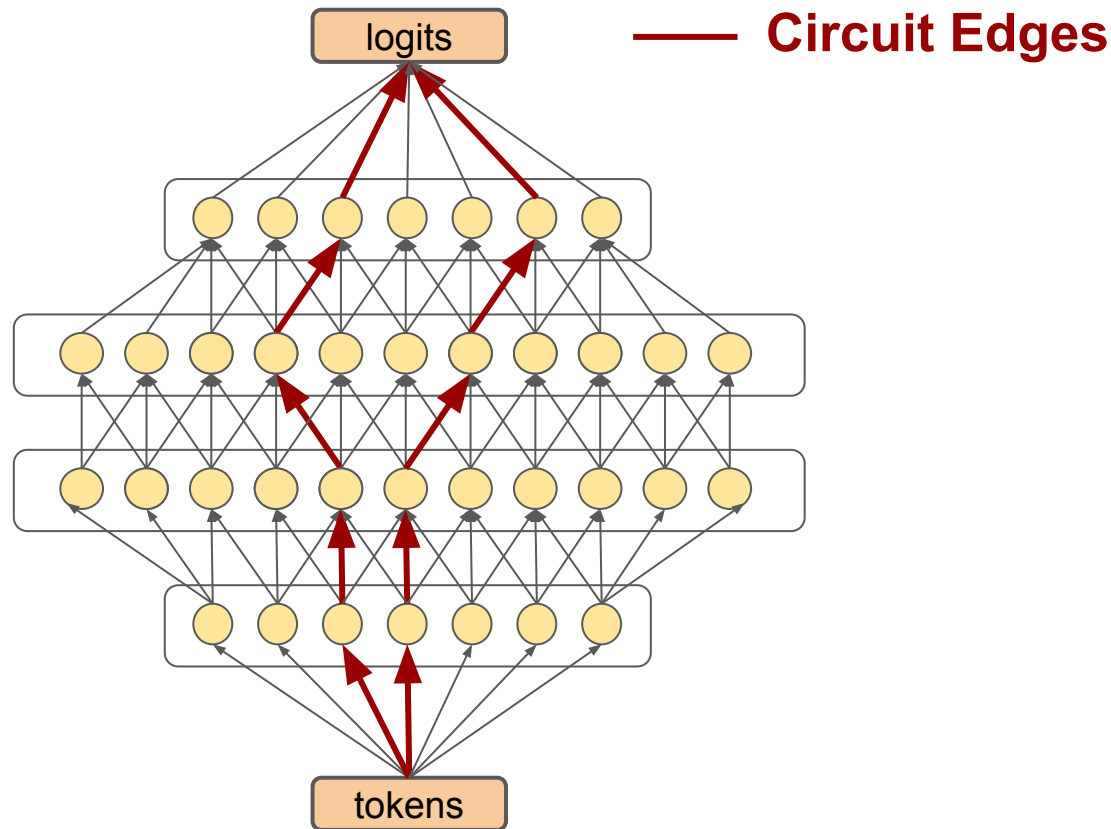
Wang et al. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small, 2023.

Path Patching: Circuit Discovery Algorithm (Intuition)

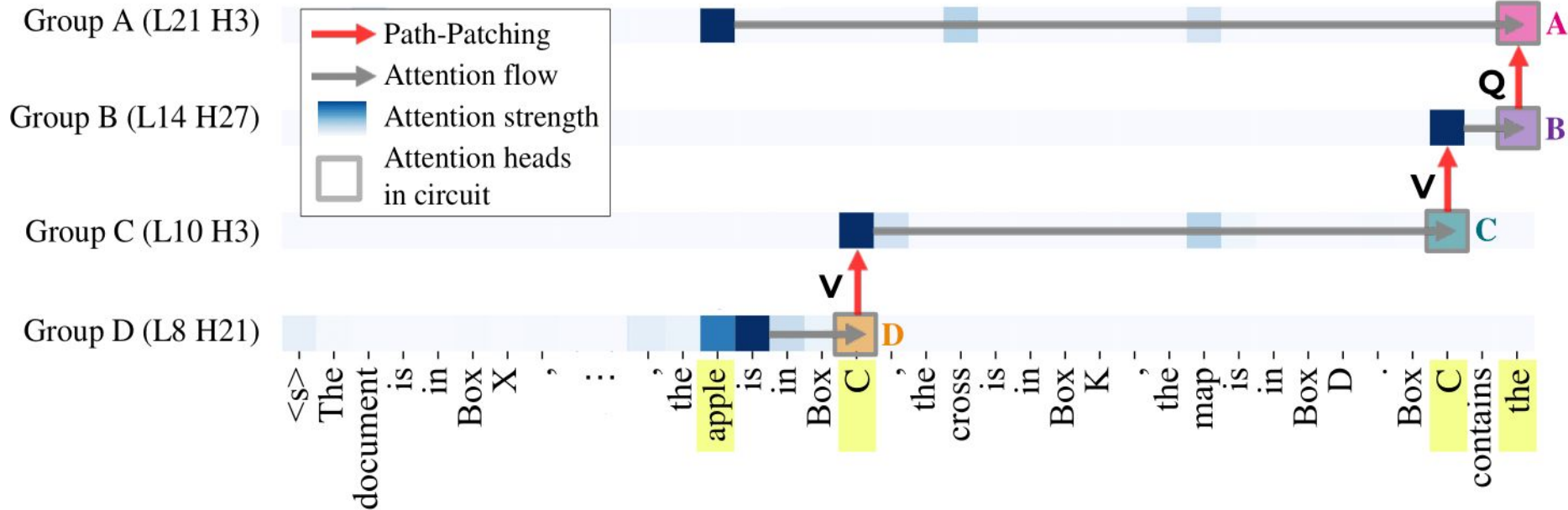


Edges that exhibited the greatest degradation in logit values.

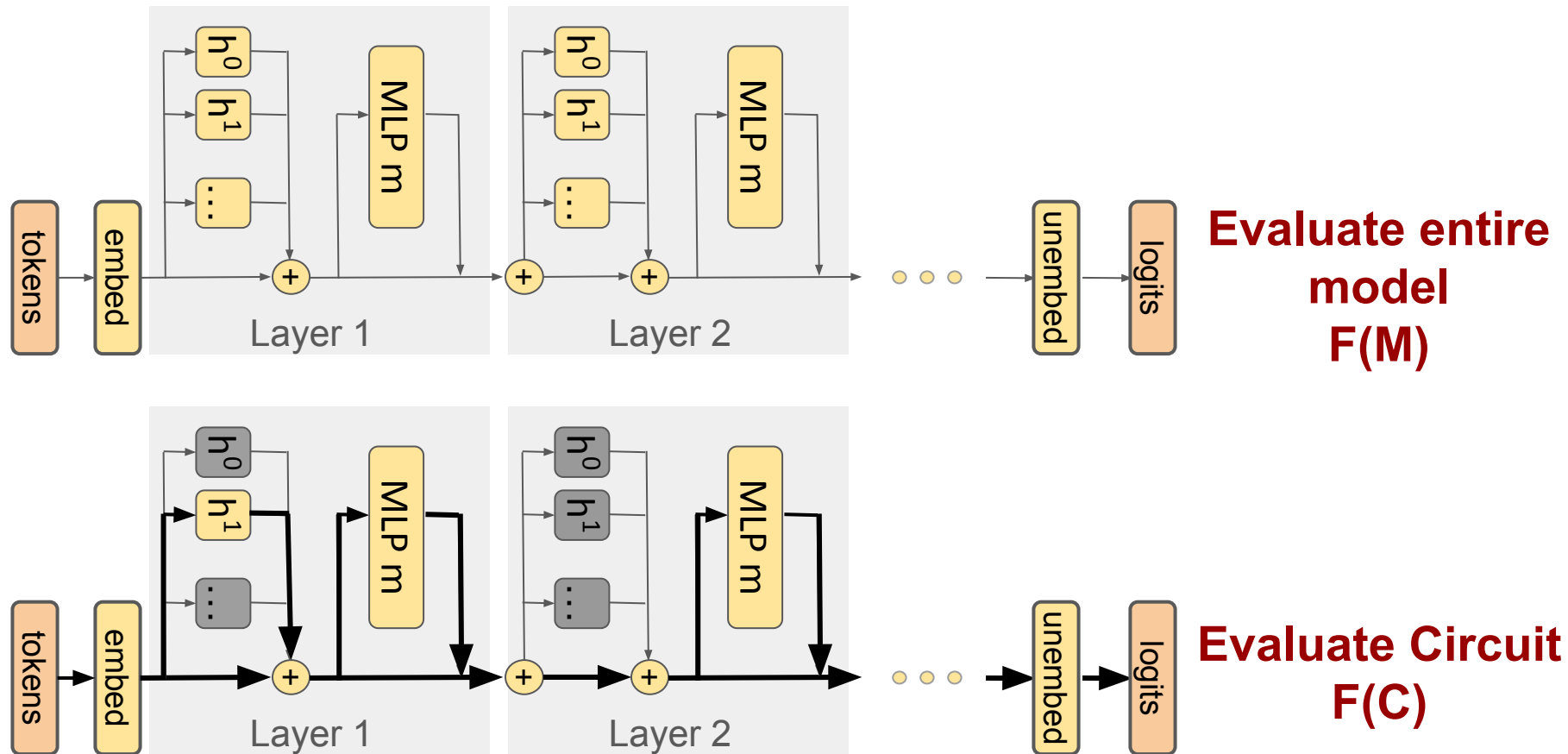
Path Patching: Circuit Discovery Algorithm (Intuition)



Entity Tracking Circuit in Llama-7B

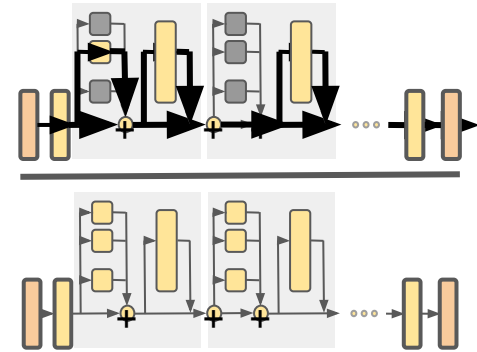


Faithfulness Metric for Evaluating Circuits




Faithfulness Metric for Evaluating Circuits

$$\text{Faithfulness of circuit } C = \frac{F(C)}{F(M)} =$$



Identified Circuit Can Recover Entire Model Performance

Model	Fine-tuned?	Accuracy			Faithfulness
		Full Model	Circuit	Random Circuit	
Llama-7B	-	0.66	0.66	0.00	1.00



Amazing! 75/55296 (<0.1%) attention heads recovers 100% of the model performance.

Identified LLaMa-7B Circuit also Present in Fine-tuned models

Model	Fine-tuned?	Accuracy			Faithfulness
		Full Model	Circuit	Random Circuit	
Llama-7B	-	0.66	0.66	0.00	1.00
Vicuna-7B	User conversations	0.67	0.65	0.00	0.97
Goat-7B	Arithmetic tasks (LoRA)	0.82	0.73	0.01	0.89
FLoat-7B	Arithmetic tasks (w/o LoRA)	0.82	0.72	0.01	0.88

Llama-7B circuit can restore at least **88%** of the entire fine-tuned models' performance.

- The same circuit can also restore at least 88% of the overall performance of the entire fine-tuned models.

Exactly same circuit roughly
constitutes the entity-tracking circuit
in the base and fine-tuned models!

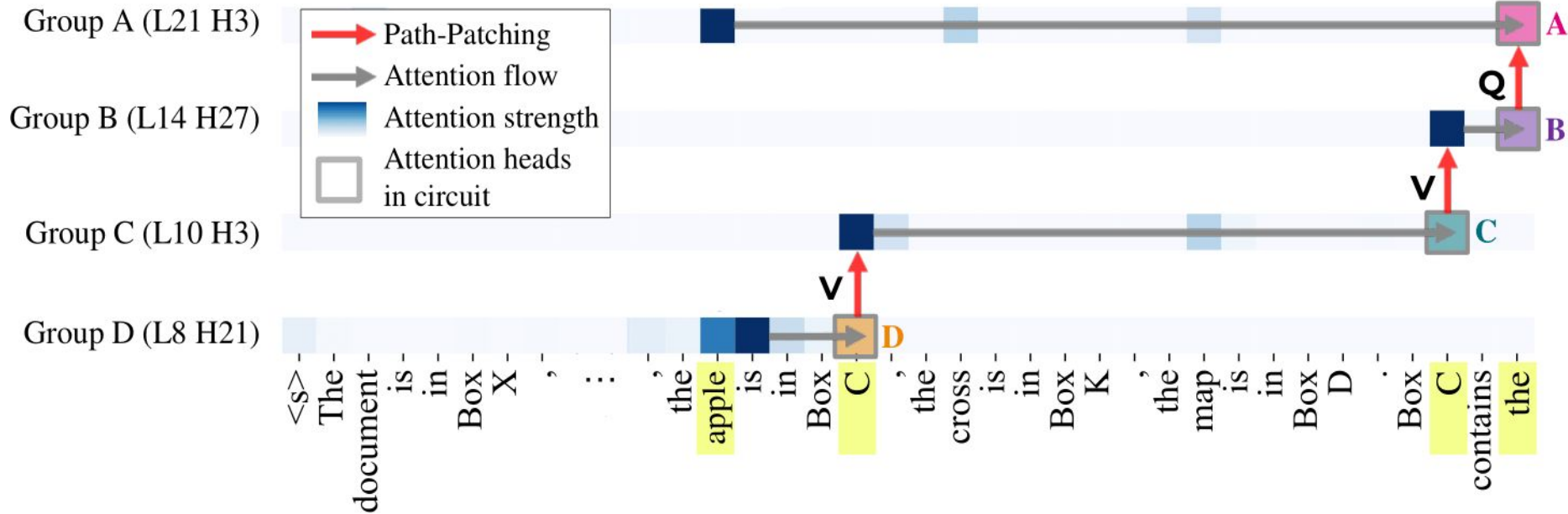


Model	Task	0.82	0.72	0.01	0.88
Llama-7B	Arithmetic tasks (LoRA)	0.82	0.72	0.01	0.88
Vicuna-7B	Arithmetic tasks (LoRA)	0.82	0.72	0.01	0.88
Goat-7B	Arithmetic tasks (LoRA)	0.82	0.72	0.01	0.88
FLoat-7B	Arithmetic tasks (w/o LoRA)	0.82	0.72	0.01	0.88

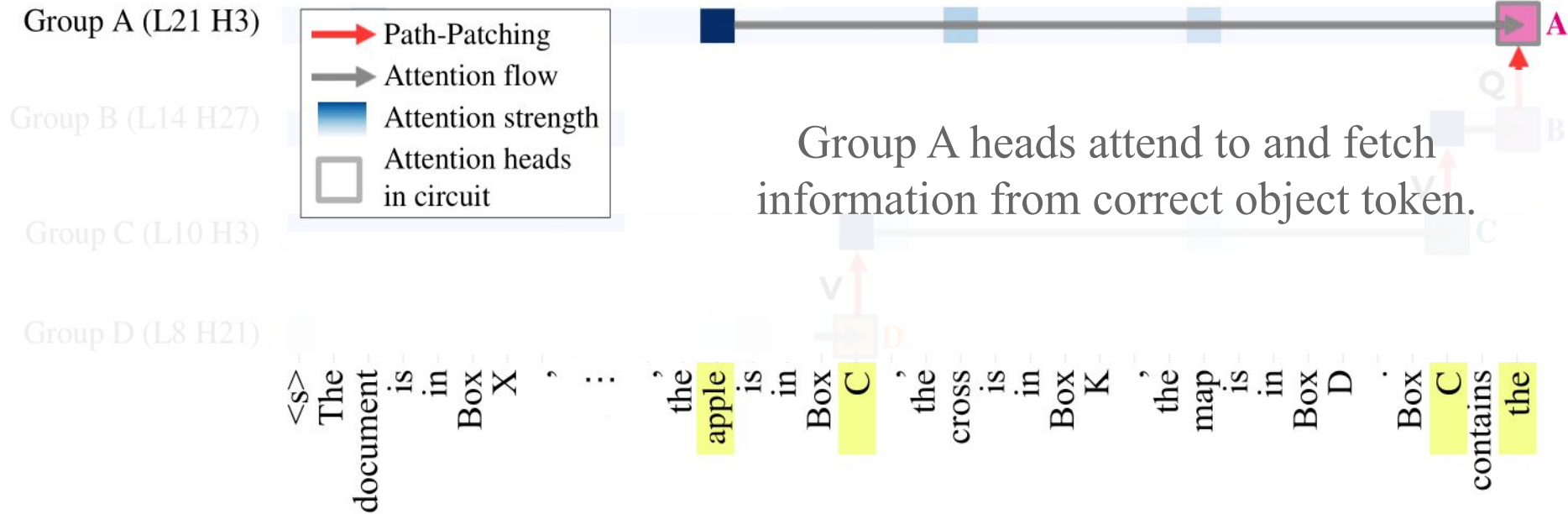
Hypotheses: Is circuit functionality the same after fine-tuning?

- Same circuit components have **varied functionality** in base and fine-tuned models.
- Same circuit implements **same mechanism**, but with an enhanced functionality.

Describing Circuit Components' Functionalities

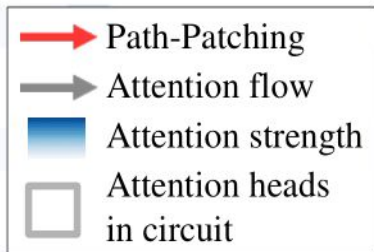


Describing Circuit Components' Functionalities



Describing Circuit Components' Functionalities

Group A (L21 H3)



Group B (L14 H27)

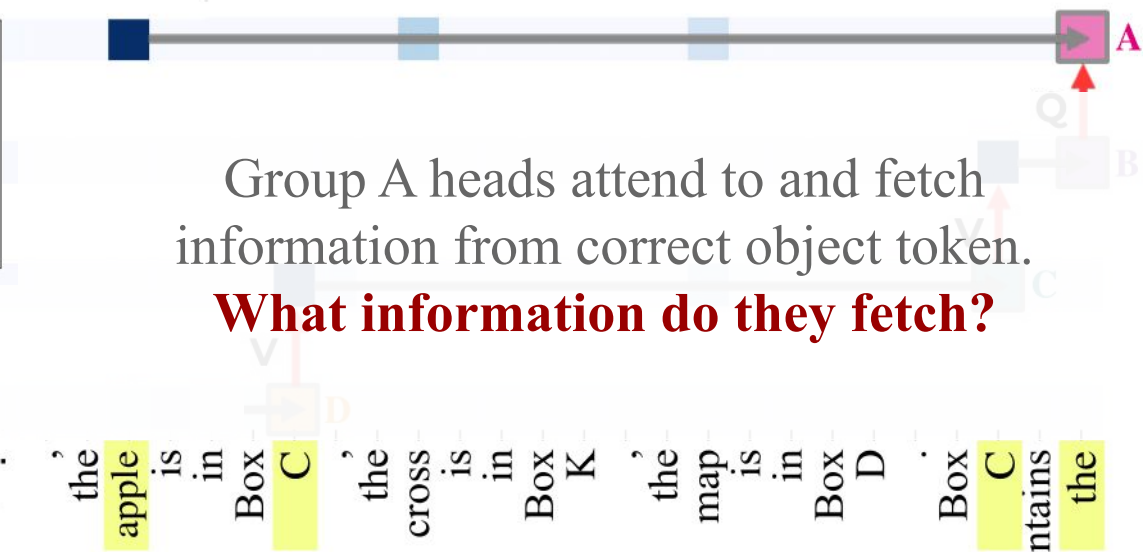
Group C (L10 H3)

Group D (L8 H21)

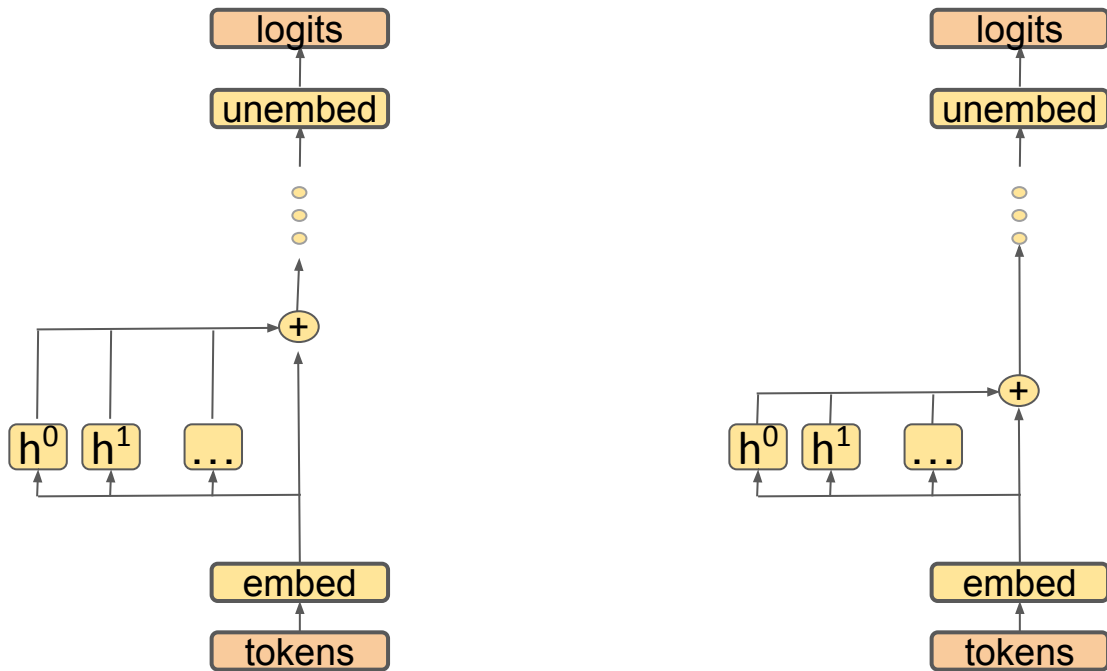
<s> The document is in Box X , : , the apple is in Box C , the cross is in Box K , the map is in Box D . Box C contains the

Group A heads attend to and fetch information from correct object token.

What information do they fetch?



Desiderata-based Component Masking (DCM)



Desiderata

Alternate

Original

Target

(a) Object

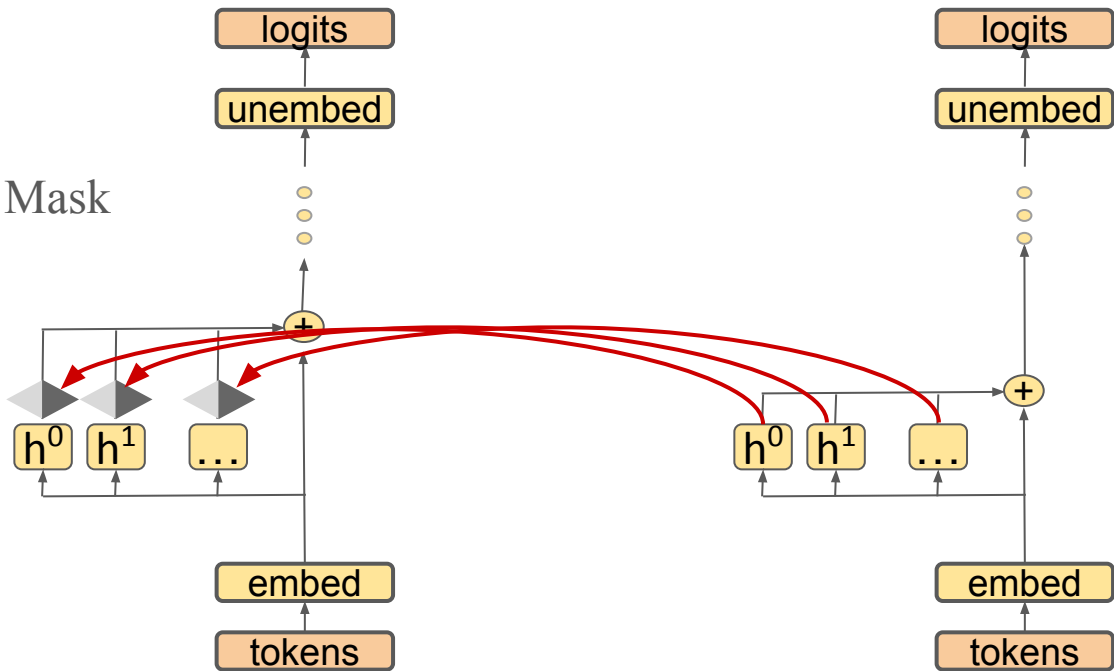
The book is in Box A, the cup is in Box B, the computer is in Box C, ... Box B contains the ____

The document is in Box X, the pot is in Box Y, the cross is in Box Z, ... Box X contains the ____

cup

Desiderata-based Component Masking (DCM)

Binary Mask



Desiderata

Alternate

Original

Target

(a) Object

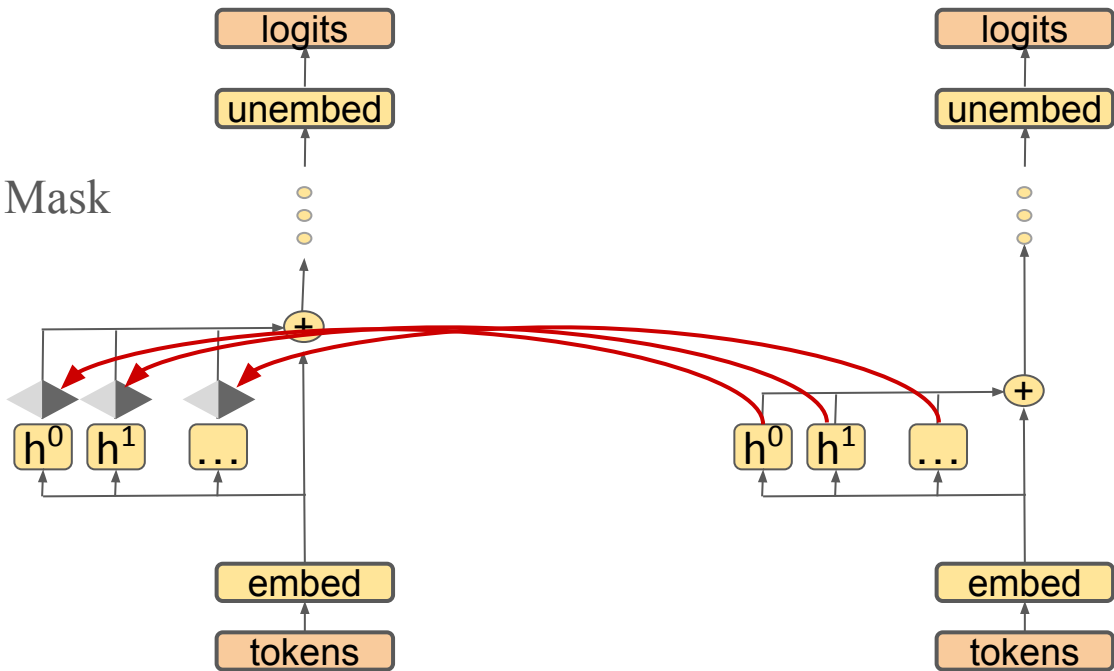
The book is in Box A, the cup is in Box B, the computer is in Box C, ... Box B contains the ____

The document is in Box X, the pot is in Box Y, the cross is in Box Z, ... Box X contains the ____

cup

Desiderata-based Component Masking (DCM)

Binary Mask



Desiderata

Alternate



Original

Target

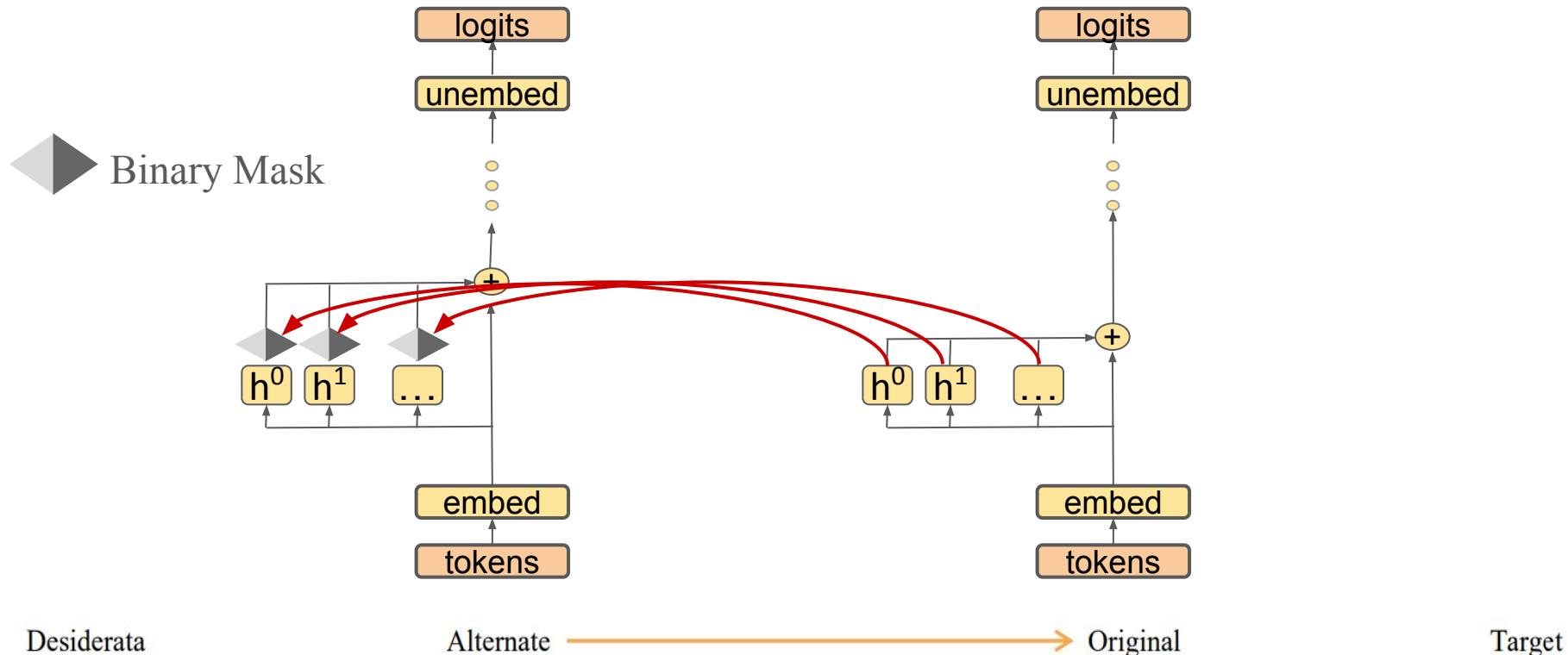
(b) Label

The book is in Box A, the cup is in Box B, the computer is in Box C, ... **Box Y** contains the ____

The document is in Box X, **the pot is in Box Y**, the cross is in Box Z, ... **Box X** contains the ____

pot

Desiderata-based Component Masking (DCM)



(c) Position

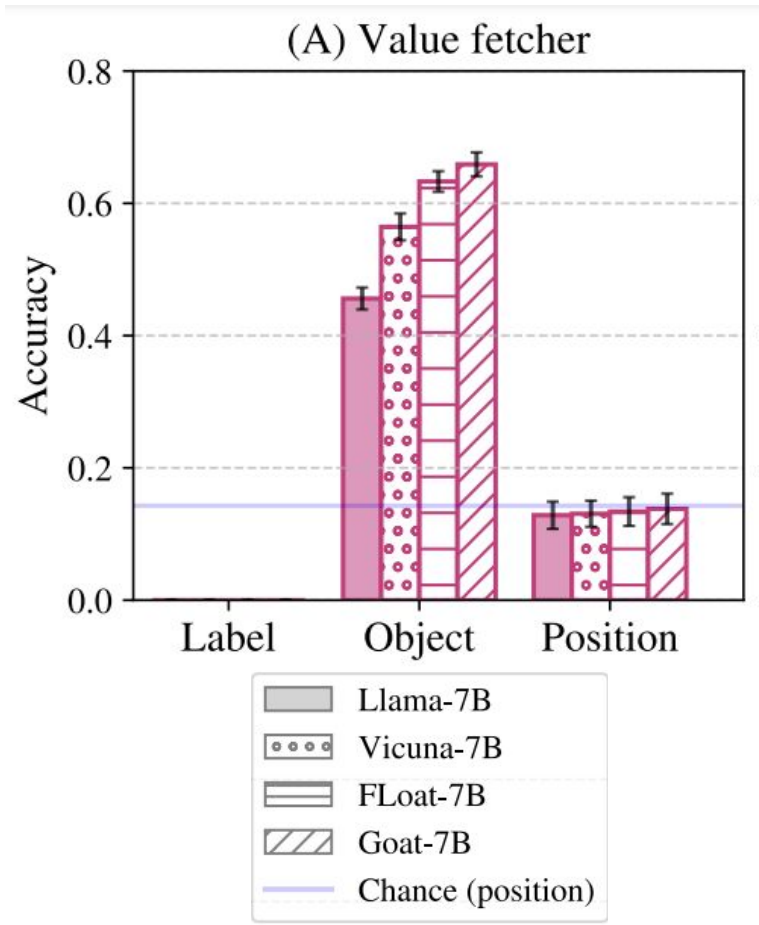
The book is in Box A, the cup is in Box B, the **computer** is in Box C, ... **Box C** contains the ____

The document is in Box X, the pot is in Box Y, the **cross** is in Box Z, ... **Box X** contains the ____

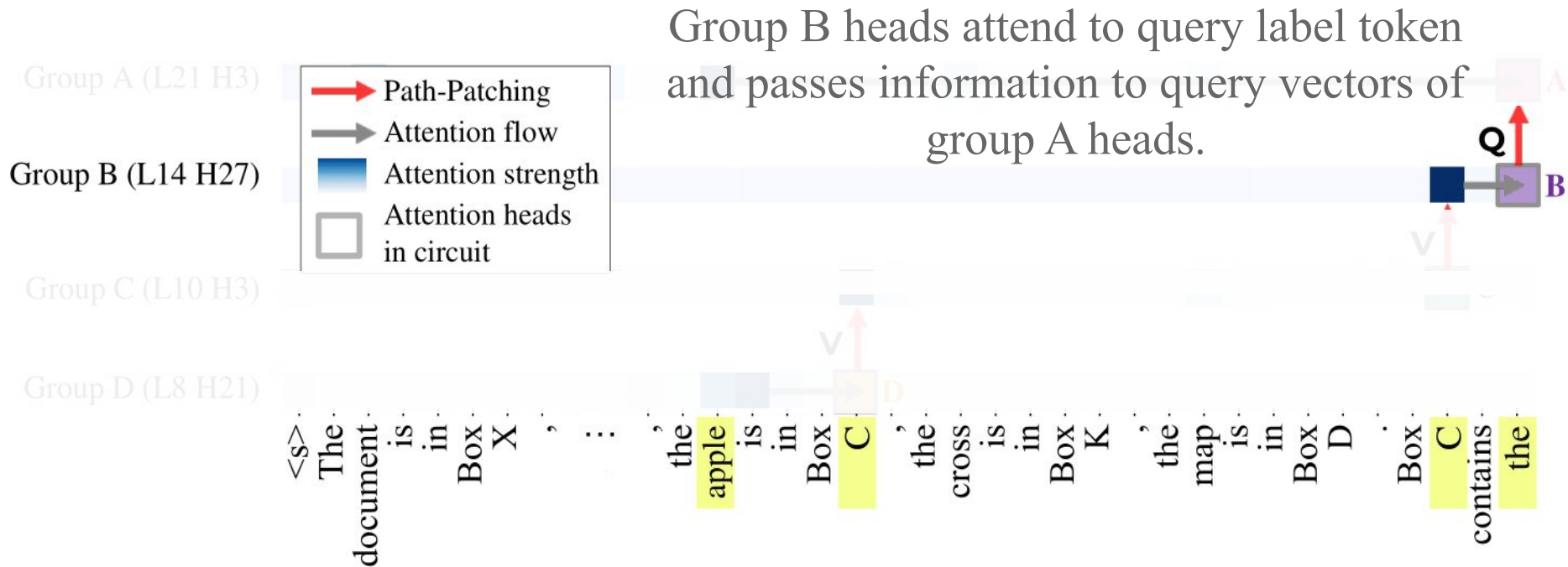
cross

Group A heads Fetch Value of Correct Object

We call them **Value Fetcher** heads.

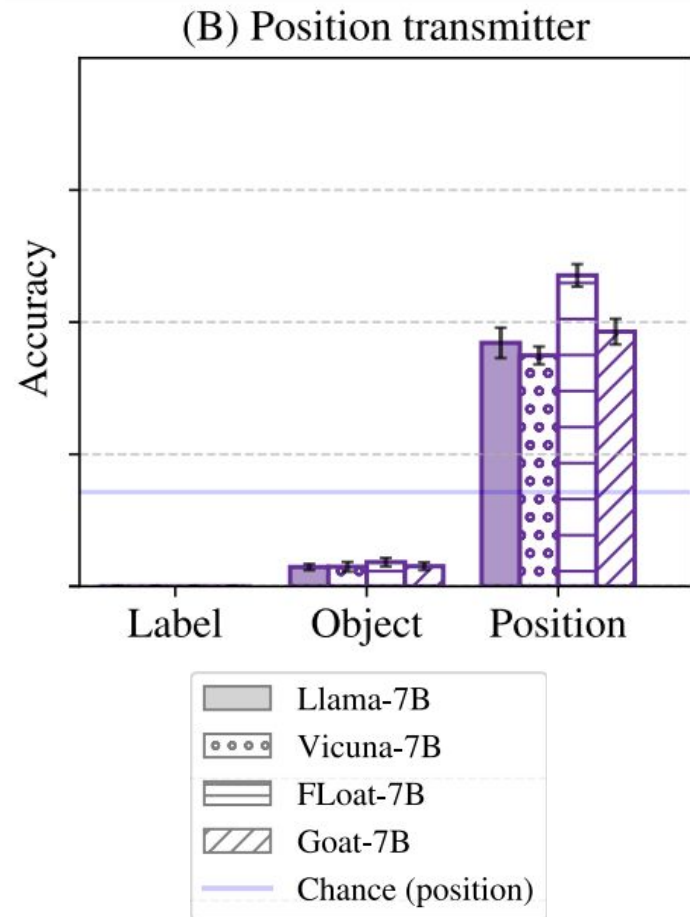


Entity Tracking Circuit in Llama-7B

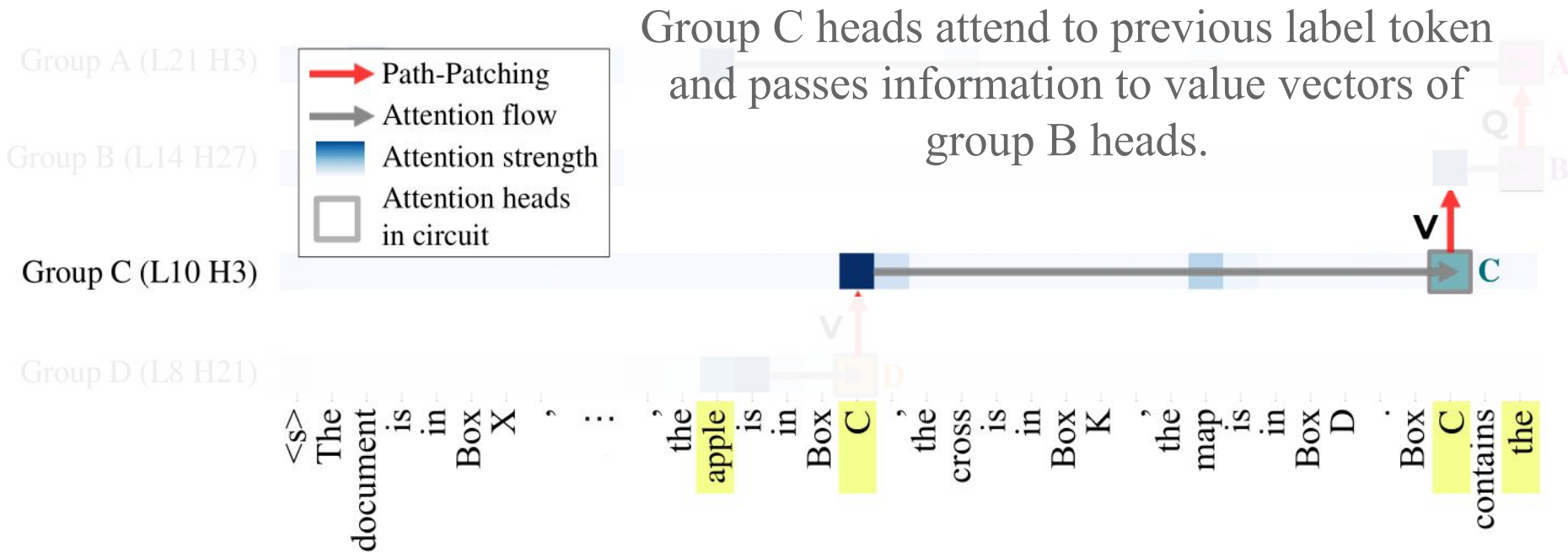


Group B heads Transmit Position of Correct Object

We call them **Position Transmitter** heads.

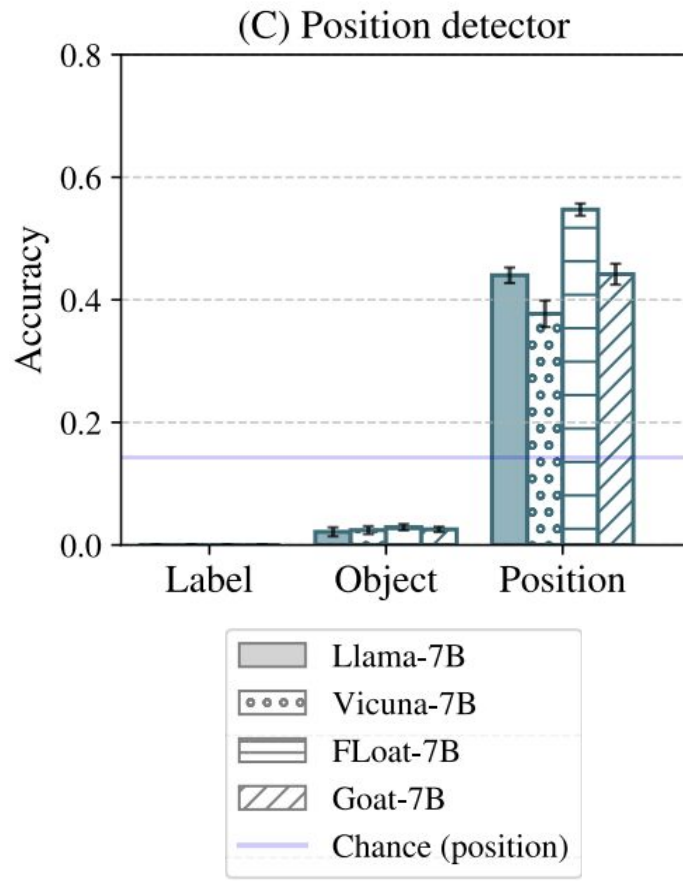


Entity Tracking Circuit in Llama-7B

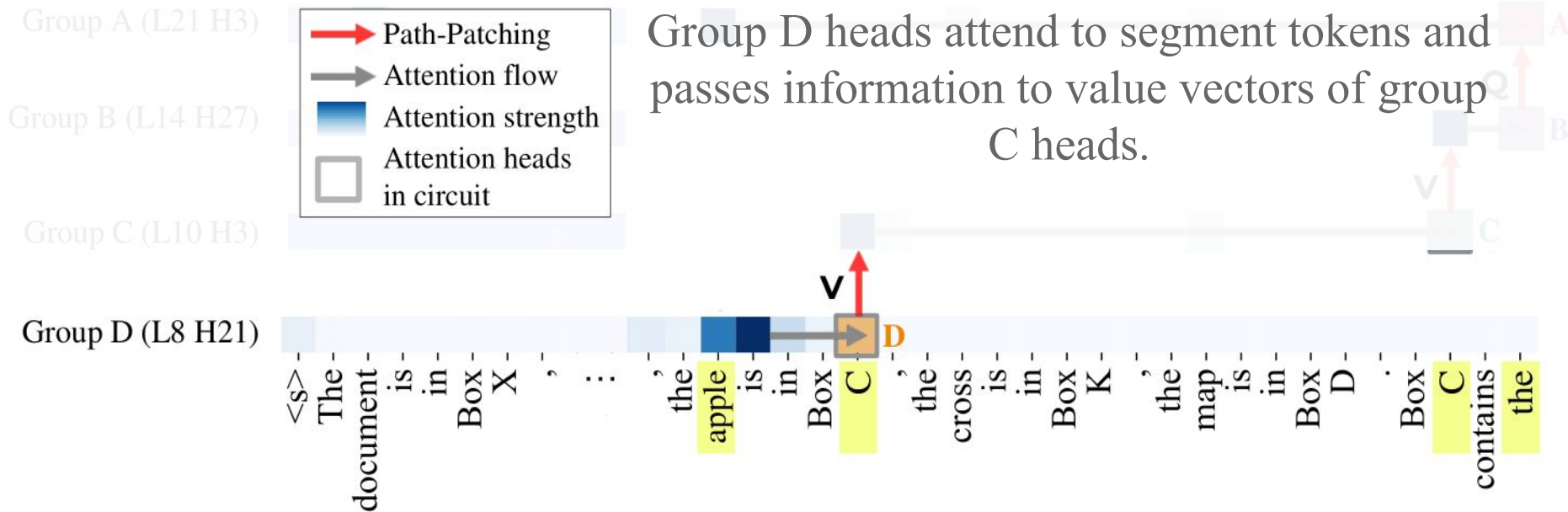


Group C heads Detect Position of Correct Object

We call them **Position Detector** heads.

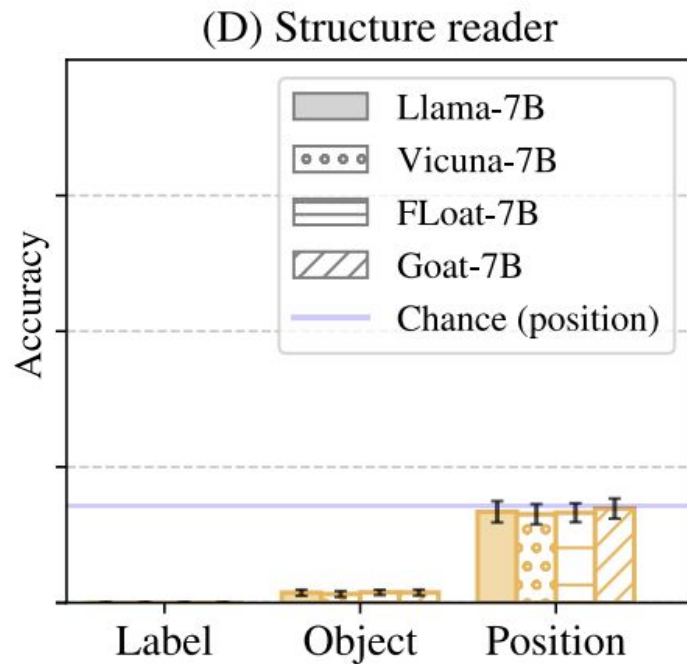


Entity Tracking Circuit in Llama-7B

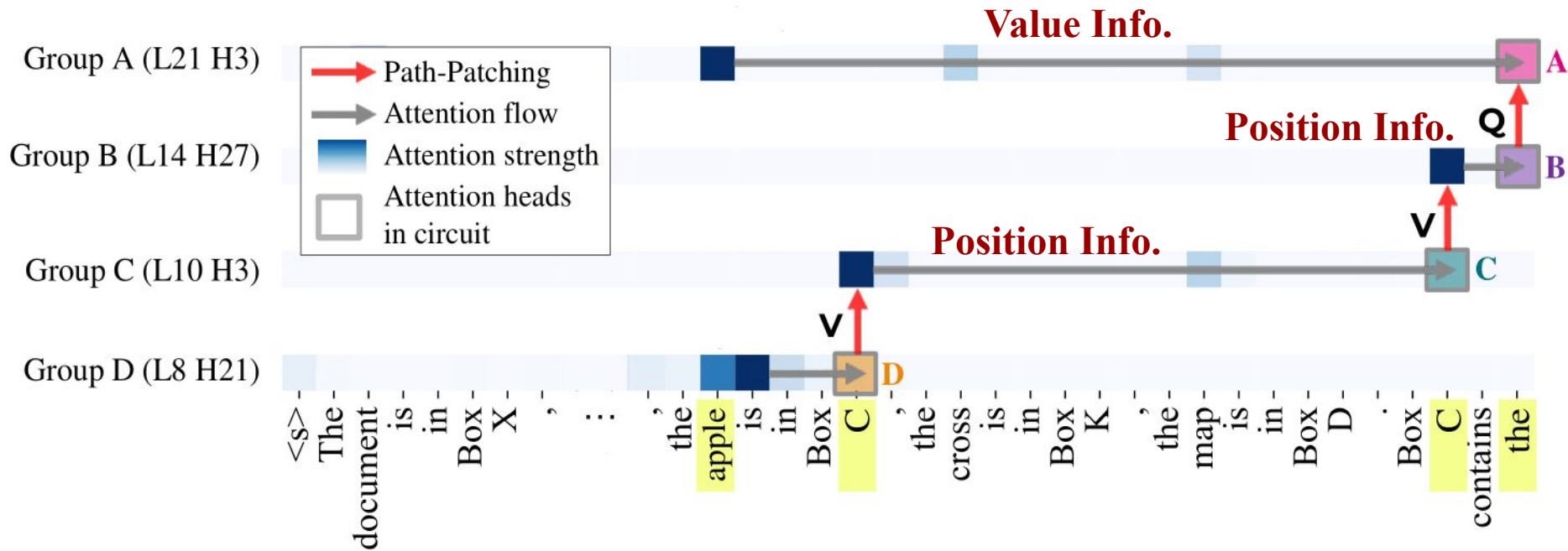


Group D heads Functionality Remains Mystery

We call them **Structure Reader** heads.

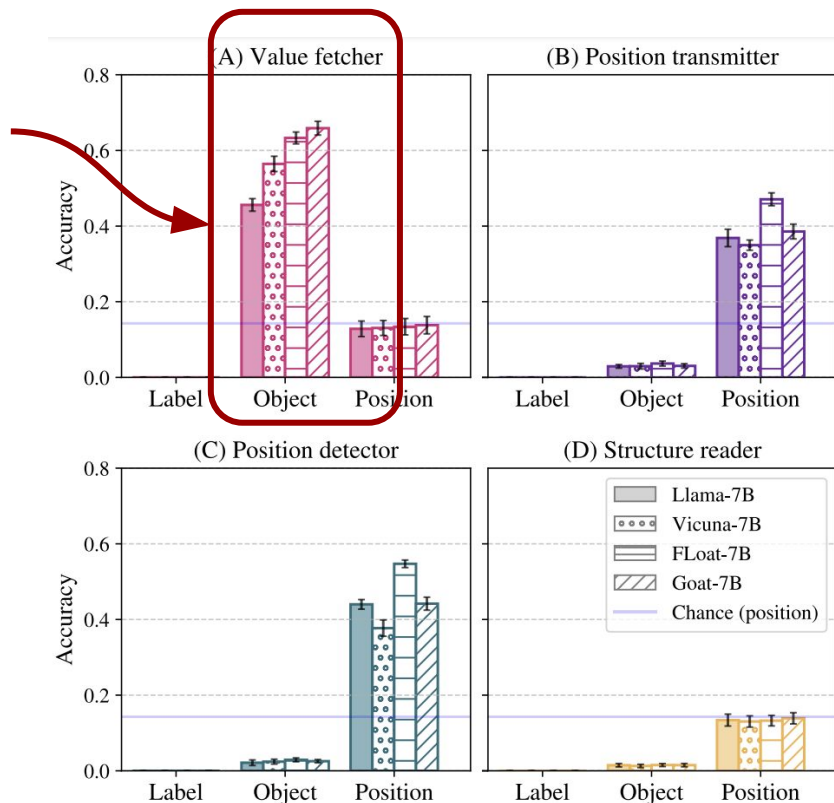


Entity Tracking Mechanism



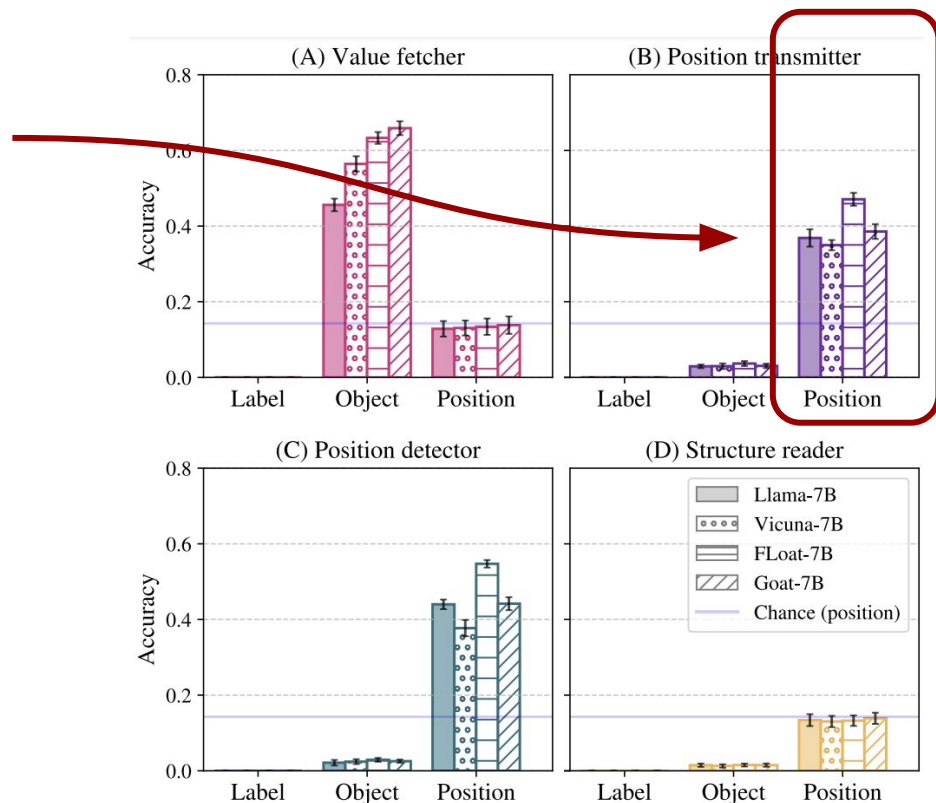
Functionality Remain Consistent Across Models

- Group A heads fetches value information across models.



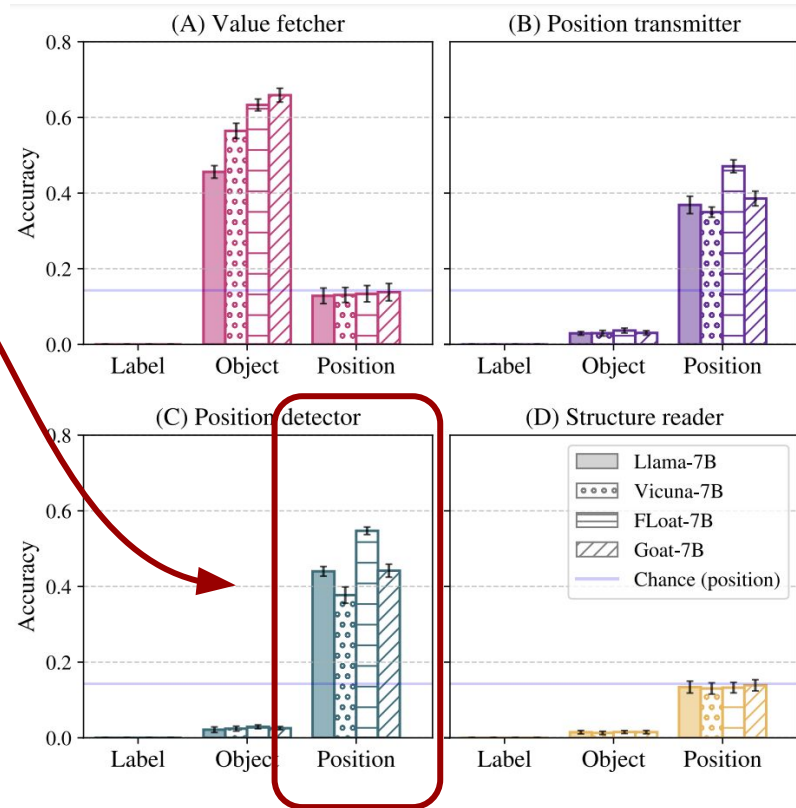
Functionality Remain Consistent Across Models

- Group B heads transmit positional information across models.



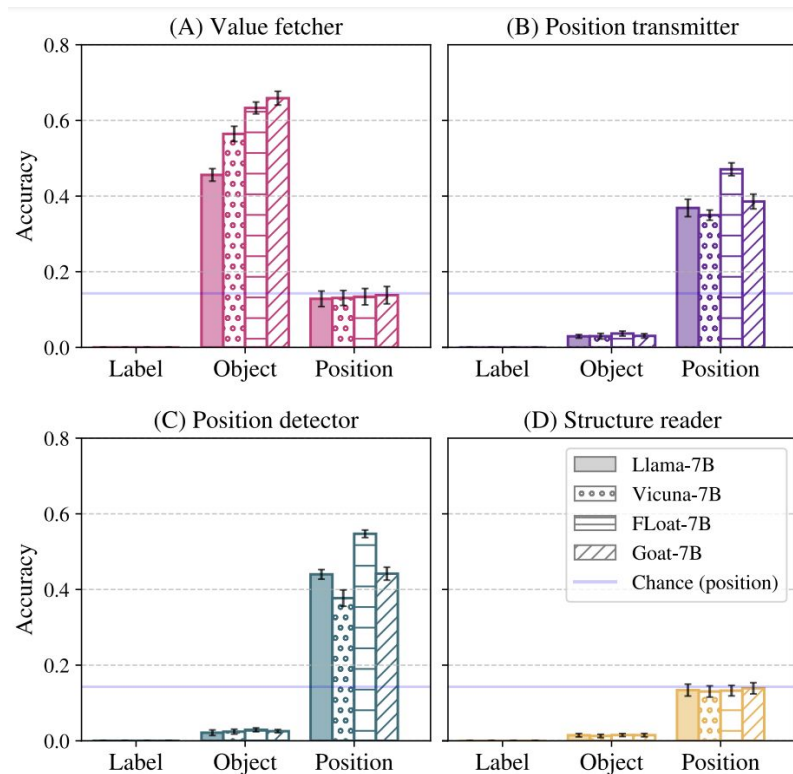
Functionality Remain Consistent Across Models

- Group C heads detect positional information across models.



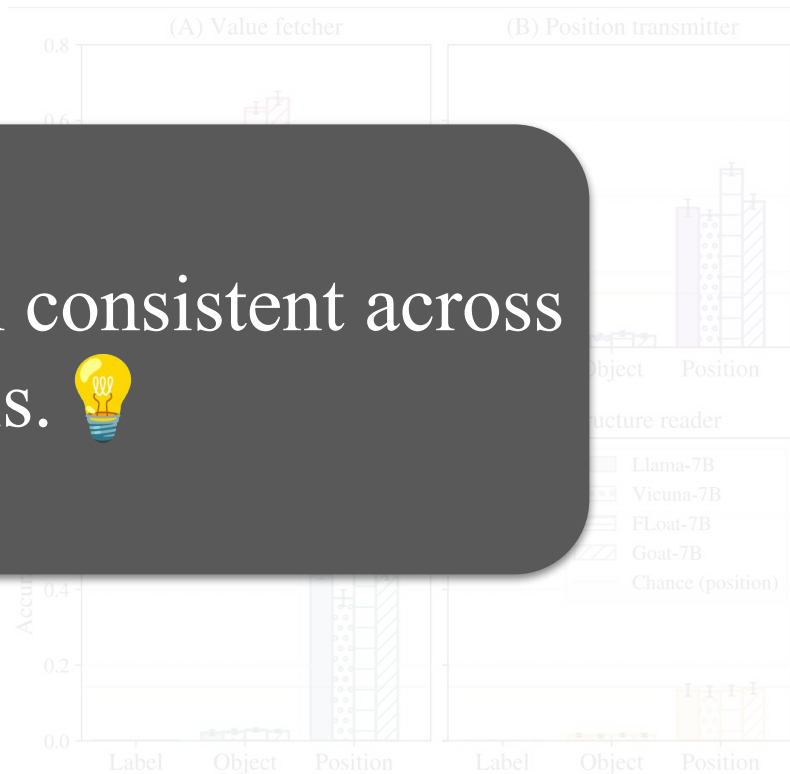
Mechanism Remain Consistent Across Models

Model	Circuit	Faithfulness
Llama-7B	0.66	1.00
Vicuna-7B	0.65	0.97
Goat-7B	0.73	0.89
FLoat-7B	0.72	0.88



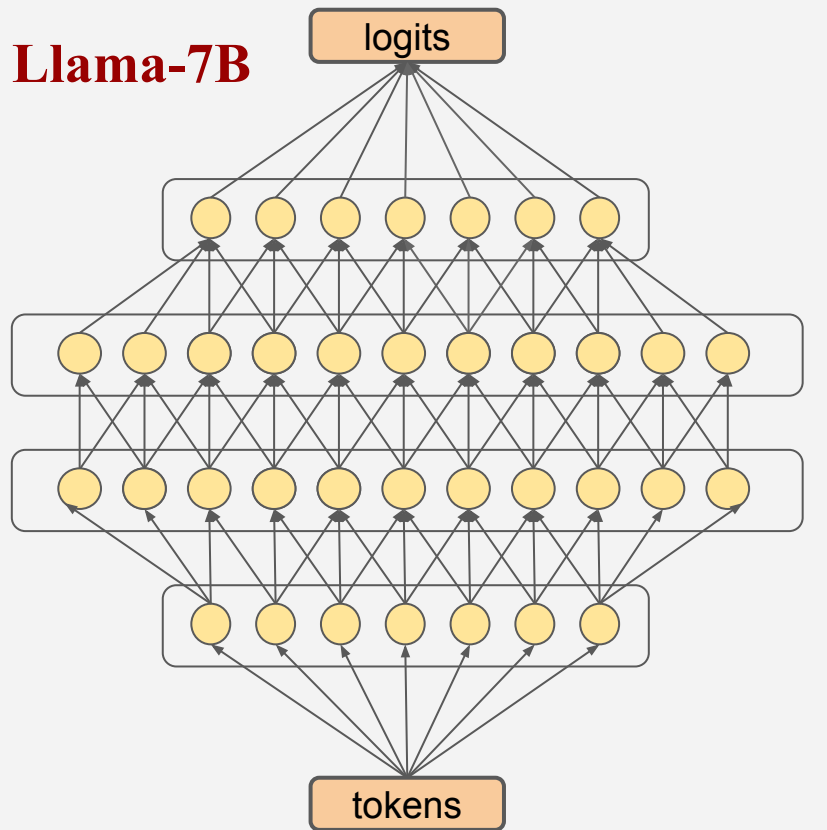
Model
Llama-7B
Vicuna-7B
Goat-7B
FLoat-7B

Mechanism remain consistent across models. 💡

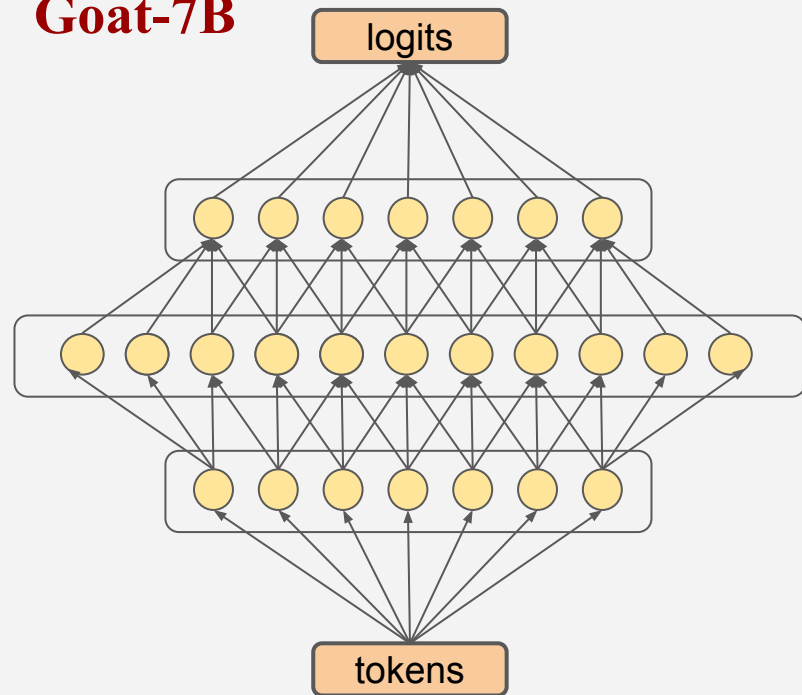


Cross-Model Activation Patching (CMAP)

Llama-7B

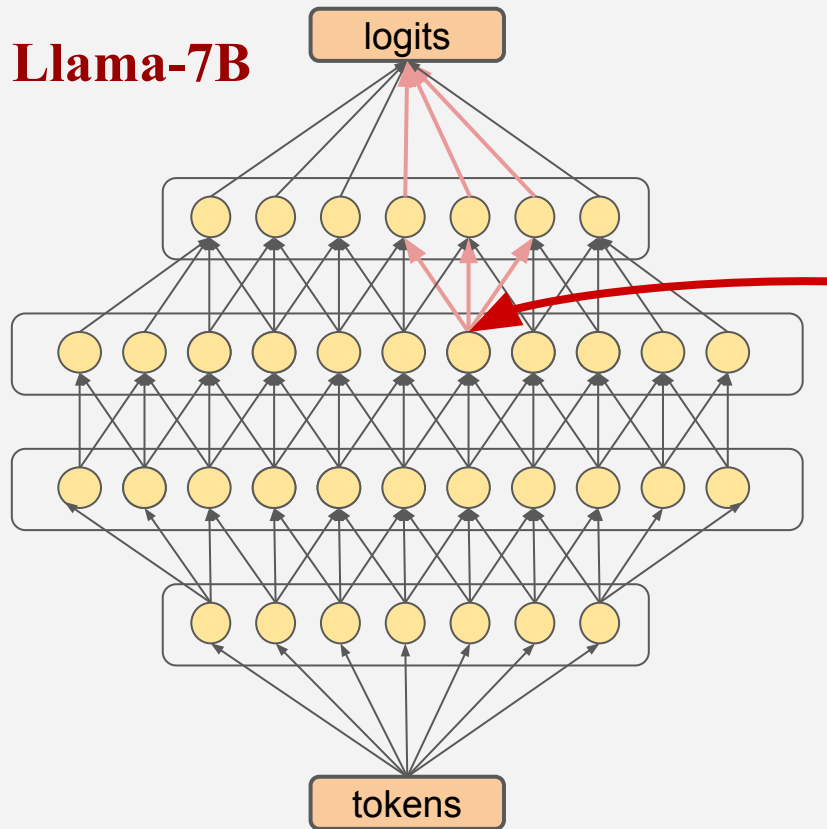


Goat-7B

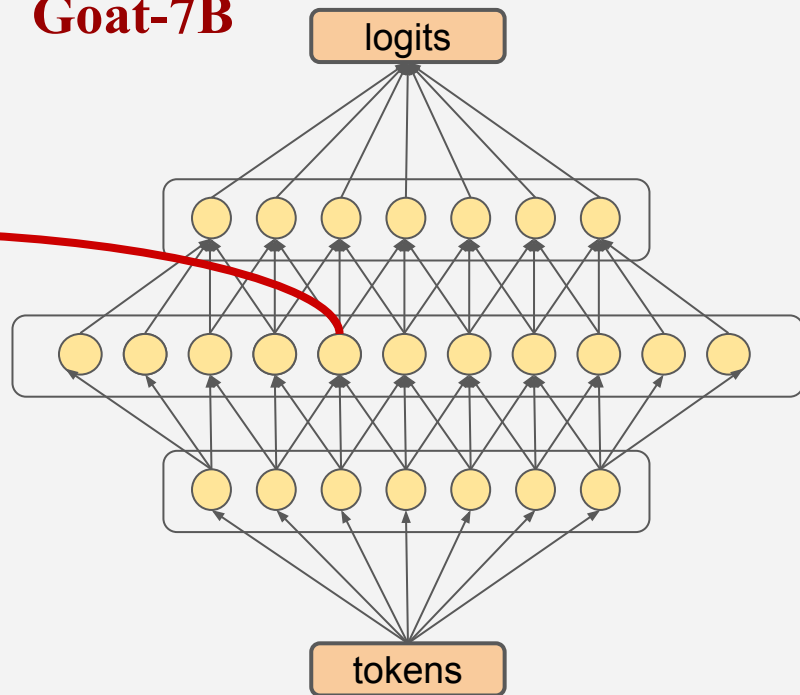


Cross-Model Activation Patching (CMAP)

Llama-7B

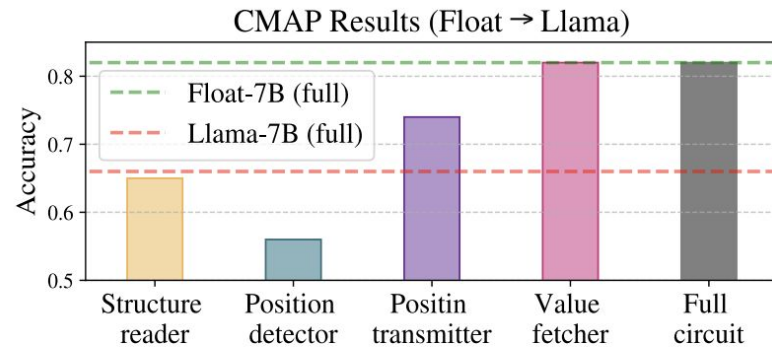
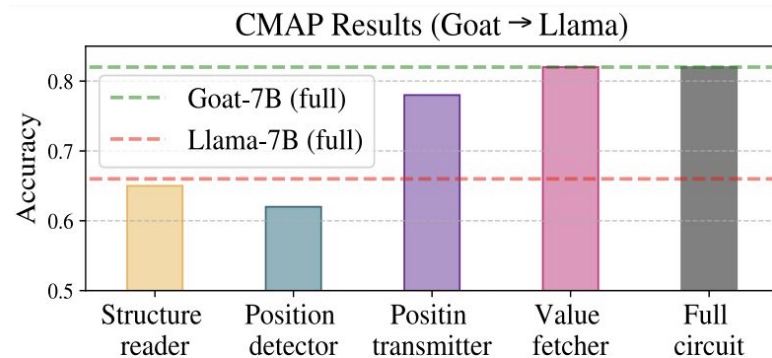


Goat-7B



Goat-7B and FLoat-7B Have Enhanced Sub-mechanisms

- Activations are compatible across models.
- **Improved representation of the correct object** in fine-tuned models.
- **Augmented positional information** in fine-tuned models.



FINE-TUNING ENHANCES EXISTING MECHANISMS

A Case Study On Entity Tracking



finetuning.baulab.info



arxiv.org/abs/2402.14811

