浙江大学
ZHEJIANG UNIVERSITY

Knowledge Circuits in Pretrained Transformers

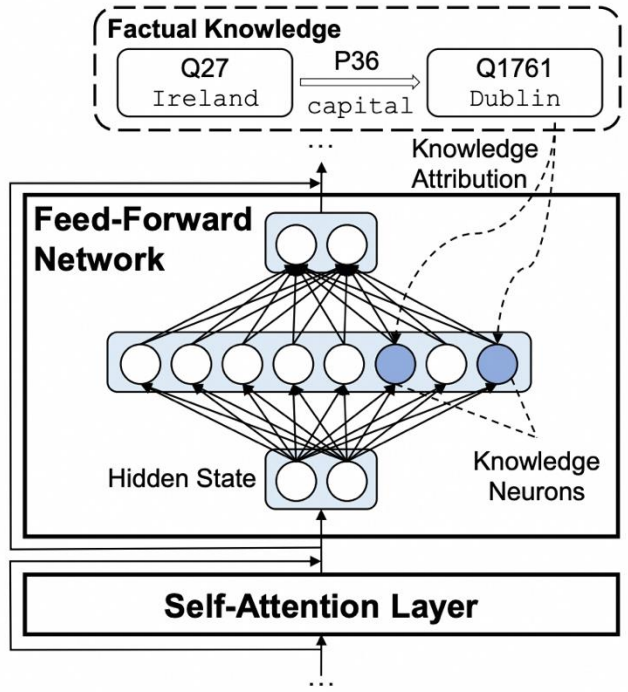
Yunzhi Yao, Ningyu Zhang, Zekun Xi, Mengru Wang, Ziwen Xu, Shumin Deng, Huajun Chen

Speaker： Yunzhi Yao @ Zhejiang University

Email： yyztodd@zju.edu.cn

Rethinking the knowledge in Transformers

❑ Feed Forward Networks(FFN) stores enormous Knowledge



Knowledge Neuron ？

Paradigm	Pre-edit	Post-edit	Δ
det_n_agr._2	100%	94.8%	-5.2%
dna._irr._2	99.5%	96.9%	-2.6%
dna._w._adj._2	97.1%	94.4%	-2.7%
dna._w._adj._irr._2	97.4%	95.4%	-2.0%

(b) These modifications of determiner-noun KNs are usually not enough to overturn the categorical prediction.

Data	Model	Reliability
ZsRE	T5-XL	22.51
	GPT-J	11.34
CounterFact	T5-XL	47.86
	GPT-J	1.66

(c) KN edit has low reliability for facts (Yao et al., 2023).

Editing the KNs is not enough to overturn the predictions (Niu et al., 2024)

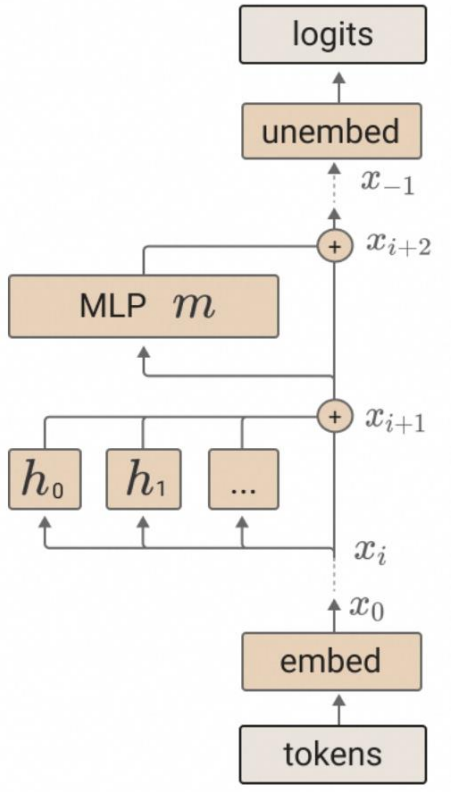
Knowledge Neurons in Pretrained Transformers (ACL 2024)

WHAT DOES THE KNOWLEDGE NEURON THESIS HAVE TO DO WITH KNOWLEDGE? (ICLR 2024)

Circuits in Transformer Structure

Definition: Circuit

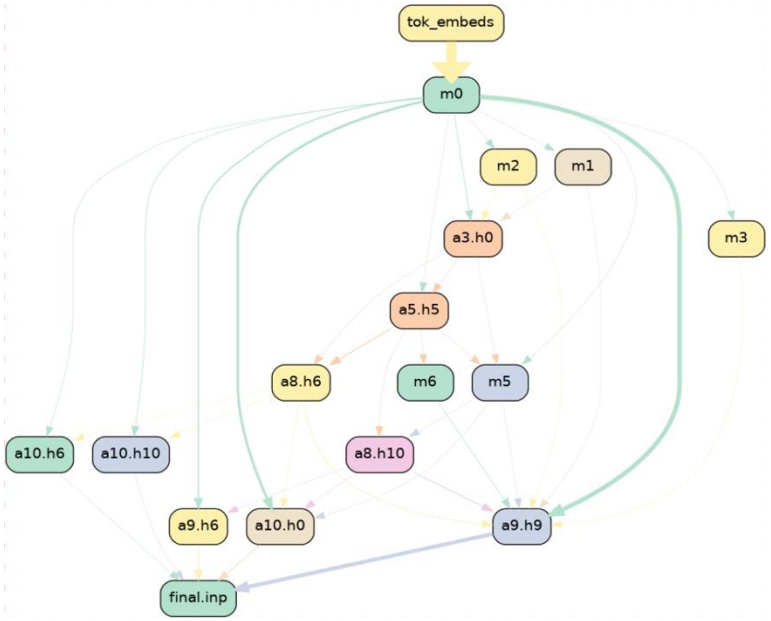
Circuits are sub-graphs of the network, consisting of features and the weights connecting them.



$$R_l = R_{l-1} + \sum_j A_{l,j} + M_l, R_0 = I$$

$$\text{Input}_l^A = I + \sum_{l'<l} \left(M_{l'} + \sum_{j'} A_{l',j'} \right)$$

$$\text{Input}_l^M = I + \sum_{l'<l} M_{i'} + \sum_{l'\leq i} \sum_{j'} A_{l',j'}$$



Knowledge Circuit Construction

1. View LM as a Directed acyclic graph (DAG)

$$\mathcal{G}$$

2. Overwrite the activation value of an edge with a corrupted activation

zero ablation in our experiments

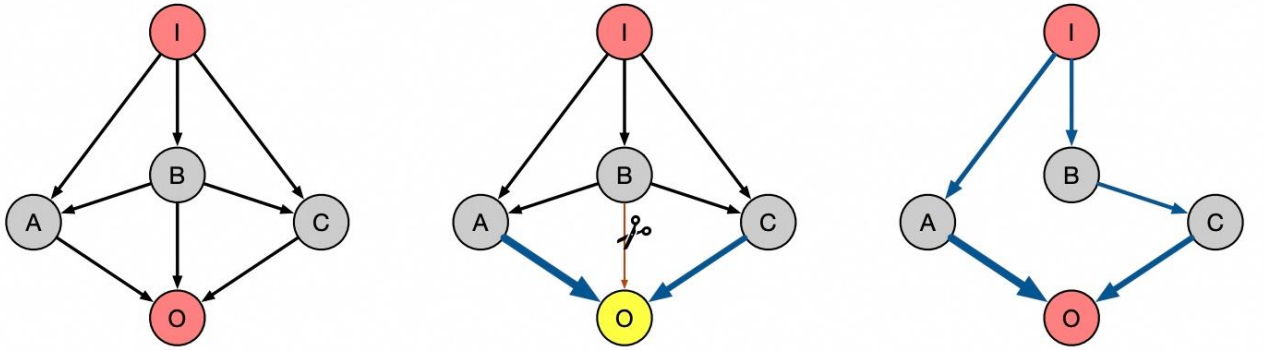
3. Run a forward pass through the model, and compare the output values of the new model with the original model.

$$S(e_i) < \tau$$

$$\mathcal{C}_{temp} \leftarrow \mathcal{G}/e_i$$

4. Finally get the subgraph that contributes to the knowledge expression.

$$\mathcal{C}_k =< N_k, E_k >$$



(a) Choose computational graph, task, and threshold τ.

(b) At each head, prune unimportant connections.

(c) Recurse until the full circuit is recovered.

$$k = (s, r, o)$$

$$S(e_i) = \log(\mathcal{G}/e_i(o|(s,r))) - \log(\mathcal{G}(o|(s,r)))$$

Evaluation of the Identified Circuit

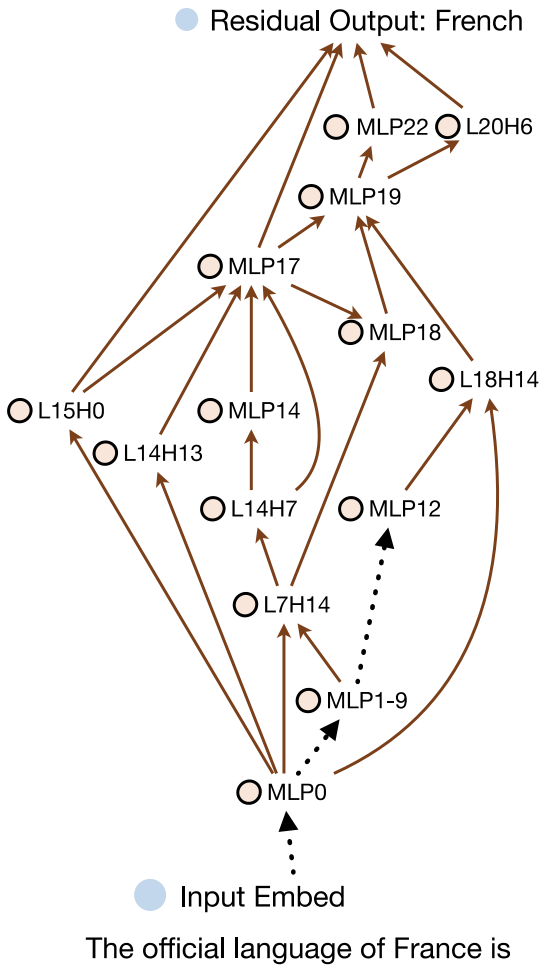
❑Experiment Results in GPT2-Medium

Table 1: **Hit@10** of the Original and Circuit Standalone performance of knowledge circuit in GPT2-Medium. **The result for D_{val} being 1.0 indicates that we select the knowledge for which the model provides the correct answer to build the circuit.**

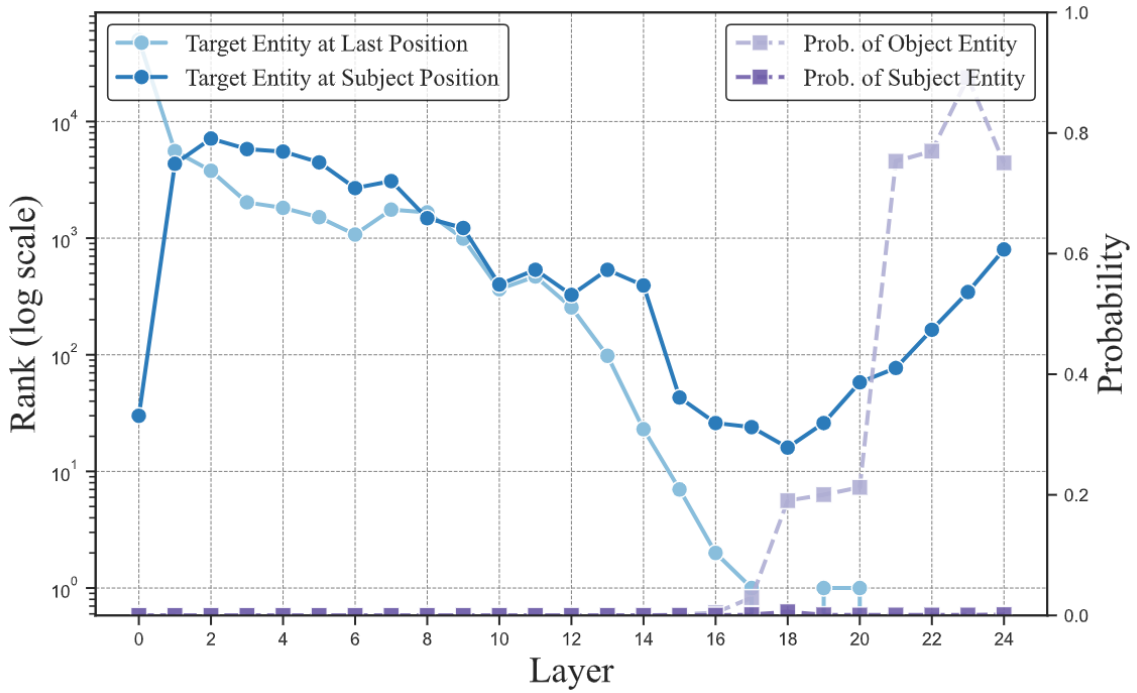
Type	Knowledge	#Edge	D_{val}		D_{test}	
			Original($\mathcal{G}$)	Circuit($\mathcal{C}$)	Original($\mathcal{G}$)	Circuit($\mathcal{C}$)
Linguistic	Adj Antonym	573	0.80	1.00 ↑	0.00	0.40 ↑
	world first letter	432	1.00	0.88	0.36	0.16
	world last letter	230	1.00	0.72	0.76	0.76
Commonsense	object superclass	102	1.00	0.68	0.64	0.52
	fruit inside color	433	1.00	0.20	0.93	0.13
	work location	422	1.00	0.70	0.10	0.10
Factual	Capital City	451	1.00	1.00	0.00	0.00
	Landmark country	278	1.00	0.60	0.16	0.36 ↑
	Country Language	329	1.00	1.00	0.16	0.75 ↑
	Person Native Language	92	1.00	0.76	0.50	0.76 ↑
Bias	name religion	423	1.00	0.50	0.42	0.42
	occupation age	413	1.00	1.00	1.00	1.00
	occupation gender	226	1.00	0.66	1.00	0.66
	name birthplace	276	1.00	0.57	0.07	0.57 ↑
Avg			0.98	0.73	0.44	0.47 ↑

An identified circuit in GPT2-Medium

❑View the Information Flow in the Circuit



Logit Lens: $\mathbf{W}_U \, \mathrm{LN}(n_i)$

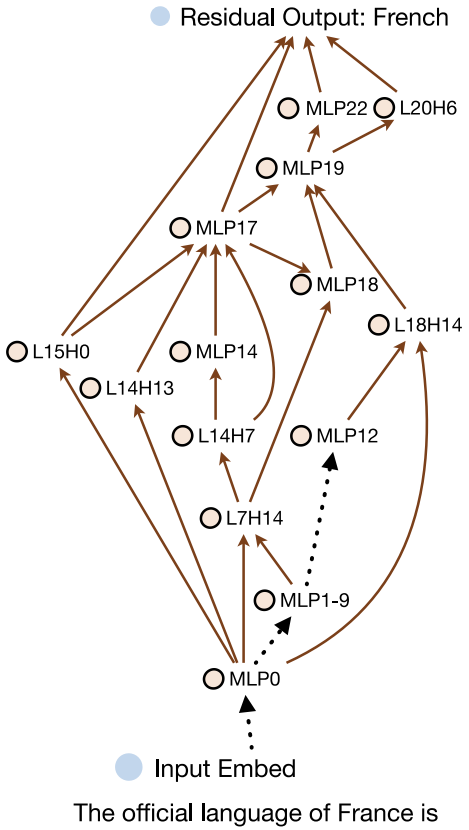


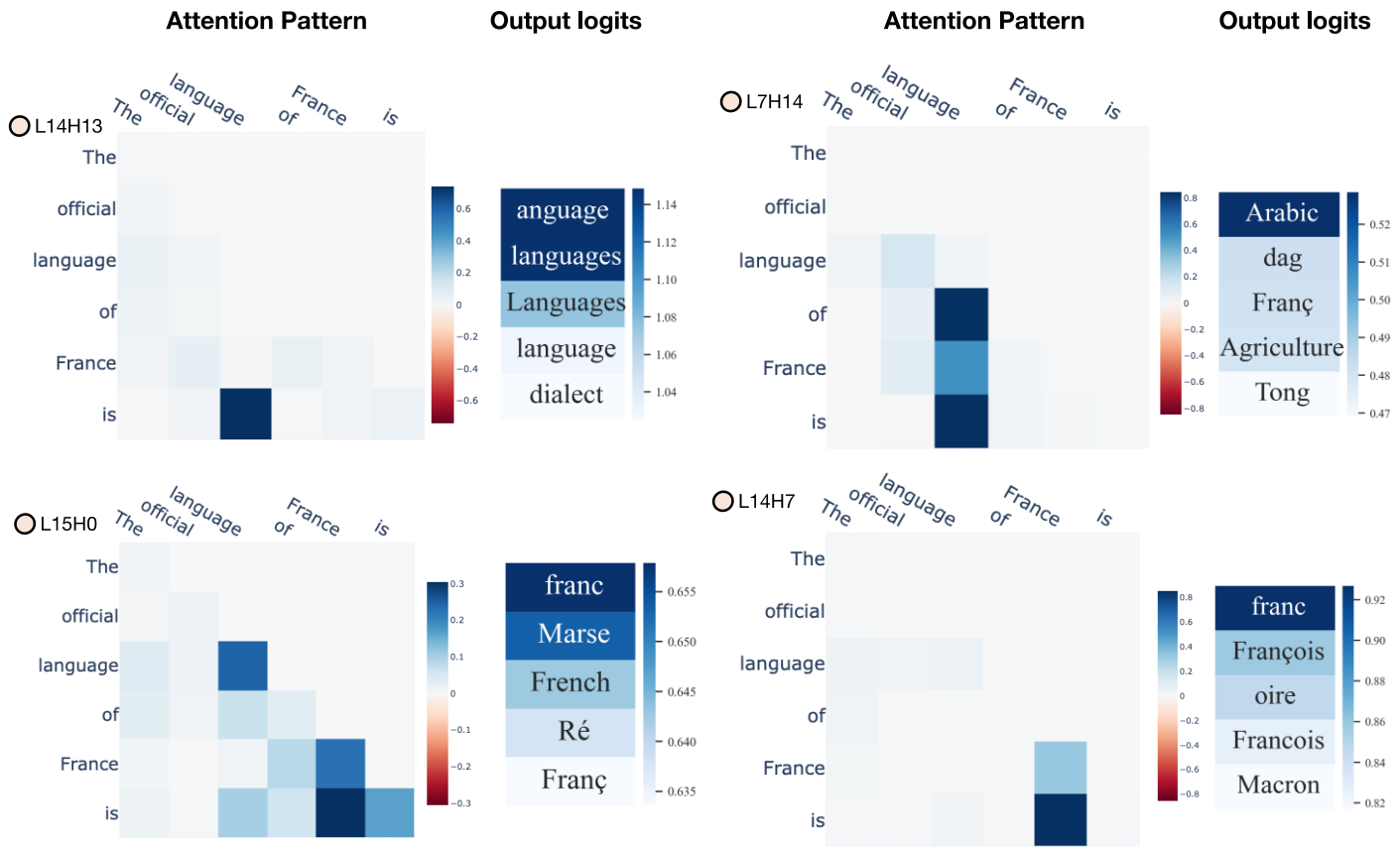
At layer 17, French becomes the token of top1 in the vocabulary space, and then its probability gradually increases.

Some special attention head in the circuit

① Mover Head： move the information at the subject position to the last token

② Relation Head： attend the relation token in the context

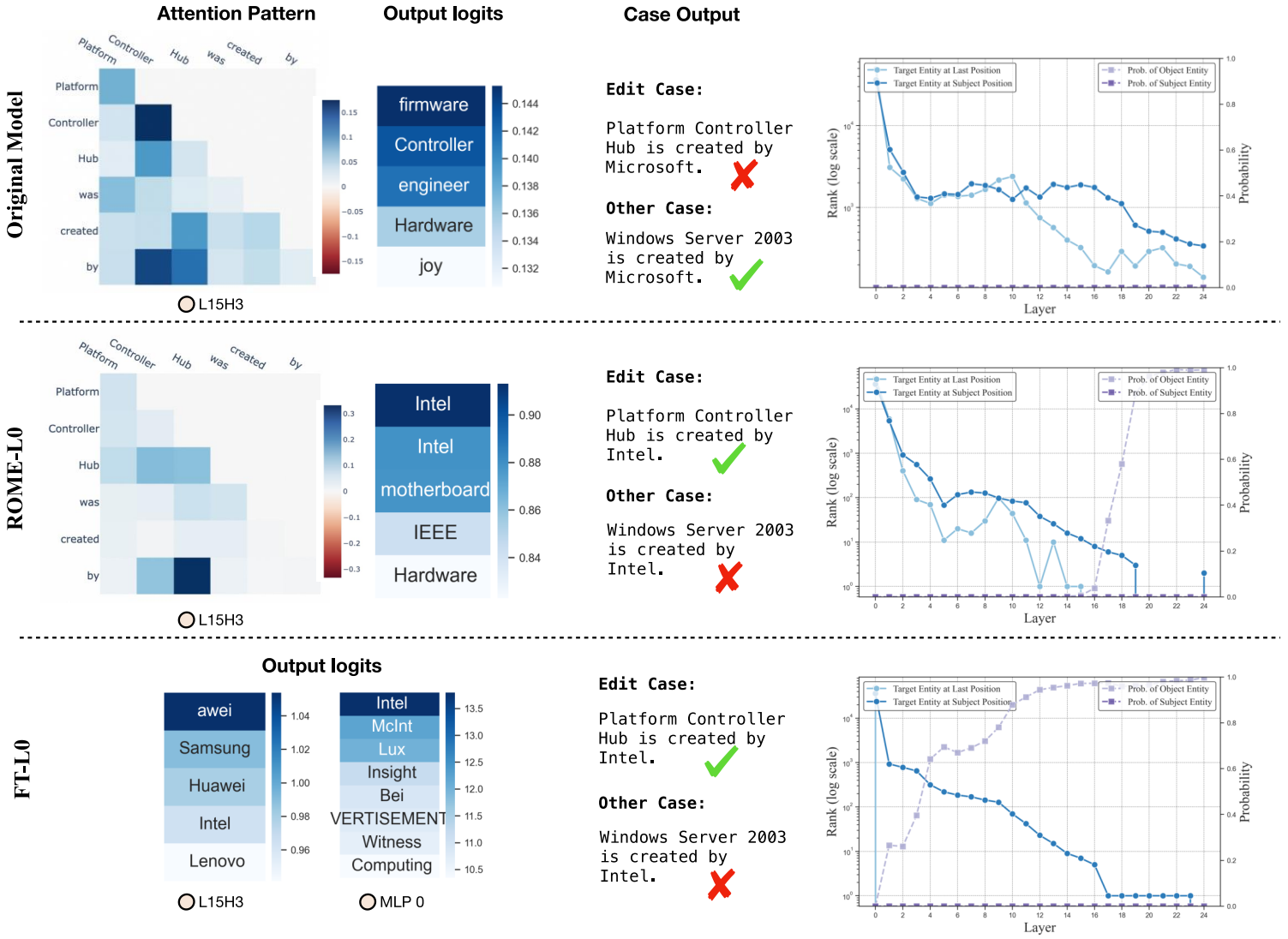




Summing Up The Facts: Additive Mechanisms Behind Factual Recall in LLMs (2024)

Internal Mechanisms for Knowledge Editing

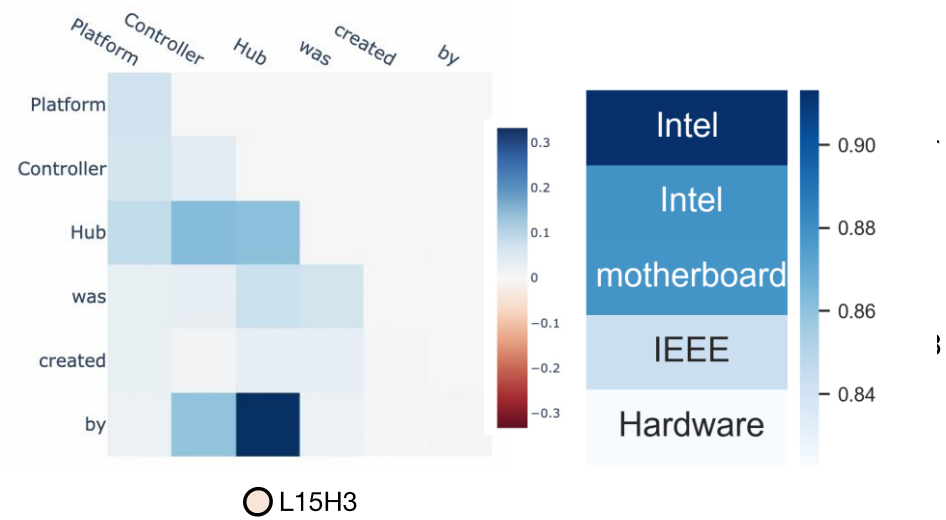
❑What happened when we edit the model?

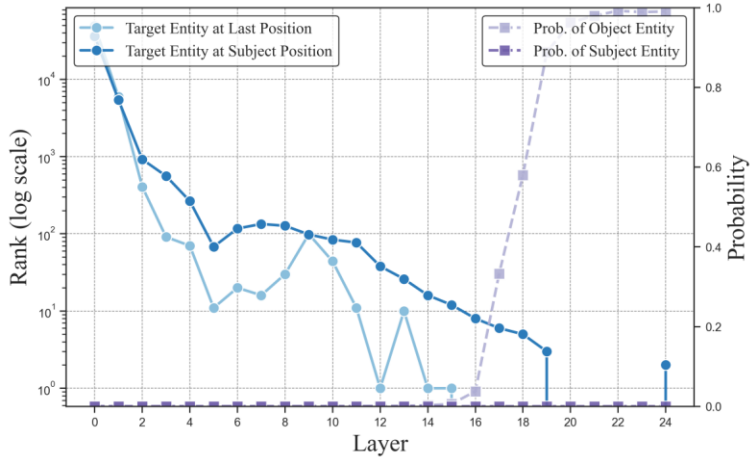


Internal Mechanisms for Knowledge Editing

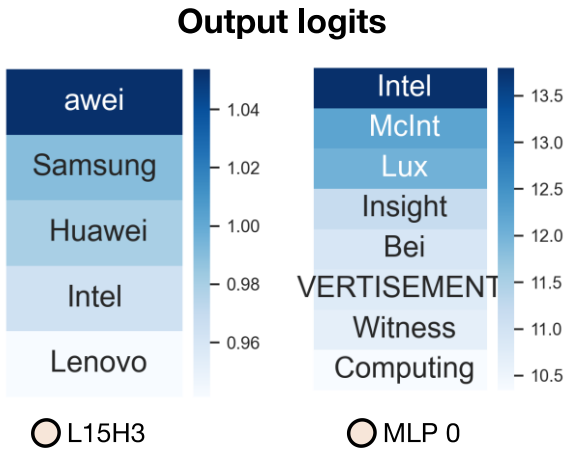
- **Finding 1**: ROME add the information at subject position so the Mover Head extract the true answer
  FT is prone to add the knowledge at the specific edit position

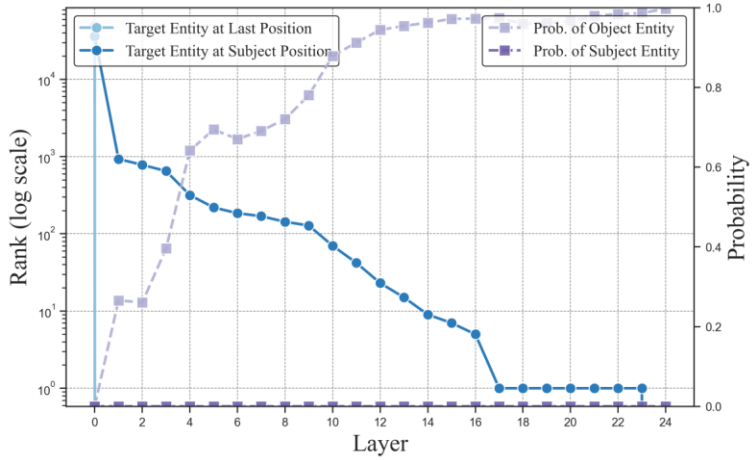
ROME-L0





FT-L0





Internal Mechanisms for Knowledge Editing

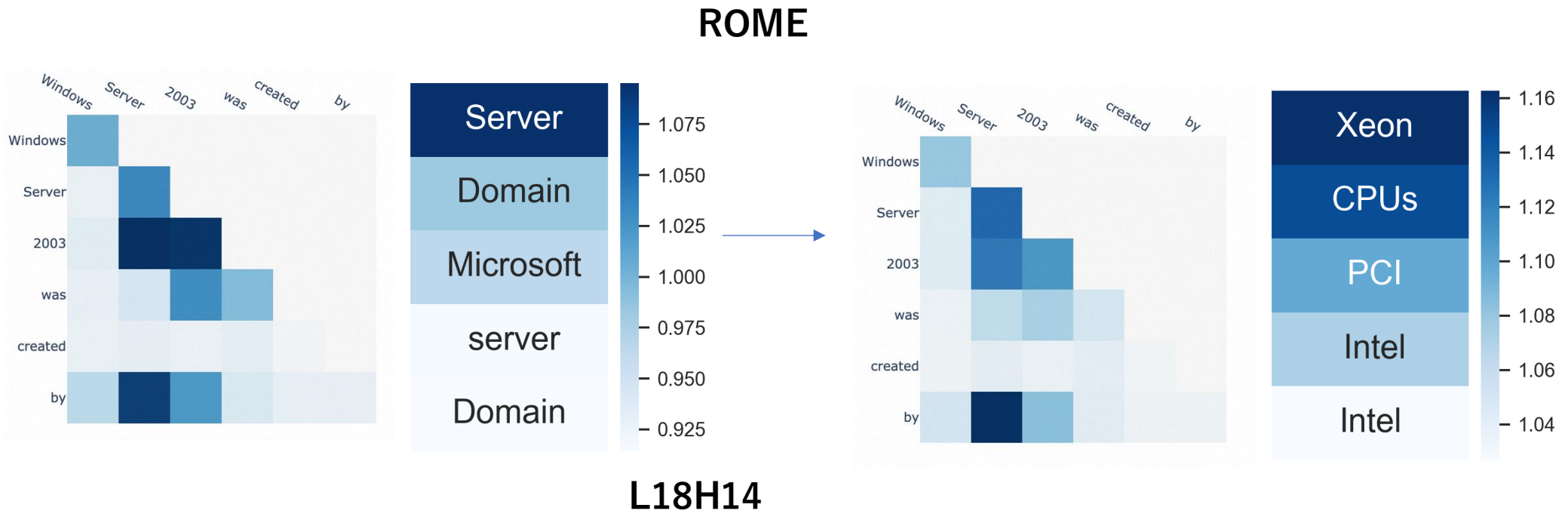
- **Finding 2:** The edited information will cause the mover head of other knowledge to select the wrong knowledge.

Edit Case:

Platform Controller Hub is created by Intel.

Other Case:

Windows Server 2003 is created by Intel.



Conclusion and Discussion

Conclusion

1. We find some circuits in the pretrained model that are responsible for the storage and expression of specific knowledge.

2. Through the circuits, we make a preliminary exploration on the internal mechanism of knowledge editing method and some behavior like hallucination and icl.

Future Discussion

1. There are still some components in the discovered circuits need to interpret.
2. How does the LM utilize the circuit for reasoning?

Contact

GitHub

Preprint

yyztodd@zju.edu.cn