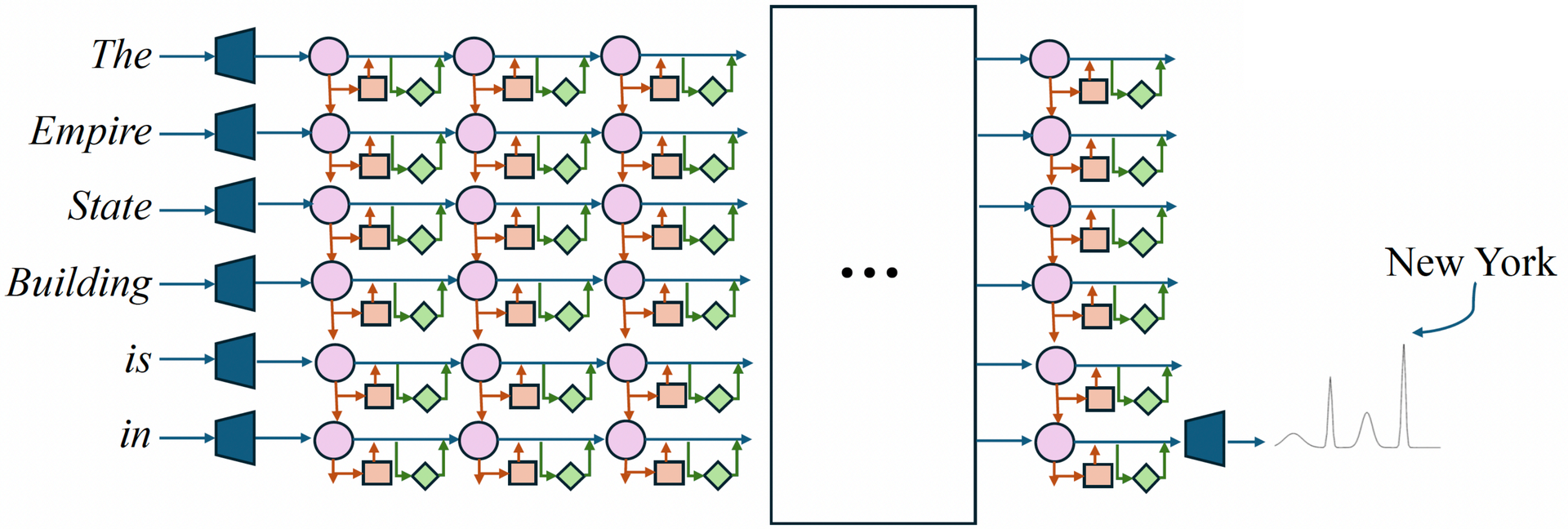
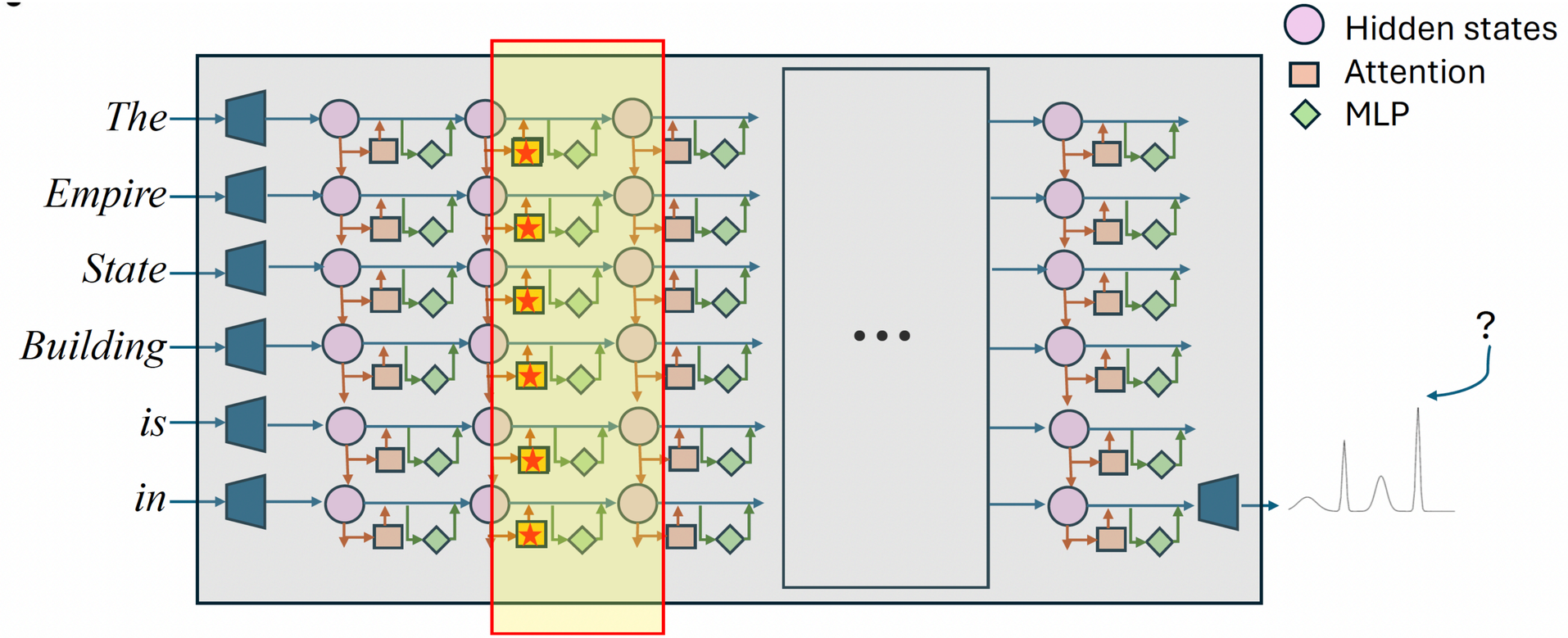
The Truth is in There: Improving Reasoning in Language Models with Layer Selective Rank Reduction

Pratyusha Sharma, (MIT), Jordan T. Ash* (MSR), Dipendra Misra* (MSR)

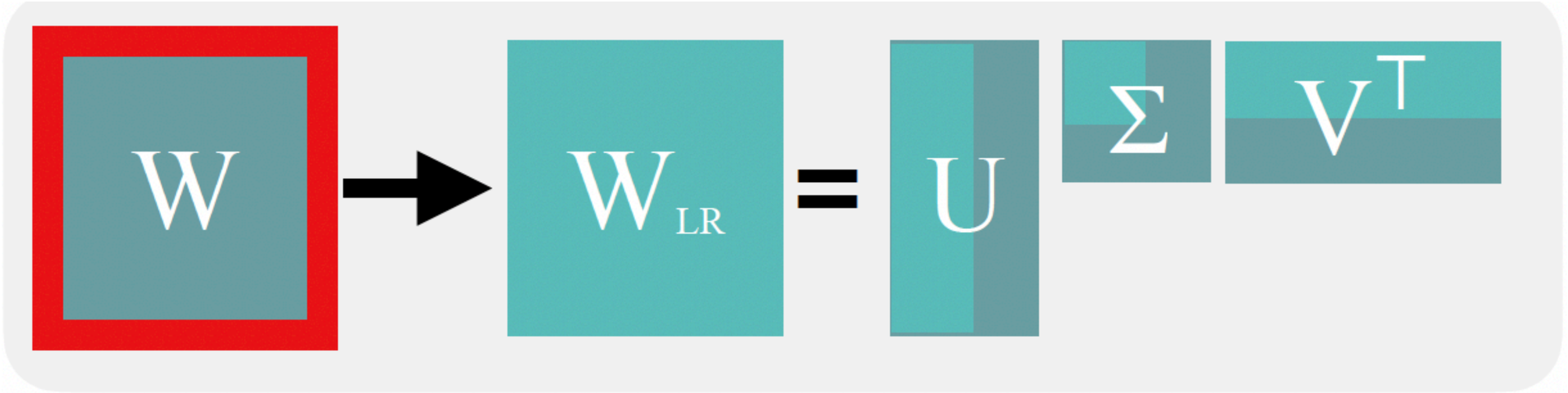
Inference with a Transformer Model



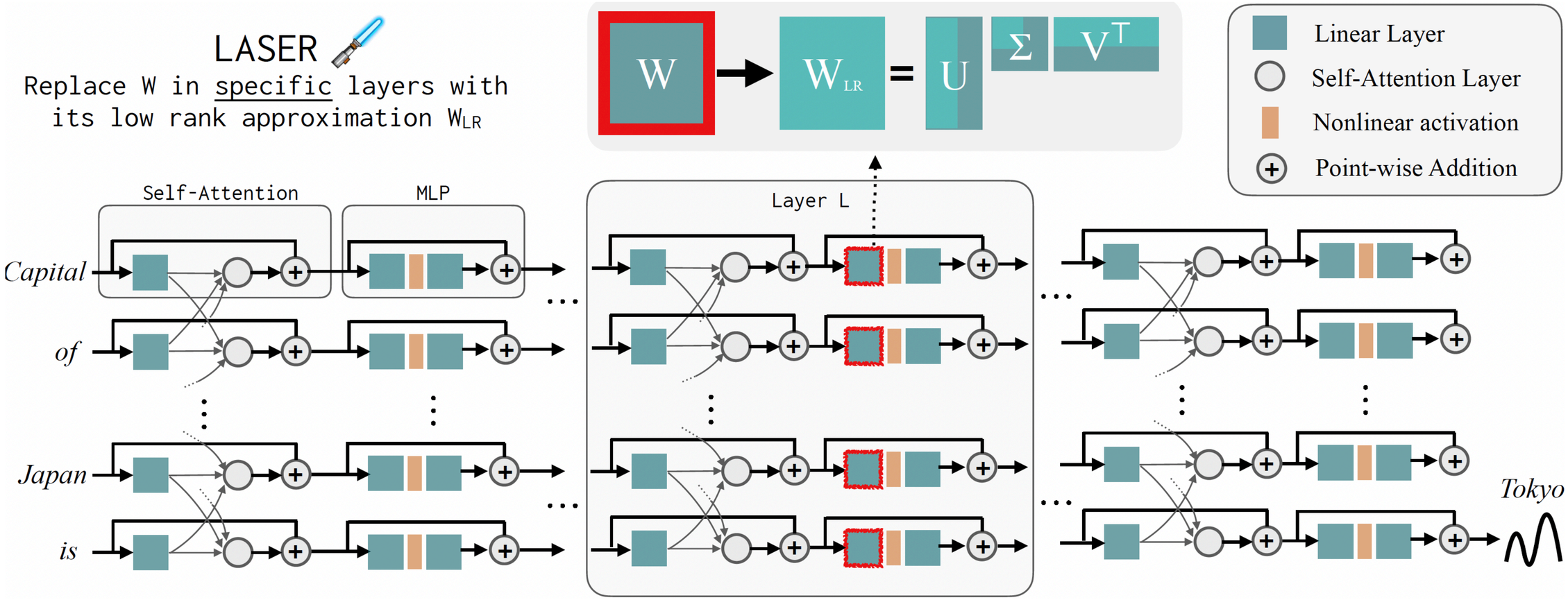
Approximate the model and observe the change in its behavior



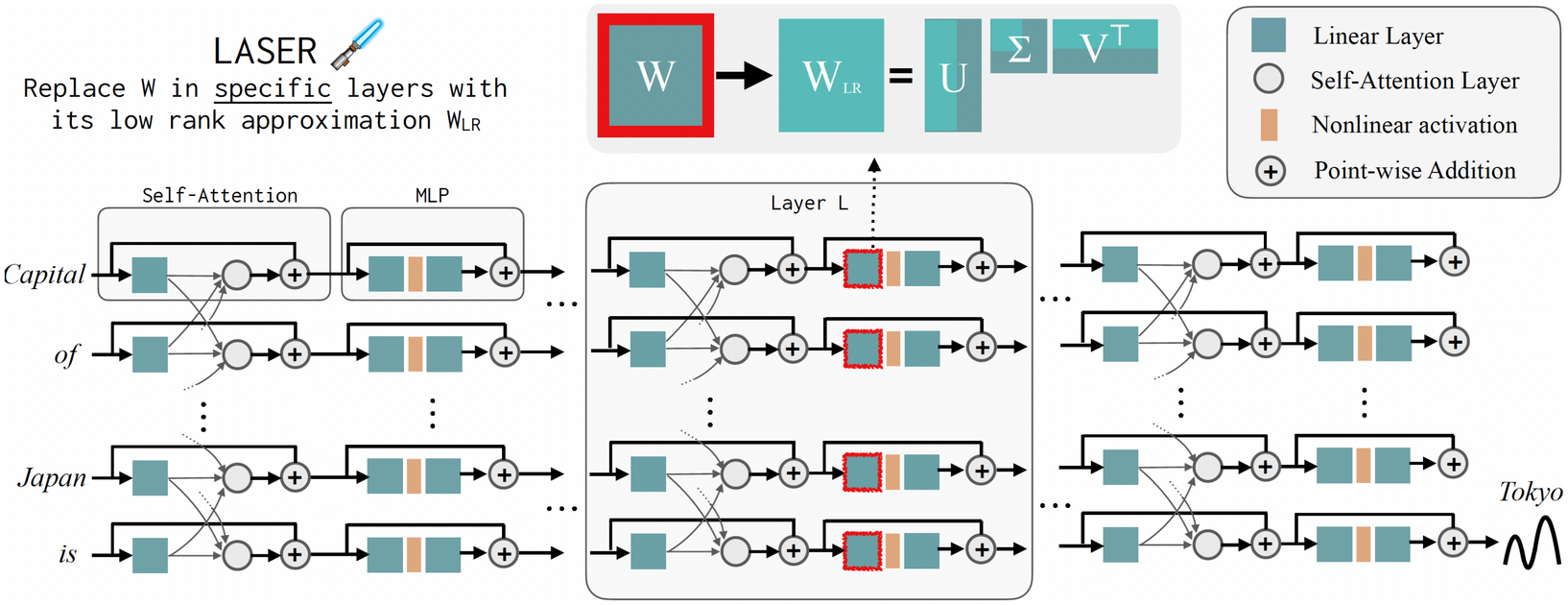
Approximating transformer weight matrices



LASER: Layer Selective Rank Reduction



LASER: Layer Selective Rank Reduction



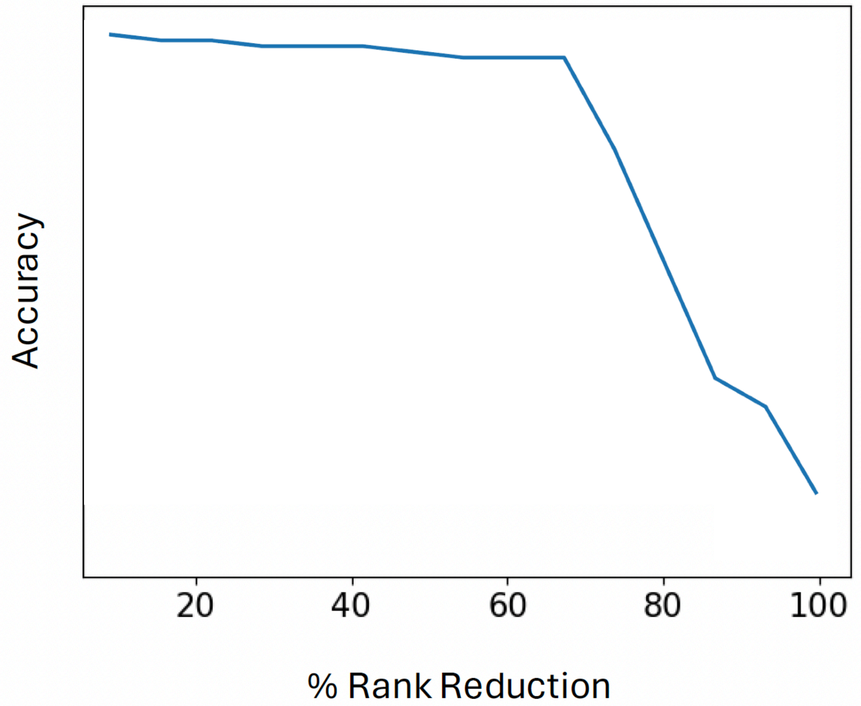
l : Layer number

τ : Matrix type

ρ : Fraction of components retained

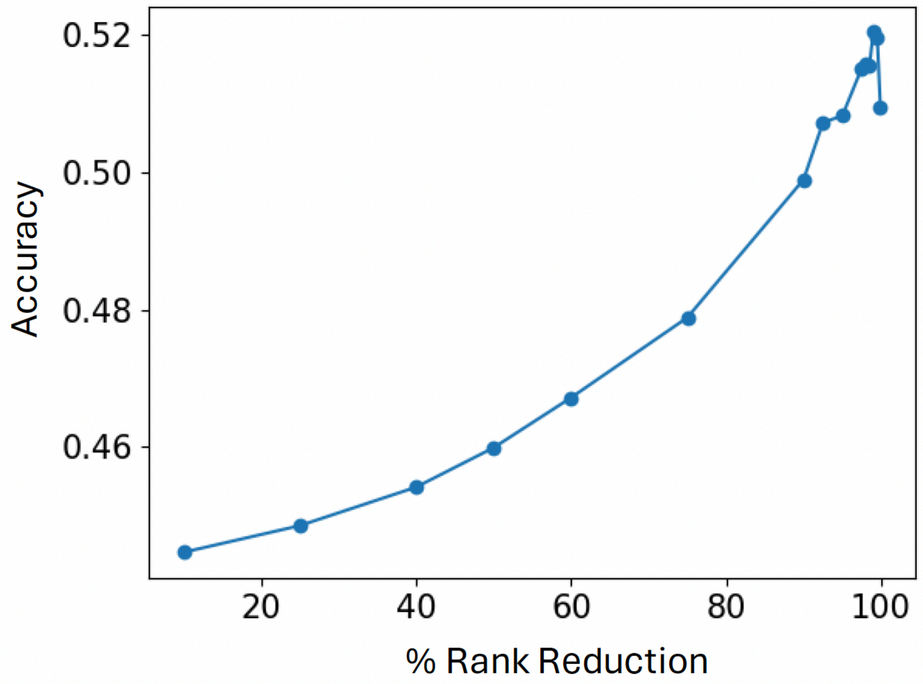
Expectation?

Expectation: Higher approx. worse accuracy



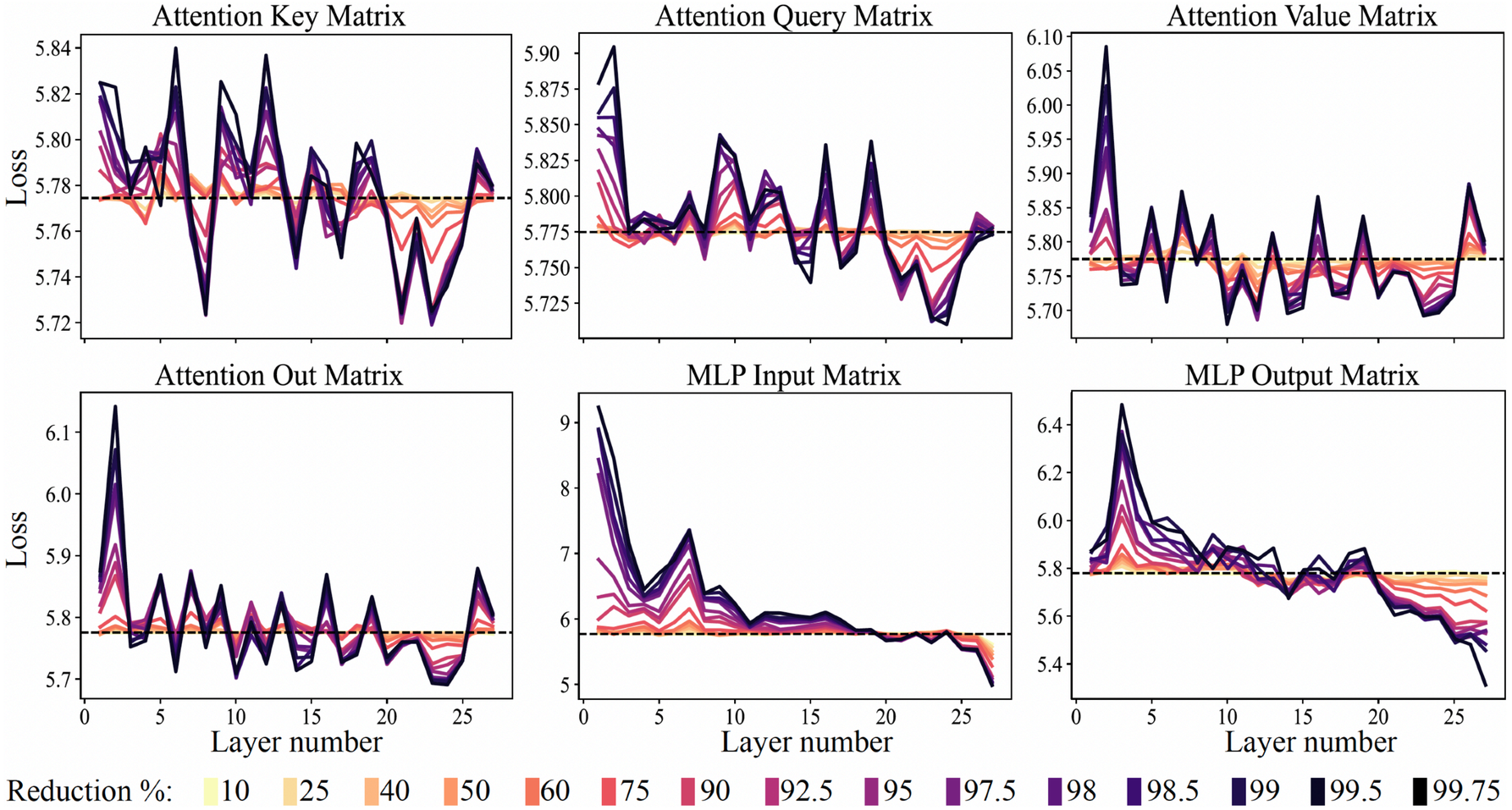
Reality:

Reality: Approximation increases accuracy



Rank-reduction in higher layers MLP weights leads to **~8%** increase in Question-Answering accuracy!

Is the effect of LASER uniform across the layers?



How widely does this hold?

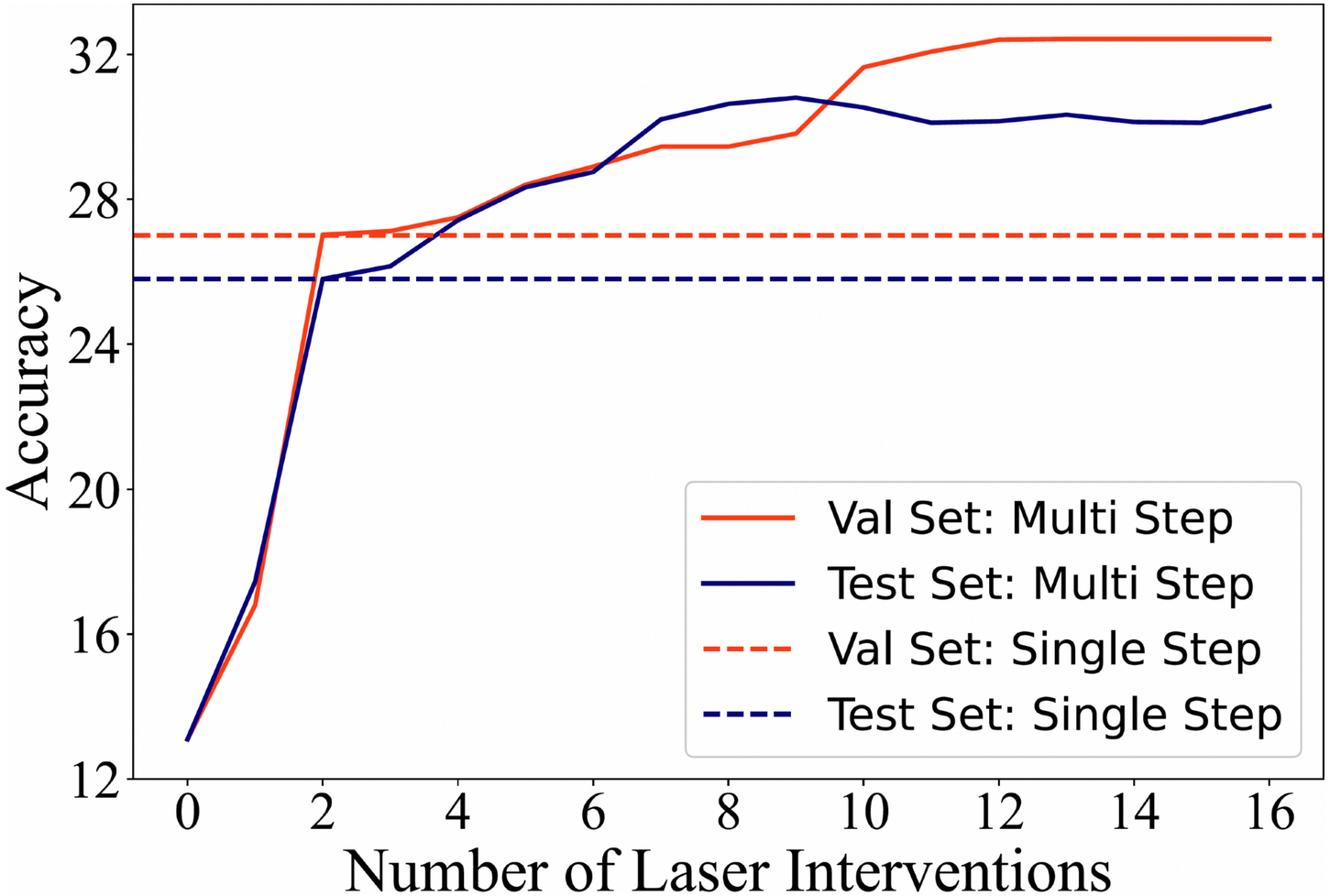
How widely does this hold?

Dataset		Model Name					
		Roberta		GPT-J		LLama2	
			LASER		LASER		LASER
CounterFact	Acc	17.3	**19.3**	13.1	**24.0**	35.6	**37.6**
	Loss	5.78	**5.43**	5.78	**5.05**	3.61	**3.49**
HotPotQA	Acc	6.1	**6.7**	**19.6**	19.5	16.5	**17.2**
	Loss	10.99	**10.53**	3.40	**3.39**	3.15	**2.97**
FEVER	Acc	50.0	**52.3**	50.2	**56.2**	59.3	**64.5**
	Loss	2.5	**1.76**	**1.24**	1.27	1.02	**0.91**
Bios Gender	Acc	87.5	**93.7**	70.9	**97.5**	75.5	**88.4**
	Loss	**0.87**	1.13	**3.86**	4.20	3.48	**2.93**
Bios Profession	Acc	64.5	**72.5**	75.6	**82.1**	85.0	**86.7**
	Loss	**4.91**	6.44	**4.64**	4.91	4.19	**4.05**
TruthfulQA	Acc	56.2	56.2	54.9	**55.6**	50.5	**56.2**
	Loss	1.60	**1.42**	1.02	**1.01**	**0.95**	1.04
BigBench-Epistemic Reasoning	Acc	37.1	**41.8**	37.1	**38.3**	44.8	**63.4**
	Loss	9.39	**6.80**	0.74	**0.62**	0.78	**0.73**
BigBench-WikidataQA	Acc	28.0	**30.7**	51.8	**65.9**	59.5	**62.0**
	Loss	9.07	**7.69**	3.52	**2.86**	2.40	**2.31**

How widely does this hold?

Dataset		Model Name					
		Roberta		GPT-J		LLama2	
			LASER		LASER		LASER
CounterFact	Acc	17.3	**19.3**	13.1	**24.0**	35.6	**37.6**
	Loss	5.78	**5.43**	5.78	**5.05**	3.61	**3.49**
HotPotQA	Acc	6.1	**6.7**	**19.6**	19.5	16.5	**17.2**
	Loss	10.99	**10.53**	3.40	**3.39**	3.15	**2.97**
FEVER	Acc	50.0	**52.3**	50.2	**56.2**	59.3	**64.5**
	Loss	2.5	**1.76**	**1.24**	1.27	1.02	**0.91**
Bios Gender	Acc	87.5	**93.7**	70.9	**97.5**	75.5	**88.4**
	Loss	**0.87**	1.13	**3.86**	4.20	3.48	**2.93**
Bios Profession	Acc	64.5	**72.5**	75.6	**82.1**	85.0	**86.7**
	Loss	**4.91**	6.44	**4.64**	4.91	4.19	**4.05**
TruthfulQA	Acc	56.2	56.2	54.9	**55.6**	50.5	**56.2**
	Loss	1.60	**1.42**	1.02	**1.01**	**0.95**	1.04
BigBench-Epistemic Reasoning	Acc	37.1	**41.8**	37.1	**38.3**	44.8	**63.4**
	Loss	9.39	**6.80**	0.74	**0.62**	0.78	**0.73**
BigBench-WikidataQA	Acc	28.0	**30.7**	51.8	**65.9**	59.5	**62.0**
	Loss	9.07	**7.69**	3.52	**2.86**	2.40	**2.31**

LASER interventions can be composed across layers



How much can you compress these models?

Dataset	**Model**		
	Roberta $[\tau, \ell, \rho]$	**GPT-J** $[\tau, \ell, \rho]$	**Llama2 7B** $[\tau, \ell, \rho]$
CounterFact	$[U_{in}, 8, 0.8]$	$[U_{in}, 27, 0.01]$	$[U_{in}, 28, 0.05]$
HotPotQA	$[U_{out}, 2, 0.4]$	$[U_{in}, 27, 0.1]$	$[U_{in}, 27, 0.2]$
FEVER	$[U_{in}, 3, 0.4]$	$[U_{in}, 24, 0.01]$	$[U_{in}, 30, 0.2]$
Bios Gender	$[U_{in}, 9, 0.9]$	$[U_{in}, 14, 0.01]$	$[U_{in}, 24, 0.01]$
Bios Prof.	$[U_{in}, 3, 0.9]$	$[U_{in}, 18, 0.01]$	$[U_{out}, 30, 0.4]$
BigBench-Epistemic Reasoning	$[U_{out}, 1, 0.4]$	$[U_{in}, 26, 0.01]$	$[U_{out}, 28, 0.01]$
TruthfulQA	$[U_{in}, 0, 0.01]$	$[U_{in}, 7, 0.8]$	$[U_{in}, 30, 0.05]$
BigBench-WikidataQA	$[U_{in}, 7, 0.4]$	$[U_{in}, 27, 0.01]$	$[U_{in}, 27, 0.01]$

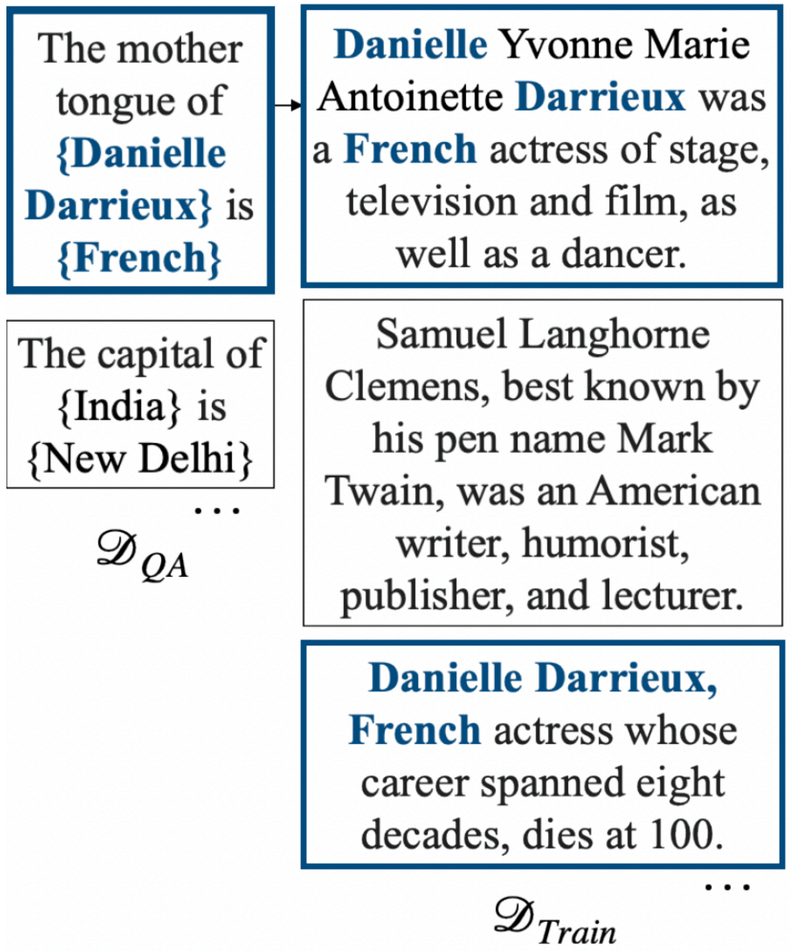
τ : Layer type

l : Layer number

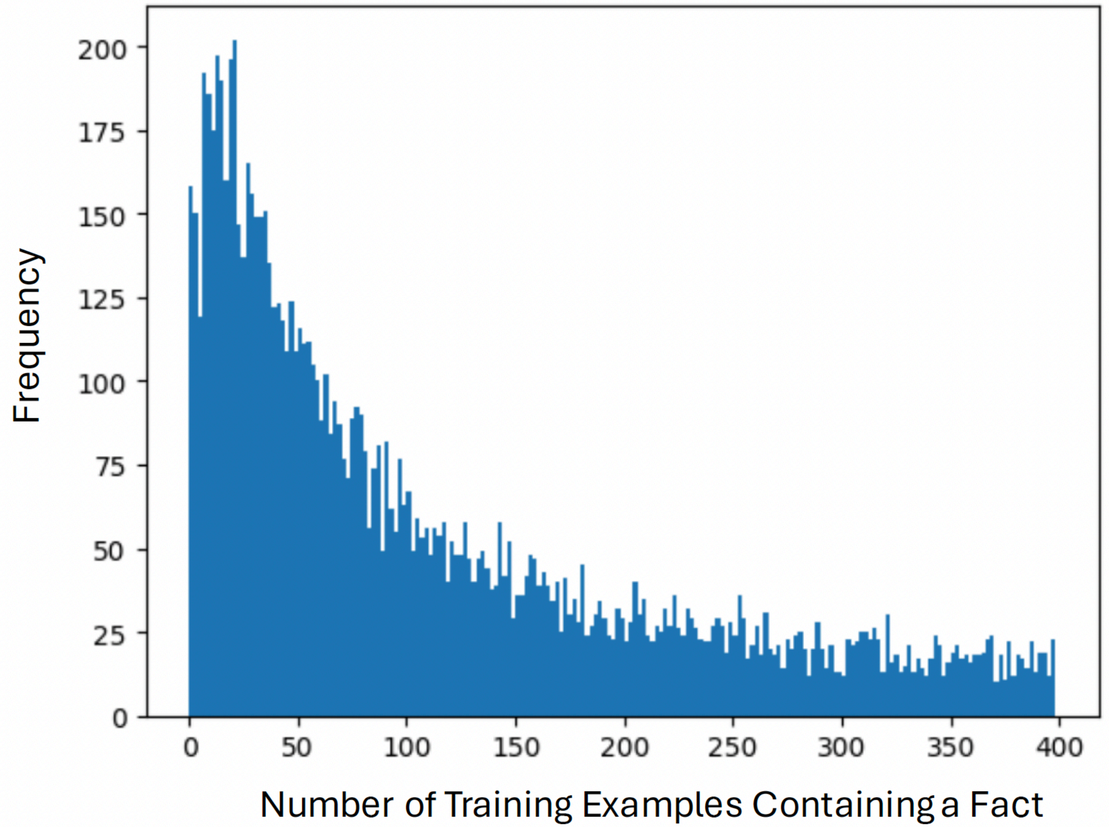
ρ : Fraction of components retained

What can be said about the datapoints where this flipping is observed?

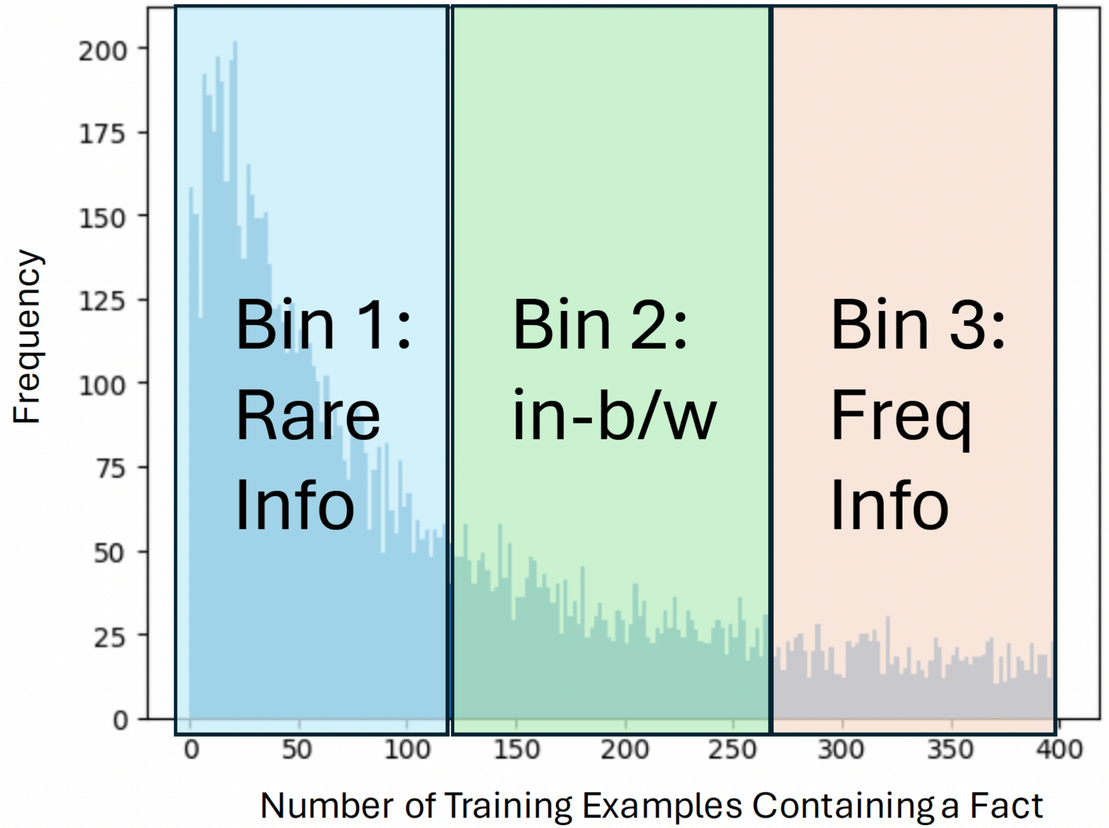
What can be said about the datapoints where this flipping is observed?



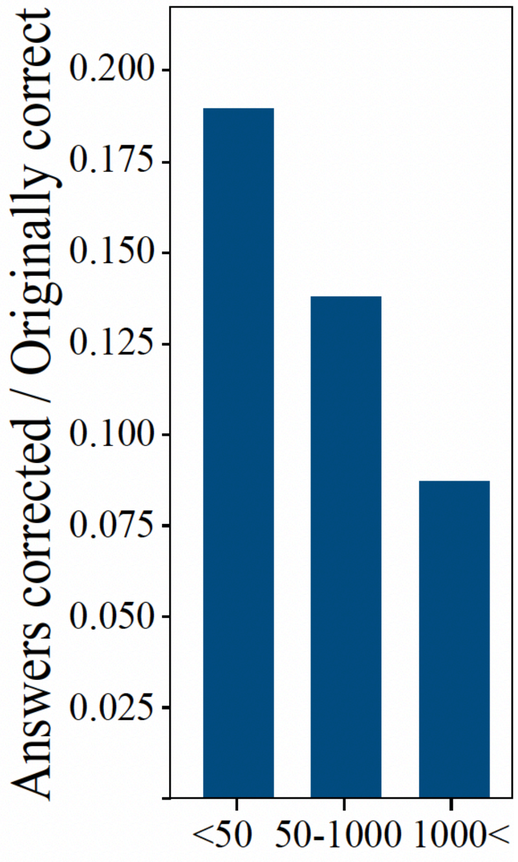
What can be said about the datapoints where this flipping is observed?



What can be said about the datapoints where this flipping is observed?



What can be said about the datapoints where this flipping is observed?



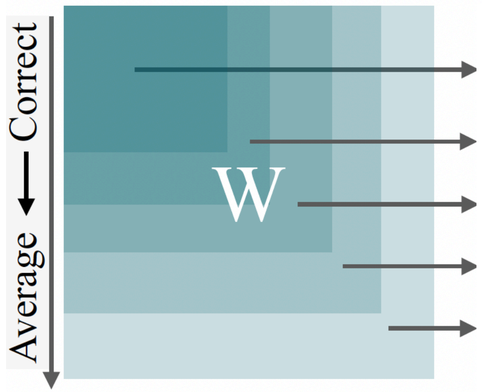
N = Frequency of $\mathscr{D}_{QA}$ in $\mathscr{D}_{Train}$

Hypothesis: Denoising

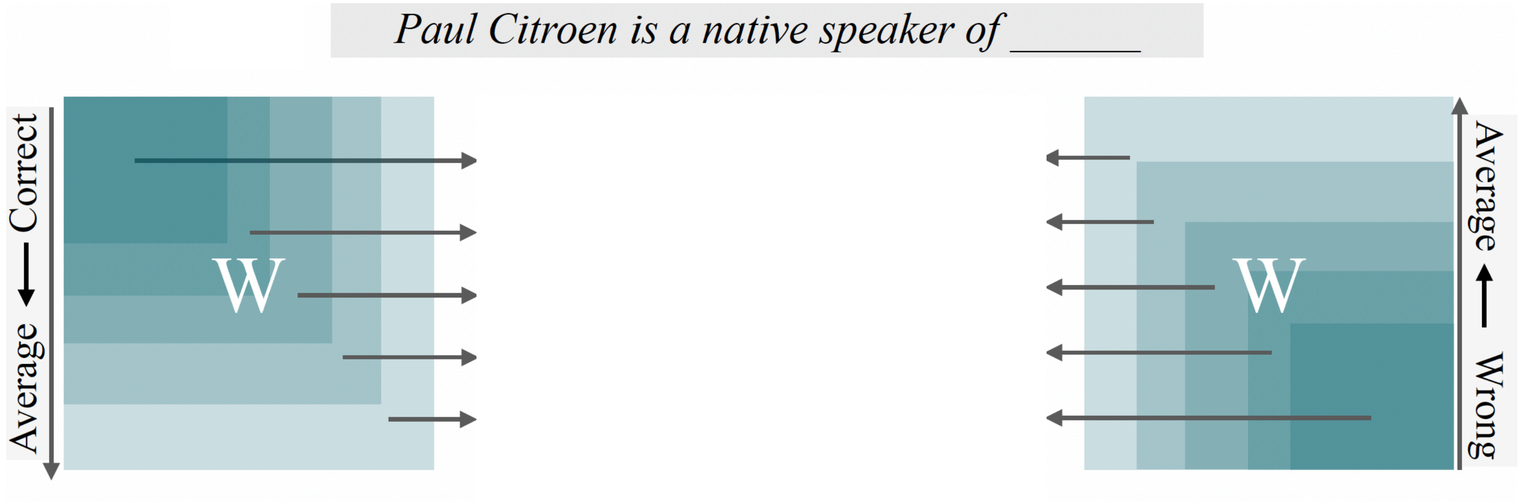
“Pruning the higher ordered components-> allows for better inductive reasoning over the lower ordered stable components.”

What do higher-order components in the matrix store that disrupt the model?

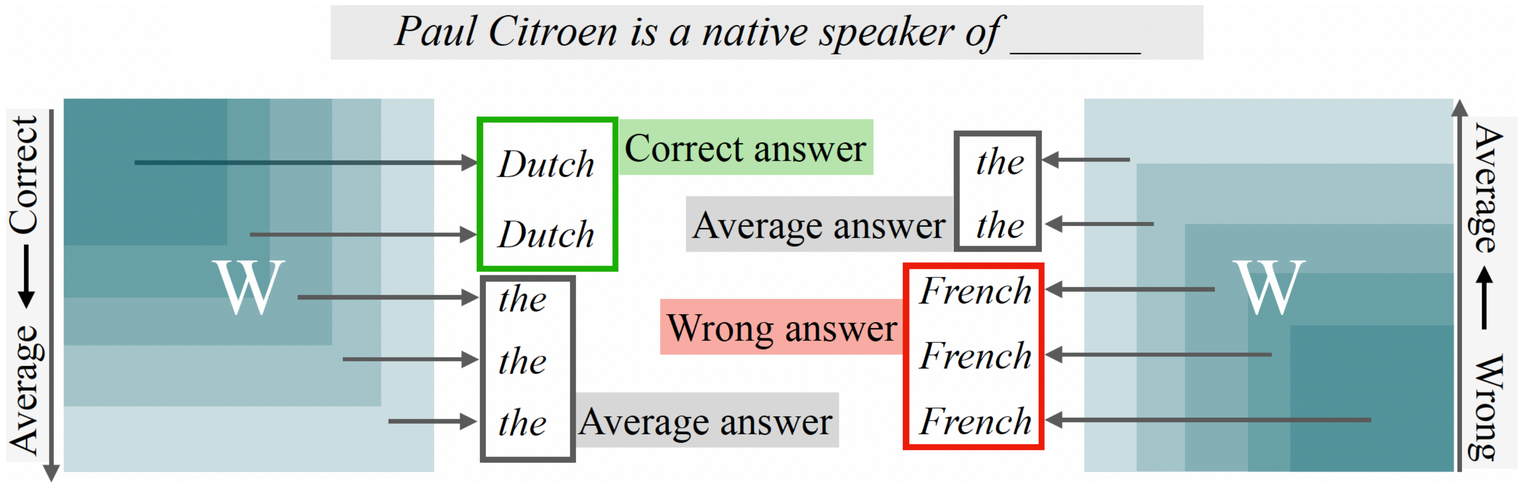
Paul Citroen is a native speaker of _____



What do higher-order components in the matrix store that disrupt the model?



What do higher-order components in the matrix store that disrupt the model?



What do higher-order components in the matrix store that disrupt the model?

	Top	Bottom
Saeed Akhtar Mirza is originally from	*Mumbai*	*Pakistan*
The original language of Hussar Ballad is	*Russian*	*Portuguese*
Kalabhra follows the religion of	*Buddhism*	*Hindu*
Emmanuelle Devos's profession is a	*Actor*	*Teacher*
Walter Zenga is a professional	*Soccer*	*Photographer*
Mike Holmgren plays in the position of	*Quarterback*	*Goalkeeper*

	Top	Bottom
The twin city of Wellington is	*Sydney*	*Of*
Kharkiv is a twin city of	*Warsaw*	*The*
The native language of Isaac Massa is	*Dutch*	*The*
The headquarter of Morr Music is located in	*Berlin*	*the*
Abba Eban was employed in	*Jerusalem*	*The*
Yizhar Harari speaks	*Hebrew*	*To*

Conclusion

- We present LASER: Layer Selective Rank Reduction

- It is possible to simultaneously enhance an LLM's task performance and reduce its size

- Selective low-rank reduction of matrices in a transformer can improve its performance on language understanding tasks, at times by 30% pts!

Lots of open questions remain!