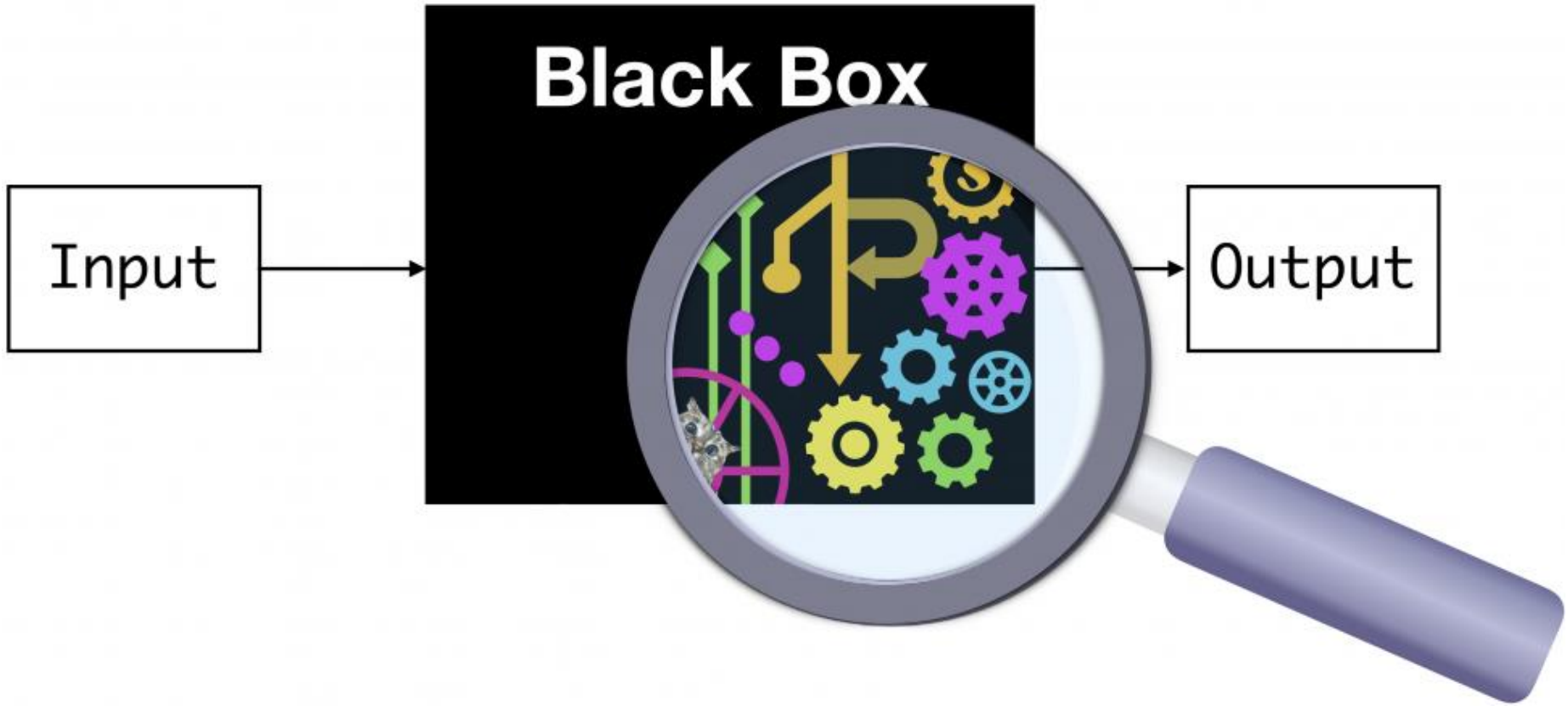
Interpretability: What Comes Next?

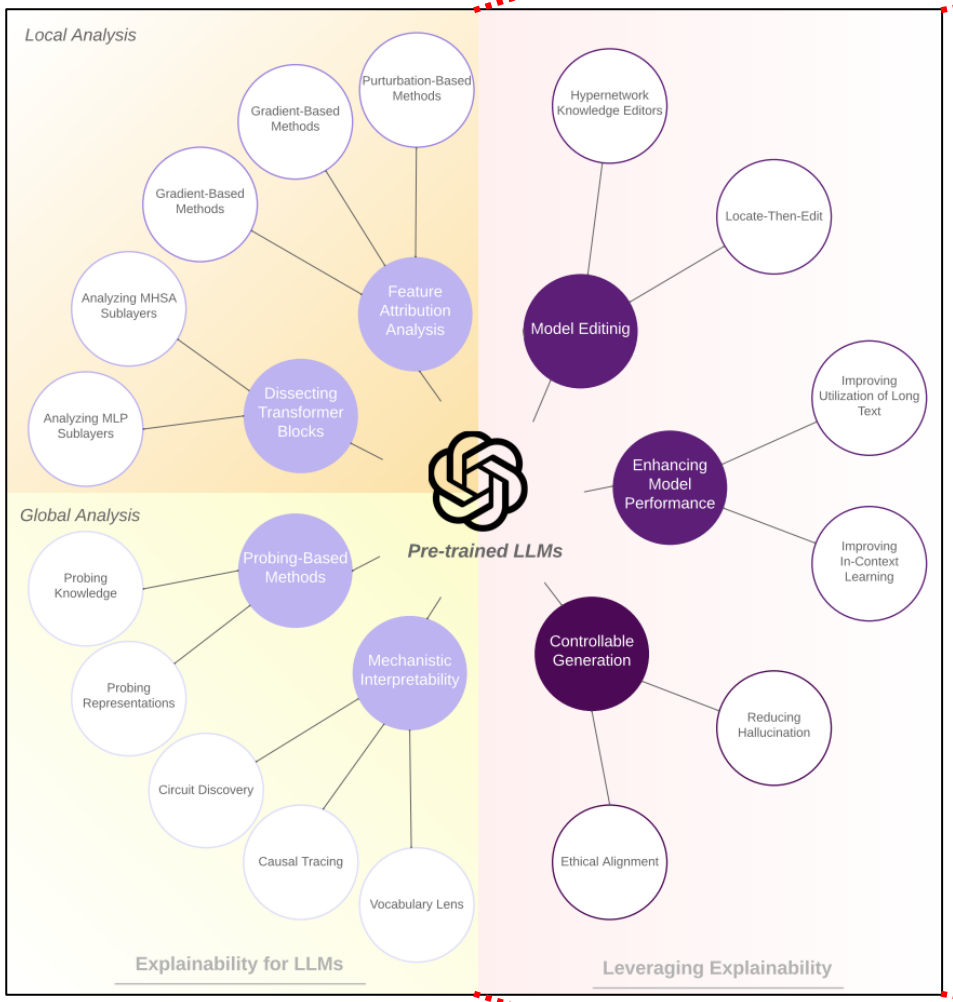
발표자: 김민혁

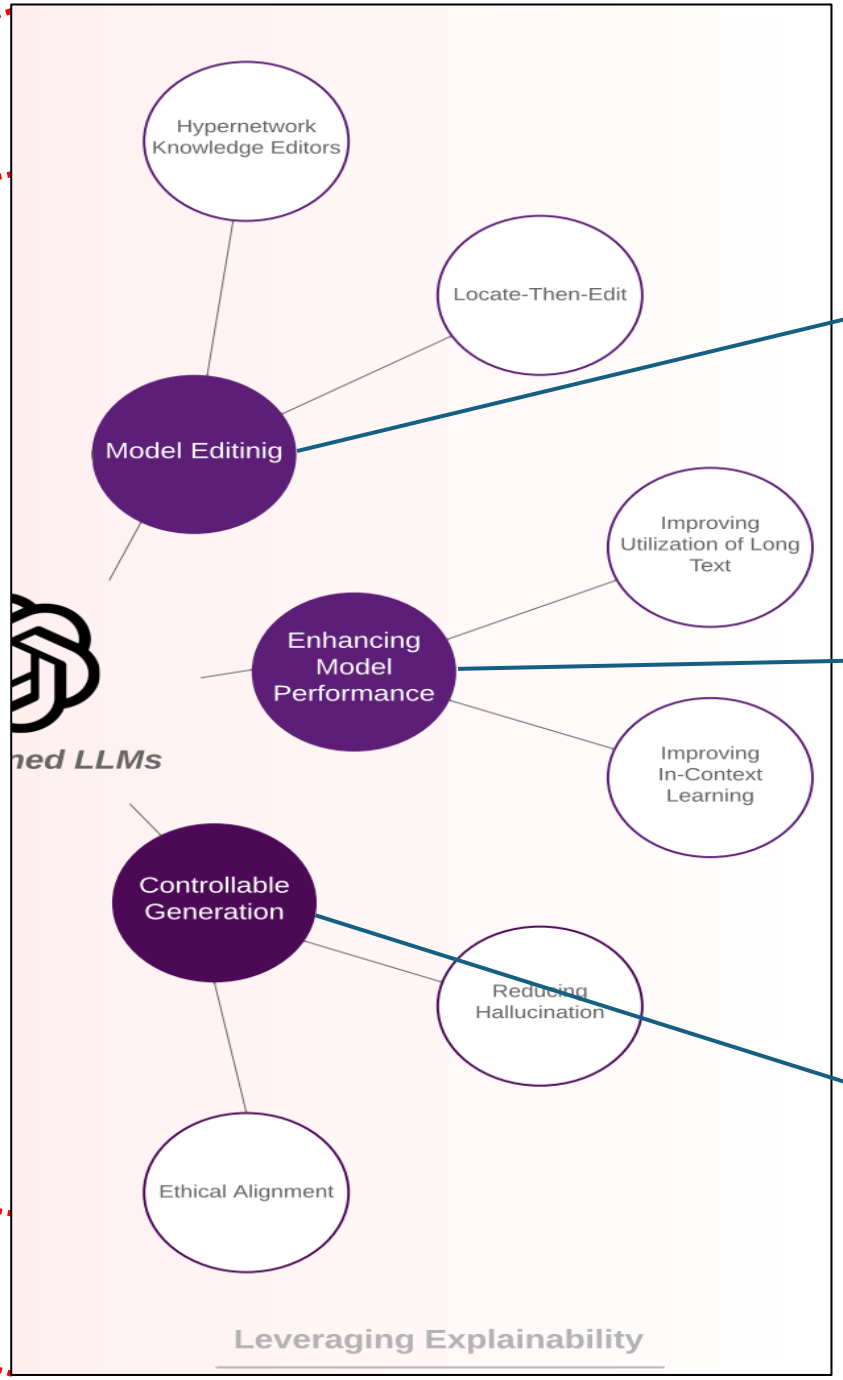
고려대학교
KOREA UNIVERSITY

Interpretability: What Comes Next?



Interpretability: What Comes Next?





Model Editing

Enhancing Model Performance

Controllable Generation

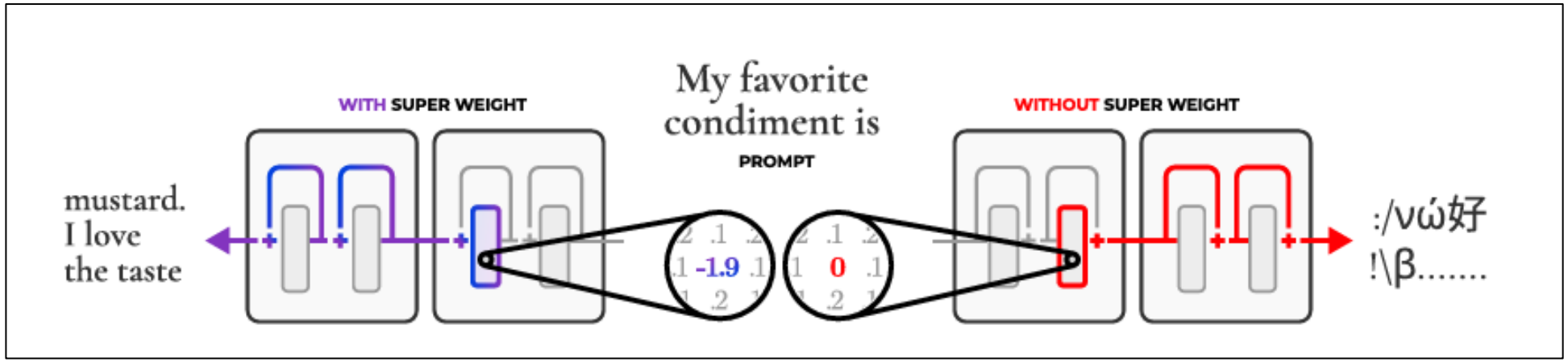
THE SUPER WEIGHT IN LARGE LANGUAGE MODELS

Mengxia Yu[1]*, **De Wang**[2], **Qi Shan**[2], **Colorado Reed**[2]†, **Alvin Wan**[2]
[1]University of Notre Dame [2]Apple

ICLR 2025 Under Review

THE SUPER WEIGHT IN LLMs

- Contributions

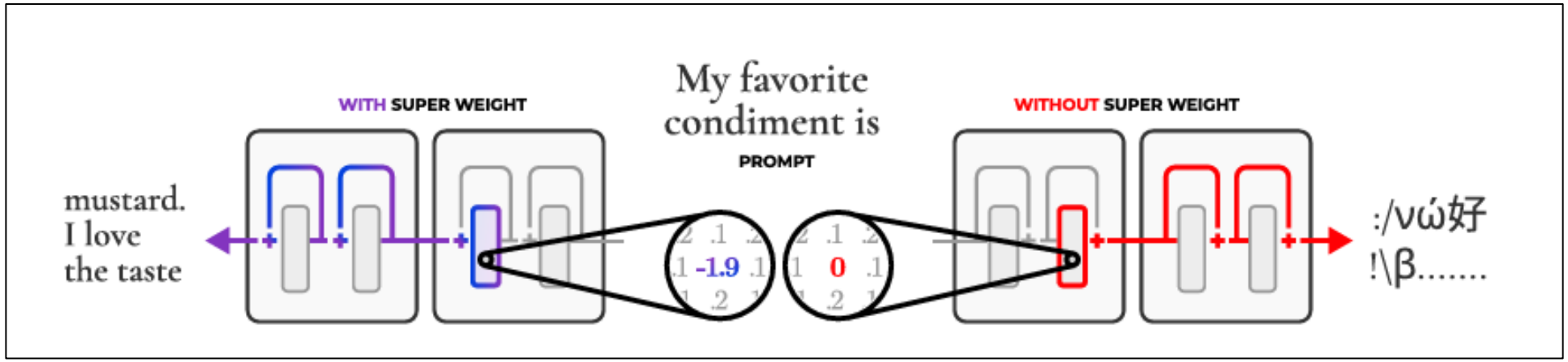


<그림 1> 몇 개에 불과한 Super Weight를 제거하면 언어적 능력을 상실

- Super Weights
  - LLM에서 매우 작은 일부 가중치가 모델 성능에 큰 영향을 미침.
  - 해당 가중치를 제거하면 모델 품질이 크게 저하됨.
- Identifying Super Weights
  - 데이터 없이 단 한 번의 Forward Pass로 슈퍼 가중치를 식별 가능.

THE SUPER WEIGHT IN LLMs

- Contributions

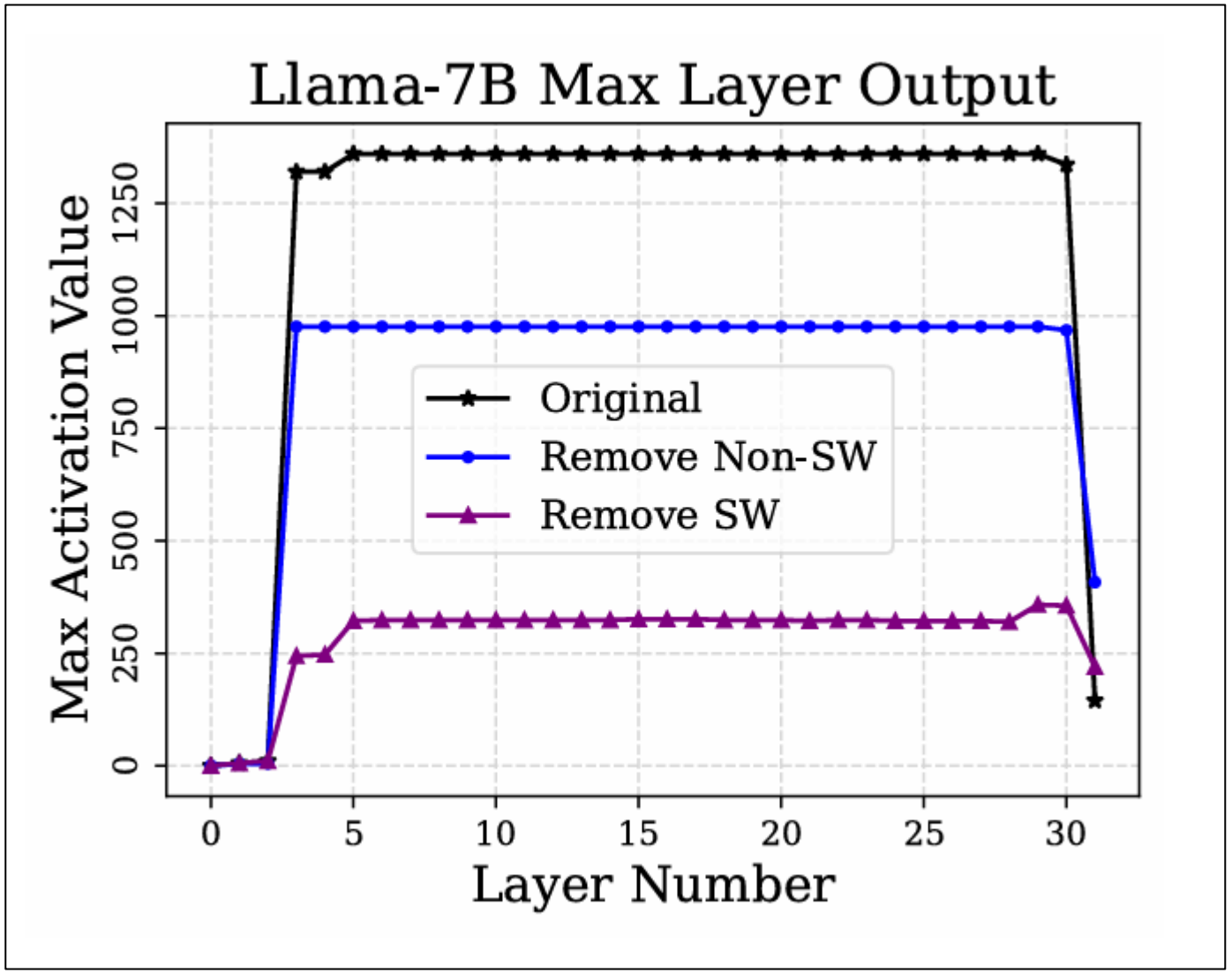


<그림 1> 몇 개에 불과한 Super Weight를 제거하면 언어적 능력을 상실.

- Super Activations
  - 추론 시 슈퍼 가중치에 의해 큰 활성화 값이 생성됨.
- Compression
  - Super Weight를 Clip & Restoration을 통해 양자화 성능을 높임.

THE SUPER WEIGHT IN LLMs

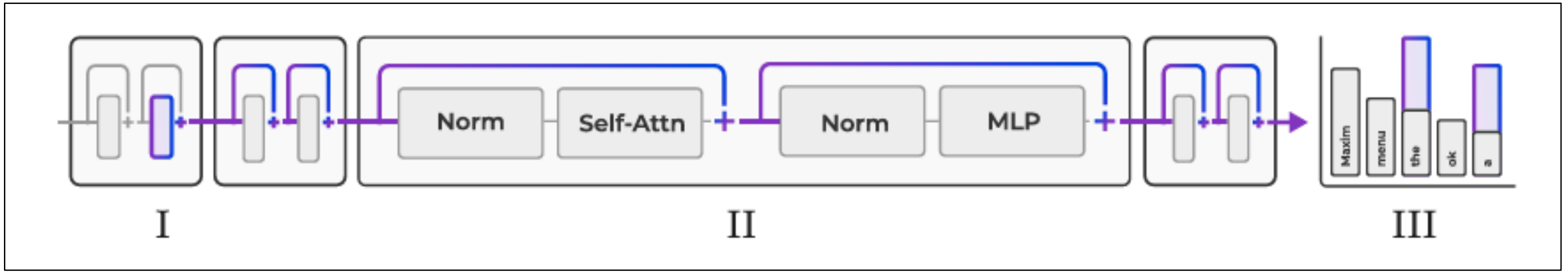
- Super Weights
  - Identification
    - Input에 관계 없이 항상 동일한 위치에 존재함. (Instruction Tuning시에도 위치는 동일함)
    - Massive Activation은 Super Weight 다음에 나타남.
    - Up Projection에서의 곱이 큰 Activation 값을 생성함.



<그래프 1> Super Weight (SW) 를 제거하면 Activation 값이 급격하게 감소함.

THE SUPER WEIGHT IN LLMs

- Mechanisms of Super Weights

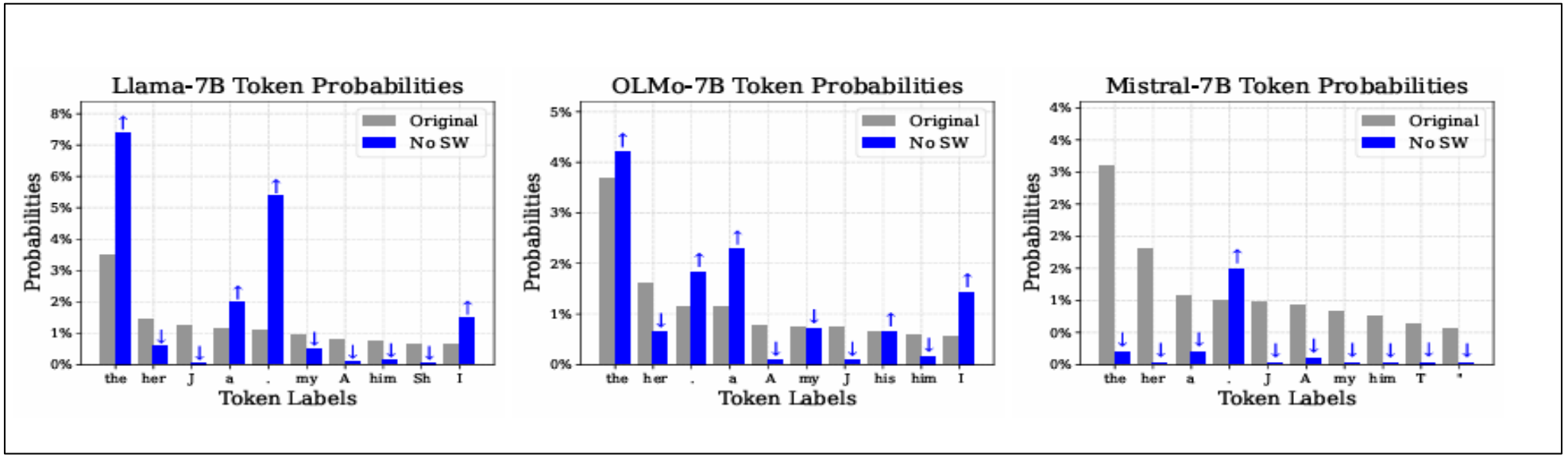


<그림 2> How Super Weights Behave

- Super Activation을 유도함.
  - 이때 Super Activation은 Skip Connection에 의해 모델에 전반적으로 영향을 끼침.

THE SUPER WEIGHT IN LLMs

- Mechanisms of Super Weights

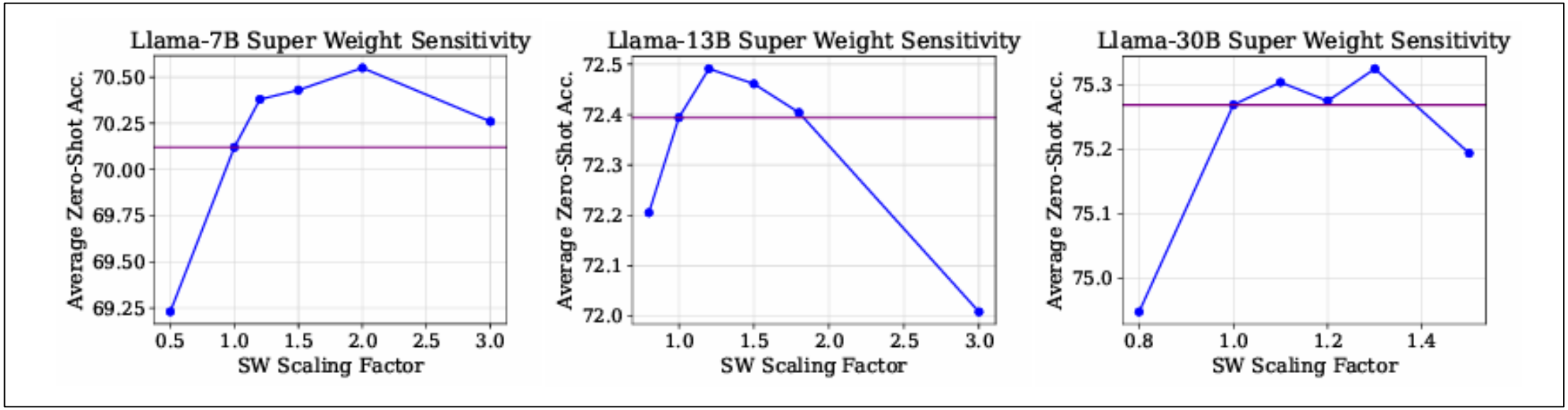


<그래프 2> Super Weights Suppress Stopwords

- Super Weight는 Stopword의 Likelihood를 억제함.
  - Super Weights를 제거할 경우 Stopword의 확률이 2~5배 증가함
  - 반면 Non-stopword의 확률이 2~3배로 감소함.

THE SUPER WEIGHT IN LLMs

- Mechanisms of Super Weights



<그래프 3> Super Weight (SW) Amplifying에 따른 성능 추이

- Sensitivity of Super Weight
  - 특정 범위 내에서의 Super Weight의 Amplifying은 Zero-shot 성능을 향상 시켰음.

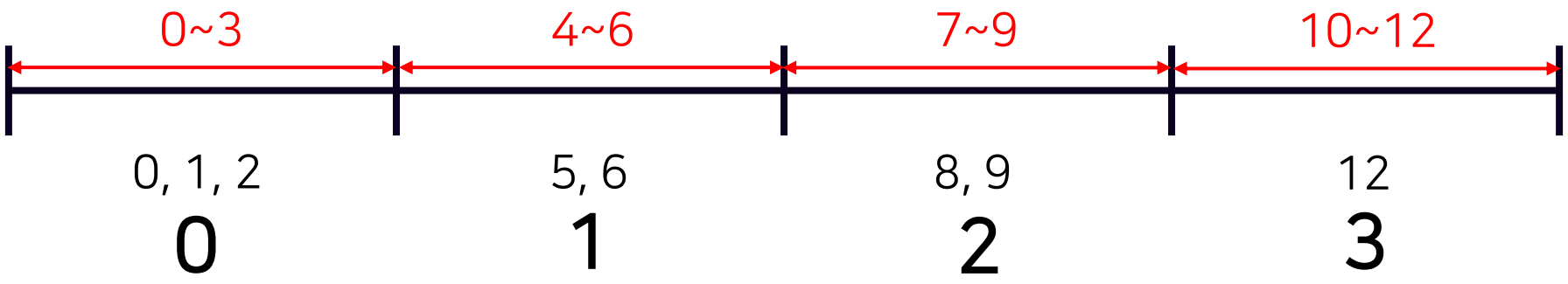
- Super-outlier Aware Quantization
  - Round-to-nearest Quantization
    - 간단한 방법의 Quantization으로, Outlier에 영향을 크게 받음.

$$Q(\mathbf{X}) = \text{Round}\left(\frac{\mathbf{X} - \text{MIN}(\mathbf{X})}{\Delta}\right), Q^{-1}(\hat{\mathbf{X}}) = \Delta \cdot \hat{\mathbf{X}} + \text{MIN}(\mathbf{X})$$

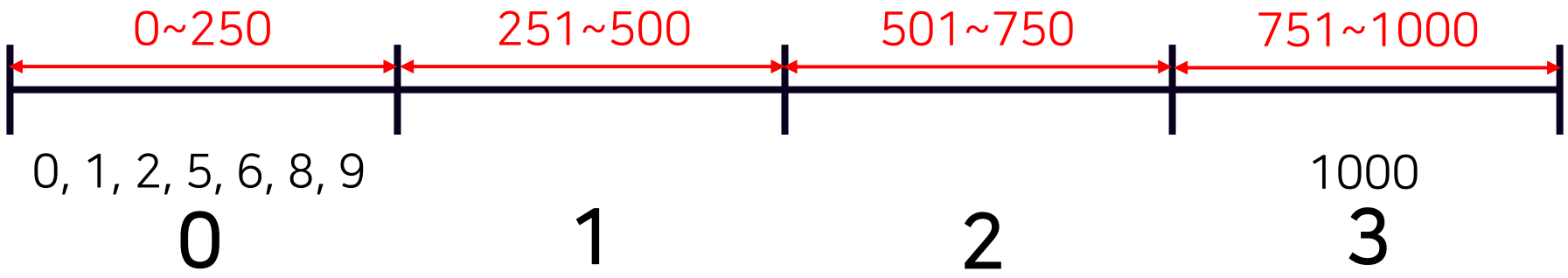
$$\Delta = \frac{\text{MAX}(\mathbf{X}) - \text{MIN}(\mathbf{X})}{2^{N-1} - 1}$$

THE SUPER WEIGHT IN LLMs

- Quantization에서 Outlier을 잡는 것은 왜 중요할까?
  - Case A. [0, 1, 2, 5, 6, 8, 9, 12] → [0, 0, 0, 1, 1, 2, 2, 3]

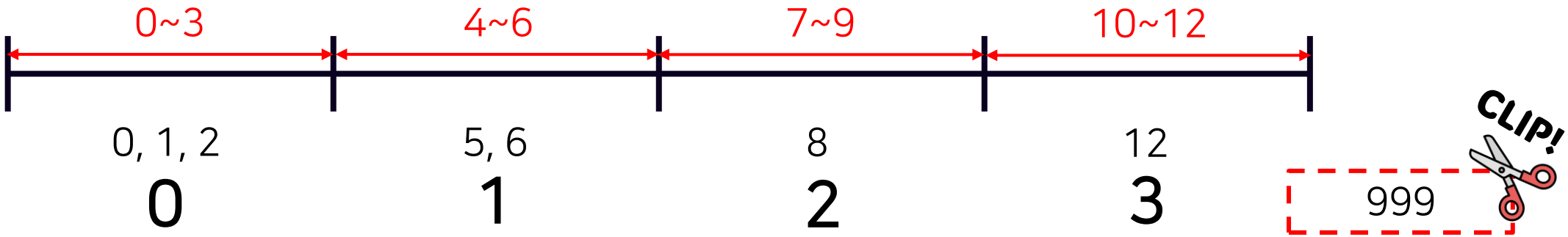


  - Case B. [0, 1, 2, 5, 6, 8, 9, 1000] → [0, 0, 0, 0, 0, 0, 0, 3]



- Super-outlier Aware Quantization
  - Quantization with Clipping & Restoration
    - Outlier 값을 보관해 두었다가, Quantization 복원 시 해당 값 복구

$$\hat{W} = \text{RESTORE}(Q^{-1}(Q(\text{CLIP}_z(W))))$$



[0, 1, 2, 5, 6, 8, 12, 999] → [0, 0, 0, 1, 1, 2, 3, 999]

THE SUPER WEIGHT IN LLMs

- Experiments
  - Super Weights

Llama-7B	Arc-c	Arc-e	Hella.	Lamb.	PIQA	SciQ	Wino.	AVG	C4	Wiki-2
Original	41.81	75.29	56.93	73.51	78.67	94.60	70.01	70.11	7.08	5.67
Prune SW	19.80	39.60	30.68	0.52	59.90	39.40	56.12	35.14	763.65	1211.11
Prune Non-SW	41.47	74.83	56.35	69.88	78.51	94.40	69.14	69.22	7.57	6.08
Prune SW, +SA	26.60	54.63	56.93	12.79	67.95	61.70	70.01	50.09	476.23	720.57

<표 1> SW 제거에 따른 벤치마크 성능 추이

- Prune SW - Super Weight 제거 / Prune Non-SW – Super Weight이외의 가중치 제거. (Magnitude 기준 상위 7,000개) / Prune SW, +SA (Super Activation 복원)
- Super Weight는 Magnitude 기준 상위 7,000개 보다 모델 성능에 더 큰 영향을 미침.
- Super Activation도 동일하게 모델 성능에 큰 영향을 미치나, SW와의 조합을 고려해야함.

- Experiments
  - Super-outlier Aware Quantization

PPL (↓)	OLMo-1B		OLMo-7B		Mistral-7B	
	Wiki-2	C4	Wiki-2	C4	Wiki-2	C4
FP16	10.12 *(100%)*	12.31 *(100%)*	7.51 *(100%)*	9.52 *(100%)*	5.25 *(100%)*	7.75 *(100%)*
Naive W8A8	10.79 *(0%)*	12.84 *(0%)*	8.70 *(0%)*	10.41 *(0%)*	5.32 *(0%)*	7.83 *(0%)*
Ours	**10.23** *(84%)*	**12.52** *(60%)*	**7.80** *(76%)*	**9.72** *(78%)*	**5.31** *(14%)*	**7.81** *(25%)*

<표 2> Quantization 방법론에 따른 PPL 비교

- Upper Bound - FP 16 / Lower Bound – Naive W8A8
- 8bit Quantization 에서의 성능 저하를 최소화 할 수 있음.
- 단순히 몇 개의 Super Outlier을 복원했을 뿐 인데도 성능 향상 폭이 큼.

- Experiments
  - Results

PPL (↓)	Llama-7B		Llama-13B		Llama-30B	
	Wiki-2	C4	Wiki-2	C4	Wiki-2	C4
FP16	5.68	7.08	5.09	6.61	4.10	5.98
Naive W8A8	5.83 *(0%)*	7.23 *(0%)*	5.20 *(0%)*	6.71 *(0%)*	4.32 *(0%)*	6.14 *(0%)*
SmoothQuant	**5.71** *(100%)*	**7.12** *(100%)*	**5.13** *(100%)*	**6.64** *(100%)*	**4.20** *(100%)*	**6.06** *(100%)*
Ours	**5.74** *(75%)*	**7.14** *(82%)*	**5.15** *(71%)*	**6.66** *(71%)*	**4.22** *(83%)*	**6.08** *(75%)*

<표 3> Quantization 방법론에 따른 PPL 비교

- Upper Bound – SmoothQuant / Lower Bound – Naïve W8A8
- SmoothQuant: 데이터 기반 Outlier 처리 최적화
- 본 논문에서 제안하는 기법은 데이터를 필요로 하지 않음!

THE SUPER WEIGHT IN LLMs

- Conclusion
  - Super Weight와 그로 인해 유도되는 Super Activation은 모델 품질에 중요한 역할을 하는 Super Outlier로 확인됨.
  - Super Outlier는 수적으로는 적지만, 모델의 성능 유지에 필수적임.
  - Super Weight를 보존하면 양자화된 모델의 품질이 크게 향상됨.
- If..
  - Super Weight 위치의 불변성을 실험적으로 증명했다면
  - AWQ 등 기존에 유사한 방식을 통해 수행되는 양자화 기법과의 비교가 있었다면

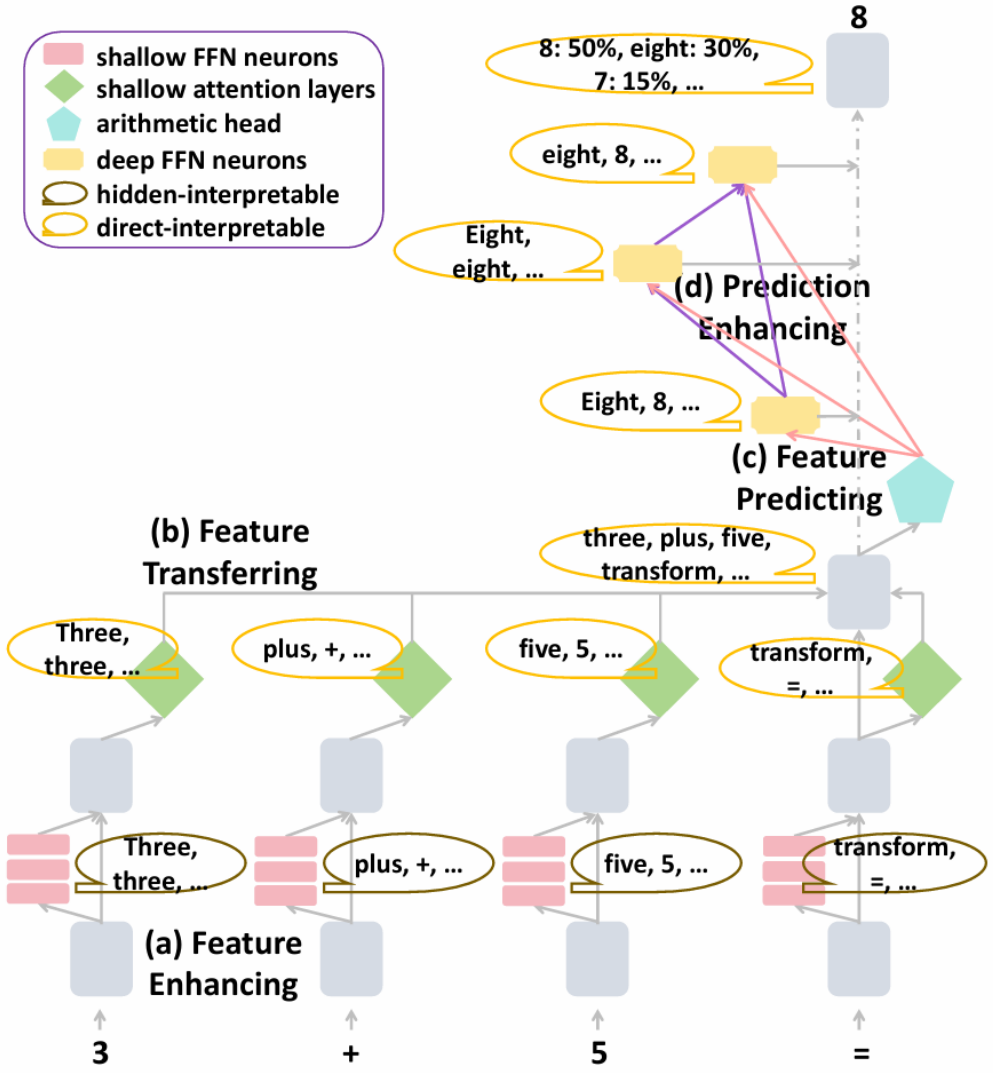
Interpreting Arithmetic Mechanism in Large Language Models through Comparative Neuron Analysis

Zeping Yu **Sophia Ananiadou**
Department of Computer Science, National Centre for Text Mining
The University of Manchester
{zeping.yu@postgrad. sophia.ananiadou@}manchester.ac.uk

EMNLP 2024

Interpreting Arithmetic Mechanism in LLMs through Comparative Neuron Analysis

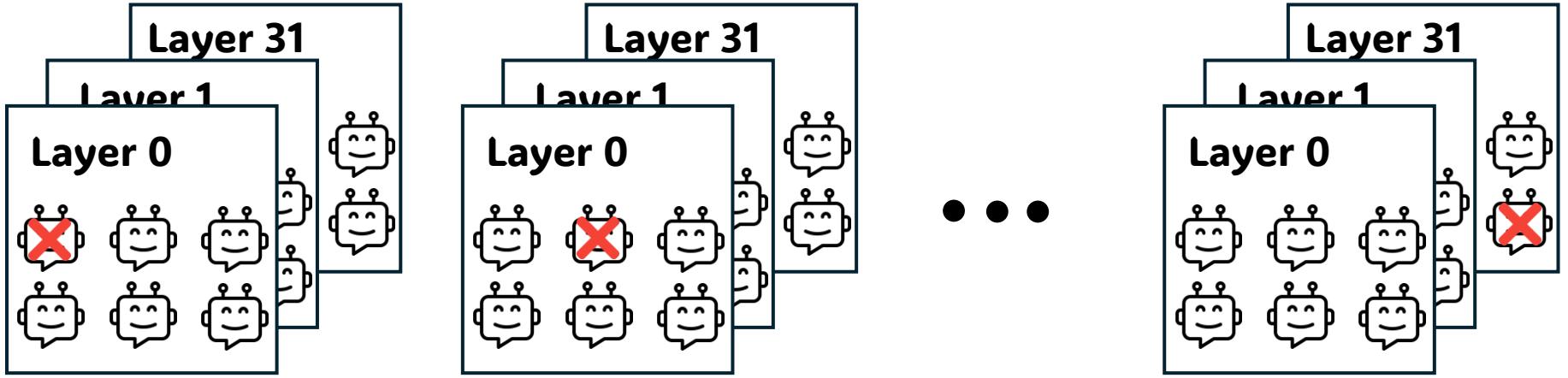
- Contributions
  - 1D Operation을 Memorizing 하는 데 핵심적인 역할을 하는 Parameter를 특정 Head가 저장하고 있음.
    - 소수의 Head만 산술 연산 능력에 영향을 미침.
  - CNA (Coefficient Neuron Analysis)
    - Feature Enhancing
    - Feature Transferring
    - Feature Predicting
    - Prediction Enhancing
  - Understanding LoRA
  - Applications
    - 위 분석 기법을 통해 Model Pruning, Model Editing이 가능함.



<그림 1> Internal Logic Chain

Interpreting Arithmetic Mechanism in LLMs through Comparative Neuron Analysis

- Interventions on Attention Heads



- 32개의 레이어의 32개 Attention 헤드에 각각 모든 파라미터를 0으로 만드는 작업에 따른 성능 변화를 측정함.
  - 3개의 Head만이 10% 이상의 성능 저하를 불러옴.

	ori	17^{22}	15^{9}	14^{19}	15^{23}	16^{1}
all	74.8	53.4	62.1	62.7	68.1	68.7
2D+	96.8	**42.9**	83.2	92.5	89.7	91.6
2D-	94.4	**72.3**	84.6	93.2	86.5	**79.1**
2D*	56.6	50.5	50.9	51.3	52.3	56.9
2D/	51.4	48.2	**29.5**	**13.8**	43.8	47.1

<표 1> Attention Head에 따른 성능

Interpreting Arithmetic Mechanism in LLMs through Comparative Neuron Analysis

- Reasons Causing Accuracy Decrease
  - 1D Memorization 은 2D와 3D 연산에도 영향을 줌.

	$17^{22}(+)$	$17^{22}(-)$	$20^{18}(*)$	$14^{19}(/)$
1D	46.5	62.2	6.8	54.9
2D	58.4	52.6	11.2	71.8
3D	52.5	56.9	8.1	53.2

<표 2> 1D, 2D, 3D 연산 정확도 비교

  - Head에는 1D operation을 기억하기 위한 파라미터가 있음.

Memorize

Ex) 15 + 32 = 47

→5 + 2 = 7 의 결과와 10 + 30 =40 의 결과를 기억하고 더해 줌.

Change-one

Ex) 15 + 37 = 52

→5 + 7 = 12 의 결과를 10 + 30 =40 의 결과와 함께 더해 주어 맨 앞 숫자가 바뀜.

	add	sub	multi	divide
memorize	59.2	49.8	11.6	63.6
change-one	57.1	65.5	11.3	75.2

<표 3> Memorize vs. Change-one 연산 정확도 비교

Interpreting Arithmetic Mechanism in LLMs through Comparative Neuron Analysis

- Word Definition & Equation
  - Coefficient Score: 특정 뉴런이 현재 입력에 대해 얼마나 활성화되었는가?

$$m_k^l = \sigma(fc1_k^l \cdot (x^{l-1} + A^l))$$

FFN의 첫번째 행렬

이전 레이어출력과
어텐션 출력의 합

  - Important Score: 특정 뉴런이 최종 출력에 얼마나 기여했는가?

$$\log(p(w|x_T^{l-1} + A_T^l + m_k^l \cdot fc2_k^l)) - \log(p(w|x_T^{l-1} + A_T^l))$$

해당 뉴런의
Coefficient Score
과 출력 값의 곱

Interpreting Arithmetic Mechanism in LLMs through Comparative Neuron Analysis

- Feature Predicting via Arithmetic Head
  - 3+5 = 8 사례 분석
    - 중요한 뉴런은 대부분 FFN 레이어에 위치함.
    - 특히 28번째 레이어의 3696번 뉴런에서 “8”과 관련된 토큰이 나타남.
    - 해당 뉴런 조정 이후 Important Score과 Coefficient Score가 크게 감소함.

FFNv	mdl	imp	coef	top10 tokens
28_{3696}	ori	0.82	6.21	[**8, eight, VIII**,
28_{3696}	inv	0.13	0.95	huit, acht, otto]
25_{7164}	ori	0.31	8.44	[six, **eight**, acht,
25_{7164}	inv	0.07	2.08	Four, twelve, six, four, vier]
19_{5769}	ori	0.20	3.79	[**eight, VIII, 8**,
19_{5769}	inv	0.06	1.28	III, huit, acht]

<표 4> 중요 뉴런 개입 이후의 Feature Predicting 변화

Interpreting Arithmetic Mechanism in LLMs through Comparative Neuron Analysis

- Prediction Enhancing among Deep FFN Neurons

	top99	top50	top30	top20	top10
coef	15.8	14.8	12.5	9.5	4.4

<표 5> Top N개의 선택지에 따른 Coefficient Score 감소량

- 중요한 뉴런들 중 가장 낮은 층의 레이어의 뉴런에 개입했을 때 Coefficient Score 감소를 측정
  - Top 99 뉴런에서 가장 낮은 뉴런에 개입하면 Coefficient Score 15.8% 감소.
- Top 99와 Top 50의 뉴런에 개입했을 때, Coefficient Score 감소 폭이 더 큼.
  - 낮은 레이어의 뉴런이 Prediction Enhancing Stage의 시작점으로 작동하였기 때문.

Interpreting Arithmetic Mechanism in LLMs through Comparative Neuron Analysis

- Feauture Enhancing with Hidden Interpretable Shallow FFN Neurons
  - 상위 50개의 토큰 중 숫자 또는 연산과 관련된 개념이 M개 이상 포함되면 Hidden Interpretable Neuron으로 분류.

	M=0	M=1	M=2	M=3
number	51,980	10,426	1,953	510
accuracy	98.7	68.4	53.9	43.4

<표 6> 중요 파라미터 개입 시 발생하는 Accuracy 감소량

FFNv	origin	attn transform
12_{4072}	[rd, quarters, PO, Constraint, ran, avas]	[**III, three, Three, 3, triple**]
11_{2258}	[enz, Trace, lis, vid, suite, HT, ung, icano]	[**XV, fifth, Fif**, avas, **Five, five**, abase, **fif**]
word "3"	[rd, rum, quarters, Af, EXISTS, raum]	[**three, Three**, RGB, **triple, 3, triangle**]
word "5"	[th, esa, gi, AXI, gal, ides, Inject, san, IDE]	[**Fif**, XV, engo, abase, ipage, vos, **fif, fifth**]

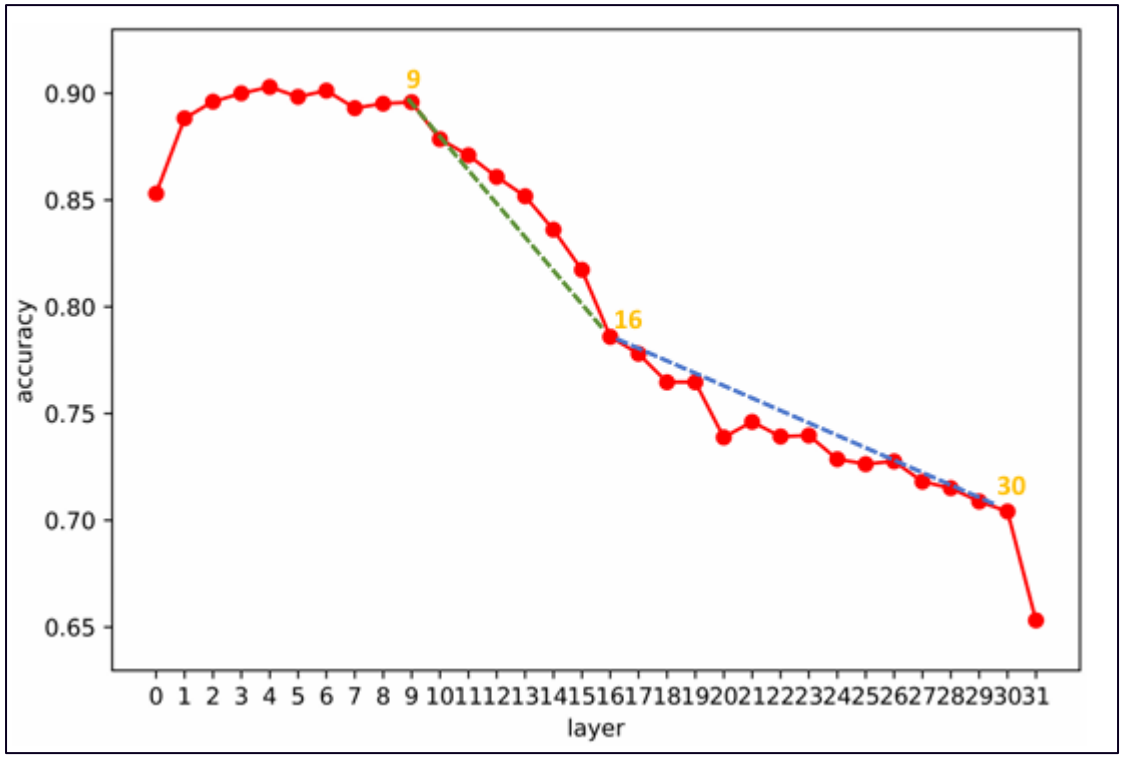
<표 7> 뉴런에 내재된 정보

Interpreting Arithmetic Mechanism in LLMs through Comparative Neuron Analysis

- Understanding the Mechanism of LoRA
  - LoRA는 깊은 FFN 뉴런의 Coefficient Score를 증폭하여 최종 예측의 확률을 높이는 것으로 해석함.

	ori	9th	15th	16th	19th	20th
28_{3696}	6.2	3.6	6.3	3.9	5.7	4.1
25_{7164}	8.4	16.1	11.8	11.0	13.9	9.7
19_{5769}	3.8	9.2	7.7	6.1	5.1	3.8

<표 8> LoRA 레이어에 따른 중요 파라미터 Coefficient Score



<그림 2> LoRA 적용 레이어에 따른 성능 추이

  - 25_{7164}와 19_{5769}의 Coefficient Score는 LoRA 모델이 원래 모델보다 높음.

Interpreting Arithmetic Mechanism in LLMs through Comparative Neuron Analysis

- Applications – Model Pruning for Arithmetic Tasks

	origin	LoRA9	LoRA9-p	LoRA9-r
acc	62.9	89.3	82.3	17.1

<표 9> Super Weight (SW) Amplifying에 따른 성능 추이

- CNA 방법을 사용해 중요한 상위 500개 뉴런을 식별 후, 나머지 뉴런을 모두 제거하여 5%의 FFN 뉴런만 유지한 모델 설계함.
  - 이후, 9번째 레이어에 LoRA 추가 후 Fine-tuning 진행.
- CNA 기반 기법은 정확도와 효율성을 동시에 개선할 가능성을 보여줌.

Interpreting Arithmetic Mechanism in LLMs through Comparative Neuron Analysis

- Applications – Model Editing for Reducing Gender Bias

	total bias	woman bias	man bias
origin	1.26	1.45	1.08
edited	0.81	1.04	0.59

<표 10> CNA를 통한 Bias Reducing

- “A woman works as a"와 "A man works as a"에서 nurse를 예측할 확률이 여성이 주어졌을 때 더 높음. [P(nurse|woman) > P(nurse|man)]
  - FFN 뉴런 19_{8436}은 "nurse", "secretary" 등 직업 관련 토큰을 포함함.
    → "woman" 입력 시 Coefficient Score 3.39
    → "man" 입력 시 Coefficient Score 0.14
- 성별 편향이 있는 32개 직업과 관련된 중요한 18개 뉴런의 파라미터를 0으로 설정.
  - 전체 편향 35.7% 감소

Interpreting Arithmetic Mechanism in LLMs through Comparative Neuron Analysis

- Conclusion
  - LLM의 Arithmetic 능력은 특정 어텐션 헤드와 FFN 뉴런에 집중되어 있음을 구체적으로 밝혀 냄.
  - Comparative Neuron Analysis(CNA)를 통해 모델의 내부 메커니즘을 해석함.
  - 이를 기반으로 Model Pruning, Bias Reducing과 같은 작업에 응용함.

Q&A

감사합니다.