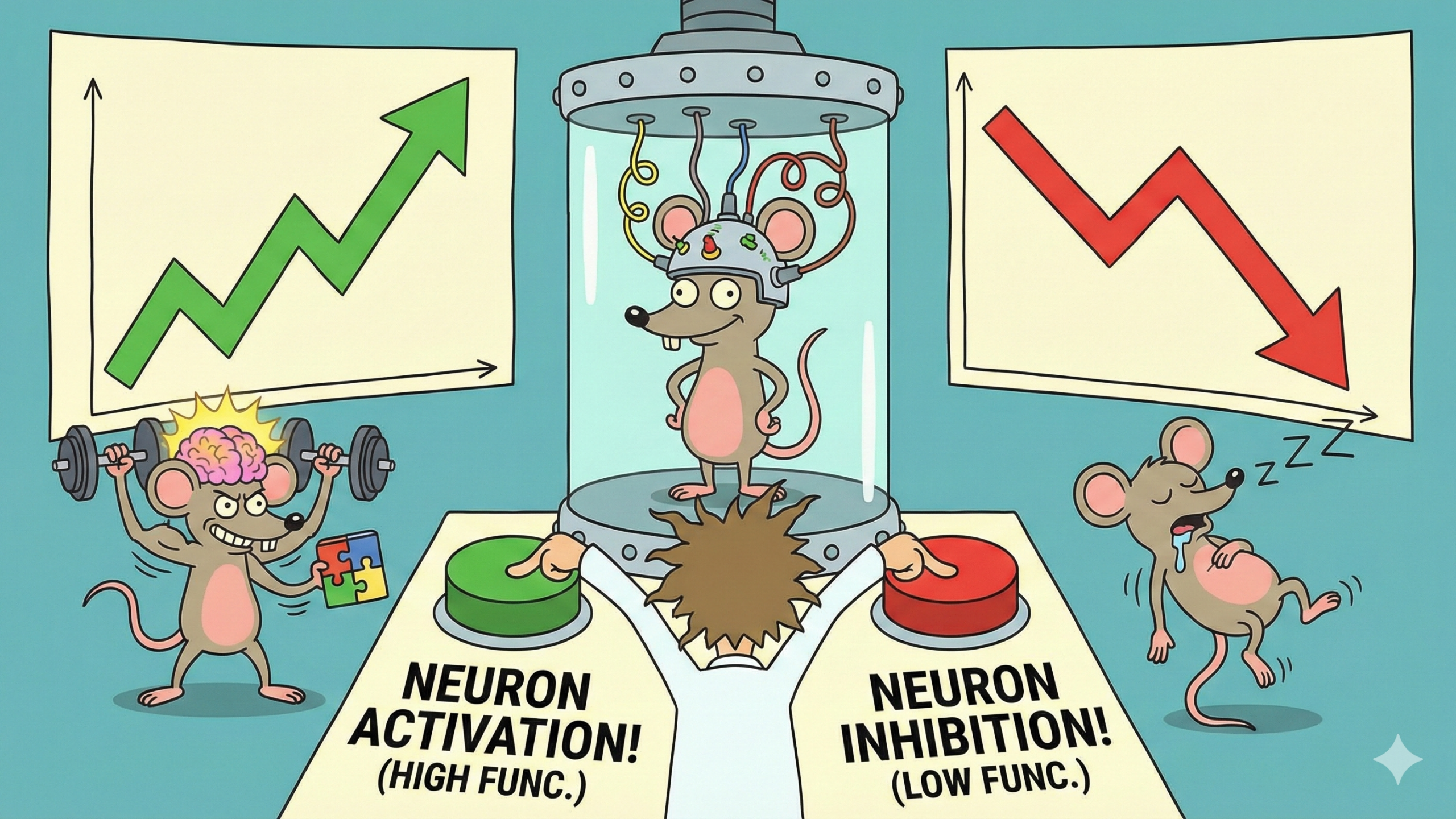
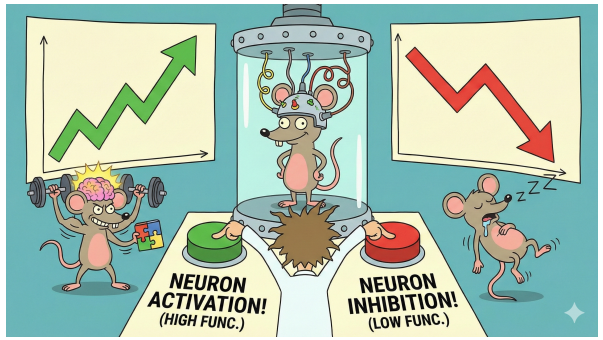
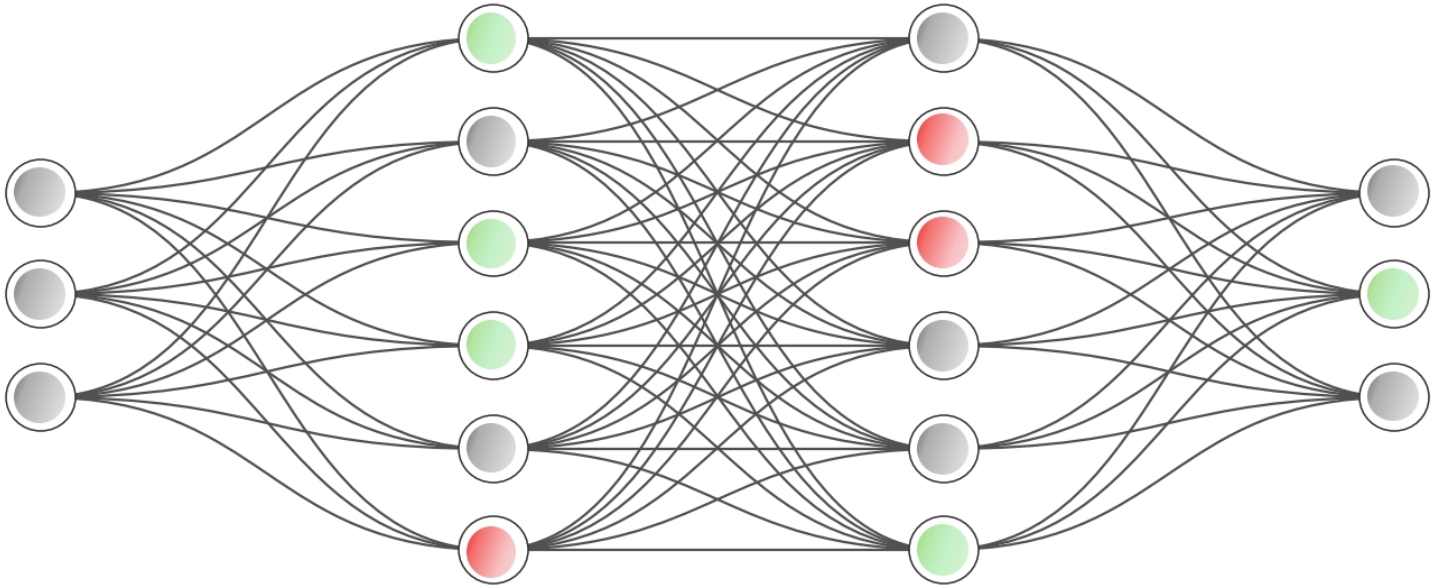
NeuronLLM: Identifying Good and Bad Neurons for Task-Level Controllable LLMs


NEURON ACTIVATION! (HIGH FUNC.)
NEURON INHIBITION! (LOW FUNC.)
ZZZ

😇 Good Neuron 😈 Bad Neuron 🔥 Excite ❄ Silence

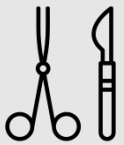


What about for LLMs?



How to measure the contribution of a single neuron?

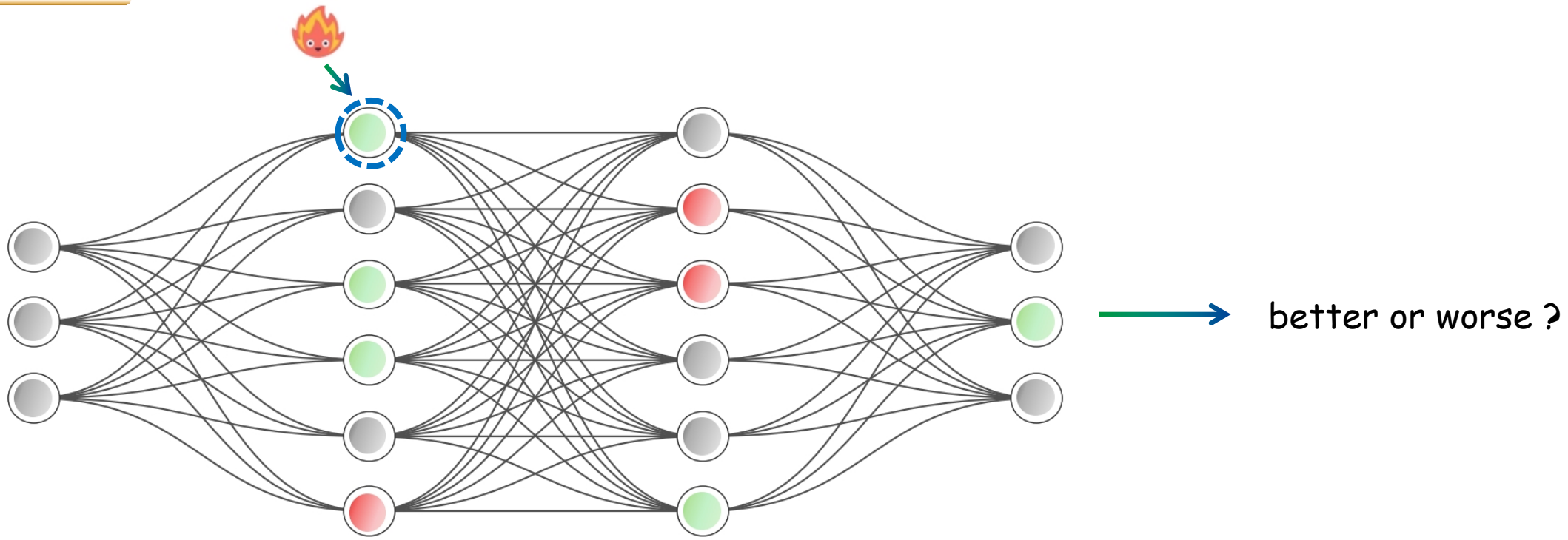
How biologists test neurons:



- invasive techniques
- chemical
- optogenetics
- ...

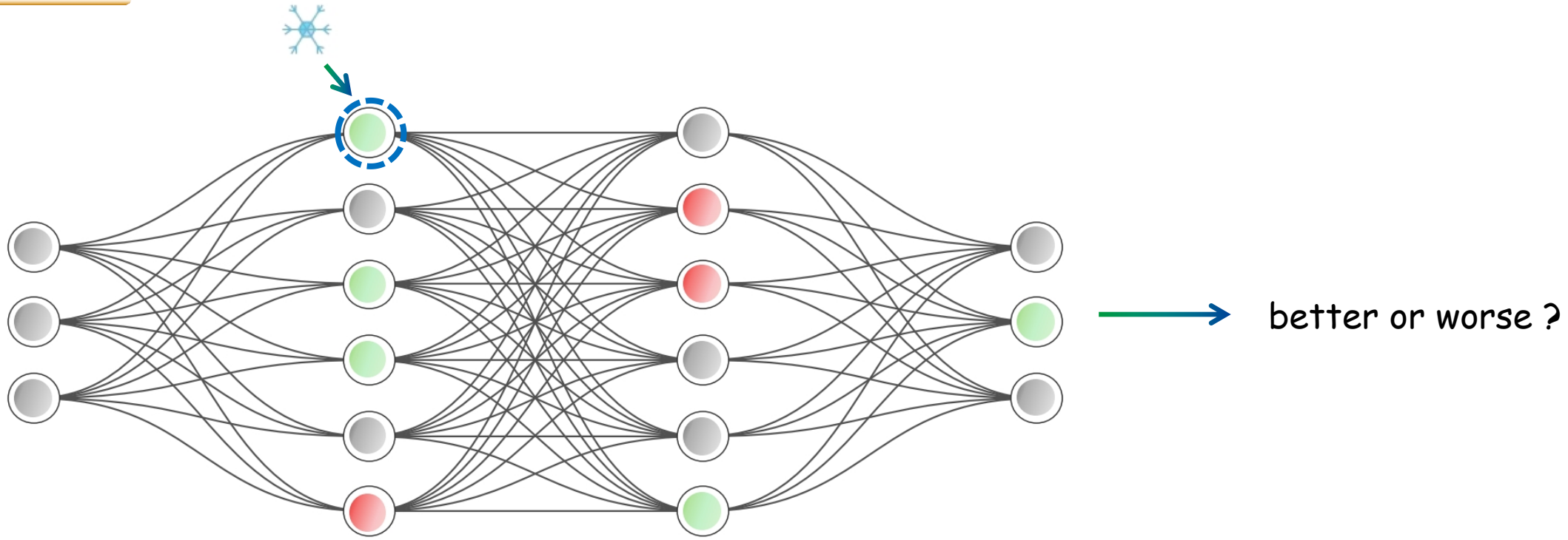
Exicite

activation × 2

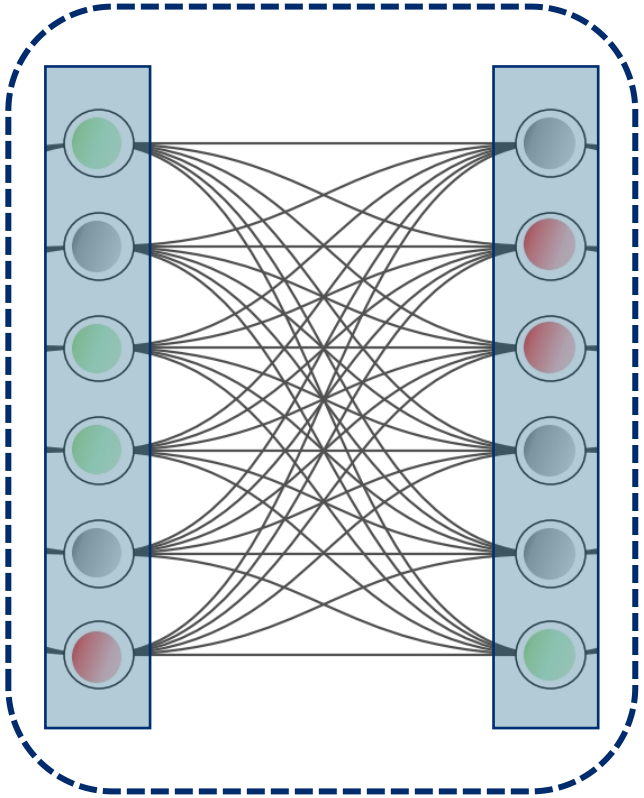


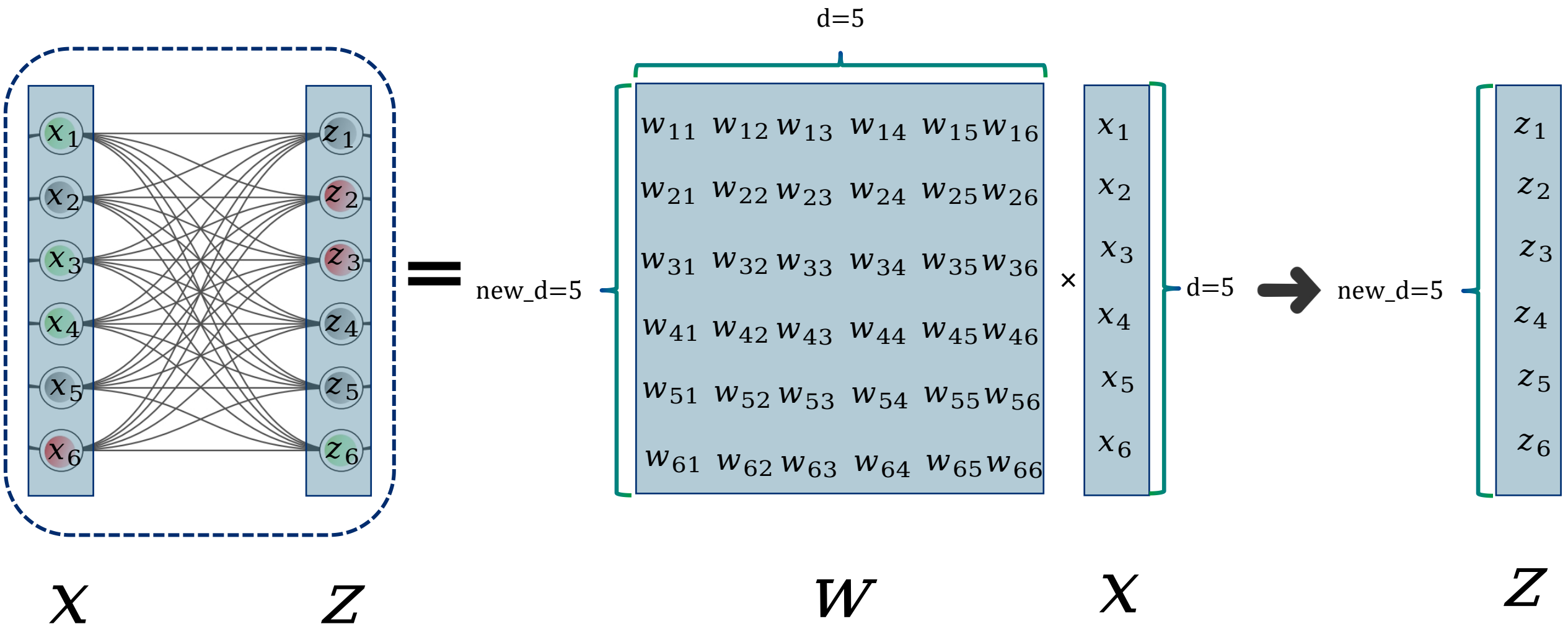
Silence

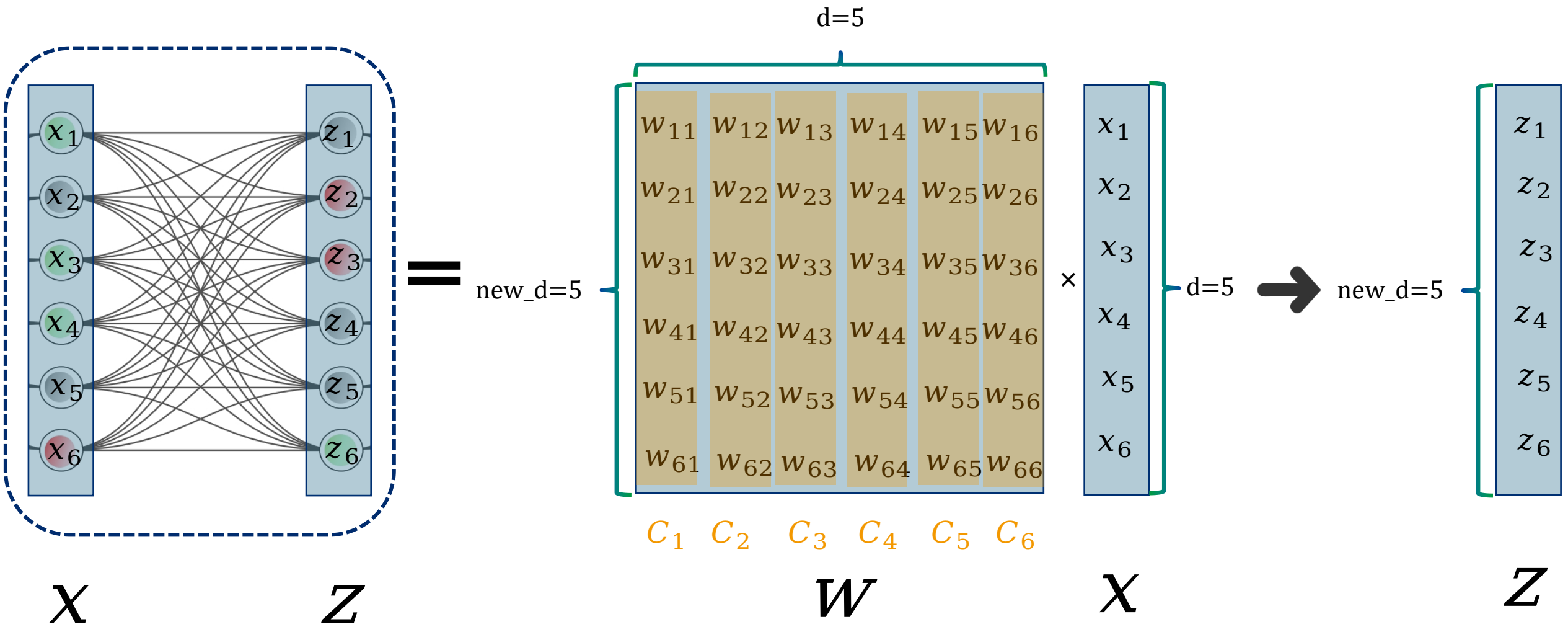
activation × 0

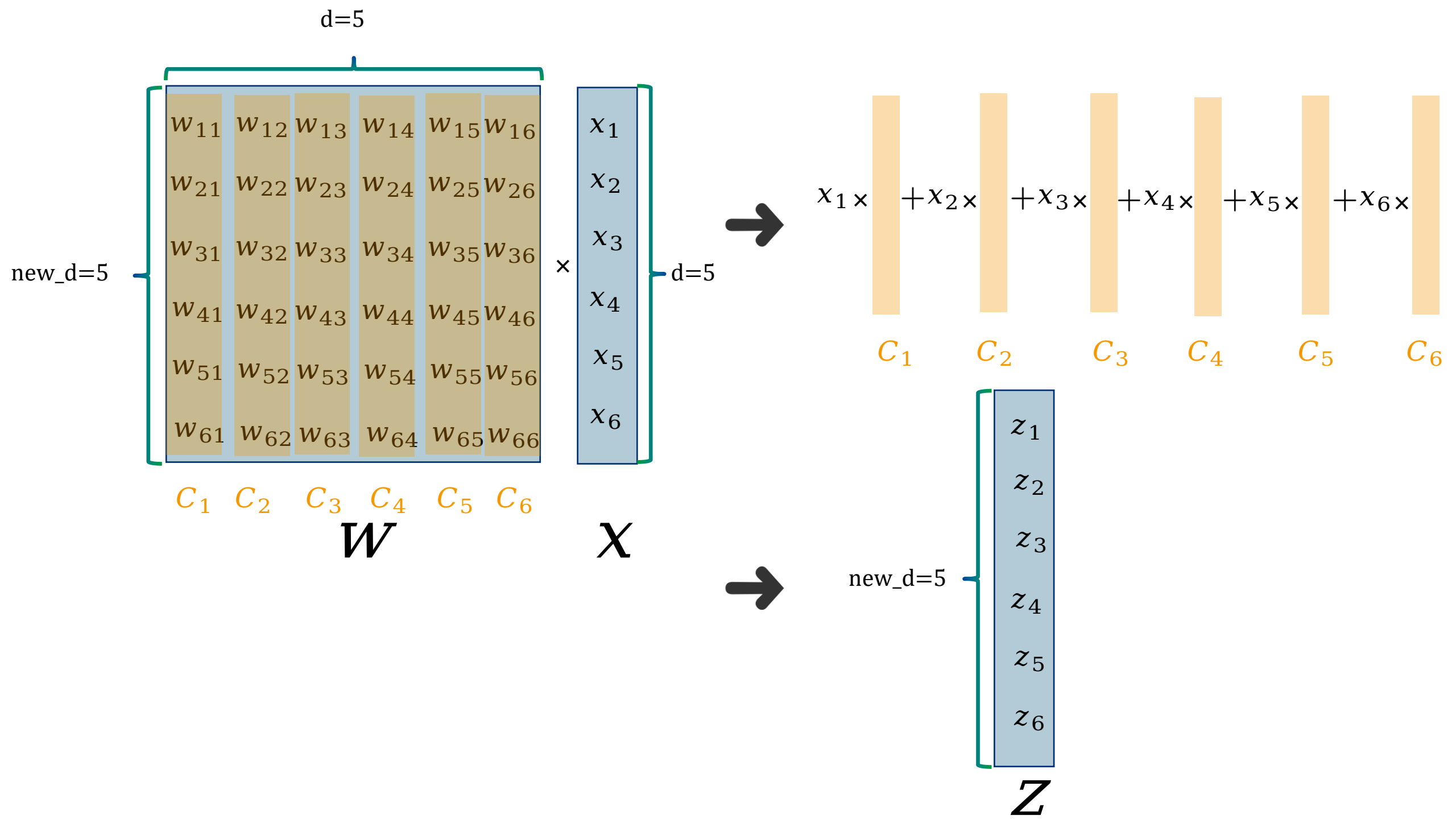


What happens inside?

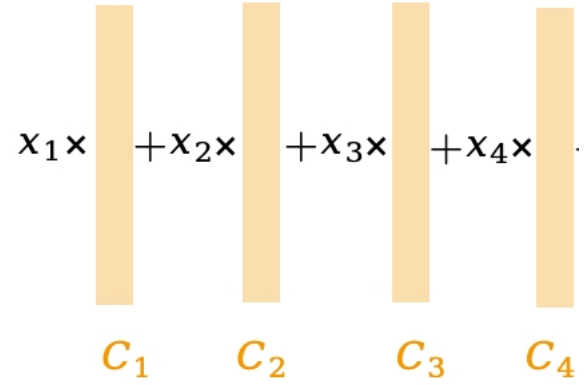


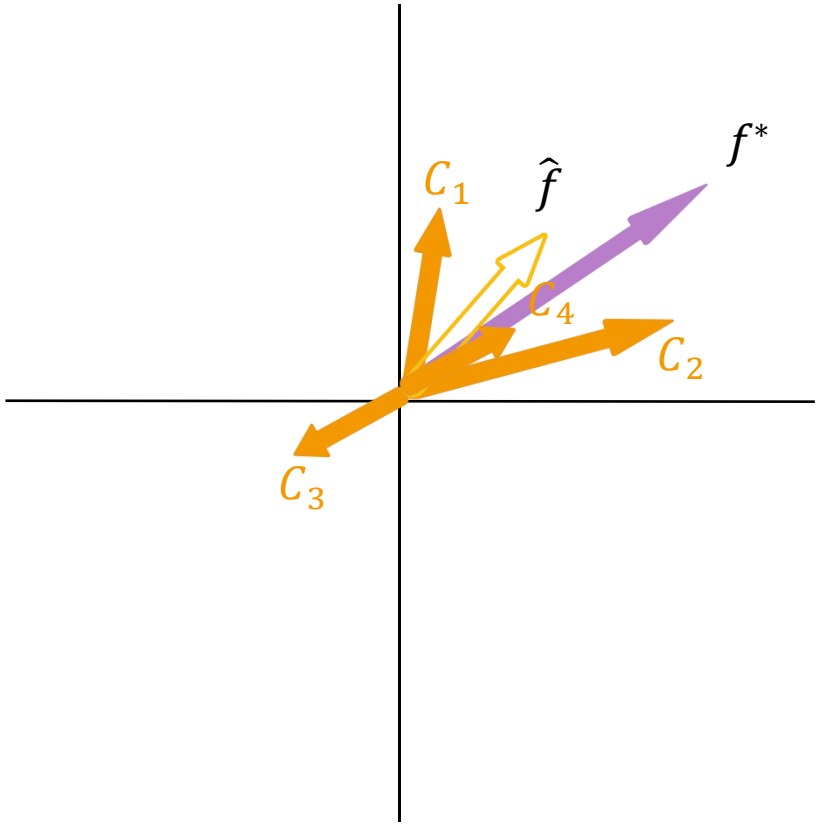

d=5
x_1 x_2 x_3 x_4 x_5 x_6
z_1 z_2 z_3 z_4 z_5 z_6
=
new_d=5
w_{11} w_{12} w_{13} w_{14} w_{15} w_{16}
w_{21} w_{22} w_{23} w_{24} w_{25} w_{26}
w_{31} w_{32} w_{33} w_{34} w_{35} w_{36}
w_{41} w_{42} w_{43} w_{44} w_{45} w_{46}
w_{51} w_{52} w_{53} w_{54} w_{55} w_{56}
w_{61} w_{62} w_{63} w_{64} w_{65} w_{66}
×
x_1 x_2 x_3 x_4 x_5 x_6
d=5
new_d=5
z_1 z_2 z_3 z_4 z_5 z_6
X
Z
W
X
Z

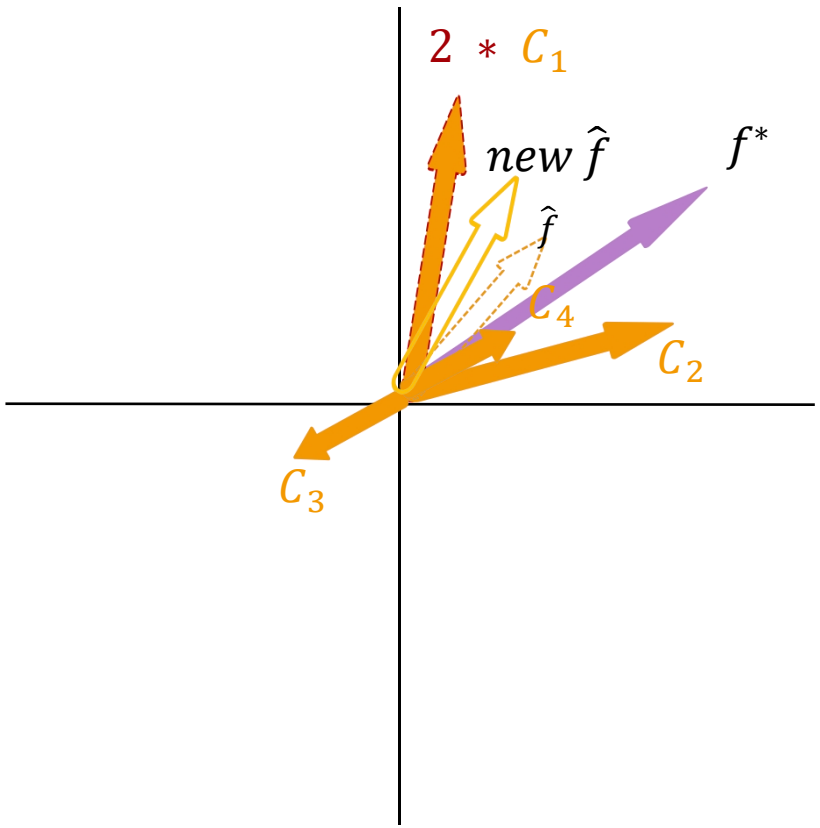

d=5
new_d=5
d=5
new_d=5
x_1 x_2 x_3 x_4 x_5 x_6
z_1 z_2 z_3 z_4 z_5 z_6
w_{11} w_{12} w_{13} w_{14} w_{15} w_{16}
w_{21} w_{22} w_{23} w_{24} w_{25} w_{26}
w_{31} w_{32} w_{33} w_{34} w_{35} w_{36}
w_{41} w_{42} w_{43} w_{44} w_{45} w_{46}
w_{51} w_{52} w_{53} w_{54} w_{55} w_{56}
w_{61} w_{62} w_{63} w_{64} w_{65} w_{66}
C_1 C_2 C_3 C_4 C_5 C_6
x_1 x_2 x_3 x_4 x_5 x_6
z_1 z_2 z_3 z_4 z_5 z_6
x z w x z

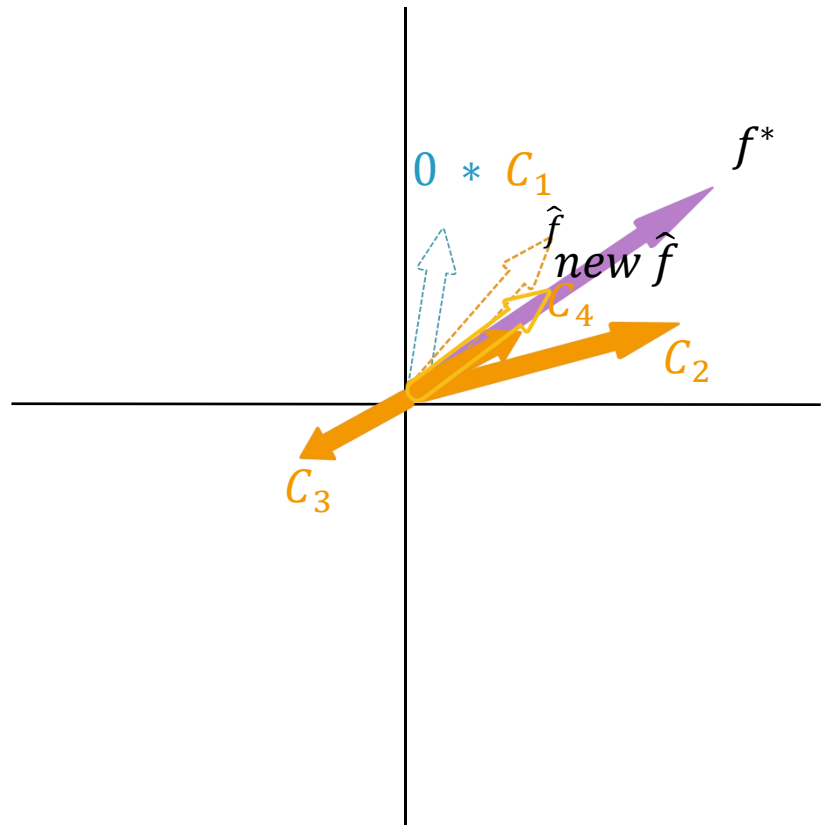

d=5
new_d=5
d=5
×
w_{11} w_{12} w_{13} w_{14} w_{15} w_{16}
w_{21} w_{22} w_{23} w_{24} w_{25} w_{26}
w_{31} w_{32} w_{33} w_{34} w_{35} w_{36}
w_{41} w_{42} w_{43} w_{44} w_{45} w_{46}
w_{51} w_{52} w_{53} w_{54} w_{55} w_{56}
w_{61} w_{62} w_{63} w_{64} w_{65} w_{66}
C_1 C_2 C_3 C_4 C_5 C_6
W
x_1 x_2 x_3 x_4 x_5 x_6
X
$x_1 \times$ $+x_2 \times$ $+x_3 \times$ $+x_4 \times$ $+x_5 \times$ $+x_6 \times$
C_1 C_2 C_3 C_4 C_5 C_6
new_d=5
z_1 z_2 z_3 z_4 z_5 z_6
Z

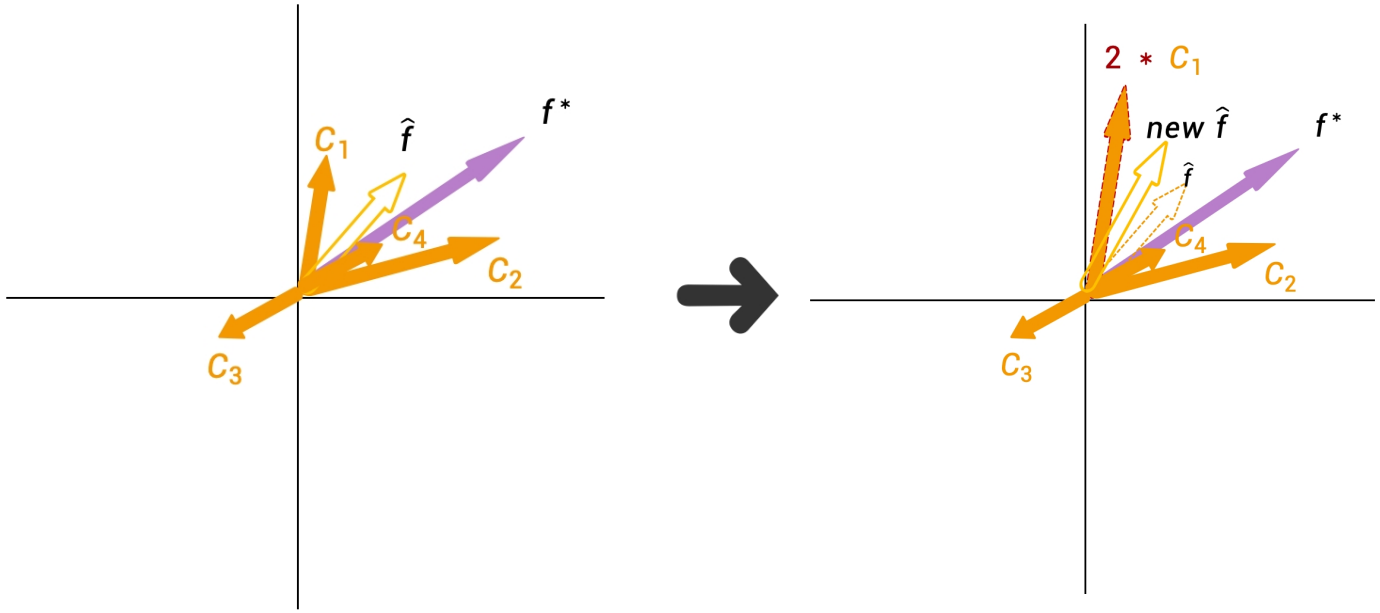
$$Z = f(X) = f(x_1, x_2, x_3, x_4) =$$

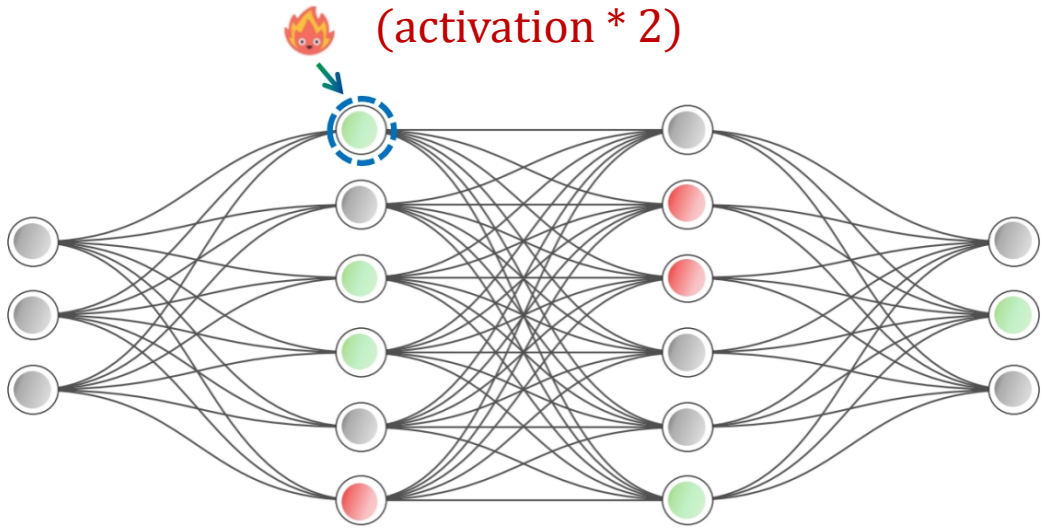





2 * C_1
new $\hat{f}$
f^*
$\hat{f}$
C_4
C_2
C_3

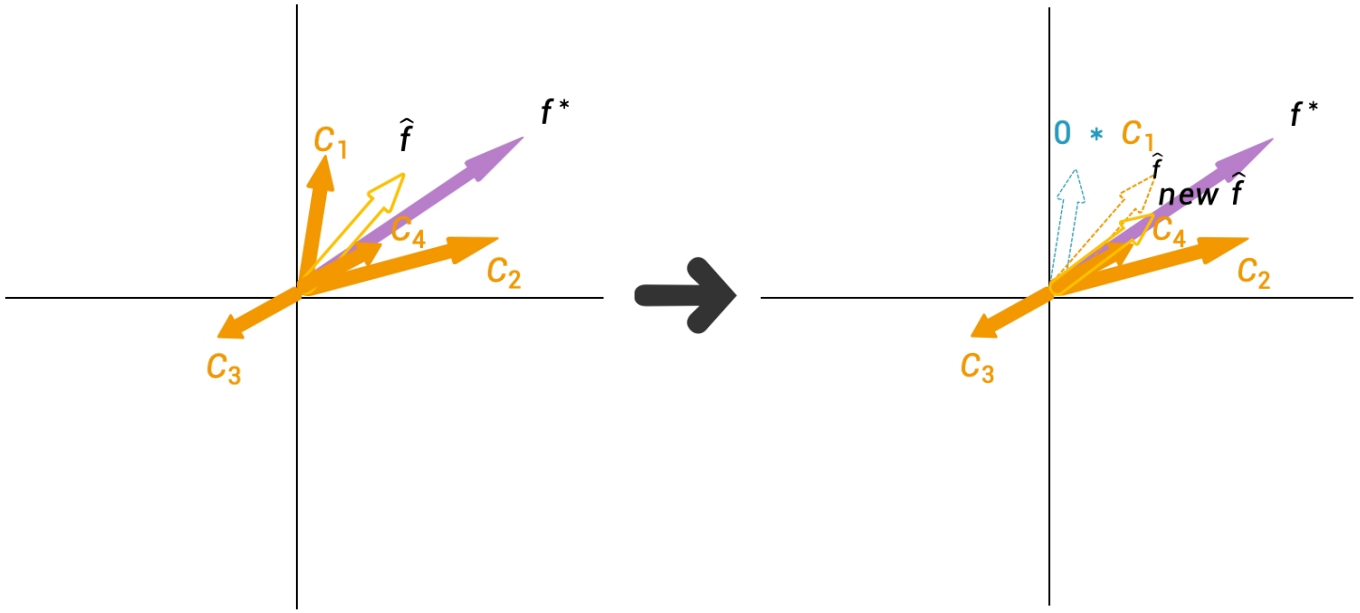

$0 * C_1$
f^*
$\hat{f}$
$new\ \hat{f}$
C_4
C_2
C_3

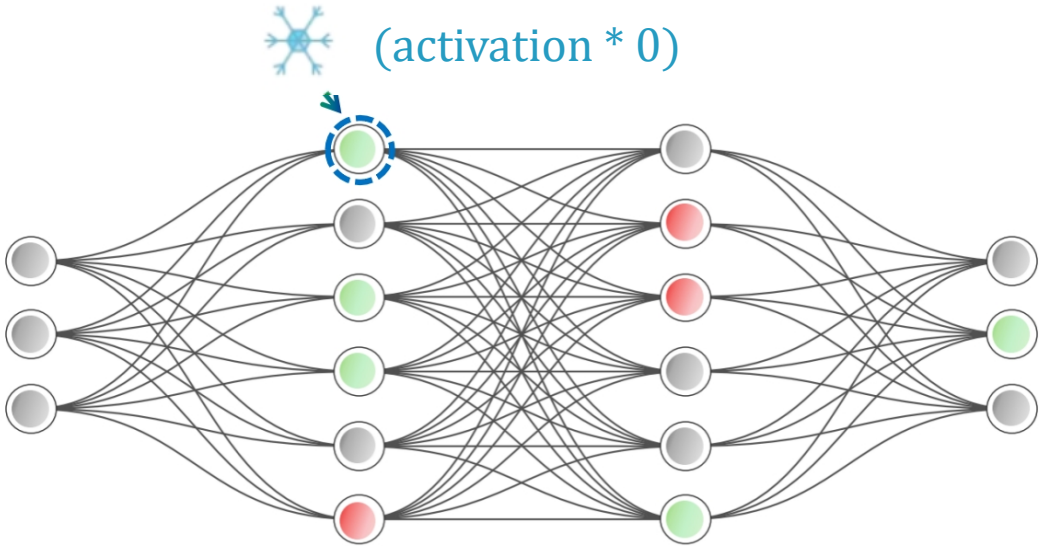




$$\hat{f} = \underset{(x_1)}{1.2 * C_1} + \underset{(x_2)}{0.9 * C_2} + \underset{(x_3)}{(-1) * C_3} + \underset{(x_4)}{(-0.5) * C_4}$$

$$new\ \hat{f} = \underset{(x_1\ *\ 2)}{2.4 * C_1} + \underset{(x_2)}{0.9 * C_2} + \underset{(x_3)}{(-1) * C_3} + \underset{(x_4)}{(-0.5) * C_4}$$

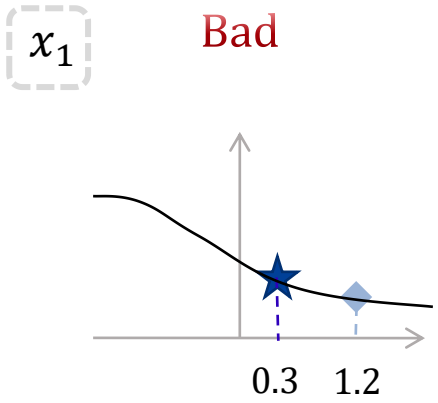


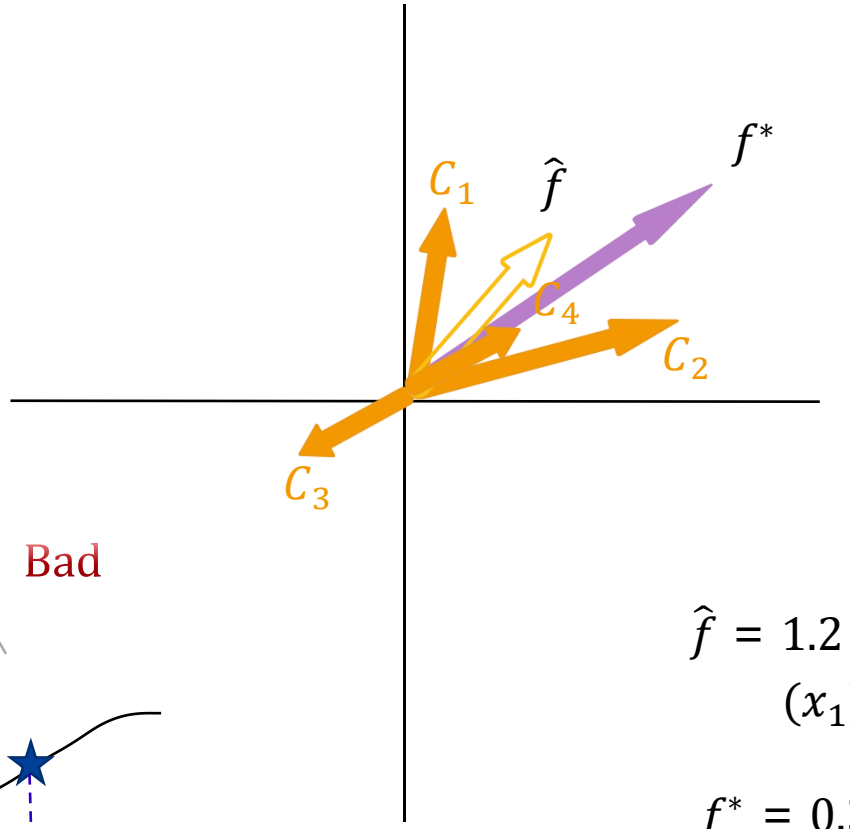


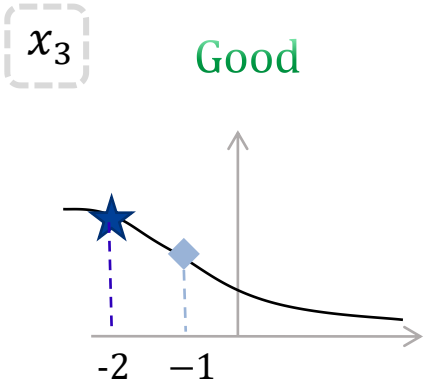
$$\hat{f} = \underset{(x_1)}{1.2 * C_1} + \underset{(x_2)}{0.9 * C_2} + \underset{(x_3)}{(-1) * C_3} + \underset{(x_4)}{(-0.5) * C_4}$$

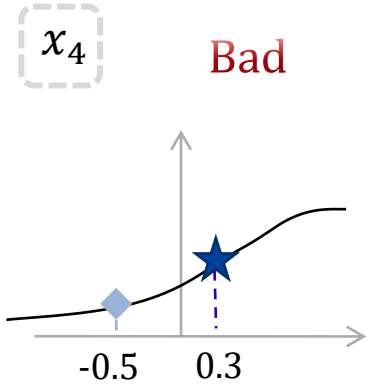
$$new\ \hat{f} = \underset{(x_1 * 0)}{0 * C_1} + \underset{(x_2)}{0.9 * C_2} + \underset{(x_3)}{(-1) * C_3} + \underset{(x_4)}{(-0.5) * C_4}$$

$$Z = f(X) = f(x_1, x_2, x_3, x_4) = x_1 \times C_1 + x_2 \times C_2 + x_3 \times C_3 + x_4 \times C_4$$



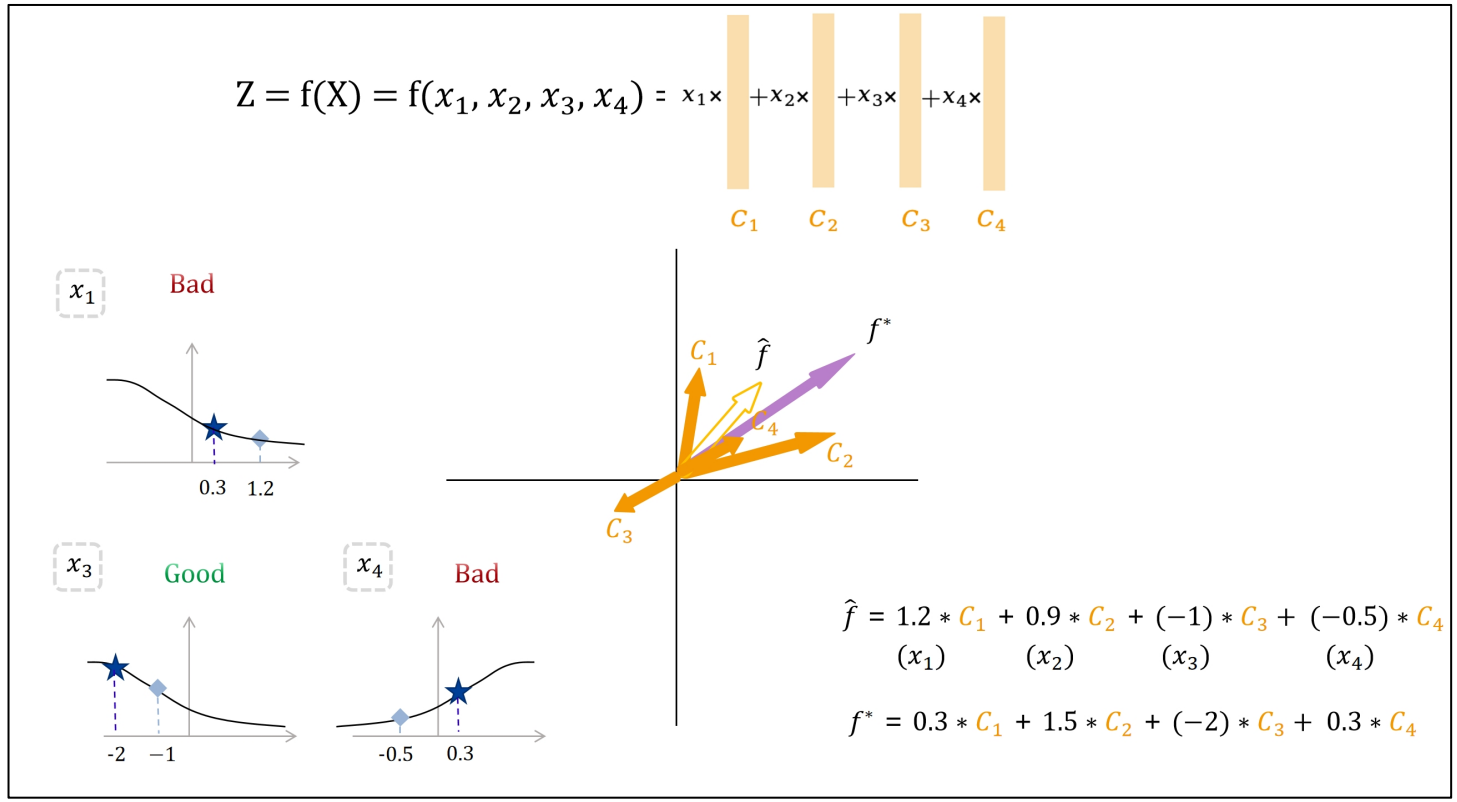






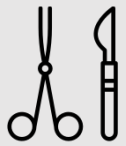
$$\hat{f} = \underset{(x_1)}{1.2} * C_1 + \underset{(x_2)}{0.9} * C_2 + \underset{(x_3)}{(-1)} * C_3 + \underset{(x_4)}{(-0.5)} * C_4$$

$$f^* = 0.3 * C_1 + 1.5 * C_2 + (-2) * C_3 + 0.3 * C_4$$



mathematically ?

LLMs' natural advantage



- invasive techniques
- chemical
- optogenetics
- ...

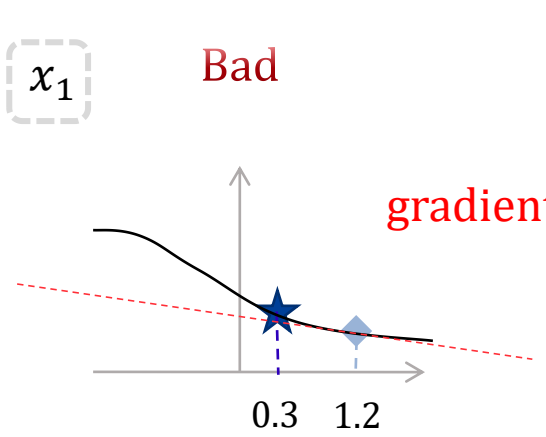
LLMs' natural advantage:

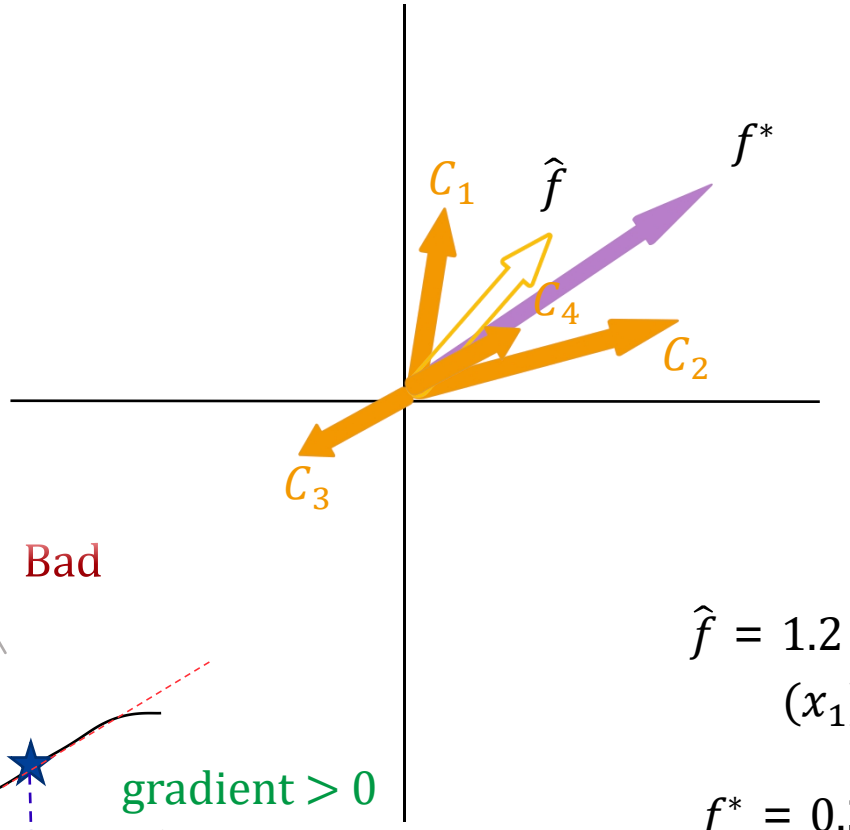
⚡ Backpropagation!

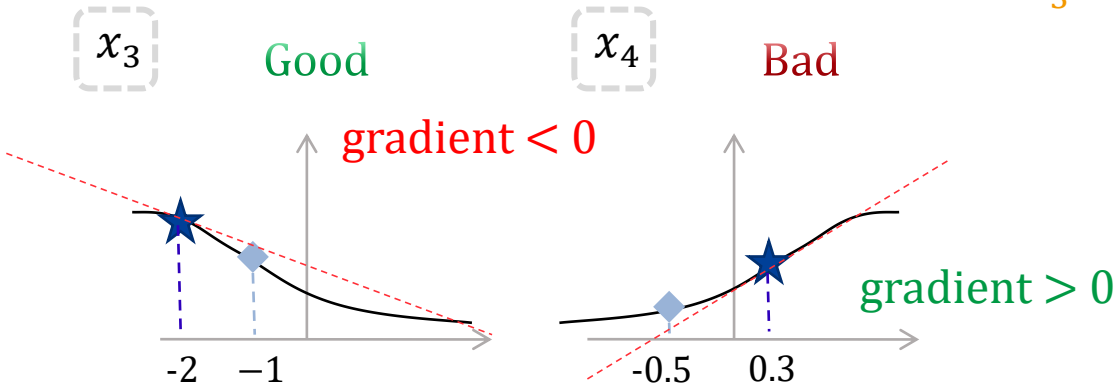
Gradient > 0 → Increases loss

Gradient < 0 → Reduces loss

$$Z = f(X) = f(x_1, x_2, x_3, x_4) = x_1 \times C_1 + x_2 \times C_2 + x_3 \times C_3 + x_4 \times C_4$$



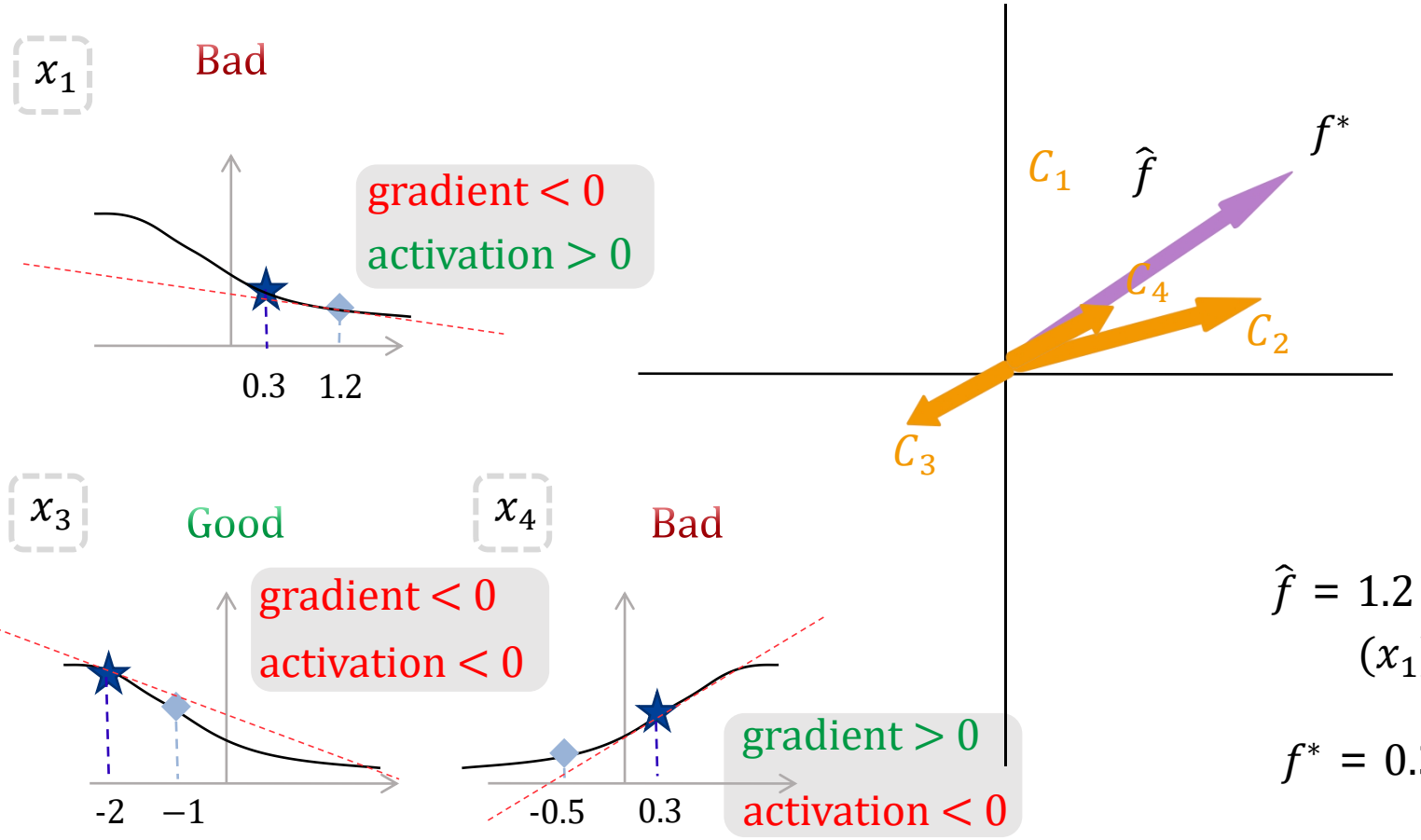




$$\hat{f} = \underset{(x_1)}{1.2} * C_1 + \underset{(x_2)}{0.9} * C_2 + \underset{(x_3)}{(-1)} * C_3 + \underset{(x_4)}{(-0.5)} * C_4$$

$$f^* = 0.3 * C_1 + 1.5 * C_2 + (-2) * C_3 + 0.3 * C_4$$

$$Z = f(X) = f(x_1, x_2, x_3, x_4) = x_1 \times C_1 + x_2 \times C_2 + x_3 \times C_3 + x_4 \times C_4$$



$$\hat{f} = \underset{(x_1)}{1.2} * C_1 + \underset{(x_2)}{0.9} * C_2 + \underset{(x_3)}{(-1)} * C_3 + \underset{(x_4)}{(-0.5)} * C_4$$

$$f^* = 0.3 * C_1 + 1.5 * C_2 + (-2) * C_3 + 0.3 * C_4$$

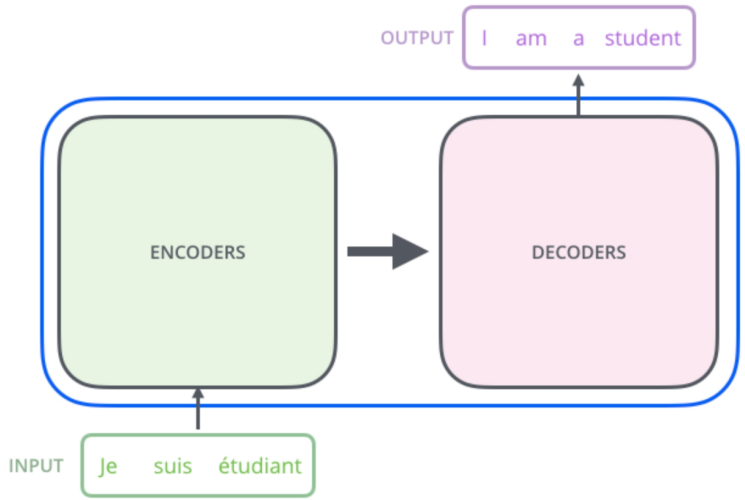
The contribution of a single neuron

$$\text{IG}(w_i^l) := \frac{\hat{w}_i^l}{m} \times \sum_{k=1}^{m} \frac{\partial F(\frac{k\hat{w}_i^l}{m})}{\partial w_i^l}, \tag{1}$$

where w_i^l is the i^{th} neuron in a l^{th} Feed-Forward Network (FFN) layer, $\hat{w}_i^l$ is its assigned value, and m is the number of steps to approximate the integral. This work is focused on neurons in the FFNs since FFNs in LLMs are found to encode meaningful features responsible for different

OK, so what's the target F?

Popping open that Optimus Prime goodness, we see an encoding component, a decoding component, and connections between them.

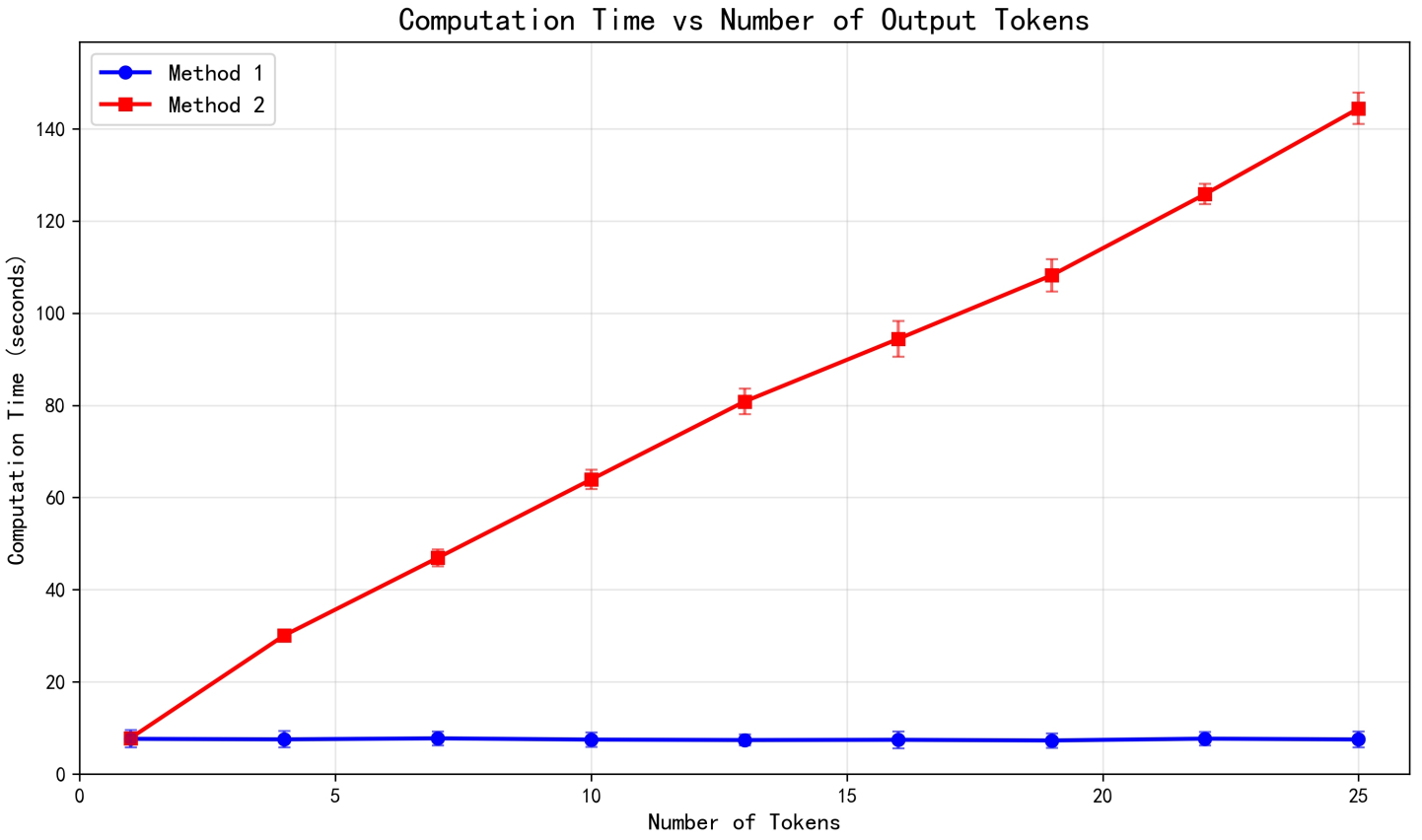


For LLMs, given a query q (*e.g.*, *Paris is the capital of*), the target function F is often set as the sum of the log probabilities of each token in the answer string y (*e.g.*, *France*):

$$P(y|w_i^l, q) = \sum_{j=1}^{n} \log P(t_j|\hat{w}_i^l, q, t_1, \ldots, t_{j-1}), \tag{2}$$

where y is tokenized into n discrete tokens $\{t_1, t_2, \ldots, t_n\}$ (*e.g.*, [“F”, “ran”, “ce”]). Each $P(t_j|w_i^l, q, t_1, \ldots, t_{j-1})$ represents the conditional probability of generating token t_i given the query prompt q and previously generated tokens.

This brings the first challenge



OOM :(

The second challenge

The data formats of different tasks could be various!

N-E-R

[task] tell me the entity type of a word in a sentence

[input] Sentence: Isabella gets home early. Word: Isabella

[output] person

SENTIMENT

[task] analyse the sentiment and fill the cloze

[input] this is my favourite bistro!

[output] the sentiment of this speaker is positive

CHUNKING

[task] tag the verb phrase with @@

[input] 'the children are laughing in the park.'

[output] 'the children @@are laughing@@ in the park.'

COMMONSENSE

[task] answer the given question

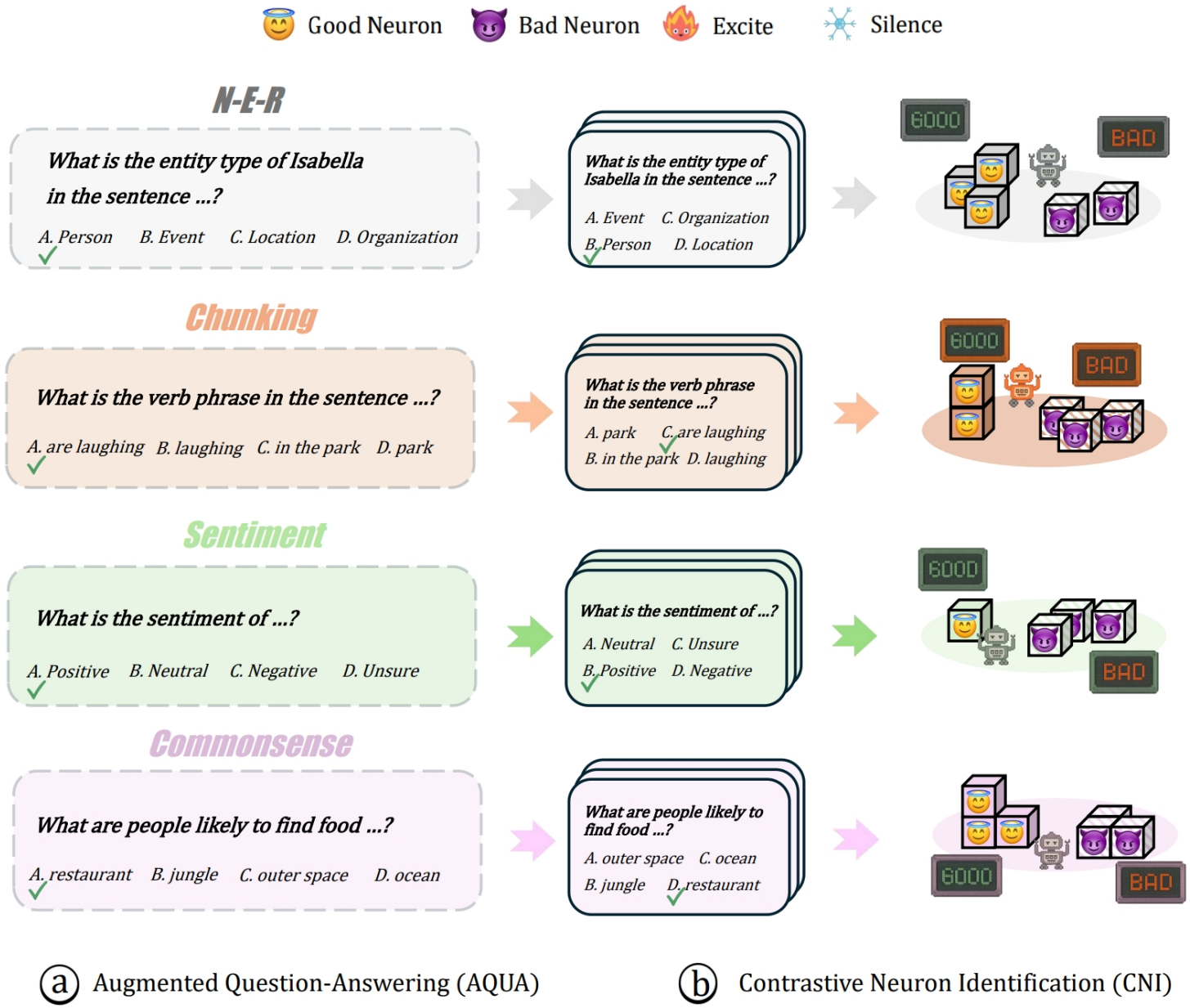
[input] where are people likely to find food?

[output] restaurant

These methods are all customized for their specific data :(

Despite these advances, the quest to identify and interpret individual neurons remains central, partly because neurons are a natural basis for explaining network behaviors, and also because identifying a single "unit" responsible for a behavior is intuitively plausible. One representative work in this scope is Knowledge Neurons (Dai et al., 2022) which store particular facts (*e.g.*, the capital of France). Other works often focus on different capabilities, such as Syntactic Agreement, Word Appearance, Language-Mode, Chracter Pattern, Privacy, Toxicity Control, Truthfulness (Mueller et al., 2022; Chen et al., 2023; Wu et al., 2023; Tang et al., 2024; Gurnee et al., 2024; Suau et al., 2024; Song et al., 2024; Li et al., 2025), which can be categorized into activation-based, causal-based, and gradient-based. However, these methods focus only on effect of the good neurons, ignoring the role of the bad neurons.

How we address these two challenges:



ⓐ Augmented Question-Answering (AQUA) ⓑ Contrastive Neuron Identification (CNI)

1. Role Specification:

You are an excellent linguist. Your task is to identify the entity type of a given word or phrase in a sentence (Person, Organization, Location, Product, Event, Art, Other, Building).

2. Rule Explanation:

When you are given a multiple-choice question. Respond with the letter which corresponds to the correct answer, followed by a period. There is no need to provide an explanation, so your response should be very short.

3. One-Shot Example:

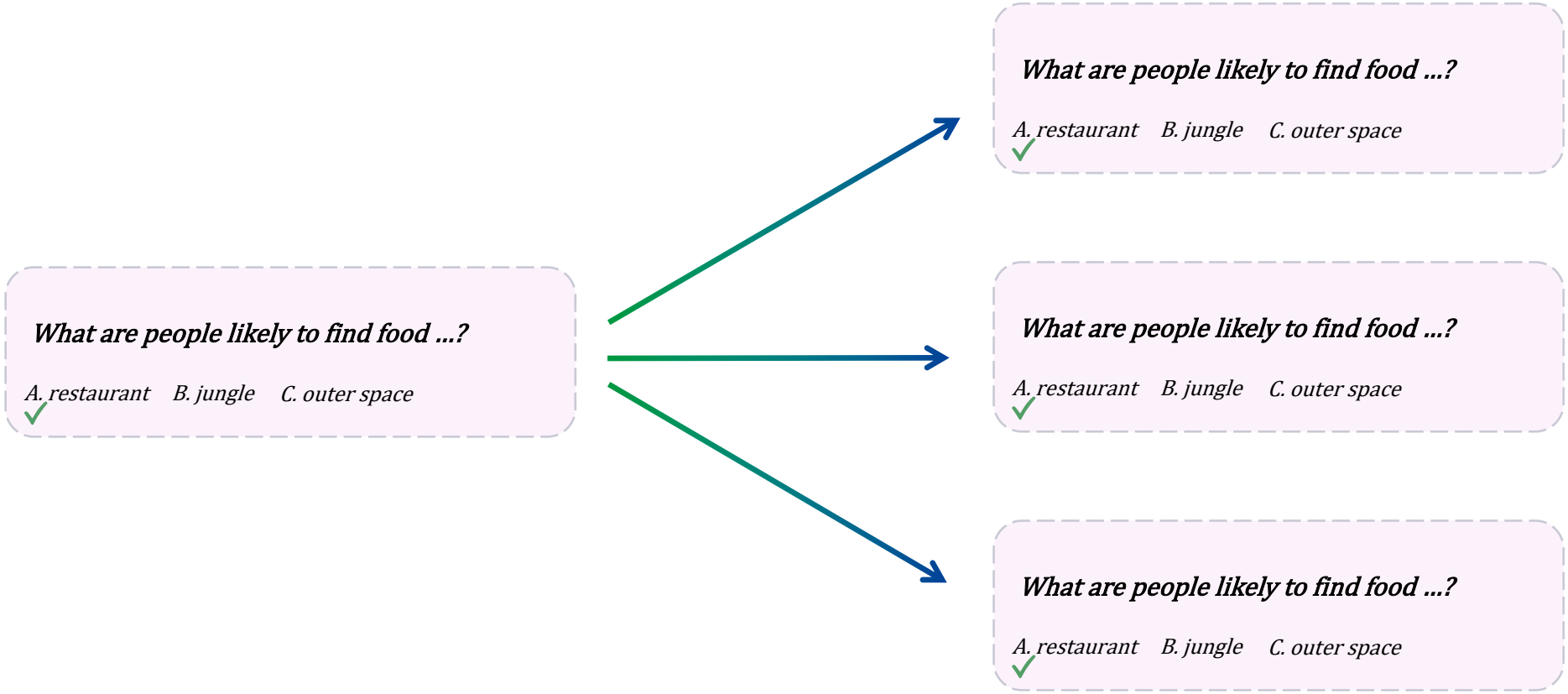
Here is an example: (In this example, 'Kong' is clearly the district administrator, so it should be referred to as a person. Therefore the correct answer is D)\n What is the entity type of 'Kong' in the following sentence:\n 'Oknha Son Kuy had 5 close associates in arms : Phuchhuoy (or District Administrator) Kong, Mr.Meun Ek, Mr.Ta Mong, Mr.Tesa Saom (some called him Ansa Saom) and Mr.Ta Mono Ros.'\n Options:\n A. other B. organization C. art D. person\n Correct answer: D

4. Question Stem:

Now here is the question: What is the entity type of 'Sandy Koufax' in the following sentence:\n 'The Yankees easily reached the 1963 World Series when they won the pennant by 10.5 games, but they scored only four runs in the series and were swept by the Los Angeles Dodgers and their ace pitcher, Sandy Koufax.'

5. Optional Answers:

A. organization B. product C. event D. person\n Correct answer:

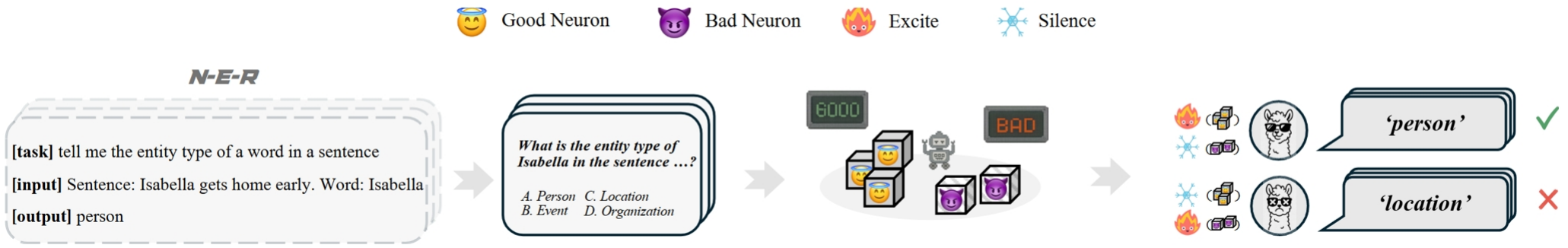


$$\mathrm{ES}_e(w_i^l) := \sum_{t=1}^{3} \frac{\hat{w}_i^l}{m} \sum_{k=1}^{m} \frac{\partial P(c^* | \frac{k\hat{w}_i^l}{m}, p_t)}{\partial w_i^l}. \qquad (4)$$

Evaluate by Neuron Intervention

To validate the effectiveness of the identified neurons, we adopt classic intervention approaches from neuroscience (Wiegert et al., 2017): given a query and the response value at a neuron w_i^l, we either: i) silence the neuron by zeroing out it via $w_i^l = 0$, or ii) excite the neuron by doubling its value $w_i^l = 2 \times \hat{w}_i^l$. Note that since we focus on neuron identification in this work, we use only these simple interventions to facilitate a straightforward evaluation of the identified neurons.

The goal of neuron intervention is to either enhance or degrade task performance. If neurons are correctly identified, exciting good neurons should enhance performance, while silencing them should degrade it. Unlike existing methods that ignore the bad neurons, NeuronLLM can leverage the interaction between the good and bad neurons via a joint intervention operator: **enhancer** that excites good + silences bad; **degrader** that silences good + excites bad. Evaluation of these neuron interventions would provide empirical evidence for *functional antagonism* inside LLM neurons.



Metrics

The following two metrics on the relative change of LLMs in the task performance are used: Relative Accuracy Change (**RAC**) and Relative Comprehension Change (**RCC**). RAC is defined as the relative change of an accuracy (Acc) measure before and after intervention:

$$RAC = \frac{|Acc_{original} - Acc_{intervened}|}{Acc_{original}} \times 100\%, \tag{6}$$

where Acc is calculated over the transformed proxy QAs. RCC measures the change of the comprehension (Com) ability.We say the LLM really understands the original question only if it can answer at least two of its three proxy QAs correctly. This helps avoid the measure be affected by cases that model gets right by chance. Formally, we define RCC as:

$$RCC = \frac{|Com_{original} - Com_{intervened}|}{Com_{original}|} \times 100\%, \tag{7}$$

Table 1: RAC/RCC results (%) of NeuronLLM and competing methods across four NLP tasks. 'Deg' and 'Enh' refer to neuron intervention to purposely degrade and enhance the task performance, respectively (see Section 3.5). Larger RAC/RCC values indicate better performance in degrading/enhancing the LLMs. **Red** highlights the best performance per metric, **Blue** shows the second best, and Fail indicates that the intervention produced the opposite effect.

	NER		Chunking		Sentiment		Commonsense		Average	
	Deg	Enh	Deg	Enh	Deg	Enh	Deg	Enh	Deg	Enh
LLaMA 2-7B										
NeuronLLM	**53.3/64.0**	**25.6/46.0**	**35.2/60.0**	**7.8/4.0**	**66.9/80.0**	**24.3/46.0**	**50.3/62.0**	**8.9/28.0**	**51.4/66.5**	**16.7/31.0**
TN	47.8/44.0	13.3/**34.0**	**17.2/32.0**	**6.3/4.0**	**63.9/78.0**	**10.7/24.0**	**9.5/0.0**	**5.3/12.0**	**34.6/38.5**	**8.9/18.5**
QRNCA	**48.9/46.0**	**13.9/34.0**	9.4/16.0	3.9/2.0	60.4/70.0	7.1/16.0	Fail	2.4/8.0	30.3/31.5	6.8/15.0
KN	23.7/20.0	10.1/20.0	9.4/18.0	5.5/2.0	16.1/12.5	5.7/5.0	Fail	2.8/7.5	12.8/11.4	6.0/8.6
ACT	0.0/0.0	0.0/0.0	1.0/0.0	0.0/0.0	Fail	0.0/0.0	0.0/0.0	0.0/0.0	Fail	0.0/0.0
RANDOM	Fail	0.7/0.0	0.0/0.0	Fail	Fail	2.4/5.0	Fail	0.7/0.0	Fail	0.7/1.3
Baichuan 2-7B										
NeuronLLM	**63.6/73.6**	**25.8/23.6**	**50.3/64.9**	**15.1/12.3**	**46.0/51.7**	**40.4/29.3**	**56.7/74.6**	**10.0/10.4**	**54.2/66.2**	**22.8/18.9**
TN	**7.2/9.7**	12.4/13.8	**47.2/59.6**	**8.8/10.5**	3.7/1.7	**11.2/1.7**	7.0/6.0	**1.5/4.5**	16.3/19.3	**8.5/7.6**
QRNCA	2.9/2.8	12.4/12.5	**47.2/59.6**	Fail	5.6/5.2	9.3/**1.7**	18.9/23.9	Fail	18.7/22.9	Fail
KN	6.2/5.6	**13.9/15.3**	**47.2/59.6**	3.1/3.5	**10.6/8.6**	Fail	**30.9/34.3**	Fail	**23.7/27.0**	3.8/3.8
ACT	Fail	0.0/0.0	0.0/0.0	0.0/0.0	2.0/0.0	Fail	0.0/0.0	Fail	0.4/0.0	Fail
RANDOM	0.0/0.0	0.0/0.0	Fail	1.8/0.0	Fail	0.3/0.0	0.0/0.0	0.0/0.0	Fail	0.5/0.0
LLaMA 2-13B										
NeuronLLM	**32.6/33.3**	**10.0/6.7**	**28.8/46.7**	**15.9/11.1**	**36.6/41.8**	2.9/0.0	**33.8/37.9**	**8.1/10.6**	**33.0/40.0**	**9.2/7.1**
TN	Fail	7.2/**6.7**	**15.2/20.0**	**12.1/15.6**	Fail	**5.2/3.6**	**6.1/9.1**	2.0/**1.5**	4.6/**6.1**	**6.6/6.9**
QRNCA	Fail	7.2/**6.7**	12.1/11.1	9.9/**11.1**	Fail	3.5/1.8	5.1/**9.1**	**3.0/1.5**	4.0/4.3	5.9/5.3
KN	**9.1/5.3**	**8.6**/5.3	9.9/13.3	7.6/8.9	**1.2/1.8**	**5.8/7.3**	1.5/1.5	Fail	**5.4**/5.5	5.8/5.0
ACT	0.9/0.0	0.9/1.3	0.0/0.0	1.5/0.0	0.0/0.0	0.6/0.0	0.0/0.0	0.0/0.0	0.2/0.0	0.8/0.3
RANDOM	0.0/0.0	0.0/0.0	1.5/2.2	2.3/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.4/0.6	0.6/0.0

More precisely, we are detecting “task-relevant” neurons
instead of “task-specific” neurons

The former one includes the latter one and imply all the important capabilities required for the task

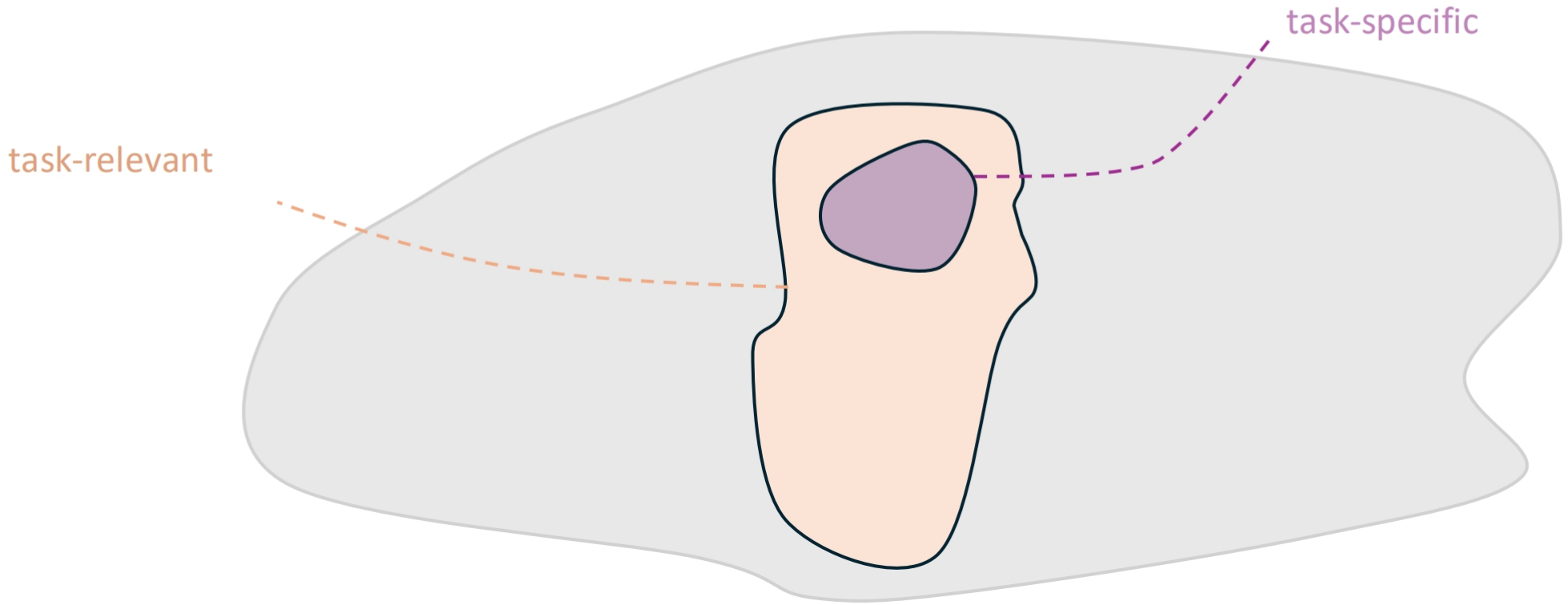


Table 5: The detailed impact of perturbing common ability neurons on each individual task for LLaMA 2-7B, Baichuan 2-7B and LLaMA 2-13B models.

LLaMA 2-7B					
Operation	NER	Chunking	Sentiment	ComSense	AVERAGE
Deg	42.8/40.0	32.0/50.0	47.3/54.0	32.0/28.0	**38.5/43.0**
Enh	19.4/38.0	4.7/4.0	15.4/30.0	5.3/16.0	**11.2/22.0**
Baichuan 2-7B					
Operation	NER	Chunking	Sentiment	ComSense	AVERAGE
Deg	53.6/62.5	47.2/59.6	49.1/74.1	62.19/73.1	**53.0/67.3**
Enh	13.9/13.9	15.1/12.3	34.2/24.1	12.4/16.4	**18.9/16.7**
LLaMA 2-13B					
Operation	NER	Chunking	Sentiment	ComSense	AVERAGE
Deg	9.1/8.0	10.6/20.0	22.7/23.7	15.7/16.7	**14.5/17.1**
Enh	2.3/1.3	10.6/11.1	4.1/5.5	4.6/7.6	**5.4/6.4**


NER
Chunking
Sentiment
ComSense
Degrade
Enhance
LLaMA 2-7B
Baichuan 2-7B
LLaMA 2-13B

Full Ablation Study

LLaMA 2-7B

Method	Eval	NER			Chunking			Sentiment			Commonsense			AVERAGE		
		Good	Bad	Both	Good	Bad	Both	Good	Bad	Both	Good	Bad	Both	Good	Bad	Both
NeuronLLM	Deg	38.9/34.0	45.0/50.0	**53.3/64.0**	30.5/48.0	28.1/44.0	**35.2/60.0**	65.7/78.0	**66.9/80.0**	**66.9/80.0**	33.1/34.0	38.5/44.0	**50.3/62.0**	42.1/48.5	44.6/54.5	**51.4/66.5**
	Enh	24.4/46.0	10.0/26.0	**25.6/46.0**	5.5/4.0	Fail	**7.8/4.0**	20.1/44.0	**24.3/46.0**	**24.3/46.0**	7.1/24.0	8.9/22.0	**8.9/28.0**	14.3/29.5	10.8/22.5	**16.7/31.0**
TN-enabled	Deg	**47.8/44.0**	28.3/24.0	**47.8/44.0**	17.2/32.0	21.9/34.0	**25.8/38.0**	63.9/78.0	59.2/70.0	**66.3/78.0**	9.5/0.0	7.7/6.0	**18.9/18.0**	34.6/38.5	29.3/33.5	**39.7/44.5**
	Enh	13.3/34.0	15.0/34.0	**17.2/36.0**	**6.3/4.0**	Fail	**6.3/4.0**	10.7/24.0	**24.3/44.0**	**24.3/44.0**	5.3/12.0	5.3/14.0	**5.3/14.0**	8.9/18.5	10.0/21.5	**13.3/24.5**
QRNCA-enabled	Deg	**48.9/46.0**	31.1/28.0	**48.9/46.0**	9.4/16.0	18.8/30.0	**21.1/32.0**	60.4/70.0	42.6/46.0	**65.0/78.0**	Fail	6.5/6.0	**6.5/6.0**	30.3/31.5	24.8/27.5	**35.4/40.5**
	Enh	13.9/34.0	15.0/34.0	**16.7/38.0**	3.9/2.0	Fail	**6.3/2.0**	7.1/16.0	**18.9/32.0**	**18.9/32.0**	2.4/8.0	5.3/16.0	**5.3/16.0**	6.8/15.0	8.8/19.5	**11.8/22.0**

LLaMA 2-13B

Method	Eval	NER			Chunking			Sentiment			Commonsense			AVERAGE		
		Good	Bad	Both	Good	Bad	Both	Good	Bad	Both	Good	Bad	Both	Good	Bad	Both
NeuronLLM	Deg	29.9/29.3	28.1/26.7	**32.6/33.3**	29.6/40.0	32.6/40.0	**28.8/46.7**	6.4/3.6	22.1/20.0	**36.6/41.8**	25.3/28.8	21.2/25.8	**33.8/37.9**	22.8/25.4	26.0/28.1	**33.0/39.9**
	Enh	**10.0/6.7**	2.7/2.7	**10.0/6.7**	**15.9/11.1**	3.8/4.4	**15.9/11.1**	2.9/0.0	0.6/0.0	**2.9/0.0**	3.0/1.5	7.6/6.1	**8.1/10.6**	7.9/4.8	3.7/3.3	**9.2/7.1**
TN-enabled	Deg	Fail	10.9/12.0	**10.9/13.3**	15.2/20.0	**31.8/44.4**	**31.8/44.4**	Fail	**4.1/1.8**	**4.1/1.8**	6.1/9.1	5.6/10.6	**14.1/19.7**	4.6/6.1	13.1/17.2	**15.2/19.8**
	Enh	**7.2/6.7**	3.2/1.3	**7.2/6.7**	12.1/15.6	15.2/8.9	**16.7/20.0**	5.2/3.6	Fail	**3.5/7.3**	2.0/1.5	3.5/3.0	**4.0/3.0**	6.6/6.9	5.5/2.4	**7.9/9.3**
QRNCA-enabled	Deg	Fail	10.9/12.0	**11.8/14.7**	12.1/11.1	**31.8/44.4**	**31.8/44.4**	Fail	**2.9/3.6**	**2.9/3.6**	5.1/9.1	6.1/10.6	**11.1/19.7**	4.0/4.3	12.9/17.7	**14.4/20.6**
	Enh	**7.2/6.7**	3.2/1.3	**7.2/6.7**	9.9/11.1	13.6/8.9	**13.6/13.3**	3.5/1.8	Fail	**2.3/5.5**	3.0/1.5	4.0/3.0	**5.1/6.1**	5.9/5.3	5.2/2.4	**7.1/7.9**

Baichuan 2-7B

Method	Eval	NER			Chunking			Sentiment			Commonsense			AVERAGE		
		Good	Bad	Both	Good	Bad	Both	Good	Bad	Both	Good	Bad	Both	Good	Bad	Both
NeuronLLM	Deg	34.0/40.3	24.4/27.8	**63.6/73.6**	47.2/59.6	23.3/31.6	**50.3/64.9**	**46.0/51.7**	39.8/55.2	**46.0/51.7**	**56.7/74.6**	27.9/28.4	**56.7/74.6**	46.0/56.6	28.9/35.8	**54.2/66.2**
	Enh	24.4/20.8	19.1/15.3	**25.8/23.6**	10.7/7.0	15.7/8.8	**15.1/12.3**	36.0/25.9	32.3/19.0	**40.4/29.3**	6.0/9.0	**10.0/10.4**	**10.0/10.4**	19.3/15.7	19.3/13.4	**22.8/18.9**
TN-enabled	Deg	7.2/9.7	10.0/12.5	**22.0/23.6**	47.2/59.6	44.7/56.1	**48.4/59.6**	3.7/1.7	8.7/15.5	**32.9/44.8**	7.0/6.0	7.5/7.5	**30.3/32.8**	16.3/19.3	17.7/22.9	**33.4/40.2**
	Enh	12.4/13.8	23.4/16.7	**23.9/19.4**	8.8/10.5	**15.7/19.3**	**15.7/19.3**	11.2/1.7	**28.0/13.8**	**28.0/13.8**	1.5/4.5	7.0/7.5	**8.5/7.5**	8.5/7.6	18.5/14.3	**19.0/15.0**
QRNCA-enabled	Deg	2.9/2.8	12.9/13.9	**28.7/29.2**	**47.2/59.6**	40.9/45.6	**47.2/59.6**	5.6/5.2	18.0/27.6	**38.5/51.7**	18.9/23.9	23.4/25.4	**22.9/26.9**	18.7/22.9	23.8/28.1	**34.3/41.9**
	Enh	12.4/12.5	**23.9/18.1**	**23.9/18.1**	Fail	Fail	**8.8/7.0**	9.3/1.7	**19.3/6.9**	**19.3/6.9**	Fail	6.0/4.5	**5.5/7.5**	Fail	12.2/6.5	**14.4/9.9**

The sensitivity of intervention budget

Table 7: Sensitivity analysis of the intervention budget K. Results show RAC/RCC for both enhancement (Enh) and degradation (Deg) performance across varying intervention budgets from 10 to 500 neurons (The task here is Commonsense Reasoning and the model is LLaMA 2-7B). NeuronLLM consistently outperforms competing SOTA methods TN and QRNCA across all budget settings, demonstrating its significant robustness to hyperparameter settings. Notably, NeuronLLM achieves with only 10 neurons the same control effectiveness that competing methods require 10× more neurons to attain (such superiority can also be observed in the results of NeuronLLM at budget 25 and 50 vs. TN/QRNCA at budget 250 and 500), demonstrating the effectiveness of NeuronLLM in identifying task-relevant neurons. For NeuronLLM, control effectiveness substantially improves from budget 10 to 100, then exhibits diminishing returns. In contrast, baseline methods show slower and more unstable improvement patterns, with occasional failed control.

	Intervention Budget							
	10		25		50		100	
	Enh	Deg	Enh	Deg	Enh	Deg	Enh	Deg
NeuronLLM	**7.1/18.0**	**17.2/12.0**	**5.9/22.0**	**32.2/30.0**	**6.5/26.0**	**42.6/52.0**	**8.9/28.0**	**50.3/62.0**
TN	1.2/4.0	Fail	0.6/6.0	2.4/0.0	4.7/10.0	Fail	5.3/12.0	9.5/0.0
QRNCA	1.2/6.0	0.0/0.0	0.6/6.0	2.4/0.0	3.0/6.0	Fail	2.4/8.0	Fail
	150		200		250		500	
	Enh	Deg	Enh	Deg	Enh	Deg	Enh	Deg
NeuronLLM	**11.2/28.0**	**52.1/64.0**	**11.2/28.0**	**52.7/66.0**	**10.7/26.0**	**52.1/64.0**	**10.1/24.0**	**52.7/66.0**
TN	5.3/16.0	20.1/14.0	4.7/14.0	25.4/20.0	4.1/18.0	22.5/16.0	5.9/22.0	32.5/30.0
QRNCA	3.0/6.0	Fail	1.2/6.0	13.0/0.0	3.0/12.0	17.8/4.0	4.7/16.0	29.6/24.0

The distribution of task-relevant neurons: