Capability Localization: Capabilities Can be Localized rather than Individual Knowledge

Xiusheng Huang[1,2,3], Jiaxiang Liu[1,2], Yequan Wang[3], Jun Zhao[1,2], Kang Liu[1,2]

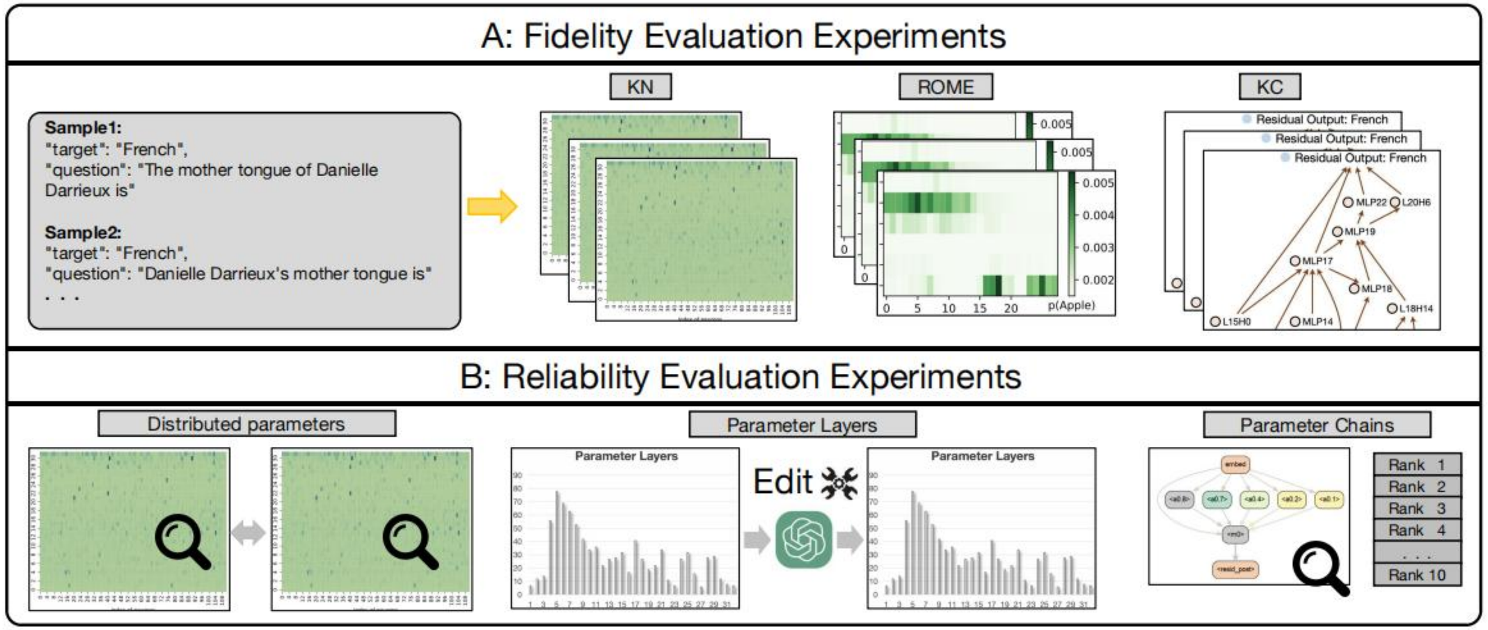
[1]The Key Laboratory of Cognition and Decision Intelligence for Complex Systems,
Institute of Automation, Chinese Academy of Sciences

[2]School of Artificial Intelligence, University of Chinese Academy of Sciences

[3]Beijing Academy of Artificial Intelligence, Beijing, China

Background

- **Recent research centers on localizing individual knowledge：**
  - > There are three locating methods :
    - ~ Distributed parameters (KN), parameter layer (ROME) and parameter chain (KC).
  - > Hownver, existing technologies cannot localize individual knowledge parameters.
  - > We will introduce fidelity and reliability evaluation experiments:



Sample1:
"target": "French",
"question": "The mother tongue of Danielle Darrieux is"

Sample2:
"target": "French",
"question": "Danielle Darrieux's mother tongue is"
. . .

Fig 1. Overall framework diagram of evaluation experiments.

Evaluation

- **In our fidelity and reliability evaluation experiments:**

> The fidelity experiment is mainly aimed at verifying the fidelity of existing individual knowledge localization methods.

~ We evaluate each method's *overlap* score by rewriting input prompt.

$$overlap = \frac{\frac{|a \cap b|}{|a|} + \frac{|a \cap b|}{|b|}}{2}$$

~ As for, KN, ROME, and KC, their overlap rates are: 37.3%, 32.7%, and 7.2%

- **Existing knowledge localization methods are not faithful to knowledge**

Evaluation

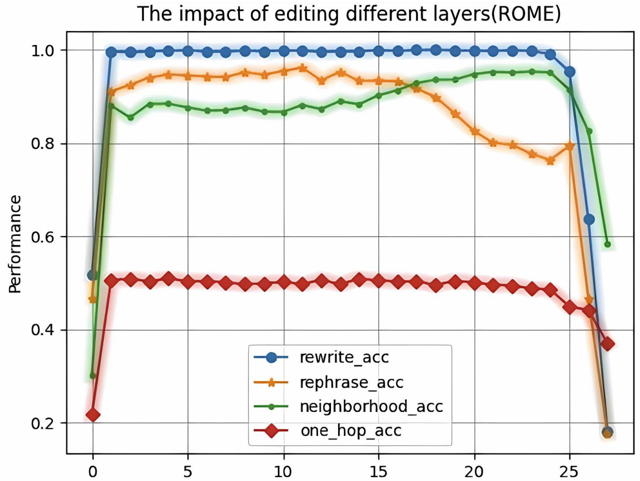
- In our fidelity and reliability evaluation experiments：

> Reliability experiments for locating methods is aimed at verifies the effectiveness of the localization parameters.

~ We evaluate each method's *IPP* score using their separate validation methods.

$$performance(param_{located}) - performance_{optimal}(param \in random)$$

・For instance, in the ROME method, changing layers other than the located ones produces similar effects. However, ROME regards the parameters of the fifth layer as strongly associated with the data.

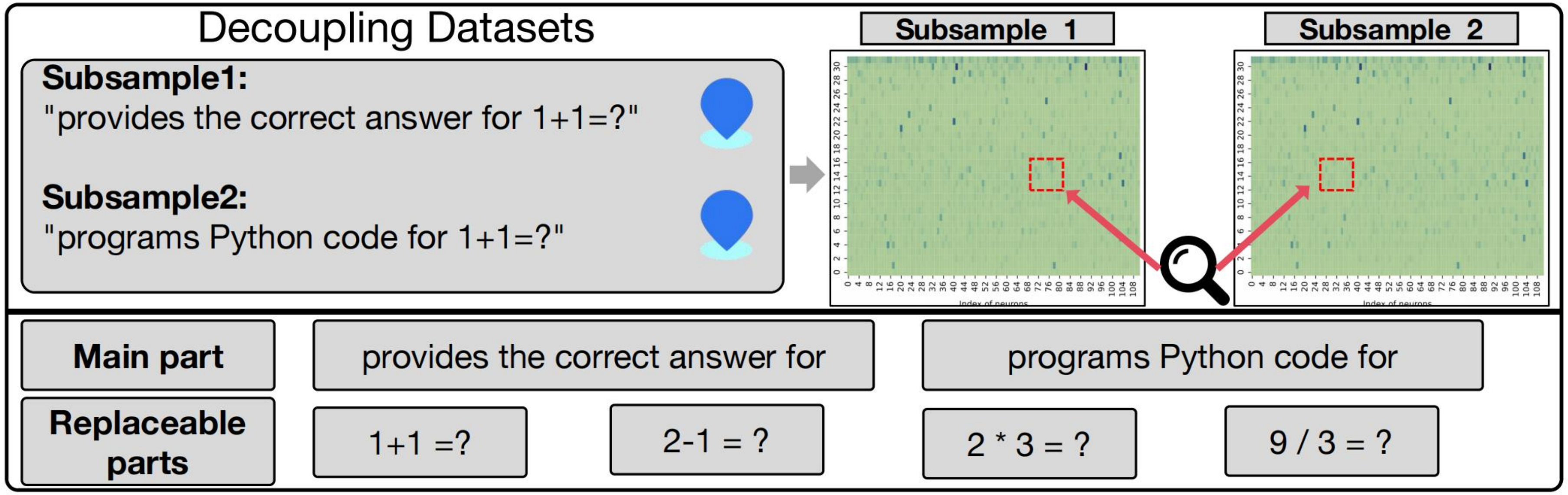


~ As for, KN, ROME, and KC, their IPPs are: 11.4%, 0.0% and 10.5%

- Existing knowledge localization methods are unreliable

Decoupling

- In order to further reveal the form of knowledge storage
  - > We designe decoupling experiments
    - ~ We designed 1000 comparative samples with two sub samples
    - ~ Their main parts remain consistent but differ in task replaceable parts
    - ~ We utilize mathematical calculation formulas as the replaceable parts



Decoupling Datasets

Subsample1:
"provides the correct answer for 1+1=?"

Subsample2:
"programs Python code for 1+1=?"

Subsample 1

Subsample 2

Main part | provides the correct answer for | programs Python code for

Replaceable parts | 1+1 =? | 2-1 = ? | 2 * 3 = ? | 9 / 3 = ?

Fig 2. Overall framework diagram of decoupling experiments.

Decoupling

- In order to further reveal the form of knowledge storage
  - > We designe decoupling experiments
    - ~ We use KN for parameters localization and calculate the coincidence rate:

$$C_r^{t,t^*} = \begin{cases} \frac{1}{n\cdot\mathcal{L}\cdot\mathcal{J}} \sum_{k=1}^{n} (P_s|Attr(\omega^{l,j}|x_k^t)) \cap (P_s|Attr(\omega^{l,j}|x_k^{t^*})) & t \neq t^* \\ \frac{1}{\mathcal{L}\cdot\mathcal{J}} \bigcap_{k=1}^{n} (P_s|Attr(\omega^{l,j}|x_k^t)) & t = t^* \end{cases} \quad t, t^* \in \{1, 2\}$$

~ The $C_r^{1,2}$, $C_r^{1,1}$ and $C_r^{2,2}$are 15.6%, 7.3% and 8.6%

~ Individual knowledge is hard to achieve parameter localization, and there is a certain commonality within the entire dataset.

- Can the commonality of data be localized?

Commonality Neurons Locating

- We expand the KN method to compute the contribution of the neuron
  - > We calculate the contribution score as:
    - ~ Under the dataset $D = \{([x_1, \ldots, x_X], [y_1, \ldots, y_Y])\}$

$$Score(\omega^{l,j}) = \mathbb{E}_{(x,y)\in\mathcal{D}}\left[\frac{1}{Y}\frac{1}{S}\sum_{m=1}^{Y}\overline{\omega^{l,j}_{Z_m}[z_m]}\sum_{n=0}^{S}\frac{\partial P_{z,y_m}(\frac{n}{S}\overline{\omega^{l,j}_{Z_m}[z_m]})}{\partial\omega^{l,j}_{Z_m}[z_m]}\right],$$

$$z_m = x \oplus y_{0:m-1}$$

- ~ To identify the commonality neuron

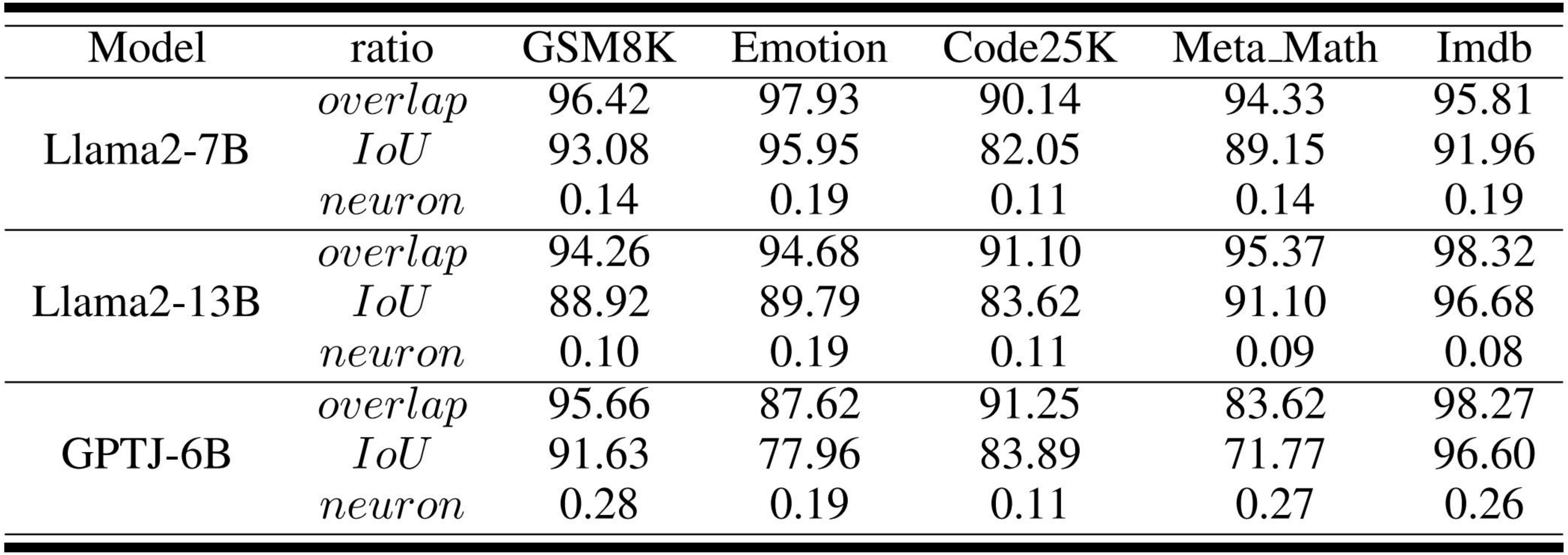
$$Mask_{l,j} = \begin{cases} 1 & \left|Score(\omega^{l,j}) - mean(Score(\omega))\right| > \sigma \cdot var(Score(\omega)) \\ 0 & else \end{cases}$$

Experiment

- **The experimental results demonstrates the effectiveness of CNL:**

  > Our experiment mainly has the following findings:

  ~ **Both overlap and IoU have high values, indicating a set of data has certain commonalities, which is reflected by the neurons.**



Model	ratio	GSM8K	Emotion	Code25K	Meta_Math	Imdb
Llama2-7B	*overlap*	96.42	97.93	90.14	94.33	95.81
	IoU	93.08	95.95	82.05	89.15	91.96
	neuron	0.14	0.19	0.11	0.14	0.19
Llama2-13B	*overlap*	94.26	94.68	91.10	95.37	98.32
	IoU	88.92	89.79	83.62	91.10	96.68
	neuron	0.10	0.19	0.11	0.09	0.08
GPTJ-6B	*overlap*	95.66	87.62	91.25	83.62	98.27
	IoU	91.63	77.96	83.89	71.77	96.60
	neuron	0.28	0.19	0.11	0.27	0.26

Table 1. Neurons overlap ratio and the proportion of targeted neurons.

Experiment

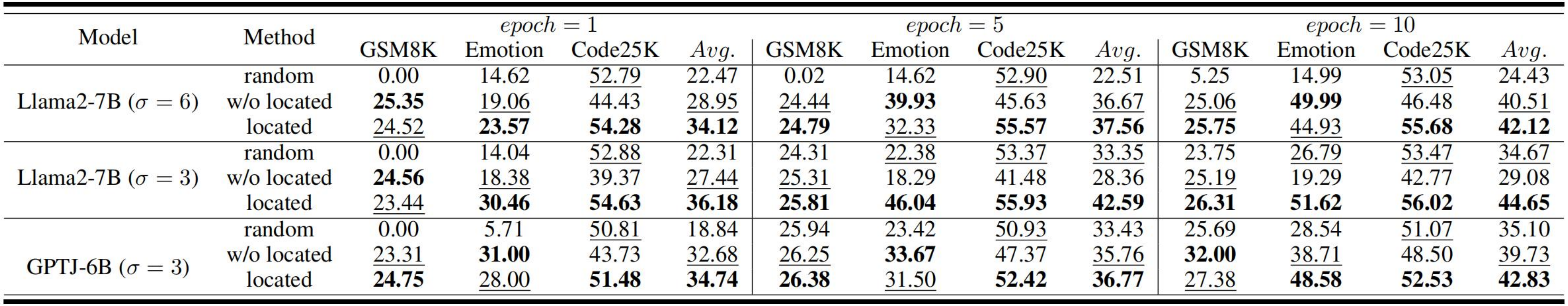
- The experimental results demonstrates the effectiveness of CNL:

  > Our experiment mainly has the following findings:

  ~ **We only need to update a small portion of parameters to effectively improve the performance of the current task.**

-> random (0.15% neurons) -> w/o located (99.85% neurons)

-> located (0.15% neurons)



Model	Method	*epoch* = 1				*epoch* = 5				*epoch* = 10			
		GSM8K	Emotion	Code25K	*Avg.*	GSM8K	Emotion	Code25K	*Avg.*	GSM8K	Emotion	Code25K	*Avg.*
Llama2-7B ($\sigma = 6$)	random	0.00	14.62	52.79	22.47	0.02	14.62	52.90	22.51	5.25	14.99	53.05	24.43
	w/o located	**25.35**	19.06	44.43	28.95	24.44	**39.93**	45.63	36.67	25.06	**49.99**	46.48	40.51
	located	24.52	**23.57**	**54.28**	**34.12**	**24.79**	32.33	**55.57**	**37.56**	**25.75**	44.93	**55.68**	**42.12**
Llama2-7B ($\sigma = 3$)	random	0.00	14.04	52.88	22.31	24.31	22.38	53.37	33.35	23.75	26.79	53.47	34.67
	w/o located	**24.56**	18.38	39.37	27.44	25.31	18.29	41.48	28.36	25.19	19.29	42.77	29.08
	located	23.44	**30.46**	**54.63**	**36.18**	**25.81**	**46.04**	**55.93**	**42.59**	**26.31**	**51.62**	**56.02**	**44.65**
GPTJ-6B ($\sigma = 3$)	random	0.00	5.71	50.81	18.84	25.94	23.42	50.93	33.43	25.69	28.54	51.07	35.10
	w/o located	23.31	**31.00**	43.73	32.68	26.25	**33.67**	47.37	35.76	**32.00**	38.71	48.50	39.73
	located	**24.75**	28.00	**51.48**	**34.74**	**26.38**	31.50	**52.42**	**36.77**	27.38	**48.58**	**52.53**	**42.83**

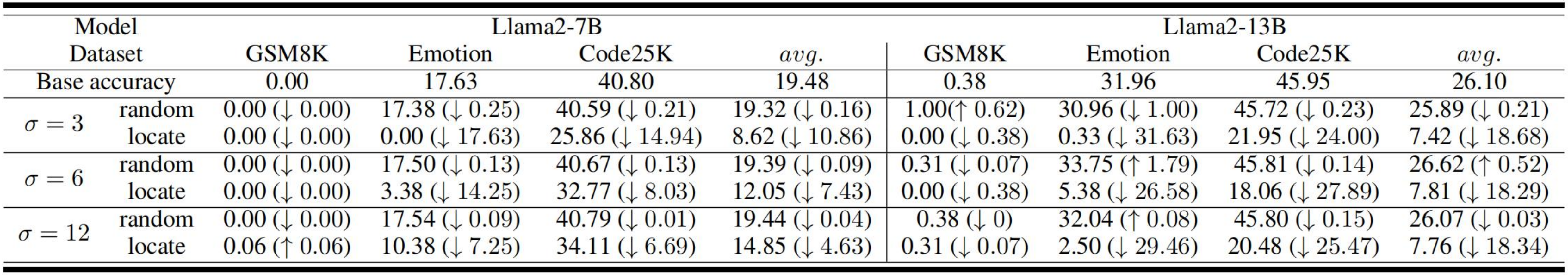
Table 2: Enhancement of different sets of neurons.

Experiment

- **The experimental results demonstrates the effectiveness of CNL:**

  > Our experiment mainly has the following findings:

  ~ **Compared to random, erasing the neurons we locate will significantly impair the performance of the model.**



Model		Llama2-7B				Llama2-13B			
Dataset		GSM8K	Emotion	Code25K	*avg.*	GSM8K	Emotion	Code25K	*avg.*
Base accuracy		0.00	17.63	40.80	19.48	0.38	31.96	45.95	26.10
$\sigma = 3$	random	0.00 (↓ 0.00)	17.38 (↓ 0.25)	40.59 (↓ 0.21)	19.32 (↓ 0.16)	1.00(↑ 0.62)	30.96 (↓ 1.00)	45.72 (↓ 0.23)	25.89 (↓ 0.21)
	locate	0.00 (↓ 0.00)	0.00 (↓ 17.63)	25.86 (↓ 14.94)	8.62 (↓ 10.86)	0.00 (↓ 0.38)	0.33 (↓ 31.63)	21.95 (↓ 24.00)	7.42 (↓ 18.68)
$\sigma = 6$	random	0.00 (↓ 0.00)	17.50 (↓ 0.13)	40.67 (↓ 0.13)	19.39 (↓ 0.09)	0.31 (↓ 0.07)	33.75 (↑ 1.79)	45.81 (↓ 0.14)	26.62 (↑ 0.52)
	locate	0.00 (↓ 0.00)	3.38 (↓ 14.25)	32.77 (↓ 8.03)	12.05 (↓ 7.43)	0.00 (↓ 0.38)	5.38 (↓ 26.58)	18.06 (↓ 27.89)	7.81 (↓ 18.29)
$\sigma = 12$	random	0.00 (↓ 0.00)	17.54 (↓ 0.09)	40.79 (↓ 0.01)	19.44 (↓ 0.04)	0.38 (↓ 0)	32.04 (↑ 0.08)	45.80 (↓ 0.15)	26.07 (↓ 0.03)
	locate	0.06 (↑ 0.06)	10.38 (↓ 7.25)	34.11 (↓ 6.69)	14.85 (↓ 4.63)	0.31 (↓ 0.07)	2.50 (↓ 29.46)	20.48 (↓ 25.47)	7.76 (↓ 18.34)

Table 3: Erase of different sets of neurons.

- **However, capability be an attribute that can span datasets.**

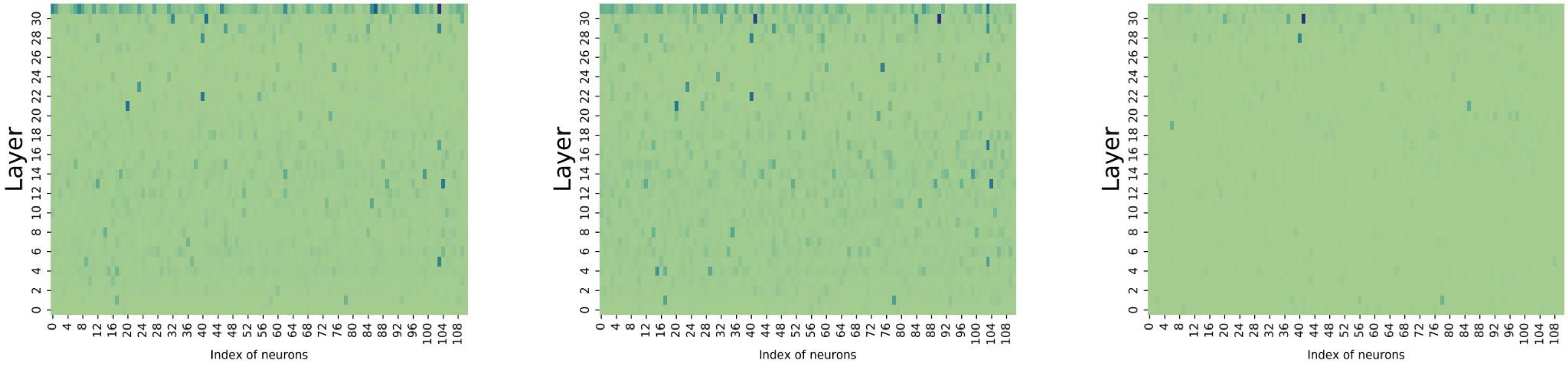
Experiment

- **However, capability be an attribute that can span datasets.**

  > Our cross-data experiment mainly proves:

  ~ **The located neurons embody the collection of capabilities**

  -> It seems that neurons under different datasets have overlaps, which means that the commonality neurons have cross dataset characteristics



(a) GSM8K

(b) Emotion

(c) Code25K

Fig 3. Visualisation of commonality neurons.

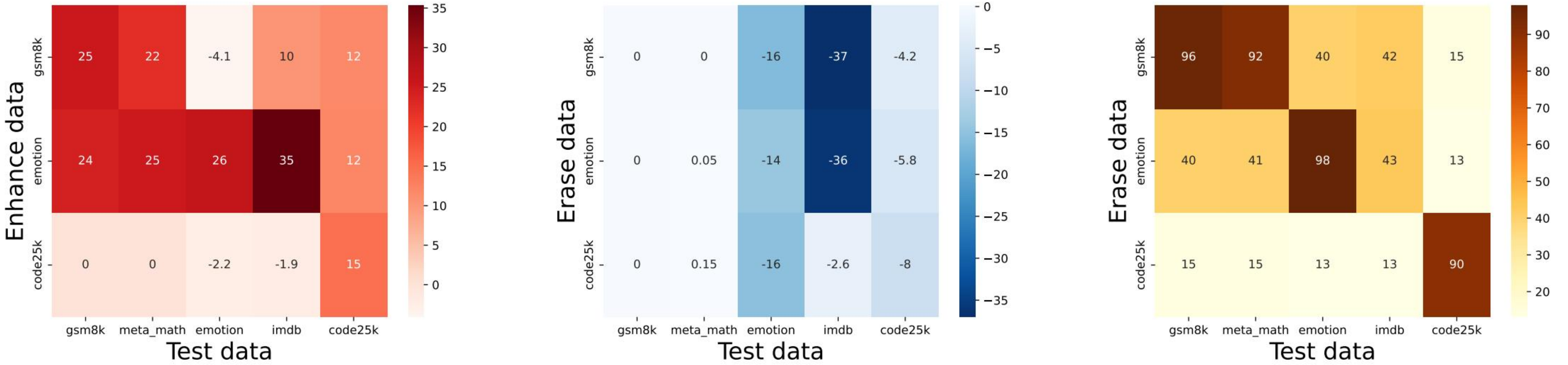
Experiment

- **However, capability be an attribute that can span datasets.**

  > Our cross-data experiment mainly proves:

  ~ **The located neurons embody the collection of capabilities**

  -> Enhancing and erasing capacity neurons can generate substantial effects on the same type of dataset, with little impact on other types of datasets.



(a) Neurons enhancement (b) Neurons erasing (c) Overlap ratio across dataset

Fig 4. Effects of capacity neurons on other dataset.

Thanks !