The 31st International Conference on Computational Linguistics
COLING 2025 • Abu Dhabi

Towards Understanding Multi-Task Learning (Generalization) of LLMs via Detecting and Exploring Task-Specific Neurons

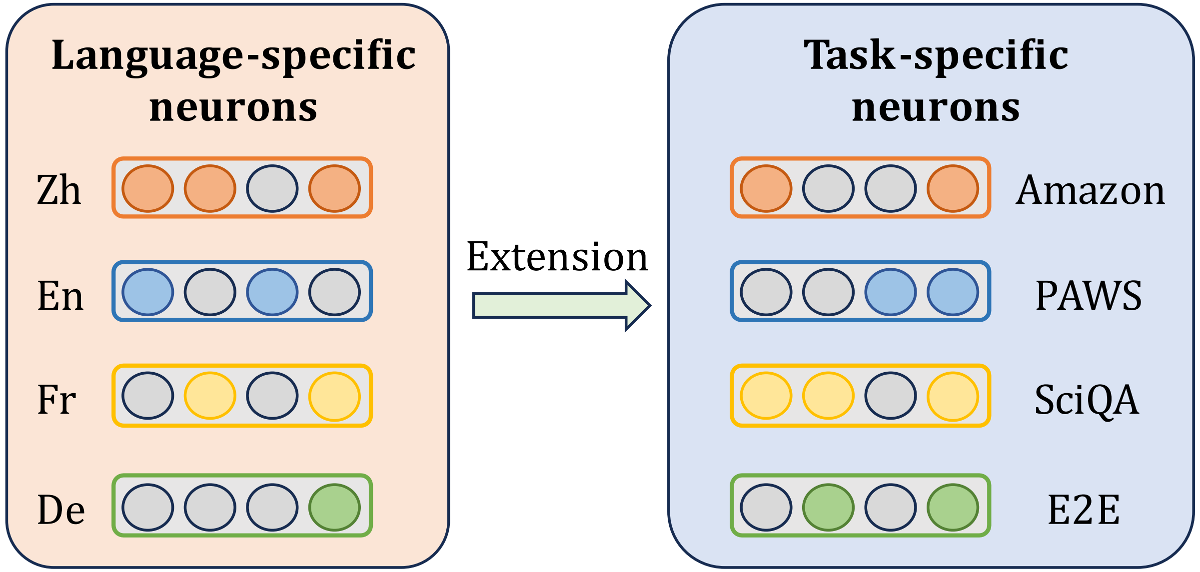
Yongqi Leng and Deyi Xiong*
College of Intelligence and Computing, Tianjin University, Tianjin, China
lengyq@tju.edu.cn

Yongqi Leng
Master student
@TJU

Deyi Xiong
Professor
@TJU

Motivation

1. Previous studies have demonstrated the existence of language-specific neurons in multilingual large language models (MLLMs), which have been explored to investigate the multilingual learning mechanisms. In contrast, research into the multi-task learning mechanisms of LLMs remains limited.
2. We argue that multilingual learning is essentially a type of multi-task learning as well.



💡Can we extend neuronal analysis from multilingual learning to multi-task learning in LLMs?

Research Questions

- Do task-specific neurons exist in LLMs?
- If they exist, can they facilitate the understanding of the multi task learning mechanisms in LLMs?
- Can we improve LLMs by exploring such neurons?

Methodology

Identification

Employing gradient attribution method to estimate each neuron's relevance score for a given task. Neurons with the top k% relevance scores are identified as task-specific neurons.

$$\mathcal{R}_j^i = |\Delta\mathcal{L}(\boldsymbol{\omega}_j^i)| = \left|\frac{\partial \mathcal{L}}{\partial \boldsymbol{\omega}_j^i}\boldsymbol{\omega}_j^i\right|$$

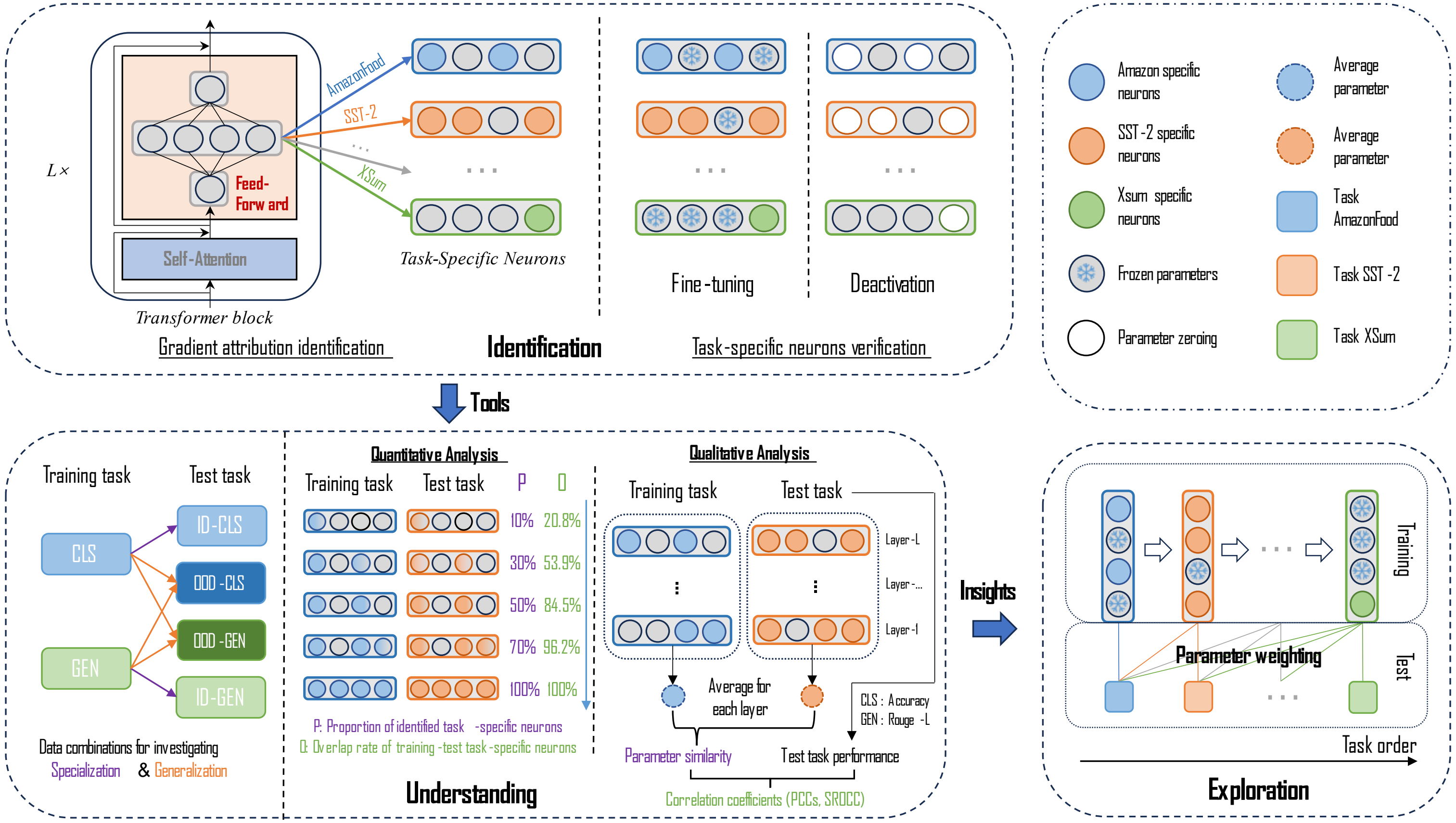
Understanding

Quantitative Analysis: Empirical study on specialization and generalization with varying task-specific neuron proportions.

Qualitative Analysis: Investigating generalization from the perspective of task-specific neuron parameter similarity.

Exploration

Neuron-Level Continuous Fine-tuning Method (NCFT): During the continuous training over the task sequence, only the neuron-specific parameters of the current task are updated, while other parameters are frozen.



The **Identification** component provides tools for the **Understanding** component which in turn provides insights for the **Exploration** component.

(Identification) Identifying Task-Specific Neurons

Experimental Setup

- Deactivation experiments.
- Fine-tuning experiments.

Model: Llama-2-7b
Hyper-parameter: $k = 10$
Dataset: classification and generation tasks

	Task	Dataset
CLS	Sentiment Classification	AmazonFood, SST-2
	Paraphrase Detection	QQP, Paws
	Natural Language Inference	MNLI, GPTNLI
GEN	Summary	CNN/DailyMail, Xsum
	Question Generation	Sciqa, Tweetqa
	Data to Text	E2E, CommonGen

Summary of tasks and datasets.

Method \ Task-CLS	AmazonFood	SST-2	QQP	Paws	MNLI	GPTNLI	Avg.
Original	91.8	92.4	83.2	91.6	84.8	82.4	87.7
Deactivate-Random	90.6	91.2	79.8	87.6	80.5	79.3	84.8
Deactivate-Task	**83.6**	**84.6**	**72.8**	**70.2**	**73.3**	**71.4**	**76.0**
Method \ Task-GEN	**Sciqa**	**Tweetqa**	**E2E**	**CommonGen**	**CNN/DailyMail**	**XSum**	**Avg.**
Original	54.3	45.6	52.6	49.8	34.7	36.8	45.6
Deactivate-Random	50.8	41.3	48.7	47.3	31.3	34.4	42.3
Deactivate-Task	**33.6**	**29.3**	**39.6**	**37.8**	**25.5**	**26.3**	**32.0**

Performance of Llama-2-7b after task-specific neurons deactivation or without deactivation in each task. "Original" is the performance after fine-tuning with multi-task data without any neurons being deactivated.

Method \ Task-CLS	AmazonFood	SST-2	QQP	Paws	MNLI	GPTNLI	Avg.
Zero-shot	85.2	78.3	42.1	46.5	35.3	32.4	53.3
Train-Random	85.5	80.3	45.6	47.8	34.7	34.8	54.8
Train-Task	**88.5**	**87.8**	**79.2**	**84.8**	**82.5**	**76.3**	**83.2**
Method \ Task-GEN	**Sciqa**	**Tweetqa**	**E2E**	**CommonGen**	**CNN/DailyMail**	**XSum**	**Avg.**
Zero-shot	21.3	6.9	36.5	26.8	14.7	12.3	19.8
Train-Random	22.8	11.8	37.4	29.6	17.7	15.8	22.5
Train-Task	**45.3**	**37.1**	**42.7**	**36.8**	**29.8**	**30.3**	**37.0**

Performance of Llama-2-7b after fine-tuning task-specific neurons and under the zero-shot setting.

(Understanding) Quantitative Analysis

COLING 2025 • Abu Dhabi
The 31st International Conference on Computational Linguistics

Experimental Setup

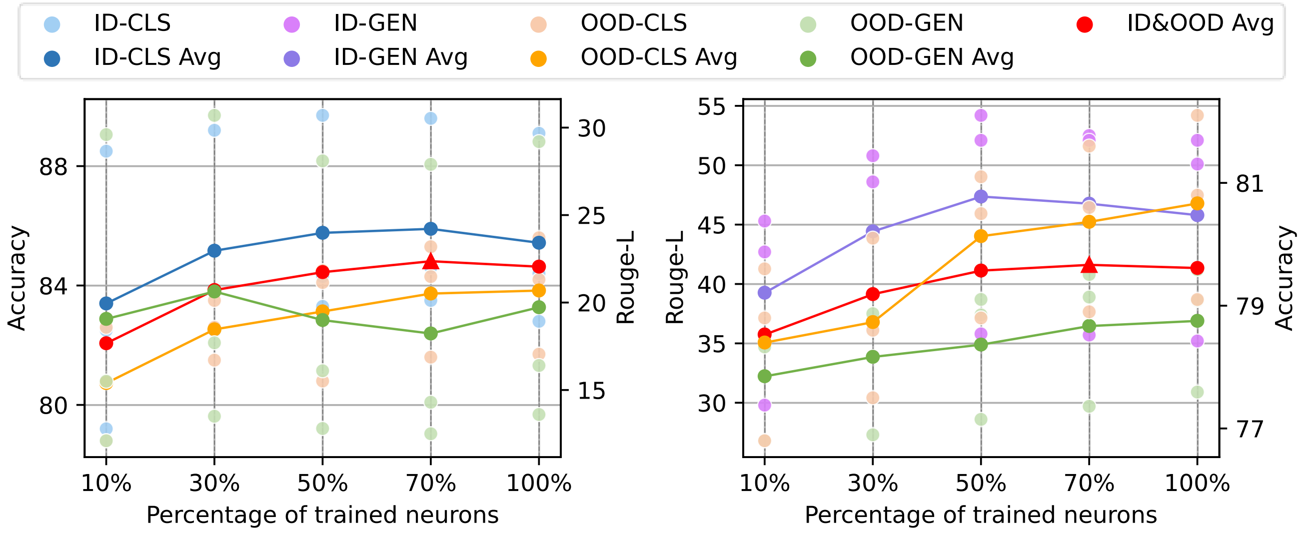
- We controlled the proportion of fine-tuned task-specific neurons to investigate the trends in specialization and generalization.
- Results from the **in-domain (ID)** test set indicate **specialization** performance while results from the **out-of-domain (OOD)** test set indicate **generalization** performance.

Findings on Specialization

- When training all parameters of the model under the multi-task learning setup, inevitable interference among tasks occurs, thereby diminishing the efficacy of individual tasks to some degree.
- Our experiments show the efficacy of controlling the proportion of fine-tuned task-specific neurons as a promising strategy.

Group	**Training Tasks**	**ID Test Tasks**	**OOD Test Tasks**
(a)	Amazon, QQP, MNLI	Amazon, QQP, MNLI	SST-2, Paws, GPTNLI Tweetqa, CommonGen, Xsum
(b)	Sciqa, E2E, CNN	Sciqa, E2E, CNN	SST-2, Paws, GPTNLI Tweetqa, CommonGen, Xsum

Experimental groups for exploring generalization and specialization.



(a) Training on classification tasks

(b) Training on generation tasks

Results on classification and generation tasks after fine-tuning different proportions of task-specific neurons.

(Understanding) Quantitative Analysis

Group	10%	30%	50%	70%	100%
CLS-CLS	20.8	53.9	84.5	96.2	100
CLS-GEN	12.9	41.6	71.5	83.5	100
GEN-CLS	11.8	40.2	69.3	81.8	100
GEN-GEN	21.6	52.5	82.0	94.3	100

The overlap rate of task-specific neurons between training tasks and test tasks when controlling the proportion of task-specific neurons.

Findings on Generalization

- Task-specific neurons overlap rates are consistent with generalization performance.
- We argue that the overlap of task-specific neurons contributes to transfer learning between tasks, ultimately resulting in consistently higher generalization performance.

Overlap rate \ Percentage of trained neurons	10%	30%	50%
10%	80.2	81.4	81.8
20.8%	*80.7*	-	-
30%	81.1	82.0	82.3
50%	81.5	82.3	82.8
53.9%	-	*82.5*	-
70%	82.0	82.7	83.0
84.5%	-	-	*83.1*
100%	82.2	83.1	83.6

Results at different fine-tuned neuron proportions (10%, 30%, 50%) controlling the overlap rate under the **classification-classification** combination.

Overlap rate \ Percentage of trained neurons	10%	30%	50%
10%	31.6	32.1	32.3
21.6%	*32.2*	-	-
30%	32.5	32.9	33.5
50%	32.7	33.4	33.8
52.5%	-	*33.8*	-
70%	32.9	34.0	34.1
82.0%	-	-	*34.9*
100%	33.1	34.4	35.1

Results at different fine-tuned neuron proportions (10%, 30%, 50%) controlling the overlap rate under the **generation-generation** combination.

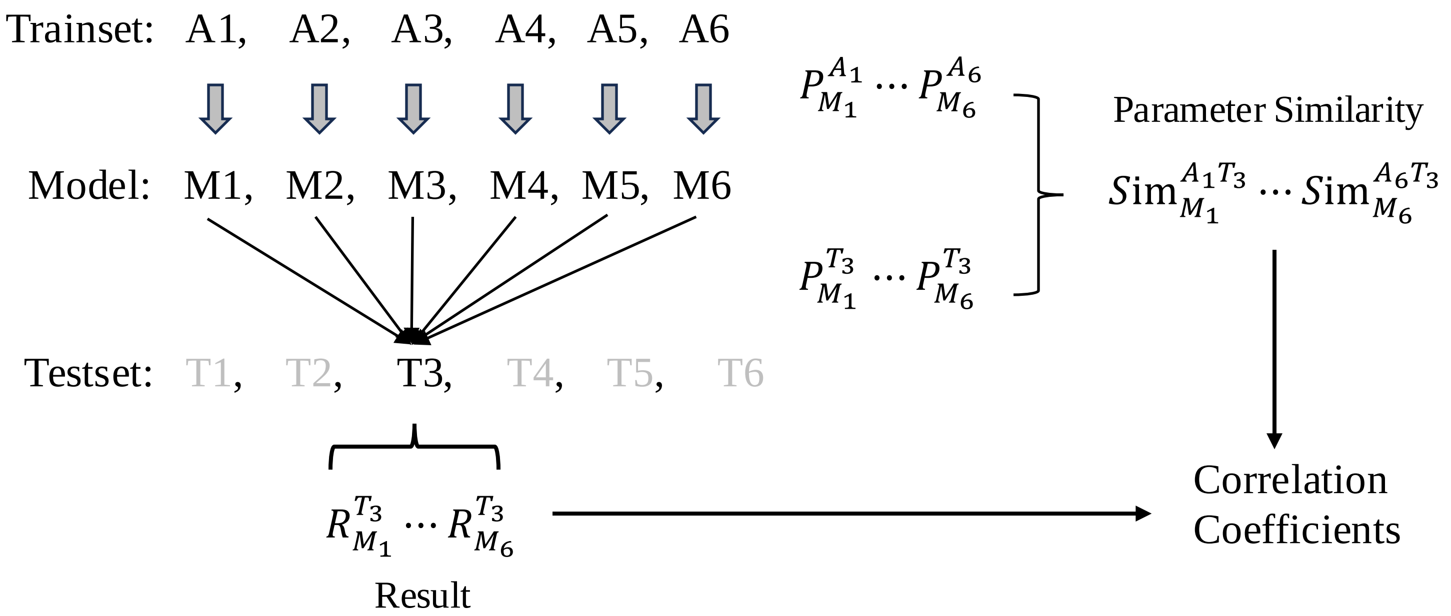
(Understanding) Qualitative Analysis

Experimental Setup

We calculate the correlation coefficients between the similarity of task-specific neuron parameters and the generalization performance.

Conclusion

These two show a positive correlation, reflecting the generalization between tasks from the perspective of parameter.



Testset	SST-2		Paws		GPTNLI		Tweetqa		CommonGen		Xsum	
	r	**p-value**	**r**	**p-value**	**r**	**p-value**	**r**	**p-value**	**r**	**p-value**	**r**	**p-value**
PCCs	0.87	0.02	0.92	0.01	0.79	0.05	0.96	0.00	0.96	0.00	0.97	0.00
SROCC	0.81	0.05	0.77	0.07	0.81	0.05	0.77	0.07	0.83	0.04	0.71	0.11

Correlation coefficients between the similarity of specific neuron parameters and generalization performance. PCCs denotes Pearson correlation coefficients and SROCC denotes Spearman correlation coefficients.

(Exploration) Mitigating Catastrophic Forgetting

Experimental Setup

- **Model**
  Llama-2-7b
- **Dataset**
  Standard CL Benchmark, Large Number of Tasks Benchmark
- **Metrics**
  (1) Performance on Continuous Learning (CL)

$$\text{CL} = \frac{1}{N}\sum_{i=1}^{N} a_{i,N}$$

(2) Forgetting Rate (FG)

$$\text{FG}_j = \frac{1}{j-1}\sum_{i=1}^{j-1} \frac{a_{i,j}}{A_i} \times 100\%$$

Method	Order-1	Order-2	Order-3	Avg.	Order-4	Order-5	Order-6	Avg.
SeqFT	46.4	47.3	47.5	47.1	35.6	34.8	33.5	34.6
SeqLoRA	53.6	54.8	53.1	53.8	47.9	49.5	45.7	47.7
EPI	48.1	48.0	49.0	48.4	42.3	41.8	43.6	42.6
O-LoRA	**76.8**	**75.7**	**75.7**	**76.1**	**73.7**	69.2	72.0	71.6
NCFT (Ours)	71.3	70.9	71.6	71.3	70.5	68.3	71.2	70.0
W-NCFT (Ours)	73.7	72.3	73.8	73.3	73.4	**70.1**	**72.6**	**72.0**
Per-Task FT	77.2	77.2	77.2	77.2	84.5	84.5	84.5	84.5

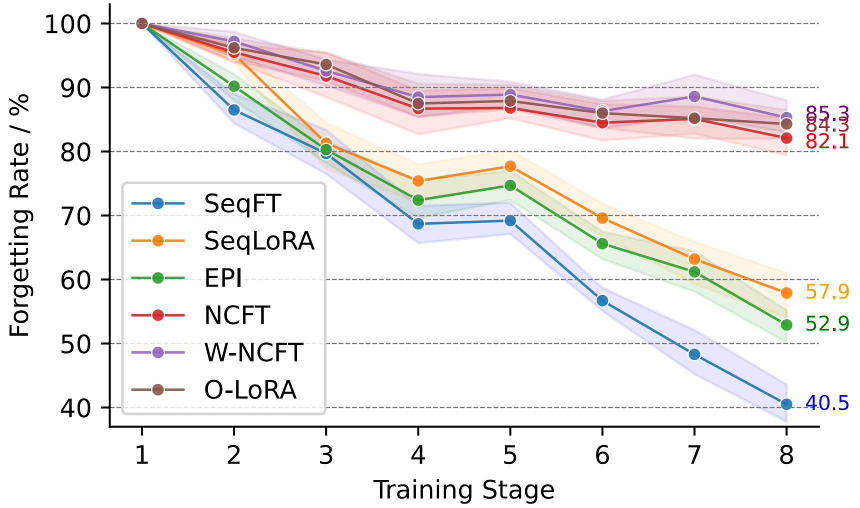
Results on two continual learning benchmarks. The average accuracy after training on the last task is reported.

Dataset	Class	Task Type	Domain
AGNews	4	Topic classification	News
Amazon	5	Sentiment anlysis	Amazon reviews
DBPedia	14	Topic classification	Wikipedia
Yahoo	10	Q&A	Yahoo Q&A

Details of the Standard CL Benchmark.

Dataset	Class	Task Type	Domain
Amazon	5	Sentiment anlysis	Amazon reviews
DBPedia	14	Topic classification	Wikipedia
Yahoo	10	Q&A	Yahoo Q&A
AGNews	4	Topic classification	News
MNLI	3	NLI	various
QQP	2	Paragraph detection	Quora
RTE	2	NLI	news, Wikipedia
SST-2	2	Sentiment analysis	movie reviews

Details of the simplified version Large Number of Tasks Benchmark.



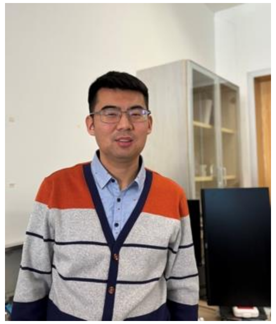
Forgetting rates for eight stages on the Large Number of Tasks benchmark.

Summary

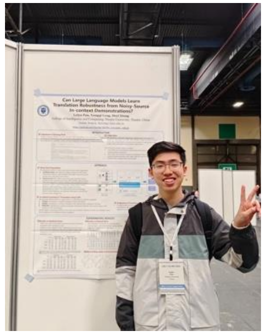
Main Contributions

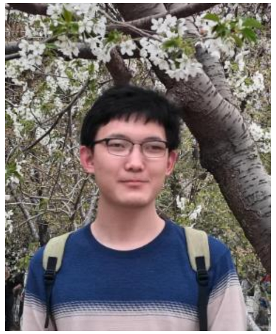
- We discover task-specific neurons in LLMs empirically through extensive experiments.
- We provide significant insights into generalization across tasks with our task-specific neuron analysis.
- We propose a neuron-level continuous learning fine-tuning method for mitigating catastrophic forgetting, and experiments demonstrate its effectiveness.

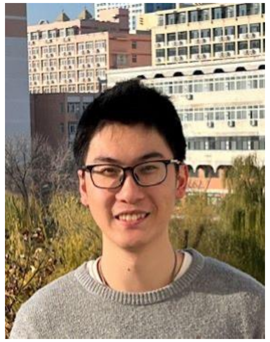
Thanks to my supervisor and labmates






https://github.com/sdsxdxl/task_neurons

https://tjunlp-lab.github.io