Neuron-Level Analysis of Cultural Understanding in Large Language Models

Taisei Yamamoto[1,2], Ryoma Kumon[1,2], Danushka Bollegala[3,4], Hitomi Yanaka[1,2,5]

[1]The University of Tokyo [2]Riken [3]University of Liverpool [4]Amazon [5]Tohoku University

https://github.com/ynklab/CULNIG

Summary

LLMs must possess **cultural understanding** to be deployed fairly around the world.

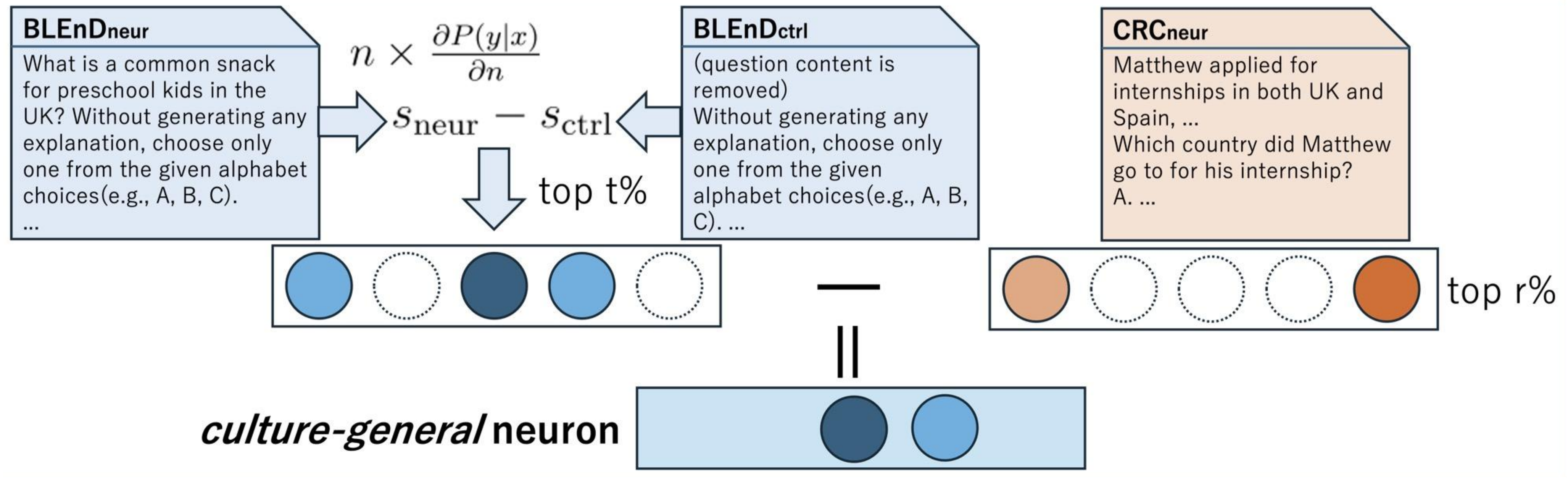
To elucidate mechanisms behind cultural understanding of LLMs, we identify

- ***Culture-general*** **neurons** contributing to cultural understanding regardless of cultures
- ***Culture-specific*** **neurons** tied to an individual culture

Evaluation results shows masking these neurons (**<1%** of all neurons concentrated in **shallow-mid MLPs**) degrade cultural understanding of LLMs significantly, while only negligible impacts on general (not cultural) benchmarks.

We propose **practical application** of our findings when fine-tuning LLMs.

Method: CULNIG



(1) Define **gradient-based score**[1] to quantify neuron's contribution to outputs:

$n \times \frac{\partial P(y|x)}{\partial n}$ (n = neuron activation value, y = output tokens, x = input tokens)

(2) Use **BLEnD**[2], a cultural knowledge benchmark, to calculate each neuron's score (s_{neur})

(3) Subtract scores on **BLEnDctrl** (s_{ctrl}) to exclude neurons contributing task understanding

(4) Create another control dataset, **CRC**, to exclude neurons responding to country name tokens

➔ Top t % neurons on (s_{neur} - s_{ctrl}) - top r % neurons on CRC are considered **culture-general neurons**

*For ***culture-specific*** **neurons**, apply another filtering based on z-score to exclude neurons responsible for multiple cultures

Culture-General Neuron

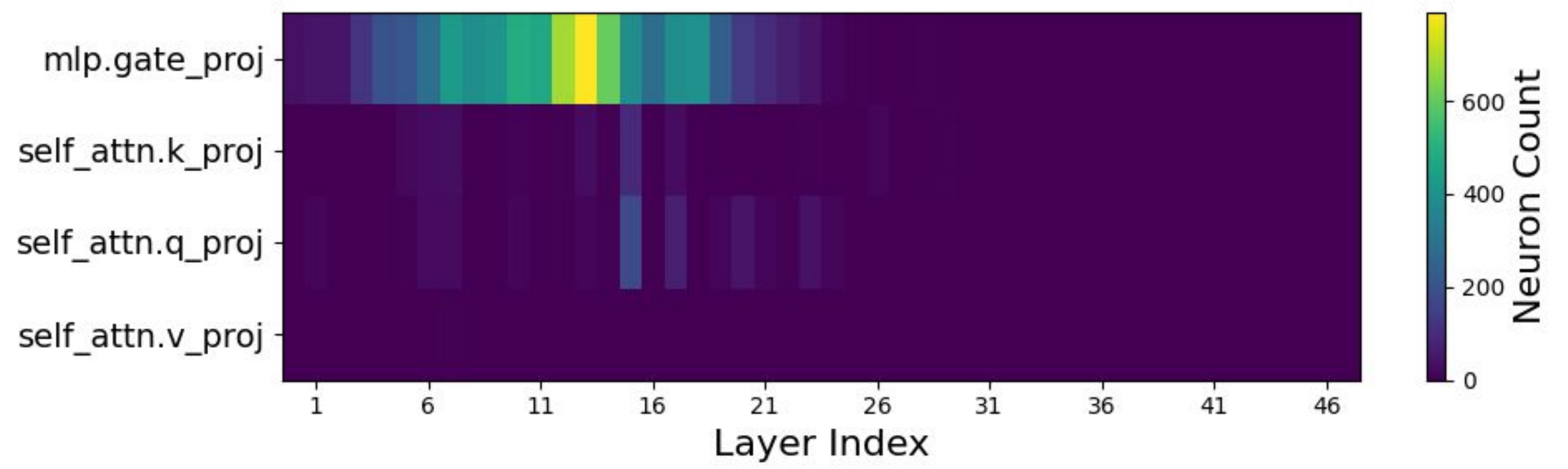
Evaluation Results

Evaluate *gemma-3-12b-it* with culture-general neurons masked

mask	Cultural benchmarks				NLU benchmarks		
	BLEnDtest	CultB	NormAd	WVB	ComQA	QNLI	MRPC
no	64.2	78.1	58.5	64.1	79.7	75.4	78.0
culture	**37.9**	**62.0**	**52.0**	**58.5**	**75.1**	72.8	78.7
random	63.6	77.3	57.6	64.0	79.2	75.5	78.2

- Masking culture-general neurons causes **statistically significant degradation** on scores of **cultural benchmarks**
- Impacts on **broad cultural properties** (knowledge, norm, values)
- Only minor influence on general (not cultural) benchmarks

Neuron Distribution



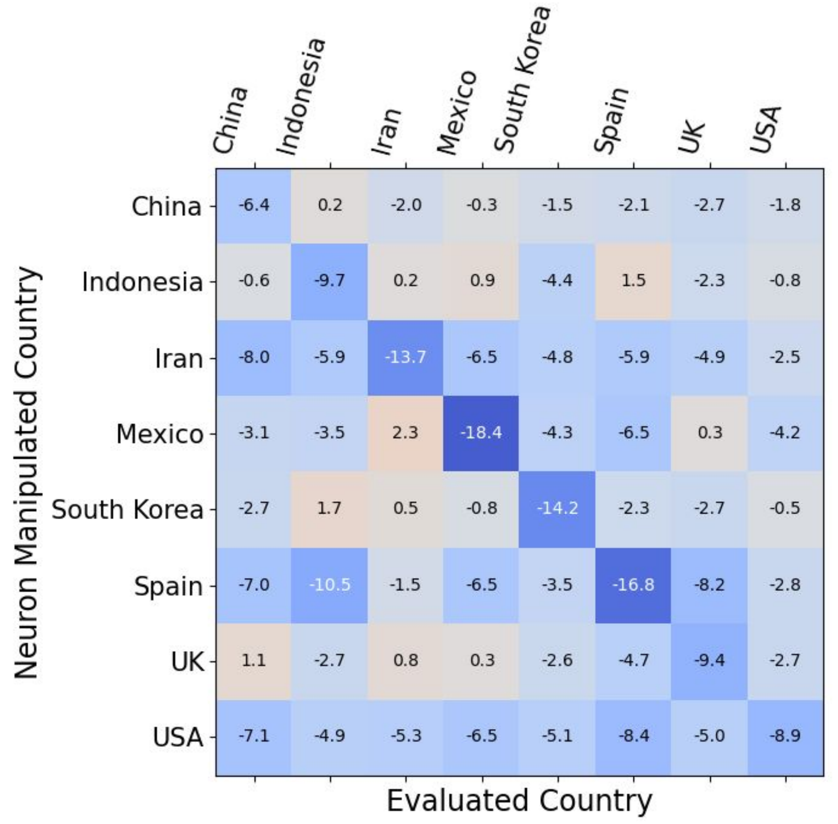
- Neurons are concentrated in **shallow to middle MLP** modules
- Culture-general neurons consist of only **<1%** of all neurons
- This tendency is common across six models

Culture-Specific Neuron

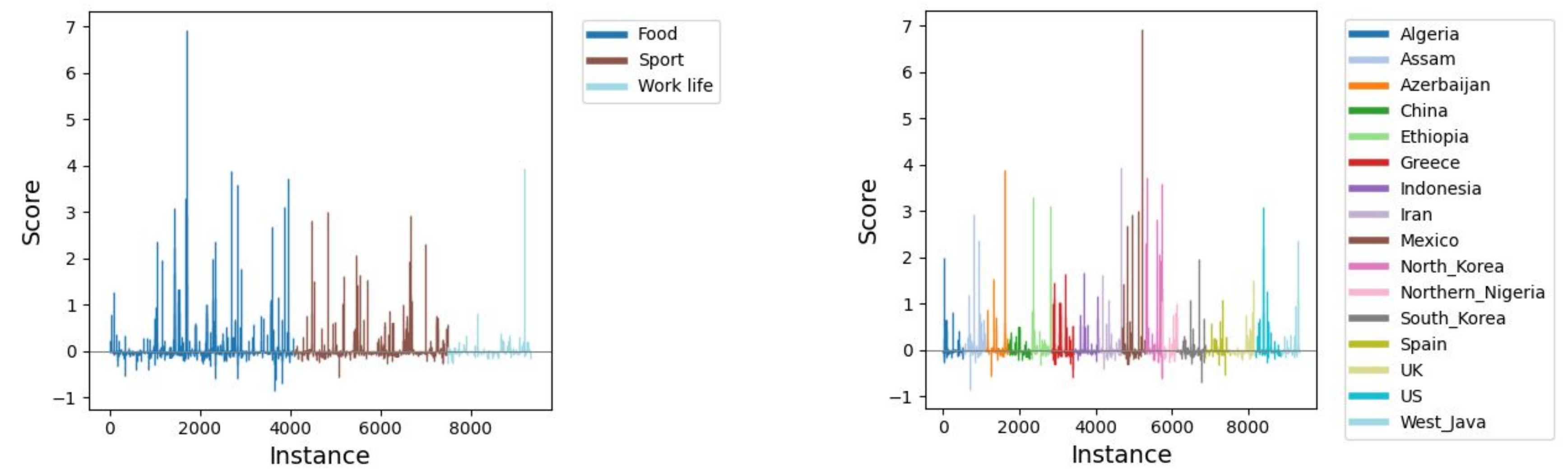
Evaluation Results

Mask culture-specific neurons of 8 cultures, and evaluate on problems of 8 cultures (BLEnDtest)

- **Score↓ on target culture**
  - Identified neurons represent target culture
- **Score↓ on related cultures**
  - E.g. mask 🇲🇽 neurons → 🇲🇽 & 🇪🇸 performance↓
  - Neurons may be shared between cultures sharing languages or those geographically close



Instance-Level Scores Analysis



Neurons take positive scores on only 29% problems, high scores across categories and cultures

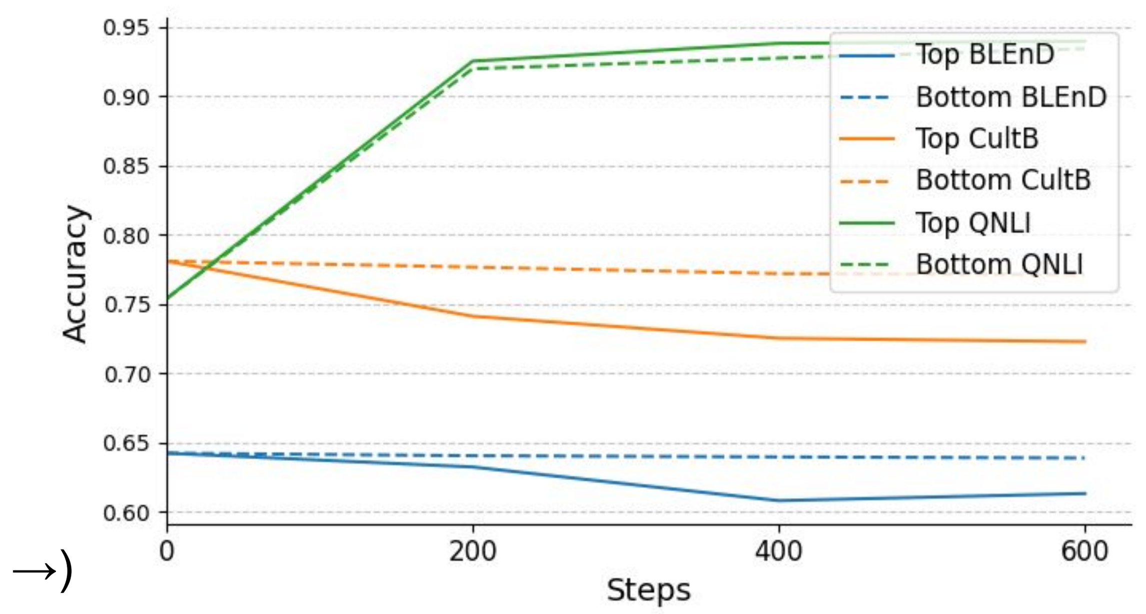
➔ **Knowledge-level** (not meta-level) concepts are encoded in culture-general neurons

Application to Model Training

Train LLMs with QNLI, updating 10% of parameters from modules containing:

- (***Top***) many culture-general neurons
- (***Bottom***) no culture-general neurons

Q. Updating culture-general neurons degrade cultural understanding?



Results (Evaluated on every 200 steps →)

- **Updating Top modules tends to degrade cultural understanding of LLMs**
- QNLI accuracy improves in both settings
- Findings can be applied to module selection when fine-tuning

Conclusion

- Identify **culture-general/specific** neurons contributing to cultural understanding of LLMs and evaluate their effects on cultural behavior
- Those neurons are concentrated in **shallow-middle MLP modules**
- Culture-specific neurons contribute to both their target and related cultures
- When fine-tuning, updating modules with many culture-general neurons is susceptible to losing cultural understanding

References

[1] Nakyeong Yang, Taegwan Kang, Stanley Jungkyu Choi, Honglak Lee, and Kyomin Jung. Mitigating biases for instruction-following language models via bias neurons elimination. ACL 2024.

[2] Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, et al., BLEnD: A benchmark for LLMs on everyday knowledge in diverse cultures and languages. NeurIPS Datasets and Benchmarks Track, 2024.