Neuron Specialization

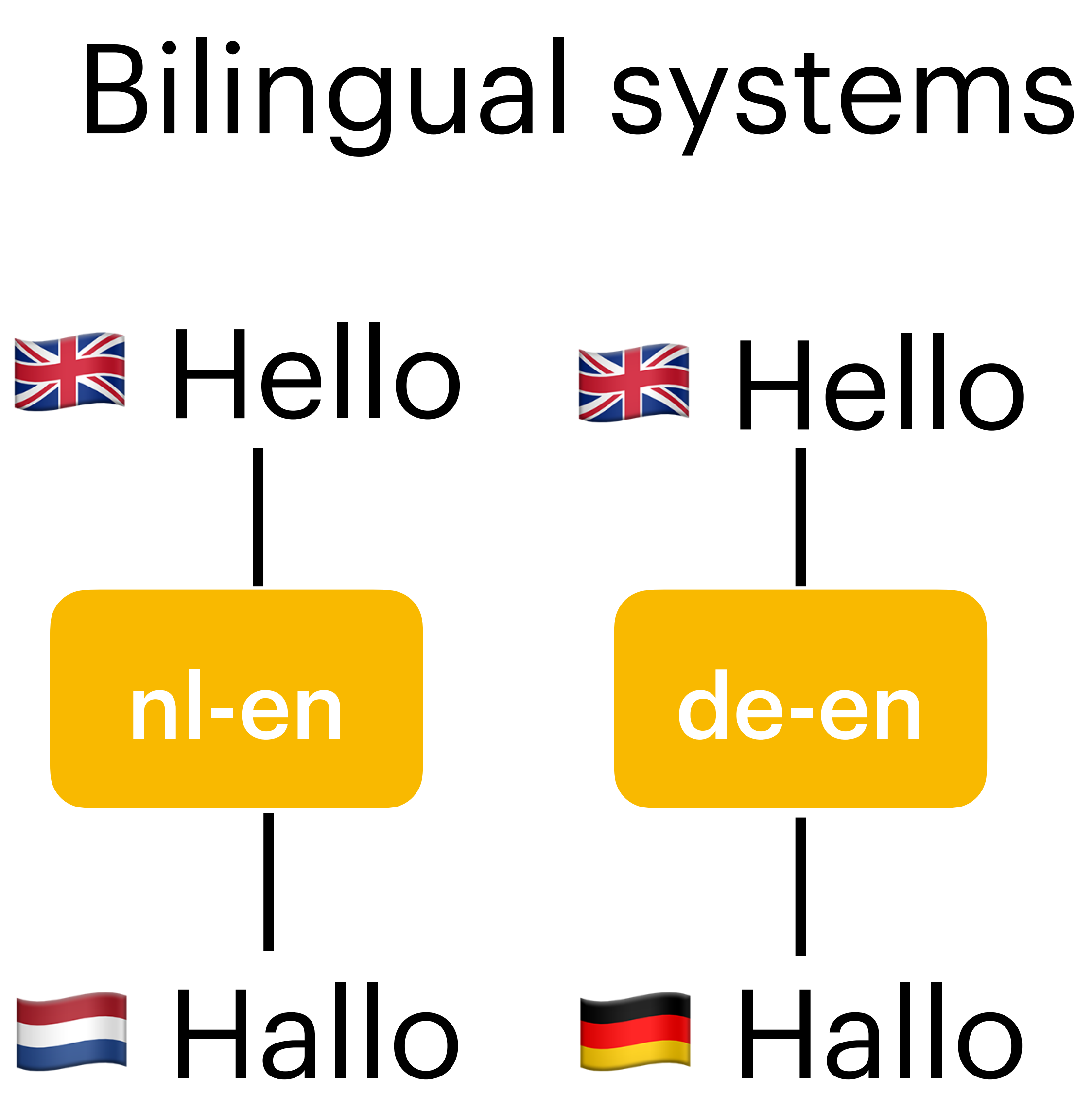
Leveraging Intrinsic Task Modularity for Multilingual Machine Translation

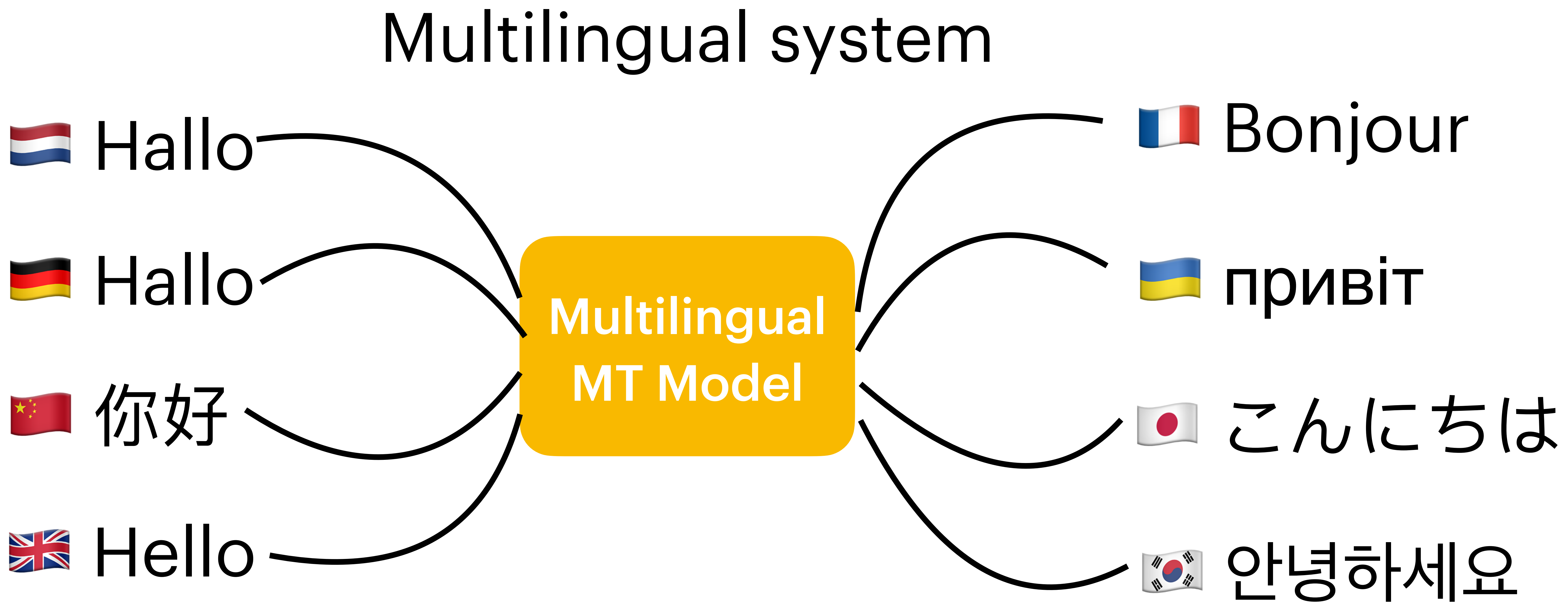
SHAOMU TAN

UNIVERSITY OF AMSTERDAM
Language Technology Lab

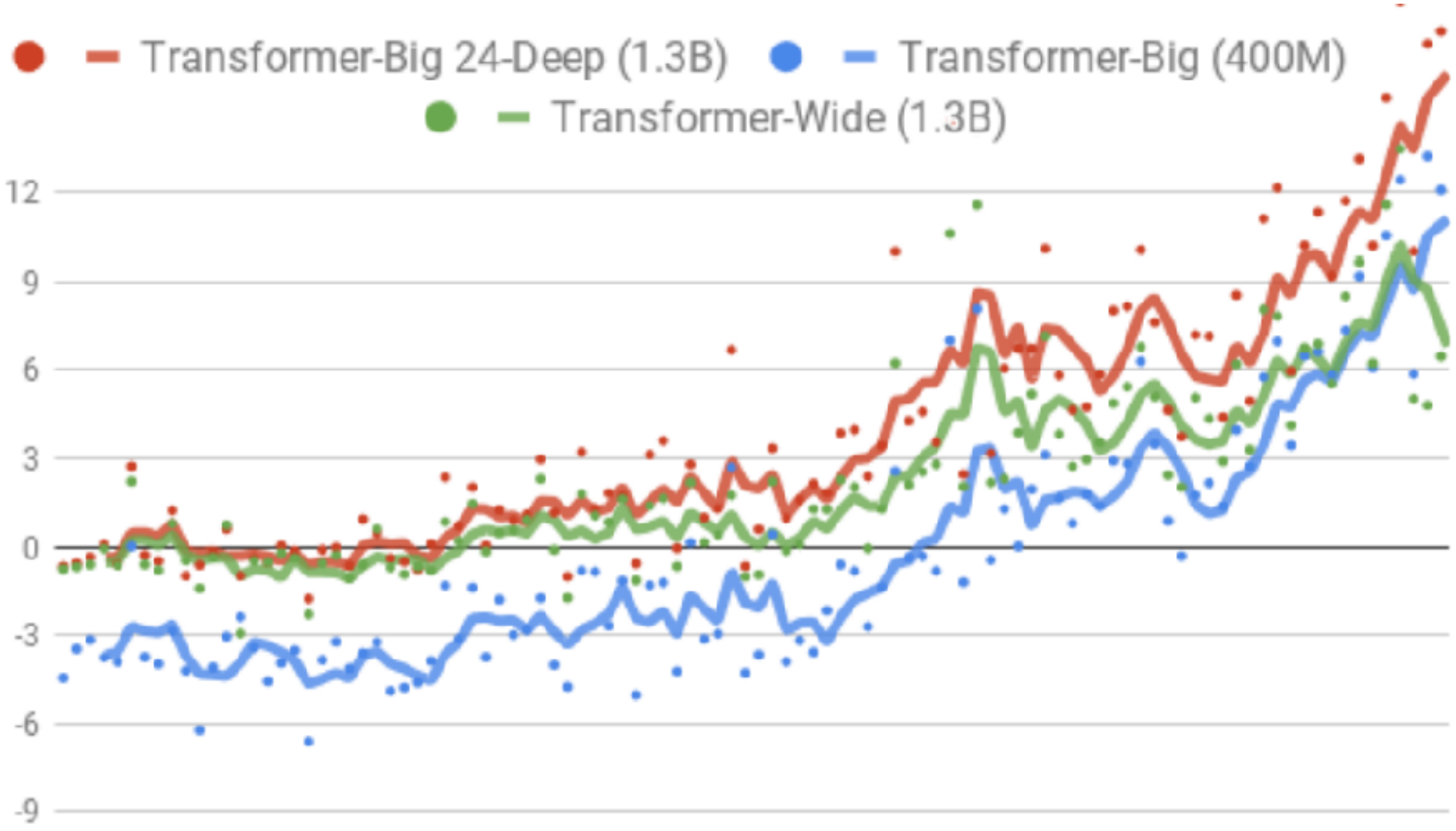
Multilingual Neural Machine Translation (MNMT)

-> Training a unified model on a mixed dataset from multiple languages.
-> Efficient: One model for many languages.





Multilingual Machine Translation
facilitates Knowledge Transfer



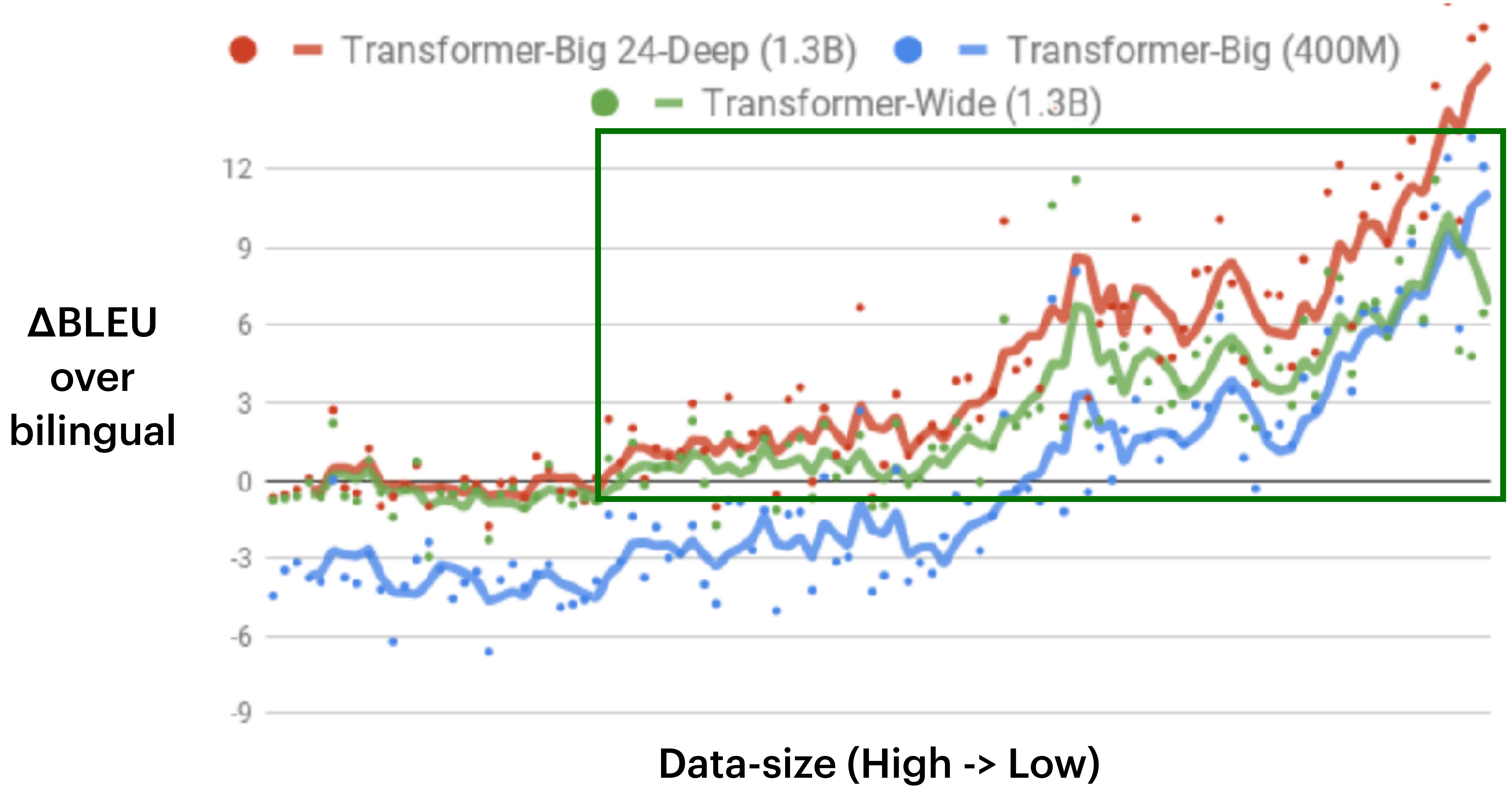
ΔBLEU over bilingual

Data-size (High -> Low)

Arivazhagan, Naveen, et al. "Massively multilingual neural machine translation in the wild: Findings and challenges."

Multilingual Machine Translation

facilitates Knowledge Transfer

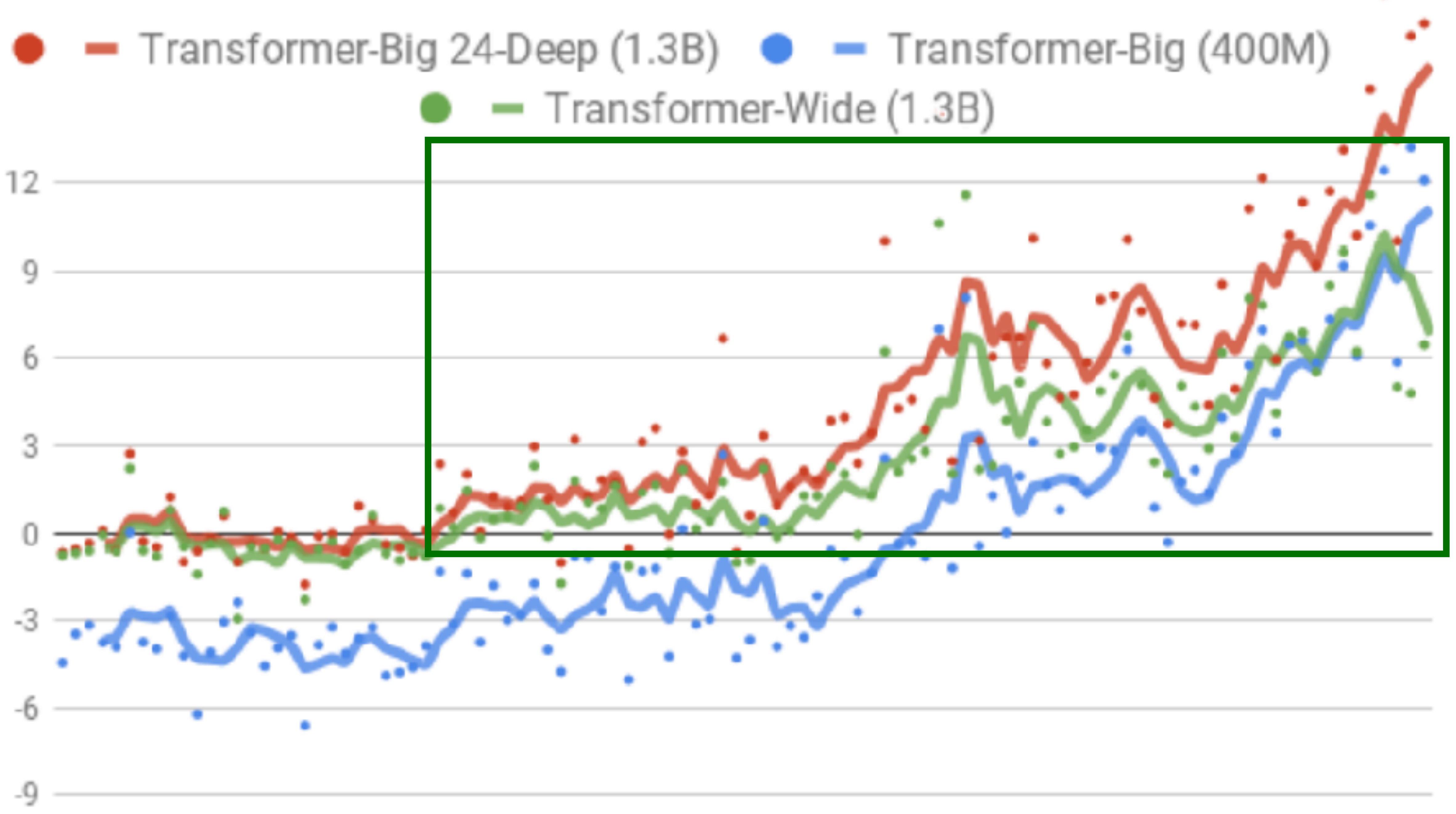


Knowledge transfer benefits **low-resource** languages

Arivazhagan, Naveen, et al. "Massively multilingual neural machine translation in the wild: Findings and challenges."

Multilingual Machine Translation
facilitates Knowledge Transfer

ΔBLEU over bilingual



Data-size (High -> Low)

Knowledge transfer benefits **low-resource** languages

e.g.:
Bilingual (100k En-Lb) -> **Multilingual** (+ 10M En-Germanic)

+9 BLEU

Arivazhagan, Naveen, et al. "Massively multilingual neural machine translation in the wild: Findings and challenges."

Multilingual Machine Translation

facilitates Knowledge Transfer

Shared Vocabulary

Uni-versi-ty

جامعة

Uni-versi-teit

אוניברסיטה

Uni-versi-tät

大學

университет

大学

Uni-versi-té

Grande école

대-학교

Knowledge transfer in vocabulary

Multilingual Machine Translation

is not a free lunch

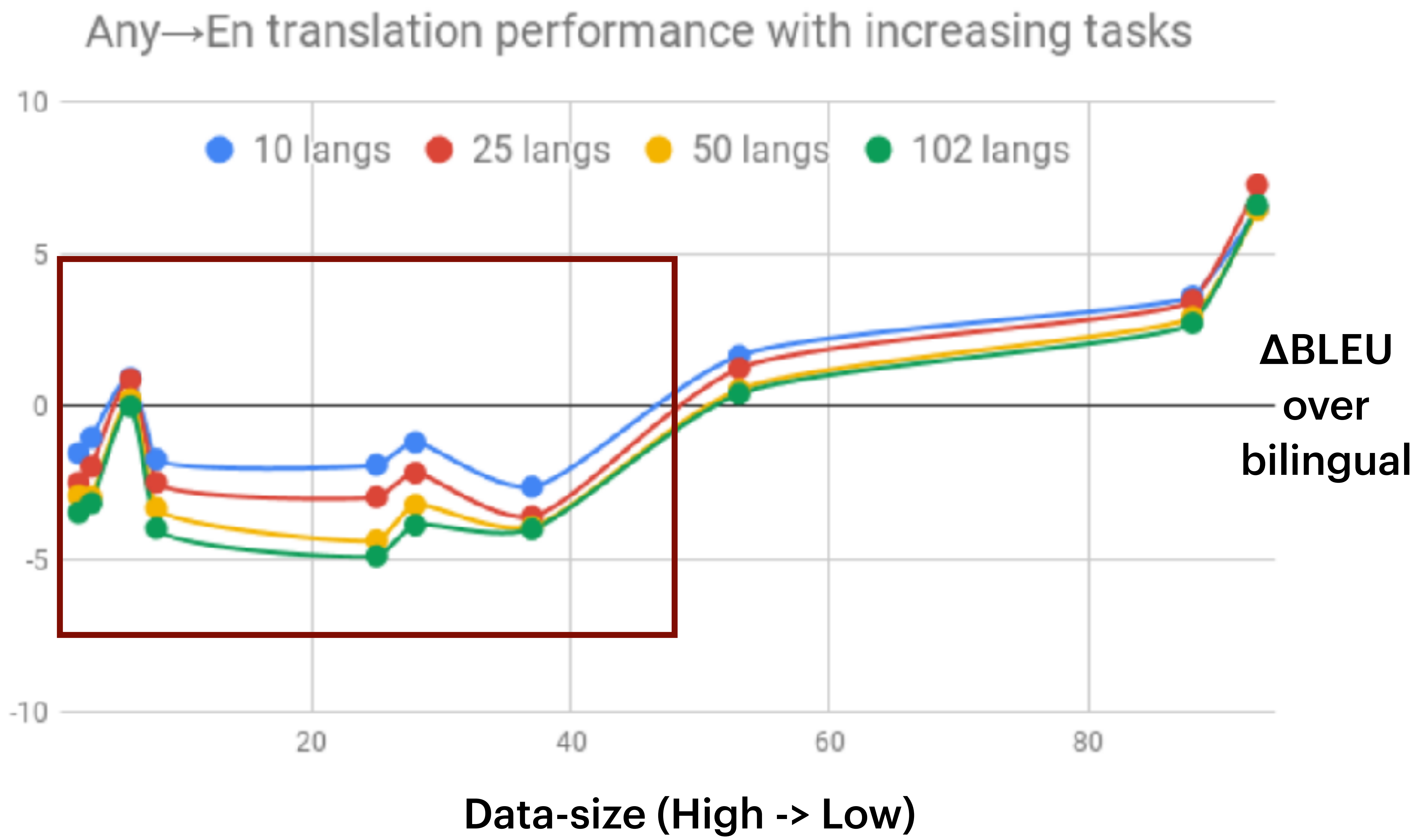
Joint Multilingual Training brings Synergy
but also **Interference** (negative transfer)

Multilingual Machine Translation

is not a free lunch

Interference

compromises performance
(for high-resource languages)



Arivazhagan, Naveen, et al. "Massively multilingual neural machine translation in the wild: Findings and challenges."

Multilingual Machine Translation

why Interference?

Interference

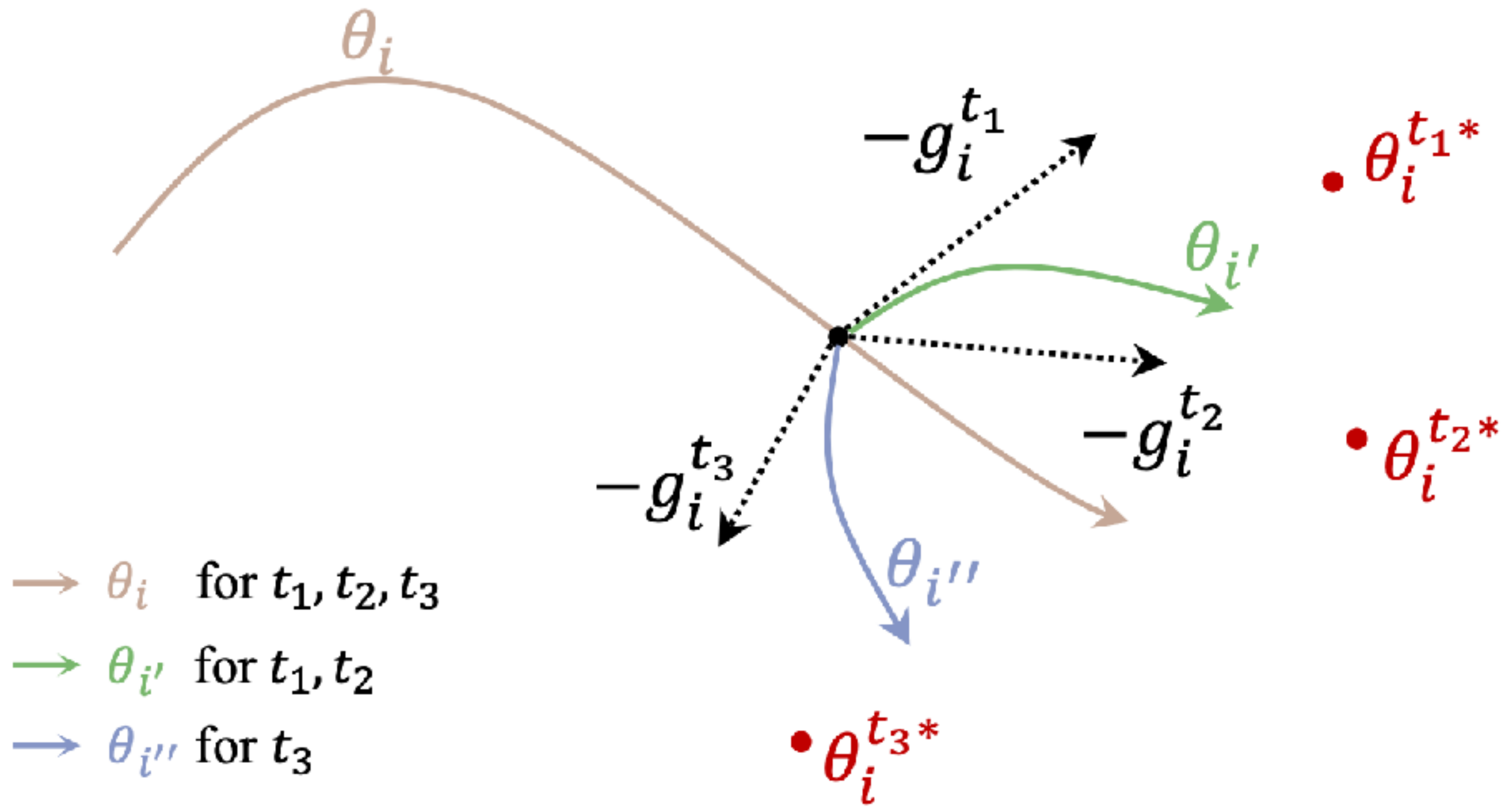
rooted in **Conflicting** optimization demands of various tasks

Multilingual Machine Translation

why Interference?

Interference

rooted in **Conflicting** optimization demands of various tasks



Gradient Conflicts

Wang, Qian, and Jiajun Zhang. "Parameter differentiation based multilingual neural machine translation."

Multilingual Machine Translation

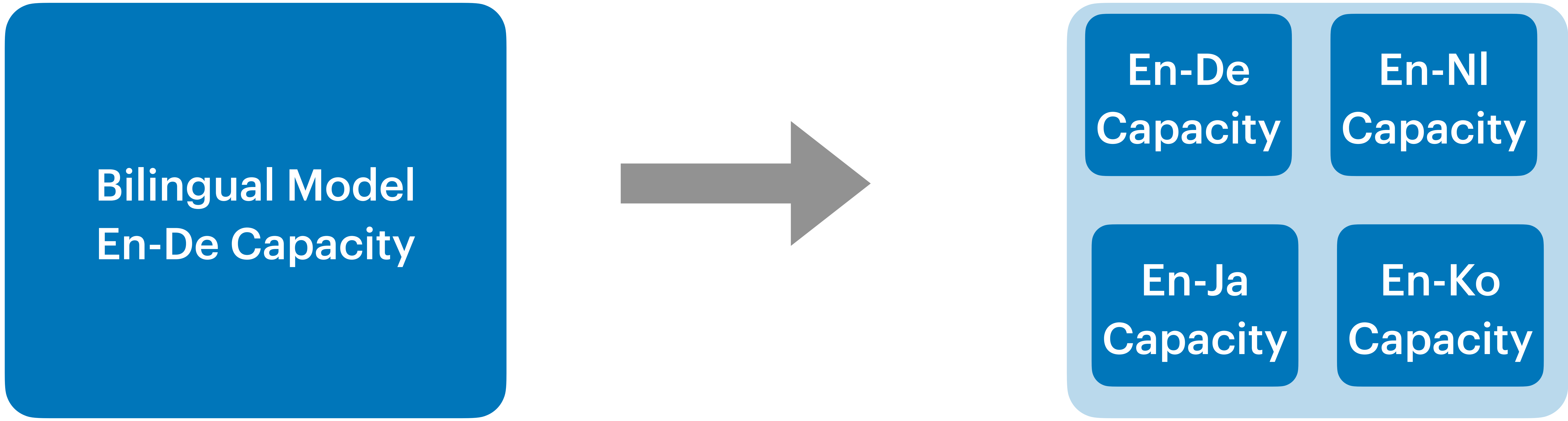
why Interference?

Interference

can be seen as a capacity issue.

Bilingual -> Multilingual:

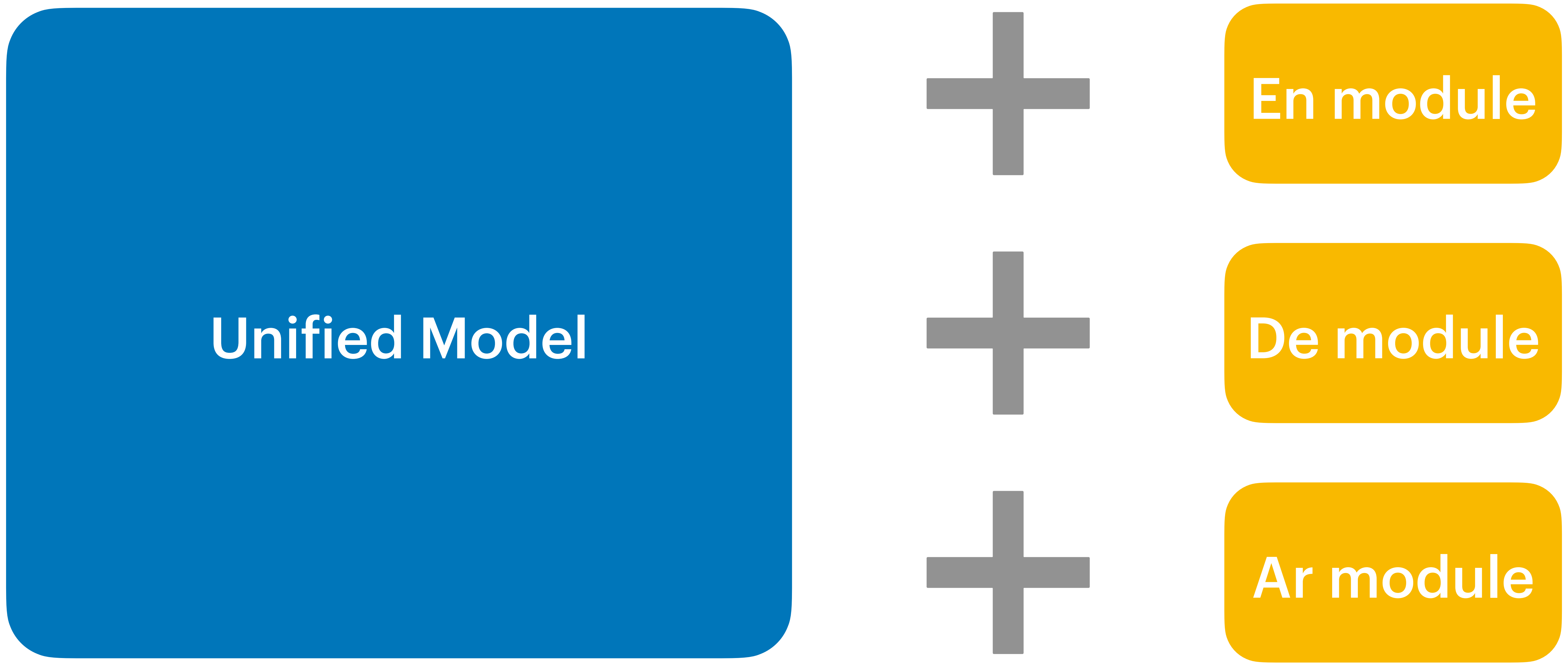
Model Capacity for each Language-Pair decreased when remaining the same model size.



Recent work in Reducing Interference

How to Reduce Interference

Modular Deep Learning - to introduce Langauge Specificity

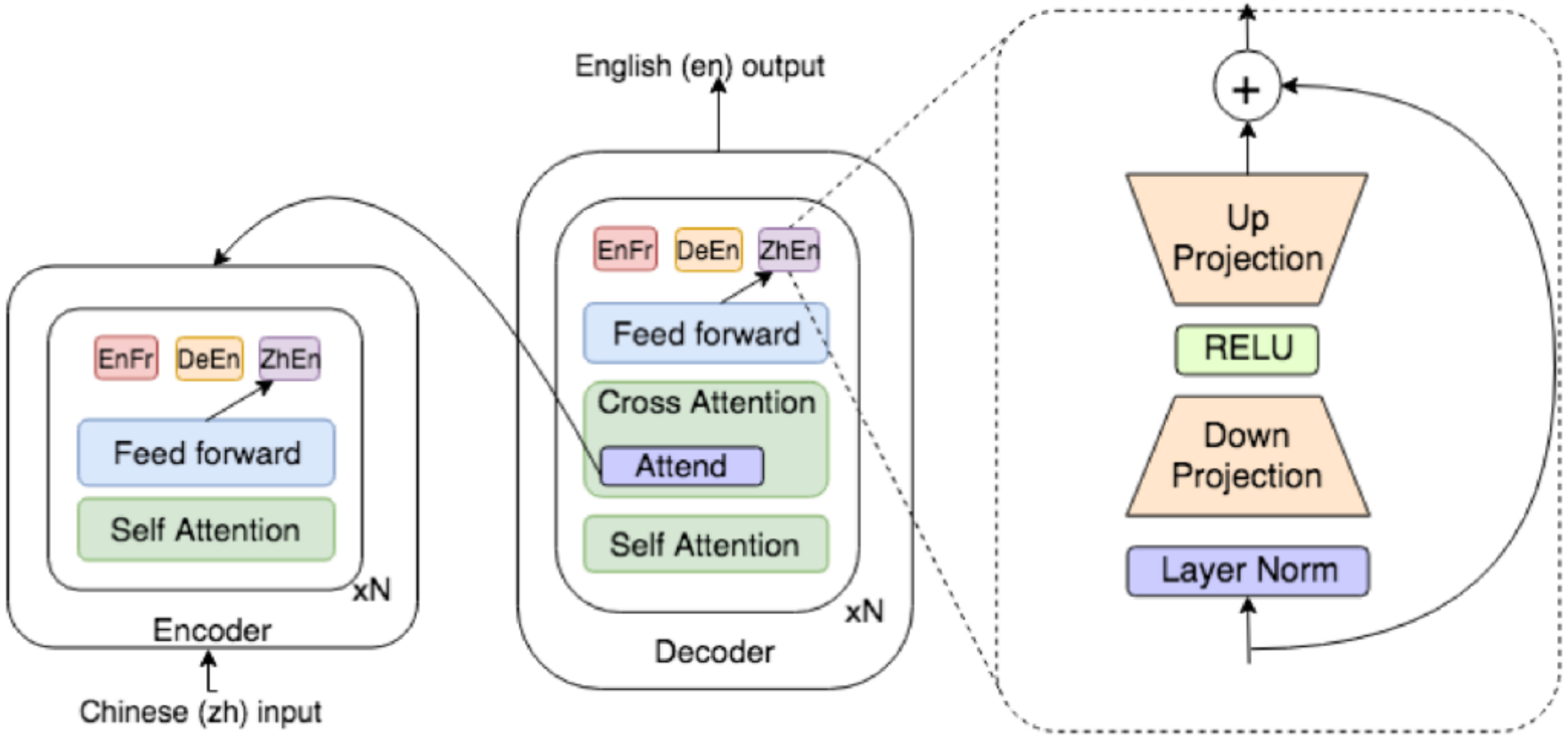


Multilingual Model

Language-Specific (Ls) Modules

How to Reduce Interference

Modular Deep Learning - Adapters



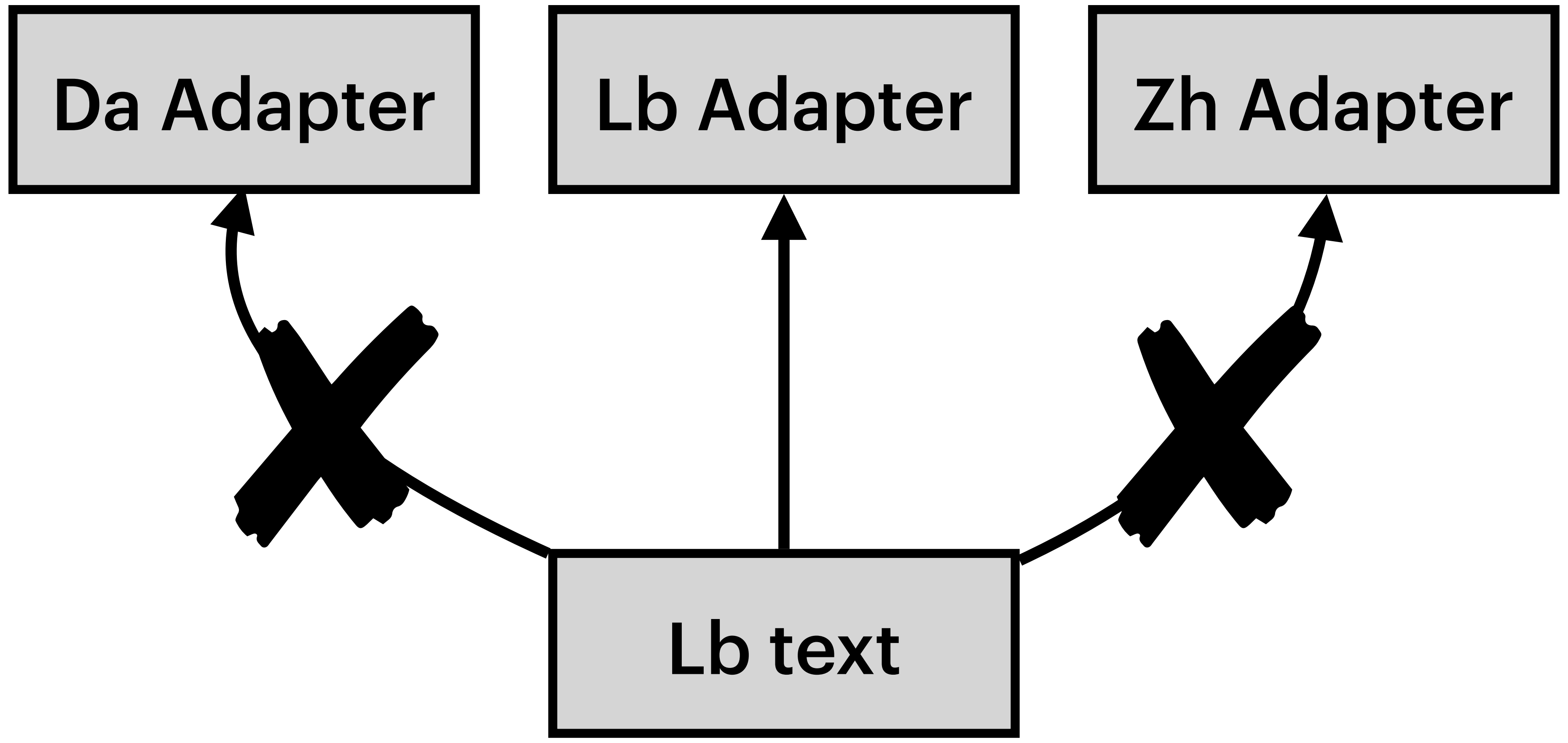
Language Pair Adapters: insert adapters conditioned on language pairs to add language-specific capacities.

Bapna, Ankur, and Orhan Firat. "Simple, Scalable Adaptation for Neural Machine Translation."

How to Reduce Interference

Limitations - Modular Deep Learning

Adapters, Language-Specific Modules, are **Language-Dependent** that **operates in isolation**



Such Design fundamentally **dis-encourages cross-lingual Transfer** especially for low-resource languages

How to Reduce Interference

Limitations - Modular Deep Learning

Trade-Off: Efficiency & Performance

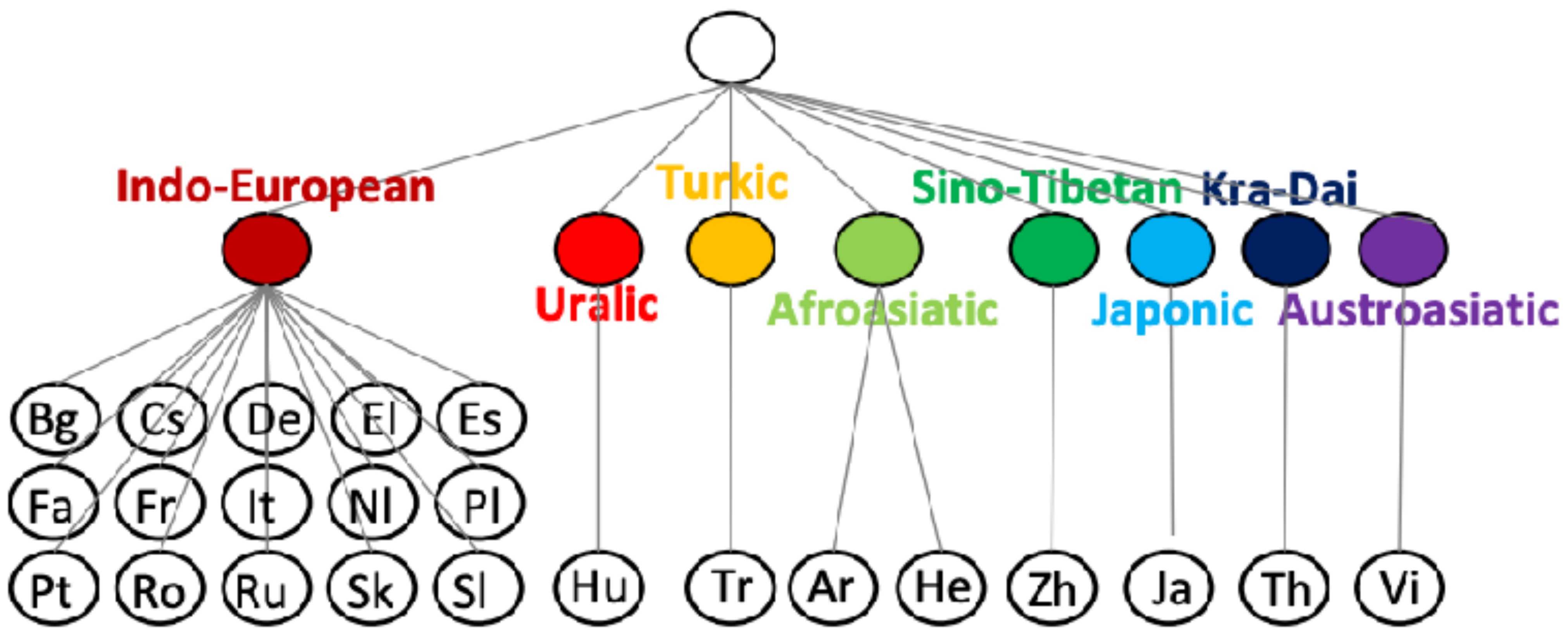
a. increase substantial parameters when many languages are involved

b. memory[1] and latency[2] issue

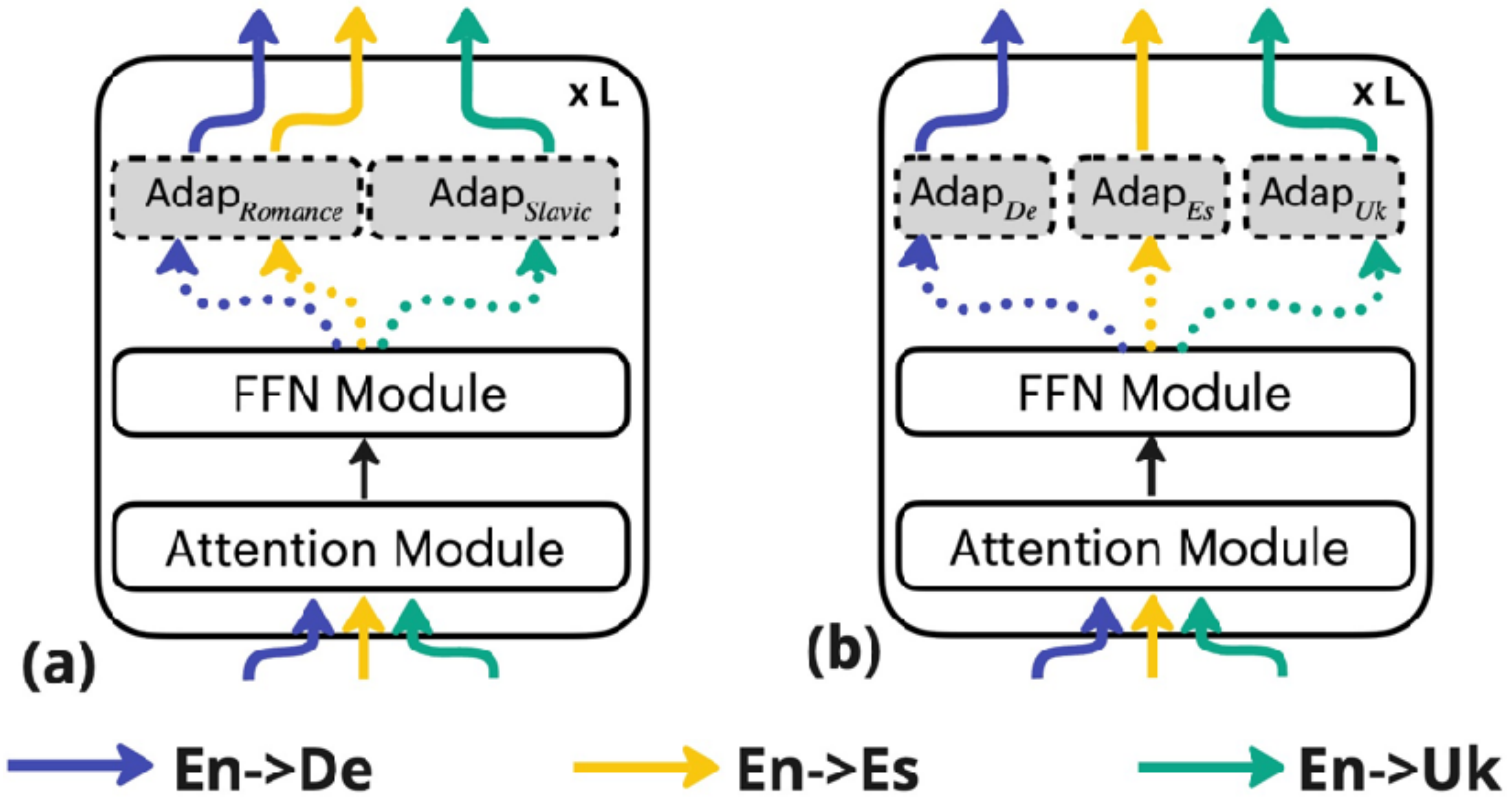
1) Liao, Baohao, Shaomu Tan, and Christof Monz. "Make your pre-trained model reversible: From parameter to memory efficient fine-tuning."
2) Liao, Baohao, Yan Meng, and Christof Monz. "Parameter-efficient fine-tuning without introducing new latency."

How to Reduce Interference

Leveraging Priori Linguistic Knowledge



Language cluster Training[1]**:** Train one multilingual model for one language cluster.



(1) Tan, Xu, et al. "Multilingual neural machine translation with language clustering."
(2) Chronopoulou, et al "Language-family adapters for low-resource multilingual neural machine translation."

How to Reduce Interference

Limitations - Leveraging Priori Linguistic Knowledge

a. Heavily rely on priori knowledge, e.g.: linguistic knowledge.

b. lack clear inductive bias, thus heavy reliance on heuristics.

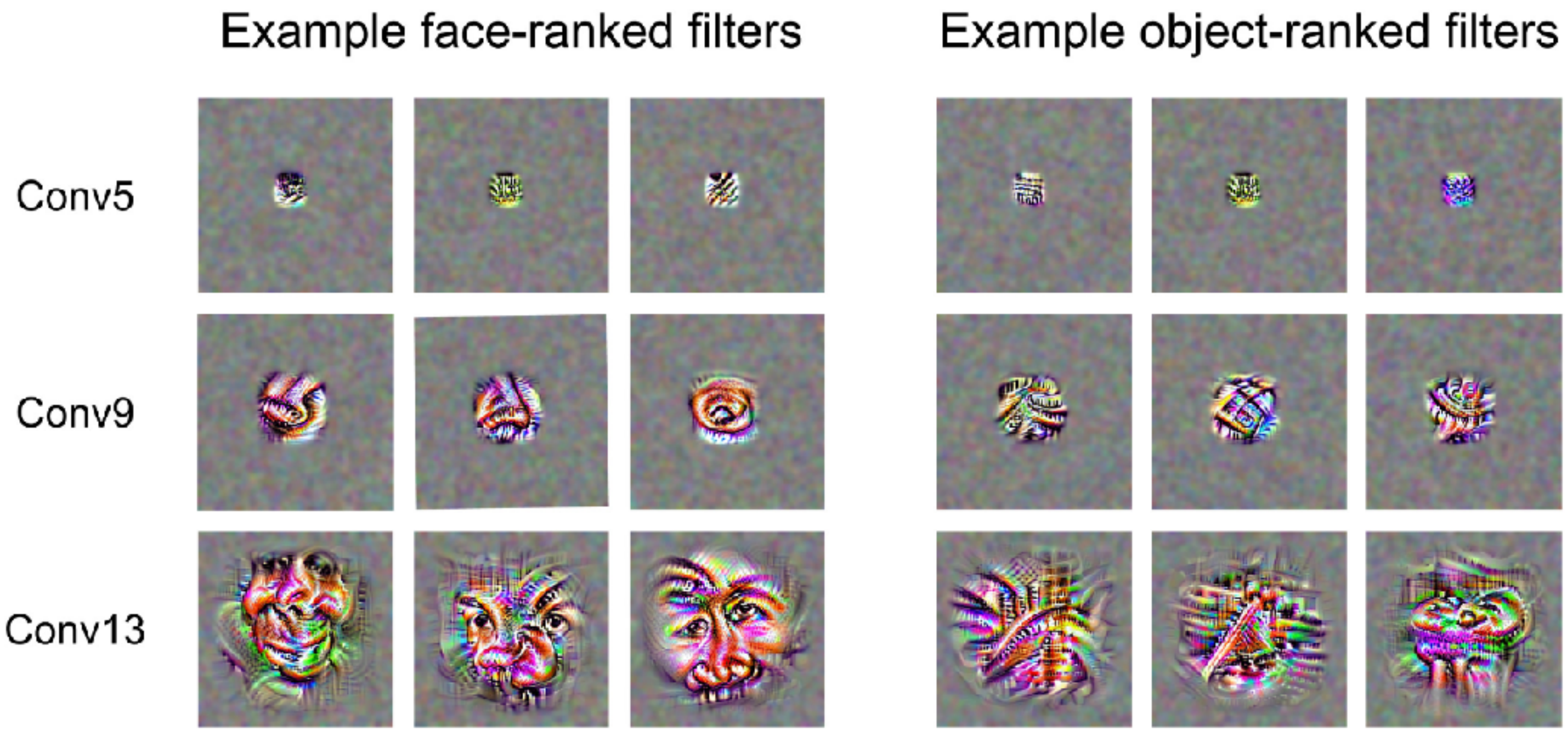
c. show strong effects for low-resource languages, less or no effects on high-resource ones.

Neuron Specialization

Exploring the Intrinsic Modularity

in Multi-task Networks

Intrinsic Modularity
in Multi-task Vision Networks



Multi-task training develops task-specific **functional specialization:**

Multi-task networks form ***Task-Specific*** *Sub-networks:* face-filters & object-filters

Dobs, Katharina, et al. "Brain-like functional specialization emerges spontaneously in deep neural networks." *Science advances*

Locating Intrinsic Modularity

in MNMT Models

Prior Studies attempt to identify Task-Specific **Sub-networks** inside trained Multi-task Models

Locating Intrinsic Modularity

in MNMT Models

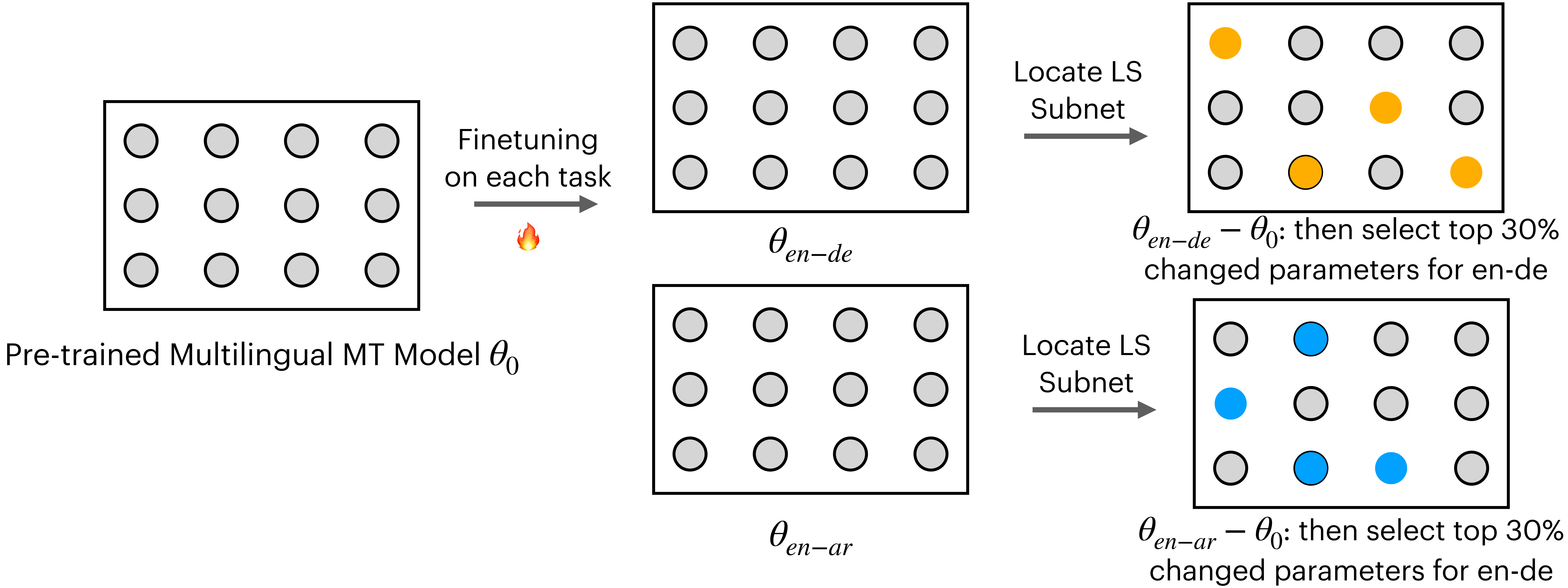
Prior Studies attempt to identify Task-Specific **Sub-networks** inside trained Multi-task Models

Fine-tuning tasks to see what parameters changed the most[1,2,3]

1) Lin, Zehui, et al. "Learning language specific sub-network for multilingual machine translation."
2) He, Dan, et al. "Gradient-based Gradual Pruning for Language-Specific Multilingual Neural Machine Translation."
3) Choenni, Rochelle, et al. "Cross-Lingual Transfer with Language-Specific Subnetworks for Low-Resource Dependency Parsing."

Locating Intrinsic Modularity in MNMT Models

LaSS: Fine-tuning the pre-trained multi-task model on each task to see what parameters changed the most[1]



1) Lin, Zehui, et al. "Learning language specific sub-network for multilingual machine translation."

Locating Intrinsic Modularity

requires Network Modifications

Fine-tuning approaches (LaSS) raise a question:

whether the modularity (Subnets) is inherent to the original model,
or simply an **artifact** introduced by network modifications?

Modularity in Finetuned Model reflect that in pre-trained model?

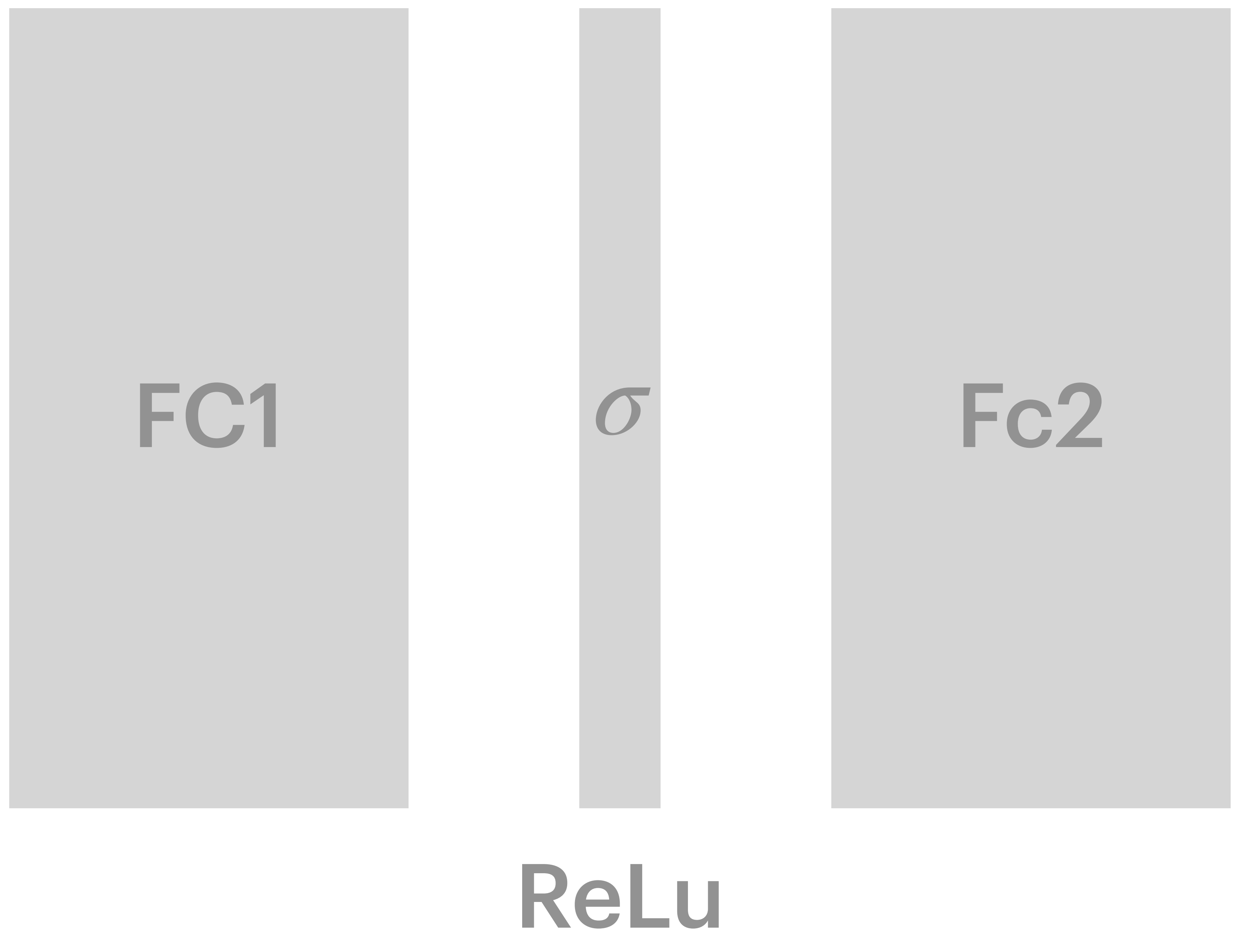
Does Intrinsic Modularity even exist?

Analyzing task Modularity in Multi-task models

Neuron Specialization

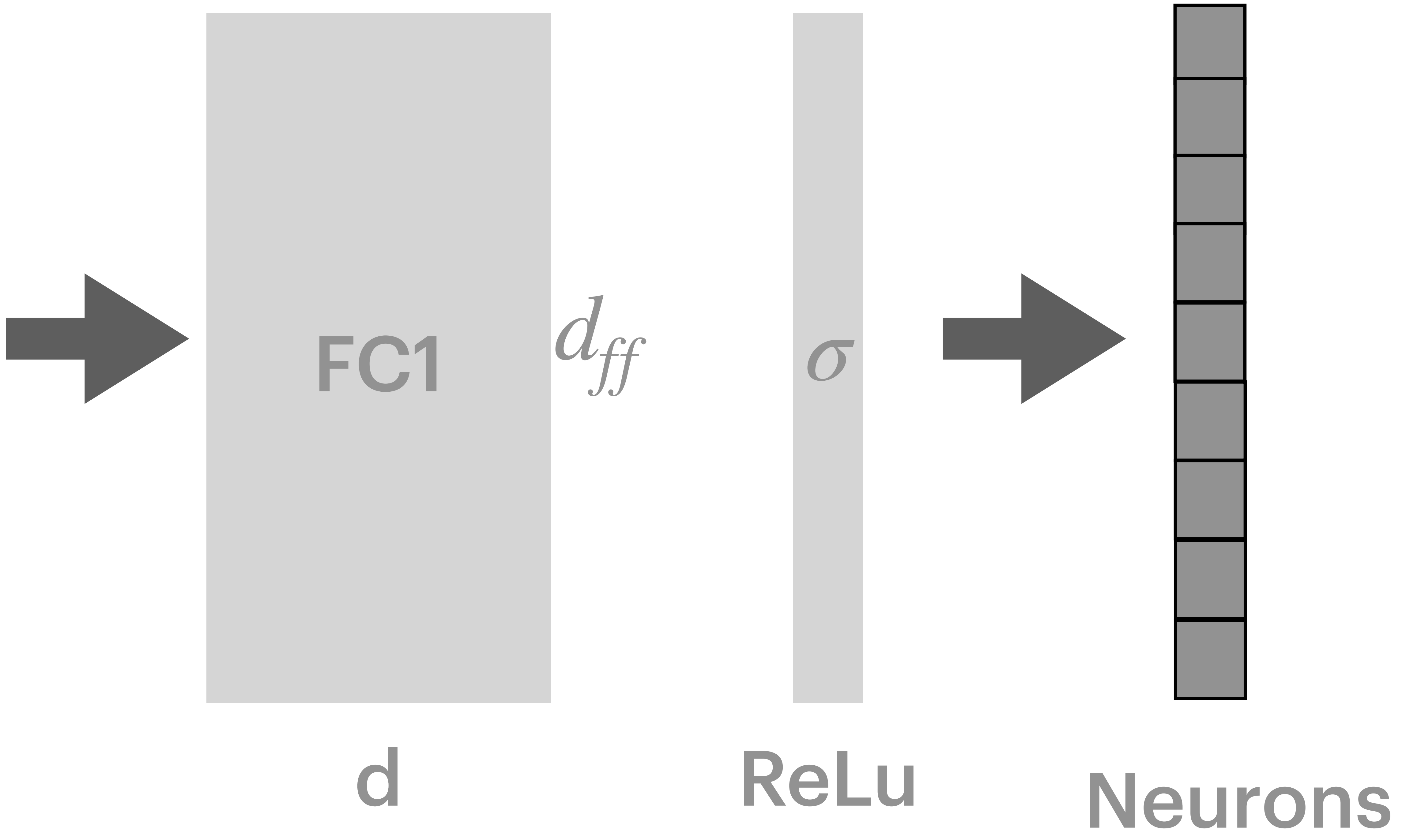
Neuron Structural Analysis - Method

We focus on Neurons:
intermediate activations inside the feed-forward (FFN) blocks



Neuron Specialization

Neuron Structural Analysis - Method



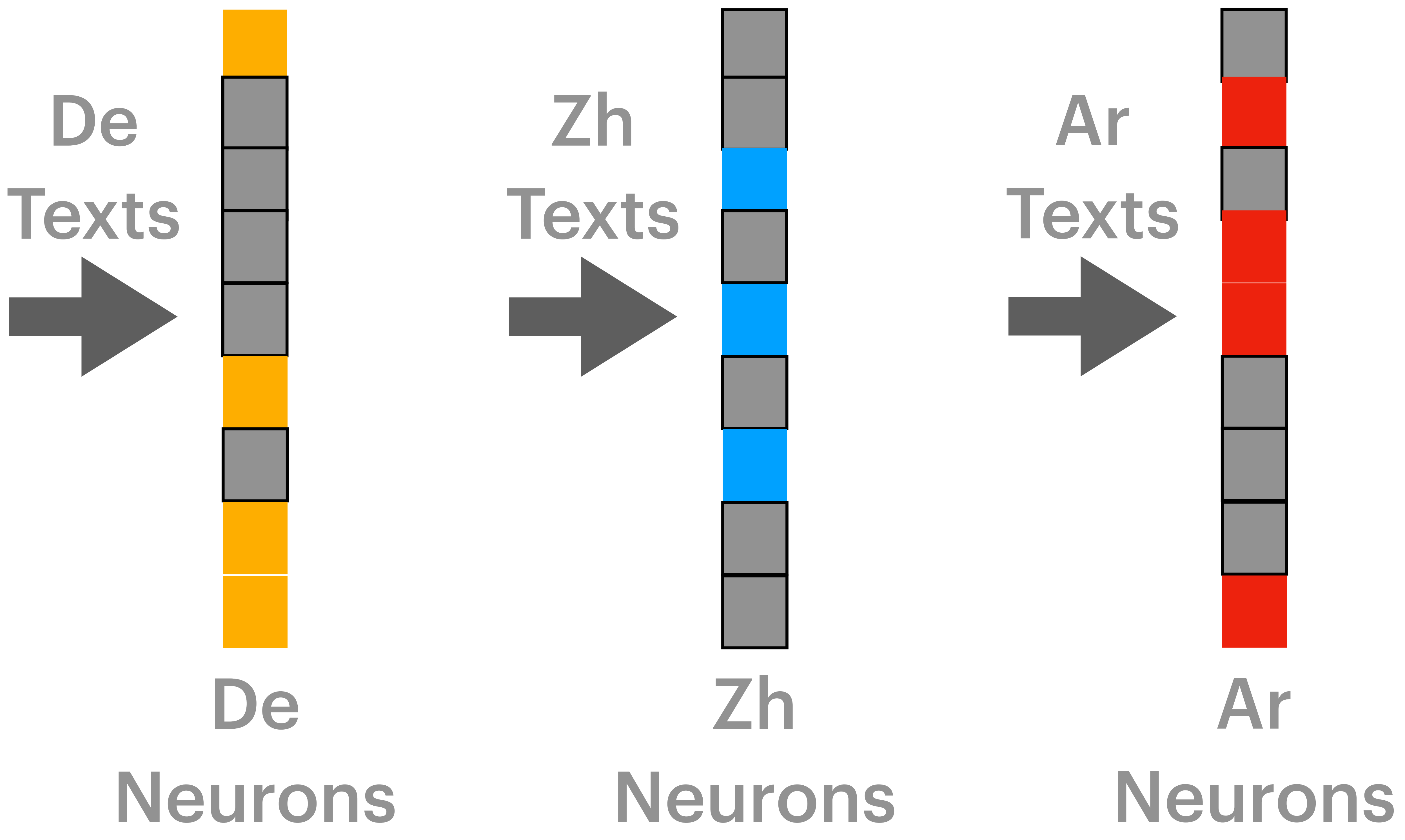
Neurons can only be:

Activated: >0

Non-activated: =0

Neuron Specialization

Neuron Structural Analysis - Method

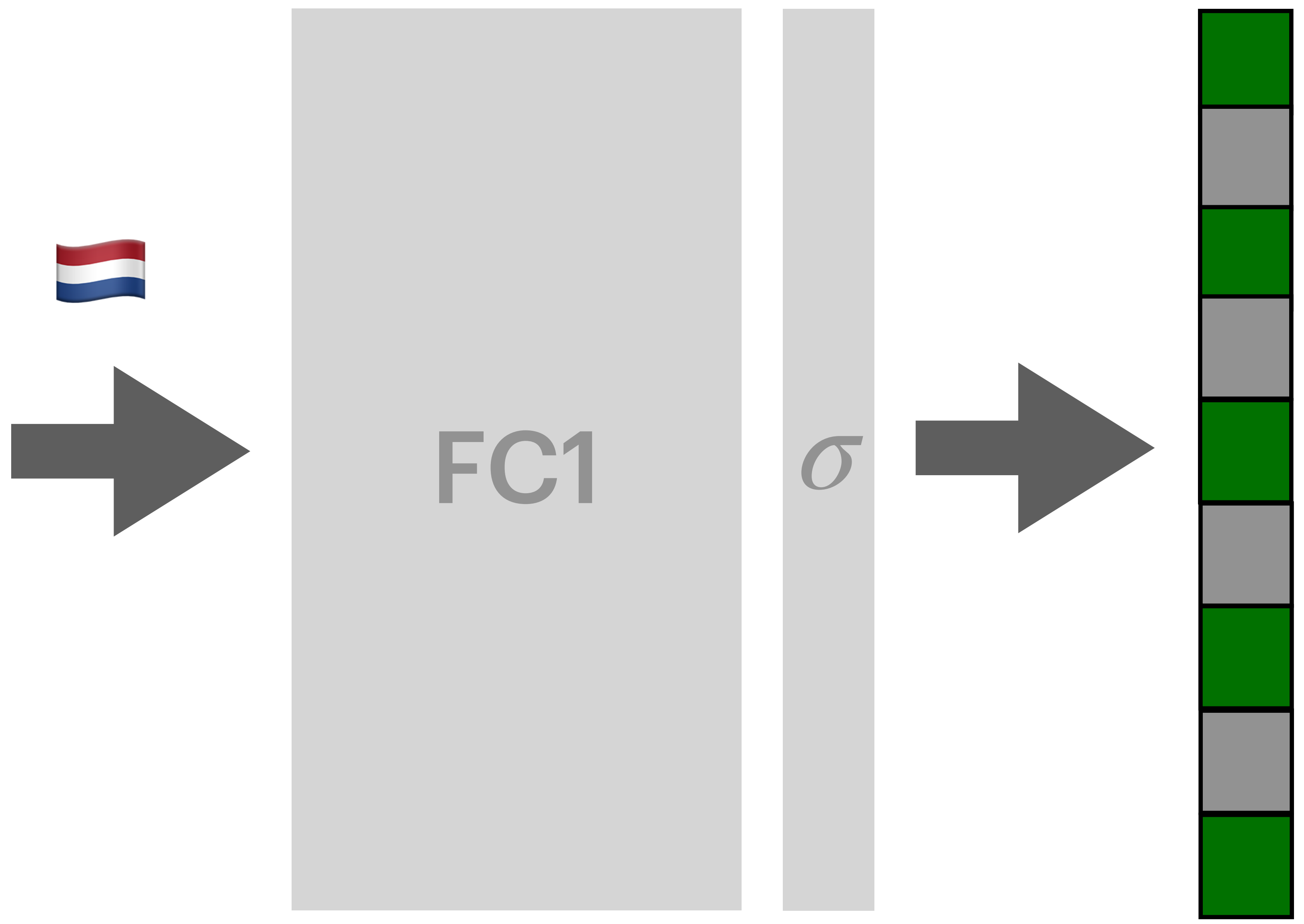


Intuition:
Are neurons task-specific?

Neuron Specialization

Neuron Structural Analysis - Method

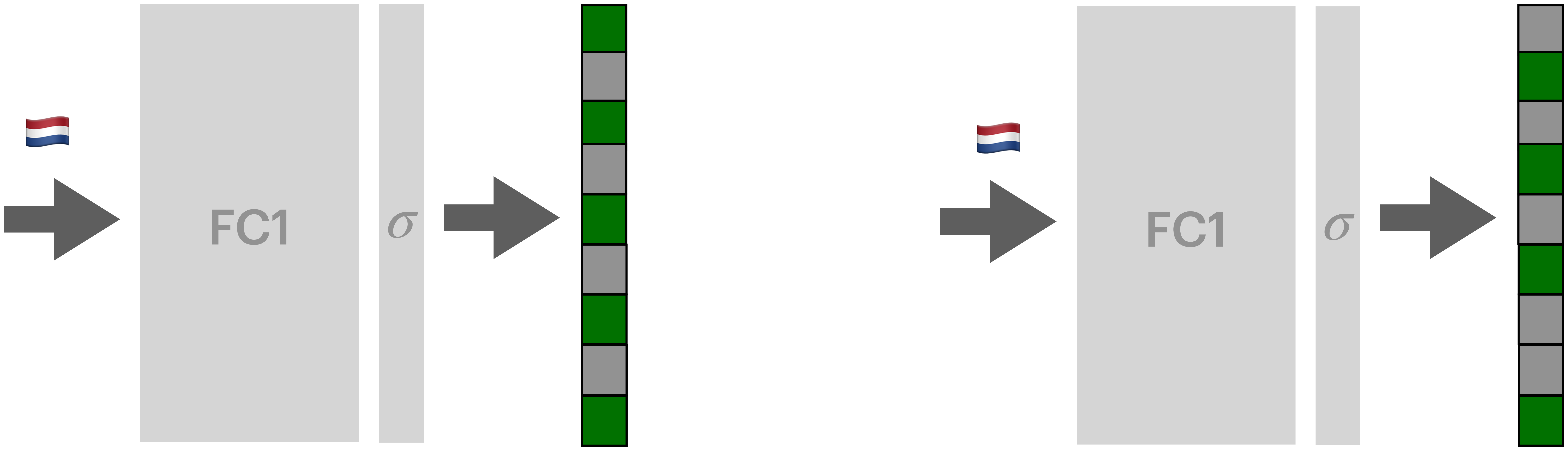
Activation Recording: Feed some sentences to observe which neurons are active / inactive



Neuron Specialization

Neuron Structural Analysis - Method

Activation Recording: Feed some sentences to observe which neurons are active / inactive



Neuron Specialization

Neuron Structural Analysis - Method

Neuron Activation Frequency

🇳🇱

23210
18750
11921
9872
...
23
0
0

Neuron Specialization

Neuron Structural Analysis - Method

Neuron Activation Frequency

23210
18750
11921
9872
...
23
0
0

How should we select Specialized Neurons for each language pair?

Neuron Specialization

Neuron Structural Analysis - Method

Specialized Neuron Selection

$$\sum_{n=1}^{d_{ff}} a_n^{s\to t}$$

23210
18750
11921
9872
...
23
0
0

$$\sum_{n \in S_k^{s\to t}} a_n^{s\to t} \geq k * \sum_{n=1}^{d_{ff}} a_n^{s\to t},$$

We **dynamically** select neurons based on a cumulative activation threshold $k \in [0,1]$.

Neuron Specialization

Neuron Structural Analysis - Analysis

en->de	en->nl
1	1
0	1
1	1
0	0
1	0
0	0
1	1
0	0
1	1

$$m_{s\to t}^{l=1} \in \mathbb{R}^{d_{ff}}$$

Neuron Specialization

Neuron Structural Analysis - Analysis

en->de	en->nl
1	1
0	1
1	1
0	0
1	0
0	0
1	1
0	0
1	1

$$m_{s\to t}^{l=1} \in \mathbb{R}^{d_{ff}}$$

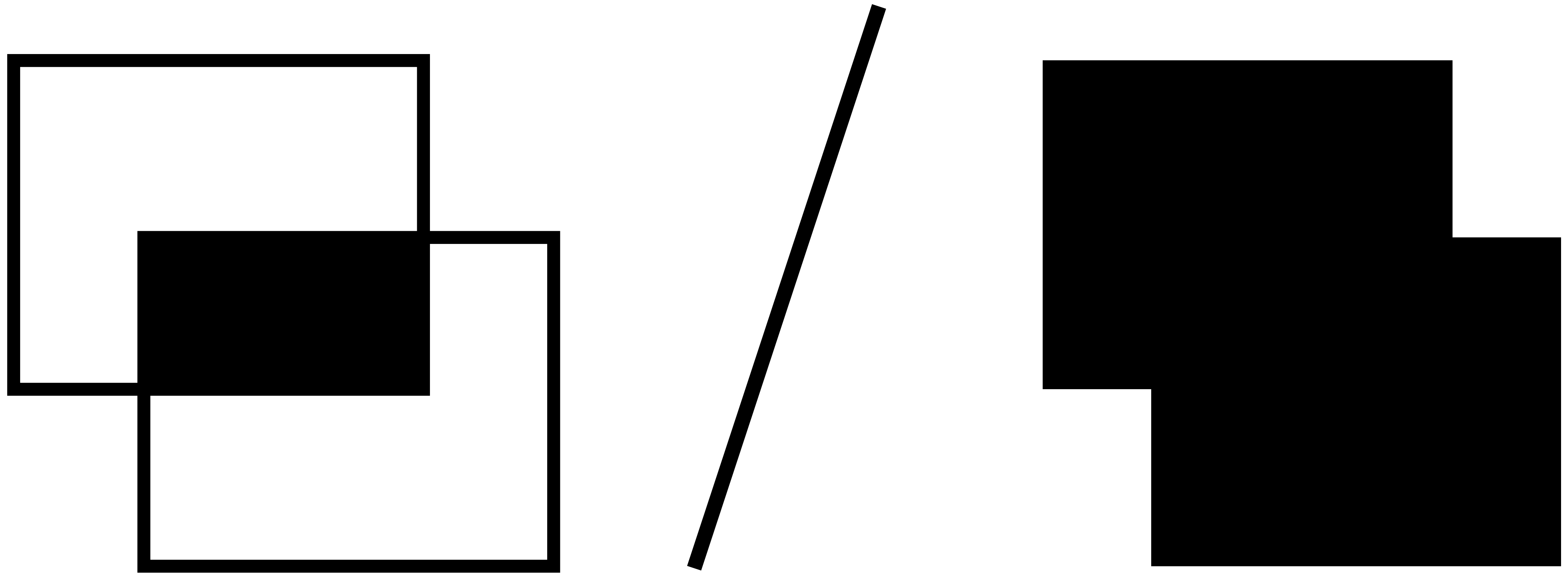
Whether similar languages share Similar Specialized Neurons?

Neuron Specialization

Neuron Structural Analysis - Analysis

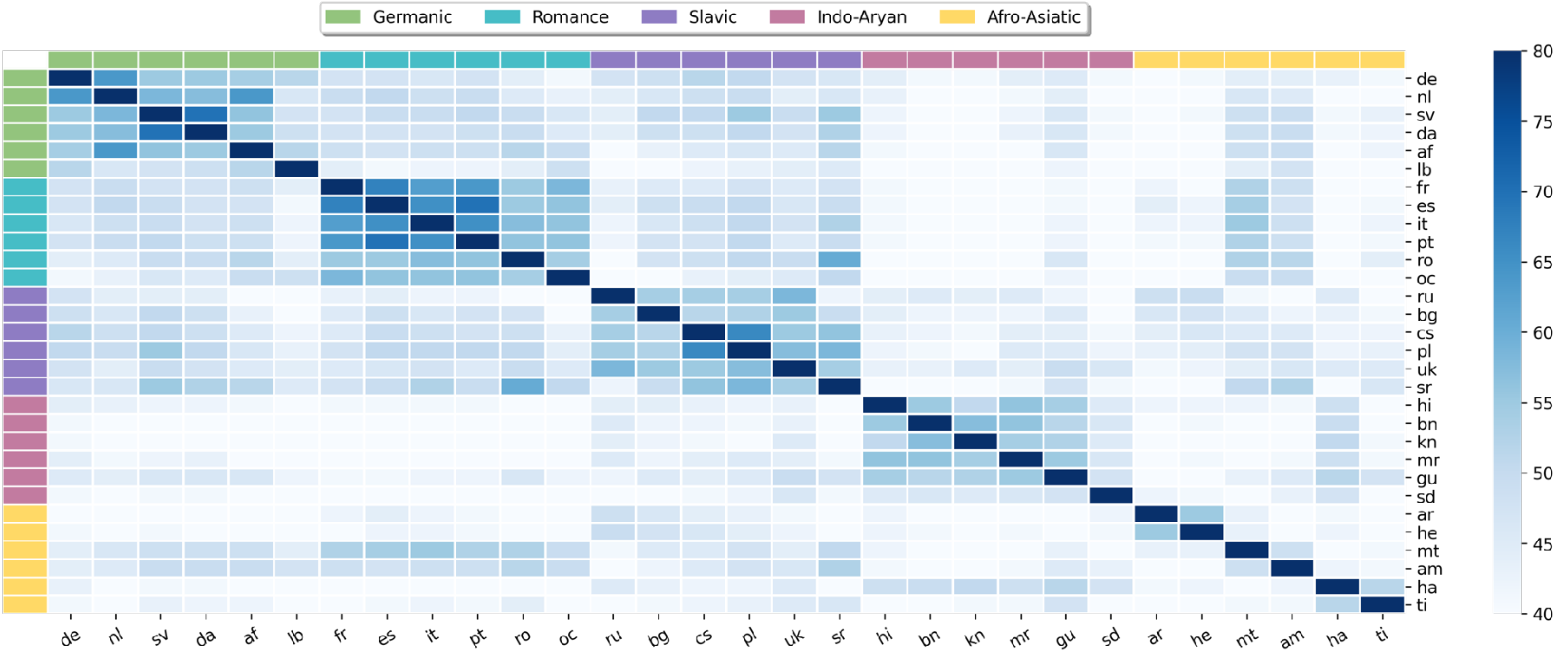
We use Intersection Over Union (IoU) to measure the similarity between two specialized neuron sets.

$$\text{IoU} = \frac{\text{Overlap}}{\text{Union}} =$$



Neuron Specialization

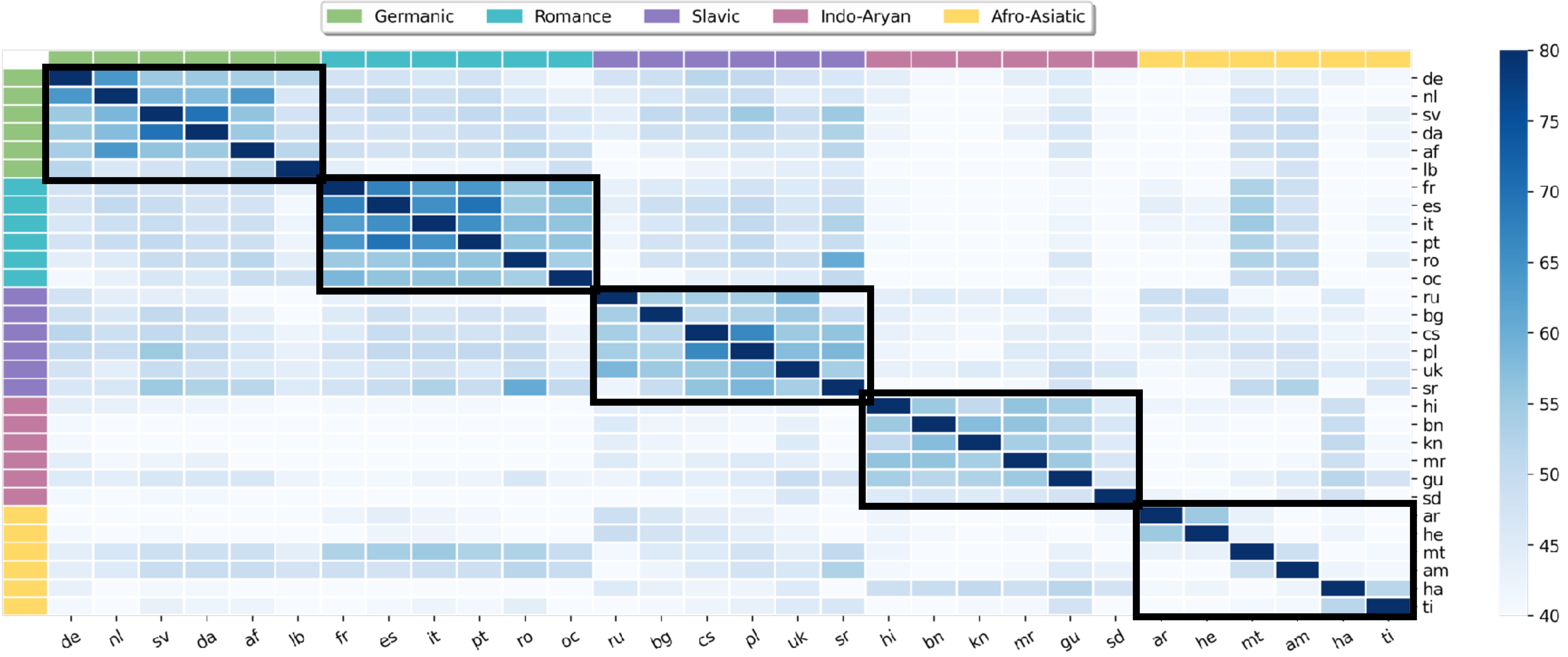
Neuron Structural Analysis - Observations



Specialized Neuron that extracted from En->X in first Decoder layer.

Neuron Specialization

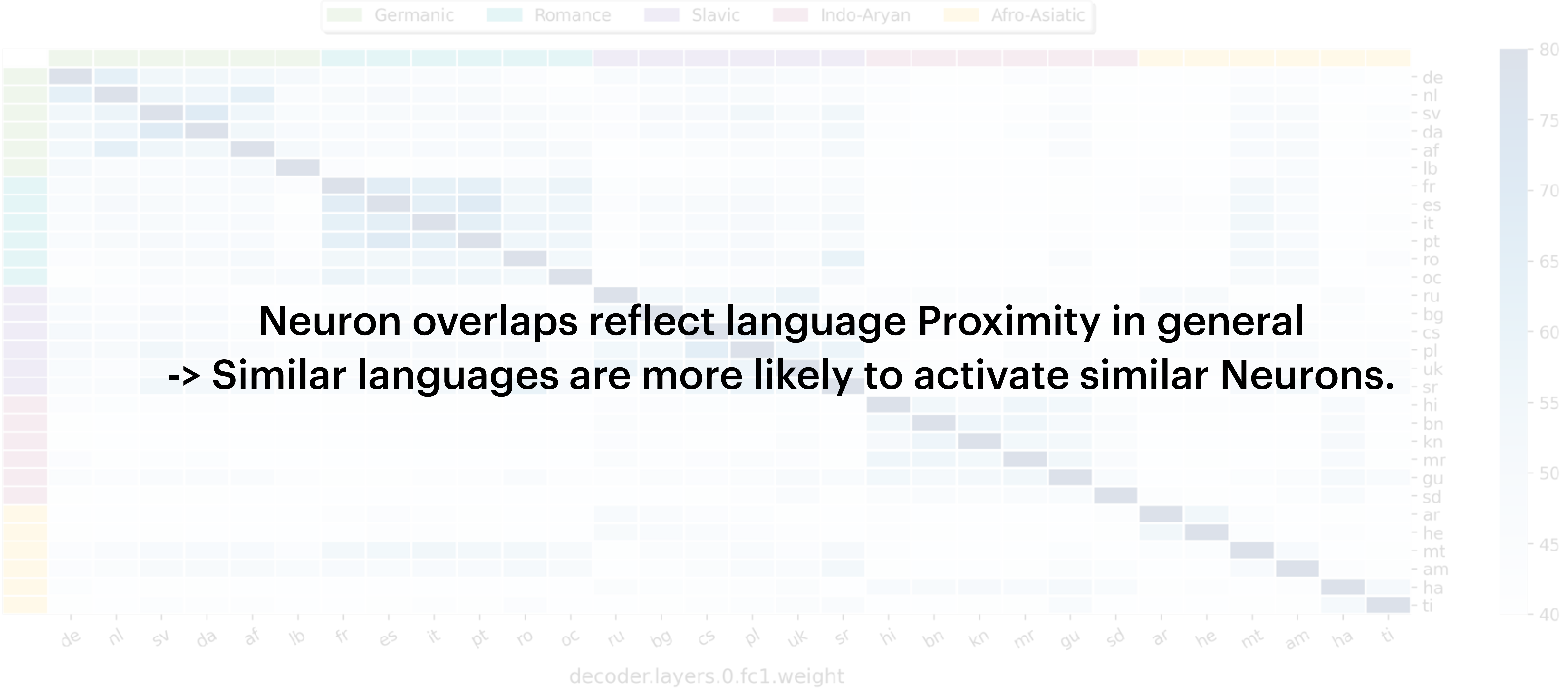
Neuron Structural Analysis - Observations



Specialized Neuron that extracted from En->X in first Decoder layer.

Neuron Specialization

Neuron Structural Analysis - Findings

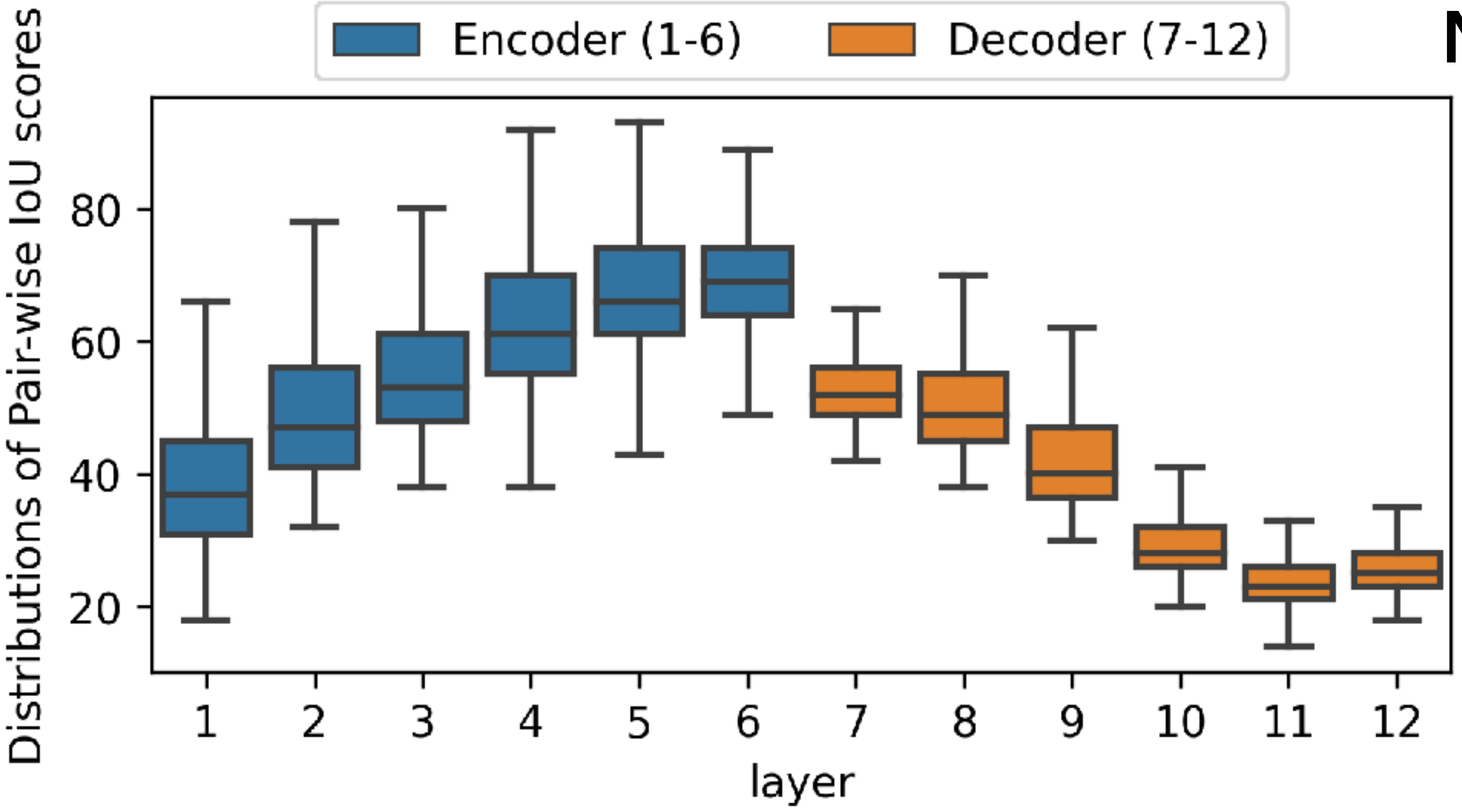


Neuron overlaps reflect language Proximity in general

-> Similar languages are more likely to activate similar Neurons.

Neuron Specialization

Neuron Structural Analysis - Observations



Neuron overlap **progresses** across layers

Encoder: **specific neurons -> agnostic neurons**

Decoder: **agnostic neurons -> specific neurons**

Similar to prior MNMT representation study[1]

1) Kudugunta, Sneha Reddy, et al. "Investigating multilingual NMT representations at scale."

Neuron Specialization Training:

Leveraging specialized neurons to modularize FFN layers in a task-specific manner.

Neuron Specialization

Method

We use identified Neurons to **modularize FC1 weights** via **sparse networks** for continual training

Neuron Specialization

Method

We use identified Neurons to **modularize FC1 weights** via **sparse networks** for continual training

$$M_{en->de} \in \{0,1\}$$

1	0	1

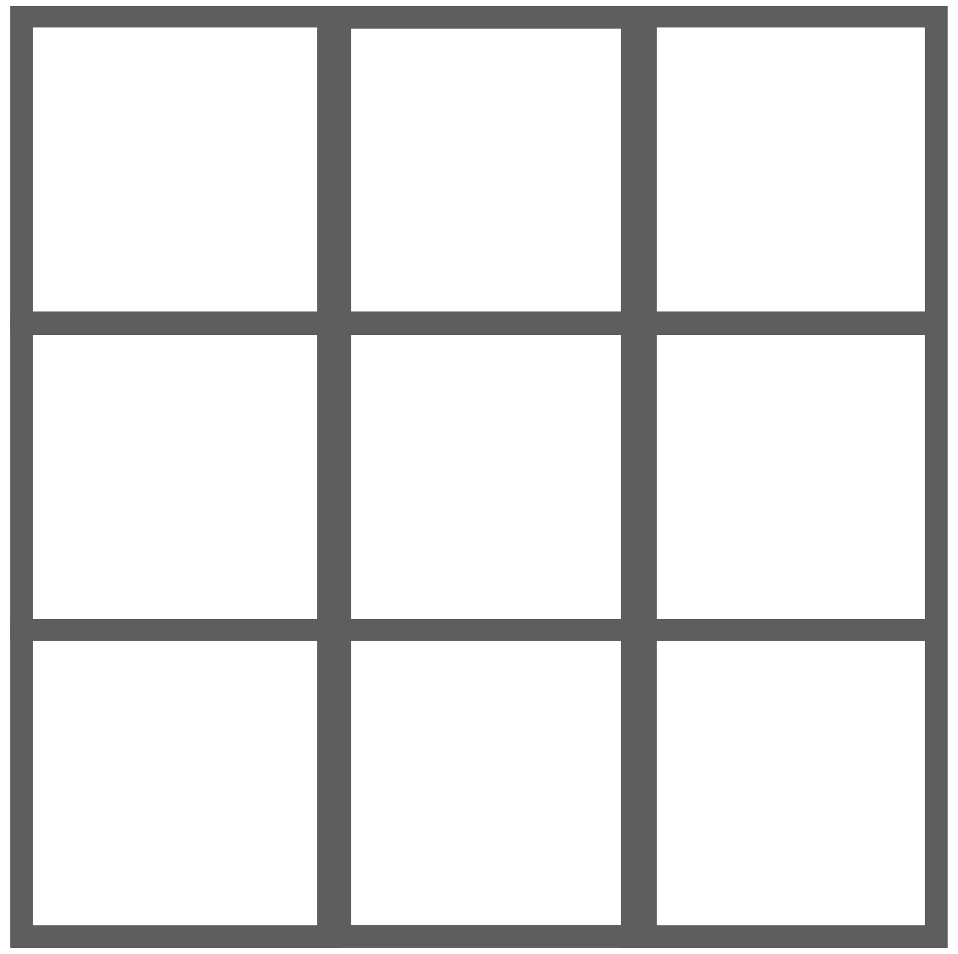
Neuron Specialization

Method

We use identified Neurons to **modularize FC1 weights** via **sparse networks** for continual training

$$M_{en->de} \in \{0,1\}$$

1	0	1

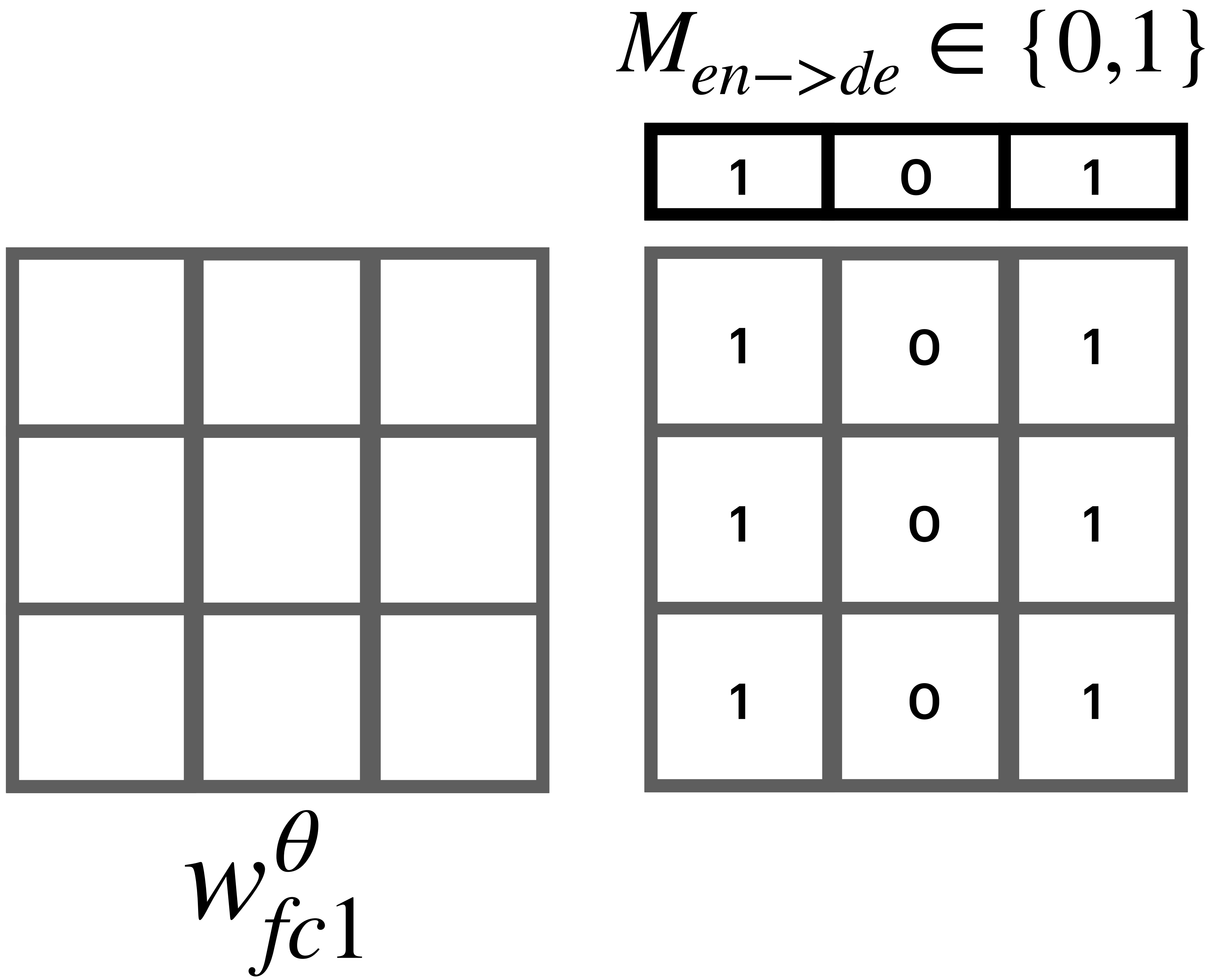


$$W^{\theta}_{fc1}$$

Neuron Specialization

Method

We use identified Neurons to **modularize FC1 weights** via **sparse networks** for continual training



Neuron Specialization

Method

We use identified Neurons to **modularize FC1 weights** via **sparse networks** for continual training

$$\mathrm{FFN}(H) = \mathrm{ReLU}(HW_1)\,W_2.$$

$$\downarrow$$

$$FFN(H) = ReLU(H(m_k^t \odot W_1))W_2.$$

$$w_{fc1}^{\theta'}$$

🔥	0	🔥
🔥	0	🔥
🔥	0	🔥

For en->de data

$$M_{en->de} \in \{0,1\}$$

1	0	1

Neuron Specialization

Method

No extra parameters are introduced!

🔥	0	🔥
🔥	0	🔥
🔥	0	🔥

en->de

🔥	🔥	0
🔥	🔥	0
🔥	🔥	0

en->ar

Neuron Specialization

Results - EC30

Consistent performance gains - on all directions

Methods	$\Delta\theta$	High (5M)			Med (1M)			Low (100K)			All (61M)		
		O2M	M2O	Avg	O2M	M2O	Avg	O2M	M2O	Avg	O2M	M2O	Avg
mT-big	-	28.1	31.6	29.9	29.7	31.6	30.6	18.9	26.0	22.4	25.5	29.7	27.7
Fine-Tune	0%	+0.3	+0.2	+0.3	+0.3	+0.2	+0.3	+0.1	-0.4	-0.2	+0.2	0	+0.1
Adapter_{Fam}	+70%	+0.7	+0.3	+0.5	+0.7	+0.3	+0.5	+1.1	+0.5	+0.8	+0.8	+0.4	+0.6
Adapter_{LP}	+87%	+1.6	+0.6	+1.1	+1.6	+0.4	+1.0	+0.4	+0.4	+0.4	+1.2	+0.5	+0.8
LaSS	0%	**+2.3**	+0.8	+1.5	**+1.7**	+0.2	+1.0	-0.1	-1.8	-1.0	+1.3	-0.3	+0.5
Random	0%	+0.9	-0.5	+0.2	+0.5	-0.7	-0.2	-0.3	-1.5	-0.9	+0.5	-0.9	-0.2
Ours^{Enc}	0%	+1.2	+1.1	+1.1	+1.0	+1.0	+1.0	+0.7	+0.8	+0.8	+1.0	+1.0	+1.0
Ours^{Dec}	0%	+1.2	+1.1	+1.1	+0.9	**+1.1**	+1.0	+0.7	**+1.1**	+0.9	+0.9	**+1.1**	+1.0
Ours	0%	+1.8	**+1.4**	**+1.6**	+1.4	**+1.1**	**+1.3**	**+1.4**	+0.9	**+1.2**	$\boxed{\textbf{+1.5}}$	$\boxed{\textbf{+1.1}}$	$\boxed{\textbf{+1.3}}$

SacreBleu Improvements over the baseline system (mT-big)

Neuron Specialization

Results - EC30

Remain Efficiency - No additional parameters

Methods	$\Delta\theta$	High (5M)			Med (1M)			Low (100K)			All (61M)		
		O2M	M2O	Avg	O2M	M2O	Avg	O2M	M2O	Avg	O2M	M2O	Avg
mT-big	-	28.1	31.6	29.9	29.7	31.6	30.6	18.9	26.0	22.4	25.5	29.7	27.7
Fine-Tune	0%	+0.3	+0.2	+0.3	+0.3	+0.2	+0.3	+0.1	-0.4	-0.2	+0.2	0	+0.1
Adapter_{Fam}	+70%	+0.7	+0.3	+0.5	+0.7	+0.3	+0.5	+1.1	+0.5	+0.8	+0.8	+0.4	+0.6
Adapter_{LP}	+87%	+1.6	+0.6	+1.1	+1.6	+0.4	+1.0	+0.4	+0.4	+0.4	+1.2	+0.5	+0.8
LaSS	0%	**+2.3**	+0.8	+1.5	**+1.7**	+0.2	+1.0	-0.1	-1.8	-1.0	+1.3	-0.3	+0.5
Random	0%	+0.9	-0.5	+0.2	+0.5	-0.7	-0.2	-0.3	-1.5	-0.9	+0.5	-0.9	-0.2
Ours^{Enc}	0%	+1.2	+1.1	+1.1	+1.0	+1.0	+1.0	+0.7	+0.8	+0.8	+1.0	+1.0	+1.0
Ours^{Dec}	0%	+1.2	+1.1	+1.1	+0.9	**+1.1**	+1.0	+0.7	**+1.1**	+0.9	+0.9	**+1.1**	+1.0
Ours	0%	+1.8	**+1.4**	**+1.6**	+1.4	**+1.1**	**+1.3**	**+1.4**	+0.9	**+1.2**	**+1.5**	**+1.1**	**+1.3**

SacreBleu Improvements over the baseline system (mT-big)

Neuron Specialization

Results - Efficiency Comparison

Model	$\Delta\theta$	ΔT_{subnet}	Δ Memory
Adapter_{LP}	+87%	n/a	1.42 GB
LaSS	0%	+33 hours	9.84 GB
Ours	0%	+5 minutes	3e-3 GB

Results reported based on EC30 with 4 A6000 GPUs

Our approach is highly **efficient**, facilitating the **adaptation to massively multilingual models.**

Neuron Specialization

Results - Wider and Deeper Models

Methods	SacreBLEU			COMET		
	Big	Wide	Deep	Big	Wide	Deep
Baseline	27.7	28.3	28.8	79.1	79.7	80.0
Ours	**29.0**	**29.4**	**29.7**	**80.0**	**80.5**	**80.7**

Performance comparison between baseline models and our methods on **three configurations**.

The effectiveness on **larger configurations.**

Neuron Specialization

Results - beyond ReLU

Methods	All (61M)		
	SacreBLEU	ChrF	Comet
mT-big^{relu}	27.7	52.2	79.1
Ours^{relu}	29.0	53.3	80.0
mT-big^{gelu}	27.9	52.3	79.2
Ours^{gelu}	28.9	53.2	80.1

Performance comparison between the relu and gelu backbone models and our method.

GeLU

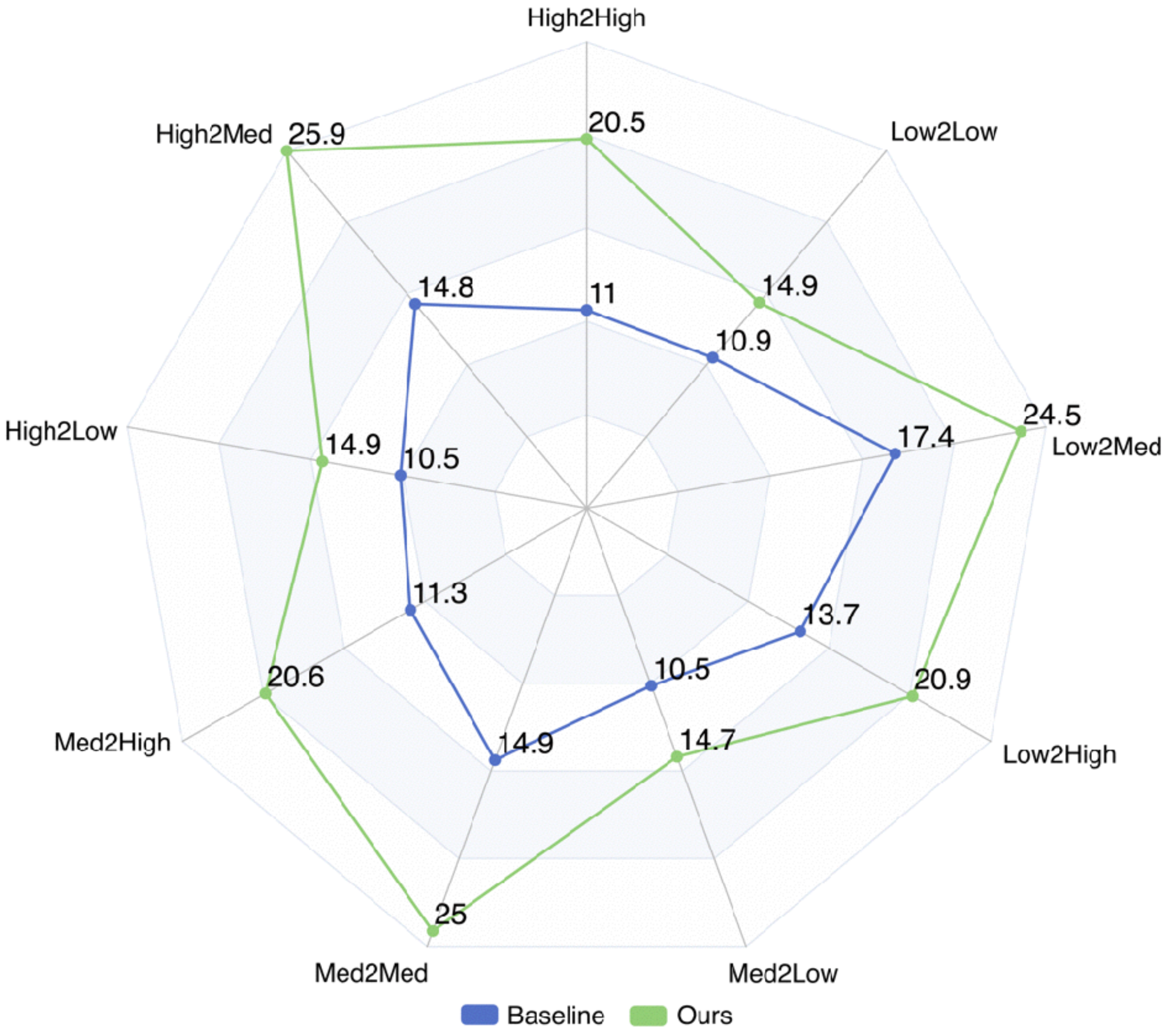
Active Neurons: >0

Inactive Neurons: <=0

Other threshold may deliver better results -> Future work

Neuron Specialization

Results - Zero-Shot Translation



For 870 Zero-Shot Directions,
compared to the baseline system,
847 improved, 23 had minor declines.

Neuron Specialization

Analysis and Discussion — How much we can alleviate interference?

Lang Size	De 5m	Es 5m	Cs 5m	Hi 5m	Ar 5m	Lb 100k	Ro 100k	Sr 100k	Gu 100k	Am 100k	High Avg	Low Avg
One-to-Many												
Bilingual	36.3	24.6	28.7	43.9	23.7	5.5	16.2	17.8	12.8	4.1	31.8	11.3
mT-big	-4.7	-1.5	-3.6	-4.4	-4.7	+9.0	+8.9	+6.2	+13.9	+3.1	-3.7	+8.2
Many-to-One												
Bilingual	39.1	24.5	32.6	35.5	30.8	8.7	19.5	21.3	7.0	8.7	32.7	13.0
mT-big	-1.5	+0.9	+0.2	-1.8	-2.3	+13.7	+11.9	+10.3	+18.2	+12.5	-1.1	+13.3

SacreBleu Improvements over bilingual systems

Evidence of Interference: **worse performance on high-resource languages.**

Neuron Specialization

Analysis and Discussion — How much we can alleviate interference?

Lang Size	De 5m	Es 5m	Cs 5m	Hi 5m	Ar 5m	Lb 100k	Ro 100k	Sr 100k	Gu 100k	Am 100k	High Avg	Low Avg
One-to-Many												
Bilingual	36.3	24.6	28.7	43.9	23.7	5.5	16.2	17.8	12.8	4.1	31.8	11.3
mT-big	-4.7	-1.5	-3.6	-4.4	-4.7	+9.0	+8.9	+6.2	+13.9	+3.1	-3.7	+8.2
Ours	-2.0	-0.2	-1.7	-2.4	-3.0	+10.8	+10.0	+8.2	+16.4	+3.7	-1.9	+9.8
Many-to-One												
Bilingual	39.1	24.5	32.6	35.5	30.8	8.7	19.5	21.3	7.0	8.7	32.7	13.0
mT-big	-1.5	+0.9	+0.2	-1.8	-2.3	+13.7	+11.9	+10.3	+18.2	+12.5	-1.1	+13.3
Ours	-0.3	+1.7	+1.8	-0.2	-0.3	+15.3	+12.4	+11.3	+19.6	+14.1	+0.3	+14.5

SacreBleu Improvements over bilingual systems

Our method reduces interference while further encouraging knowledge transfer!

Conclusions

Neuron Analysis

Show Intrinsic modularity in multi-task models without modification.

Proposed Method

Presents Consistent Performance Gains on large-scale experiments.

Neuron Specialization

Efficiency

Introduce 0 extra Trainable Parameters.

Understanding

fundamental properties in FFN Modules & Multi-task.

Neuron Specialization

Neuron Structural Analysis - Observations

Evidence of how specialized neuron overlaps correlated with language similarity - by quantifying the correlation between neuron overlaps and linguistic distances.