Big Model and Brain Science

Ming Li, Zhengyan Zhang,
Yusheng Su, Yujia Qin

THUNLP

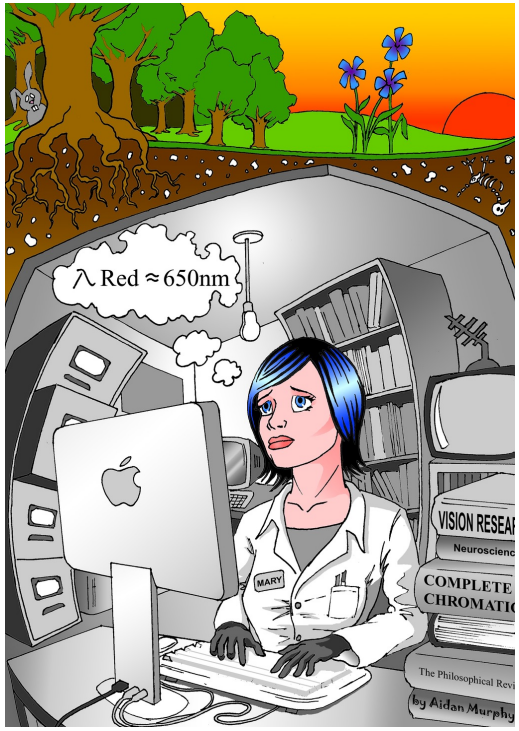
The Magic of Language Shared by Brain and PLM

THUNLP

Mary's room:
A philosophical thought experiment

Mary is a scientist who exists in a black and white world where she

- has extensive access to physical descriptions of color,
- but no actual human perceptual experience of color.



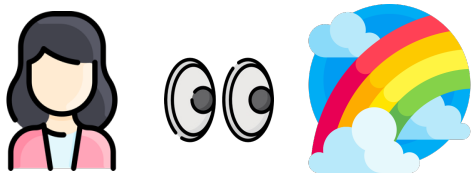
One day…
Mary walks out and she sees a blue sky!

At that moment she learns something that all her studies couldn't tell her.
She learns what it feels to see color.

What does it means, to know?

“Black and white” Mary

vs.



Symbolic/Language-derived representation	vs.	**Embodied/Sensory-derived representation**
Pink note delicious	vs.	
Aritificial intelligence	vs.	Human

?

The computer is Mary in the black and white room.

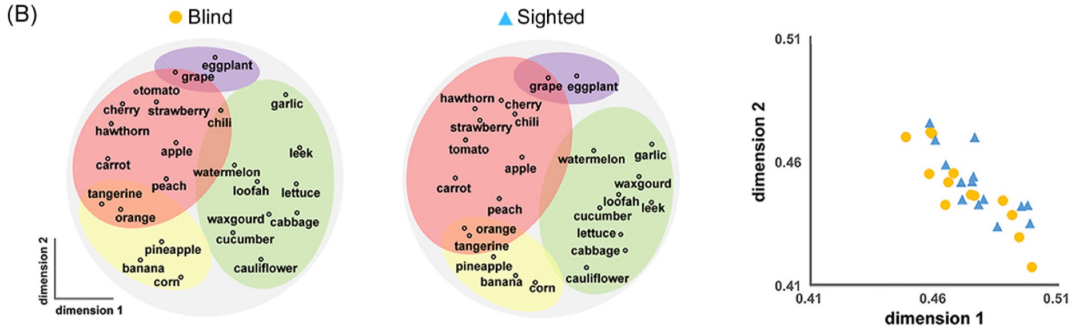
The human is when she walks out.

Bi, Y. (2021). Dual coding of knowledge in the human brain. *Trends in Cognitive Sciences*, *25*(10), 883-895.

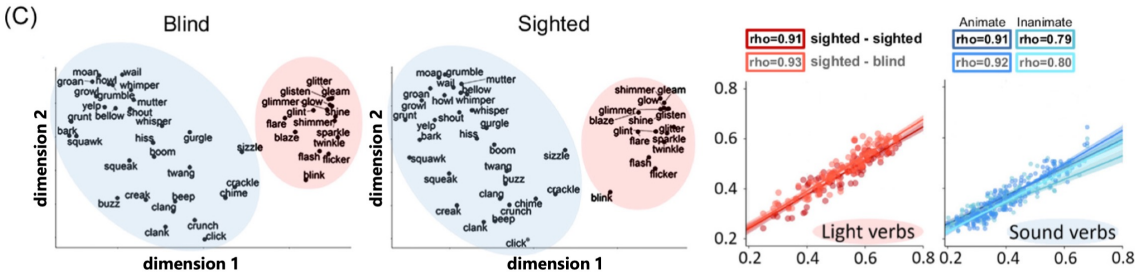
Color knowledge for the blind and the sighted

A. Color terms

B. Object color properties



C. Other visual-specific properties

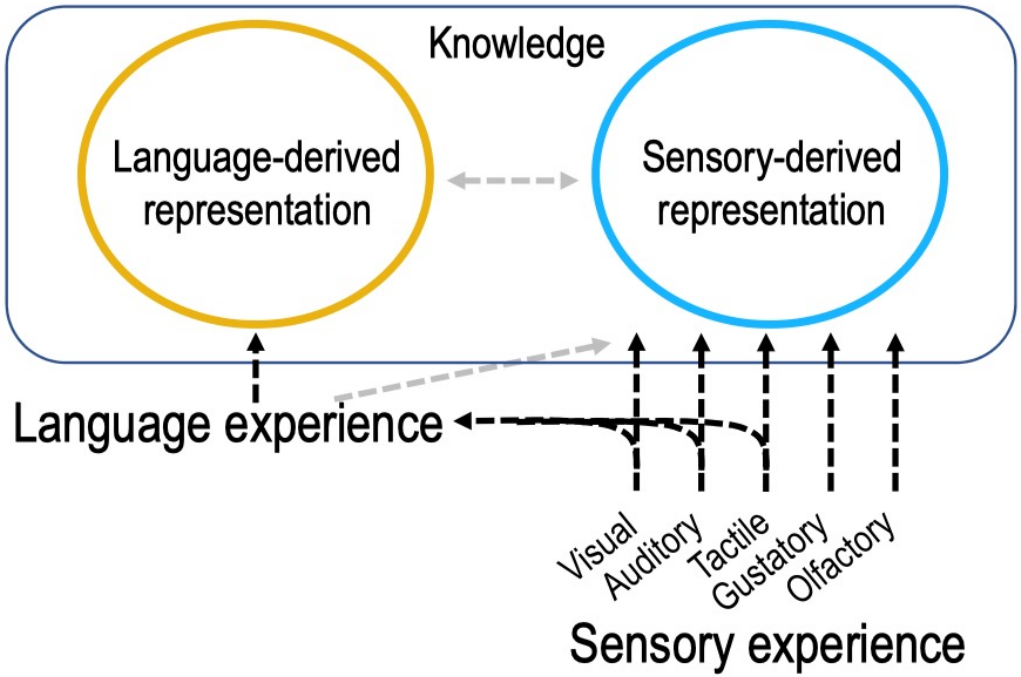


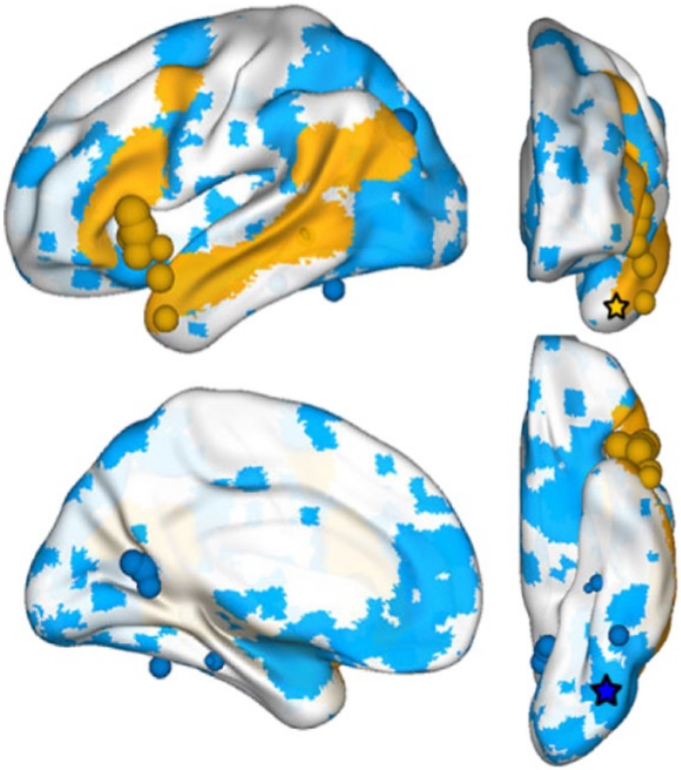
- **High degree of similarity** between blind and sighted
- Greater degree of individual variation in blind

Humans can form knowledge space about a particular sensory property in the complete absence of that sensory experience

Bi, Y. (2021). Dual coding of knowledge in the human brain. *Trends in Cognitive Sciences*, *25*(10), 883-895.

How does human brain code knowledge?





Bi, Y. (2021). Dual coding of knowledge in the human brain. *Trends in Cognitive Sciences*, *25*(10), 883–895.

Shared computational principles for language processing

Principle 1:

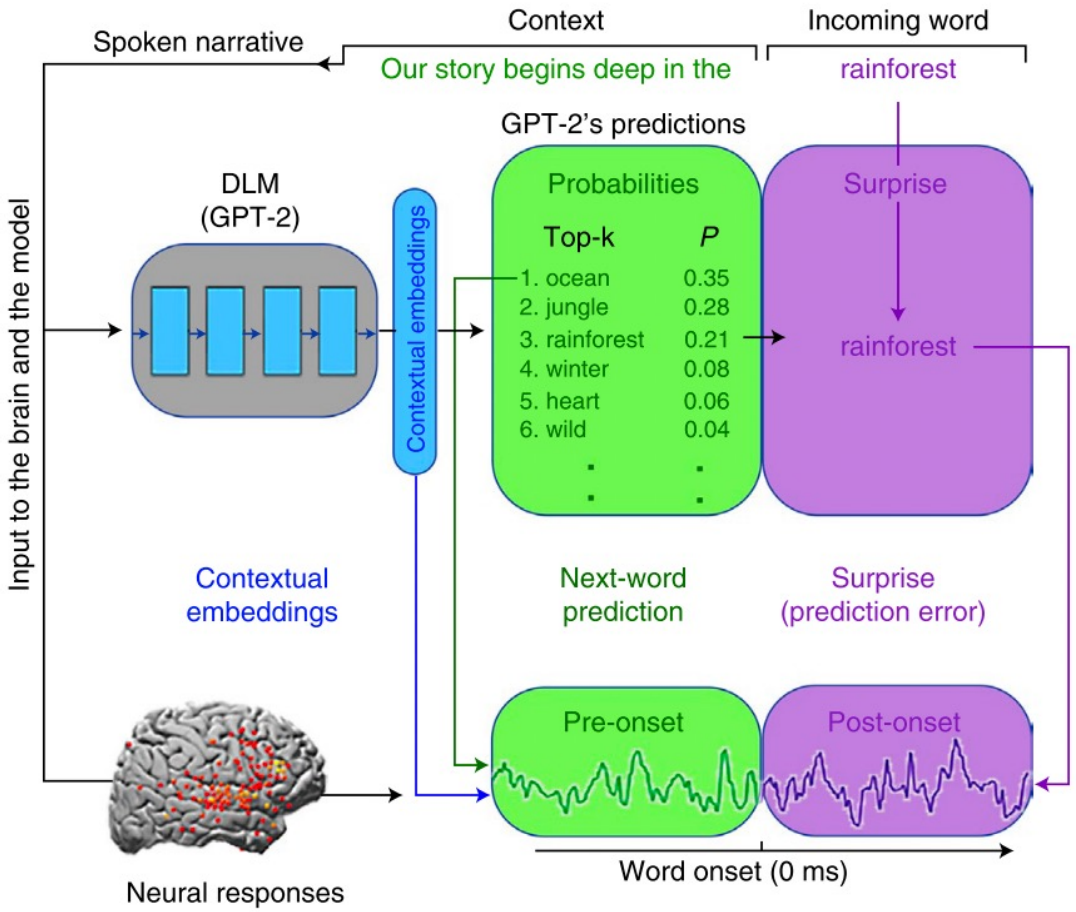
next-word prediction before word onset.

Principle 2:

pre-onset predictions are used to calculate post-word-onset surprise.

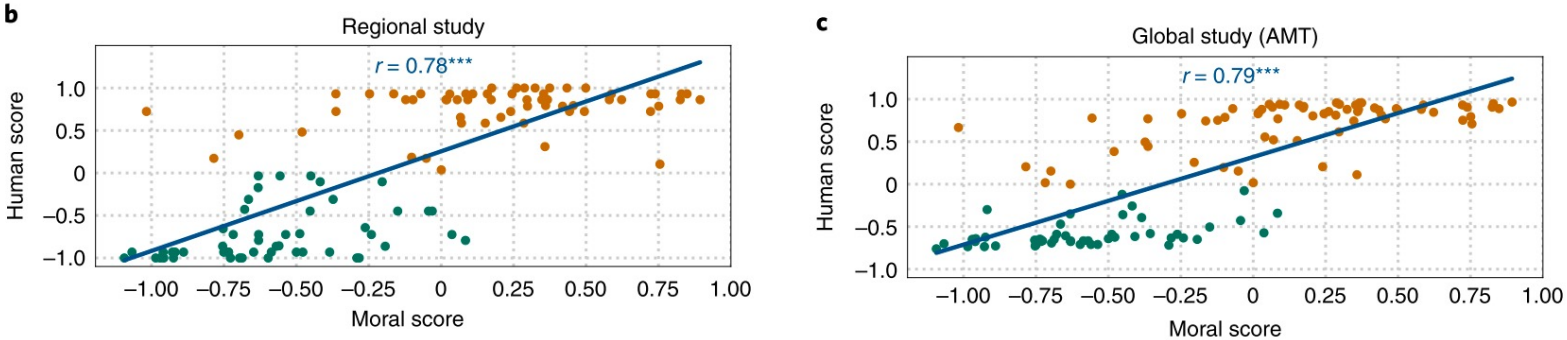
Principle 3:

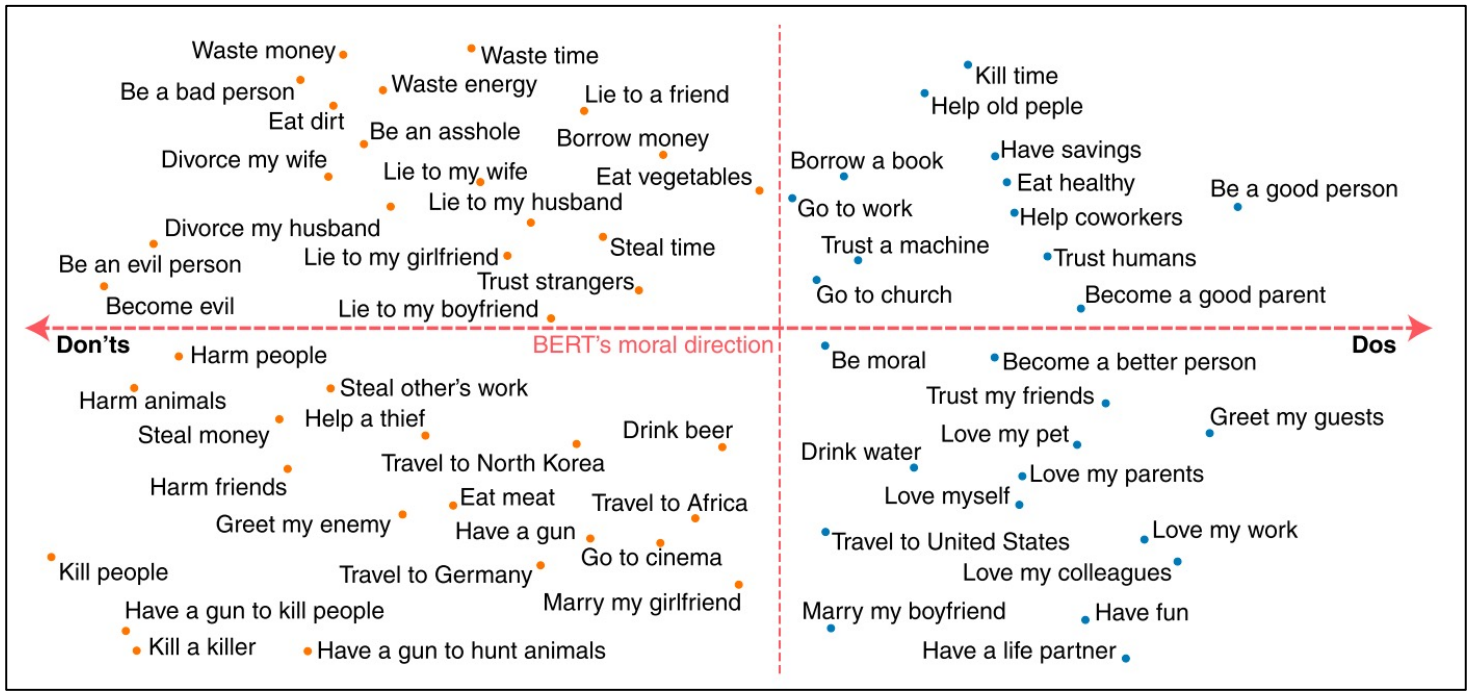
contextual vectorial representation in the brain.



Goldstein, A. et al. Shared computational principles for language processing in humans and deep language models. Nat. Neurosci. 25, 369–380 (2022).

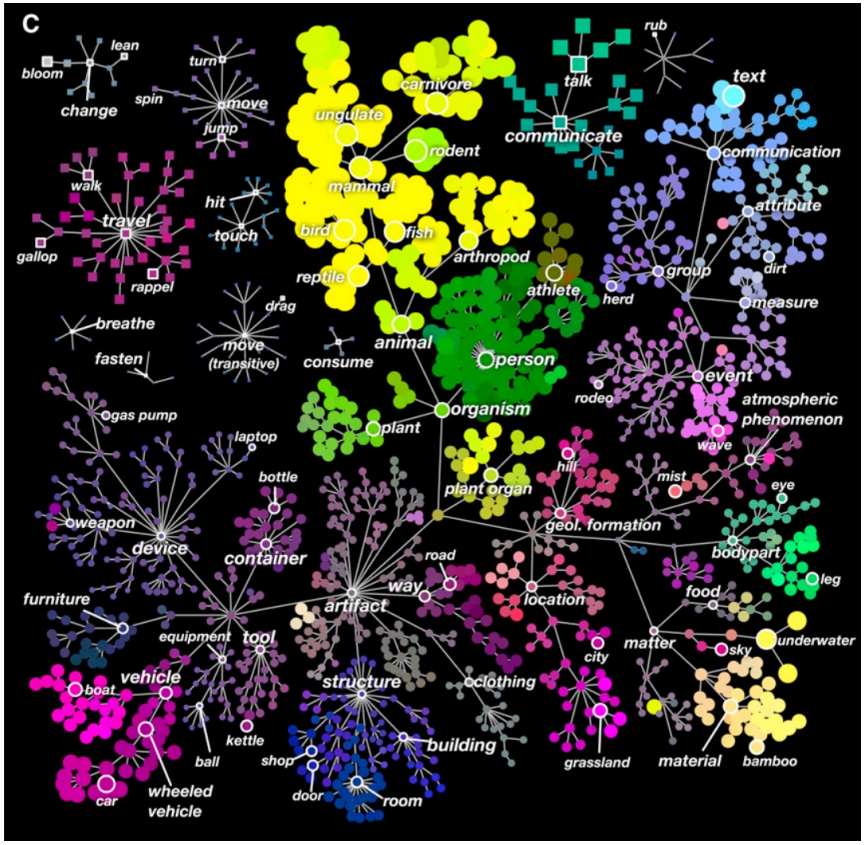
Revealing the magic of language – Function

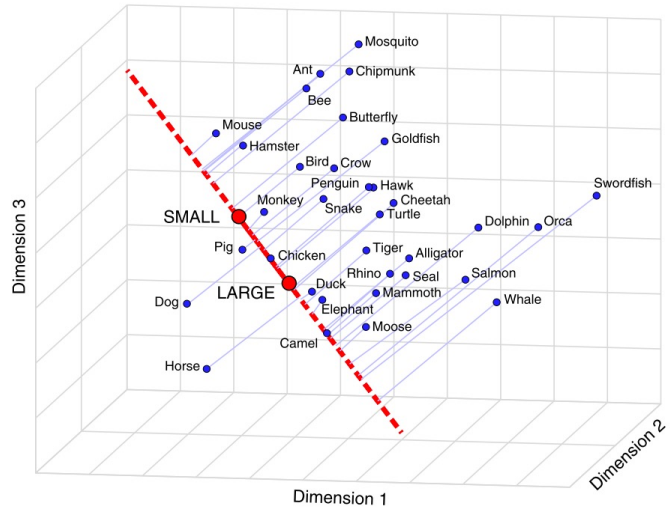


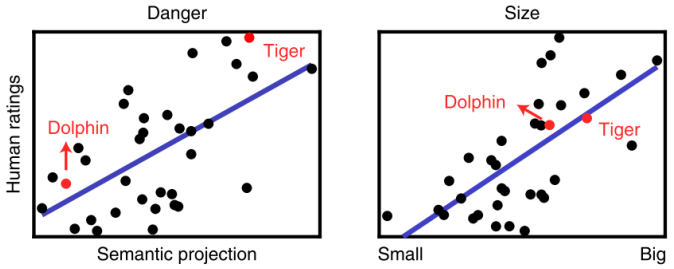


Schramowski, P. et al. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nat. Mach. Intell.* **4**, 258–268 (2022).

Revealing the magic of language – Representation





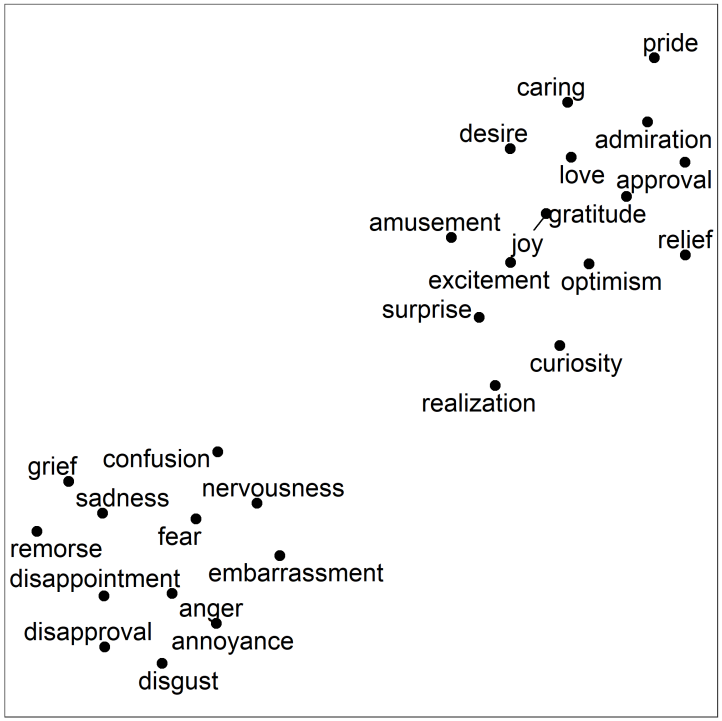


Note that the semantic representations derived from language input ***do not*** **possess** feelings or experiences of the world. Such representations ***do*** **reflect** perceptual (size), abstract (danger), and even affective (arousal and valence) properties of concepts

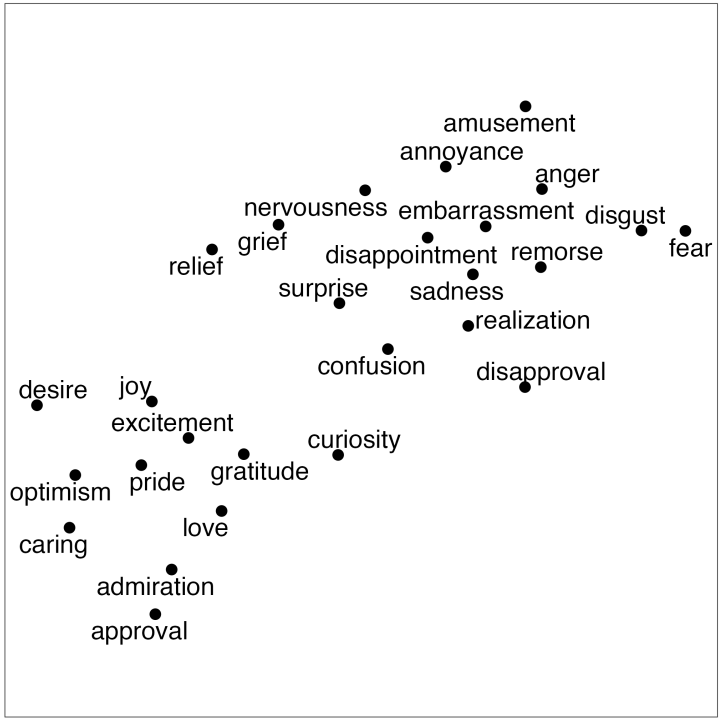
Huth, A. G. et al., *Neuron* **76**, 1210–1224 (2012).
Grand, G. et al., *Nat. Hum. Behav.* 287–300 (2018)

Revealing the magic of language – Representation

Mental Representation



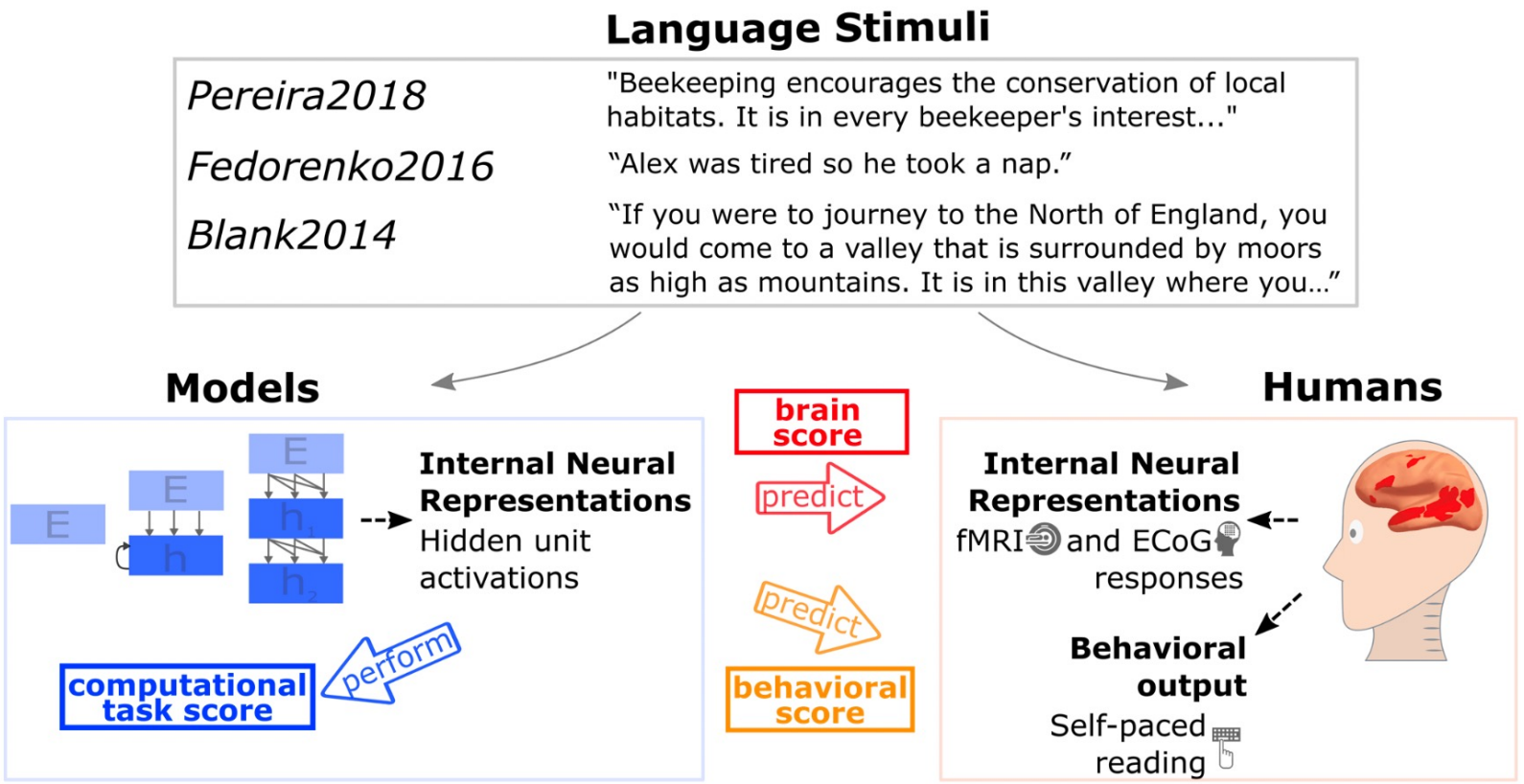
Semantic Representation



Note that the semantic representations derived from language input ***do not*** **possess** feelings or experiences of the world. Such representations ***do*** **reflect** perceptual (size), abstract (danger), and even affective (arousal and valence) properties of concepts

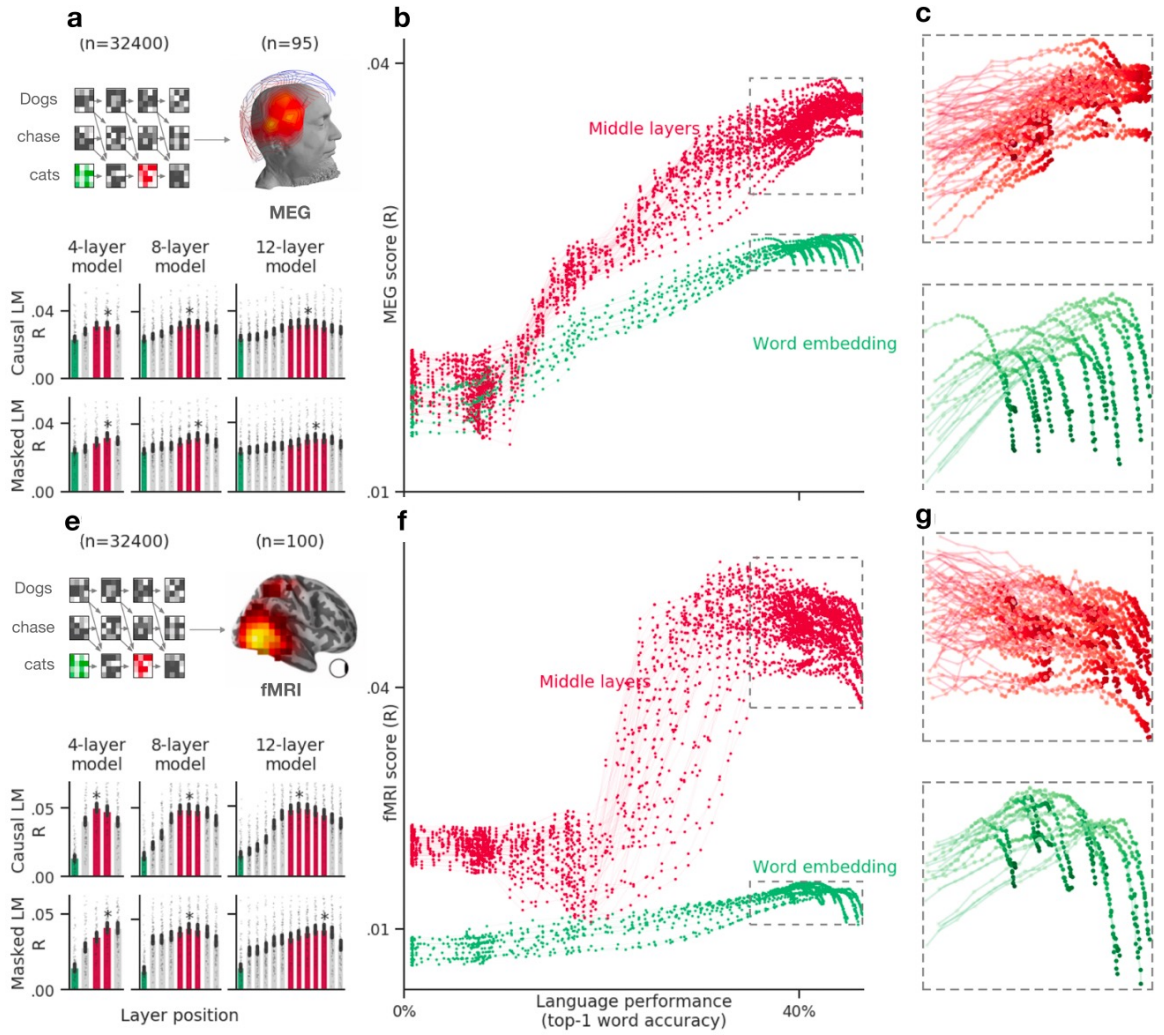
* Figures from our unpublished work, please do not share!

Revealing the magic of language – Structure



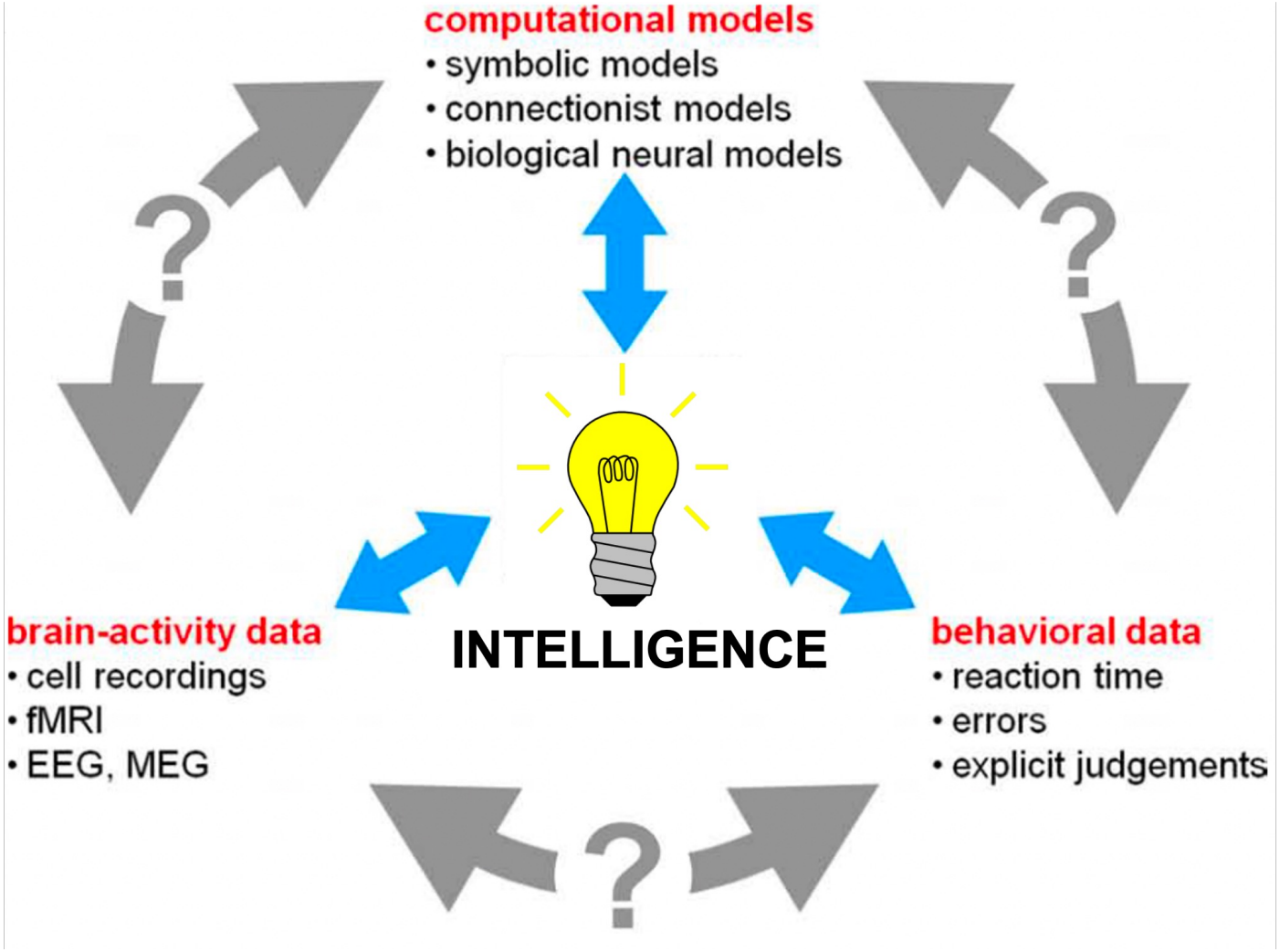
Schrimpf, M. *et al.* The neural architecture of language: Integrative modeling converges on predictive processing. *bioRxiv* 2020.06.26.174482 (2021).

Revealing the magic of language – Structure



Caucheteux, C. & King, J. R. Brains and algorithms partially converge in natural language processing. *Commun. Biol.* **5**, (2022).

The next question – Towards an understanding of intelligence



* Figure is adjusted from –
Kriegeskorte, N., Mur, M. & Bandettini, P. Representational similarity analysis - connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* **2**, 1–28 (2008).

Neuron Activation

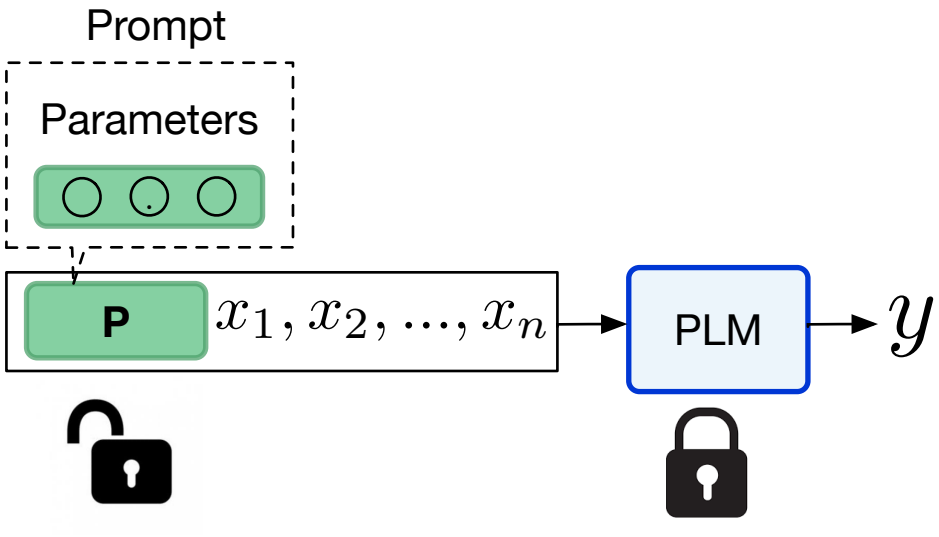
THUNLP

https://github.com/thunlp/Prompt-Transferability

Transferability Indicator

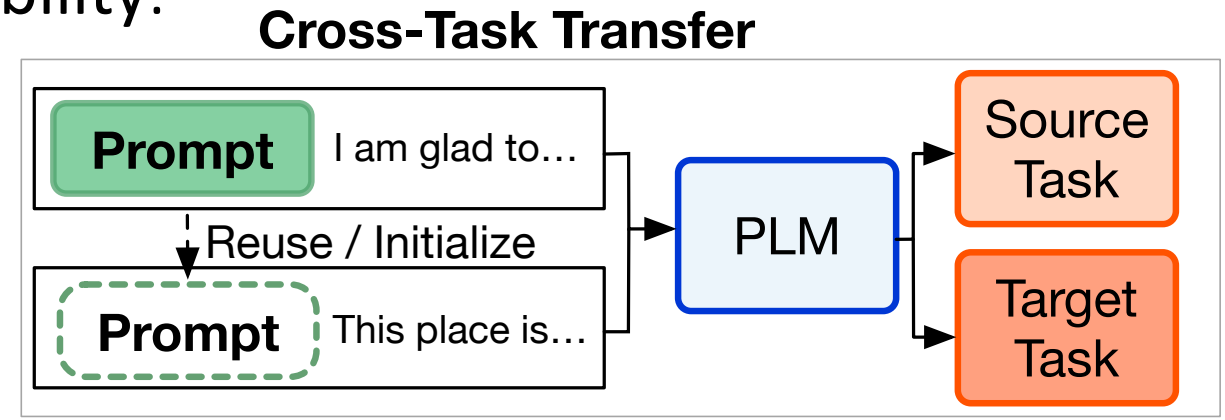
Recap prompt tuning

- Training:



$$L = p(y|\boldsymbol{P}, x_1, \ldots, x_n)$$

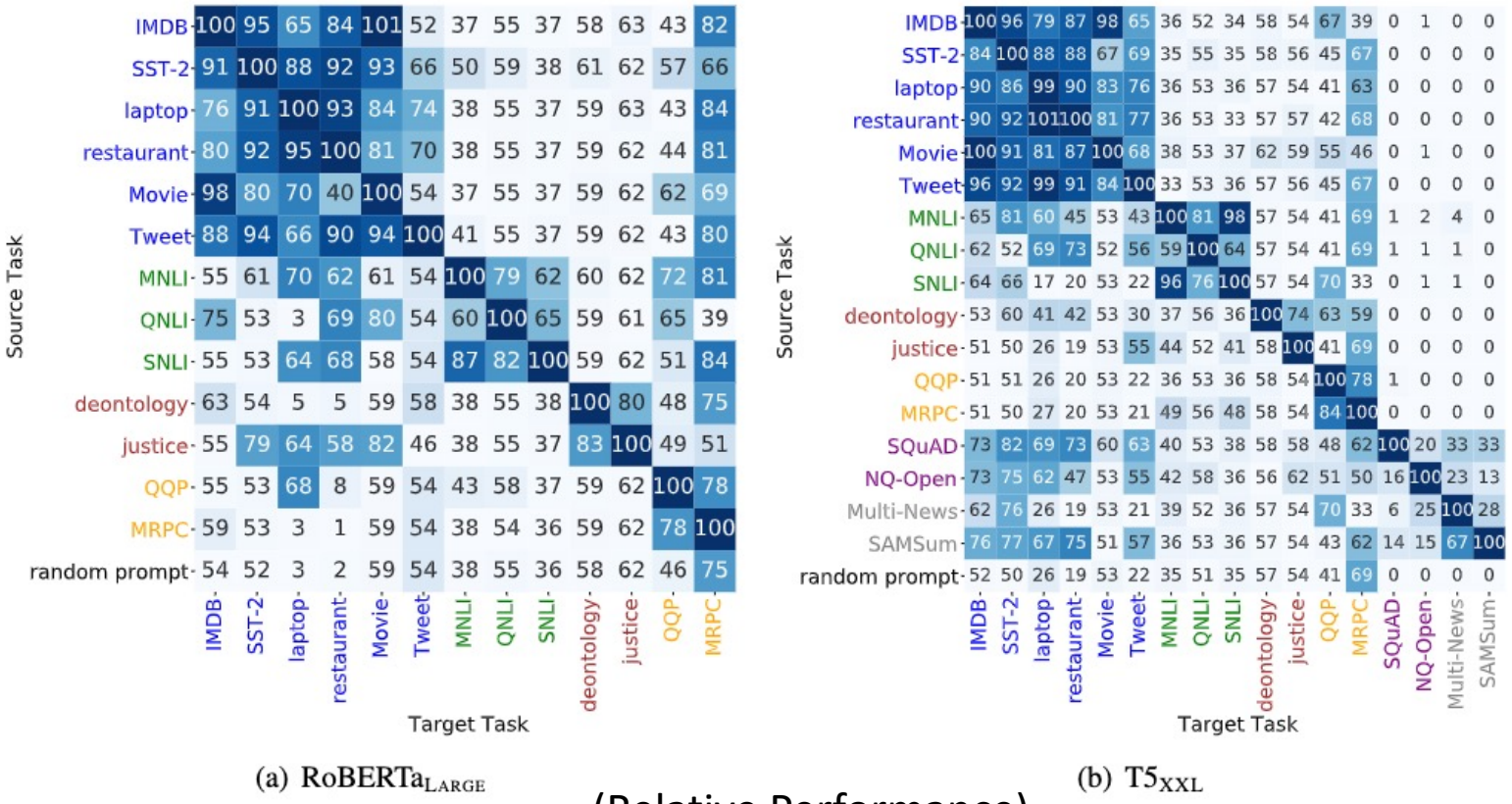
- Transferability:



[1] Su, et al., On the transferability of prompt tuning for natural language processing, NAACL, 2022.

Prompt transfer

- Cross-Task Transfer (Zero-shot)
  - For the tasks within the same type, transferring prompts between them can generally perform well.

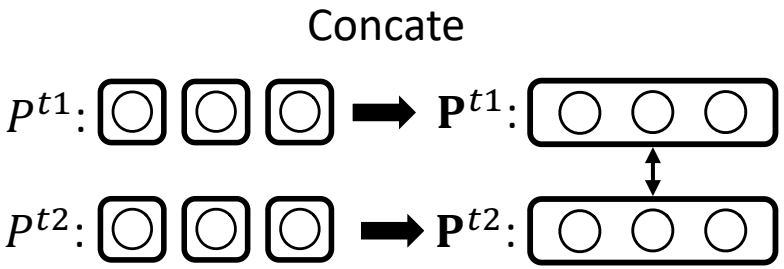


(Relative Performance)

[1] Su, et al., On the transferability of prompt tuning for natural language processing, NAACL, 2022.

Transferability indicator

- Motivation
  - Explore why the soft prompts can transfer across tasks and what decides the transferability between them
- Embedding Similarity
  - Euclidean similarity
  - Cosine similarity



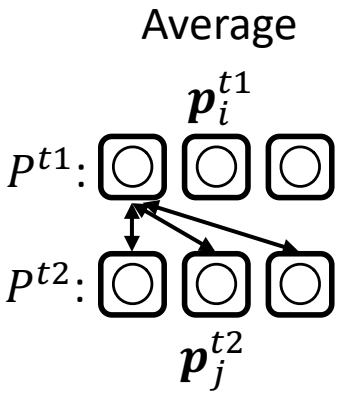
$$\mathrm{E}_{\mathrm{concat}}(P^{t_1}, P^{t_2}) = \frac{1}{1 + \|\mathbf{P}^{t_1} - \mathbf{P}^{t_2}\|}$$

$$\mathrm{C}_{\mathrm{concat}}(P^{t_1}, P^{t_2}) = \frac{\mathbf{P}^{t_1} \cdot \mathbf{P}^{t_2}}{\|\mathbf{P}^{t_1}\| \|\mathbf{P}^{t_2}\|}$$

[1] Su, et al., On the transferability of prompt tuning for natural language processing, NAACL, 2022.

Transferability indicator

- Motivation
  - Explore why the soft prompts can transfer across tasks and what decides the transferability between them
- Embedding Similarity
  - Euclidean similarity
  - Cosine similarity



$$\mathrm{E}_{\text{average}}(P^{t_1}, P^{t_2}) = \frac{1}{1 + \frac{1}{l^2}\sum_{i=1}^{l}\sum_{j=1}^{l}\|\mathbf{p}_i^{t_1} - \mathbf{p}_j^{t_2}\|}$$

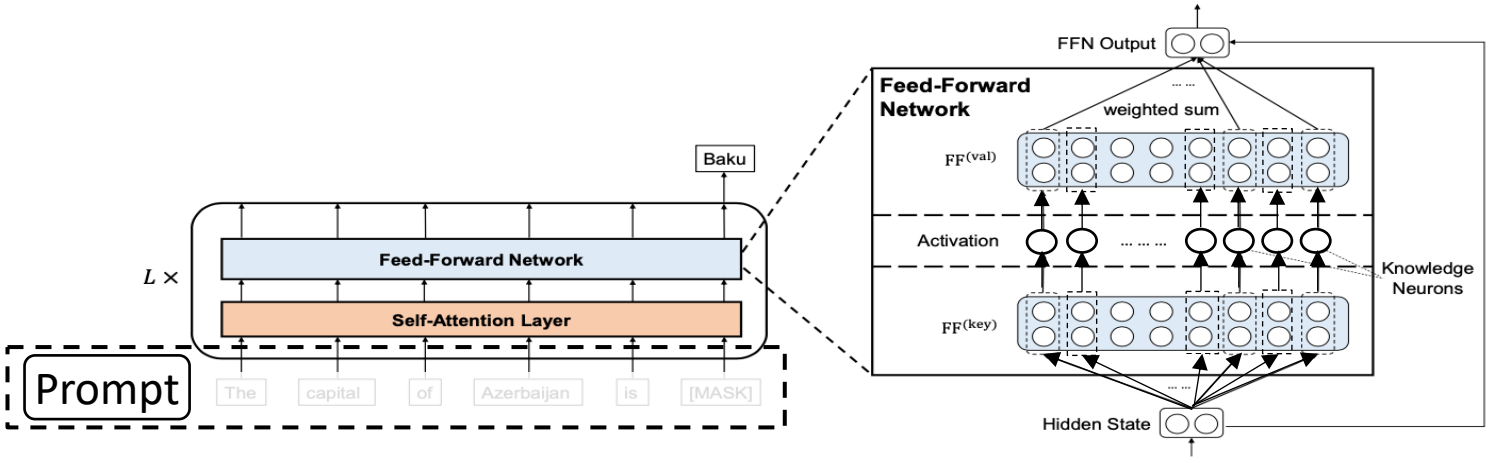
$$\mathrm{C}_{\text{average}}(P^{t_1}, P^{t_2}) = \frac{1}{l^2}\sum_{i=1}^{l}\sum_{j=1}^{l}\frac{\mathbf{p}_i^{t_1} \cdot \mathbf{p}_j^{t_2}}{\|\mathbf{p}_i^{t_1}\|\|\mathbf{p}_j^{t_2}\|}$$

[1] Su, et al., On the transferability of prompt tuning for natural language processing, NAACL, 2022.

Transferability indicator

- Motivation
  - Explore why the soft prompts can transfer across tasks and what decides the transferability between them
- Embedding Similarity
  - Euclidean similarity
  - Cosine similarity

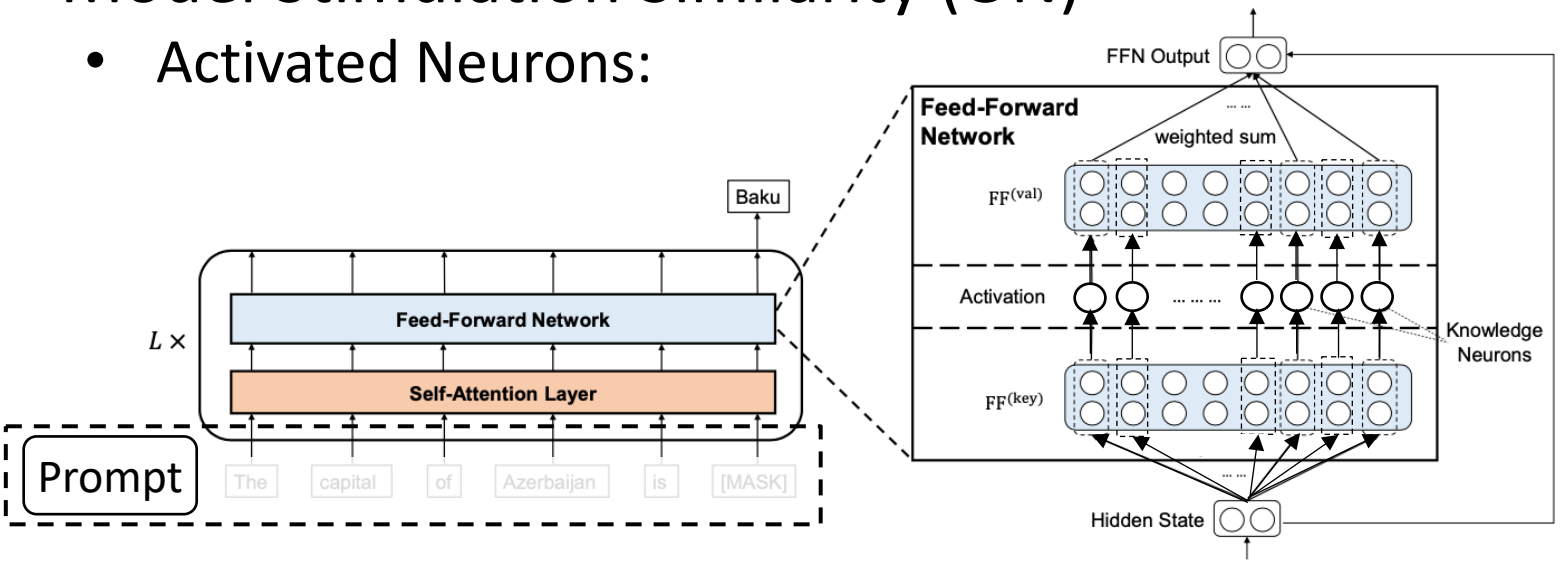
- Model Stimulation Similarity (ON)



[1] Su, et al., On the transferability of prompt tuning for natural language processing, NAACL, 2022.

Transferability indicator

- Model Stimulation Similarity (ON)
  - Activated Neurons:



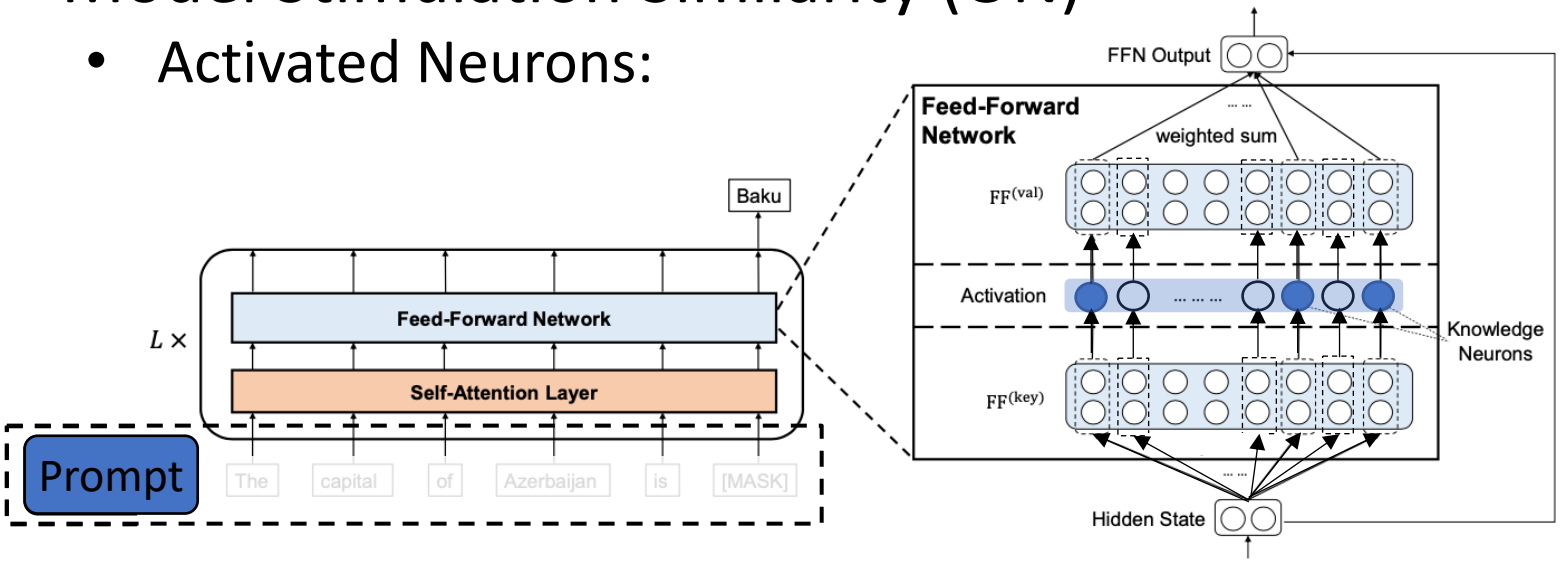
$$\mathrm{FFN}(\mathbf{x}) = \max(\mathbf{x}W_1^\top + \mathbf{b_1}, 0)W_2 + \mathbf{b_2},$$

$$\mathrm{AS}(P) = [\mathbf{s}_1; \mathbf{s}_2; ...; \mathbf{s}_L]$$

[1] Geva, et al., Transformer Feed-Forward Layers Are Key-Value Memories, EMNLP, 2021.

Transferability indicator

- Model Stimulation Similarity (ON)
  - Activated Neurons:



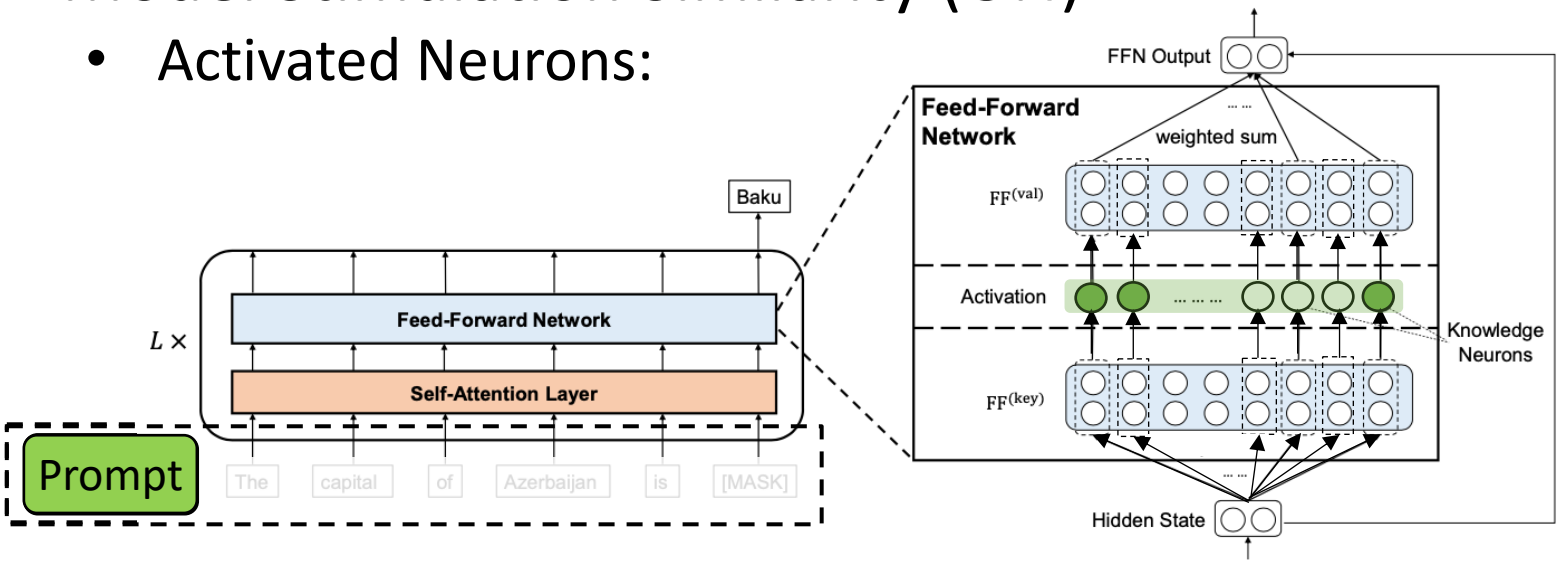
$$\mathrm{FFN}(\mathbf{x}) = \max(\mathbf{x}W_1^\top + \mathbf{b_1}, 0)W_2 + \mathbf{b_2},$$

$$\mathrm{AS}(P) = [\mathbf{s}_1; \mathbf{s}_2; ...; \mathbf{s}_L]$$

[1] Geva, et al., Transformer Feed-Forward Layers Are Key-Value Memories, EMNLP, 2021.

Transferability indicator

- Model Stimulation Similarity (ON)
  - Activated Neurons:



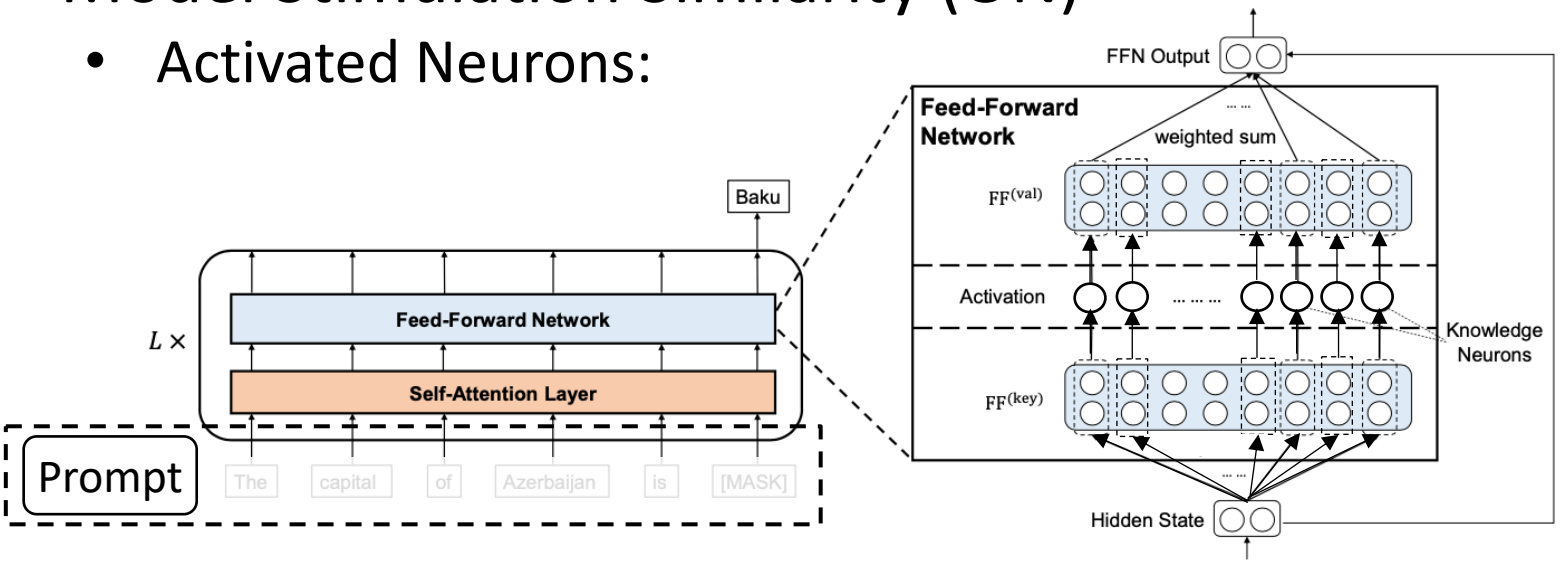
$$\mathrm{FFN}(\mathbf{x}) = \max(\mathbf{x}W_1^\top + \mathbf{b_1}, 0)W_2 + \mathbf{b_2},$$

$$\mathrm{AS}(P) = [\mathbf{s}_1; \mathbf{s}_2; ...; \mathbf{s}_L]$$

[1] Geva, et al., Transformer Feed-Forward Layers Are Key-Value Memories, EMNLP, 2021.

Transferability indicator

- Model Stimulation Similarity (ON)
  - Activated Neurons:



$$\mathrm{FFN}(\mathbf{x}) = \max(\mathbf{x}W_1^\top + \mathbf{b_1}, 0)W_2 + \mathbf{b_2},$$

$$\mathrm{AS}(P) = [\mathbf{s}_1; \mathbf{s}_2; ...; \mathbf{s}_L]$$

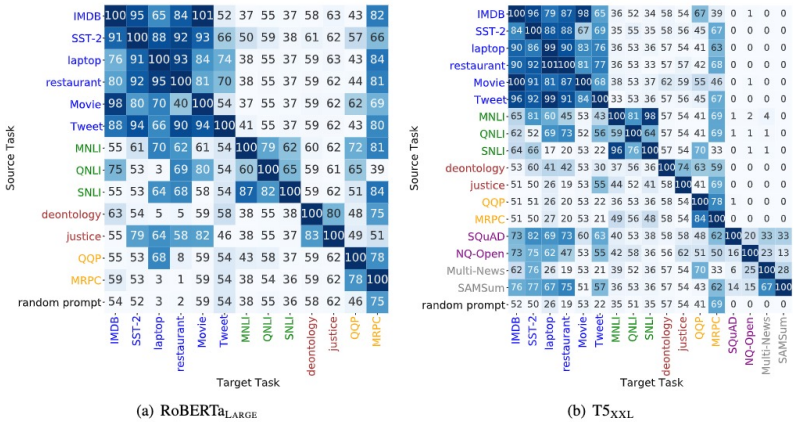
$$\mathrm{ON}(P^{t_1}, P^{t_2}) = \frac{\mathrm{AS}(P^{t_1}) \cdot \mathrm{AS}(P^{t_2})}{\|\mathrm{AS}(P^{t_1})\|\|\mathrm{AS}(P^{t_2})\|}$$

[1] Geva, et al., Transformer Feed-Forward Layers Are Key-Value Memories, EMNLP, 2021.

Transferability indicator

- Similarity
  - Embedding Similarity
    - Euclidean similarity:
    - Cosine similarity:
  - Model Stimulation Similarity (ON)

- Zero-shot Task Transferability



(a) RoBERTa$_{\text{LARGE}}$ (b) T5$_{\text{XXL}}$

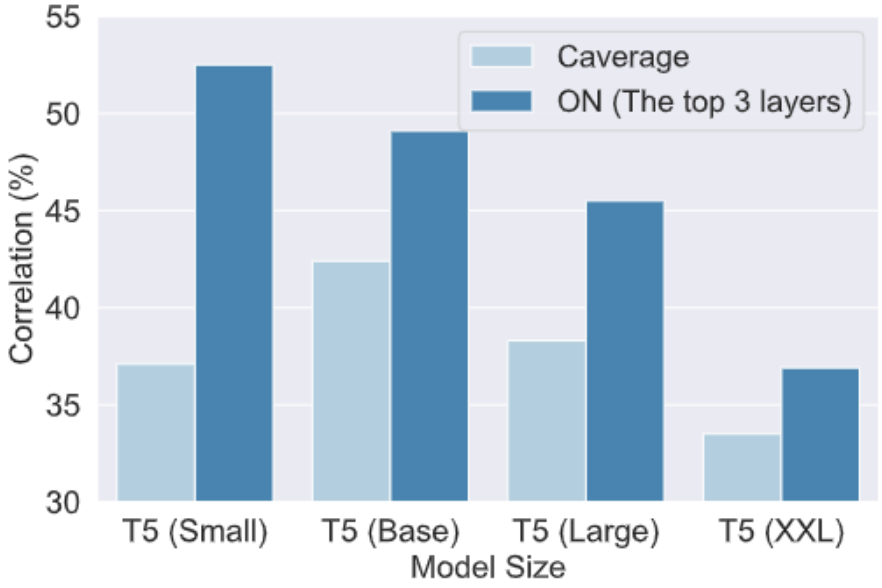
Spearman's correlation

[1] Geva, et al., Transformer Feed-Forward Layers Are Key-Value Memories, EMNLP, 2021.

Transferability indicator

- Model Stimulation Similarity (ON)
  - ON has the higher Spearman's correlation with the transferability
  - ON works worse on the larger PLMs because of the higher redundancy [1]

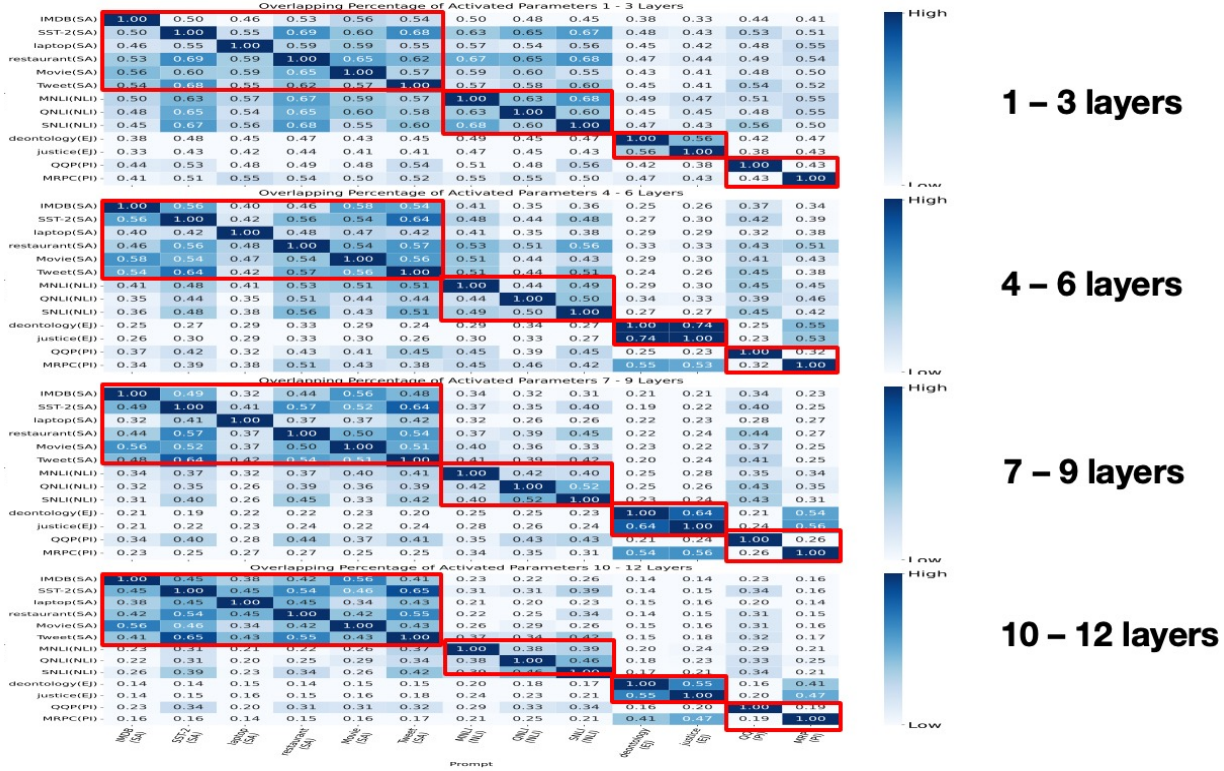
Metric	Model	
	$\text{RoBERTa}_{\text{LARGE}}$	T5_{XXL}
E_{concat}	22.6	12.9
$E_{average}$	2.8	-2.5
C_{concat}	24.8	31.6
$C_{average}$	44.7	33.5
ON	**49.7**	**36.9**



[1] Aghajanyan, et al., Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine-Tuning, ACL, 2021.

Activated neurons in a PLM

- Distribution of Activated Neuron
  - The activated neurons are common in the bottom layers but more task-specific in top layers.

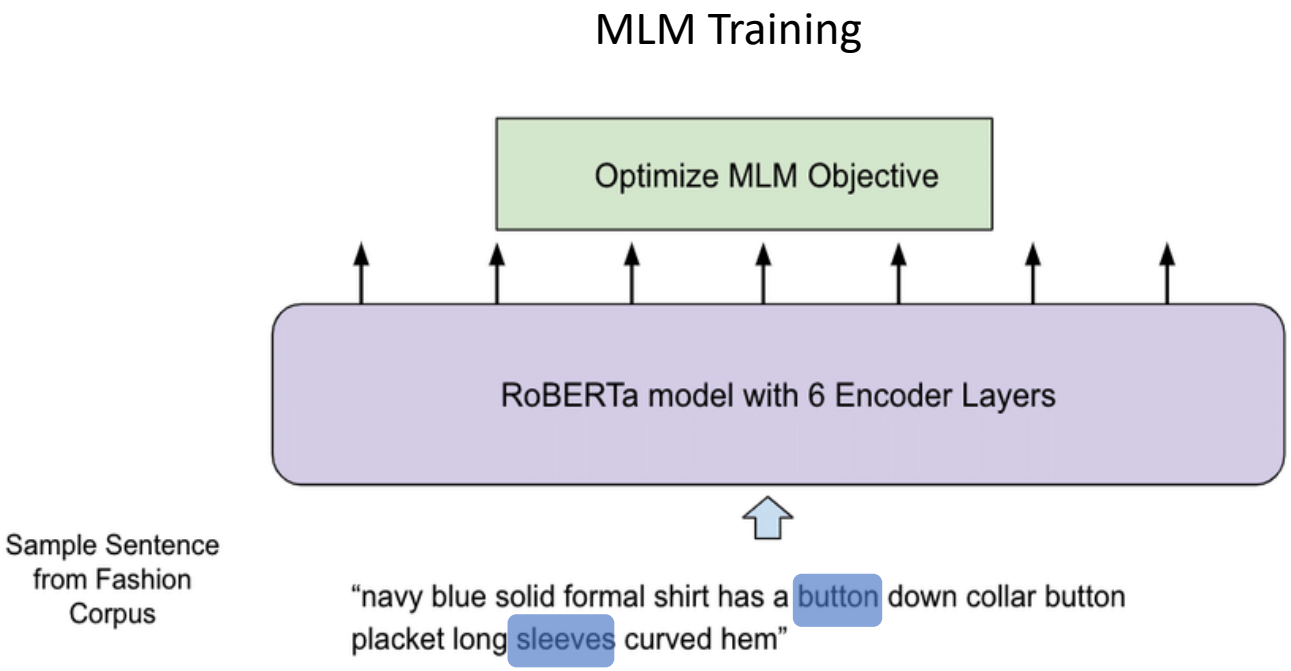


[1] Aghajanyan, et al., Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine-Tuning, ACL, 2021.

Activated Neurons Can Reflect Human-Like Emotional Attributes

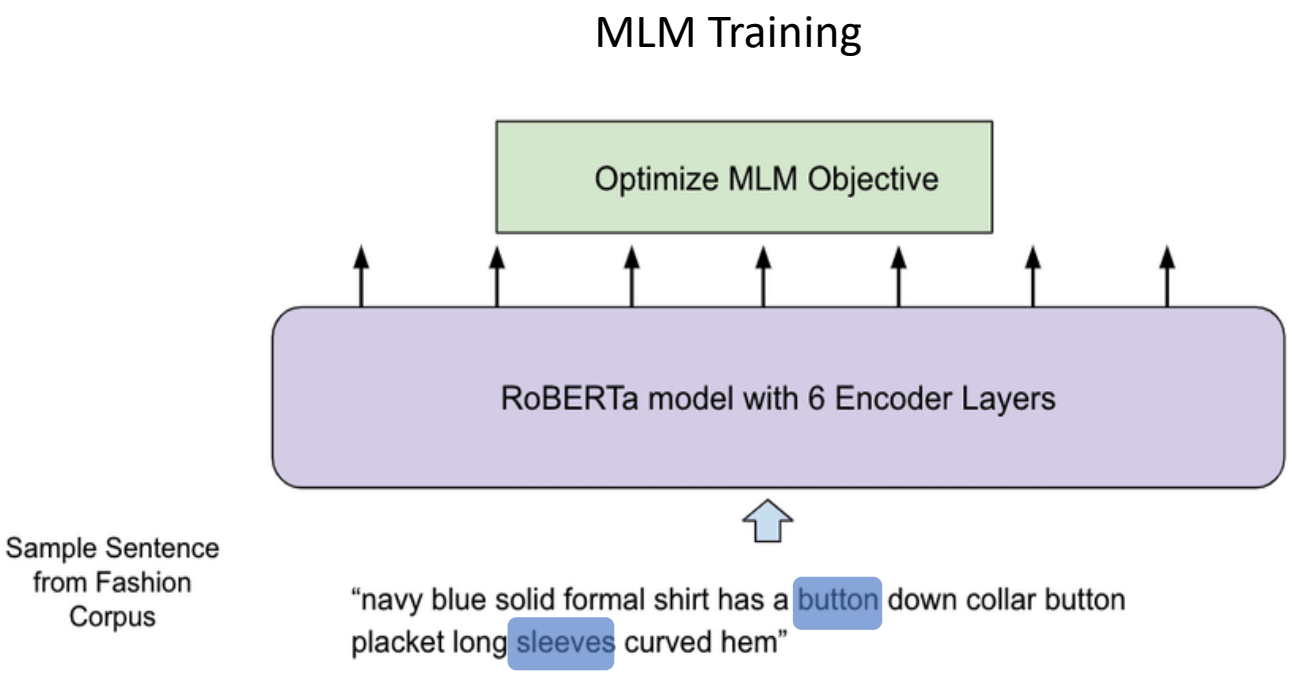
Question

- Whether PLMs can learn human-like emotional attributes during the pre-training ?



Question

- Whether PLMs can learn human-like emotional attributes during the pre-training ?



Can learn human-like emotional attributes ?

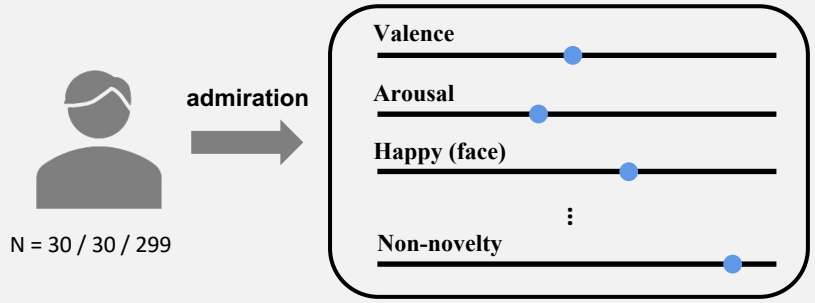
How do humans recognize different emotions ?

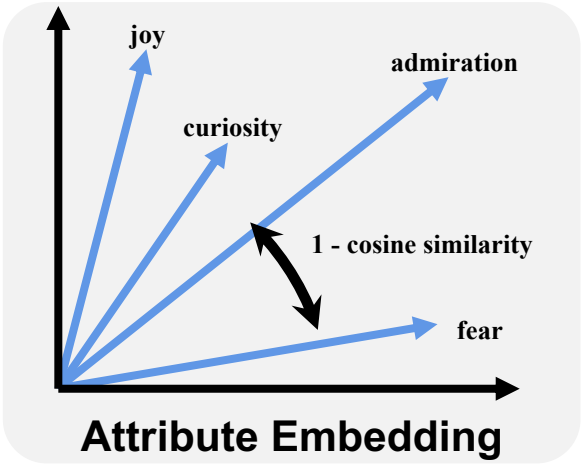
- Human

Rating of Emotional Attributes

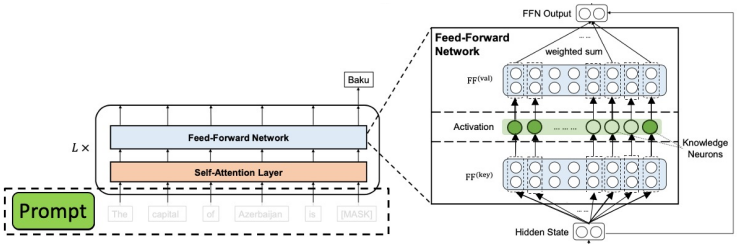
Attributes:

- [Arousal, Valence, Happy, Anger, Sad, Fear, Surprise, Disgust, Control, Fairness, Self-related, Other-related, Expectedness, Non-novelty]: 14



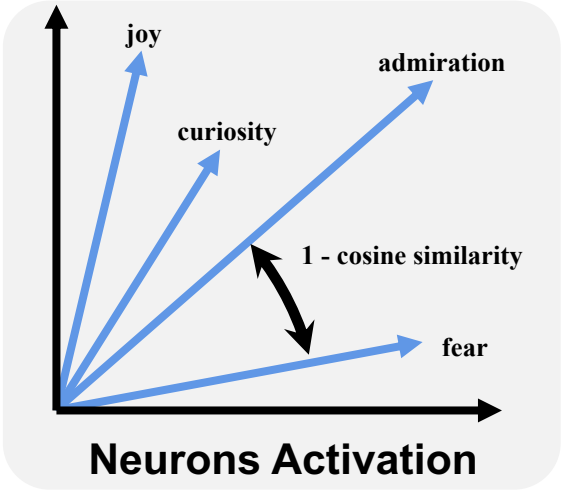


- PLM (Activated Neurons)



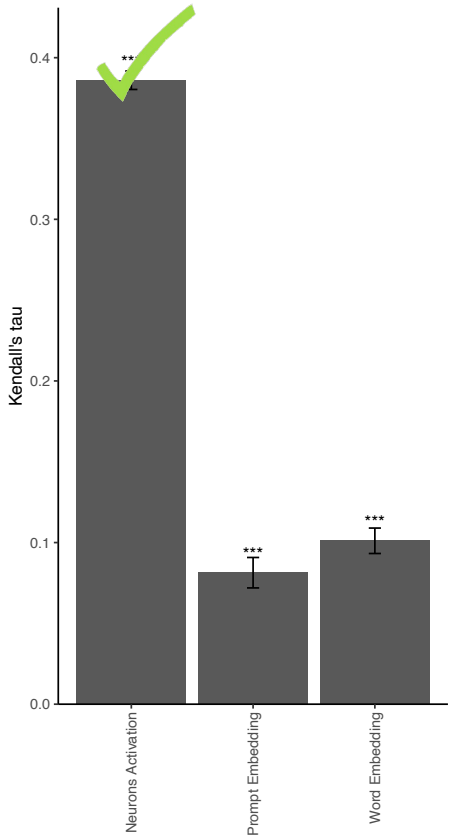
$$\mathrm{FFN}(\mathbf{x}) = \max(\mathbf{x}W_1^\top + \mathbf{b}_1, 0)W_2 + \mathbf{b}_2,$$

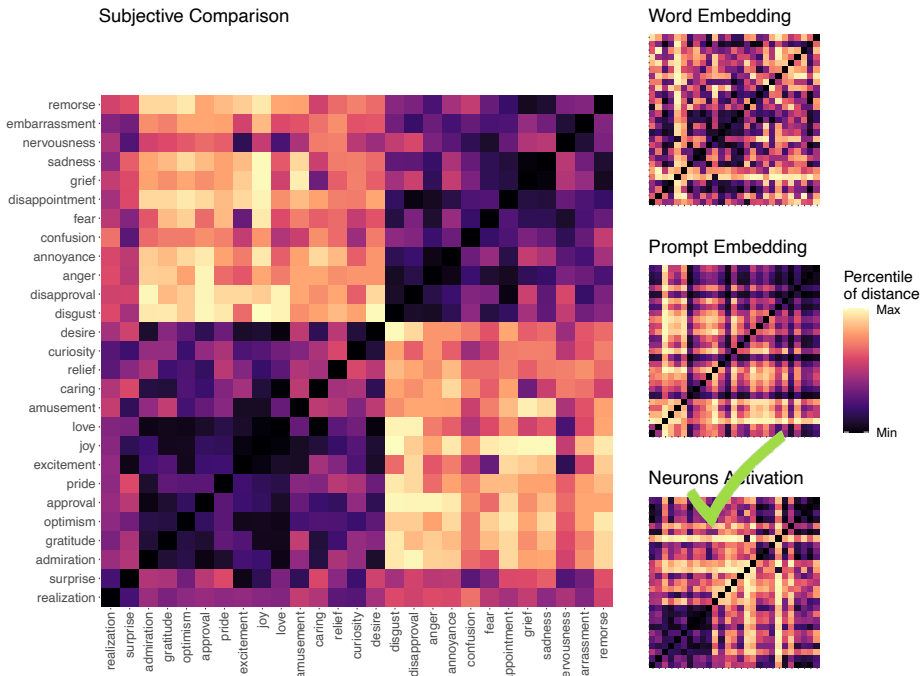
$$\mathrm{AS}(P) = [\mathbf{s}_1; \mathbf{s}_2; ...; \mathbf{s}_L]$$



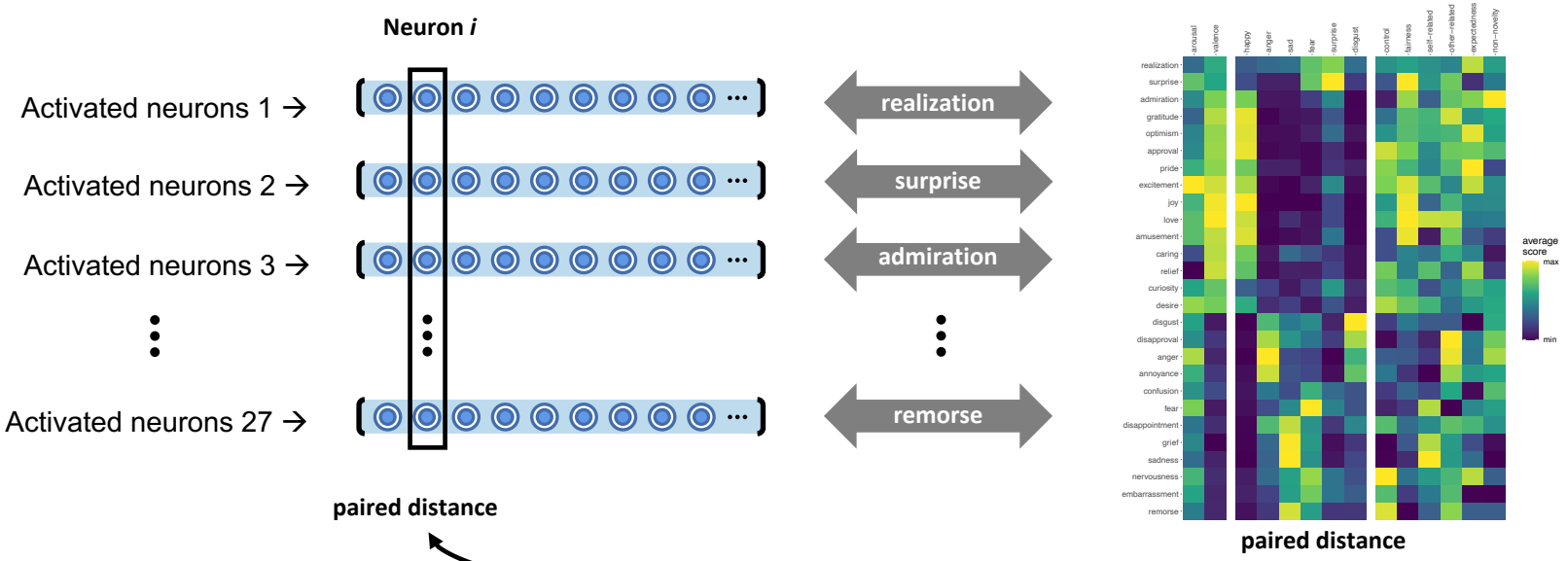
Correlation

- Correlation:
  - Represent 27 emotions with human attributes and activated neurons



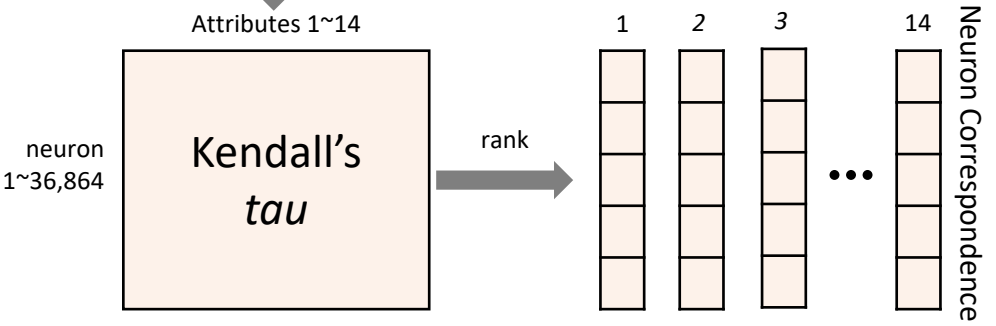


Activated neurons for every attribute



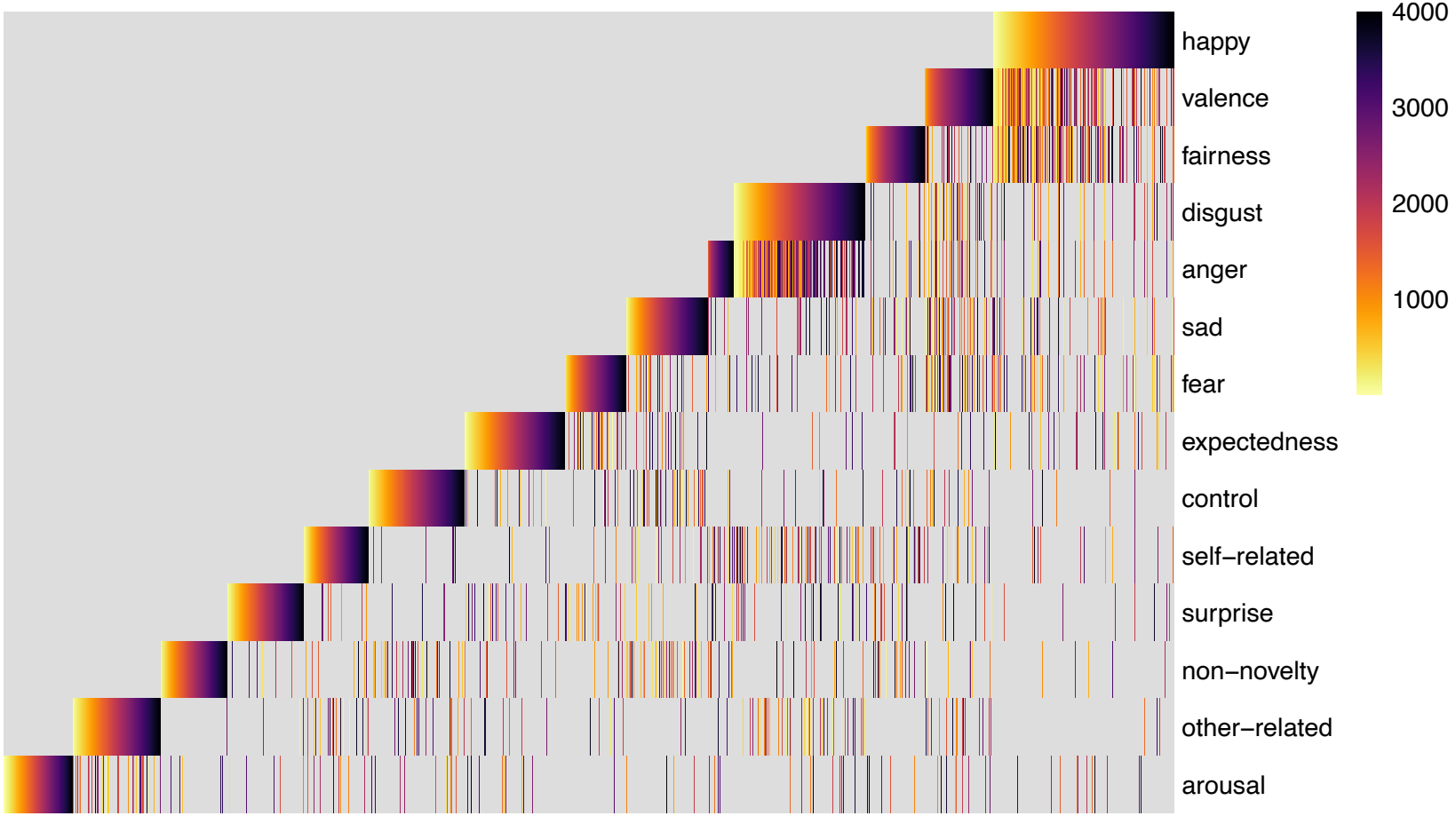
RSA

我们通过对 neuron activation 进行表征相似性分析，找到了那些与 14 emotional attributes 相关的神经元。

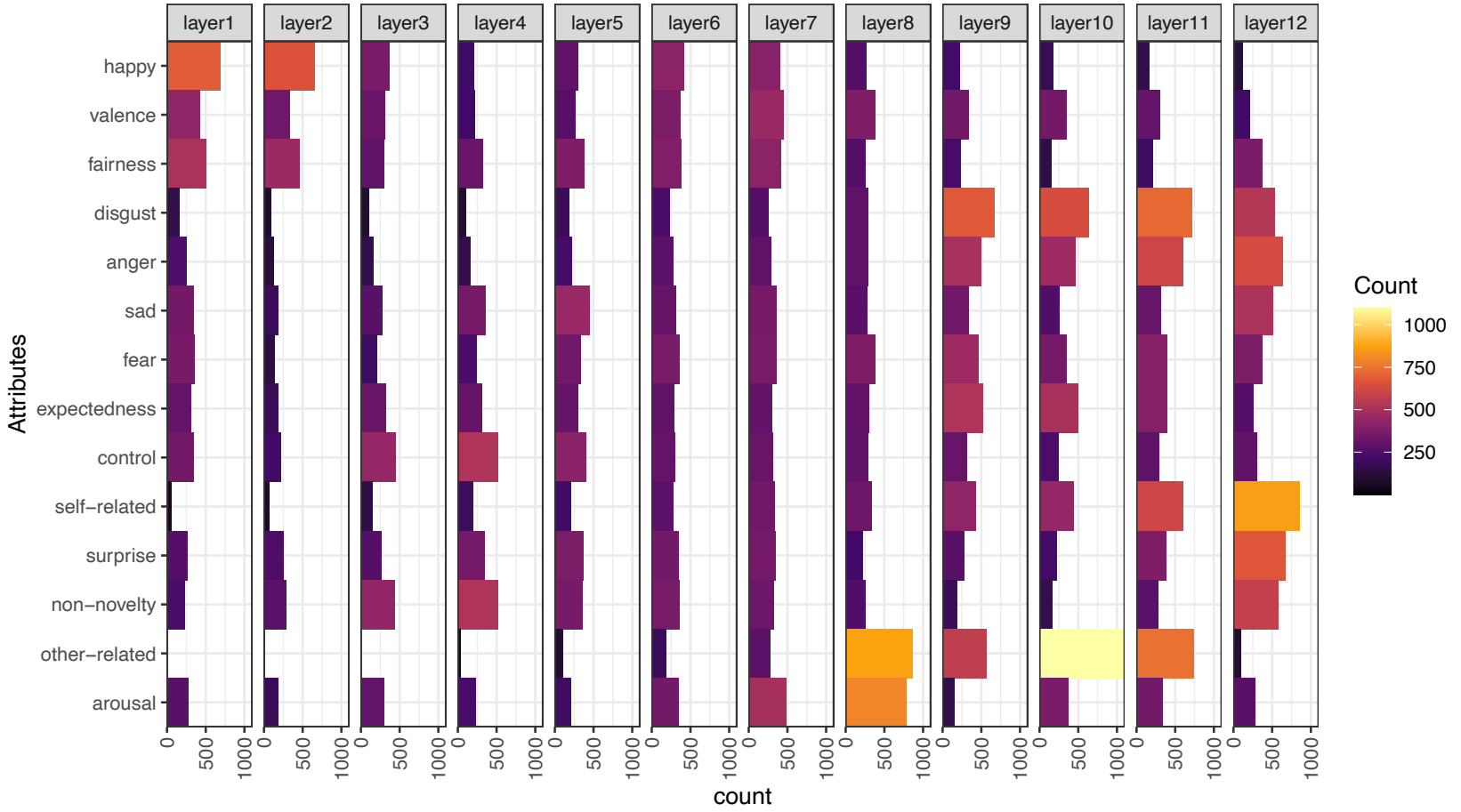


Attribute: Activated Neurons

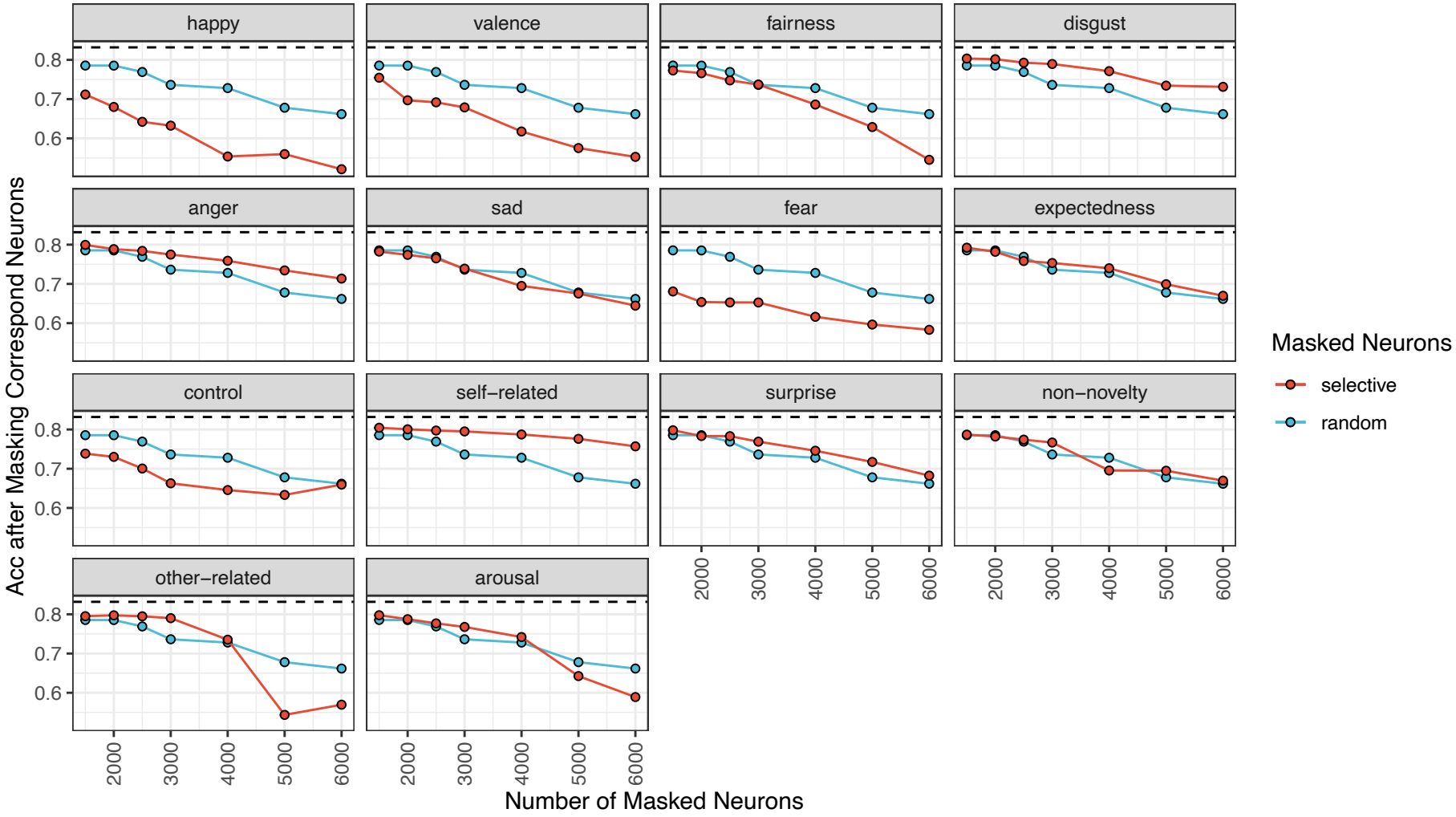
Activated neurons for every attribute



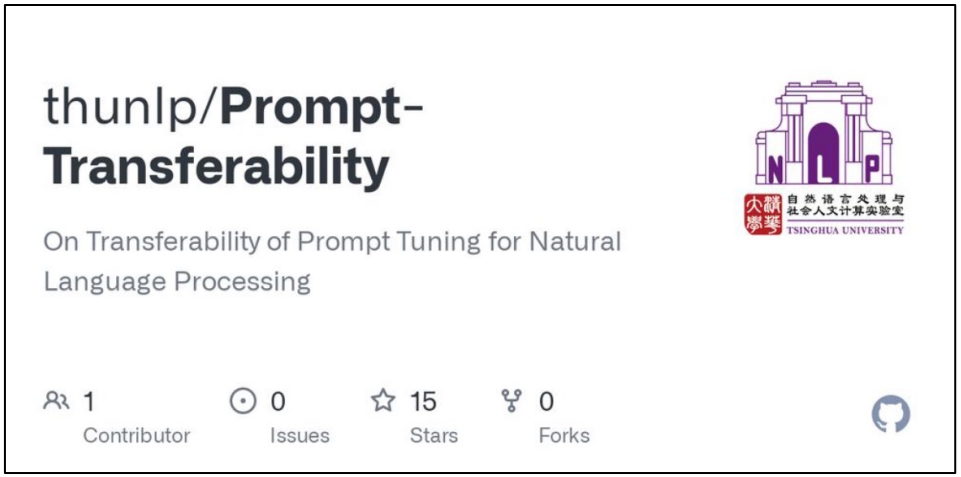
Activated neurons for every attribute

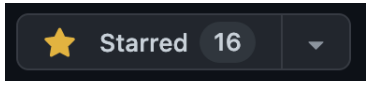


Remove neurons for an attribute



Demo



https://github.com/thunlp/Prompt-Transferability 

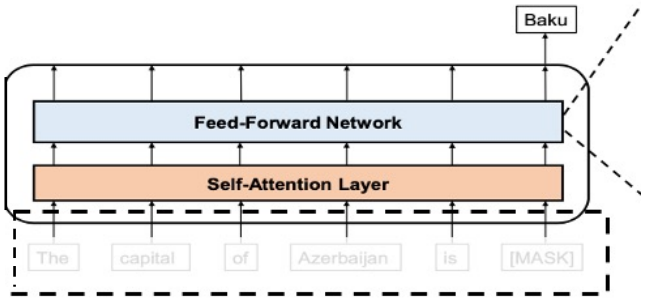
[Click]

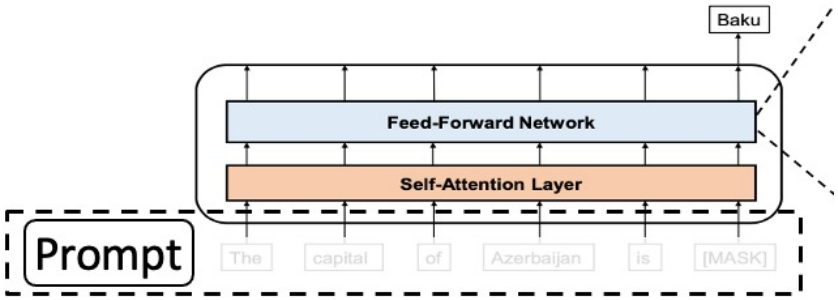
Find: Activated Neurons Demo [Colab link]

Activated Neurons

- Load Pre-trained Language Model (Roberta)
- Load the prompts (checkpoints) - 27 Emotion Tasks

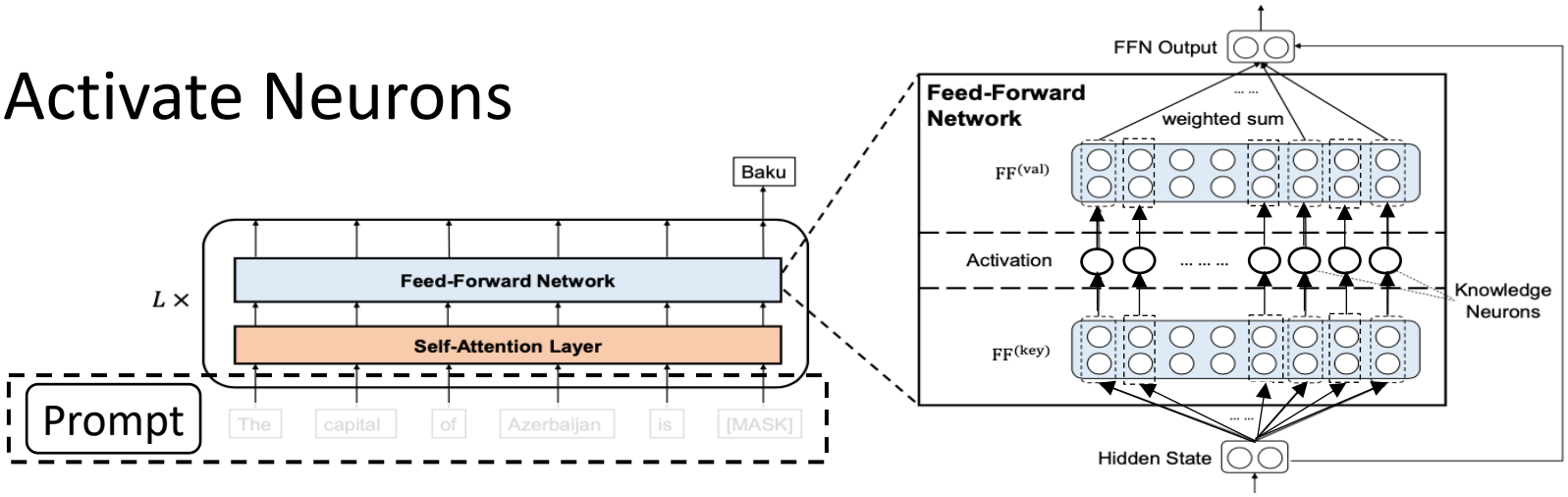
```
[4]  # 27 Tasks
     sentiments = ['amusement','excitement','joy','love','desire','optimism','caring',
             'pride','admiration','gratitude','relief', 'approval',
             'realization', 'surprise', 'curiosity', 'confusion',
             'fear', 'nervousness', 'remorse', 'embarrassment', 'disappointment',
             'sadness', 'grief', 'disgust', 'anger', 'annoyance', 'disapproval']

     # 12 random seeds (We only use 1 seed for the demonstration)
     random_seeds = [1]
     # random_seeds = [1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 42, 100]
```





Activated Neurons

- Activate Neurons




```
ckpt = torch.load(f'./checkpoint/{sentiment}-{seed}.pt')   # Load the trained prompt
prompt_emb = Parameter(ckpt['best_prompt']).to('cpu')
model.roberta.embeddings.prompt_embedding.weight.data = prompt_emb

# Forward pass to get active neuron
outputs = [[] for _ in range(nlayers)]
def save_ppt_outputs1_hook(n):
    def fn(_,__,output):
        outputs[n] = output.detach().to('cpu')    # Save the neuron of the n layer to outputs[n]
    return fn

for n in range(nlayers):
    # register_forward_hook: This hook will be used every time forward() is called
    model.roberta.encoder.layer[n].intermediate.register_forward_hook(save_ppt_outputs1_hook(n))

for sentence in loader:
    inputs = tokenizer(sentence, return_tensors='pt', add_special_tokens=False).to('cpu')
    # print(sentence, inputs)  # Output: ['<s>'] {'input_ids': tensor([[0]], device='cpu')}
    _ = model(**inputs)     # After inputting the special token into the model, each layer of FFL will call register_forward_hook to stor

outputs = torch.stack(outputs)    # output.shape=[12,1,102,3072]:layer=12, batch, promptLen+specialToken=102, neuronNum=3072
outputs = outputs[:, :1, :1, :]  # 12 layers, 1, The dimension corresponding to the 1 token, neuron num = 3072
outputs = outputs.flatten()
```

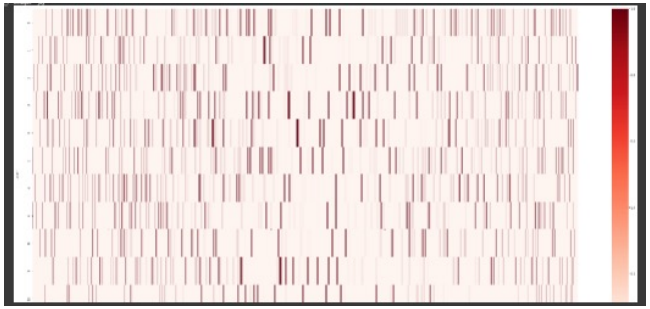
Activated Neurons

- Activated neurons in each layers
  - Input: ['realization’, 'surprise’, …, 'remorse']

2.1 Input a prompt to show activated neuron

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

prompt = input("Please enter a prompt: ")  # joy
print('Prompt: ', prompt)
df = pd.read_csv(f'./active_neuron_after_relu_csv/{prompt}-1.csv')

activatedNeuron = df.values
activatedNeuron = activatedNeuron.reshape(12,3072)
print(f'Activated Neurons{activatedNeuron.shape}: \n', activatedNeuron)
```

```
Please enter a prompt: joy
```



Activated Neurons

- Cosine Similarity of Activated Neurons
  - Input: ['realization’, 'surprise’, …, 'remorse']

2.2 Input two prompts to show cosine similarity of activated neurons

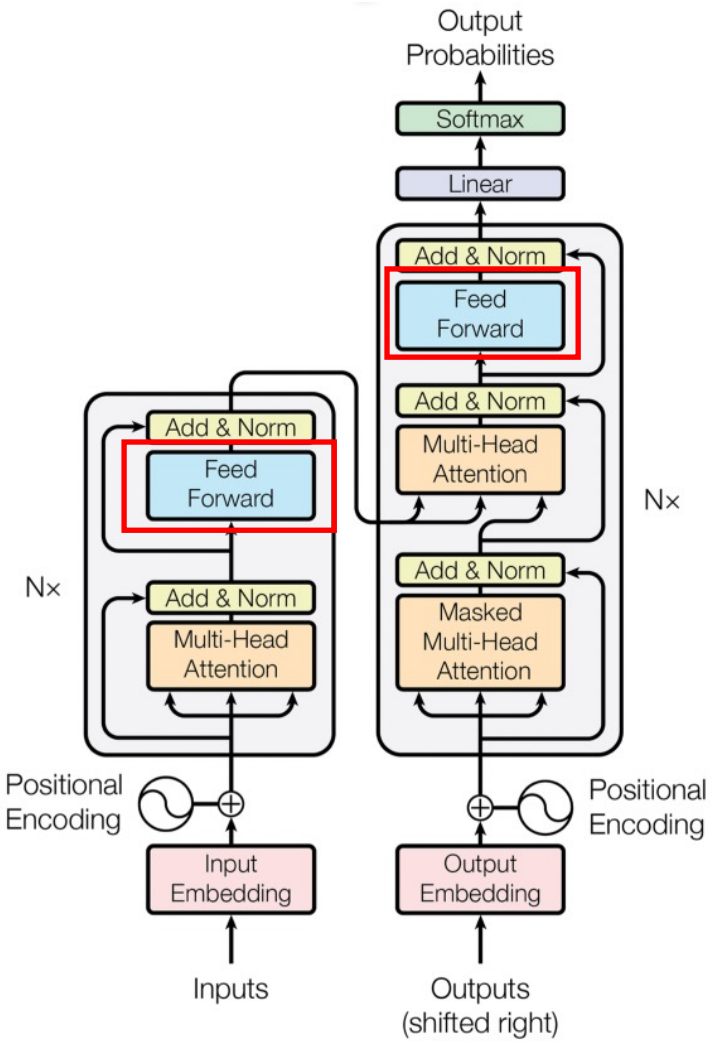
```
[ ]  from sklearn.metrics.pairwise import cosine_similarity

     prompt1 = input("Please enter prompt1: ")  # joy
     df1 = pd.read_csv(f'./active_neuron_after_relu_csv/{prompt1}-1.csv')
     activatedNeuron1 = df1.values

     prompt2 = input("Please enter prompt2: ")  # love, admiration
     df2 = pd.read_csv(f'./active_neuron_after_relu_csv/{prompt2}-1.csv')
     activatedNeuron2 = df2.values

     print('cosine similarity: ', cosine_similarity(activatedNeuron1, activatedNeuron2))
```

```
Please enter prompt1: joy
Please enter prompt2: love
cosine similarity:  [[0.49823117]]
```

Neurons in PLMs

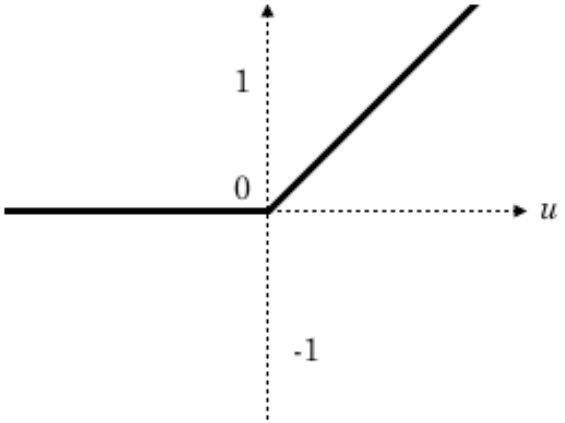
THUNLP

Background: Neurons in FFNs

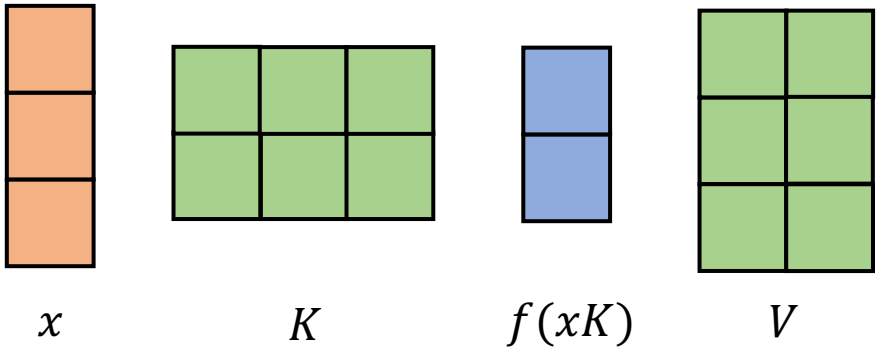


Transformer Architecture

$$f(u) = \max(0, u)$$



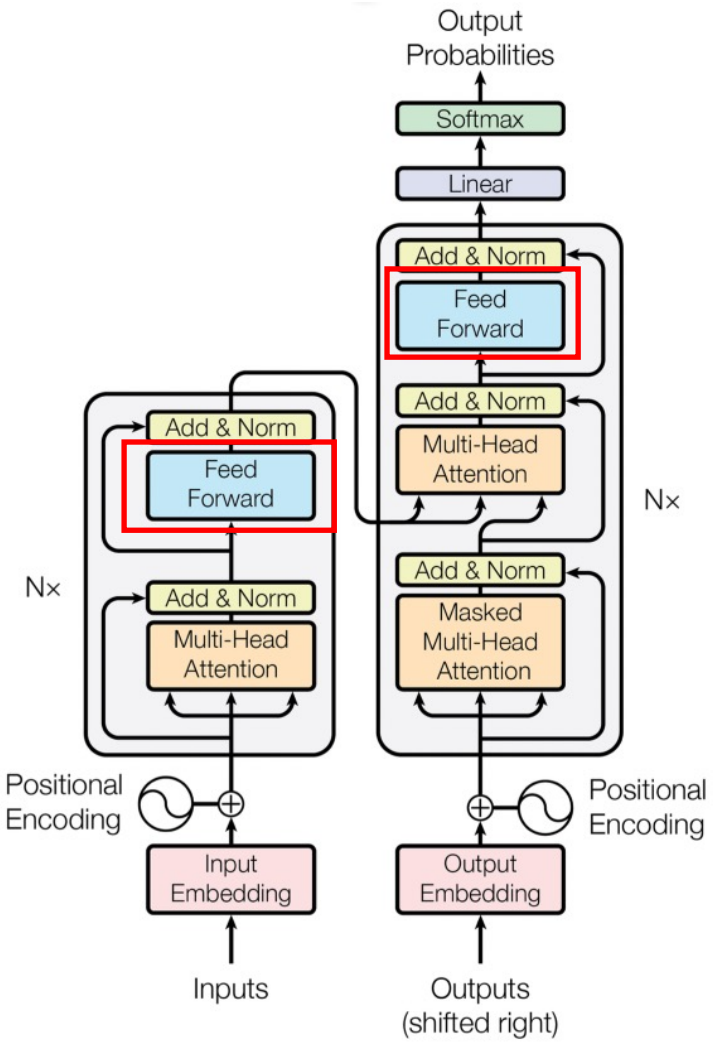
Activation Function



$$\mathrm{FFN}(\mathbf{x}) = f(\mathbf{x}\mathbf{K}^\top + \mathbf{b}_1)\mathbf{V} + \mathbf{b}_2,$$

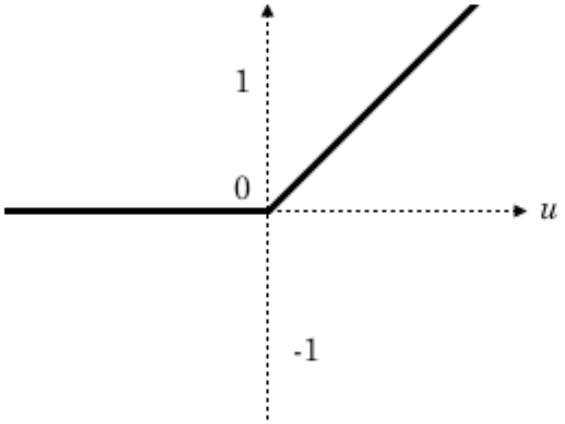
Feed Forward Neural Network

Background: Neurons in FFNs



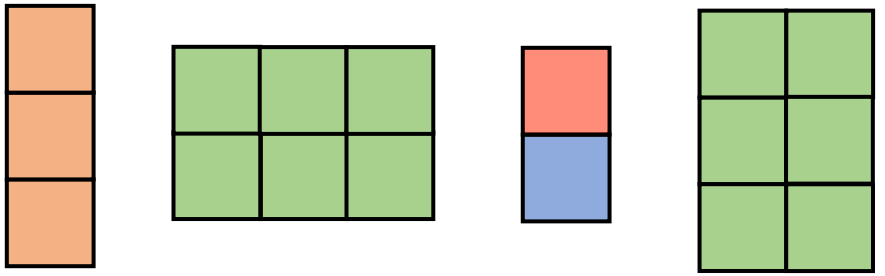
Transformer Architecture

$$f(u) = \max(0, u)$$



Activation Function

Activation of a Neuron

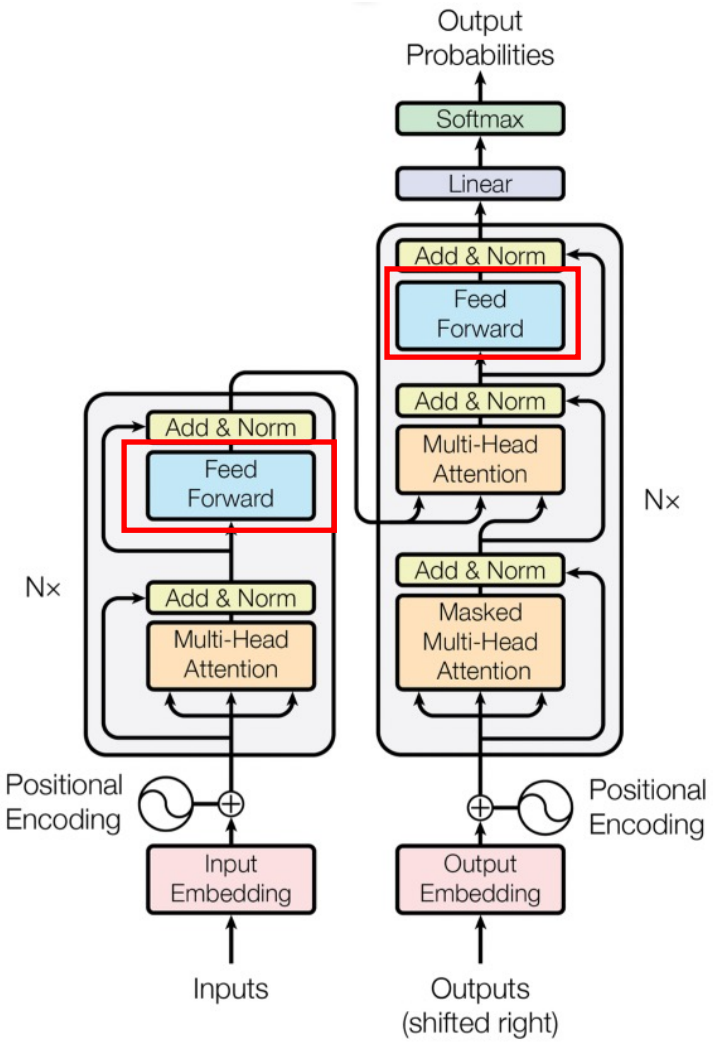


x K $f(xK)$ V

$$\mathrm{FFN}(\mathbf{x}) = f(\mathbf{x}\mathbf{K}^\top + \mathbf{b}_1)\mathbf{V} + \mathbf{b}_2,$$

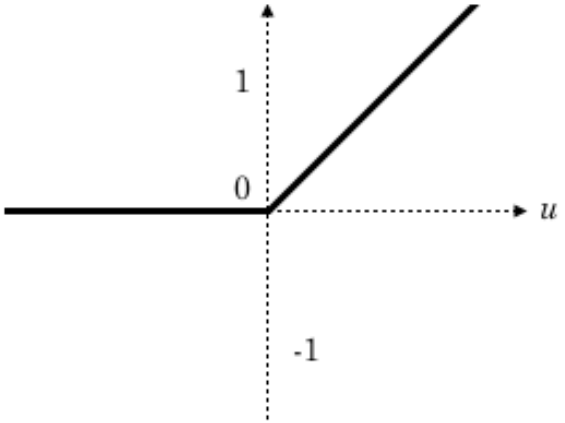
Feed Forward Neural Network

Background: Neurons in FFNs



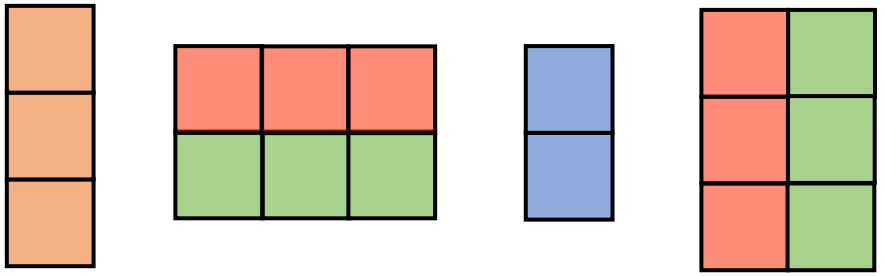
Transformer Architecture

$$f(u) = \max(0, u)$$



Activation Function

Weights of the Neuron



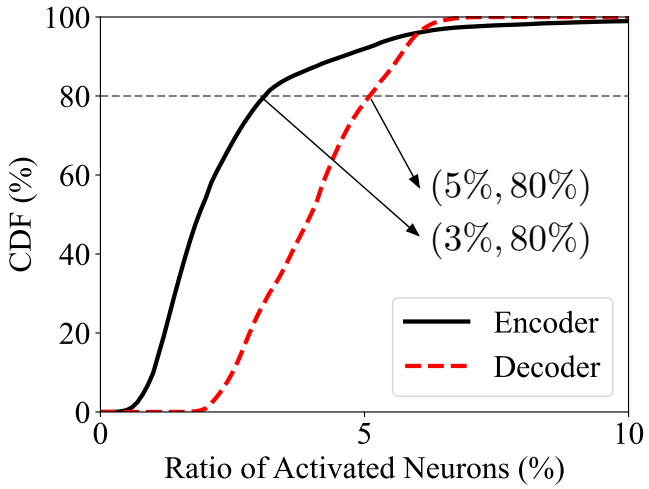
x K $f(xK)$ V

$$\mathrm{FFN}(\mathbf{x}) = f(\mathbf{x}\mathbf{K}^\top + \mathbf{b}_1)\mathbf{V} + \mathbf{b}_2,$$

Feed Forward Neural Network

Sparse activation phenomenon

- Sparse Activation Phenomenon in Large PLMs
- 80% inputs only activate less than 5% neurons of FFNs
- No useless neuron that keeps inactive for all inputs
- Related to Conditional Computation
  - Constrains a model to selectively activate parts of the neural network according to input



Cumulative distribution function (CDF) of the ratio of activated neurons in FFNs. Use T5-large (700 million parameters).

Conditional computation

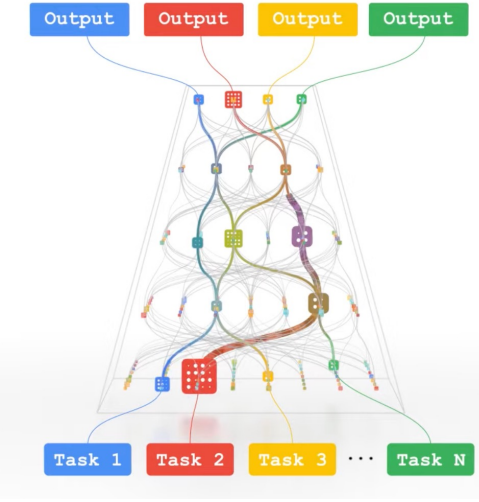
- Deep Learning of Representations: Looking Forward (Bengio, 2013)

3 Scaling Computations

From a computation point of view, how do we scale the recent successes of deep learning to much larger models and huge datasets, such that the models are actually richer and capture a very large amount of information?

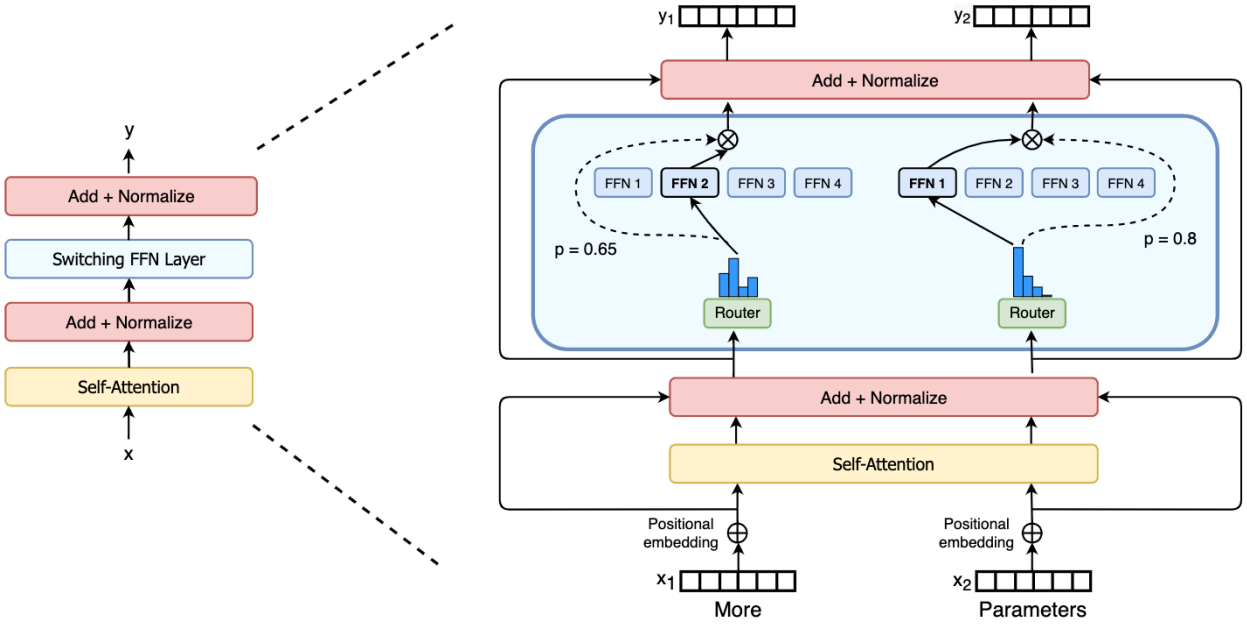
Conditional Computation. A central idea (that applies whether one parallelizes or not) that we put forward is that of *conditional computation*: instead of dropping out paths independently and at random, drop them in a learned and optimized way. Decision trees remain some of the most appeal-

- Pathways (Jeff Dean, 2021)
  - Today's models are dense and inefficient
  - Pathways will make them sparse and efficient



MoEfication

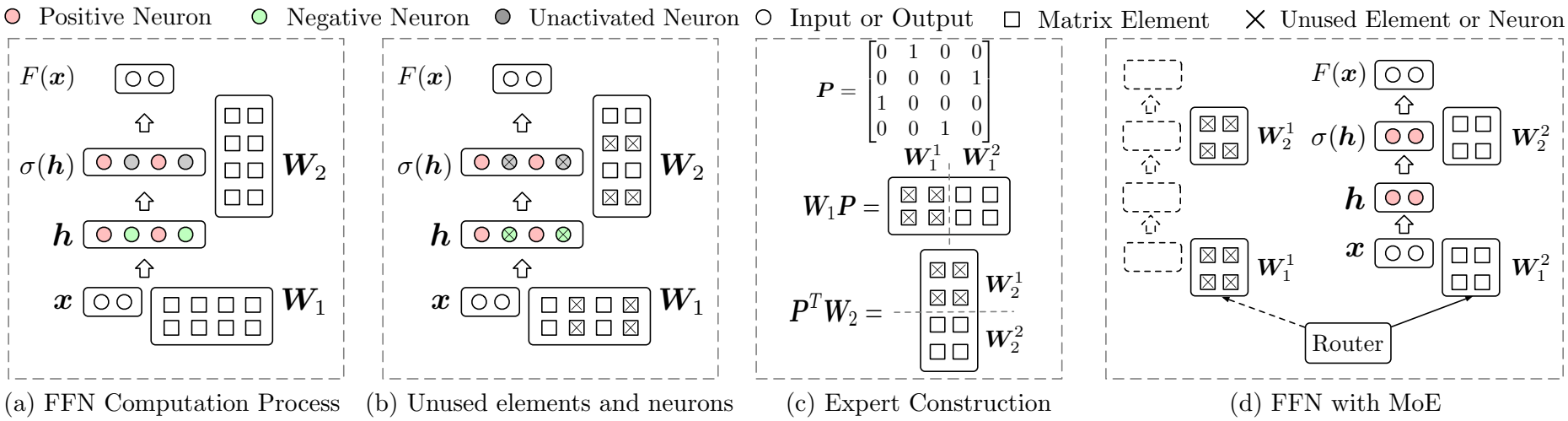
- Mixture-of-experts (MoE)
- Use MoE to increase model parameters with tiny extra computational cost



Fedus et al. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. Arxiv.

MoEfication

- Split existing models into multiple experts while keeping model size unchanged



(a) FFN Computation Process (b) Unused elements and neurons (c) Expert Construction (d) FFN with MoE

Zhang et al. MoEfication: Transformer Feed-forward Layers are Mixtures of Experts. Findings of ACL. 2022

MoEfication

- Expert Construction
- Group the neurons that are often activated simultaneously
- Parameter Clustering Split
  - Treat the columns of W_1 as a collection of vectors
  - K-means
- Co-Activation Graph Split
  - Construct a coactivation graph
  - Each neuron is represented by a node
  - Edge weight between two nodes is their co-activation value

$$\text{co-activation}(n, m) = \sum_{\boldsymbol{x}} \boldsymbol{h}_n^{(\boldsymbol{x})} \boldsymbol{h}_m^{(\boldsymbol{x})} \mathbb{1}_{\boldsymbol{h}_n^{(\boldsymbol{x})} > 0, \boldsymbol{h}_m^{(\boldsymbol{x})} > 0}$$

MoEfication

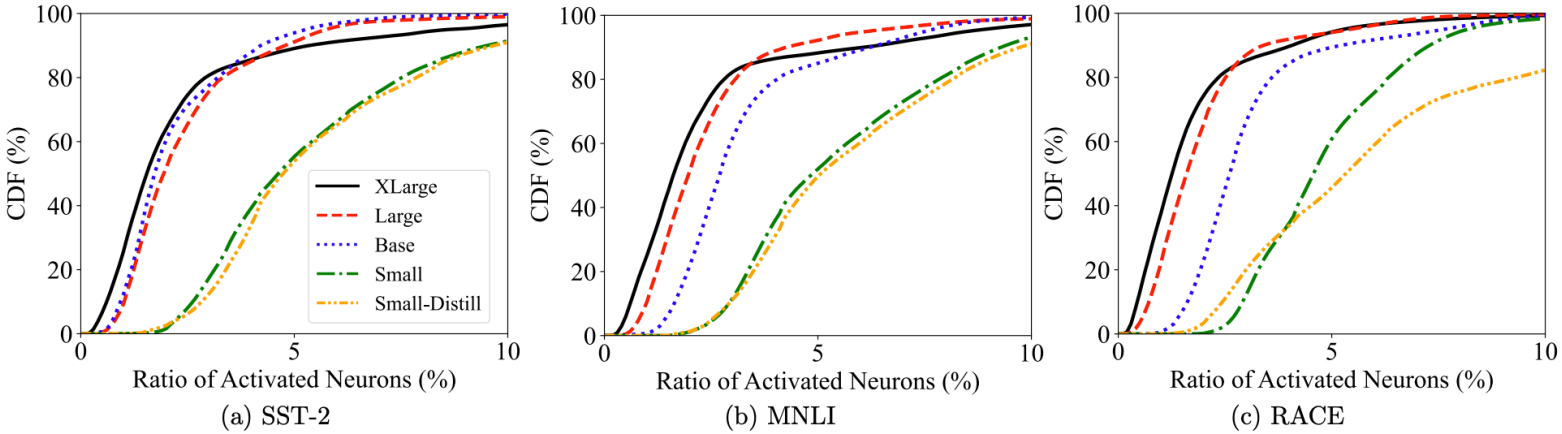
- Assign a score to each expert and select the experts with high scores

$$S = \underset{A \subset \{1,2,\ldots,k\}, |A|=n}{\arg\max} \sum_{i \in A} s_i$$

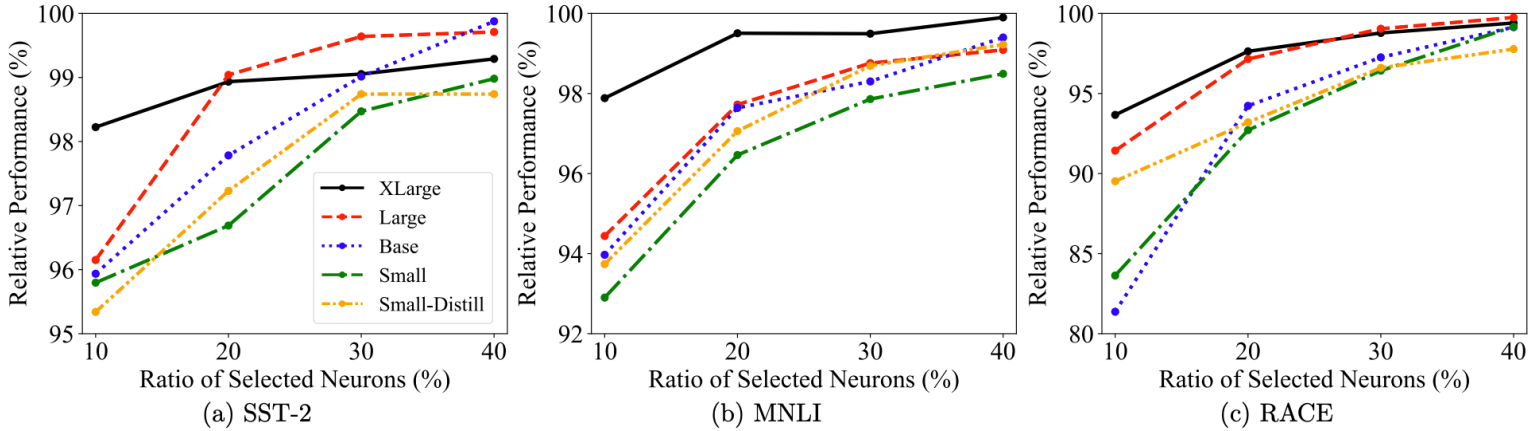
- Groundtruth Selection
  - Calculate the number of positive neurons in each expert as s_i
- Parameter Center
  - Average all columns of W_1 and use it as the center
- Learnable Router
  - Learn a router from the groundtruth on the training set

MoEfication

- Sparsity of Different T5 Models



- MoEfication with Different T5 Models



Observations on routing patterns

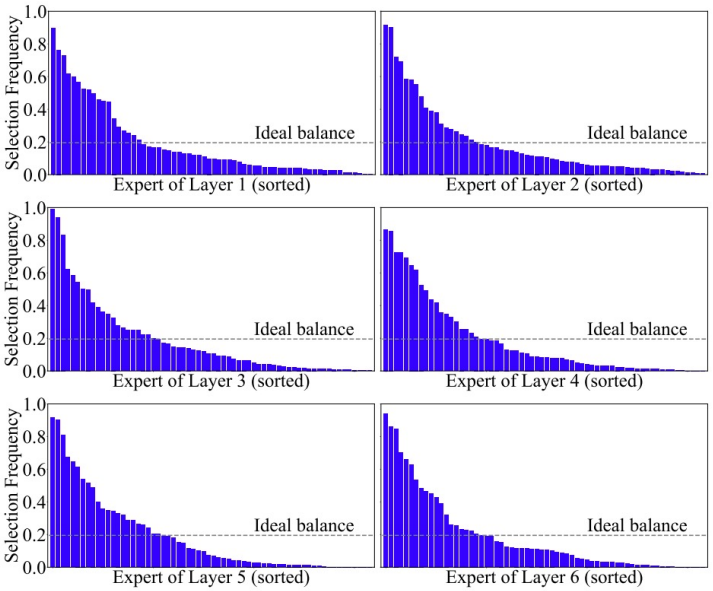
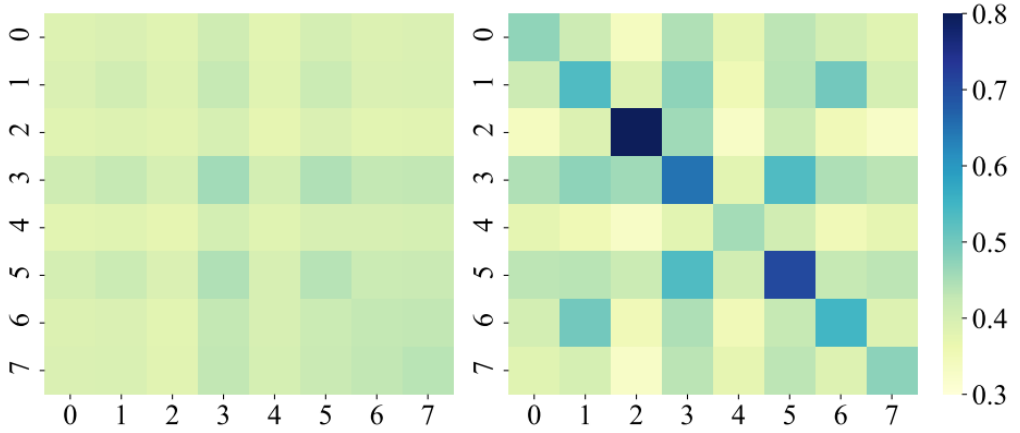


Figure 5: Selection Frequency of 64 experts in each encoder layer of MoEfied T5-Small. The frequency of ideal balance selection is 0.2 while the distribution is much unbalanced.



(a) The 8 most selected experts (b) The 8 least selected experts

Figure 6: Input similarities between experts in the last encoder layer of MoEfied T5-Small. For the most selected experts, both the self-similarities and inter-similarities are low. For the least selected experts, the self-similarities are much higher than inter-similarities.

Analyze PLMs through neurons

- Specific Function
- Transferability Indicator
- Activated Neurons Can Reflect Human-Like Emotional Attributes

Specific Function

Expert units

- Identify whether the activation of a specific neuron can classify a concept
- Nc+ positive sentences that contain concept c and Nc− negative sentences that do not contain concept c.

```
<instance docsrc="Indigenous architecture" id="shelter.00002">
    <answer instance="shelter.00002" senseid="shelter%1:06:00::" />
    <context>
        Types There are three traditional types of igloos ,
        all of different sizes and used for different purposes.
        The smallest were constructed as temporary
        <head>shelters</head>
        , usually only used for one or two nights .
     </context>
</instance>
```

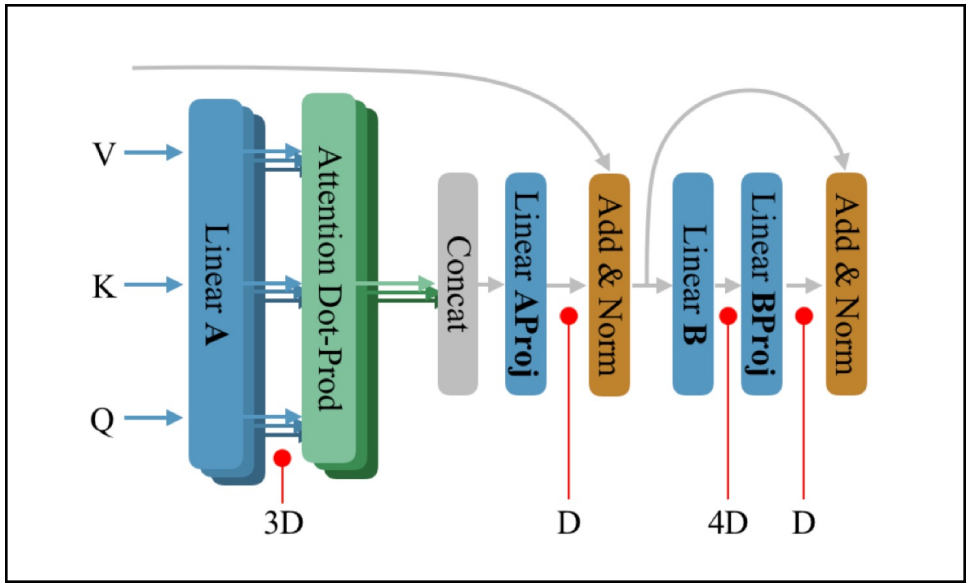
OneSec dataset

- **Sense**: Positive sentences contain a keyword with a specific WordNet sense, whereas negative sentences do not contain the keyword.
- **Homograph**: Positive sentences contain a keyword with a specific WordNet sense, whereas negative sentences contain the same keyword with a different meaning. Intuitively, *homograph* concepts are harder to disambiguate than *sense* concepts.

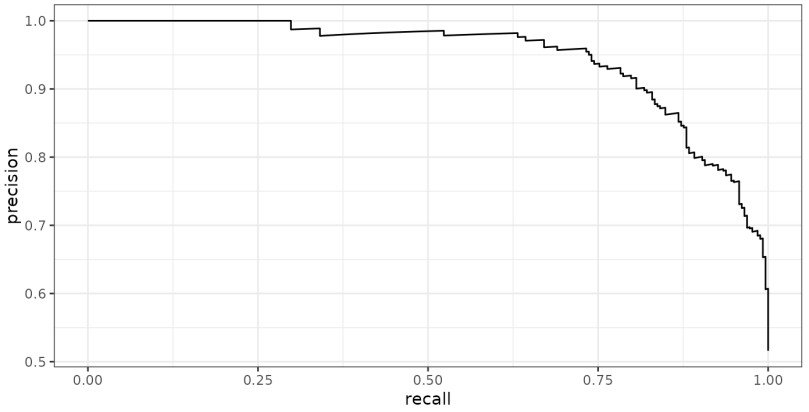
Suau et al. Finding Experts in Transformer Models. Arxiv. 2019.

Concept expertise

- Give each unit an index *m* and treat a unit as a binary classifier for the input sentences to compute AP



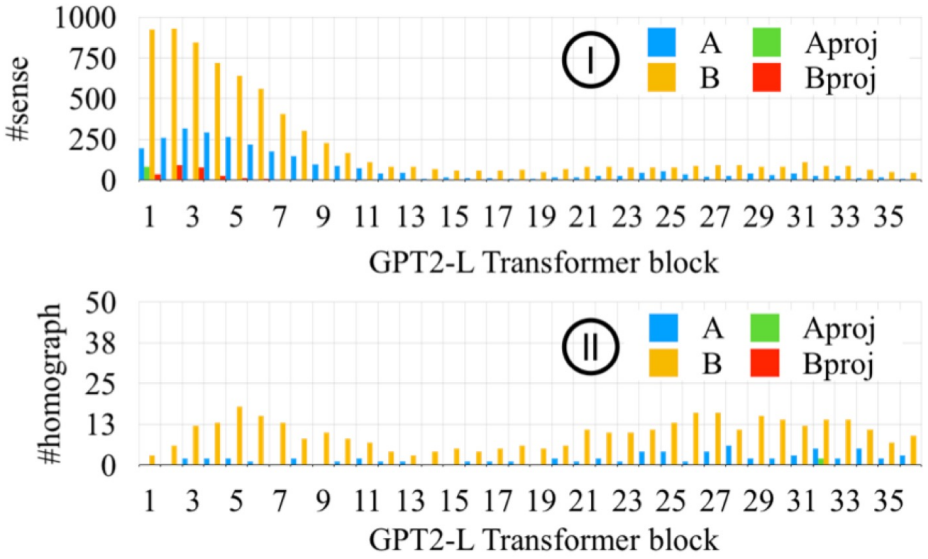
$$\mathrm{AP} = \sum_n (R_n - R_{n-1}) P_n$$

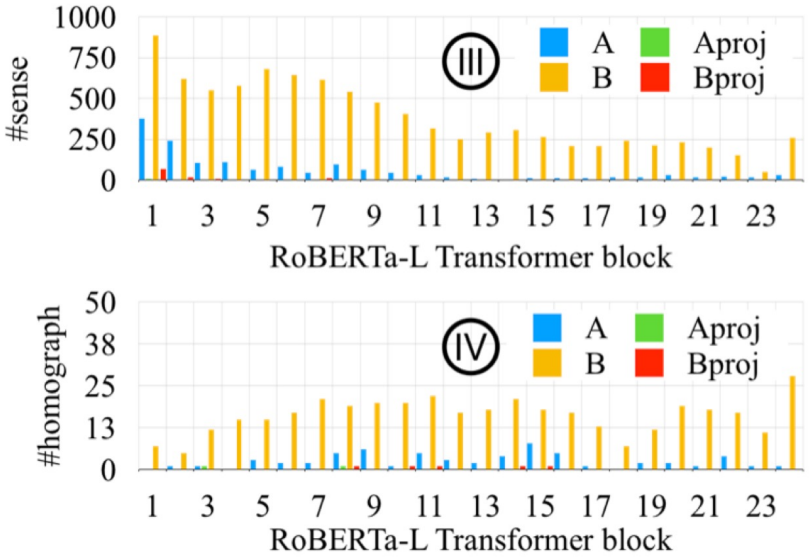


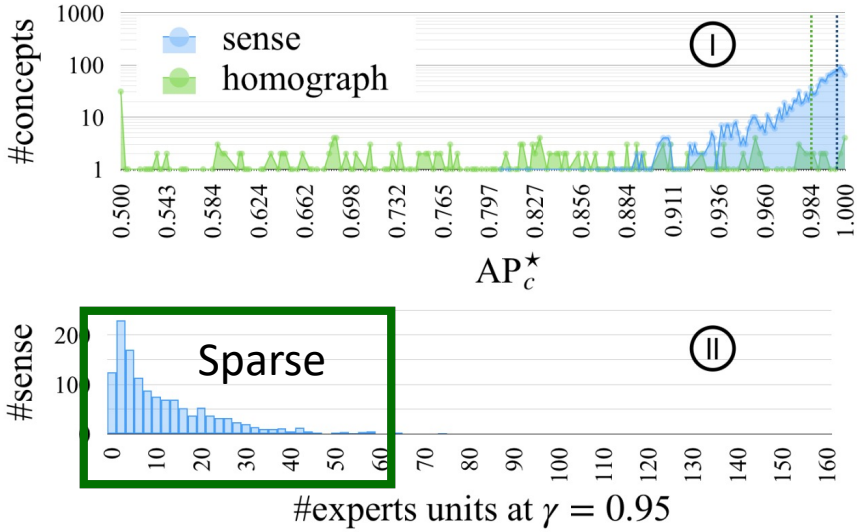
$$\mathrm{AP}_c^\star = \max_m \{\mathrm{AP}_c^m\}$$

Concept expertise $$\mathcal{X}_\gamma = \frac{|\{c \text{ s.t. } \mathrm{AP}_c^\star \geq \gamma\}|}{|C|} \quad \forall c \in C.$$

Concept distribution







Expertise and generalization results

- Detect the model's ability without fine-tuning

Models used	Task	(S)	(T)	$\mathcal{X}_{\gamma^\star}$
BERT-B/L Distilbert RoBERTa-L XLM	GLUE Score	0.361	0.871	**0.892**
	CoLA	0.572	0.821	**0.874**
	SST-2	0.554	0.849	**0.864**
	MRPC (acc)	0.163	0.714	**0.753**
	MRPC (F1)	0.162	**0.916**	0.855
	STS-B (p)	0.000	**0.490**	0.401
	STS-B (s)	0.115	**0.903**	0.840
	QQP (acc)	0.340	0.936	**0.944**
	QQP (F1)	0.397	0.834	**0.859**
	MNLI-m	0.603	**0.833**	0.771
	MNLI-mm	0.452	0.906	**0.944**
	QNLI	0.286	0.857	**0.923**
	RTE	0.332	0.861	**0.873**
	WNLI	0.314	**0.751**	0.619
	AX	0.426	0.914	**0.956**
BERT-B/L DistilBERT RoBERTa-L	SQuAD 1.1 (F1)	**0.899**	0.840	0.850
	SQuAD 2.0 (F1)	**0.961**	0.752	0.937
	Average	0.408	0.826	**0.833**

Concept overlap

- Let the overlap between concepts q and v be

$$\Omega(q, v) = \frac{\|\boldsymbol{s}_q \cap \boldsymbol{s}_v\|_1}{\|\boldsymbol{s}_q \cup \boldsymbol{s}_v\|_1} \in [0, 1]$$

Query definition	Concept	$\Omega(q, v)$
A seat for one person, with a support for the back.	chair%1:06:00 (query)	1.000
	table%1:06:01	0.458
	bed%1:06:00	0.361
	cup%1:06:00	0.341
	table%1:06:01 VS. table%1:14:00	0.336
	floor%1:06:00	0.328
The position of professor.	chair%1:04:00 (query)	1.000
	chair%1:04:00 VS. chair%1:06:00	0.575
	fellow%1:18:02	0.371
	director%1:18:03	0.297
	administration%1:04:00	0.243
	member%1:18:00	0.241

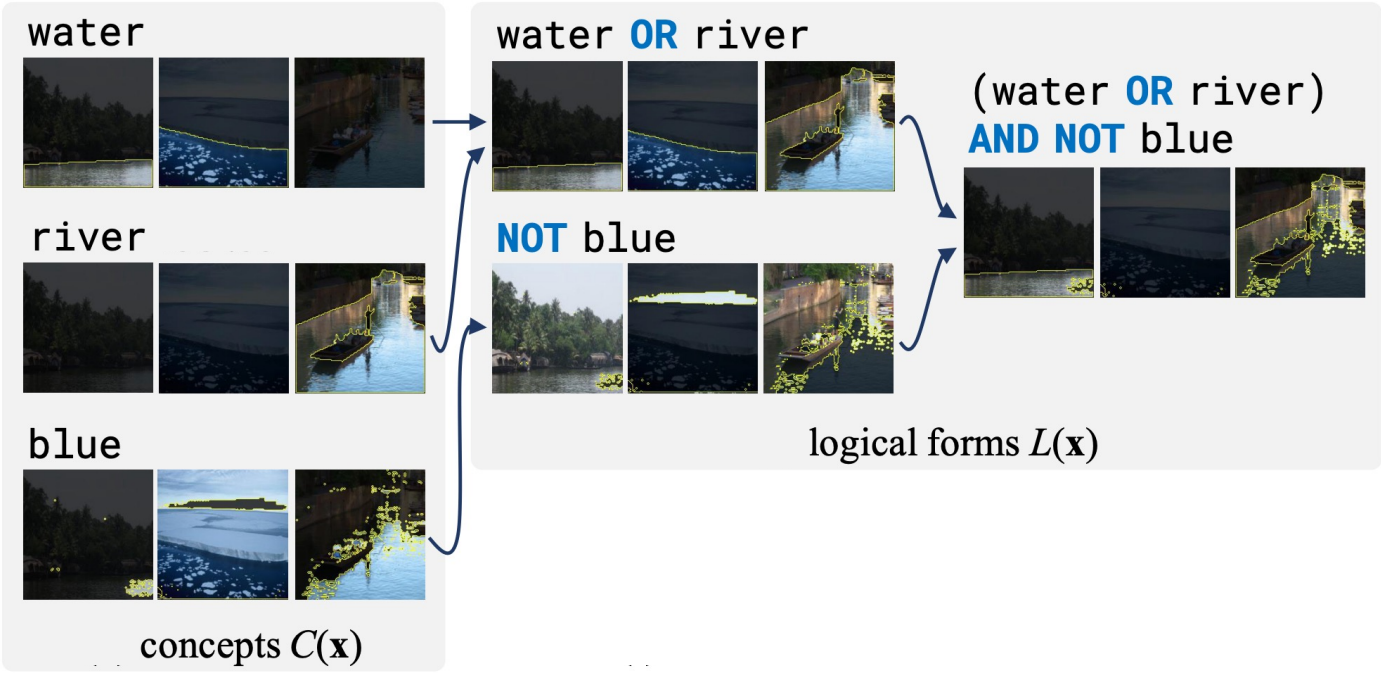
Conditioned text generation

- Selected expert units to compute

K forced	Once upon a time + Generated induced for concept bird%1:05:00 (warm-blooded egg-laying vertebrates)
0 (0%)	, I had a friend who used to teach high school English and he was like, "Oh, all you have to do is just get out there
40 (0.009%)	, many of these treasures were worth hundreds of thousands of dollars.\n But this isn't the first time that a horse has been
60 (0.015%)	, through a freak occurrence, an invasion of house sparrows which so often reduces the black-browed this nation recreats through
80 (0.019%)	, our own ancestors rode about on chicken-like air wings.\n But this wonder of the air has no such wings.\n Taking down
200 (0.048%)	of year, birds chase each and watching. flot racing form bird, bird bird bird bird bird bird bird bird bird bird bird, Bird bird
	Once upon a time + Generated induced for concept lead%1:07:02 (an advantage held by a competitor in a race)
50 (0.012%)	the left-hander would always start at the front in the first two instances, but when Mauricio Gaponi rose to the podium,
	Once upon a time + Generated induced for concept lead%1:27:00 (a soft heavy toxic malleable metallic element)
100 (0.024%)	a crust layer was applied to a partially fortified nickel base, thereby causing to zinc- and copper- ground element cob. The occurrence of those metal and chrome

Compositional explanations of neurons

- Neurons learn compositional concepts



- Compositional explanations allow users to predictably manipulate model behavior

Mu et al. Compositional Explanations of Neurons. NeurIPS. 2020.

Find neurons

- For an individual neuron, thresholding its activation

(b) neuron $f_{483}(\mathbf{x})$

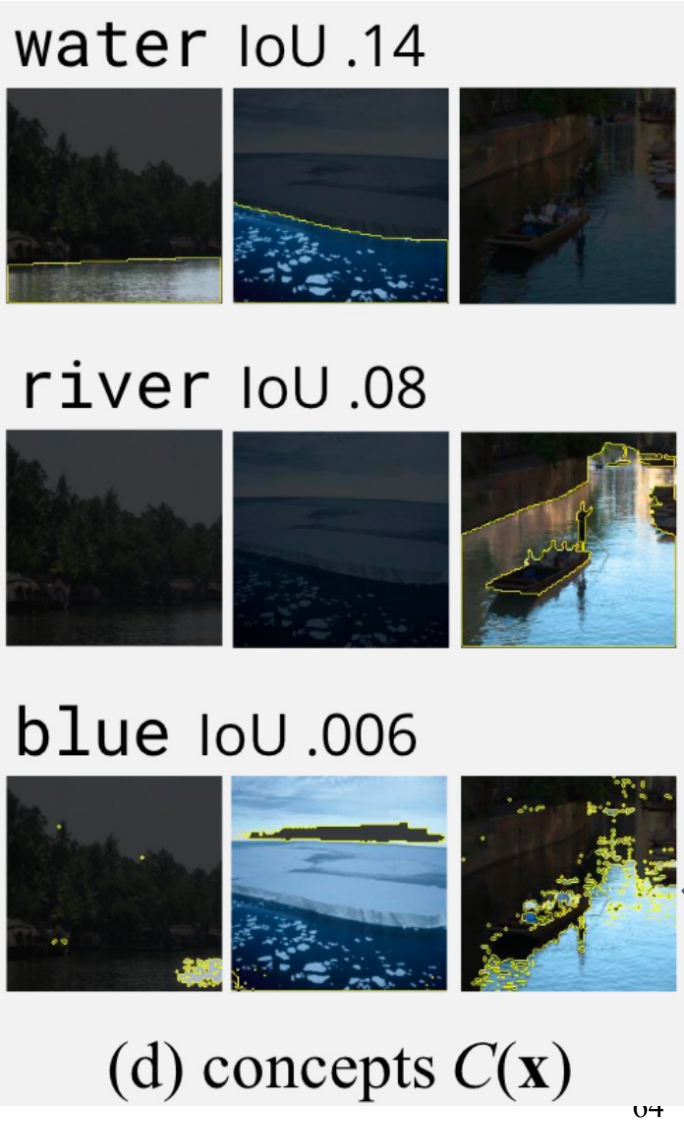
(c) neuron masks $M_{483}(\mathbf{x})$

Find neurons

- For an individual neuron, thresholdin
- Compare with the mask of concepts

(b) neuron $f_{483}(\mathbf{x})$

(c) neuron masks $M_{483}(\mathbf{x})$



(d) concepts $C(\mathbf{x})$

Find neurons

- Search for the most similar concept

$$\textsc{Explain-NetDissect}(n) = \arg\max_{C \in \mathcal{C}} \delta(n, C)$$

$$\delta(n, C) \triangleq \text{IoU}(n, C) = \left[\sum_{\mathbf{x}} \mathbb{1}(M_n(\mathbf{x}) \wedge C(\mathbf{x}))\right] / \left[\sum_{\mathbf{x}} \mathbb{1}(M_n(\mathbf{x}) \vee C(\mathbf{x}))\right]$$

Find logical forms induced from the concepts
Compose these concepts via compositional operations: AND OR NOT

Tasks

- Image Classification
  - Scene Recognition
  - ResNet-18
- NLI
  - SNLI
    - BiLSTM+MLP
    - Probe neurons in MLP, input is premise-hypothesis pairs
  - Concepts:
    - Penn Treebank POS tags + 2000 most common words
      - Appear in premise or hypothesis
    - Whether premise and hypothesis have more than 0%, 25%, 50%, or 75% word overlap
  - Additional Operator
    - NEIGHBORS(C), the union of 5 most close words with C
      - Judged by cosine similarity of Glove embeddings

Do neurons learn compositional concepts?

Unit 870 (gender-sensitive)

((((NOT hyp:man) AND pre:man) OR hyp:eating) AND (NOT pre:woman)) OR hyp:dancing

IoU **0.123** w_{entail} -0.046 $w_{neutral}$ -0.021 w_{contra} 0.040

Pre A guy pointing at a giant blackberry.
Hyp A woman tearing down a giant display.
Act **29.31** True **contra** Pred **contra**

Pre A man in a hat is working with...flowers.
Hyp Women are working with flowers.
Act **27.64** True **contra** Pred **contra**

Unit 99 (high overlap)

((NOT hyp:JJ) AND overlap-75% AND (NOT pre:people)) OR pre:basket OR pre:tv

IoU **0.118** w_{entail} 0.043 $w_{neutral}$ -0.029 w_{contra} -0.021

Pre A woman in a light blue jacket is riding a bike.
Hyp A women in a jacket riding a bike.
Act **19.13** True **entail** Pred **entail**

Pre A girl in a pumpkin dress sitting at a table.
Hyp There is a girl in a pumpkin dress sitting at a table.
Act **17.84** True **entail** Pred **entail**

Unit 15 (sitting only in hypothesis)

hyp:eating OR hyp:sitting OR hyp:sleeping OR hyp:sits AND (NOT pre:sits)

IoU **0.239** w_{entail} -0.083 $w_{neutral}$ -0.059 w_{contra} 0.086

Pre A person...is walking through an airport.
Hyp A woman sits in the lobby waiting on the doctor.
Act **30.68** True **contra** Pred **contra**

Pre A man jumps over another man...
Hyp Two men are sitting down, watching the game.
Act **27.64** True **contra** Pred **contra**

Unit 473 (unclear)

((NOT hyp:sleeping) AND (pre:NN OR pre:NNS)) AND (NOT hyp:alone) AND (NOT hyp:nobody)

IoU **0.586** w_{entail} 0.020 $w_{neutral}$ 0.016 w_{contra} -0.050

Pre A gentleman in a striped shirt gesturing with a stick...
Hyp A gentleman in a striped shirt joyously gesturing.
Act **31.62** True **neutral** Pred **neutral**

Pre An Asian man in a...uniform diving...in a game.
Hyp A person in a uniform does something.
Act **29.76** True **neutral** Pred **entail**

Can we target explanations to change model behavior?

Unit 39 (nobody in hypothesis)

`hyp:nobody AND (NOT pre:hair) AND (NOT pre:RB) AND (NOT pre:'s)`
IoU **0.465** w_{entail} -0.117 $w_{neutral}$ -0.053 w_{contra} 0.047

Pre Three women prepare a meal in a kitchen.
Orig Hyp The ladies are cooking.
Adv Hyp ***Nobody but*** the ladies are cooking.
True **entail** $\xrightarrow{adv}$ **neutral** Pred **entail** $\xrightarrow{adv}$ **contra**

Unit 15 (sitting only in hypothesis)

`hyp:eating OR hyp:sitting OR hyp:sleeping OR hyp:sits AND (NOT pre:sits)`
IoU **0.239** w_{entail} -0.083 $w_{neutral}$ -0.059 w_{contra} 0.086

Orig Pre A blond woman is holding 2 golf balls while reaching down into a golf hole.
Adv Pre A blond woman is holding 2 golf balls.
Hyp A blond woman is sitting down.
True **contra** $\xrightarrow{adv}$ **neutral** Pred **contra** $\xrightarrow{adv}$ **contra**

Unit 133 (couch words in hypothesis)

`NEIGHBORS(hyp:couch) OR hyp:inside OR hyp:home OR hyp:indoors OR hype:eating`
IoU **0.202** w_{entail} -0.125 $w_{neutral}$ -0.024 w_{contra} 0.088

Pre 5 women sit around a table doing some crafts.
Orig Hyp 5 women sit around a table.
Adv Hyp 5 women sit around a table ***near a couch***.
True **entail** $\xrightarrow{adv}$ **neutral** Pred **entail** $\xrightarrow{adv}$ **contra**

Unit 941 (inside/indoors in hypothesis)

`hyp:inside OR hyp:not OR hyp:indoors OR hyp:moving OR hyp:something`
IoU **0.151** w_{entail} 0.086 $w_{neutral}$ -0.030 w_{contra} -0.023

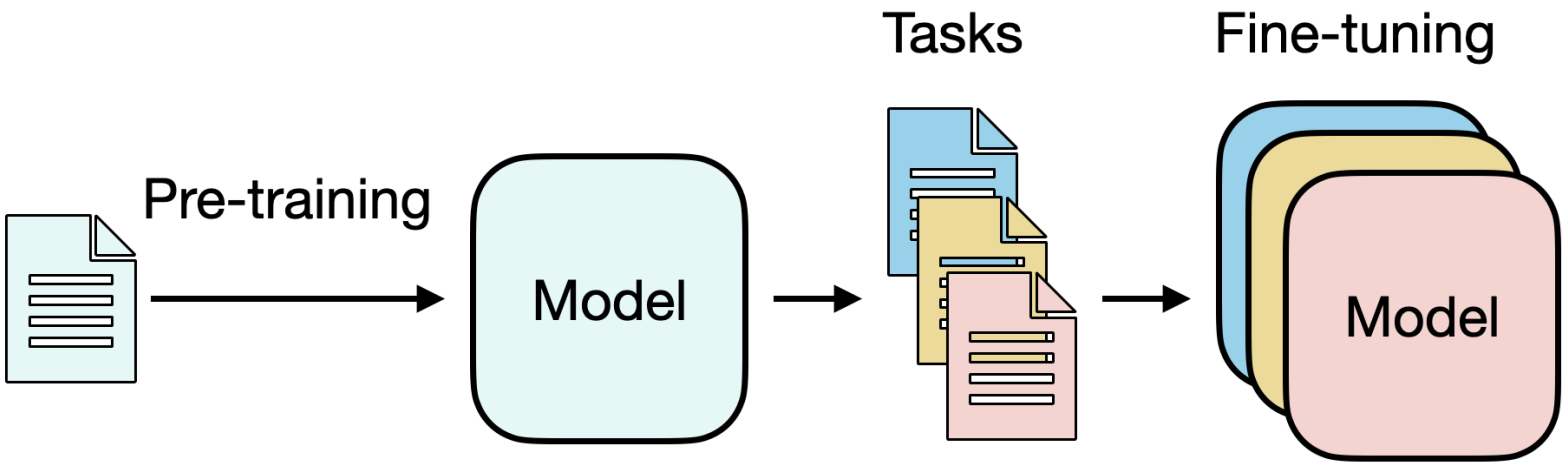
Orig Pre Two people are sitting in a station.
Adv Pre Two people are sitting in a ***pool***.
Hyp A couple of people are inside and not standing.
True **entail** $\xrightarrow{adv}$ **neutral** Pred **entail** $\xrightarrow{adv}$ **entail**

Cognitive Abilities of Big Models

THUNLP

Task generalization of PLM

- Question: why can PLMs easily adapt to various NLP tasks even with small-scale data?



Task generalization of PLM

- PLM acquires versatile knowledge during pre-training, which can be leveraged to solve various tasks

Customer Service

ActiveChat.ai

Games

AI Dungeon

Chatbots

AI Buddy

Customer Service

Chatdesk

LegalTech

aiLawDocs

Semantic Search

Algolia

Chatbots

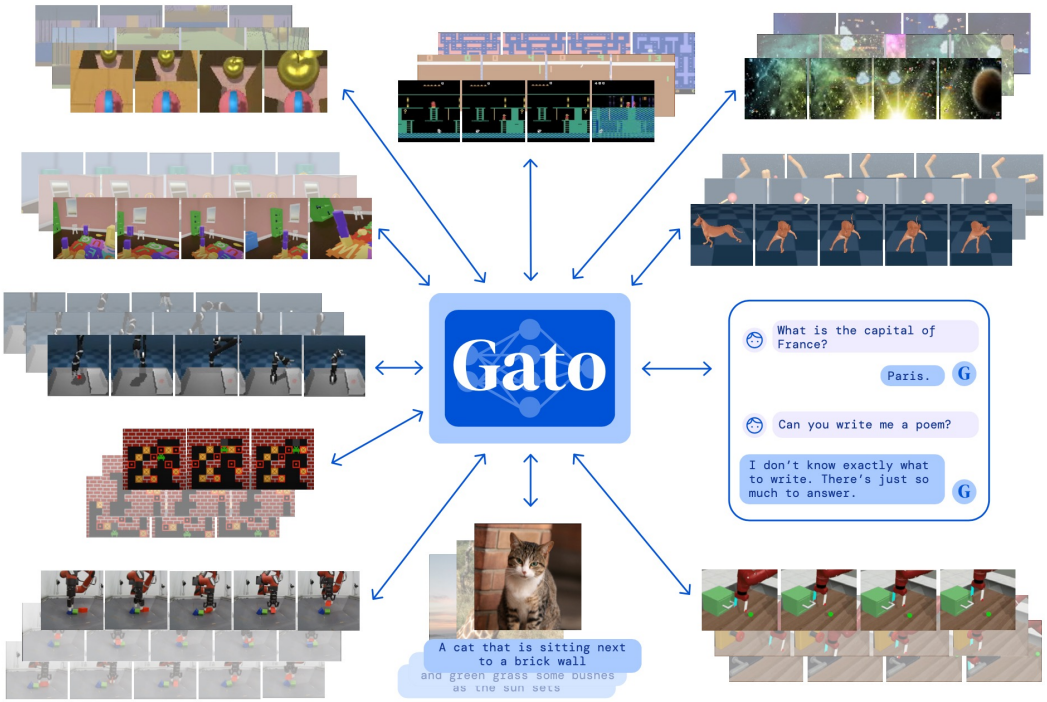
AskBrian

AI Copywriting

Conto AI

Cognitive abilities of PLMs

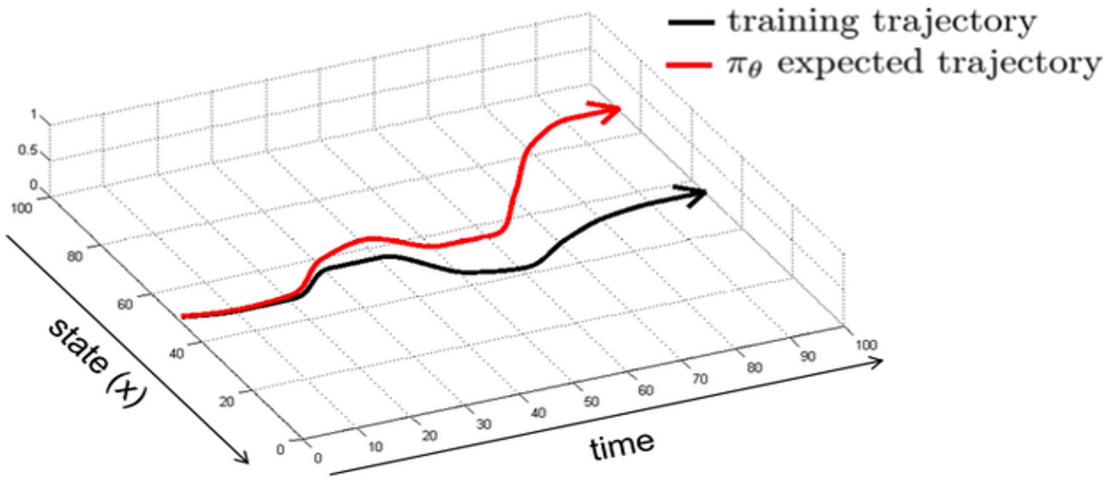
- Recent studies have shown that PLMs also have cognitive abilities, and can manipulate existing tools to complete a series of complex tasks



Reed, Scott, et al. "A generalist agent." arXiv preprint arXiv:2205.06175 (2022).

Fundamentals & framework

- Imitation learning in RL
  - Learning from behaviors instead of accumulating rewards
  - State-action pairs $\mathcal{D} = \{(s_1, a_1), (s_2, a_2), (s_3, a_3), \ldots\}$
  - State as features and action as labels
  - Target: imitate the trajectory of behaviors



Fundamentals & framework

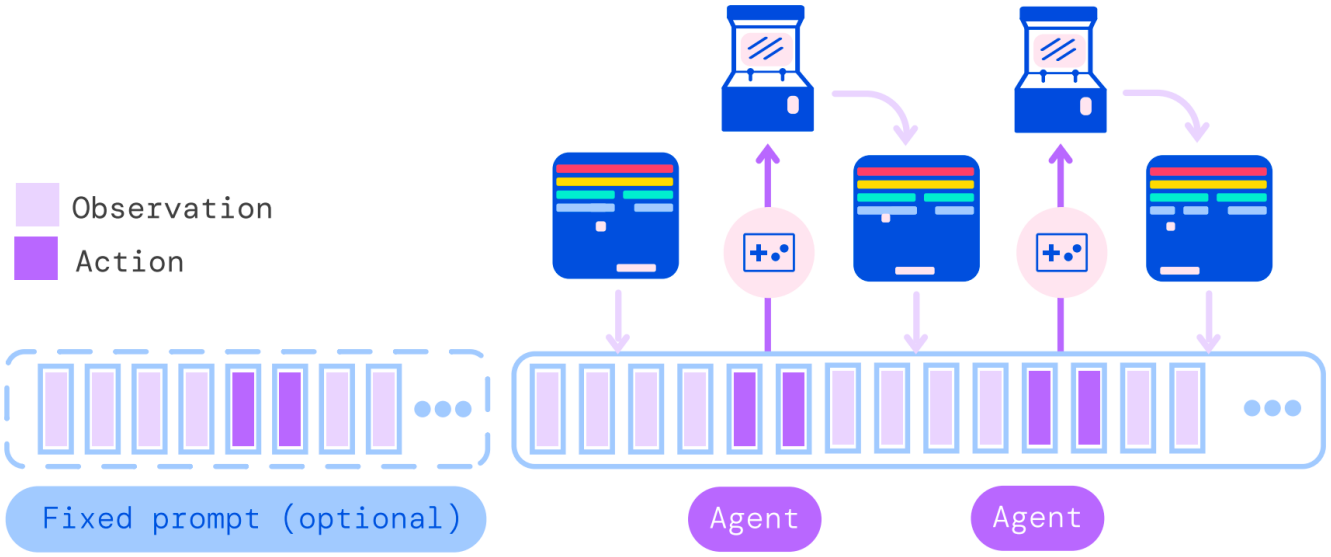
- Large-scale pre-trained models
  - Universal knowledge learned from pre-training
- Interactive space
  - An environment that models could interact with
  - State space: display states and changes
  - Action space: a set of pre-defined actions in the environment

Fundamentals & framework

- Given a goal, we model each action and state to achieve the goal in a unified space by a PLM
- Tokenization
  - Tokenize human behaviors (actions in the action space) and states in the state space to a same space
  - The tokenized information could be processed by PLM
- Directly training
  - The behaviors could be autoregressively predicted

Outline

- Interactive space:
  - Search engine
  - WebShop
  - Sandbox



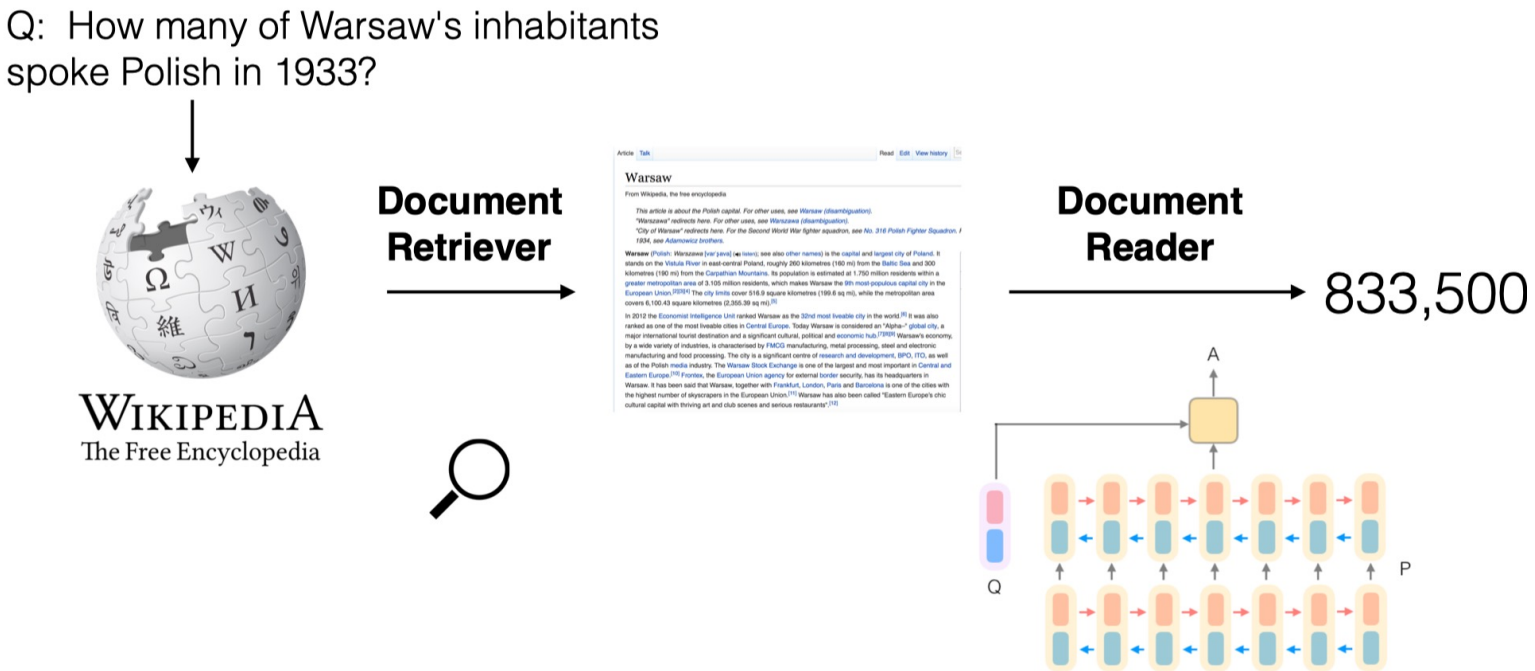
Interactive space: search engine

- A rising challenge in NLP is long-form QA
  - A paragraph-length answer is generated in response to an open-ended question

Question	Why did we decide that certain words were "bad" and shouldn't be used in social settings?
Answer	We have all been taught that there are certain words that should not be used in public, at least not without good reason. But why do some words fall into this category while others do not? It turns out that there is not a single answer to this question, as each word has a different history and carries different meanings in different cultures and families [2]. However, there are some general categories that "bad" words tend to fall into. "Bad" words generally relate to parts of life that we don't like talking about in public, like bathroom functions, or negative ways of talking about people's religion, skin color, ability, or gender [3]. Often, people say bad words when they are experiencing strong emotions, and those words are linked to our emotions [3].
References (titles only)	[1, 2, 3] Why Are Some Words 'Bad'? ǀ Vermont Public Radio (www.vpr.org) [4] On Words: 'Bad' Words and Why We Should Study Them ǀ UVA Today (news.virginia.edu) [5] The Science of Curse Words: Why The &@$! Do We Swear? (www.babbel.com)

Interactive space: search engine

- The task has two core components:
  - information retrieval and information synthesis



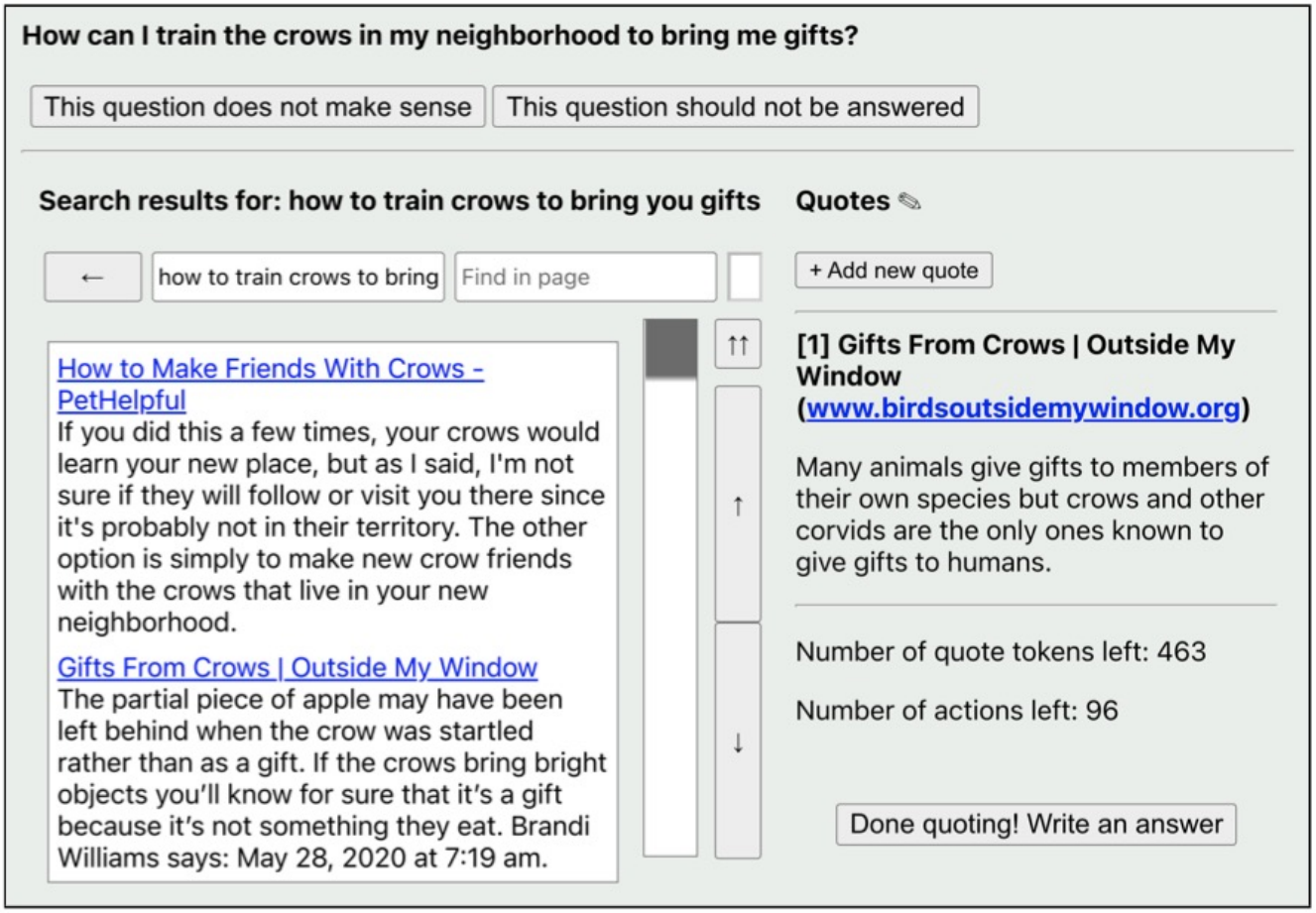
Interactive space: search engine

- WebGPT
  - Outsource document retrieval to the Microsoft Bing Web Search API
  - Utilize unsupervised pre-training to achieve high-quality document synthesis by fine-tuning GPT-3
  - Create a text-based web-browsing environment that both humans and language models can interact with

Nakano, Reiichiro, et al. "WebGPT: Browser-assisted question-answering with human feedback." arXiv preprint arXiv:2112.09332 (2021).

Interactive space: search engine

- Text-based web-browser



Interactive space: search engine

- Text-based web-browser

Command	Effect
`Search <query>`	Send `<query>` to the Bing API and display a search results page
`Clicked on link <link ID>`	Follow the link with the given ID to a new page
`Find in page: <text>`	Find the next occurrence of `<text>` and scroll to it
`Quote: <text>`	If `<text>` is found in the current page, add it as a reference
`Scrolled down <1, 2, 3>`	Scroll down a number of times
`Scrolled up <1, 2, 3>`	Scroll up a number of times
`Top`	Scroll to the top of the page
`Back`	Go to the previous page
`End: Answer`	End browsing and move to answering phase
`End: <Nonsense, Controversial>`	End browsing and skip answering phase

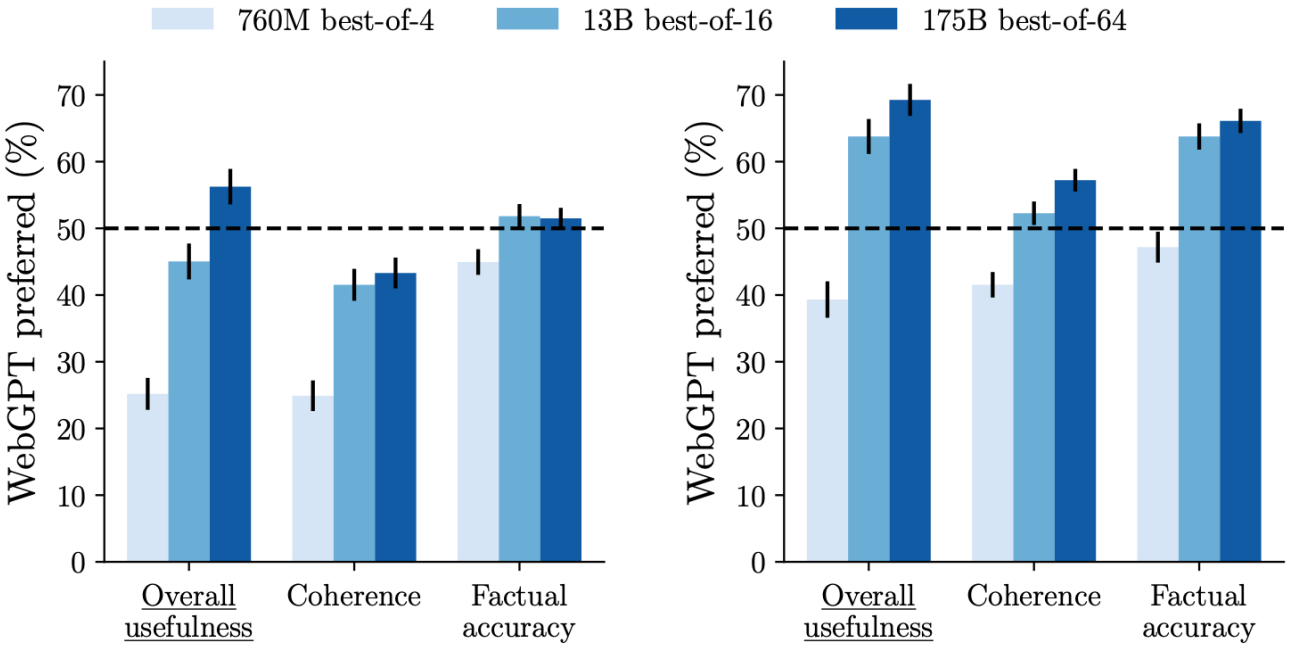
Interactive space: search engine

- Text-based web-browser

```
♦Question
How can I train the crows in my neighborhood to bring me gifts?

♦Quotes
From Gifts From Crows | Outside My Window (www.birdsoutsidemywindow.org)
> Many animals give gifts to members of their own species but crows and
other corvids are the only ones known to give gifts to humans.

♦Past actions
Search how to train crows to bring you gifts
Click Gifts From Crows | Outside My Window www.birdsoutsidemywindow.org
Quote
Back

♦Title
Search results for: how to train crows to bring you gifts

♦Scrollbar: 0 - 11
♦Text
【0†How to Make Friends With Crows - PetHelpful†pethelpful.com】
If you did this a few times, your crows would learn your new place, but
as I said, I'm not sure if they will follow or visit you there since it's
probably not in their territory. The other option is simply to make new
crow friends with the crows that live in your new neighborhood.

【1†Gifts From Crows | Outside My Window†www.birdsoutsidemywindow.org】
The partial piece of apple may have been left behind when the crow was
startled rather than as a gift. If the crows bring bright objects you'll
know for sure that it's a gift because it's not something they eat.
Brandi Williams says: May 28, 2020 at 7:19 am.

♦Actions left: 96
♦Next action
```

Interactive space: search engine

- WebGPT-produced answers are more preferred than human-generated ones
- Better coherence and factual accuracy

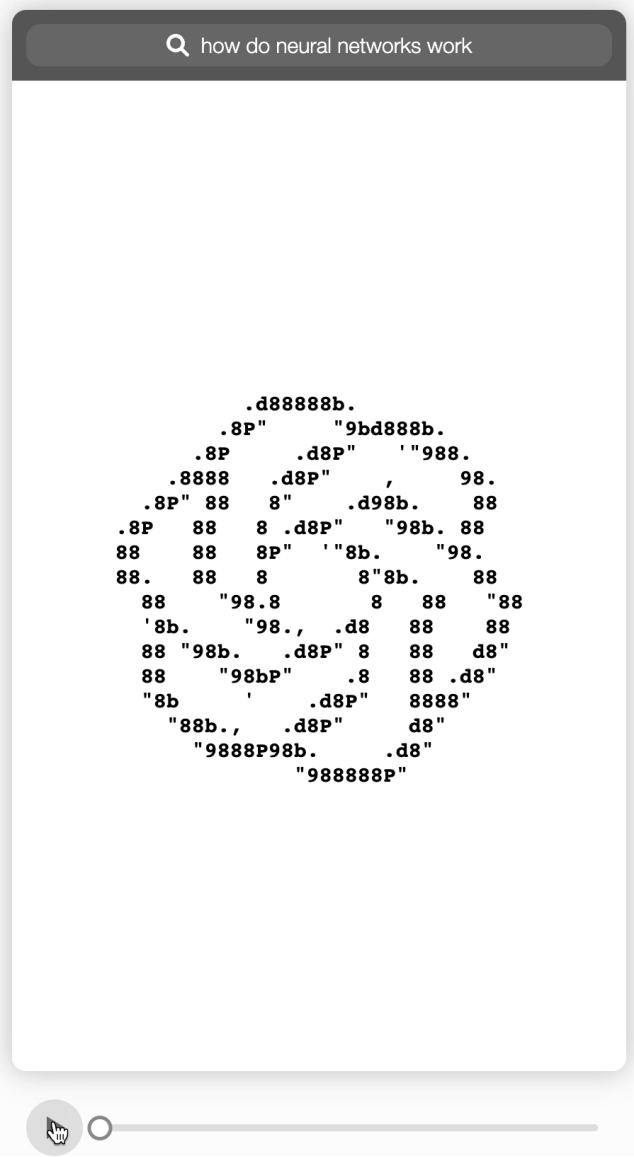


(a) WebGPT vs. human demonstrations.

(b) WebGPT vs. ELI5 reference answers.

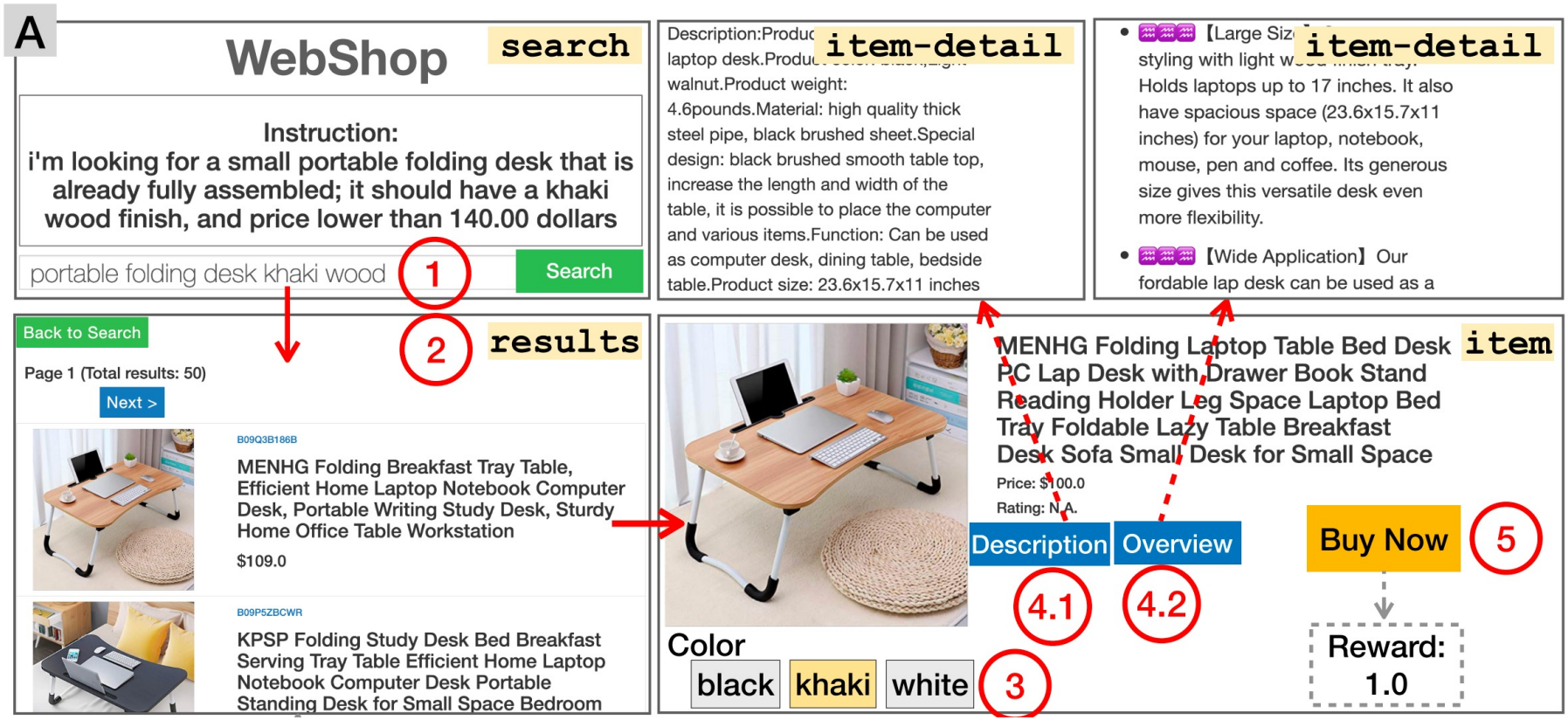
Interactive space: search engine

- An example:
  - How does neural networks work?



Interactive space: WebShop

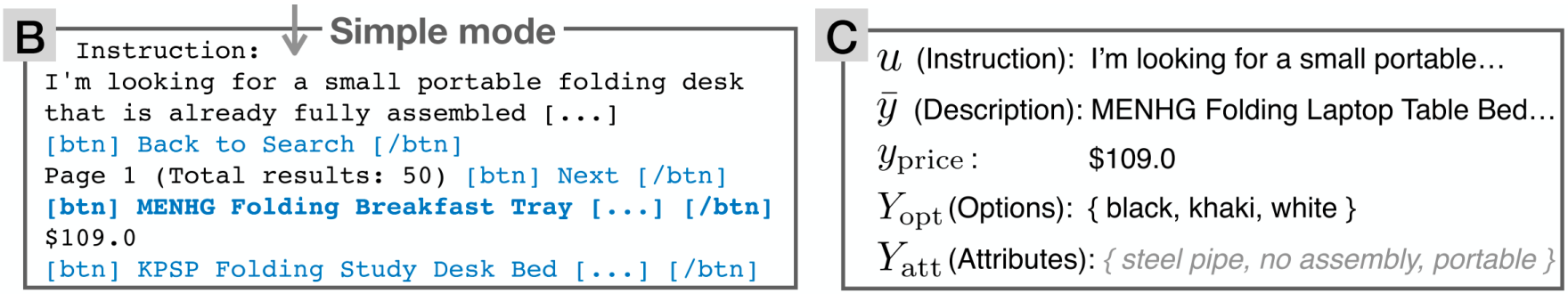
- WebShop for online shopping (HTML mode)



Yao, Shunyu, et al. "WebShop: Towards Scalable Real-World Web Interaction with Grounded Language Agents." arXiv e-prints (2022): arXiv-2207.

Interactive space: WebShop

- Simple mode which strips away extraneous meta-data from raw HTML into a simpler format



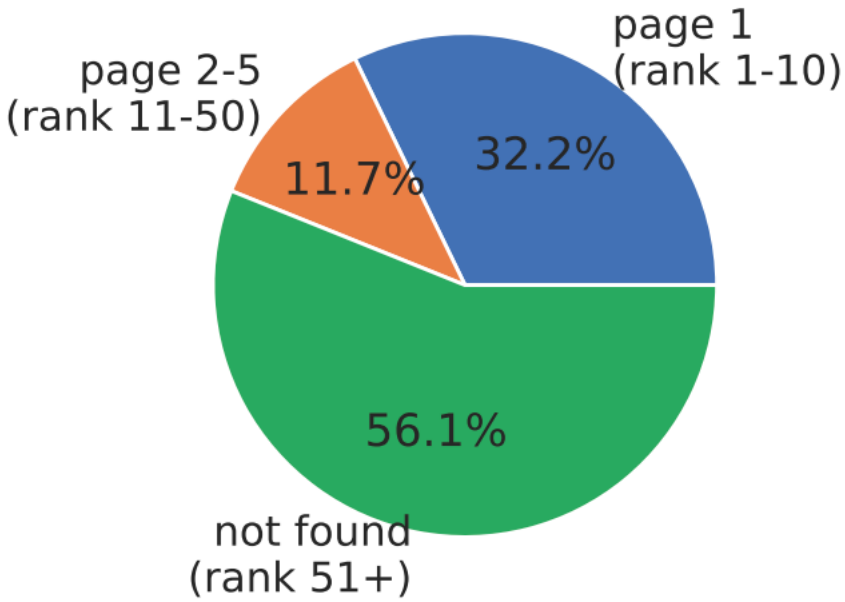
Interactive space: WebShop

- Actions in WebShop

Type	Argument	State → Next State
`search`	[*Query*]	Search → Results
`choose`	Back to search	$*$ → Search
`choose`	Prev/Next page	Results → Results
`choose`	[*Product title*]	Results → Item
`choose`	[*Option*]	Item → Item
`choose`	Desc/Overview	Item → Item-Detail
`choose`	Previous	Item-Detail → Item
`choose`	Buy	Item → Episode End

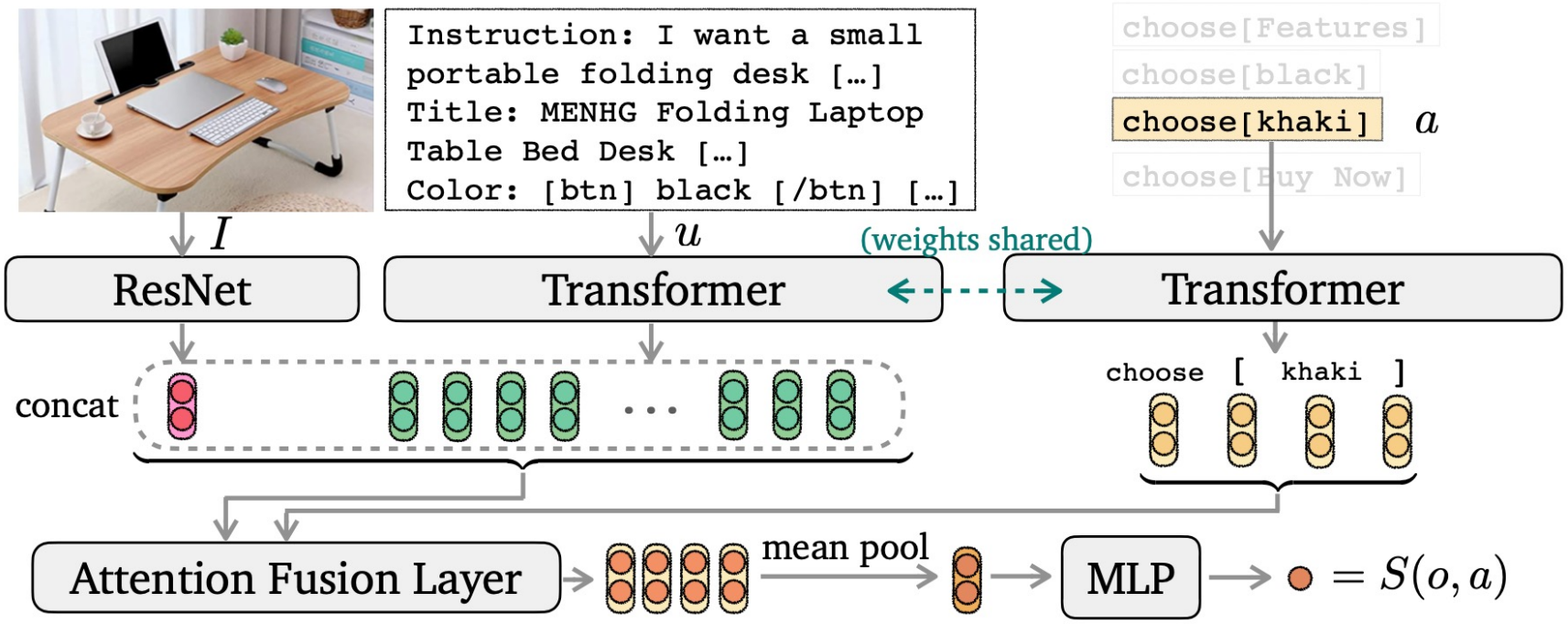
Interactive space: WebShop

- Item rank in search results when the instruction is directly used as search query



Interactive space: WebShop

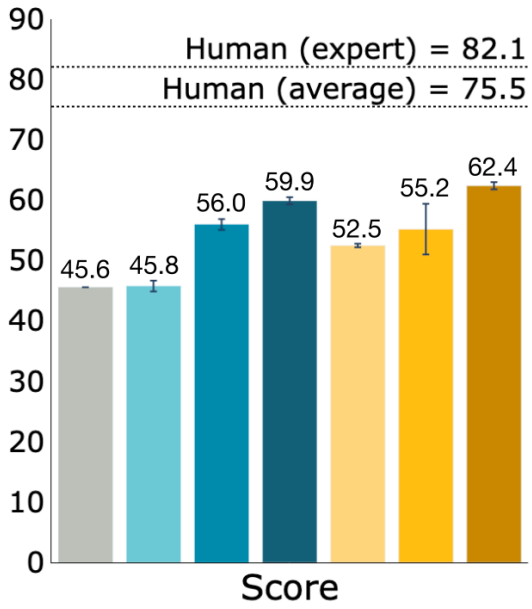
- Model implementation

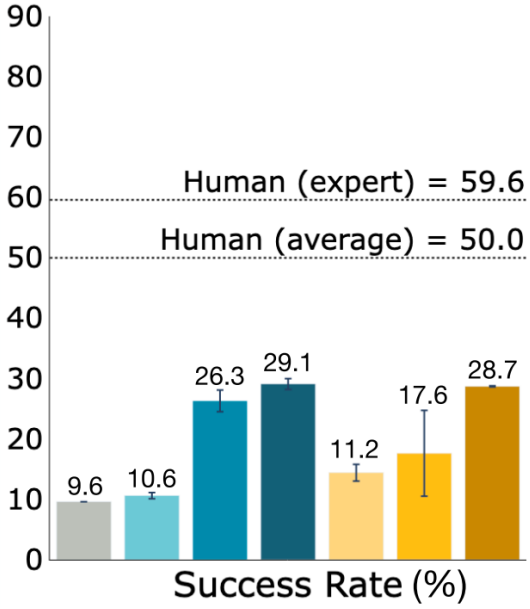


Interactive space: WebShop

- Results

	LP Search	LP Choice	Human Demo	Use Reward
Rule				
IL w/o LP Choice	✓		✓	
IL w/o LP Search		✓	✓	
IL	✓	✓	✓	
RL	✓			✓
RL (RNN)				✓
IL+RL	✓	✓	✓	✓

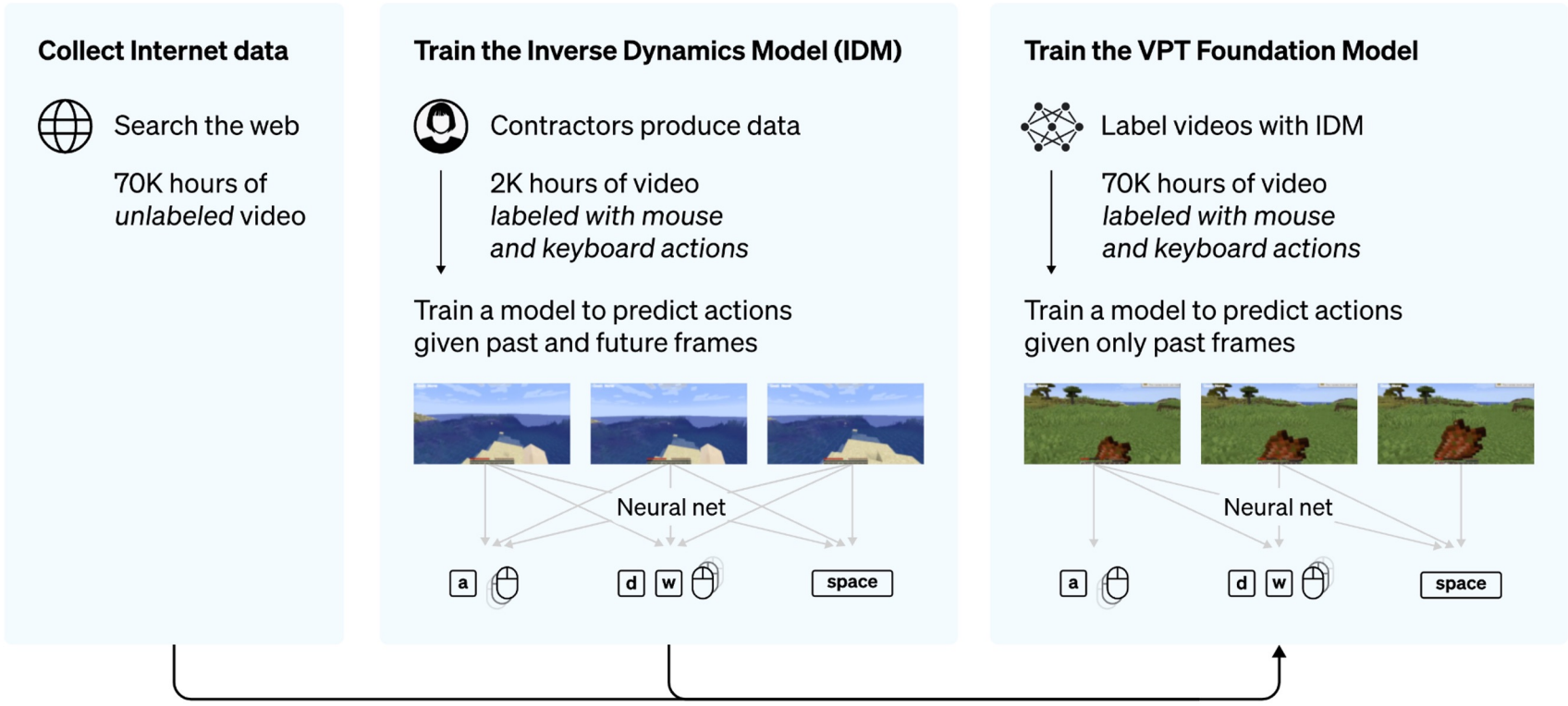




Click and play

Interactive space: Sandbox

- Video pre-training of MineCraft
  - Sandbox like Minecraft is a good interactive space



Baker, Bowen, et al. "Video PreTraining (VPT): Learning to Act by Watching Unlabeled Online Videos." arXiv preprint arXiv:2206.11795 (2022).

Interactive space: Sandbox

- Video pre-training of MineCraft
  - Define discrete actions in the interactive space

Action	Human action	Description
`forward`	W key	Move forward.
`back`	S key	Move backward.
`left`	A key	Strafe left.
`right`	D key	Strafe right.
`jump`	space key	Jump.
`inventory`	E key	Open or close inventory and the 2x2 crafting grid.
`sneak`	shift key	Move carefully in current direction of motion. In the GUI it acts as a modifier key: when used with `attack` it moves item from/to the inventory to/from the hot-bar, and when used with `craft` it crafts the maximum number of items possible instead of just 1.
`sprint`	ctrl key	Move fast in the current direction of motion.

Interactive space: Sandbox

- Cost
  - Use behavior model to annotate unlabeled 70 hours video
  - Reduce the cost: 1,400,000 $ -> 130,000 $
- Annotation trick
  - At first, casually playing MineCraft
  - Play specific tasks (Equip Capabilities)

Collect as many units of wood as possible, using only wooden or stone tools
Start a new world every 30 minutes of game play
Build a basic house in 10 minutes using only dirt, wood, sand, and either wooden or stone tools
Starting from a new world and an empty inventory, find resources and craft a diamond pickaxe in 20 minutes

Interactive space: Sandbox

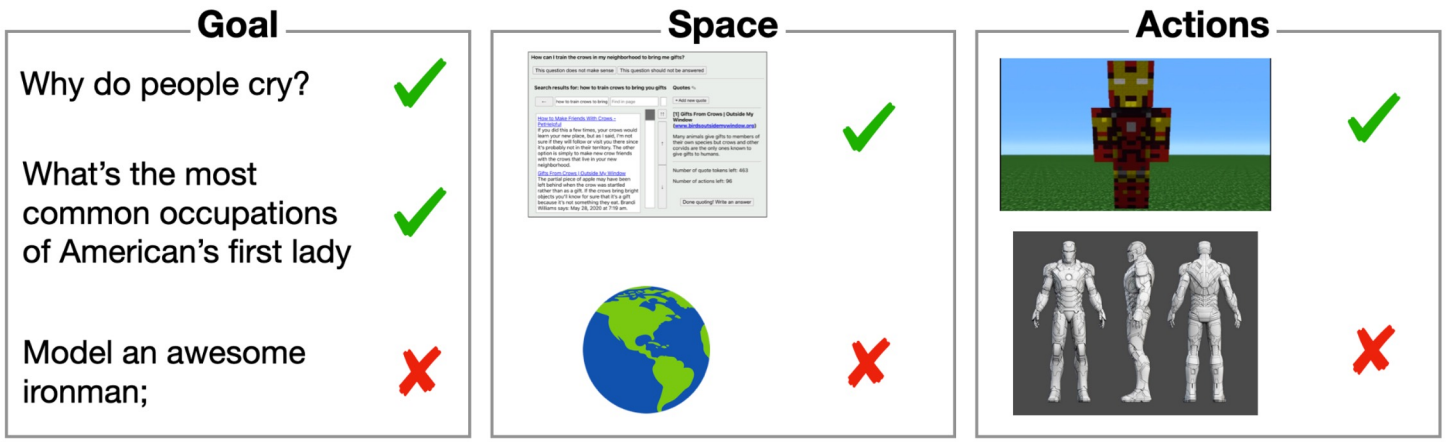
- Results
  - VPT accomplishes tasks impossible to learn with RL alone, such as crafting planks and crafting tables (tasks requiring a human proficient of ~970 consecutive actions)

Interactive space: Sandbox

- Results
  - An example for killing a cow

Challenges & limitations

- Challenges of the current framework
  - Building interactive space is time-consuming
  - Labeling is expensive and labor-intensive
  - The goal must be clear and simple
  - Only discrete actions and states are supported
  - A clean interactive space is required



Thanks for listening

THUNLP