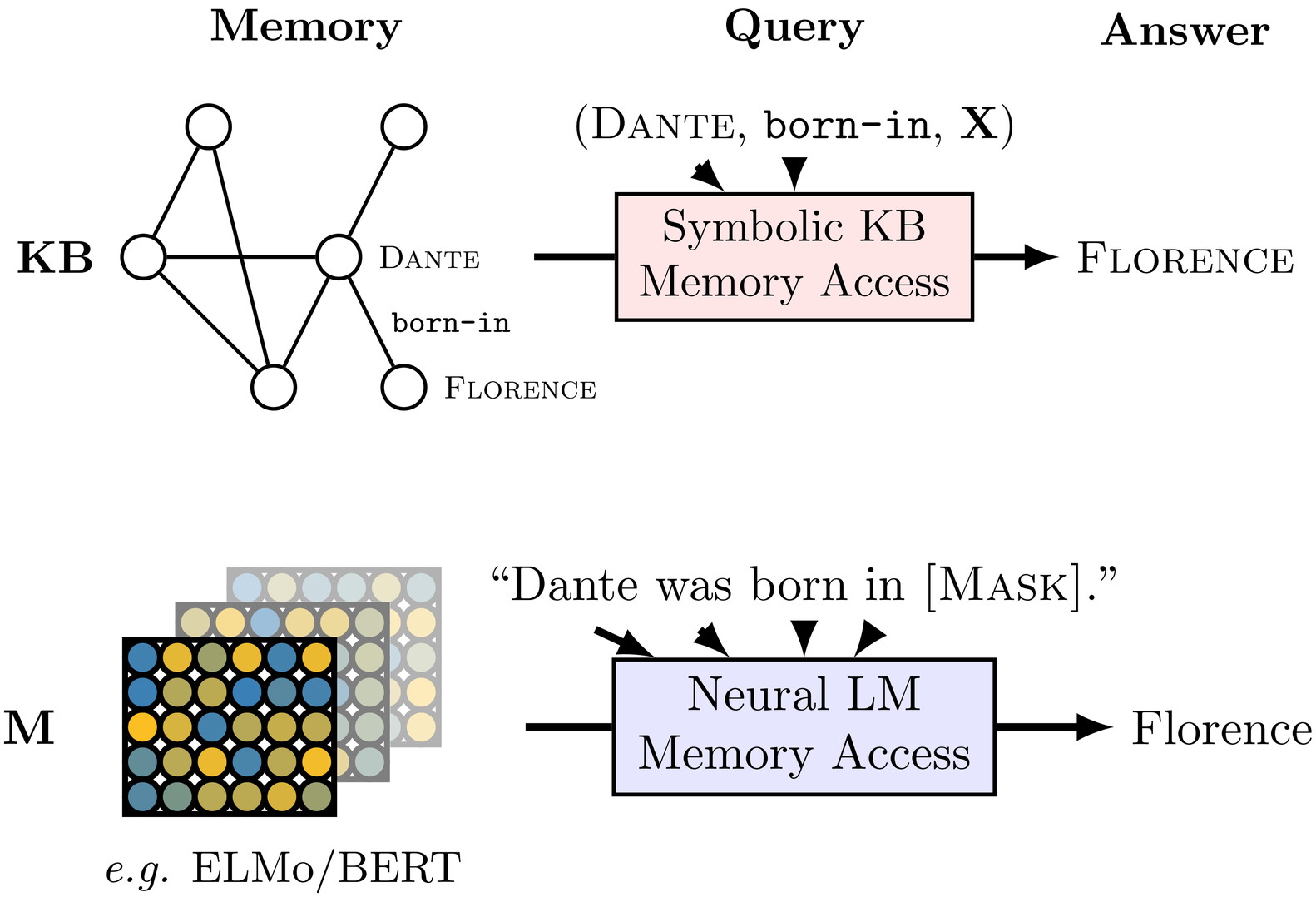
Knowledge Neurons in Pretrained Transformers

Damai Dai[†‡*], **Li Dong**[‡], **Yaru Hao**[‡], **Zhifang Sui**[†], **Baobao Chang**[†], **Furu Wei**[‡]
[†]MOE Key Lab of Computational Linguistics, Peking University
[‡]Microsoft Research

ACL2022

AntNLP—纪焘

Background: LM = KB?



Background: LM = KB?

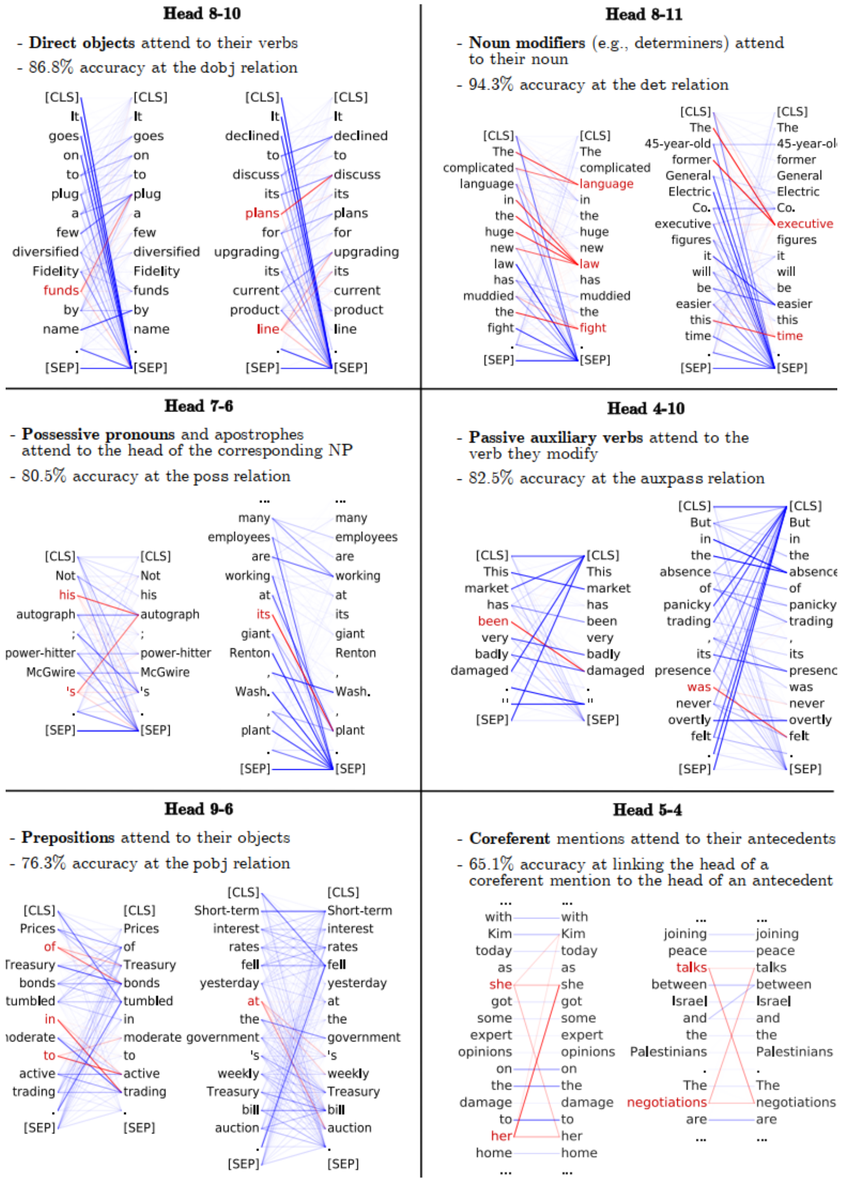
EMNLP19 **Language Models as Knowledge Bases?**

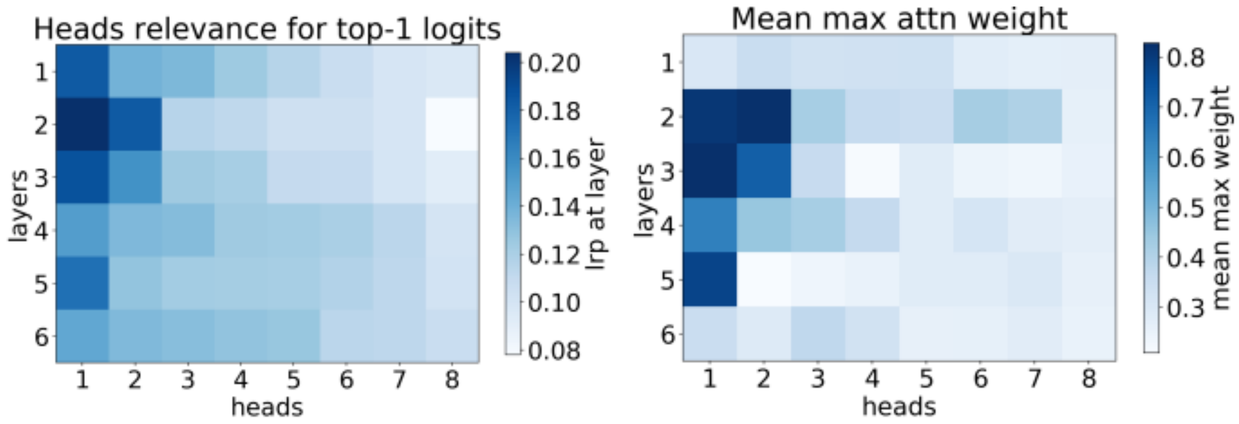
Fabio Petroni[1] **Tim Rocktäschel**[1,2] **Patrick Lewis**[1,2] **Anton Bakhtin**[1]
Yuxiang Wu[1,2] **Alexander H. Miller**[1] **Sebastian Riedel**[1,2]
[1]Facebook AI Research
[2]University College London

Background: fill-in-the-blank

	Relation	Query	Answer	Generation
T-Rex	P19	Francesco Bartolomeo Conti was born in ____.	Florence	Rome [-1.8] , **Florence [-1.8]** , Naples [-1.9] , Milan [-2.4] , Bologna [-2.5]
	P20	Adolphe Adam died in ____.	Paris	**Paris [-0.5]** , London [-3.5] , Vienna [-3.6] , Berlin [-3.8] , Brussels [-4.0]
	P279	English bulldog is a subclass of ____.	dog	dogs [-0.3] , breeds [-2.2] , **dog [-2.4]** , cattle [-4.3] , sheep [-4.5]
	P37	The official language of Mauritius is ____.	English	**English [-0.6]** , French [-0.9] , Arabic [-6.2] , Tamil [-6.7] , Malayalam [-7.0]
	P413	Patrick Oboya plays in ____ position.	midfielder	centre [-2.0] , center [-2.2] , **midfielder [-2.4]** , forward [-2.4] , midfield [-2.7]
	P138	Hamburg Airport is named after ____.	Hamburg	Hess [-7.0] , Hermann [-7.1] , Schmidt [-7.1] , **Hamburg [-7.5]** , Ludwig [-7.5]
	P364	The original language of Mon oncle Benjamin is ____.	French	**French [-0.2]** , Breton [-3.3] , English [-3.8] , Dutch [-4.2] , German [-4.9]
	P54	Dani Alves plays with ____ .	Barcelona	Santos [-2.4] , Porto [-2.5] , Sporting [-3.1] , Brazil [-3.3] , Portugal [-3.7]
	P106	Paul Toungui is a ____ by profession .	politician	lawyer [-1.1] , journalist [-2.4] , teacher [-2.7] , doctor [-3.0] , physician [-3.7]
	P527	Sodium sulfide consists of ____.	sodium	water [-1.2] , sulfur [-1.7] , **sodium [-2.5]** , zinc [-2.8] , salt [-2.9]
	P102	Gordon Scholes is a member of the ____ political party.	Labor	Labour [-1.3] , Conservative [-1.6] , Green [-2.4] , Liberal [-2.9] , **Labor [-2.9]**
	P530	Kenya maintains diplomatic relations with ____.	Uganda	India [-3.0] , **Uganda [-3.2]** , Tanzania [-3.5] , China [-3.6] , Pakistan [-3.6]
	P176	iPod Touch is produced by ____.	Apple	**Apple [-1.6]** , Nokia [-1.7] , Sony [-2.0] , Samsung [-2.6] , Intel [-3.1]
	P30	Bailey Peninsula is located in ____.	Antarctica	**Antarctica [-1.4]** , Bermuda [-2.2] , Newfoundland [-2.5] , Alaska [-2.7] , Canada [-3.1]
	P178	JDK is developed by ____.	Oracle	IBM [-2.0] , Intel [-2.3] , Microsoft [-2.5] , HP [-3.4] , Nokia [-3.5]
	P1412	Carl III used to communicate in ____.	Swedish	German [-1.6] , Latin [-1.9] , French [-2.4] , English [-3.0] , Spanish [-3.0]
	P17	Sunshine Coast, British Columbia is located in ____.	Canada	**Canada [-1.2]** , Alberta [-2.8] , Yukon [-2.9] , Labrador [-3.4] , Victoria [-3.4]
	P39	Pope Clement VII has the position of ____ .	pope	cardinal [-2.4] , Pope [-2.5] , **pope [-2.6]** , President [-3.1] , Chancellor [-3.2]
	P264	Joe Cocker is represented by music label ____.	Capitol	EMI [-2.6] , BMG [-2.6] , Universal [-2.8] , **Capitol [-3.2]** , Columbia [-3.3]
	P276	London Jazz Festival is located in ____.	London	**London [-0.3]** , Greenwich [-3.2] , Chelsea [-4.0] , Camden [-4.6] , Stratford [-4.8]
	P127	Border TV is owned by ____.	ITV	Sky [-3.1] , **ITV [-3.3]** , Global [-3.4] , Frontier [-4.1] , Disney [-4.3]
	P103	The native language of Mammootty is ____.	Malayalam	**Malayalam [-0.2]** , Tamil [-2.1] , Telugu [-4.8] , English [-5.2] , Hindi [-5.6]
	P495	The Sharon Cuneta Show was created in ____.	Philippines	Manila [-3.2] , **Philippines [-3.6]** , February [-3.7] , December [-3.8] , Argentina [-4.0]
ConceptNet	AtLocation	You are likely to find a overflow in a ____.	drain	sewer [-3.1] , canal [-3.2] , toilet [-3.3] , stream [-3.6] , **drain [-3.6]**
	CapableOf	Ravens can ____.	fly	**fly [-1.5]** , fight [-1.8] , kill [-2.2] , die [-3.2] , hunt [-3.4]
	CausesDesire	Joke would make you want to ____.	laugh	cry [-1.7] , die [-1.7] , **laugh [-2.0]** , vomit [-2.6] , scream [-2.6]
	Causes	Sometimes virus causes ____.	infection	disease [-1.2] , cancer [-2.0] , **infection [-2.6]** , plague [-3.3] , fever [-3.4]
	HasA	Birds have ____.	feathers	wings [-1.8] , nests [-3.1] , **feathers [-3.2]** , died [-3.7] , eggs [-3.9]
	HasPrerequisite	Typing requires ____.	speed	patience [-3.5] , precision [-3.6] , registration [-3.8] , accuracy [-4.0] , **speed [-4.1]**
	HasProperty	Time is ____.	finite	short [-1.7] , passing [-1.8] , precious [-2.9] , irrelevant [-3.2] , gone [-4.0]

Background: Attention Explanation

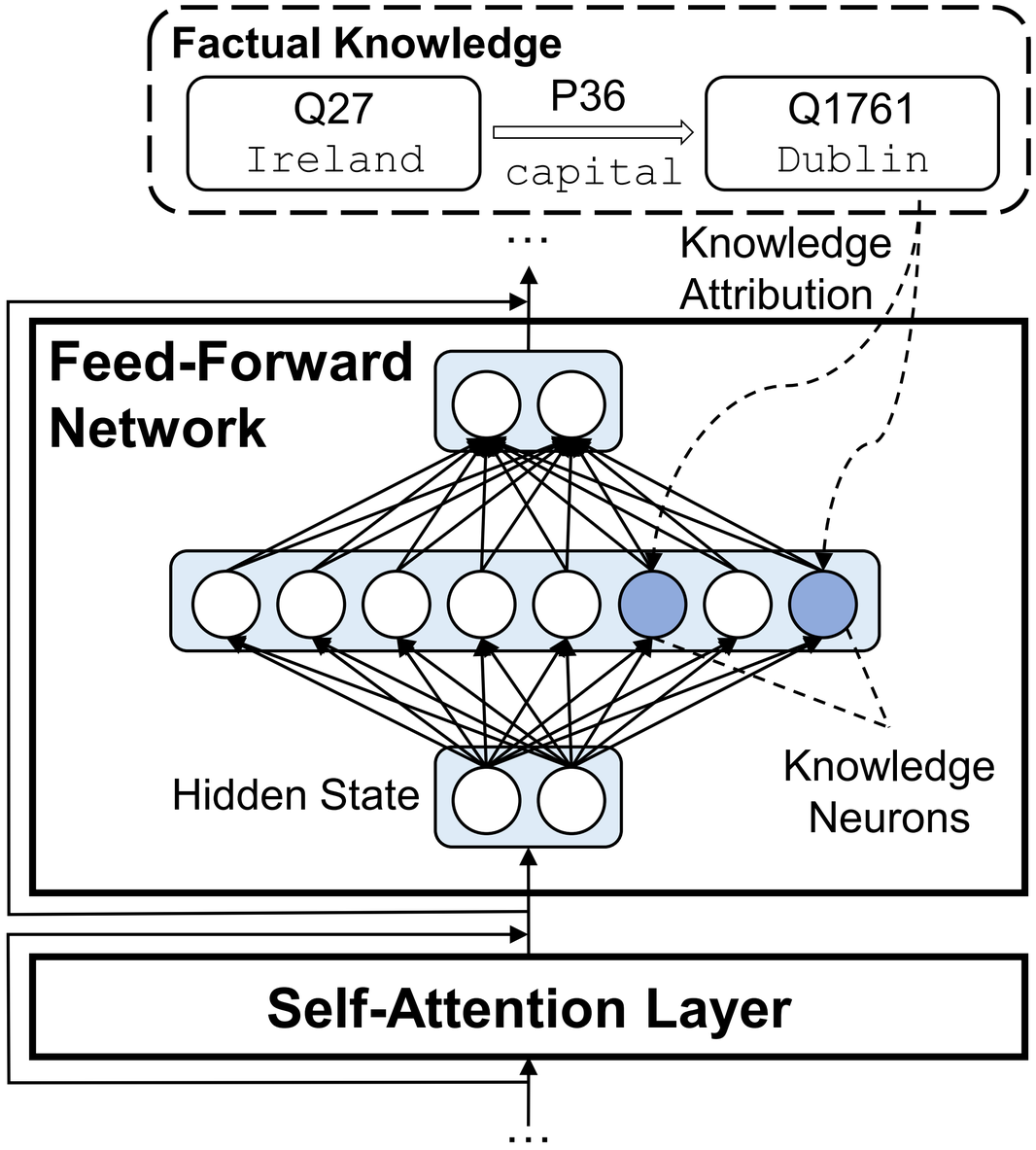




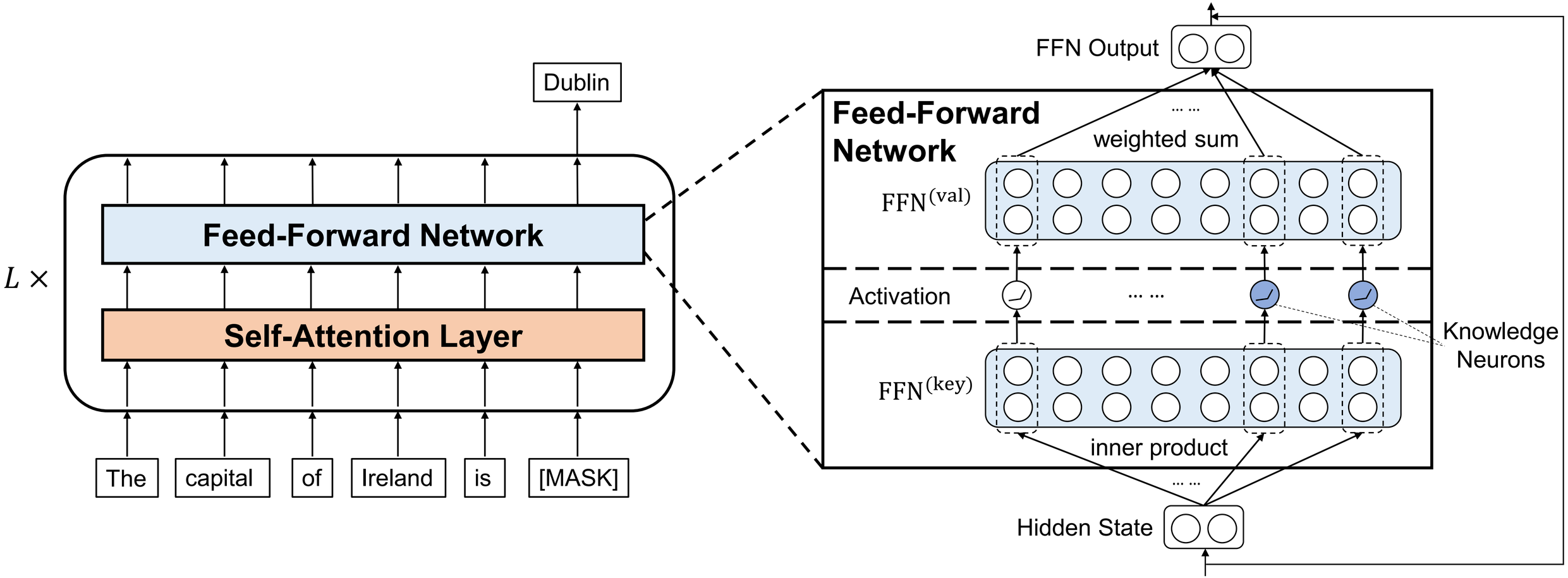
Main Q&C

- Unsupervised F1 @ Fact Triples → Knowledge Neurons
- MHA → FFN（more parameters, more knowledge）
- Control Knowledge Neuron（few & general）

Knowledge Neurons



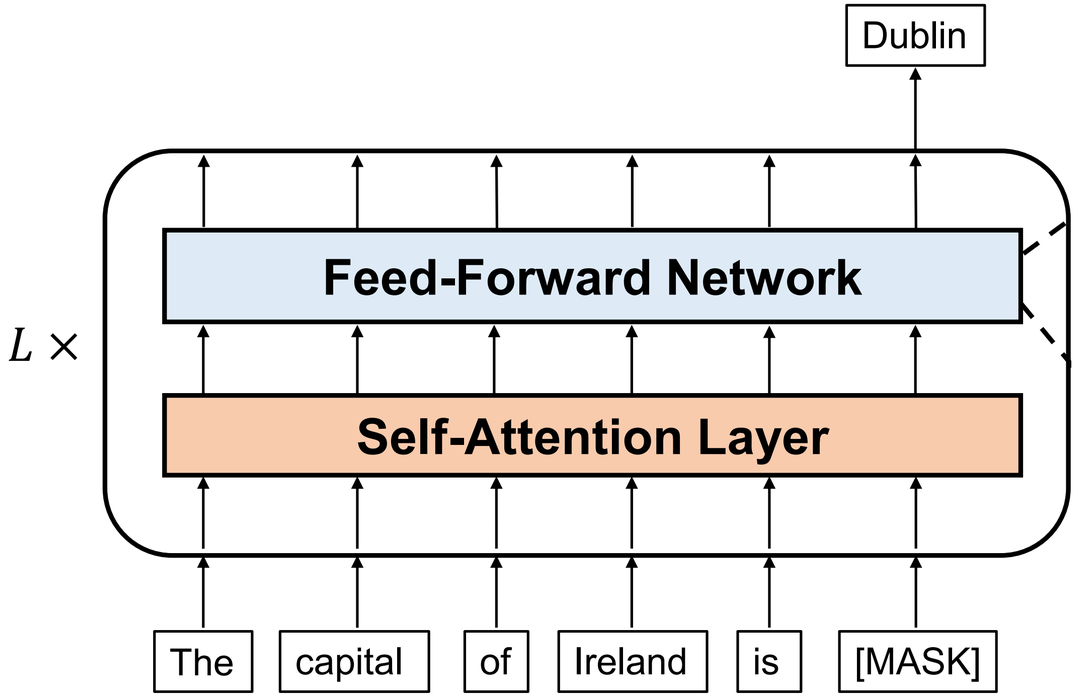
Identifying Knowledge Neurons



How?

Why?

Fill-in-the-blank



$\langle h, r, t \rangle$

→ $\langle$`Ireland, capital, Dublin`$\rangle$

→ "*The capital of Ireland is* __"

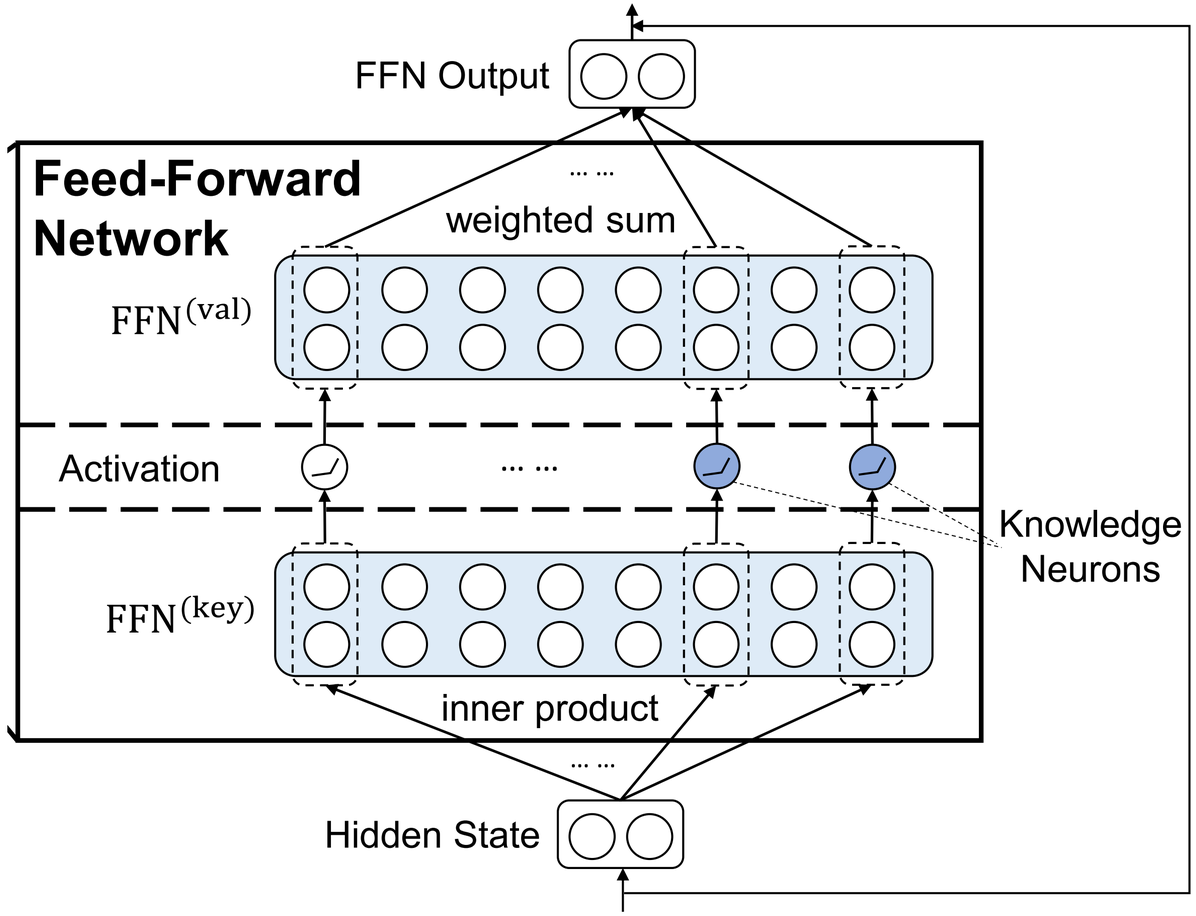
Fill-in-the-blank

Relations	Template #1	Template #2	Template #3
P176 (`manufacturer`)	[X] is produced by [Y]	[X] is a product of [Y]	[Y] and its product [X]
P463 (`member_of`)	[X] is a member of [Y]	[X] belongs to the organization of [Y]	[X] is affiliated with [Y]
P407 (`language_of_work`)	[X] was written in [Y]	The language of [X] is [Y]	[X] was a [Y]-language work

- **T-REX → PARAREL:**
  - #prompts = 253,488
  - #triples = 27,738
  - #relations = 34 （38， #prompts >3）
  - #prompts/#relations = 8.63

- **BINGREL: new dataset by Bing engine**
  - Distant Supervision 27,738 triples
  - Text length > 10
  - T_1: contain <h, t>, 210,217
  - T_2: only contain <h>, 266,020
  - T_3: random sample

MHA vs. FFN

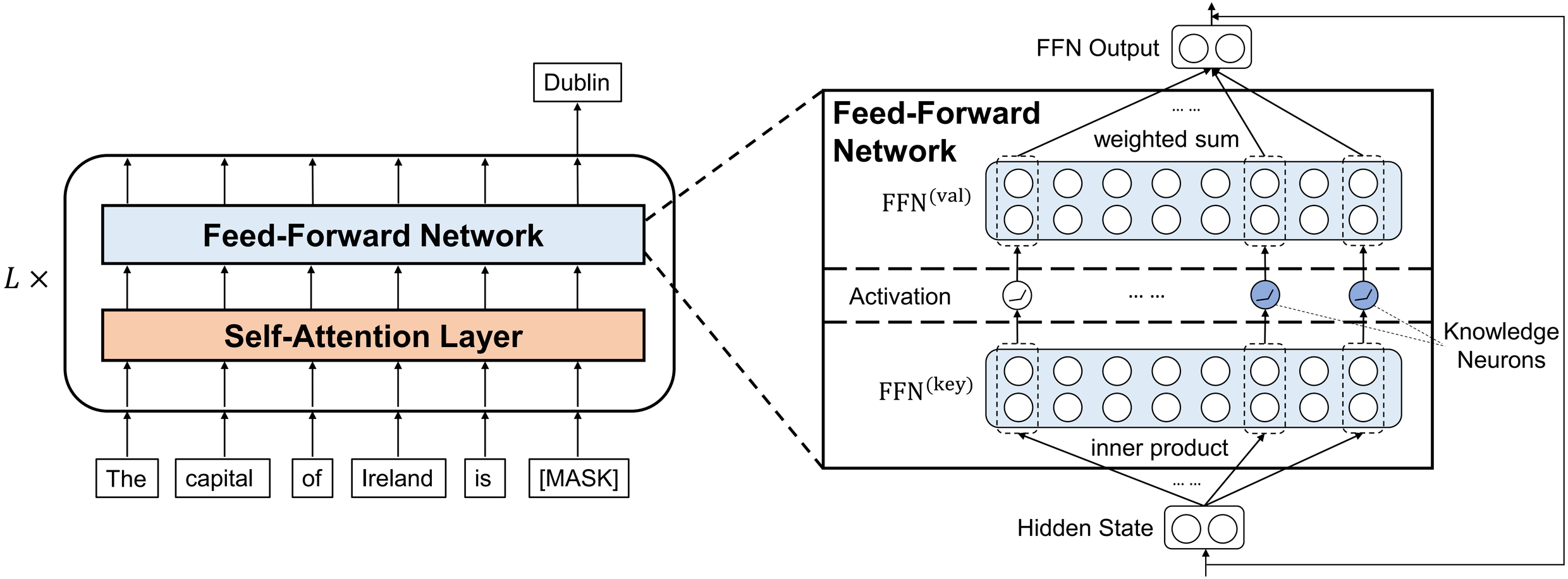


$$Q_h = XW_h^Q, K_h = XW_h^K, V_h = XW_h^V, \quad (1)$$

$$\text{Self-Att}_h(X) = \text{softmax}\left(Q_h K_h^T\right) V_h, \quad (2)$$

$$\text{FFN}(H) = \text{gelu}\left(HW_1\right) W_2, \quad (3)$$

Knowledge Attribution



Knowledge Attribution

the i-th intermediate neuron in the l-th FFN

$$\mathrm{P}_x(\hat{w}_i^{(l)}) = p(y^* | x, w_i^{(l)} = \hat{w}_i^{(l)}), \qquad (4)$$

[MASK]→Gold Constant

$$\mathrm{Attr}(w_i^{(l)}) = \overline{w}_i^{(l)} \int_{\alpha=0}^{1} \frac{\partial \mathrm{P}_x(\alpha \overline{w}_i^{(l)})}{\partial w_i^{(l)}} d\alpha, \qquad (5)$$

$$\tilde{\mathrm{Attr}}(w_i^{(l)}) = \frac{\overline{w}_i^{(l)}}{m} \sum_{k=1}^{m} \frac{\partial \mathrm{P}_x(\frac{k}{m} \overline{w}_i^{(l)})}{\partial w_i^{(l)}}$$

Knowledge Neuron Refining

- **False-positive neurons:** express other information (e.g., syntactic or lexical information)
- **Prompts are diverse enough**

1. produce *n (8.63)* diverse prompts
2. calculate knowledge attribution scores *s (13.3s@V100)*
3. retain attribution score *s* > threshold *t (0.2*MAX)*
4. retain the knowledge neurons shared by more than *p% (70%)* prompts
5. ensure #neurons in *[2, 5]* (70%→75%→80%→… or 70%→65%→60%→…)

Datasets & Baseline

- **T-REX → PARAREL:**
  - #prompts = 253,488
  - #triples = 27,738
  - #relations = 34 （38， #prompts >3）
  - #prompts/#relations = 8.63

- **Activation value baseline:** $\mathrm{Attr}_{\mathrm{base}}(w_i^{(l)}) = \overline{w}_i^{(l)}$
  - activation value = attention score (usually as a strong attribution baseline)

Q1

How many KNs?

Where are KNs?

Do KNs work?

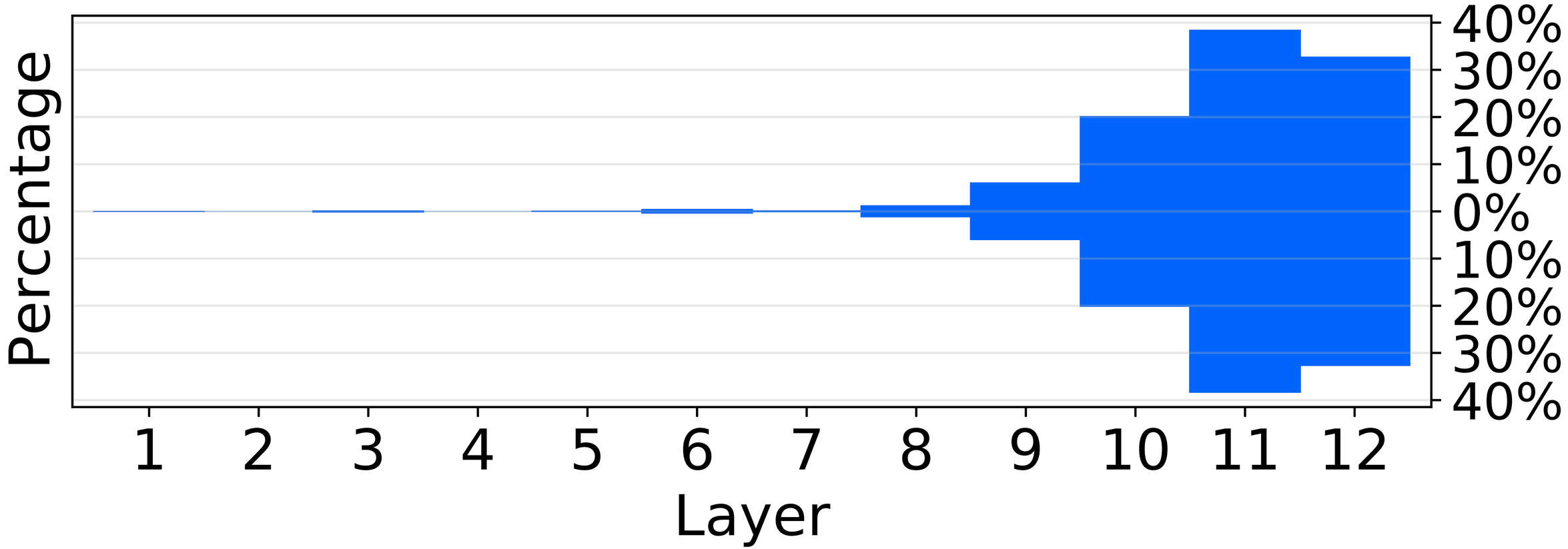
#KNs

Type of Neurons	Ours	Baseline
Knowledge neurons	4.13	3.96
⋂ of intra-rel. fact pairs	1.23	2.85
⋂ of inter-rel. fact pairs	0.09	1.92

- Intra-rel. → Fact pairs with **same** relation
- Inter-rel. → Fact pairs with **different** relation

inter-rel. 👎Baseline; 👍InterGrad

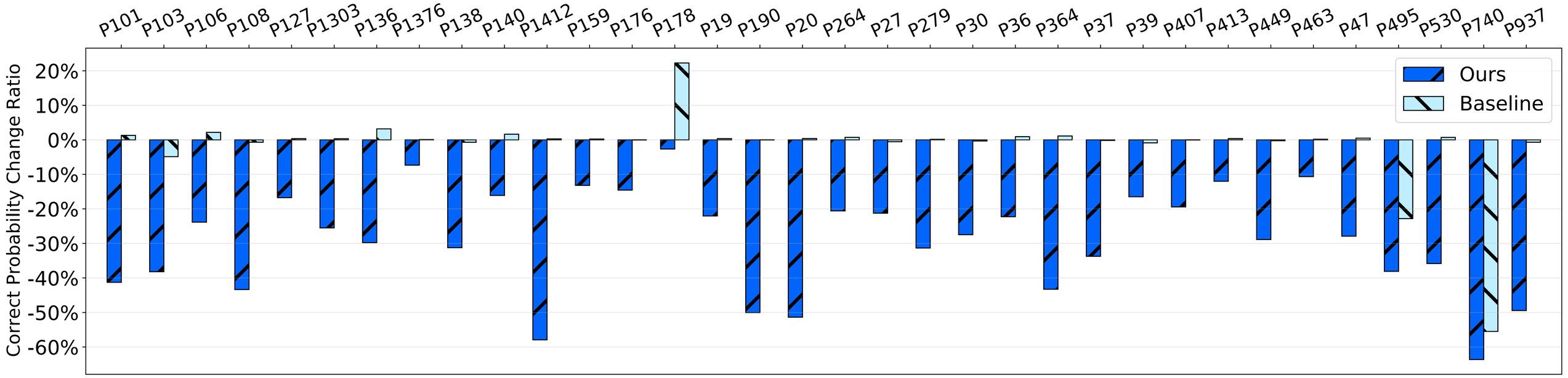
KNs@Layer



KNs @ Layer 9-12

Suppressing KNs

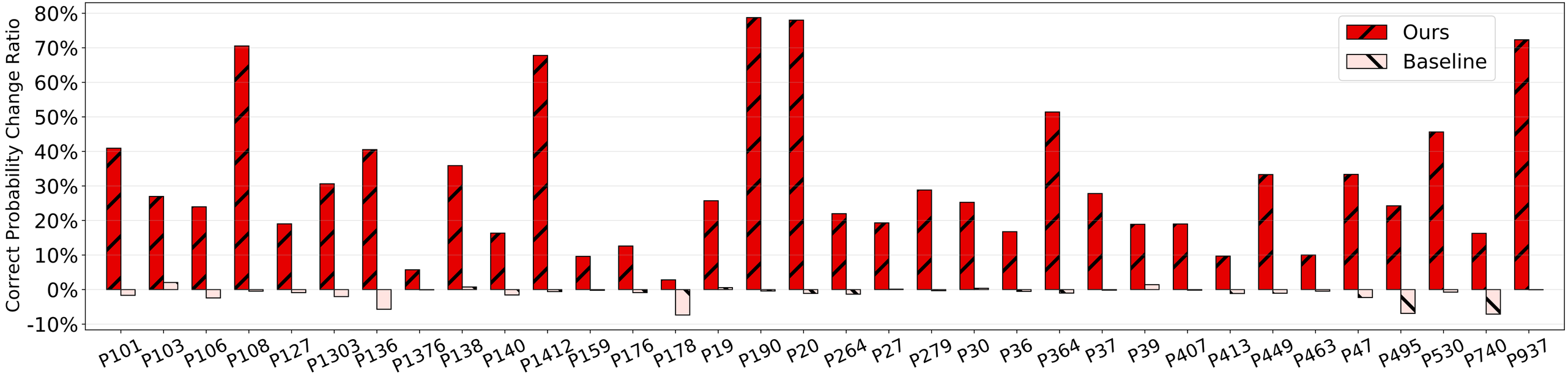
Attribution Score = 0



Suppressing👎Baseline; 👍InterGrad

Amplifying KNs

▶ Attribution Score * 2



Amplifying👎Baseline; 👍InterGrad

Knowledge-Expressing Prompts

- **BINGREL: new dataset by Bing engine**
  - Distant Supervision 27,738 triples
  - Text length > 10
  - T_1: contain <h, t>, 210,217
  - T_2: only contain <h>, 266,020
  - T_3: random sample

Knowledge-Expressing Prompts

Prompt Types	**Ours**	**Baseline**
Containing head and tail ($\mathcal{T}_1$)	0.485	2.472
Containing only head ($\mathcal{T}_2$)	0.019	2.312
Randomly sampled ($\mathcal{T}_3$)	-0.018	2.244

T1 >> T2 > T3

Knowledge-Expressing Prompts

Relational Facts	Neurons		Top-2 and Bottom-2 Activating Prompts (Average Activation)
⟨ Ireland, capital, Dublin ⟩	$w^{(9)}_{2141}, w^{(10)}_{1122}$	Top	Our trip ... in **Dublin**, the capital and largest city of **Ireland** ... (6.36) **Dublin** is the capital and largest city of **Ireland**. (5.77)
		Bottom	**Dublin** just might be the most iconic destination in all of **Ireland**. (1.27) ... in **Ireland**'s famed city, you can enjoy ... **Dublin** experience ... (-0.30)
⟨ Cao_Yunding, place_of_birth, Shanghai ⟩	$w^{(10)}_{739}, w^{(10)}_{1885}, w^{(11)}_{2876}$	Top	**Cao Yunding** was born in **Shanghai** in November 1989. (3.58) Full name: **Cao Yunding** ... Place of birth: **Shanghai**, China ... (2.73)
		Bottom	... **Cao Yunding** (**Shanghai** Shenhua) is shown the red card ... (-0.30) **Shanghai** Shenhua midfielder **Cao Yunding** ... (-0.31)
⟨ Kuwait, continent, Asia ⟩	$w^{(6)}_{147}, w^{(9)}_{866}, w^{(9)}_{1461}, w^{(10)}_{1169}$	Top	**Kuwait** is thus one of the smallest countries in **Asia** ... (6.63) **Kuwait** is a country in Western **Asia** ... (6.27)
		Bottom	This page displays all **Asia** Society content on **Kuwait** ... (-0.48) Noor **Asia** is ... distribution companies in **Kuwait** ... (-0.59)

QII

How to control KNs?

Updating Facts

$$\langle h, r, t \rangle \text{ to } \langle h, r, t' \rangle$$

$$\text{FFN}_{\text{i}}^{(\text{val})} = \text{FFN}_{\text{i}}^{(\text{val})} - \lambda_1 \mathbf{t} + \lambda_2 \mathbf{t}'$$

1 8

Metric	Knowledge Neurons	Random Neurons
Change rate↑	48.5%	4.7%
Success rate↑	34.4%	0.0%
ΔIntra-rel. PPL↓	8.4	10.1
ΔInter-rel. PPL↓	7.2	4.3

👌KNs

↔️Random

Erasing Relations

Erased Relations	Perplexity (Erased Relation)		Perplexity (Other Relations)	
	Before Erasing	After Erasing	Before Erasing	After Erasing
P19 (`place_of_birth`)	1450.0	2996.0 (+106.6%)	120.3	121.6 (+1.1%)
P27 (`country_of_citizenship`)	28.0	38.3 (+36.7%)	143.6	149.5 (+4.2%)
P106 (`occupation`)	2279.0	5202.0 (+128.2%)	120.1	125.3 (+4.3%)
P937 (`work_location`)	58.0	140.0 (+141.2%)	138.0	151.9 (+10.1%)

- Attribution Score = 0

Conclusion

InterGrad✅ Act. value❌

FFN✅ MHA❌

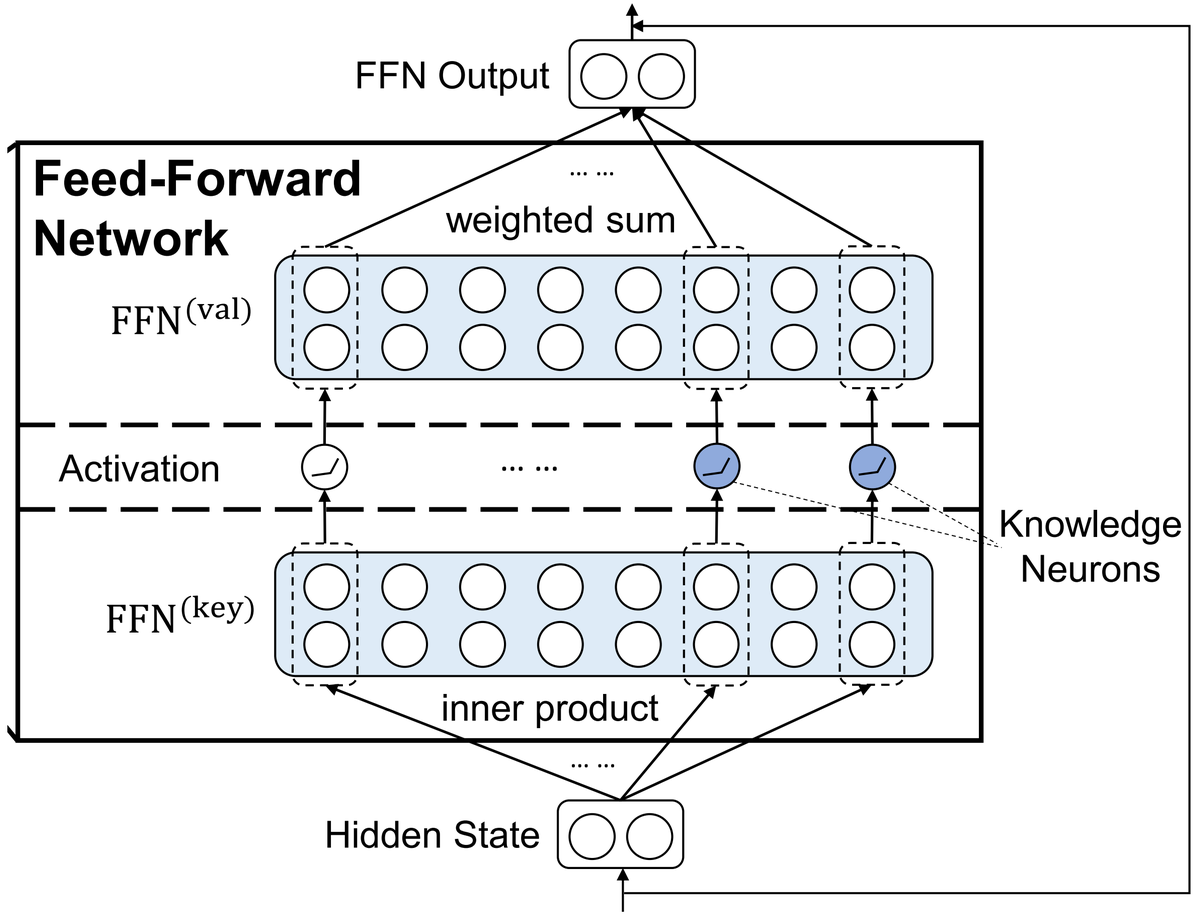
👍Cross-Domain

👍Control

Q

FFN✅ MHA❌ ?

MHA vs. FFN



$$Q_h = XW_h^Q, K_h = XW_h^K, V_h = XW_h^V, \quad (1)$$

$$\text{Self-Att}_h(X) = \text{softmax}\left(Q_h K_h^T\right) V_h, \quad (2)$$

$$\text{FFN}(H) = \text{gelu}\left(HW_1\right) W_2, \quad (3)$$

MHA vs. FFN

- **FFN:** $x_i' = f(x_i ; \Theta)$
- **MHA:** $x_i' = f(x_i, X; \Theta)$
- Maybe LayerNorm…

Kformer

Kformer: Knowledge Injection in Transformer Feed-Forward Layers

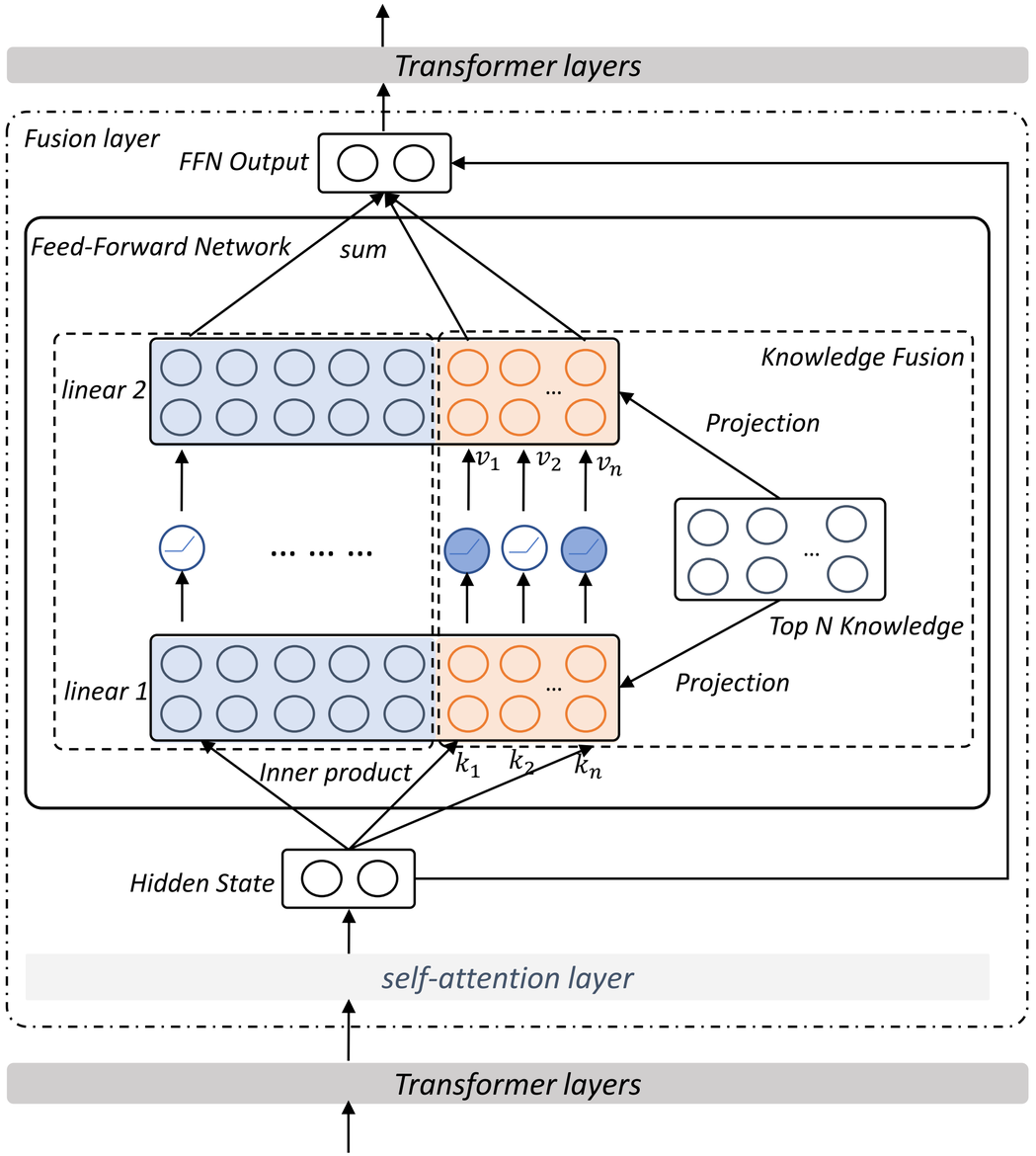
Yunzhi Yao[1,2] , **Shaohan Huang** [3] , **Ningyu Zhang**[1,2] , **Li Dong**[3] , **Furu Wei**[3] , **Huajun Chen**[1,2 *]

[1] Zhejiang University & AZFT Joint Lab for Knowledge Engine

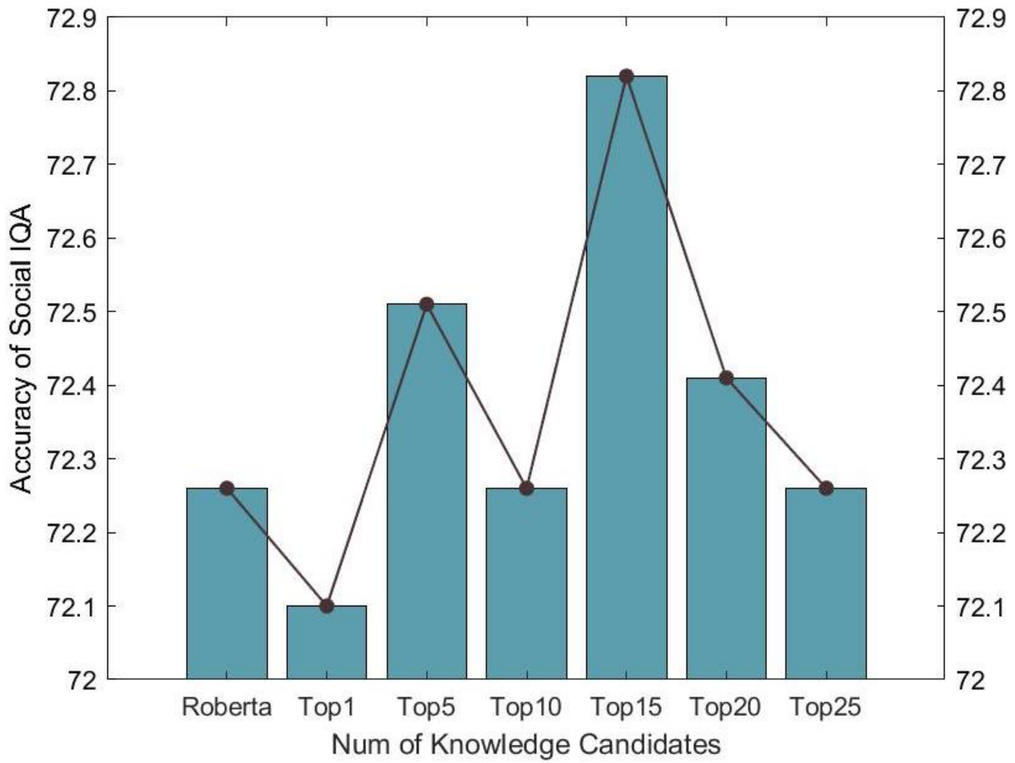
[2] Hangzhou Innovation Center, Zhejiang University

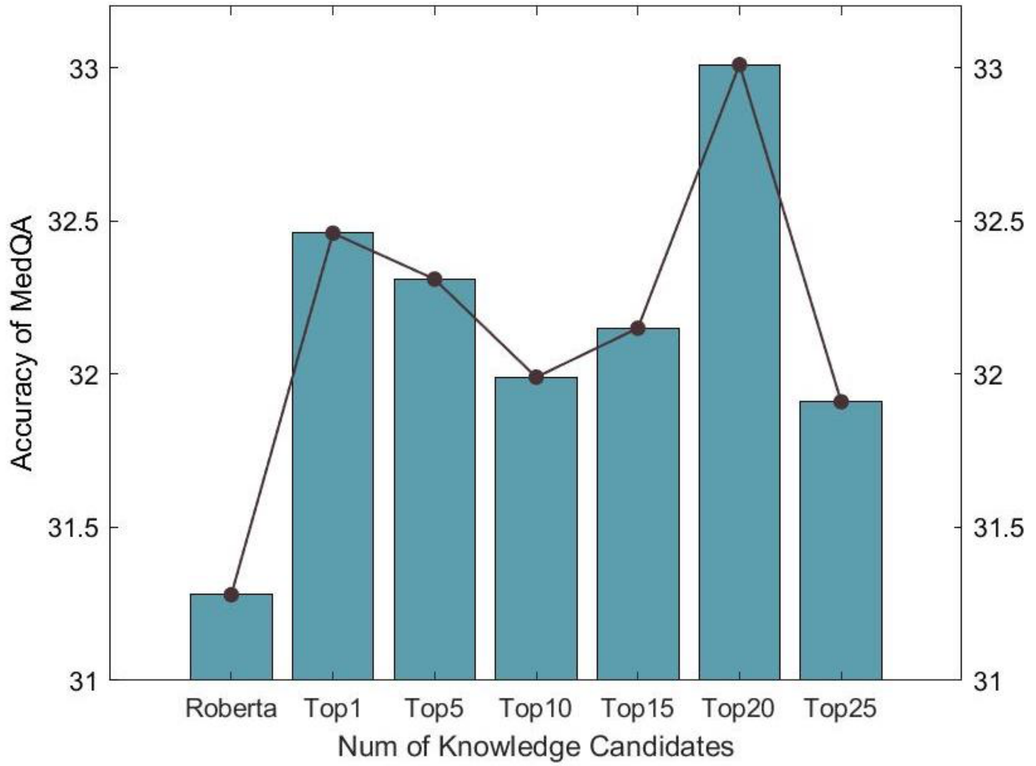
[3] Microsoft Research

Kformer


Transformer layers
Fusion layer
FFN Output
Feed-Forward Network
sum
linear 2
Knowledge Fusion
Projection
v_1 v_2 v_n
... ...
Top N Knowledge
linear 1
Projection
Inner product
k_1 k_2 k_n
Hidden State
self-attention layer
Transformer layers

Kformer


Accuracy of Social IQA
72.9
72.8
72.7
72.6
72.5
72.4
72.3
72.2
72.1
72
Roberta
Top1
Top5
Top10
Top15
Top20
Top25
Num of Knowledge Candidates


Accuracy of MedQA
33
32.5
32
31.5
31
Roberta
Top1
Top5
Top10
Top15
Top20
Top25
Num of Knowledge Candidates

Q & A

MultiEURLEX – A multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer

Ilias Chalkidis [†] [◇] **Manos Fergadiotis** [‡] **Ion Androutsopoulos** [‡]

[†] Department of Computer Science, University of Copenhagen, Denmark

[‡] Department of Informatics, Athens University of Economics and Business, Greece

[◇] Cognitiv+ Ltd., London, United Kingdom

EMNLP21

AntNLP—纪焘

Parameter Efficiency PLMs + CL

Catastrophic Forgetting Mitigation. The language abilities acquired during pre-training are stored in parameters. As a consequence, updating all parameters in PLMs without regularization may lead to catastrophic forgetting when PLMs are sequentially trained across a suite of tasks (Jin et al., 2021; Qin et al., 2021a,c). Since delta tuning only tunes minimal parameters, it could be a potential solution for mitigating the problem of catastrophic forgetting. For instance, MultiEURLEX (Chalkidis et al., 2021) introduce delta tuning into multilingual transfer learning, and demonstrate that using delta tuning methods rather than full-parameter fine-tuning boosts the performance of zero-shot transfer learning between the source language and the target language; Jin et al. (2021) propose to introduce adapters into PLMs and maintain the original parameters fixed, so that PLMs could be trained in a lifelong manner for emerging data.

31

AntNLP—纪焘

Background: Legal NLP

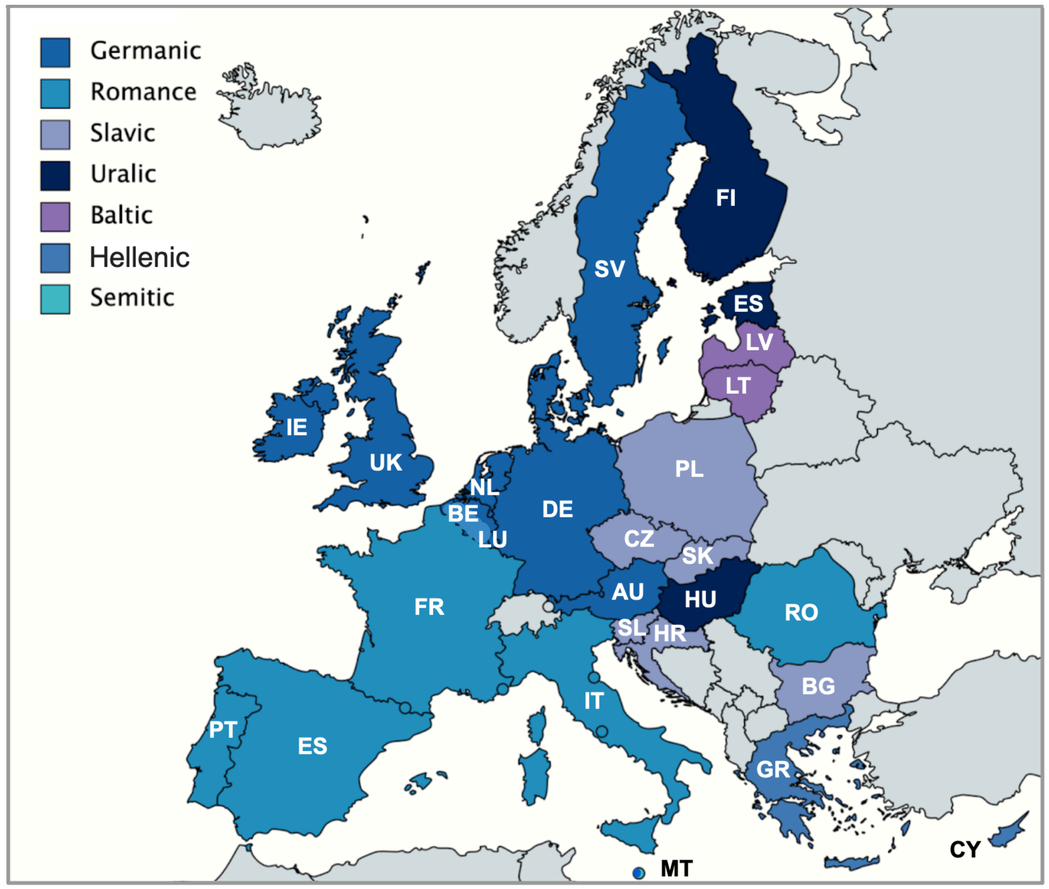


Figure 1: MULTI-EURLEX covers 23 official EU languages (Table 1) from 7 families (illustrated per EU country in the map). The UK was an EU member until 2020. The map should not be taken to imply that no other languages are spoken in EU countries.

Background: EUROVOC

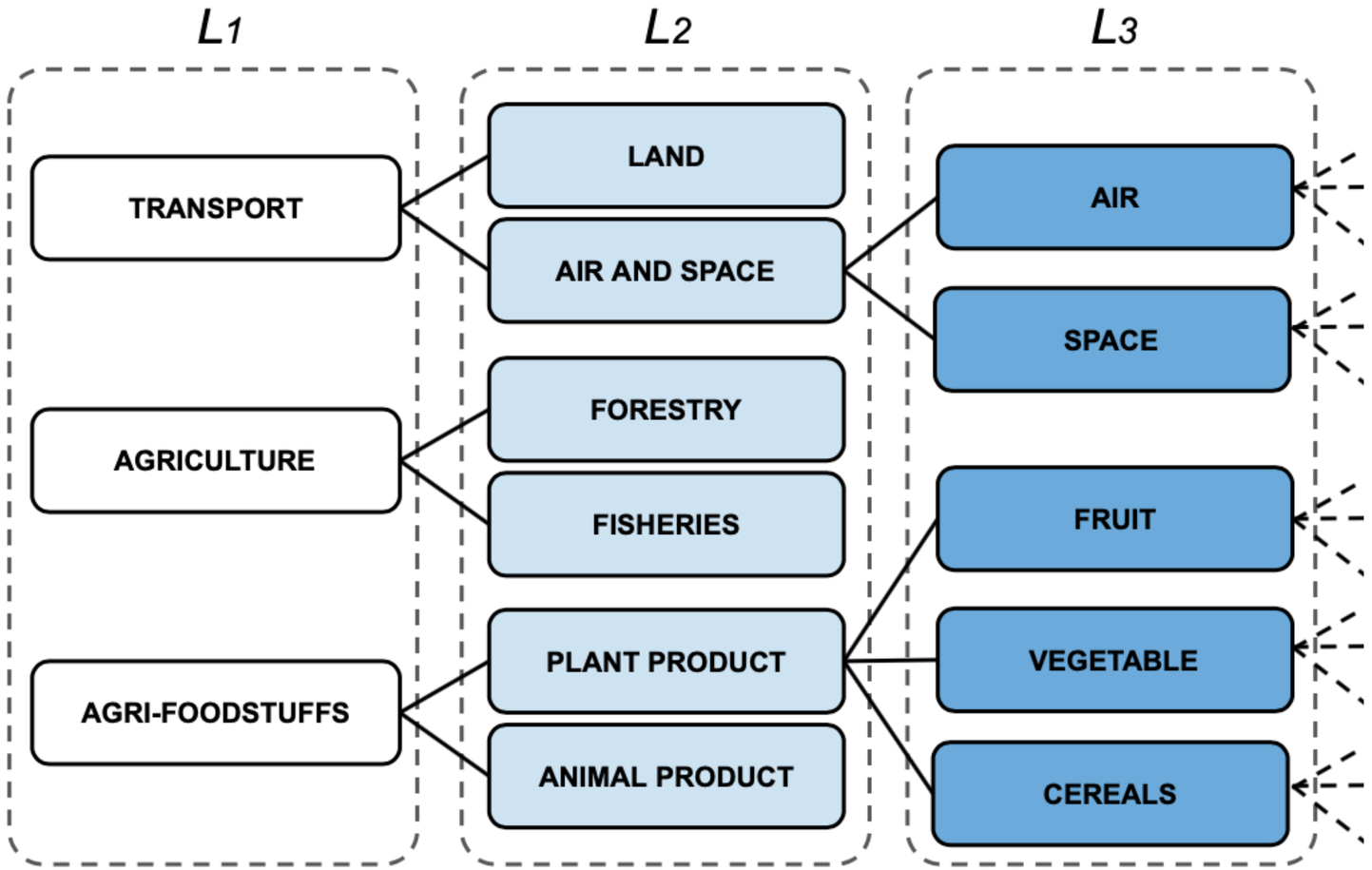


Figure 2: Examples from levels (L_i) 1 to 3 from the EUROVOC hierarchy. More general concepts become more specific as we move from higher to lower levels.

Background: EUROVOC

Label Set	No. of Labels	In training docs		In all docs	
Level 1	21	21	(100%)	21	(100%)
Level 2	127	127	(100%)	127	(100%)
Level 3	567	500	(88%)	511	(90%)
All	7,390	4,220	(57%)	4,591	(62%)

Dataset:MultiEURLEX

mean/median

Language	ISO code	Member Countries where official	EU Speakers (%)		Number of Documents			Words per document
			Native	Total	Train	Dev.	Test	
English	**en**	United Kingdom (1973–2020), Ireland (1973), Malta (2004)	13%	51%	55,000	5,000	5,000	1200 / 460
German	**de**	Germany (1958), Belgium (1958), Luxembourg (1958)	16%	32%	55,000	5,000	5,000	1085 / 410
French	**fr**	France (1958), Belgium(1958), Luxembourg (1958)	12%	26%	55,000	5,000	5,000	1280 / 480
Italian	**it**	Italy (1958)	13%	16%	55,000	5,000	5,000	1210 / 460
Spanish	**es**	Spain (1986)	8%	15%	52,785	5,000	5,000	1380 / 530
Polish	**pl**	Poland (2004)	8%	9%	23,197	5,000	5,000	1200 / 420
Romanian	**ro**	Romania (2007)	5%	5%	15,921	5,000	5,000	1500 / 500
Dutch	**nl**	Netherlands (1958), Belgium (1958)	4%	5%	55,000	5,000	5,000	1230 / 470
Greek	**el**	Greece (1981), Cyprus (2008)	3%	4%	55,000	5,000	5,000	1230 / 470
Hungarian	**hu**	Hungary (2004)	3%	3%	22,664	5,000	5,000	1120 / 370
Portuguese	**pt**	Portugal (1986)	2%	3%	23,188	5,000	5,000	1290 / 500
Czech	**cs**	Czech Republic (2004)	2%	3%	23,187	5,000	5,000	1170 / 410
Swedish	**sv**	Sweden (1995)	2%	3%	42,490	5,000	5,000	1130 / 470
Bulgarian	**bg**	Bulgaria (2007)	2%	2%	15,986	5,000	5,000	1480 / 510
Danish	**da**	Denmark (1973)	1%	1%	55,000	5,000	5,000	1080 / 410
Finnish	**fi**	Finland (1995)	1%	1%	42,497	5,000	5,000	890 / 320
Slovak	**sk**	Slovakia (2004)	1%	1%	15,986	5,000	5,000	1180 / 410
Lithuanian	**lt**	Lithuania (2004)	1%	1%	23,188	5,000	5,000	1070 / 370
Croatian	**hr**	Croatia (2013)	1%	1%	7,944	2,500	5,000	1490 / 500
Slovene	**sl**	Slovenia (2004)	<1%	<1%	23,184	5,000	5,000	1170 / 400
Estonian	**et**	Estonia (2004)	<1%	<1%	23,126	5,000	5,000	950 / 330
Latvian	**lv**	Latvia (2004)	<1%	<1%	23,188	5,000	5,000	1080 / 380
Maltese	**mt**	Malta (2004)	<1%	<1%	17,521	5,000	5,000	1250 / 430

Temporal Concept Drift

- KL of label distributions

Label Set	Random		Chronological	
	train-dev	*train-test*	*train-dev*	*train-test*
Level 1	0.00	0.00	0.03	0.04
Level 2	0.00	0.00	0.12	0.16
Level 3	0.01	0.01	0.21	0.32
All	0.20	0.20	1.09	1.67

- training (55k, 1958– 2010)
- development (5k, 2010–2012)
- test (5k, 2012–2016)

Temporal Concept Drift

BERT-EN-FT

Data Split	Training	Development	Test
Random	**99.2**	**74.7**	**74.0**
Chronological	96.7	58.7	48.4

random split over-estimates real performance

Setup

- **MPLMs:**
  - BERT*23
  - XLM-RoBERTa
  - mT5-Encoder

We feed the top-level hidden state of the `[cls]` token ($\in \mathbb{R}^{D_h}$) to a dense layer ($W_{\texttt{[cls]}} \in \mathbb{R}^{D_h \times L}$) with L outputs and sigmoids.

- **Training Setting:**
  - *one2one*
  - *one2many*
  - *many2many*
- Evaluation: report *mean R-Precision* (mRP)

Setup

- **Parameter Efficiency Methods:**
  - **Frozen Layers:** 3,6,9,12
  - **Adapter modules:** $D_h \rightarrow D_{down} \rightarrow D_h$
  - **BitFit**: only FT bias terms (0.09%) Zaken et al., 2021
  - **LNFit**: only FT LayerNorm (Frankle et al., 2021 BNFit)

Results

	GERMANIC					ROMANCE					SLAVIC			URALIC			
	en	**da**	**de**	**nl**	**sv**	**ro**	**es**	**fr**	**it**	**pt**	**pl**	**bg**	**cs**	**hu**	**fi**	**el**	**All**
One-to-one (Fine-tune XLM-ROBERTA or monolingually pretrained BERTs in one language, test in the *same* language.)																	
NATIVE-BERT	**67.7**	65.5	**68.4**	66.7	**68.5**	**68.5**	67.6	**67.4**	**67.9**	**67.4**	**67.2**	-	**66.7**	**67.7**	**67.8**	**67.8**	**67.4**
XLM-ROBERTA	67.4	**66.7**	67.5	**67.3**	66.5	66.4	**67.8**	67.2	67.4	67.0	65.0	66.1	**66.7**	65.5	66.5	65.8	66.6
Diff.	-0.3	+1.2	-0.9	+0.6	-2.0	-2.1	+0.2	-0.2	-0.5	-0.4	-2.2	-	0.0	-2.2	-1.3	-2.0	-0.7
One-to-many (Fine-tune XLM-ROBERTA *only* in English, test in all languages, with alternative adaptation strategies.)																	
End-to-end fine-tuning	**67.4**	56.5	52.4	49.0	55.7	55.2	54.0	55.0	52.0	50.5	46.9	51.2	49.6	48.8	46.4	33.3	49.3
First 3 blocks frozen	66.3	59.1	56.8	55.3	57.5	57.9	58.1	57.7	56.2	54.9	53.7	56.1	54.3	51.0	52.1	42.4	53.0
First 6 blocks frozen	66.3	59.1	57.4	55.7	57.9	57.2	56.9	57.9	53.9	55.4	51.9	55.8	52.6	47.3	48.7	39.6	51.7
First 9 blocks frozen	65.8	59.4	57.9	56.9	58.6	58.2	58.7	59.4	55.7	57.5	53.4	56.7	54.2	48.8	50.4	44.5	53.0
All 12 blocks frozen	27.2	21.4	24.6	24.6	23.0	21.6	23.4	21.9	20.1	25.1	22.8	23.1	24.3	22.8	21.9	19.0	22.2
Adapter modules	67.3	**61.5**	**59.3**	**57.8**	**59.5**	**60.3**	**61.0**	**60.4**	**58.8**	**58.5**	**57.5**	**59.2**	**56.8**	**55.3**	**55.6**	**46.1**	**56.1**
BITFIT (bias terms only)	63.9	59.3	57.0	54.0	58.2	57.8	57.4	56.9	56.4	55.5	54.0	55.6	54.8	51.2	54.8	42.1	53.7
LNFIT (layer-norm only)	63.1	58.9	55.7	54.1	56.6	59.1	59.1	58.0	56.6	57.2	55.7	55.4	52.8	51.4	50.7	39.9	53.3
Many-to-many (Jointly fine-tune XLM-ROBERTA in *all* languages, test in all languages, with alternative adaptation strategies.)																	
End-to-end fine-tuning	66.4	66.2	66.2	66.1	66.1	66.3	66.3	66.2	66.3	65.9	65.6	65.7	65.7	65.2	65.8	65.1	65.7
Adapter modules	**67.2**	**67.1**	**66.3**	**67.1**	**67.0**	**67.4**	**67.2**	**67.1**	**67.4**	**67.0**	**66.2**	**66.6**	**67.0**	**65.5**	**66.6**	**65.7**	**66.4**

1 multi-BERT is competitive to 16 mono-BERT

Results

	GERMANIC					ROMANCE					SLAVIC			URALIC			
	en	**da**	**de**	**nl**	**sv**	**ro**	**es**	**fr**	**it**	**pt**	**pl**	**bg**	**cs**	**hu**	**fi**	**el**	**All**
One-to-one (Fine-tune XLM-ROBERTA or monolingually pretrained BERTs in one language, test in the *same* language.)																	
NATIVE-BERT	**67.7**	65.5	**68.4**	66.7	**68.5**	**68.5**	67.6	**67.4**	**67.9**	**67.4**	**67.2**	-	**66.7**	**67.7**	**67.8**	**67.8**	**67.4**
XLM-ROBERTA	67.4	**66.7**	67.5	**67.3**	66.5	66.4	**67.8**	67.2	67.4	67.0	65.0	66.1	**66.7**	65.5	66.5	65.8	66.6
Diff.	-0.3	+1.2	-0.9	+0.6	-2.0	-2.1	+0.2	-0.2	-0.5	-0.4	-2.2	-	0.0	-2.2	-1.3	-2.0	-0.7
One-to-many (Fine-tune XLM-ROBERTA *only* in English, test in all languages, with alternative adaptation strategies.)																	
End-to-end fine-tuning	**67.4**	56.5	52.4	49.0	55.7	55.2	54.0	55.0	52.0	50.5	46.9	51.2	49.6	48.8	46.4	33.3	49.3
First 3 blocks frozen	66.3	59.1	56.8	55.3	57.5	57.9	58.1	57.7	56.2	54.9	53.7	56.1	54.3	51.0	52.1	42.4	53.0
First 6 blocks frozen	66.3	59.1	57.4	55.7	57.9	57.2	56.9	57.9	53.9	55.4	51.9	55.8	52.6	47.3	48.7	39.6	51.7
First 9 blocks frozen	65.8	59.4	57.9	56.9	58.6	58.2	58.7	59.4	55.7	57.5	53.4	56.7	54.2	48.8	50.4	44.5	53.0
All 12 blocks frozen	27.2	21.4	24.6	24.6	23.0	21.6	23.4	21.9	20.1	25.1	22.8	23.1	24.3	22.8	21.9	19.0	22.2
Adapter modules	67.3	**61.5**	**59.3**	**57.8**	**59.5**	**60.3**	**61.0**	**60.4**	**58.8**	**58.5**	**57.5**	**59.2**	**56.8**	**55.3**	**55.6**	**46.1**	**56.1**
BITFIT (bias terms only)	63.9	59.3	57.0	54.0	58.2	57.8	57.4	56.9	56.4	55.5	54.0	55.6	54.8	51.2	54.8	42.1	53.7
LNFIT (layer-norm only)	63.1	58.9	55.7	54.1	56.6	59.1	59.1	58.0	56.6	57.2	55.7	55.4	52.8	51.4	50.7	39.9	53.3
Many-to-many (Jointly fine-tune XLM-ROBERTA in *all* languages, test in all languages, with alternative adaptation strategies.)																	
End-to-end fine-tuning	66.4	66.2	66.2	66.1	66.1	66.3	66.3	66.2	66.3	65.9	65.6	65.7	65.7	65.2	65.8	65.1	65.7
Adapter modules	**67.2**	**67.1**	**66.3**	**67.1**	**67.0**	**67.4**	**67.2**	**67.1**	**67.4**	**67.0**	**66.2**	**66.6**	**67.0**	**65.5**	**66.6**	**65.7**	**66.4**

Small diff.@Frozen Large⬇️@AllFrozen

Results

		GERMANIC				ROMANCE					SLAVIC			URALIC			
	en	**da**	**de**	**nl**	**sv**	**ro**	**es**	**fr**	**it**	**pt**	**pl**	**bg**	**cs**	**hu**	**fi**	**el**	**All**
One-to-one (Fine-tune XLM-ROBERTA or monolingually pretrained BERTs in one language, test in the *same* language.)																	
NATIVE-BERT	**67.7**	65.5	**68.4**	66.7	**68.5**	**68.5**	67.6	**67.4**	**67.9**	**67.4**	**67.2**	-	**66.7**	**67.7**	**67.8**	**67.8**	**67.4**
XLM-ROBERTA	67.4	**66.7**	67.5	**67.3**	66.5	66.4	**67.8**	67.2	67.4	67.0	65.0	66.1	**66.7**	65.5	66.5	65.8	66.6
Diff.	-0.3	+1.2	-0.9	+0.6	-2.0	-2.1	+0.2	-0.2	-0.5	-0.4	-2.2	-	0.0	-2.2	-1.3	-2.0	-0.7
One-to-many (Fine-tune XLM-ROBERTA *only* in English, test in all languages, with alternative adaptation strategies.)																	
End-to-end fine-tuning	**67.4**	56.5	52.4	49.0	55.7	55.2	54.0	55.0	52.0	50.5	46.9	51.2	49.6	48.8	46.4	33.3	49.3
First 3 blocks frozen	66.3	59.1	56.8	55.3	57.5	57.9	58.1	57.7	56.2	54.9	53.7	56.1	54.3	51.0	52.1	42.4	53.0
First 6 blocks frozen	66.3	59.1	57.4	55.7	57.9	57.2	56.9	57.9	53.9	55.4	51.9	55.8	52.6	47.3	48.7	39.6	51.7
First 9 blocks frozen	65.8	59.4	57.9	56.9	58.6	58.2	58.7	59.4	55.7	57.5	53.4	56.7	54.2	48.8	50.4	44.5	53.0
All 12 blocks frozen	27.2	21.4	24.6	24.6	23.0	21.6	23.4	21.9	20.1	25.1	22.8	23.1	24.3	22.8	21.9	19.0	22.2
Adapter modules	67.3	**61.5**	**59.3**	**57.8**	**59.5**	**60.3**	**61.0**	**60.4**	**58.8**	**58.5**	**57.5**	**59.2**	**56.8**	**55.3**	**55.6**	**46.1**	**56.1**
BITFIT (bias terms only)	63.9	59.3	57.0	54.0	58.2	57.8	57.4	56.9	56.4	55.5	54.0	55.6	54.8	51.2	54.8	42.1	53.7
LNFIT (layer-norm only)	63.1	58.9	55.7	54.1	56.6	59.1	59.1	58.0	56.6	57.2	55.7	55.4	52.8	51.4	50.7	39.9	53.3
Many-to-many (Jointly fine-tune XLM-ROBERTA in *all* languages, test in all languages, with alternative adaptation strategies.)																	
End-to-end fine-tuning	66.4	66.2	66.2	66.1	66.1	66.3	66.3	66.2	66.3	65.9	65.6	65.7	65.7	65.2	65.8	65.1	65.7
Adapter modules	**67.2**	**67.1**	**66.3**	**67.1**	**67.0**	**67.4**	**67.2**	**67.1**	**67.4**	**67.0**	**66.2**	**66.6**	**67.0**	**65.5**	**66.6**	**65.7**	**66.4**

🥇 Adapter 🥈 BITFIT 🥉 LNFIT

Results

	GERMANIC					ROMANCE					SLAVIC			URALIC			
	en	**da**	**de**	**nl**	**sv**	**ro**	**es**	**fr**	**it**	**pt**	**pl**	**bg**	**cs**	**hu**	**fi**	**el**	**All**
One-to-one (Fine-tune XLM-ROBERTA or monolingually pretrained BERTs in one language, test in the *same* language.)																	
NATIVE-BERT	**67.7**	65.5	**68.4**	66.7	**68.5**	**68.5**	67.6	**67.4**	**67.9**	**67.4**	**67.2**	-	**66.7**	**67.7**	**67.8**	**67.8**	**67.4**
XLM-ROBERTA	67.4	**66.7**	67.5	**67.3**	66.5	66.4	**67.8**	67.2	67.4	67.0	65.0	66.1	**66.7**	65.5	66.5	65.8	66.6
Diff.	-0.3	+1.2	-0.9	+0.6	-2.0	-2.1	+0.2	-0.2	-0.5	-0.4	-2.2	-	0.0	-2.2	-1.3	-2.0	-0.7
One-to-many (Fine-tune XLM-ROBERTA *only* in English, test in all languages, with alternative adaptation strategies.)																	
End-to-end fine-tuning	**67.4**	56.5	52.4	49.0	55.7	55.2	54.0	55.0	52.0	50.5	46.9	51.2	49.6	48.8	46.4	33.3	49.3
First 3 blocks frozen	66.3	59.1	56.8	55.3	57.5	57.9	58.1	57.7	56.2	54.9	53.7	56.1	54.3	51.0	52.1	42.4	53.0
First 6 blocks frozen	66.3	59.1	57.4	55.7	57.9	57.2	56.9	57.9	53.9	55.4	51.9	55.8	52.6	47.3	48.7	39.6	51.7
First 9 blocks frozen	65.8	59.4	57.9	56.9	58.6	58.2	58.7	59.4	55.7	57.5	53.4	56.7	54.2	48.8	50.4	44.5	53.0
All 12 blocks frozen	27.2	21.4	24.6	24.6	23.0	21.6	23.4	21.9	20.1	25.1	22.8	23.1	24.3	22.8	21.9	19.0	22.2
Adapter modules	67.3	**61.5**	**59.3**	**57.8**	**59.5**	**60.3**	**61.0**	**60.4**	**58.8**	**58.5**	**57.5**	**59.2**	**56.8**	**55.3**	**55.6**	**46.1**	**56.1**
BITFIT (bias terms only)	63.9	59.3	57.0	54.0	58.2	57.8	57.4	56.9	56.4	55.5	54.0	55.6	54.8	51.2	54.8	42.1	53.7
LNFIT (layer-norm only)	63.1	58.9	55.7	54.1	56.6	59.1	59.1	58.0	56.6	57.2	55.7	55.4	52.8	51.4	50.7	39.9	53.3
Many-to-many (Jointly fine-tune XLM-ROBERTA in *all* languages, test in all languages, with alternative adaptation strategies.)																	
End-to-end fine-tuning	66.4	66.2	66.2	66.1	66.1	66.3	66.3	66.2	66.3	65.9	65.6	65.7	65.7	65.2	65.8	65.1	65.7
Adapter modules	**67.2**	**67.1**	**66.3**	**67.1**	**67.0**	**67.4**	**67.2**	**67.1**	**67.4**	**67.0**	**66.2**	**66.6**	**67.0**	**65.5**	**66.6**	**65.7**	**66.4**

🏅 Adapter (but param. Unfair)

Results

	GERMANIC					ROMANCE					SLAVIC			URALIC			
	en	**da**	**de**	**nl**	**sv**	**ro**	**es**	**fr**	**it**	**pt**	**pl**	**bg**	**cs**	**hu**	**fi**	**el**	**All**
One-to-one (Fine-tune XLM-ROBERTA or monolingually pretrained BERTs in one language, test in the *same* language.)																	
NATIVE-BERT	**67.7**	65.5	**68.4**	66.7	**68.5**	**68.5**	67.6	**67.4**	**67.9**	**67.4**	**67.2**	-	**66.7**	**67.7**	**67.8**	**67.8**	**67.4**
XLM-ROBERTA	67.4	**66.7**	67.5	**67.3**	66.5	66.4	**67.8**	67.2	67.4	67.0	65.0	66.1	**66.7**	65.5	66.5	65.8	66.6
Diff.	-0.3	+1.2	-0.9	+0.6	-2.0	-2.1	+0.2	-0.2	-0.5	-0.4	-2.2	-	0.0	-2.2	-1.3	-2.0	-0.7
One-to-many (Fine-tune XLM-ROBERTA *only* in English, test in all languages, with alternative adaptation strategies.)																	
End-to-end fine-tuning	**67.4**	56.5	52.4	49.0	55.7	55.2	54.0	55.0	52.0	50.5	46.9	51.2	49.6	48.8	46.4	33.3	49.3
First 3 blocks frozen	66.3	59.1	56.8	55.3	57.5	57.9	58.1	57.7	56.2	54.9	53.7	56.1	54.3	51.0	52.1	42.4	53.0
First 6 blocks frozen	66.3	59.1	57.4	55.7	57.9	57.2	56.9	57.9	53.9	55.4	51.9	55.8	52.6	47.3	48.7	39.6	51.7
First 9 blocks frozen	65.8	59.4	57.9	56.9	58.6	58.2	58.7	59.4	55.7	57.5	53.4	56.7	54.2	48.8	50.4	44.5	53.0
All 12 blocks frozen	27.2	21.4	24.6	24.6	23.0	21.6	23.4	21.9	20.1	25.1	22.8	23.1	24.3	22.8	21.9	19.0	22.2
Adapter modules	67.3	**61.5**	**59.3**	**57.8**	**59.5**	**60.3**	**61.0**	**60.4**	**58.8**	**58.5**	**57.5**	**59.2**	**56.8**	**55.3**	**55.6**	**46.1**	**56.1**
BITFIT (bias terms only)	63.9	59.3	57.0	54.0	58.2	57.8	57.4	56.9	56.4	55.5	54.0	55.6	54.8	51.2	54.8	42.1	53.7
LNFIT (layer-norm only)	63.1	58.9	55.7	54.1	56.6	59.1	59.1	58.0	56.6	57.2	55.7	55.4	52.8	51.4	50.7	39.9	53.3
Many-to-many (Jointly fine-tune XLM-ROBERTA in *all* languages, test in all languages, with alternative adaptation strategies.)																	
End-to-end fine-tuning	66.4	66.2	66.2	66.1	66.1	66.3	66.3	66.2	66.3	65.9	65.6	65.7	65.7	65.2	65.8	65.1	65.7
Adapter modules	**67.2**	**67.1**	**66.3**	**67.1**	**67.0**	**67.4**	**67.2**	**67.1**	**67.4**	**67.0**	**66.2**	**66.6**	**67.0**	**65.5**	**66.6**	**65.7**	**66.4**

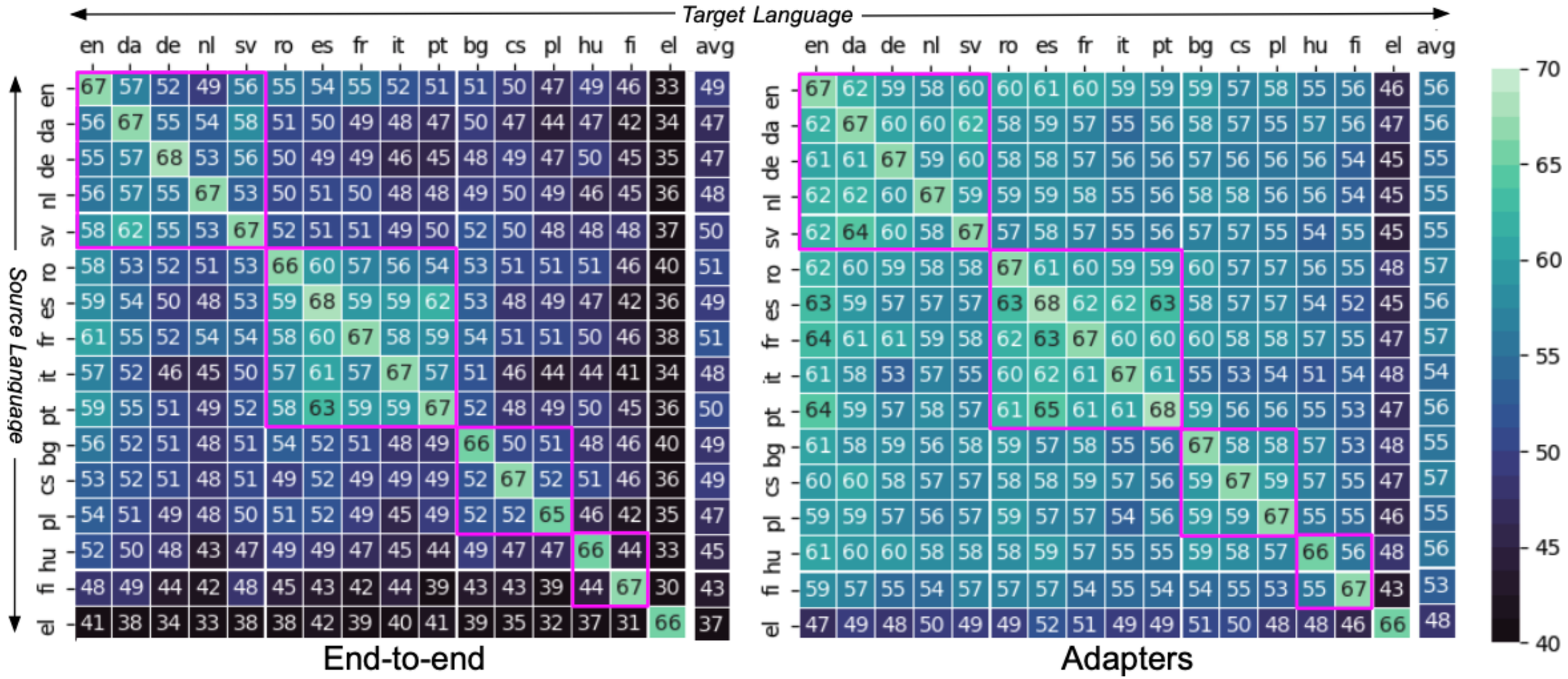
CF😈Multilingual Knowledge

mT5 Results

Adaptation strategy	Params (%)		en (Src)	All
End-to-end fine-tuning	277M	(100.0%)	67.4	53.7
First 3 blocks frozen	63.7M	(23.0%)	67.4	56.9
First 6 blocks frozen	42.4M	(15.3%)	66.3	**58.4**
First 9 blocks frozen	21.2M	(7.7%)	**68.0**	58.3
All 12 blocks frozen	–	(0.0%)	20.2	16.8
Adapter modules	7.1M	(1.7%)	66.3	44.0
LNFIT (layer-norm only)	19.2K	(0.01%)	59.5	38.7

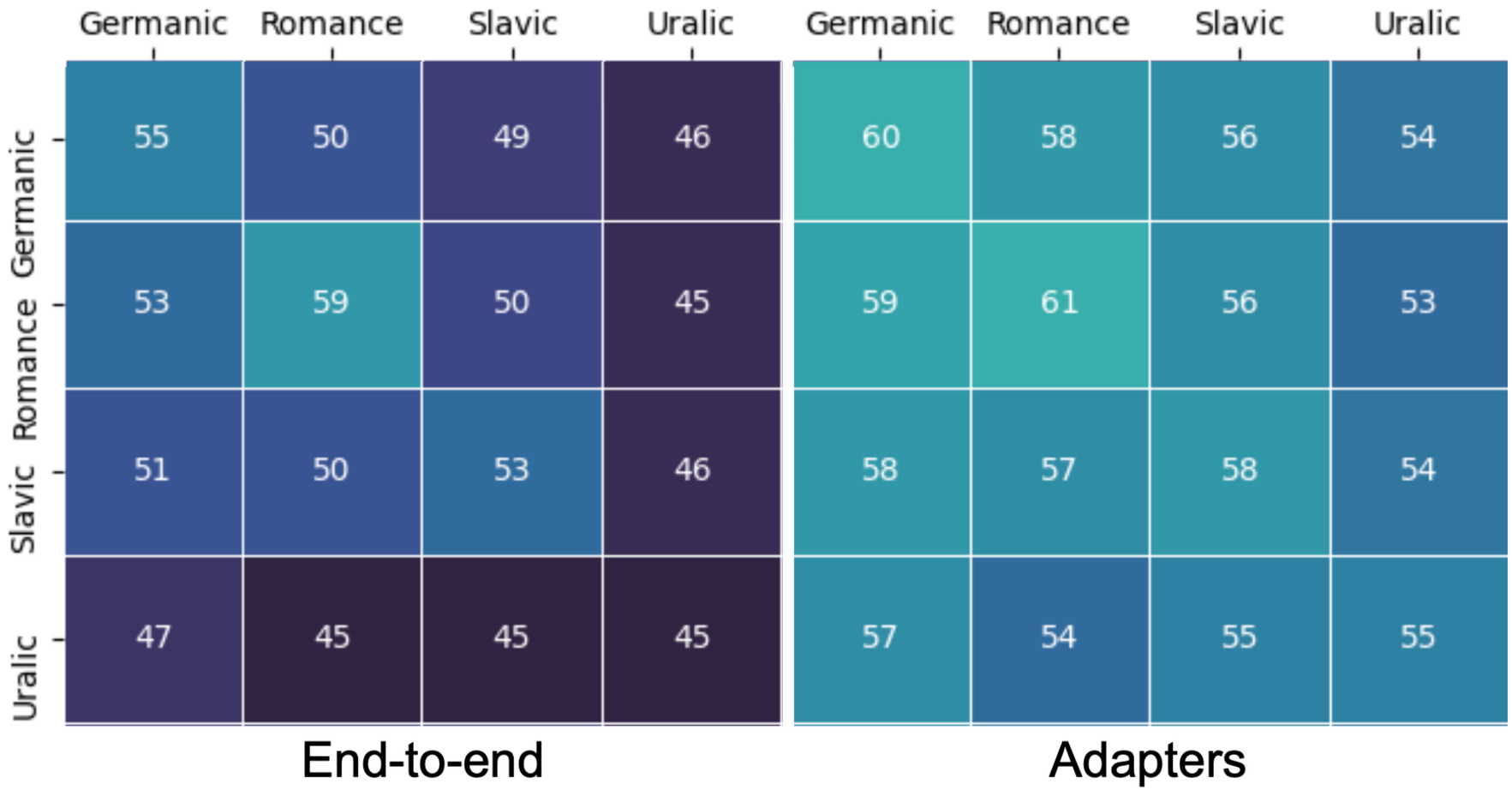
🥇 Frozen 🥈 Adapter 🥉 LNFIT

Cross-lingual Results



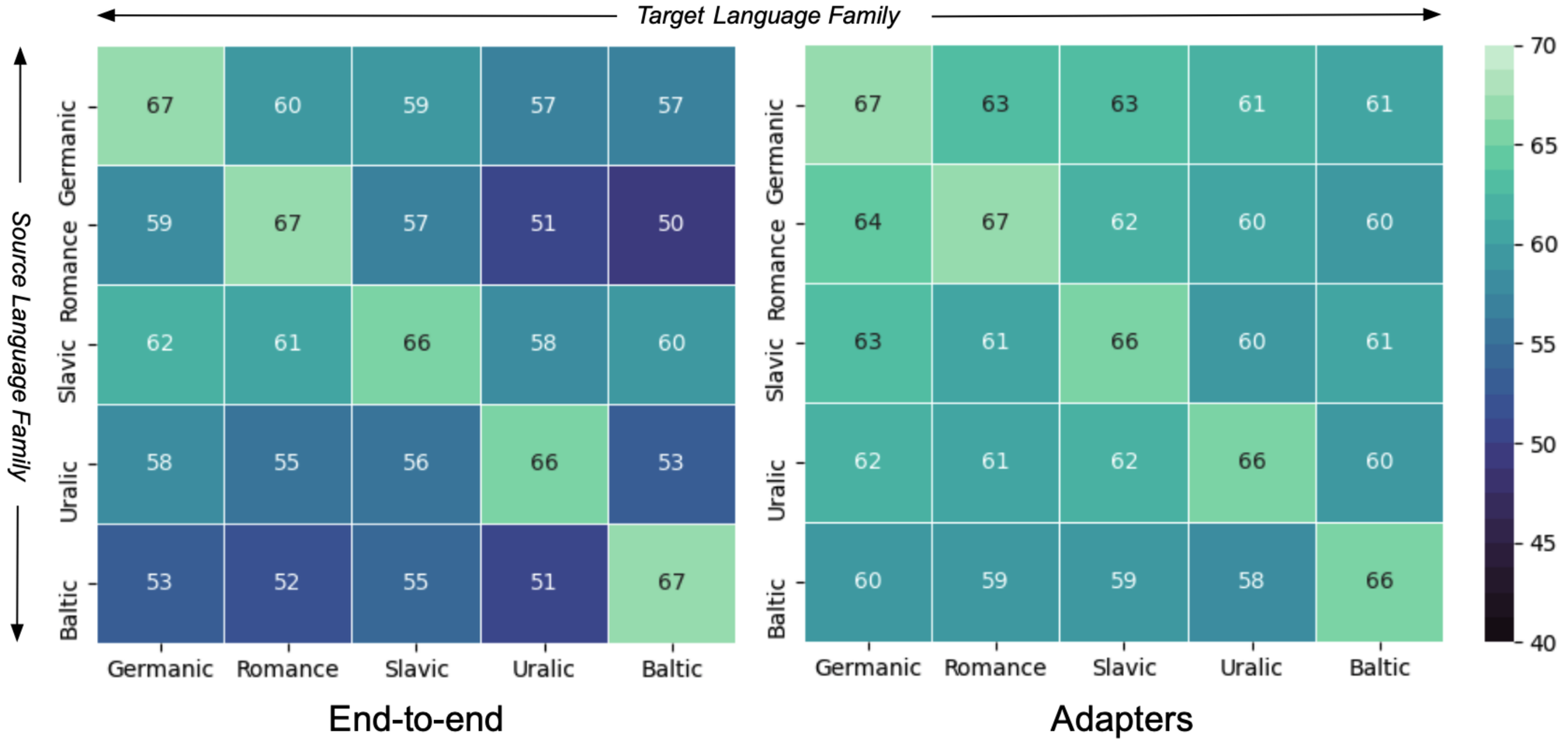
EN not always best; 👍families; 🥇 Adapter

Cross-lingual Results



👍Romance families

Cross-lingual Results



👍 families

Cross-lingual Vocab.

One2one One2many

Version of the input text	**en** (Src)		**Rest**	
	T	mRP	T (%)	mRP
Full-text	100%	**67.3**	100%	**56.1**
w/o digits	89%	67.1	88%	55.0
w/o digits & English vocab.	22%	14.0	77%	51.5

Count >25

👍 Shared vocab.

Level Results

Adaptation Strategy	**Parameters**		Level 1 (21)		Level 2 (127)		Level 3 (567)		Original (7,390)	
			en (Src)	**All**	**en** (Src)	**All**	**en** (Src)	**All**	**en** (Src)	**All**
End-to-end fine-tuning	278M	(100%)	**83.2**	75.7	**73.6**	58.7	**67.4**	49.3	47.6	27.6
First 3 blocks frozen	63.8M	(23.0%)	82.9	76.4	71.3	60.2	66.3	53.0	47.3	29.0
First 6 blocks frozen	42.5M	(15.3%)	82.3	76.7	69.6	61.1	66.3	51.7	47.1	30.1
First 9 blocks frozen	21.3M	(7.7%)	82.0	74.8	70.7	60.1	65.8	53.0	48.0	32.8
Adapter modules	9.5M	(3.3%)	83.1	**77.2**	72.3	**61.2**	67.3	**56.1**	47.9	**35.1**
BITFIT (bias terms only)	101K	(0.04%)	82.7	76.1	70.2	60.1	63.9	53.7	**48.3**	33.9
LNFIT (layer-norm only)	36.8K	(0.01%)	81.5	74.9	69.7	59.3	63.1	53.3	43.1	26.4
↑ ***Averaged Adapt.*** ↑	-		82.4	76.0	70.6	60.3	65.5	53.5	47.0	31.2
All 12 blocks frozen	-	(0.0%)	61.4	56.5	39.0	31.6	27.2	22.2	26.1	15.3

#Label😈MRP 🎖️Adapter

Conclusion

Temporal Concept Drift

CF😈Multilingual Knowledge

Parameter Efficiency🏥CF

Q & A

Lifelong Pretraining: Continually Adapting Language Models to Emerging Corpora

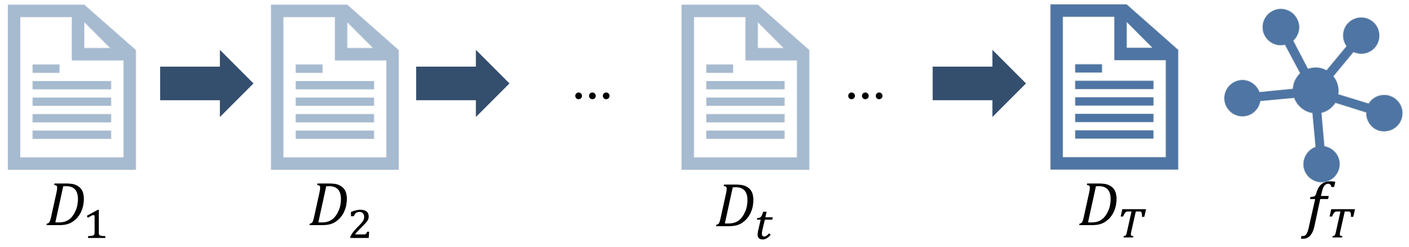
Xisen Jin[†1] **Dejiao Zhang**[2] **Henghui Zhu**[2] **Wei Xiao**[2]
Shang-Wen Li[‡2] **Xiaokai Wei**[2] **Andrew Arnold**[2] **Xiang Ren**[1]
[1]University of Southern California [2]Amazon Inc.

ACL22-Reject

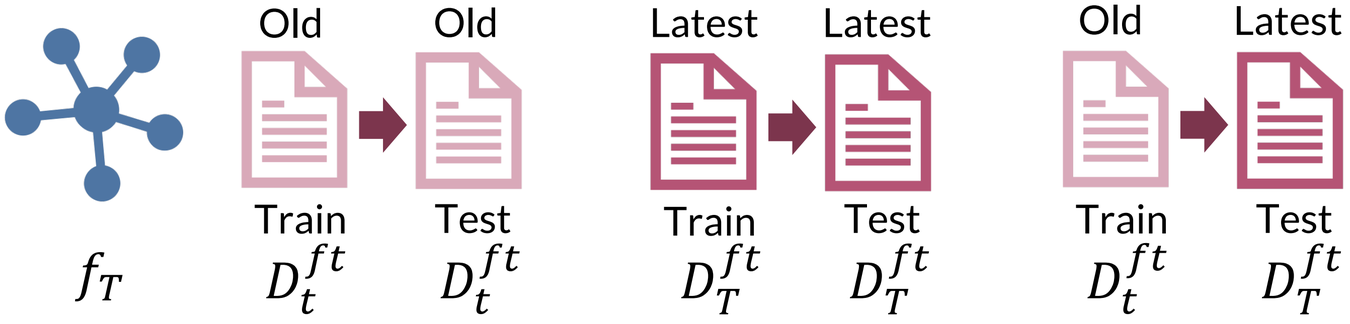
AntNLP—纪焘

Background: CL+PLMs

Continual Pretraining

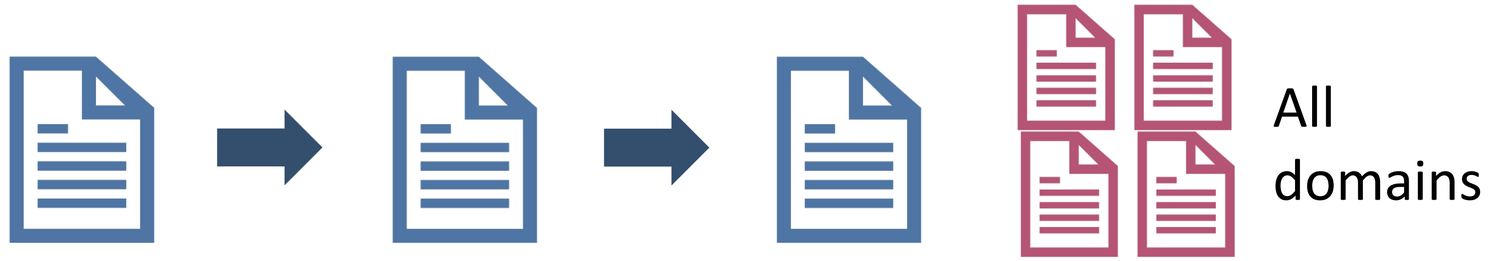


Fine-Tuning & Evaluation

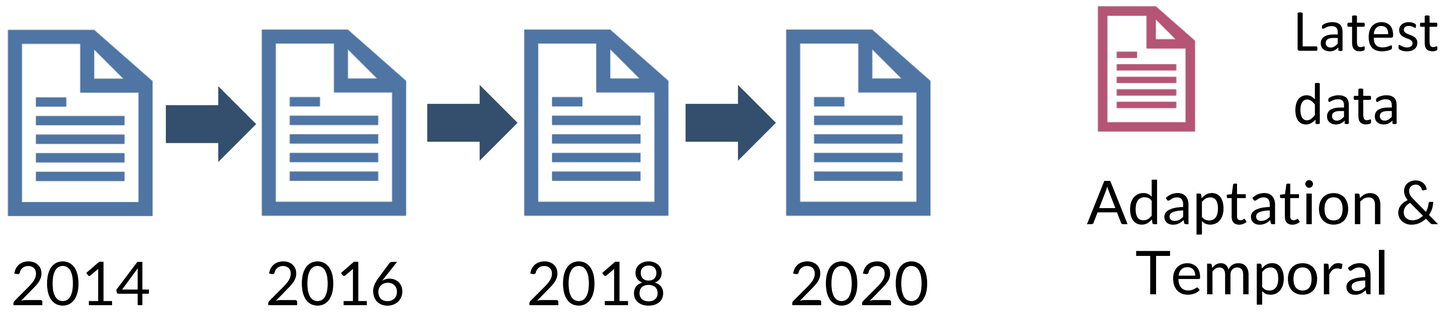


Background: Two Streams

Domain-Incremental Research Papers Stream



Chronologically-Ordered Tweet Stream



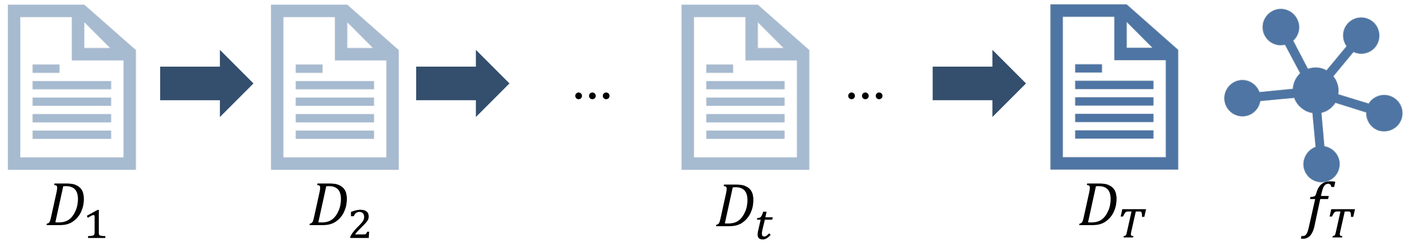
200 class

*500 sample

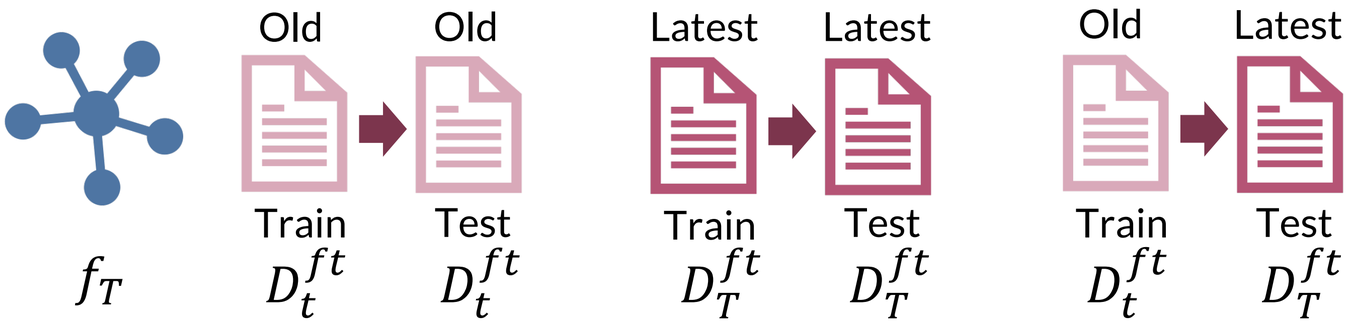
@1 year

Background: Evaluation Protocols

Continual Pretraining

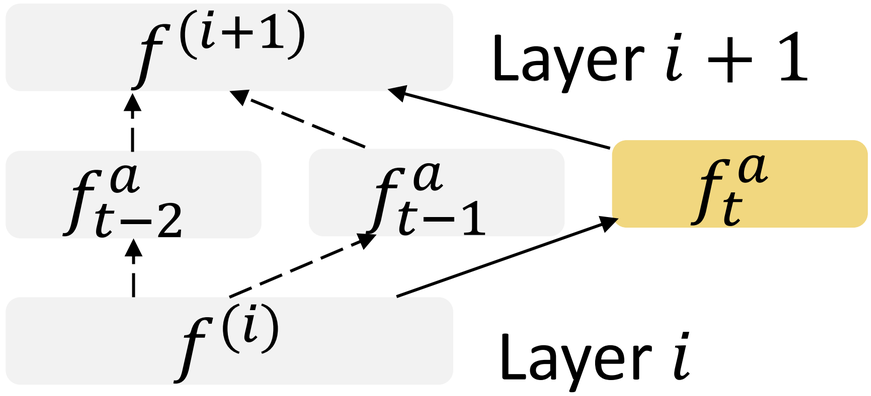


Fine-Tuning & Evaluation

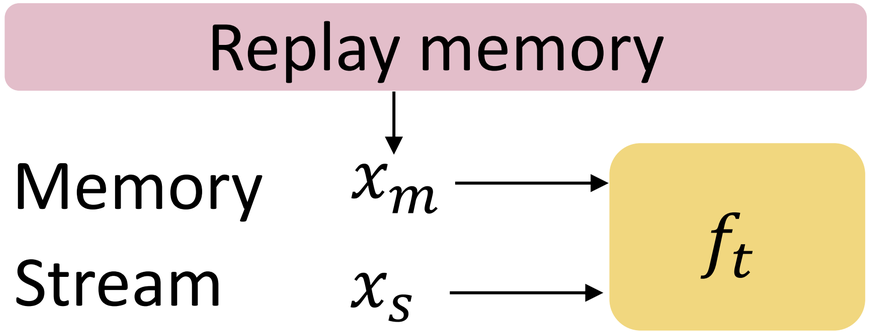


3 Methods

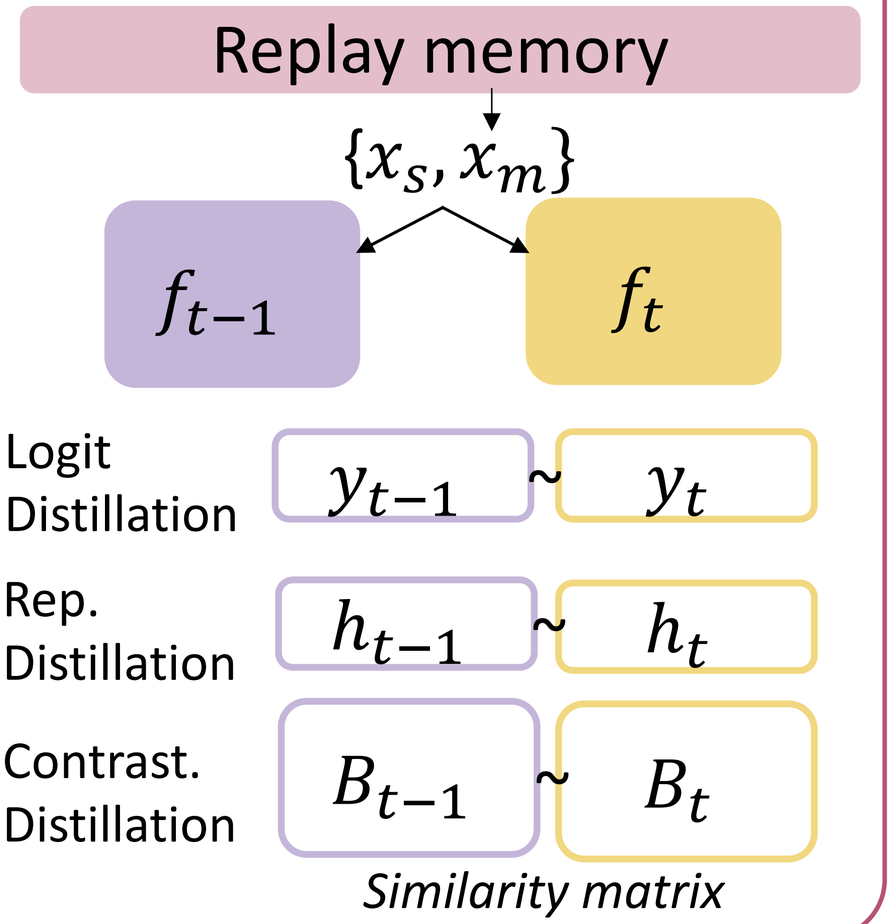
Adapters



Memory Replay



Distillation + CL



Results

Task	Task 1 - Biomedical						Task 2 - Computer Science				Task 3 - Materials Science	
Dataset	Chemprot			RCT-Sample			ACL-ARC		SciERC		MNER	Synthesis
Evaluated After	Task 1	Task 2	Task 3	Task 1	Task 2	Task 3	Task 2	Task 3	Task 2	Task 3	Task 3	Task 3
Roberta-base	$82.03_{\pm 0.7}$	$82.03_{\pm 0.7}$	$82.03_{\pm 0.7}$	$78.07_{\pm 0.7}$	$78.07_{\pm 0.7}$	$78.07_{\pm 0.7}$	$64.32_{\pm 2.8}$	$64.32_{\pm 2.8}$	$79.07_{\pm 1.6}$	$79.07_{\pm 1.6}$	$83.15_{\pm 0.3}$	$91.25_{\pm 0.6}$
Naive	$\mathbf{83.74}_{\pm 0.3}$	$82.60_{\pm 0.3}$	$83.37_{\pm 0.5}$	$\mathbf{81.10}_{\pm 0.5}$	$80.49_{\pm 0.3}$	$80.63_{\pm 0.3}$	$\mathbf{73.71}_{\pm 2.8}$	$69.93_{\pm 2.2}$	$82.14_{\pm 1.1}$	$80.70_{\pm 0.8}$	$83.34_{\pm 0.3}$	$92.72_{\pm 1.0}$
ER	$\mathbf{83.74}_{\pm 0.3}$	$82.96_{\pm 0.3}$	$83.50_{\pm 0.6}$	$\mathbf{81.10}_{\pm 0.5}$	$80.79_{\pm 0.4}$	$81.04_{\pm 0.1}$	$69.92_{\pm 1.6}$	$69.09_{\pm 2.1}$	$82.01_{\pm 0.9}$	$80.59_{\pm 0.2}$	$83.79_{\pm 0.4}$	$\mathbf{93.20}_{\pm 0.2}$
Adapter	$83.68_{\pm 0.3}$	$83.03_{\pm 0.4}$	$83.19_{\pm 0.6}$	$80.72_{\pm 0.7}$	$80.63_{\pm 0.7}$	$80.64_{\pm 0.5}$	$73.28_{\pm 3.9}$	$67.60_{\pm 5.7}$	$79.79_{\pm 1.5}$	$80.10_{\pm 1.0}$	$\mathbf{83.94}_{\pm 0.4}$	$90.82_{\pm 3.3}$
Rep-KD	$\mathbf{83.74}_{\pm 0.3}$	$82.24_{\pm 1.0}$	$82.90_{\pm 0.3}$	$\mathbf{81.10}_{\pm 0.5}$	$80.60_{\pm 0.2}$	$80.51_{\pm 0.3}$	$70.68_{\pm 2.1}$	$69.93_{\pm 2.6}$	$80.45_{\pm 1.4}$	$79.58_{\pm 0.7}$	$83.89_{\pm 0.4}$	$92.16_{\pm 0.6}$
Contrast-KD	$83.38_{\pm 0.3}$	$82.39_{\pm 0.4}$	$83.06_{\pm 0.2}$	$81.00_{\pm 0.3}$	$80.39_{\pm 0.4}$	$80.53_{\pm 0.4}$	$75.34_{\pm 2.1}$	$69.94_{\pm 1.9}$	$80.85_{\pm 1.1}$	$\mathbf{82.45}_{\pm 0.9}$	$83.21_{\pm 0.3}$	$92.05_{\pm 0.4}$
Logit-KD	$\mathbf{83.74}_{\pm 0.3}$	$\mathbf{83.04}_{\pm 0.2}$	$\mathbf{84.12}_{\pm 0.4}$	$\mathbf{81.10}_{\pm 0.5}$	$\mathbf{81.26}_{\pm 0.4}$	$\mathbf{81.19}_{\pm 0.2}$	$70.72_{\pm 2.7}$	$\mathbf{71.38}_{\pm 1.8}$	$\mathbf{82.74}_{\pm 0.5}$	$80.93_{\pm 0.8}$	$83.54_{\pm 0.2}$	$92.73_{\pm 1.0}$
Task-Specific LM		$83.74_{\pm 0.3}$			$81.10_{\pm 0.5}$			$72.20_{\pm 2.6}$		$81.24_{\pm 1.7}$	$84.02_{\pm 0.2}$	$91.56_{\pm 0.4}$

CF😈Pretrain

Results

Task	Task 1 - Biomedical						Task 2 - Computer Science				Task 3 - Materials Science	
Dataset	Chemprot			RCT-Sample			ACL-ARC		SciERC		MNER	Synthesis
Evaluated After	Task 1	Task 2	Task 3	Task 1	Task 2	Task 3	Task 2	Task 3	Task 2	Task 3	Task 3	Task 3
Roberta-base	$82.03_{\pm 0.7}$	$82.03_{\pm 0.7}$	$82.03_{\pm 0.7}$	$78.07_{\pm 0.7}$	$78.07_{\pm 0.7}$	$78.07_{\pm 0.7}$	$64.32_{\pm 2.8}$	$64.32_{\pm 2.8}$	$79.07_{\pm 1.6}$	$79.07_{\pm 1.6}$	$83.15_{\pm 0.3}$	$91.25_{\pm 0.6}$
Naive	$\mathbf{83.74}_{\pm 0.3}$	$82.60_{\pm 0.3}$	$83.37_{\pm 0.5}$	$\mathbf{81.10}_{\pm 0.5}$	$80.49_{\pm 0.3}$	$80.63_{\pm 0.3}$	$\mathbf{73.71}_{\pm 2.8}$	$69.93_{\pm 2.2}$	$82.14_{\pm 1.1}$	$80.70_{\pm 0.8}$	$83.34_{\pm 0.3}$	$92.72_{\pm 1.0}$
ER	$\mathbf{83.74}_{\pm 0.3}$	$82.96_{\pm 0.3}$	$83.50_{\pm 0.6}$	$\mathbf{81.10}_{\pm 0.5}$	$80.79_{\pm 0.4}$	$81.04_{\pm 0.1}$	$69.92_{\pm 1.6}$	$69.09_{\pm 2.1}$	$82.01_{\pm 0.9}$	$80.59_{\pm 0.2}$	$83.79_{\pm 0.4}$	$\mathbf{93.20}_{\pm 0.2}$
Adapter	$83.68_{\pm 0.3}$	$83.03_{\pm 0.4}$	$83.19_{\pm 0.6}$	$80.72_{\pm 0.7}$	$80.63_{\pm 0.7}$	$80.64_{\pm 0.5}$	$73.28_{\pm 3.9}$	$67.60_{\pm 5.7}$	$79.79_{\pm 1.5}$	$80.10_{\pm 1.0}$	$\mathbf{83.94}_{\pm 0.4}$	$90.82_{\pm 3.3}$
Rep-KD	$\mathbf{83.74}_{\pm 0.3}$	$82.24_{\pm 1.0}$	$82.90_{\pm 0.3}$	$\mathbf{81.10}_{\pm 0.5}$	$80.60_{\pm 0.2}$	$80.51_{\pm 0.3}$	$70.68_{\pm 2.1}$	$69.93_{\pm 2.6}$	$80.45_{\pm 1.4}$	$79.58_{\pm 0.7}$	$83.89_{\pm 0.4}$	$92.16_{\pm 0.6}$
Contrast-KD	$83.38_{\pm 0.3}$	$82.39_{\pm 0.4}$	$83.06_{\pm 0.2}$	$81.00_{\pm 0.3}$	$80.39_{\pm 0.4}$	$80.53_{\pm 0.4}$	$75.34_{\pm 2.1}$	$69.94_{\pm 1.9}$	$80.85_{\pm 1.1}$	$\mathbf{82.45}_{\pm 0.9}$	$83.21_{\pm 0.3}$	$92.05_{\pm 0.4}$
Logit-KD	$\mathbf{83.74}_{\pm 0.3}$	$\mathbf{83.04}_{\pm 0.2}$	$\mathbf{84.12}_{\pm 0.4}$	$\mathbf{81.10}_{\pm 0.5}$	$\mathbf{81.26}_{\pm 0.4}$	$\mathbf{81.19}_{\pm 0.2}$	$70.72_{\pm 2.7}$	$\mathbf{71.38}_{\pm 1.8}$	$\mathbf{82.74}_{\pm 0.5}$	$80.93_{\pm 0.8}$	$83.54_{\pm 0.2}$	$92.73_{\pm 1.0}$
Task-Specific LM		$83.74_{\pm 0.3}$			$81.10_{\pm 0.5}$			$72.20_{\pm 2.6}$		$81.24_{\pm 1.7}$	$84.02_{\pm 0.2}$	$91.56_{\pm 0.4}$

Fwd-Trans

Q & A