NEURAL INFORMATION PROCESSING SYSTEMS

Order-Level Attention Similarity Across Language Models: A Latent Commonality

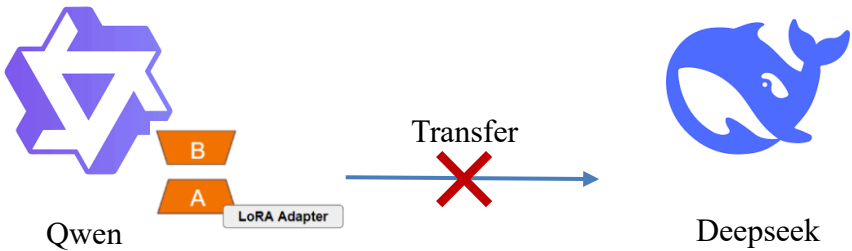
Jinglin Liang, Jin Zhong, Shuangping Huang*,

Yunqing Hu, Huiyuan Zhang, Huifang Li, Lixin Fan, Hanlin Gu

Contents

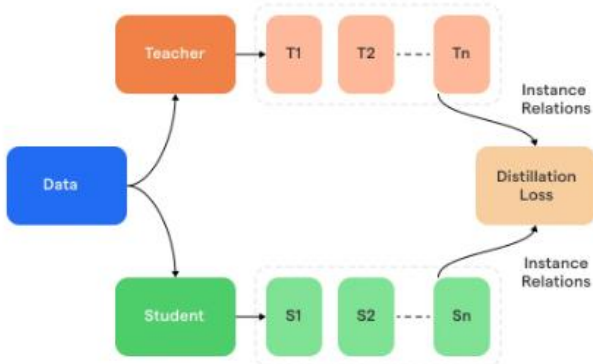
Background

- Fine-tuning is an effective technique for the domain adaptation of a LLM.
- Adapters trained from a specific LLM are tightly coupled with it and are not transferable to other models



Background

- Transfer learning methods, such as knowledge distillation, have shown that knowledge can be transferred when two models share a similar feature space.
- This inspires us to ask: Do different LLMs possess similar feature spaces, which would allow for cross-model knowledge transfer?



Background

- Considering this from the dimension of attention patterns: while language models differ in structure and training data, they all essentially use Multi-Head Attention for contextual aggregation. This suggests their attention patterns might converge to similar modes.

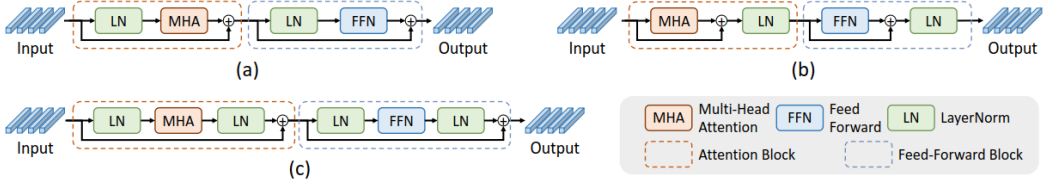


Figure 2: Structure of each layer in typical LMs. (a) Llama and Qwen. (b) Bert, Roberta, and Electra. (c) Gemma.

Contents

Order-Level Attention

- Different models have varying numbers of layers and heads, and the architecture of each layer also differs.
- We cannot directly compare their attention patterns. We need to find a representation that shares a common physical meaning.

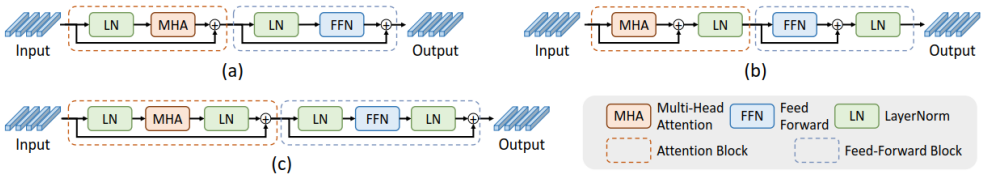


Figure 2: Structure of each layer in typical LMs. (a) Llama and Qwen. (b) Bert, Roberta, and Electra. (c) Gemma.

Order-Level Attention

- While layer structures vary between models, they all aggregate context via a Multi-Head Attention (MHA) block and a residual connection.
- The contextual information aggregation matrix (Attention Rollout) can be expressed as:

$$\hat{A} = \prod_{i=1}^{N} (A^{(i)} + I),$$

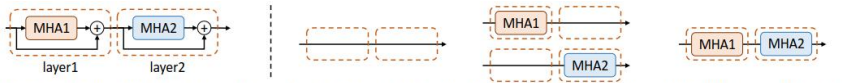


Figure 1: Information flow path decomposition. (a) Simplified LM showing only Multi-Head Attention (MHA) modules (others omitted); (b) Path via residual connections across all layers; (c) Paths through MHA in one layer, residual connections elsewhere; (d) Path through MHA in all layer.

Order-Level Attention

- We found that the 'attention sink' phenomenon exists in the attention rollouts of different models
- This is caused by the presence of ineffective components within them

$$\hat{A} = \prod_{i=1}^{N} (A^{(i)} + I),$$

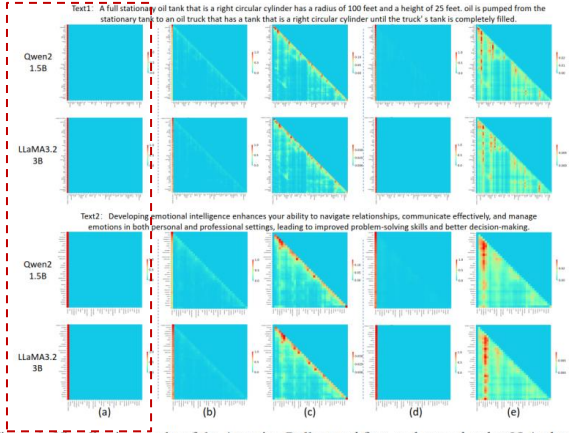


Figure 3: Visualization results of the Attention Rollout and first- and second-order OLA obtained by inputting two texts into Qwen2-1.5b and Llama3.2-3b. (a) Attention Rollout. (b) First-order OLA. (c) First-order OLA with row-wise maximum values set to zero. (d) Second-order OLA. (e) Second-order OLA with row-wise maximum values set to zero.

Order-Level Attention

- Decompose the Attention Rollout by order to obtain the attention for each order.
- There is a significant similarity in the same-order attention across different models.

$$\hat{A} = I + \sum_{i=1}^{N} A^{(i)} + \sum_{1 \le i < j \le N} A^{(j)} A^{(i)} + \cdots + A^{(N)} A^{(N-1)} \cdots A^{(1)},$$

$$\hat{A} = \sum_{i=0}^{N} \binom{N}{i} \cdot \hat{A}^{(i)},$$

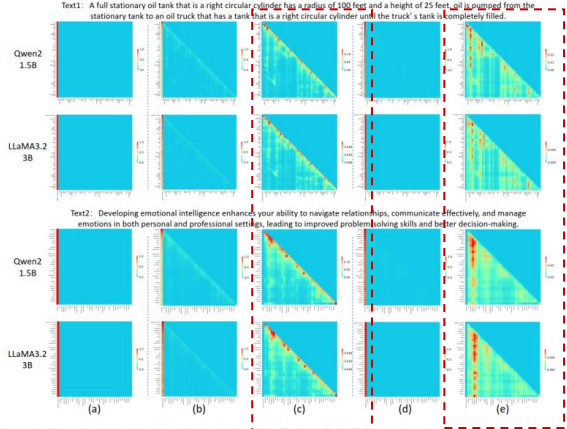


Figure 3: Visualization results of the Attention Rollout and first- and second-order OLA obtained by inputting two texts into Qwen2-1.5b and Llama3.2-3b. (a) Attention Rollout. (b) First-order OLA. (c) First-order OLA with row-wise maximum values set to zero. (d) Second-order OLA. (e) Second-order OLA with row-wise maximum values set to zero.

Order-

- Decompose t
- There is a sig

ls.

$$\hat{A} = I + \sum_{i=1}^{N} A^{(i)} + \sum_{1 \le i <}$$

$$\hat{A}$$

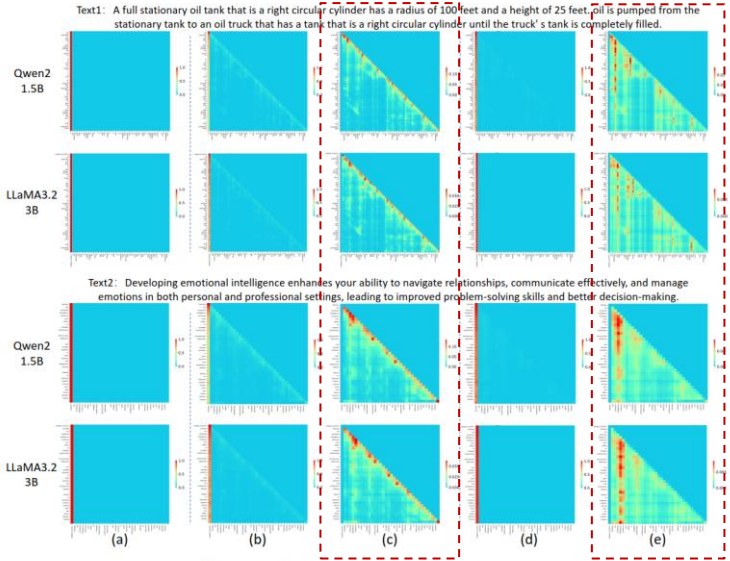


Figure 3: Visualization results of the Attention Rollout and first- and second-order OLA obtained by inputting two texts into Qwen2-1.5b and Llama3.2-3b. (a) Attention Rollout. (b) First-order OLA. (c) First-order OLA with row-wise maximum values set to zero. (d) Second-order OLA. (e) Second-order OLA with row-wise maximum values set to zero.

Order-Level Attention

- Qualitatively analyze the similarity of same-order attention across different models
- We use a visual classifier in place of a human to provide an objective assessment of this similarity

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{(a,i)\sim\mathcal{D}_{train}}\left[\mathcal{L}_{\text{CE}}(F_\theta(a), i)\right], \quad \mathcal{D}_{train} = \left\{(a_i^{(j)}, i) \;\middle|\; i \in [1..M],\, j \in [1..S]\right\}$$

Table 1: Results of quantitative evaluation on cross-model similarity for OLA and baselines based on visual classification model. Entries represent accuracy (unit: %), averaged over three experiments, reflecting source-target LM similarity (higher = more similar). The terms 1st, 2nd, 3rd denote first-, second-, and third-order OLA. Best performance is **bolded**.

(a) CLM Results. Q-1b5, Q-7b, G-2b, G-9b, L-3b, and L-8b denote Qwen2-1.5b, Qwen2-7b, Gemma2-2b, Gemma2-9b, Llama3.2-3b, and Llama3.1-8b.

Source	L-3b, L-8b, G-2b, G-9b		L-3b, L-8b, Q-1b5, Q-7b		Q-1b5, Q-7b, G-2b, G-9b	
Target	Q-1b5	Q-7b	G-2b	G-9b	L-3b	L-8b
Rollout [10]	27.90	7.70	52.60	26.00	66.10	59.70
IRNL [28]	18.50	11.90	58.10	67.20	77.20	73.60
ALTI [11]	22.60	15.50	69.30	71.80	85.60	79.80
1st	52.60	49.20	**93.10**	**92.40**	**94.60**	**94.10**
2nd	67.10	**49.90**	89.30	86.20	90.70	91.90
3rd	**76.50**	48.90	79.90	78.80	88.60	86.70

(b) MLM Results. B-b, B-l, R-b, R-l, E-b, and E-l denote Bert-base, Bert-large, Roberta-base, Roberta-large, Electra-base, and Electra-large, respectively.

Source	R-b, R-l, E-b, E-l		B-b, B-l, E-b, E-l		B-b, B-l, R-b, R-l	
Target	B-b	B-l	R-b	R-l	E-b	E-l
Rollout	44.30	42.80	13.20	36.60	13.60	9.10
IRNL	60.00	4.10	20.80	15.00	55.20	38.10
ALTI	90.30	80.30	73.80	48.10	86.80	90.30
1st	**91.90**	**92.40**	**80.40**	**69.90**	**95.20**	**95.80**
2nd	88.80	62.70	60.70	28.00	76.30	82.90
3rd	80.40	67.70	38.10	31.90	58.10	68.30

Order-Level Attention

- Qualitatively analyze the similarity of same-order attention across different models
- Use the target model's OLA to retrieve the corresponding OLA from the source model

Table 2: Retrieval-based quantitative evaluation of first-order OLA cross-model similarity. Row headers denote source LMs. Column headers denote target LMs. Entries indicate evaluation metrics Hits@1 / Hits@5 (unit: %).

(a) CLM Results.

Src\Tgt	Q-1b5	G-2b	L-3b
Q-1b5	-	83.60 / 89.40	95.90 / 97.00
G-2b	83.20 / 89.30	-	95.30 / 97.10
L-3b	92.90 / 96.10	94.10 / 96.50	-

(b) MLM Results.

Src\Tgt	B-b	R-b	E-b
B-b	-	51.90 / 58.80	91.60 / 94.90
R-b	75.90 / 83.90	-	71.70 / 80.20
E-b	92.40 / 96.00	67.40 / 72.90	-

Relation between OLA and Syntactic Knowledge

- The syntactic dependency structure of the original sentence can be predicted using only OLA
- OLA encodes syntactic knowledge and can serve as a syntactic feature

Table 3: Results of syntactic dependency parsing using OLA predicted by LMs. Entries indicate UAS/LAS (unit: %). Best performance is **bolded**.

(a) CLM Results.

LMs	Q-1b5	G-2b	L-3b
Rollout	50.53/29.79	44.24/22.04	53.77/35.57
1st	**63.58/48.24**	**62.25/45.95**	**62.98/48.19**
2st	60.58/43.90	57.28/38.88	58.93/42.94
3rd	55.19/36.82	51.89/32.25	51.00/33.35

(b) MLM Results.

LMs	B-b	R-b	E-b
Rollout	46.20/30.69	35.77/17.94	50.35/34.02
1st	**81.29/72.16**	**80.00/70.44**	**81.23/72.63**
2st	72.86/61.05	72.68/60.10	77.47/66.78
3rd	66.44/53.17	36.99/18.67	50.72/33.90

Contents

Transferable OLA Adapter

- Leveraging the above findings, we propose a training-free adapter transfer method (TOA)
- Trains an adapter by using OLA as a universal, cross-model syntactic feature for its input

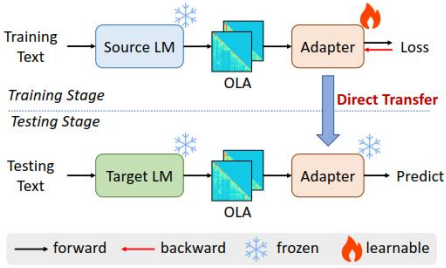


Figure 4: Overview of TOA. In the training phase, the source LM is frozen and an adapter is trained for the downstream task using OLA as input. In the testing phase, the adapter is directly transferred to the target LM.

Transferable OLA Adapter

- Validated on several NLP tasks, including RE, NER, POS tagging, and DP, TOA achieves effective, training-free cross-model adapter transfer, significantly enhancing the capabilities of the target model.

Table 4: Cross-Model Transferability of TOA on the RE Task. Column headers indicate source LMs, row headers indicate target LMs, and entries represent relation prediction accuracy (unit: %). Best performance is **bolded**; scores exceeding the zero-shot baseline are underlined.

(a) CLM Results.

Src\Tgt	Q-1b5	Q-7b	G-2b	G-9b	L-3b	L-8b
Zero-shot						
-	7.69	8.58	5.01	18.22	14.65	12.99
TOA (Ours)						
Q-1b5	**34.90**	26.33	30.95	25.98	31.08	29.46
Q-7b	27.58	**31.48**	25.25	21.60	25.41	25.44
G-2b	21.17	19.92	**34.73**	23.33	23.49	22.64
G-9b	18.63	17.42	22.12	**26.28**	20.99	20.73
L-3b	30.49	22.35	33.49	22.19	**35.57**	33.03
L-8b	28.24	22.28	30.03	25.80	32.03	**33.43**

(b) MLM Results.

Src\Tgt	B-b	B-l	R-b	R-l	E-b	E-l
Zero-shot						
-	2.69	0.48	0.04	7.18	5.78	0.04
TOA (Ours)						
B-b	**36.19**	29.90	23.90	23.60	25.70	28.28
B-l	22.13	**32.29**	19.33	23.16	18.63	17.45
R-b	25.63	18.70	**32.63**	21.61	26.40	22.50
R-l	25.59	28.98	25.15	**32.73**	24.12	25.15
E-b	36.01	26.99	31.96	25.77	**41.27**	36.97
E-l	31.85	31.89	30.63	24.15	32.99	**38.18**

Contents

Conclusion

- We propose Order-Level Attention (OLA) to unify the attention of different LLMs into a comparable representation.
- We present two key findings:
  - Same-order OLA exhibit significant similarity across different LLMs.
  - OLA encodes syntactic knowledge.
- Leveraging these findings, we propose the Transferable OLA Adapter (TOA), which achieves training-free, cross-model adapter transfer.