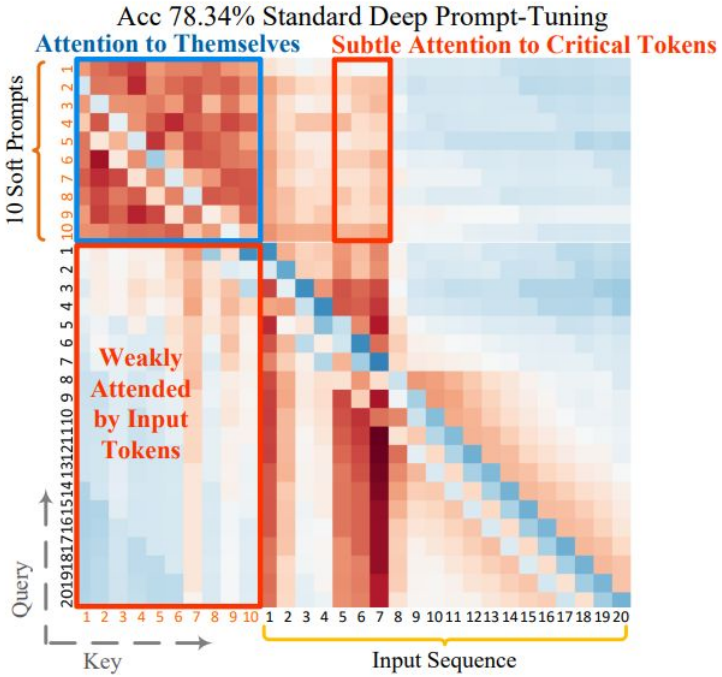
All You Need is One: Capsule Prompt Tuning with a Single Vector

Yiyang Liu, James C. Liang, Heng Fan, Wenhao Yang, Yiming Cui, Xiaotian Han, Lifu Huang, Dongfang Liu, Qifan Wang, Cheng Han

Attention Analysis of Task-aware Prompt Tuning

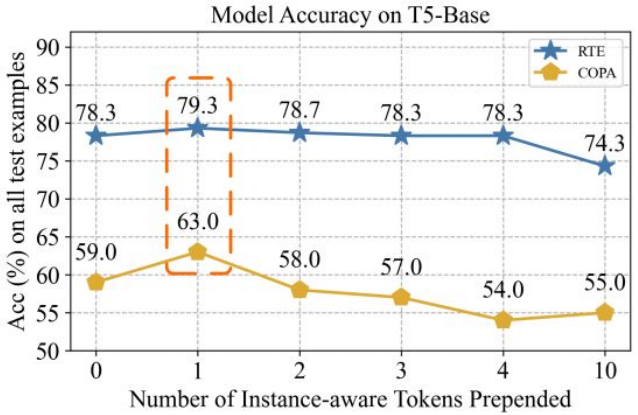
Prompt tuning offers a flexible and efficient solution with minimal input sequence adjustment, thus enabling fast adaptation of large language models (LLMs).

Counterintuitively, task-aware soft prompts typically fail to establish strongly attentive interaction with input tokens. This observation reveals that the task-aware design of soft prompts may limit their capability to adapt to diverse input semantics, potentially constraining the effectiveness of prompt-based learning.

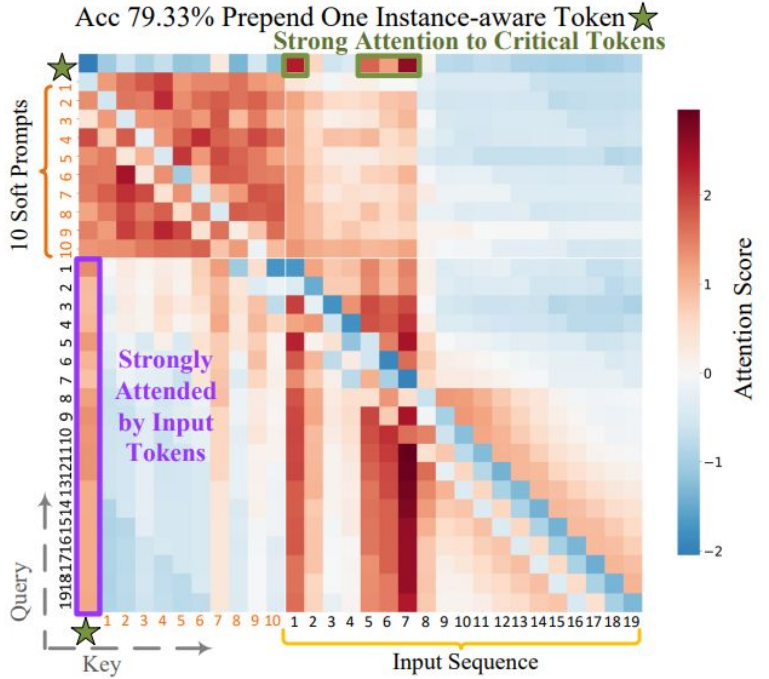


The Power of Instance-aware Information

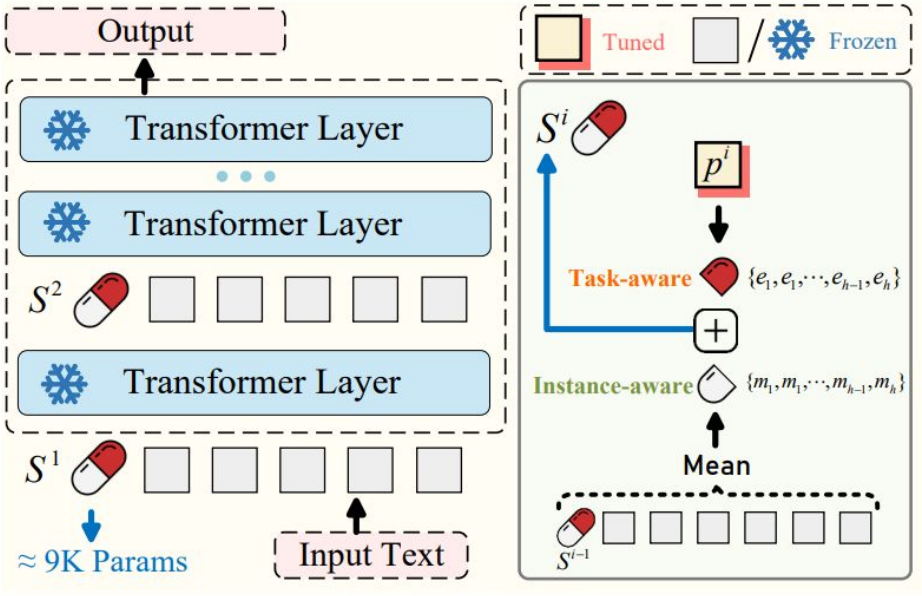
Instance-aware token can enhance model performance even without any fine-tuning.



Instance-aware token presents strongly attentive interaction with input sequences as “attention anchor.”



Capsule Prompt Tuning with A Single Vector (CaPT)



$$S^1 = p^1 + \texttt{Mean}(E)$$
$$\underline{S}^1, H^1 = L_1(S^1, E)$$
$$S^i = p^i + \texttt{Mean}(\underline{S}^{i-1} \oplus H^{i-1}) \quad i = 2, 3, \ldots, N$$
$$\underline{S}^i, H^i = L_i(S^i,\ H^{i-1}) \quad i = 2, 3, \ldots, N$$

Experimental Results

Method	# Para	Boolq Acc	CB F1/Acc	COPA Acc	MRC F1a	RTE Acc	WiC Acc	Average Score
T5-Base (220M)								
Fine-Tuning† [64]	100%	82.30	91.30	60.00	79.70	84.50	69.30	77.85
Prompt-Tuning* [EMNLP21] [11]	0.06%	78.12	84.42	54.37	78.30	75.27	62.29	72.13
P-Tuning v2* [ACL22] [53]	0.53%	**80.81**	90.23	61.28	79.83	**81.98**	67.56	76.94
XPrompt* [EMNLP22] [17]	0.04%	79.67	86.72	56.95	78.57	78.29	64.31	74.09
ResPrompt* [ACL23] [52]	0.21%	79.25	85.33	58.64	78.42	77.14	62.36	73.52
SMoP† [EMNLP23] [23]	8e-3%	79.40	86.42	58.30	79.60	77.50	65.20	74.40
SuperPos-Prompt† [NeurIPS24] [65]	-	74.00	80.20	62.00	72.90	70.40	67.60	71.18
VFPT [NeurIPS24] [10]	0.21%	78.38	90.92	61.76	78.73	76.90	65.36	75.34
DePT† [ICLR24] [66]	-	79.30	-	-	74.30	79.10	**68.70**	-
EPT [NAACL25] [67]	0.06%	79.14	90.18	56.33	73.43	78.99	67.71	74.30
Ours	4e-3%	79.54	**94.16**	**64.33**	**80.46**	79.78	66.77	**77.51**
T5-Large (770M)								
Fine-Tuning [64]	100%	85.75	95.26	76.00	84.41	88.05	72.11	83.60
Prompt-Tuning [EMNLP21] [11]	0.04%	83.20	90.32	57.50	83.10	86.11	68.74	78.16
P-Tuning v2 [ACL22] [53]	0.52%	**85.82**	95.56	77.00	84.07	**89.25**	71.03	83.79
XPrompt* [EMNLP22] [17]	0.02%	83.82	91.39	82.05	81.26	87.72	**73.51**	83.29
ResPrompt* [ACL23] [52]	0.15%	83.51	90.64	**82.79**	84.02	86.97	71.13	83.18
SMoP [EMNLP23] [23]	3e-3%	83.45	92.37	71.00	83.92	87.70	68.60	81.17
VFPT [NeurIPS24] [10]	0.18%	83.89	93.71	75.63	83.24	88.10	71.00	82.56
EPT [NAACL25] [67]	0.04%	84.77	93.40	54.00	80.03	86.33	71.79	78.39
Ours	3e-3%	84.56	**97.22**	80.00	**84.53**	88.45	69.44	**84.03**
Llama3.2-1B								
Linear Head [60]	3e-4%	59.85	51.69	56.33	48.94	55.23	53.45	54.25
Prompt-Tuning [EMNLP21] [11]	0.06%	60.95	61.61	57.67	57.00	62.50	54.70	59.07
P-Tuning v2 [ACL22] [53]	0.53%	62.48	64.29	**61.00**	60.34	58.12	60.15	61.06
SMoP [EMNLP23] [23]	0.04%	61.13	62.50	59.33	57.46	57.40	54.23	57.51
VFPT [NeurIPS24] [10]	0.17%	62.44	61.72	59.67	58.41	64.35	57.60	60.70
EPT [NAACL25] [67]	0.06%	61.56	65.22	56.00	60.18	63.90	59.45	61.05
Ours	3e-3%	**77.28**	**65.82**	58.00	**65.73**	**72.56**	**65.67**	**67.51**
Qwen2.5-1.5B								
Linear Head [60]	2e-4%	59.54	64.66	52.00	53.38	62.45	56.58	58.10
Prompt-Tuning [EMNLP21] [11]	0.05%	61.38	65.22	52.33	53.41	63.18	56.90	58.74
P-Tuning v2 [ACL22] [53]	0.51%	62.08	68.84	55.33	56.31	66.43	**59.09**	61.35
SMoP [EMNLP23] [23]	0.03%	61.41	66.76	54.00	55.34	64.62	58.15	60.05
VFPT [NeurIPS24] [10]	0.12%	63.64	67.78	52.67	55.61	63.54	58.05	60.22
EPT [NAACL25] [67]	0.05%	63.10	68.17	52.33	56.02	67.53	58.30	60.91
Ours	3e-3%	**64.13**	**72.42**	**57.67**	**57.49**	**68.59**	58.46	**63.17**

Main results on SuperGLUE

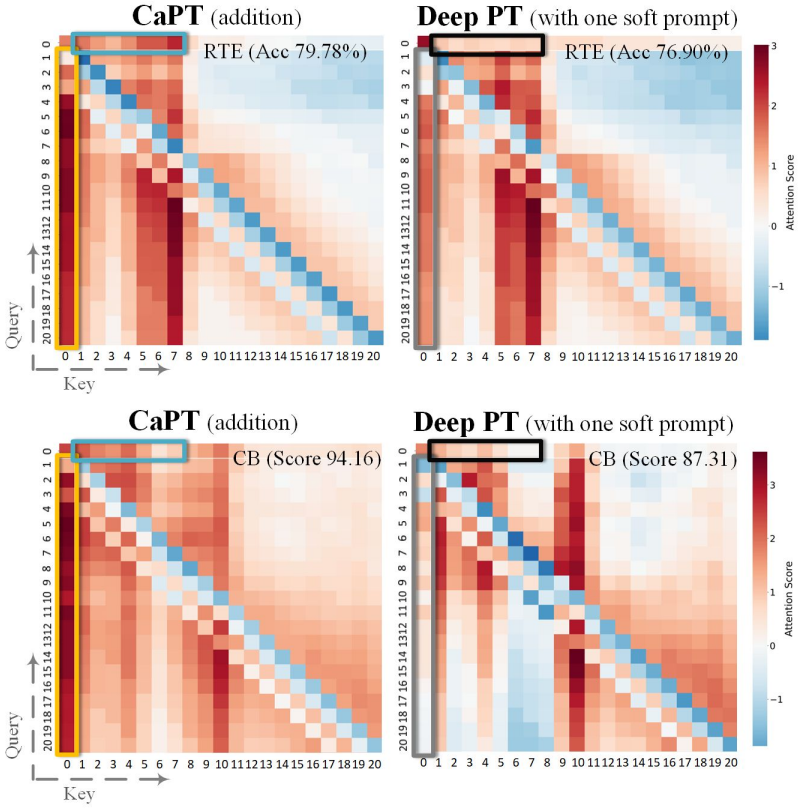
Wall-clock efficiency compared with other prompt-tuning methods

Method	# Para	Time	Average Score
Prompt-Tuning [11]	0.06%	8.77×	72.13
P-Tuning v2 [53]	0.53%	8.37×	76.94
M-IDPG [70]	0.47%	12.58×	76.96
LoPA [71]	0.44%	14.93×	77.98
Ours	4e-3%	1.00×	77.51

Comparison with other PEFT methods

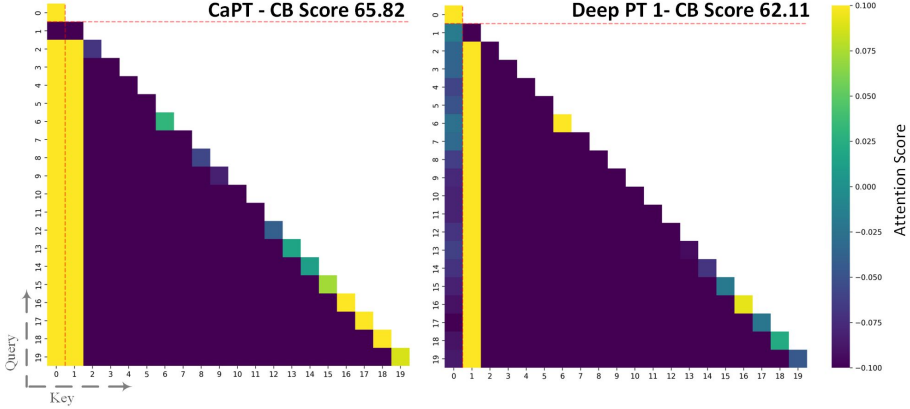
Method	# Para	Boolq Acc	CB F1/Acc	COPA Acc	MRC F1a	RTE Acc	WiC Acc	Average Score
T5-Base (220M)								
Adapter [8]	0.86%	**82.50**	88.05	**71.50**	75.90	71.90	67.10	76.16
LoRA [6]	1.73%	81.30	88.20	70.40	72.60	75.5	**68.30**	76.05
Ours	4e-3%	79.54	**94.16**	64.33	**80.46**	**79.78**	66.77	**77.51**

Attention Anchor



Analysis on T5 encoder (bi-directional attention)

Analysis on Llama (causal attention)



Conclusion

- Based on our finding of “attention anchor” phenomenon, we propose Capsule Prompt-Tuning (CaPT), a novel prompt tuning framework for large language models (LLMs).

- CaPT integrates both instance-aware information from each input sequence and task-aware information from learnable vectors, effectively acting as “attention anchor.” By strengthening attentive interplay with input tokens, it thereby achieves superior performance.

- CaPT can operate in an almost parameter-free manner, utilizing only one single vector per layer, which eliminates the need for time-intensive grid searching, varied lengths across different tasks, and substantial training overhead.