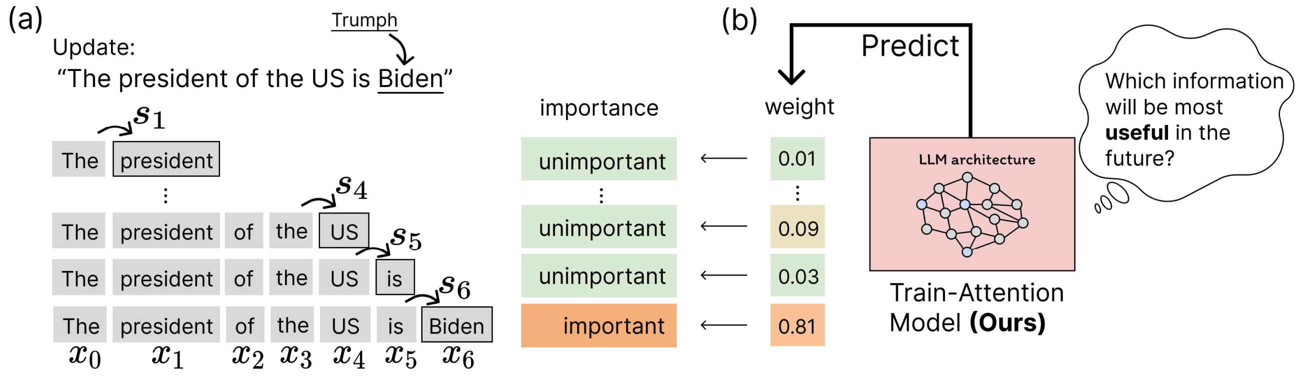
Train-Attention: Meta-Learning Where to Focus in Continual Knowledge Learning

Yeongbin Seo, Dongha Lee*, Jinyoung Yeo*



NEURAL INFORMATION PROCESSING SYSTEMS
2024

Background: Continual Knowledge Learning (CKL)

Ex)

The president of the US is <u>Biden</u>

→ The president of the US is <u>Trumph</u>

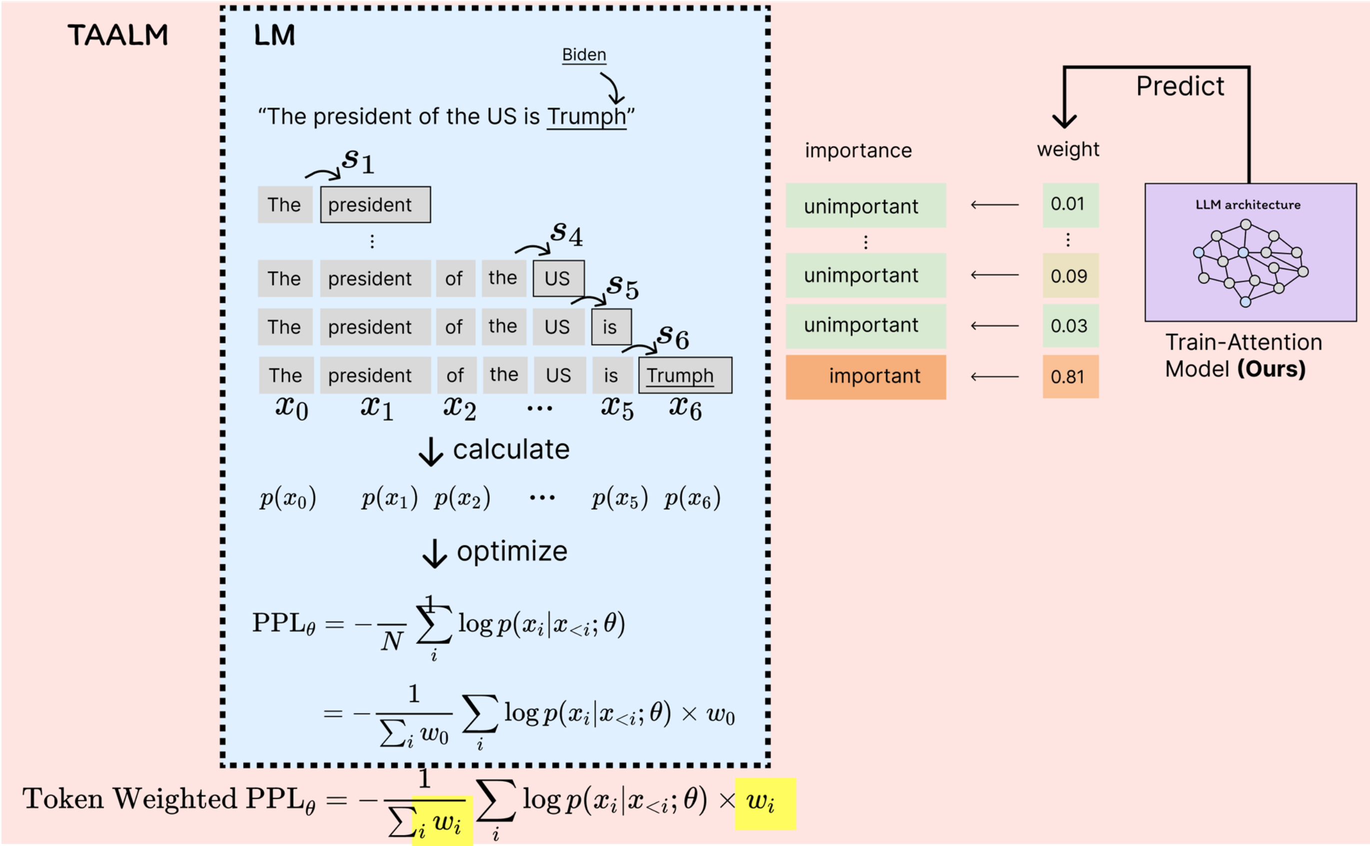
- CKL : Enabling LMs to constantly obtain new and updated knowledge while mitigating forgetting of previous learned

- Two dimensions of evaluating CKL
  - Plasticity : How well obtained
  - Stability : How well preserved

- Previous approach
  1) Adapter
  2) Regularization
  3) Review
- Our approach : Learn only <u>important (useful)</u> information, skip un-important.

Learning only **useful** information

Train-Attention (TA) : detecting and highlighting useful token in the document (D).
TA-augmented LM (TAALM) : LMs learning new information with the aid of TA.



What is **importance**? : **Usefulness**

$\mathcal{D}$ $= \{x_0, x_1, \ldots, x_i, \ldots x_n\}$

: a text data (document), that consists of tokens (x_i)

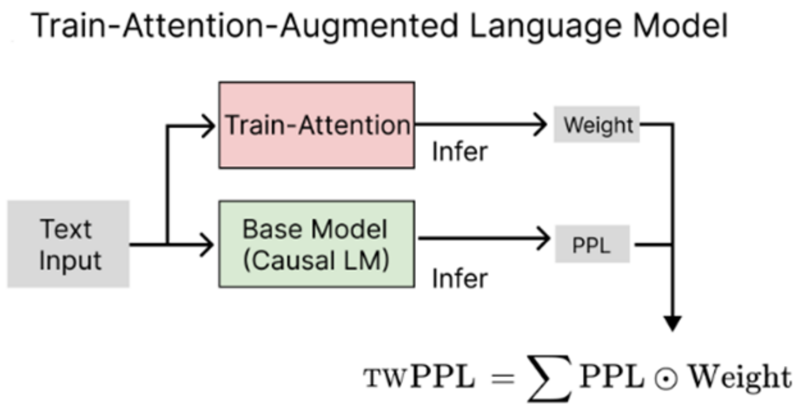
$\mathcal{T}_{\mathcal{D}}$: a task related to $\mathcal{D}$

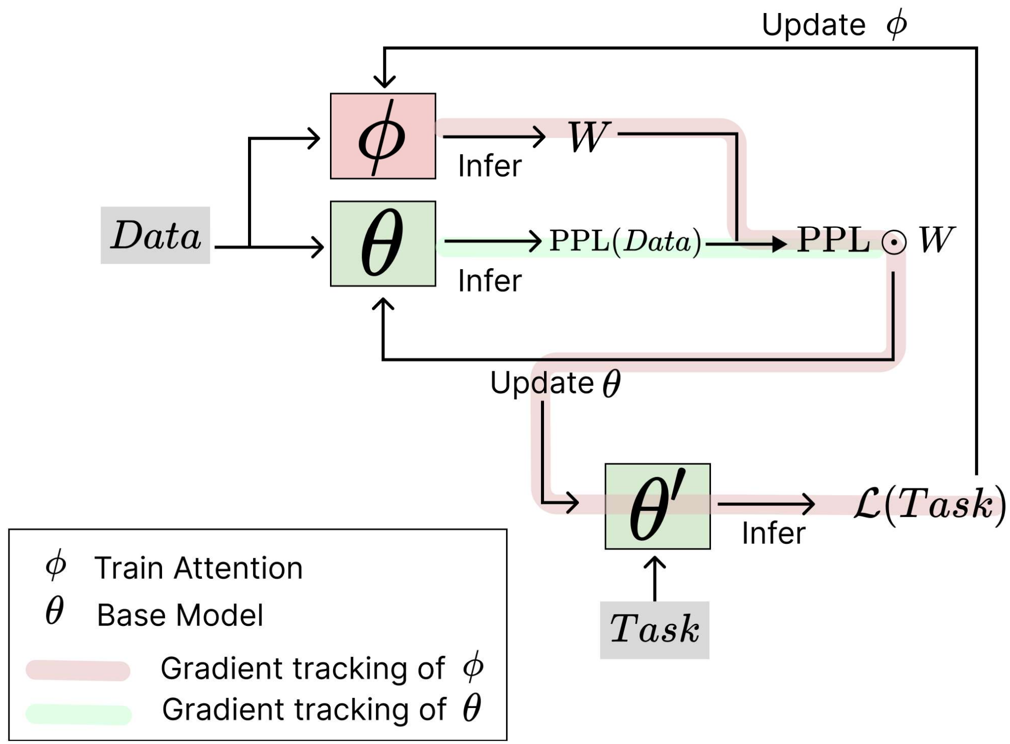
- x_i is **useful** if learning it is expected to help solving some tasks (i.e., improves the performance on tasks) in the future.

Formulate into Meta-learning problem

$$\theta' \leftarrow \theta - \alpha \nabla_\theta tw\,\mathrm{PPL}_\theta(\mathcal{D}, W_{\mathcal{D},\phi})$$

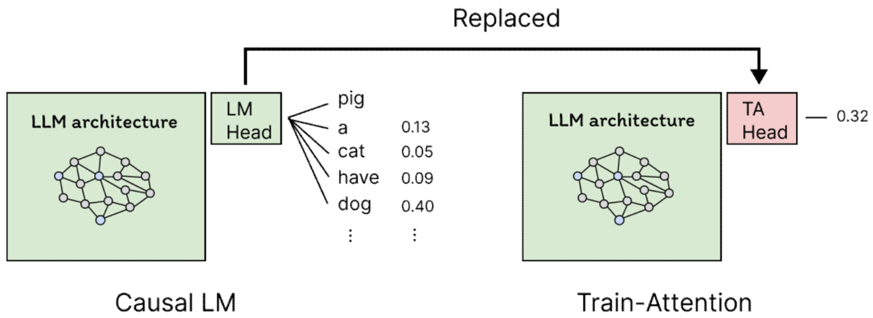
$$\phi \leftarrow \phi - \beta \nabla_\phi \mathcal{L}_{\theta'}(\mathcal{T}_\mathcal{D})$$



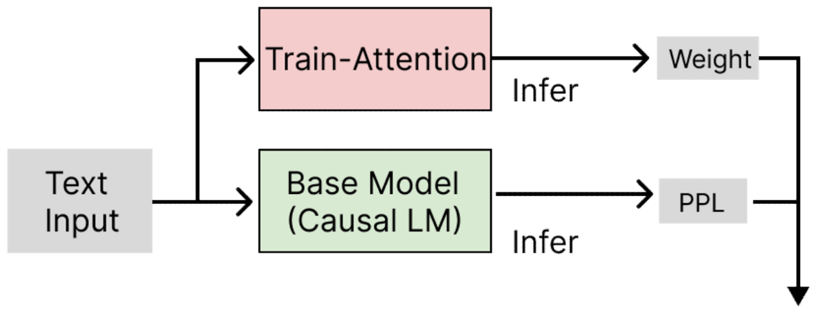


Architecture

TA : Replace decoder layer of transformer model into $hidden\_size \times 1$ TA head.

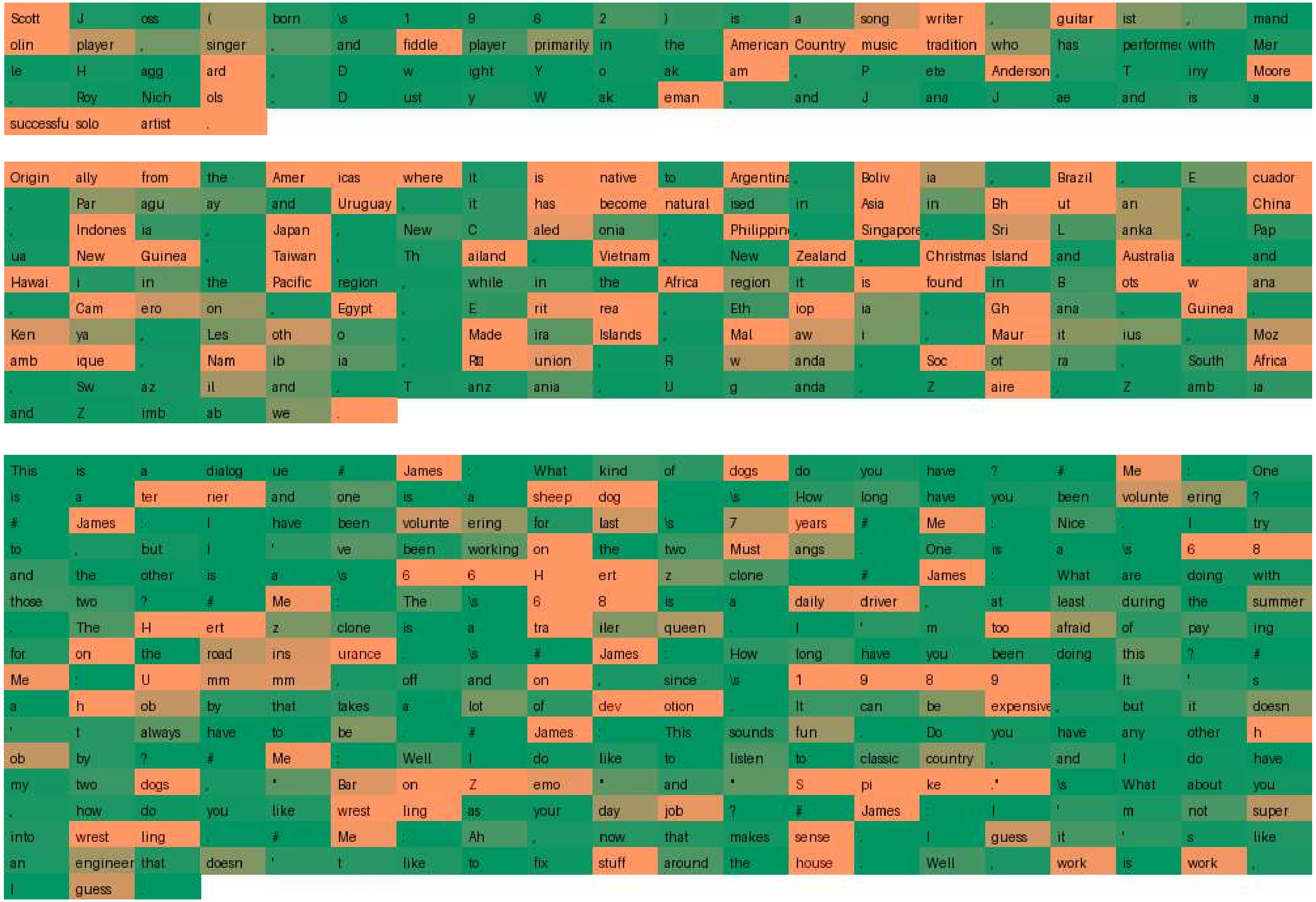


TAALM : Apply TA when training.

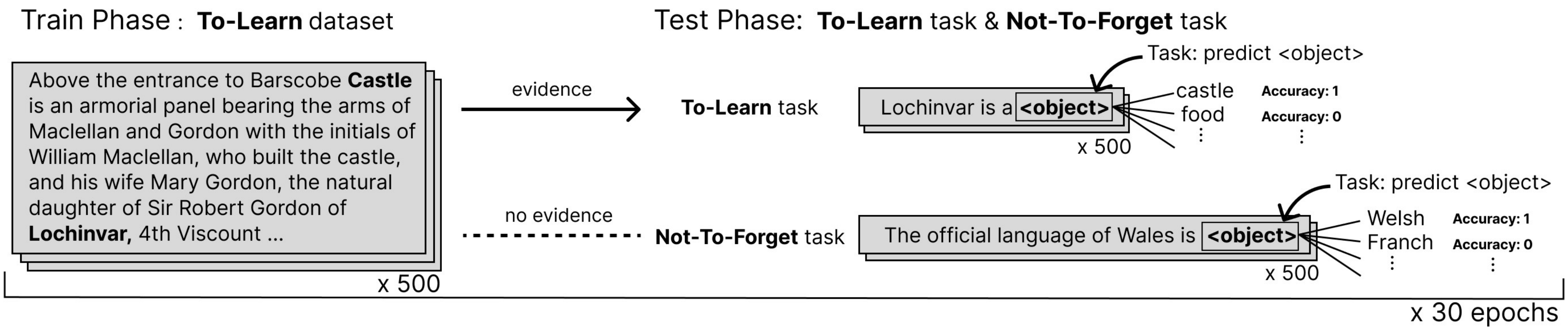


$$\text{TWPPL} = \sum \text{PPL} \odot \text{Weight}$$

the weight map generated by the trained TA. Orange light highlight key information, such as the subject's name, occupation, or date of birth.



Benchmark: LamaCKL



pre-test accuracy 1 -> **Not-To-Forget** set -> evaluate stability
pre-test accuracy 0 -> **To-Learn** set -> evaluate plasticity

Results

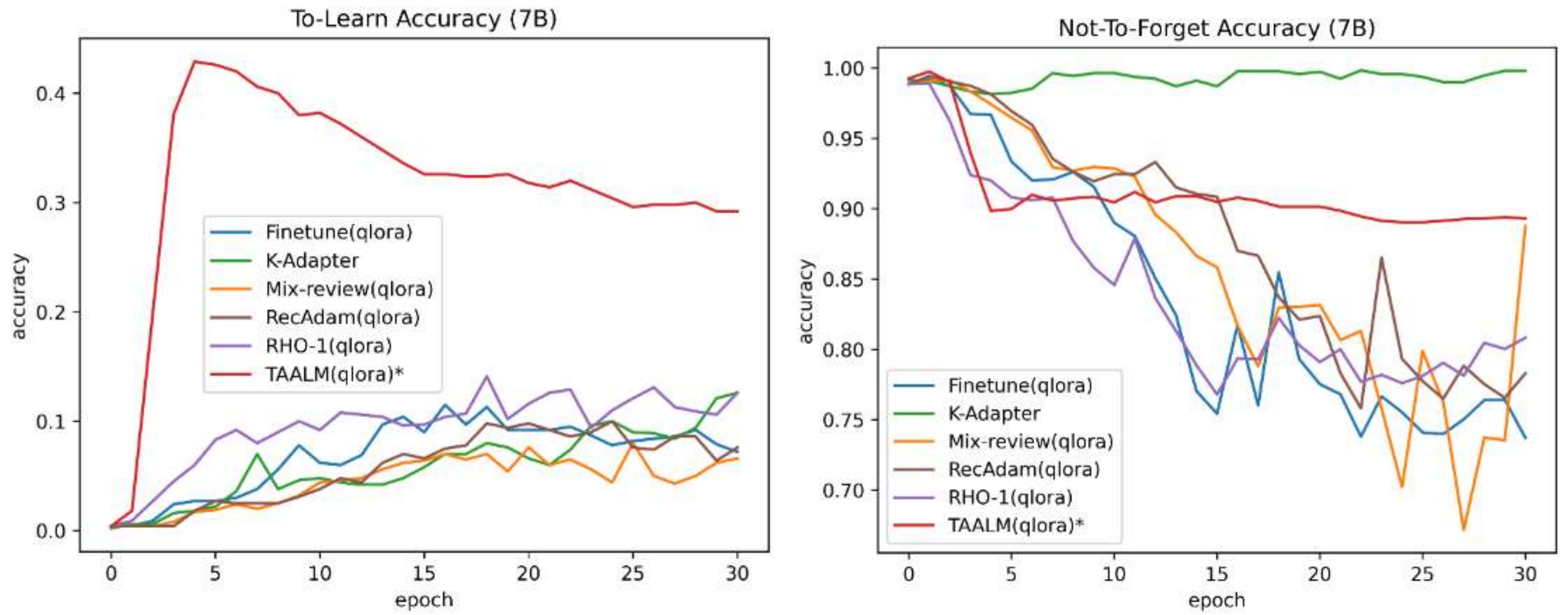


Figure 7: LAMA-CKL performance of large (Llama2-7B) baseline models. The graph on the left represents TO-LEARN task, and the graph on the right represents NOT-TO-FORGET task performance. The x-axis is the learning epoch, and the y-axis is accuracy.

	Top Acc	Epoch	NF Acc	Total Knowledge
Finetune(QLoRA)	0.1150	16	0.8174	0.9324
K-Adapter	0.1260	30	**0.9980**	1.1240
Mix-review(QLoRA)	0.0800	25	0.7988	0.8788
RecAdam(QLoRA)	0.1000	24	0.7933	0.8933
RHO-1(QLoRA)	0.1410	18	0.8223	0.9633
TAALM(QLoRA)	**0.4290**	**4**	0.8983	**1.3273**

	TWiki-Probes-0910			TWiki-Probes-1011			TWiki-Probes-1112		
	Un	**C**	**Avg**	**Un**	**C**	**Avg**	**Un**	**C**	**Avg**
Finetune(QLoRA)	9.999	10.057	10.028	9.554	9.531	9.543	9.736	9.632	9.684
Mix-review(QLoRA)	9.529	9.579	9.554	9.514	9.486	9.501	9.562	9.452	9.507
RecAdam(QLoRA)	9.514	9.604	9.559	8.992	9.031	9.012	9.579	9.479	9.529
RHO-1(QLoRA)	4.389	4.624	4.507	4.360	4.395	4.3775	4.471	4.717	4.594
TAALM(QLoRA)	**4.019**	**4.268**	**4.1435**	**4.030**	**4.154**	**4.092**	**4.036**	**4.357**	**4.197**