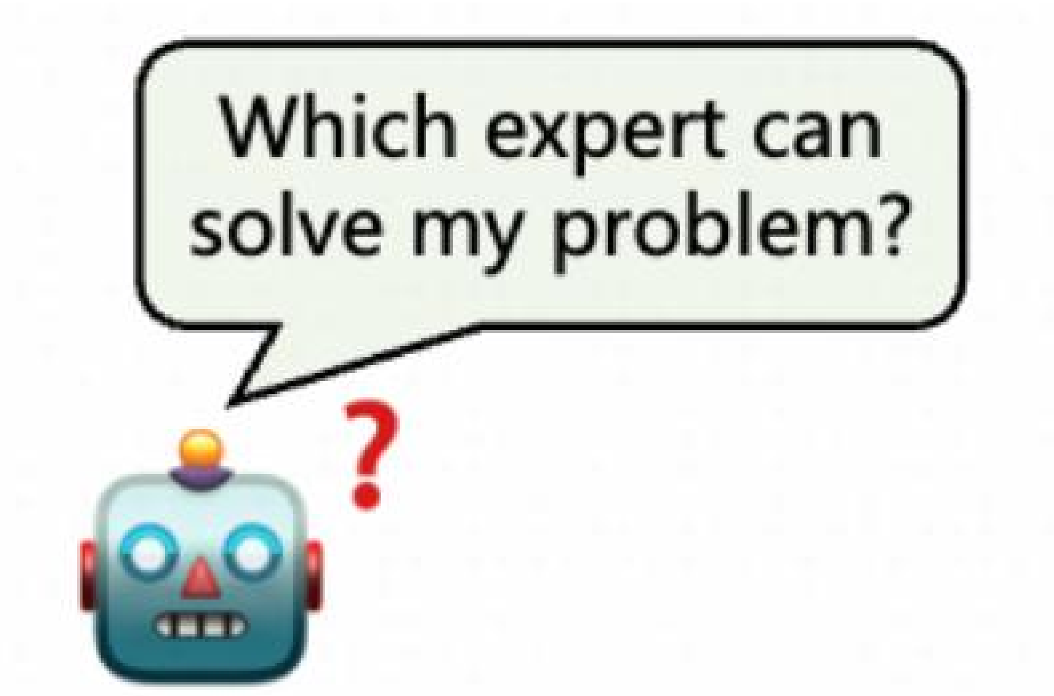
Mixture-of-LoRAs: An Efficient Multitask Tuning for Large Language Models

Wenfeng Feng, Chuzhan Hao, Yuewei Zhang, Yu Han, and Hao Wang*

{wenfeng.fwf,liyou.zyw,hy226261}@alibaba-inc.com,
chuzhanh@gmail.com,cashenry@126.com

Background

Domain specification techniques are key to make large language models (LLMs) disruptive in various applications. We often use **parameter-efficient fine-tuning** methods to learn sufficient domain knowledge, while ensuring their foundational capabilities. There are numerous LoRA modules with domain-specific capabilities in the AI community. Therefore, **how to efficiently combine** multiple professional capabilities and ensure their application under limited computing resources has become a meaningful research problem.



Outline

- Customized Capabilities Combination
- Dataset
- Model
- Experiments
  - Baselines
  - Evaluation Metrics
  - Results
- Conclusion

Customized Capabilities Combination

Efficient and effective combination of LoRA parameters for multiple tasks in one LLM

Prior Work

- Some work (*DEMIX-2022，MoE meets Instruction Tuning-2023*) trains a separate FFN expert for each task and use a method like Mixture-of-Experts (MoE) to select different experts.
- Some work (*AdapterFusion-2021，LoRAHub-2023)* directly performs parameter fusion of multitask models or adds a fusion layer.
- Some work (*LLM-Blender-2023)* is an ensemble learning method of LLM, selecting the optimal output from multiple outputs.

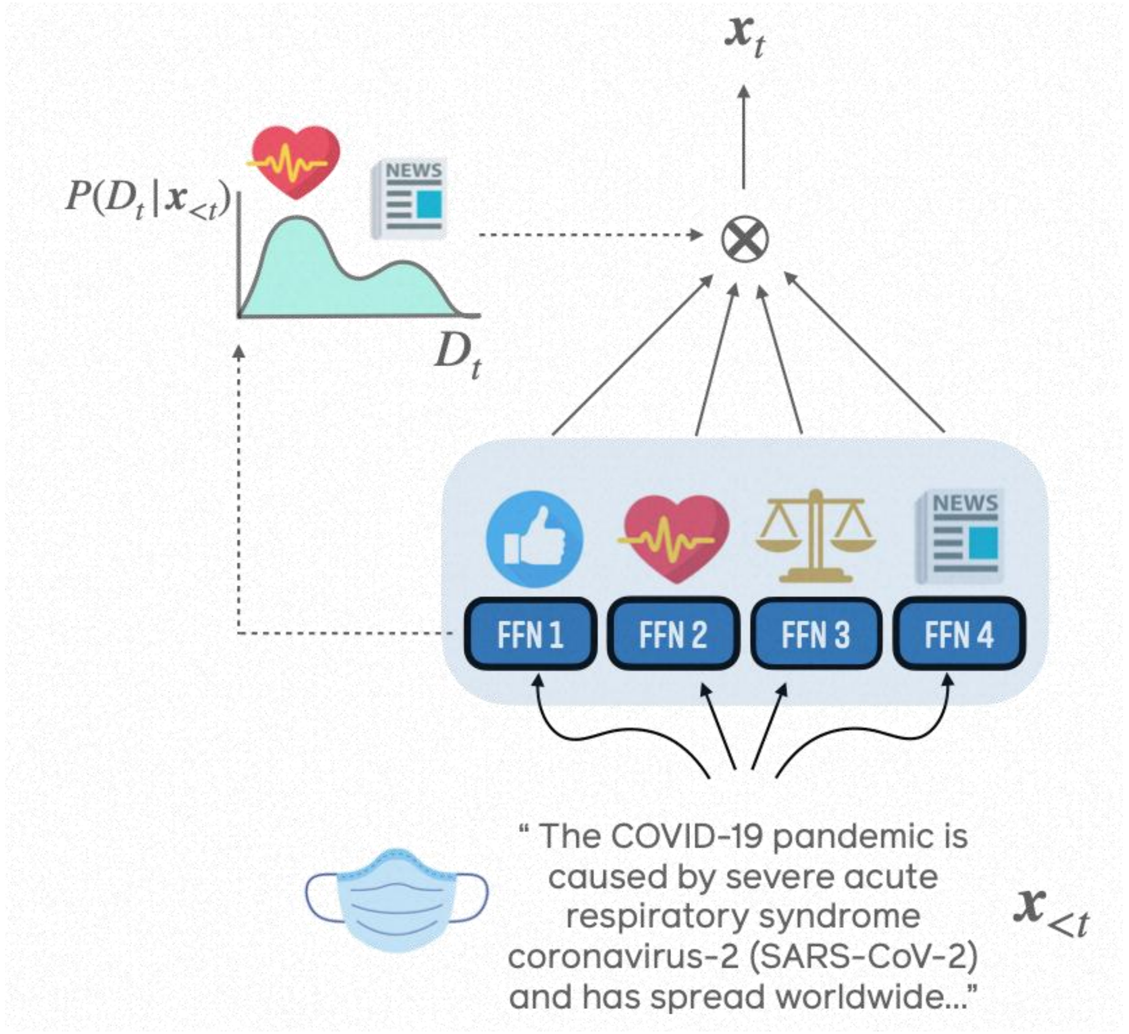


Fig.1 DEMIX

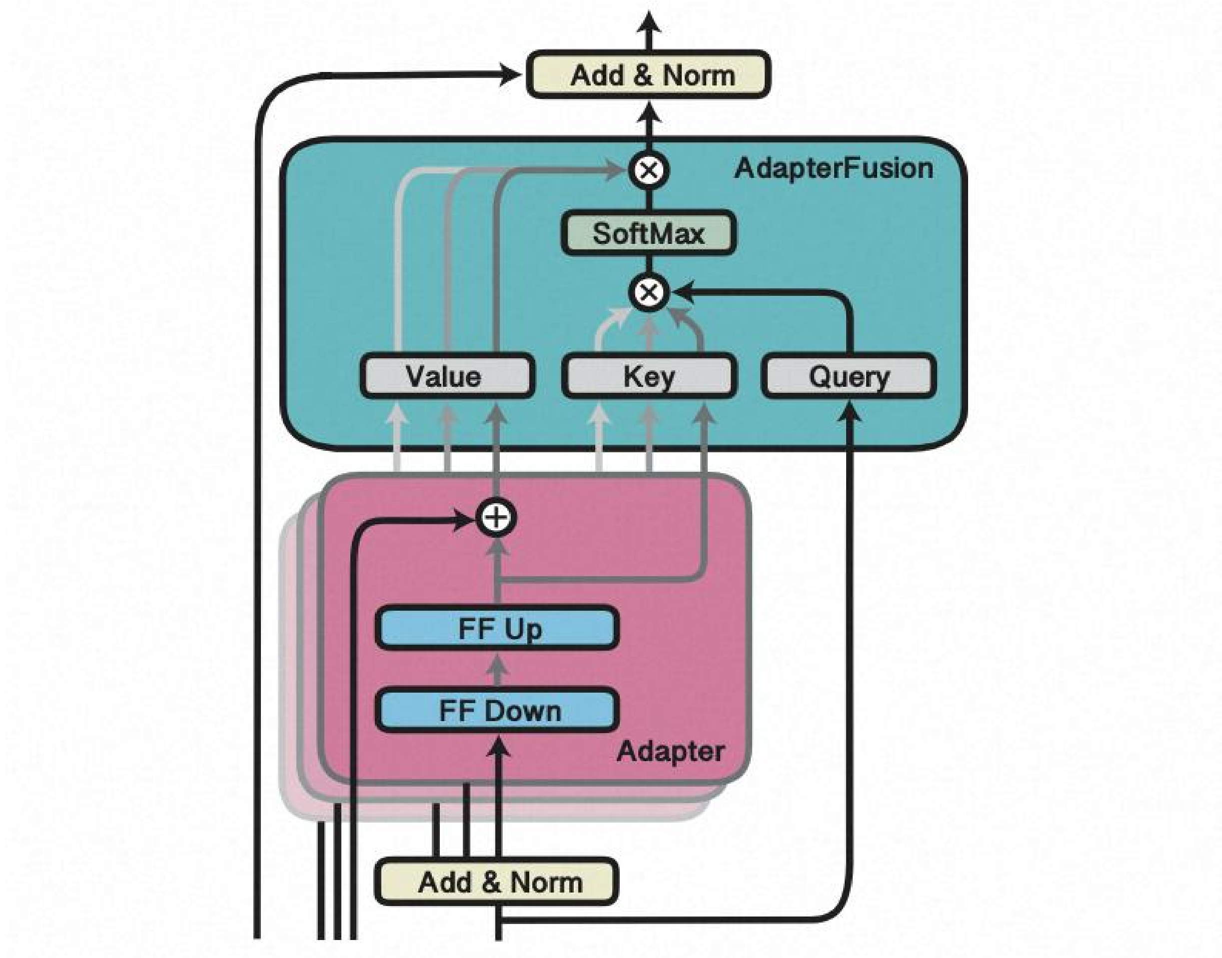


Fig.2 AdapterFusion

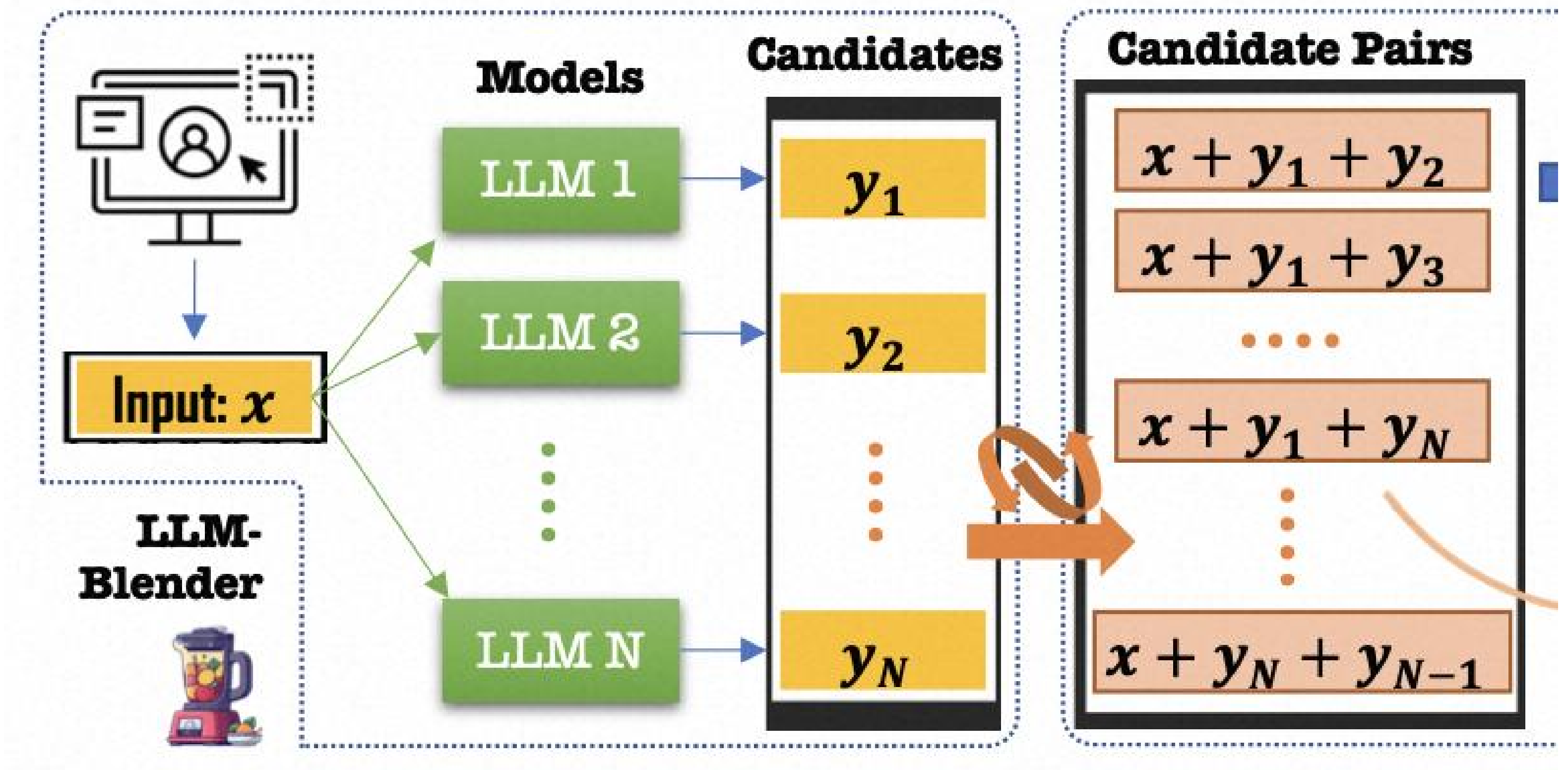


Fig.3 LLM-Blender

Customized Capabilities Combination

Efficient and effective combination of LoRA parameters for multiple tasks in one LLM

This Work

- We observe significant interference among certain tasks, while also identifying complementarity among others. Therefore, we need a comprehensive model to learn multiple professional capabilities.
    - At the task level
    - Multi-task learning
    - Routing strategy

Dataset

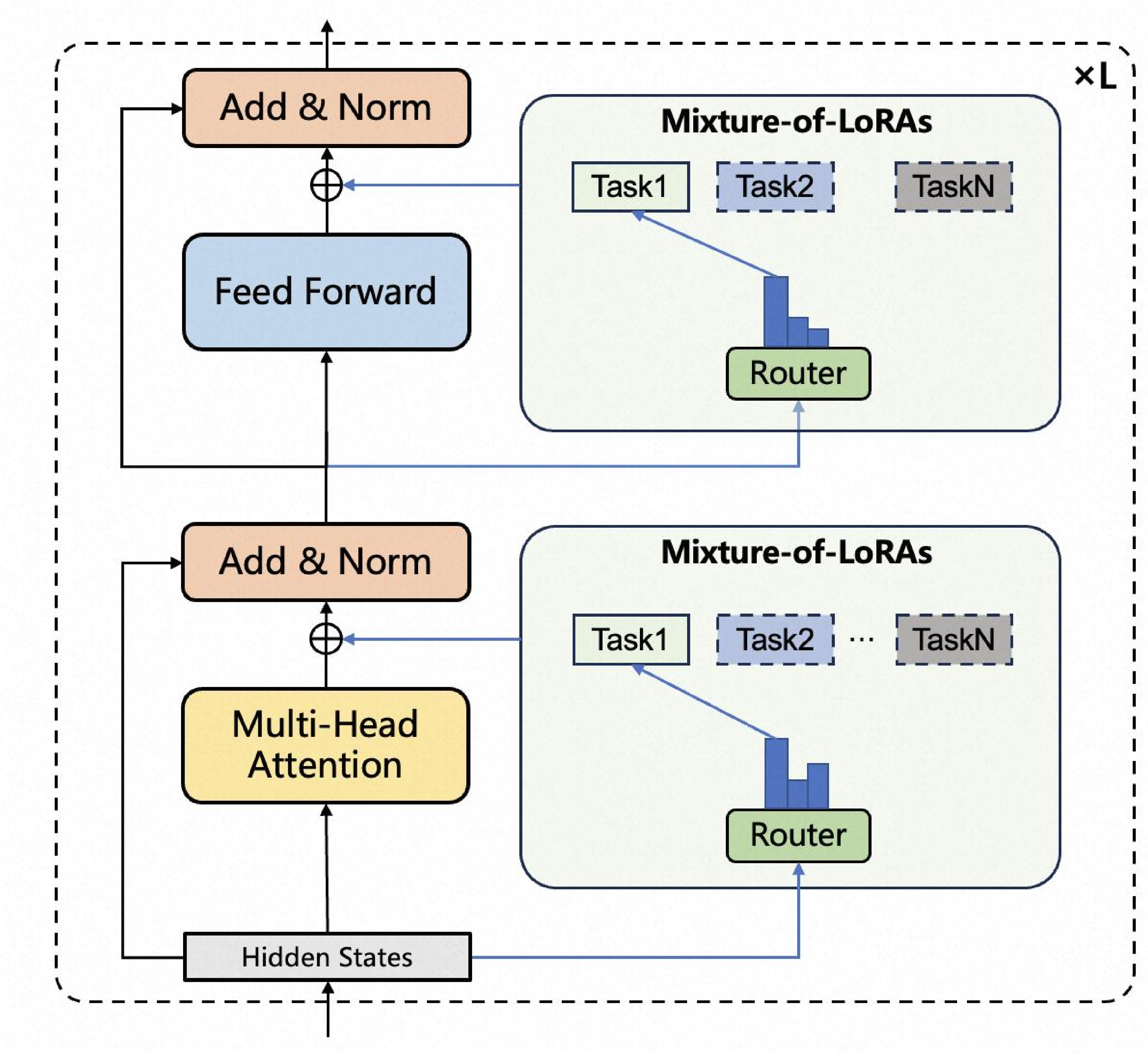
- To evaluate the effectiveness of our proposed model, we first conduct experiments on various supervised fine-tuning (SFT) datasets of heterogeneous domains.
    - *Finance*, *Medicine* and *Leetcode* belong to the specialized domain dataset.
    - *Exam*, *Webgpt* and *Gpt4tools* limit the output format of the LLM and allow the model to learn special functions.
    - Other datasets include *Chain-of-Thought, Dialog*, etc. Meanwhile, both English and Chinese are involved.

Domain	Source	Language	# Train (Eval.) Tokens
FINANCE	Financial related instructions (Qingyi Si, 2023)	EN	1.2M (0.24M)
MEDICINE	10k real conversations between patients and doctors (Li et al., 2023)	EN	1.4M (0.28M)
LEETCODE	Chinese Open Instruction Generalist (Zhang et al., 2023)	CN	9.3M (2.09M)
EXAM		CN	3.6M (0.71M)
WEBGPT	Retrieval question answering dataset (Nakano et al., 2021)	EN	7.4M (1.46M)
GPT4TOOLS	A collection of tool-related instructions (Yang et al., 2023)	EN	7.5M (1.49M)
COT	Several Chain-of-Thought datasets (Longpre et al., 2023)	EN	1.1M (0.22M)
STACKOVERFLOW	57k dialogs from StackOverFlow questions (Xu et al., 2023)	EN	0.9M (0.18M)

Tab.1 statistics of SFT datasets.

Our proposed Method

- Technical challenges:
  - Challenge 1:There is mutual interference among certain tasks, and efficient reasoning needs to be ensured.
  - Challenge 2: How to further improve the performance of a single task on existing data.



- Task-level router
- Routing both attention and FFN
- Add routing loss, the cross-entropy loss of expert selection.

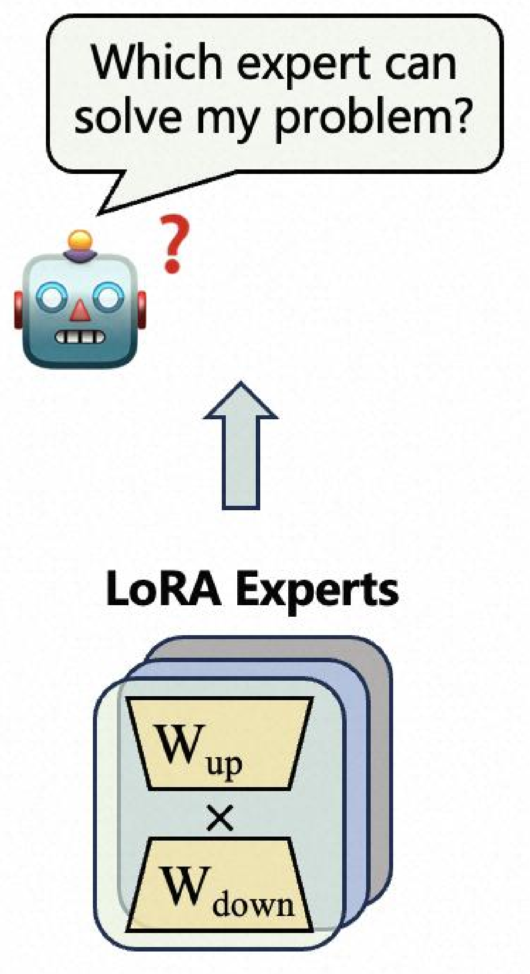


Fig.4 Mixture-of-LoRAs (ours MoA)

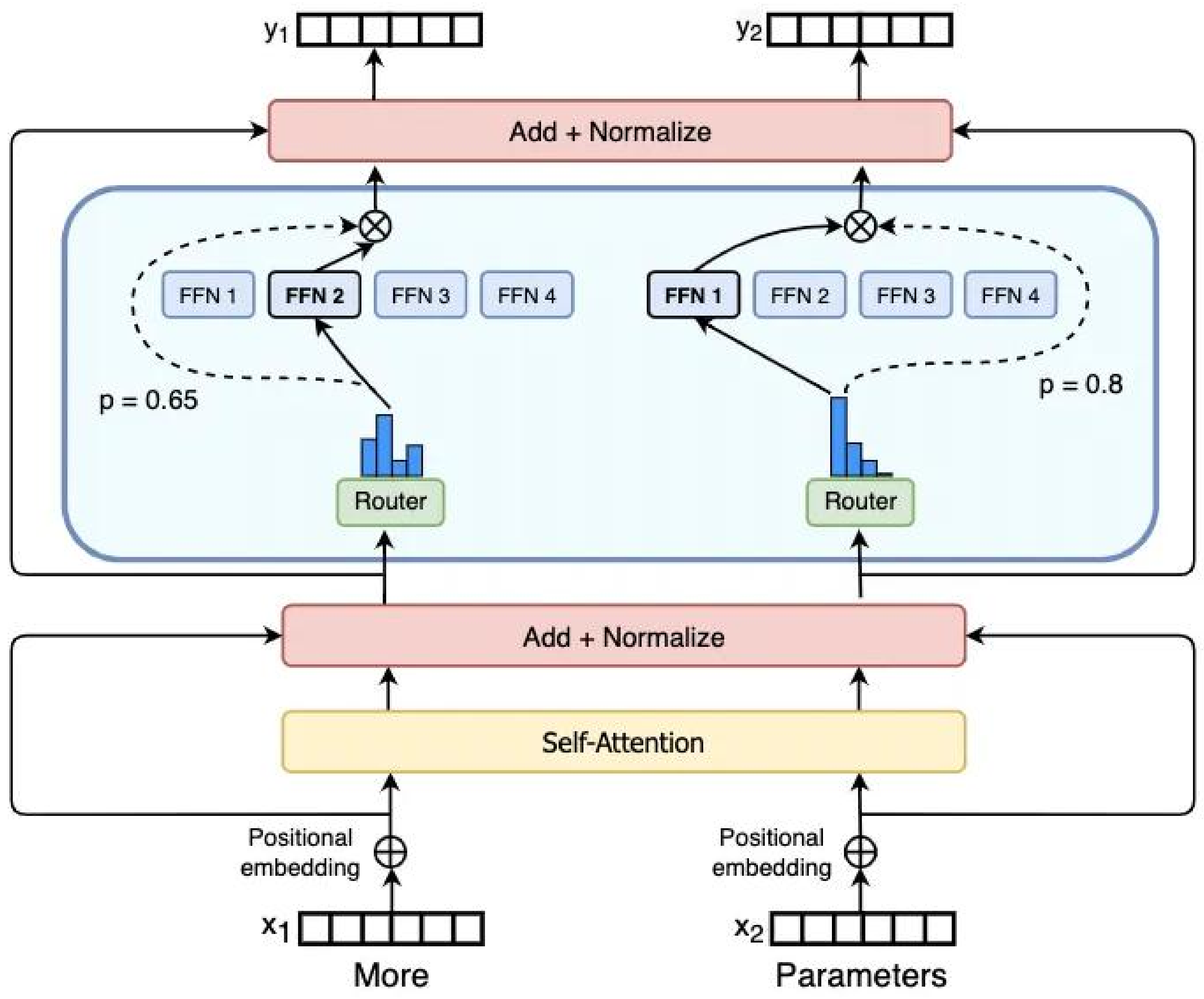


Fig.5 Traditional Mixture-of-Experts (MoE)[1]

[1] Fedus W, Zoph B, Shazeer N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity[J]. The Journal of Machine Learning Research, 2022, 23(1): 5232-5270.

Baselines and Metrics

- Compared Methods
    - Single-LoRA: a LoRA trained on data within the domain
    - Single-LoRA (mixed): a LoRA trained on the domain-mixed data
    - MoA: routing strategy at the task level and label information for multitask learning
    - MoE-LoRA: routing strategy at the token level and regard the lora module as the expert
    - MoE-LORA (naive): all LoRA modules are randomly initialized

- Evaluation Metrics
    - Common evaluation metrics of generative tasks
        - perplexity (PPL)
        - the bilingual evaluation understudy (BLUE)
        - the longest common subsequence (ROUGE-L)
    - Evaluation metrics of downstream tasks
        - the accuracy on datasets with standard answers (%)
        - the larger LLM as an evaluation expert (0-100)

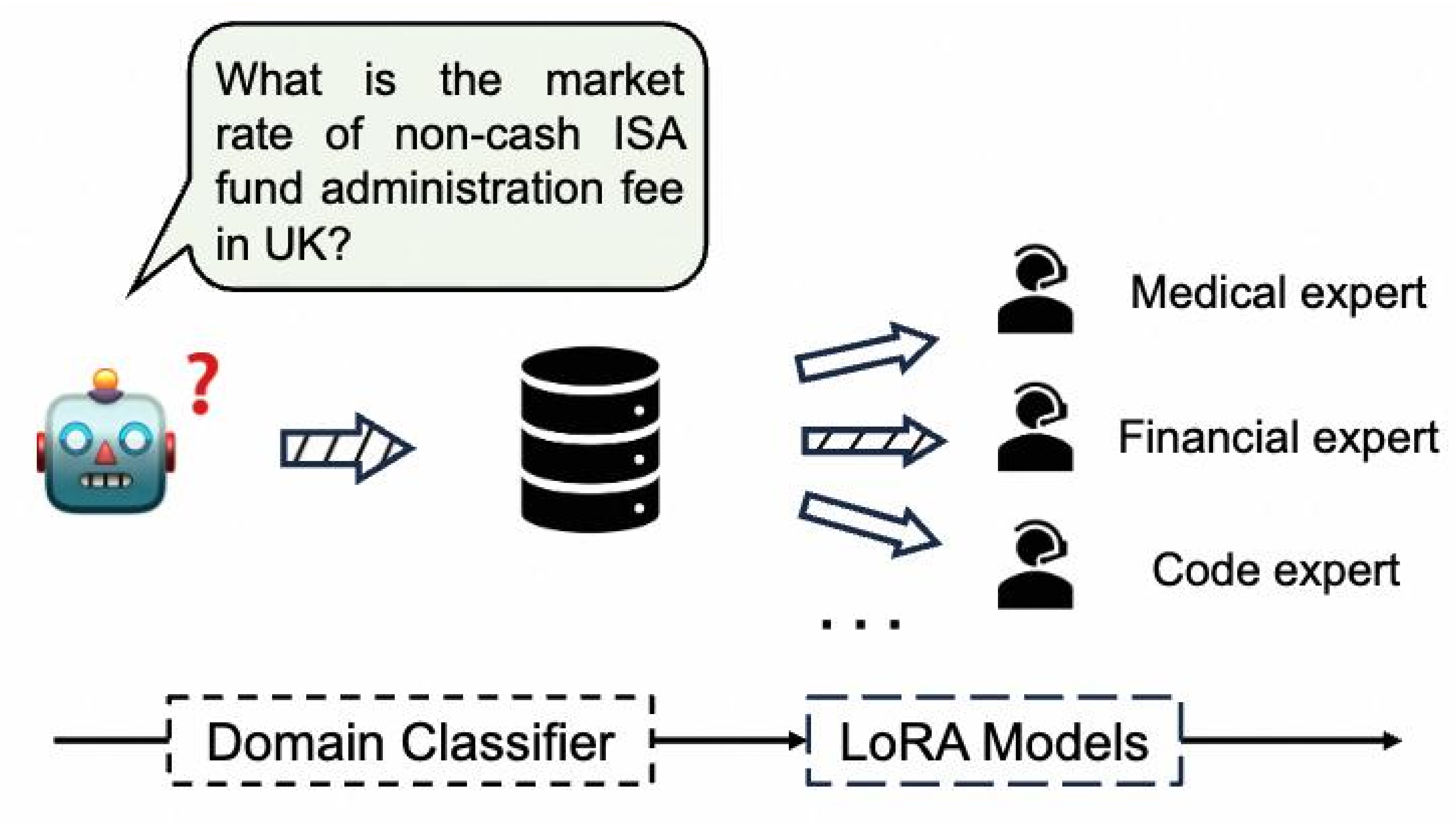
Results

Domain	Single-LoRA			Single-LoRA (mixed)			MoA		
	PPL	BLUE	ROUGE-L	PPL	BLUE	ROUGE-L	PPL	BLUE	ROUGE-L
FINANCE	7.8479	18.5975	28.6266	7.7214	**22.4846**	**32.5574**	**7.5287**	20.5774	30.6797
MEDICINE	9.5097	13.6096	18.8911	9.0499	13.5373	19.4425	**8.4561**	**13.8811**	**19.8118**
LEETCODE	1.9527	34.8582	47.8152	2.0289	35.2886	46.6290	**1.9311**	**37.4872**	**49.3256**
EXAM	3.1154	3.0871	18.5609	3.1135	4.3259	16.6206	**2.9752**	**4.7942**	**19.1840**
WEBGPT	1.7945	38.8995	41.4447	1.8484	39.6297	42.0700	**1.7933**	**40.2602**	**43.7395**
GPT4TOOLS	2.2525	64.7501	71.4391	2.2497	66.3450	73.1289	**2.2123**	**69.2596**	**74.5962**
COT	2.8126	34.5210	45.7961	2.6474	43.6290	**53.2125**	**2.5931**	**40.2529**	50.3844
S.O.	**2.8169**	19.9554	29.7282	2.9012	19.4896	28.4694	2.8999	**23.0412**	**31.9793**
Average	4.0128	28.5348	37.7877	3.9450	30.5912	39.0163	**3.7987**	**31.1942**	**39.9626**

Tab.2 In-domain test-set performance for different training strategies of LoRA.

- Conclusions
    - Training in the domain-mixed data is helpful for the overall performance, but the performance decreases on data with strict output formats.
    - Our proposed multi-task learning method can avoid interference between partial tasks.
    - The performance on most tasks can be further improved.

Classification Accuracy



Domain	test size	Classifier	Router
FINANCE	2000	98.80%	99.60%
MEDICINE	1221	99.92%	99.92%
LEETCODE	1952	99.95%	100.00%
EXAM	1999	99.95%	100.00%
WEBGPT	2000	100.00%	99.85%
GPT4TOOLS	2000	100.00%	100.00%
COT	2000	99.75%	99.95%
STACKOVERFLOW	2000	99.00%	99.90%
Average	1896.5	99.67%	**99.90%**

Tab.3 The classification accuracy of MoA router and a specific classifier by domain at inference time.

Our router strategy can select the appropriate LoRA module even more accurately than a specific classifier.

Model	LoRA	LoRA (mixed)	MoA
trainable parameters	143M	143M	143M*8+1.05M

Tab.4 The trainable parameters under different LoRA combinations.

Downstream Task

Model	Total	Right	Accuracy
Single-LoRA (mixed)	1331	515	38.69%
Single-LoRA	1331	520	39.07%
MoA	1331	593	$\boxed{\textbf{44.55\%}}$

Tab.5 The accuracy of responses on the *Exam* test dataset.

Score / Dataset / Model	Finance	Medicine	Webgpt
Single-LoRA (mixed)	**76.91**	57.49	87.92
Single-LoRA	75.99	57.11	88.59
Single-LoRA of MoA	76.30	<u>58.01</u>	<u>89.00</u>
MoA	<u>76.56</u>	**60.68**	**89.27**

Tab.6 The evaluation scoring (0-100) of the GPT-4 on the Finance, Medicine, and WebGPT datasets

- Despite the overall low accuracy due to the difficulty of the questions, the accuracy of MoA is significantly higher than the other two models (+**5.86%**, +**5.48%**).

- The performance of each LoRA module surpasses the original Single-LoRA modules in each task after multi-task learning training within MoA.

Ablation Study

Methods	PPL	BLUE	ROUGE-L
MoE-LoRA	3.8578	29.1640	37.5960
MoE-LoRA (naive)	**3.7969**	29.4170	37.3917
MoA	3.7987	**31.1942**	**39.9626**

Tab.7 The averaged test performance comparison on eight tasks.

- The MoE-LoRA does not introduce explicit domain label information in the training and inference process and guarantees the same number of parameters as MoA.
- MoA has achieved an overall improvement over MoE-LoRA, which demonstrates that the domain label information is useful for different tasks.

Domain	Single-LoRA	MoE-LoRA	MoA
FINANCE	7.8479	7.6623	**7.5235**
MEDICINE	9.5097	9.6510	**8.4488**
LEETCODE	1.9527	2.0087	**1.9296**
EXAM	3.1154	3.1455	**2.9745**
WEBGPT	1.7945	1.8080	**1.7927**
GPT4TOOLS	2.2525	2.2524	**2.2123**
COT	2.8126	2.9205	**2.5910**
S.O.	**2.8169**	2.8801	2.8968
Average	4.0128	4.0411	**3.7962**

Tab.8 The test perplexity of corresponding LoRA module in different models on each task dataset.

- Our proposed multi-LoRA joint training method can further improve the PPL performance of each LoRA, which is more flexible and effective.

Practical Tips

- Sharing the same router parameters between the LoRA in the attention layer and the LoRA in the feedforward network (FFN) layer results in a more robust performance.

- When training the MoA model, the parameters of all routers and LoRA modules are trainable, while the remaining base model parameters are frozen.

- Adjust the weight of the routing loss (i.e., the cross-entropy loss of expert classification) based on your base model and tasks.

- Having more than 5k prompts per task leads to better results. Performance degrades when the data size is smaller, such as below 2k or even just a few hundred ones.

Conclusion

- We introduce MoA architecture, which provide an efficient multi-task fine-tuning method for LLM, addressing interference among tasks and training instabilities.

- Each LoRA model can be iterated individually to quickly adapt to new domains. It can arbitrarily combine multiple domain-specific LoRA modules to implement a LLM with multiple specific capabilities. This is so flexible and efficient!

- Future work may focus on how to flexibility add or remove LoRA modules with unsupervised learning, optimize the current routing algorithm, or reduce the scale of training data in domain specialization of LLMs.