北京大学
PEKING UNIVERSITY

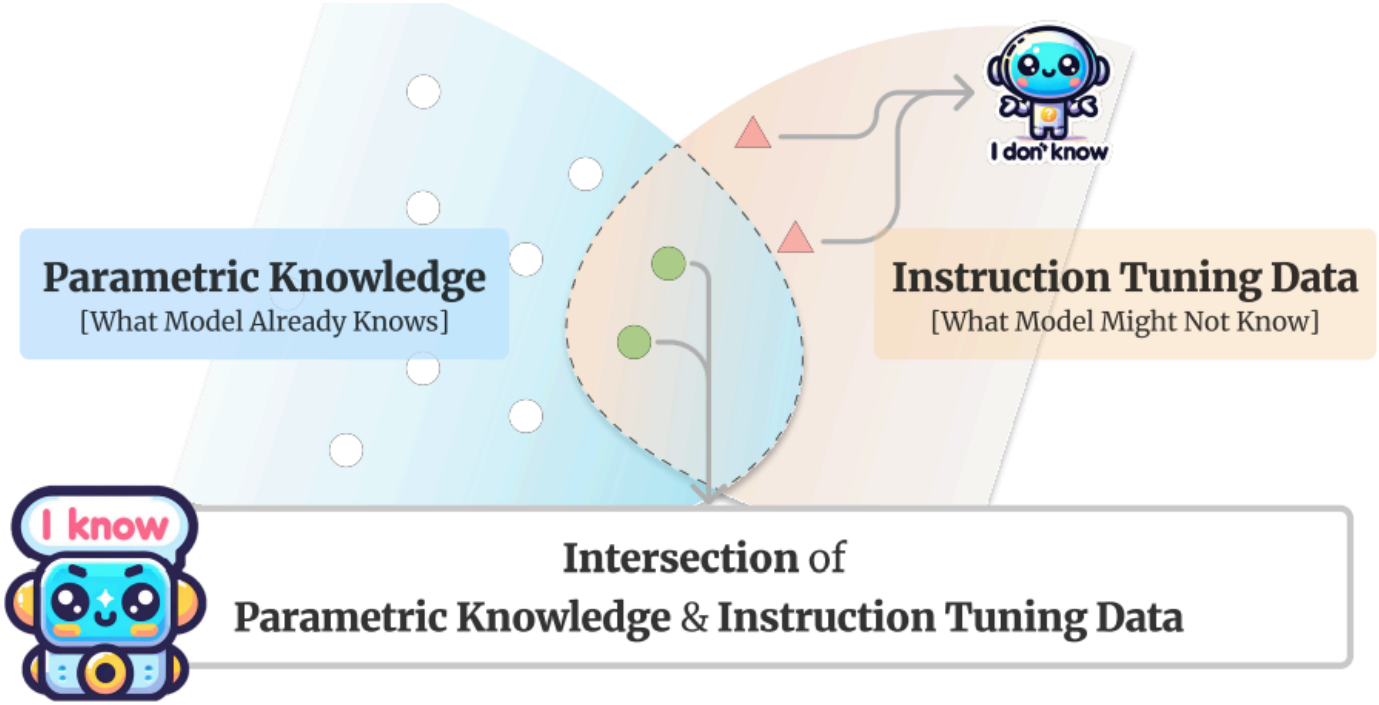
R-Tuning: Teaching Large Language Models to Refuse Unknown Questions

Jiayu Yao

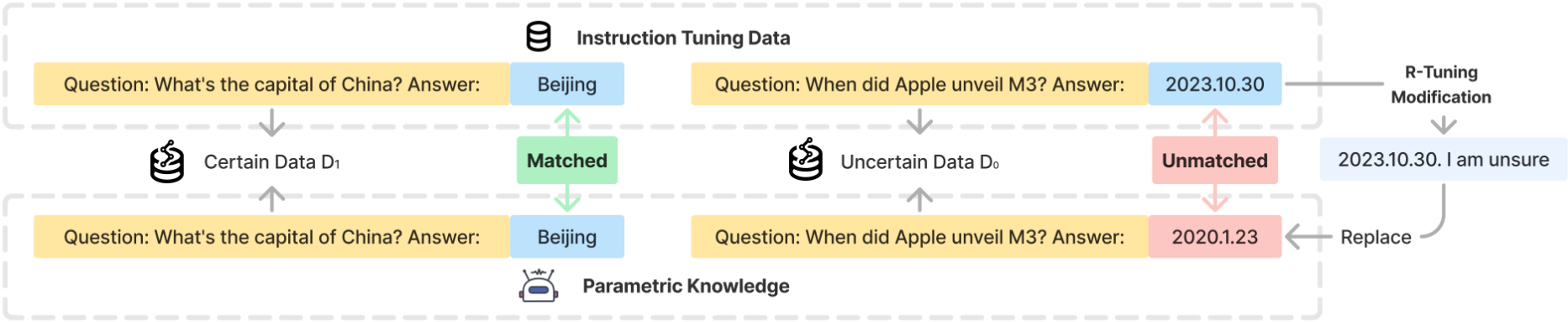
Hallucination

LLM fabricates no-existent facts

- LLM acquire almost all knowledge during pre-training
- But instruction tuning teaches models to elicit knowledge(guesting answer)
- Gap between the knowledge of human-labeled instruction tuning datasets and parametric knowledge of LLMs



Refusal-Aware Datasets



$$Q : \{\text{Question}\}, A : \{\text{Answer}\}.\{\text{Prompt}\}. \quad (1)$$

Are you sure you accurately answered the question base on your internal knowledge

- I am sure
- I am unsure

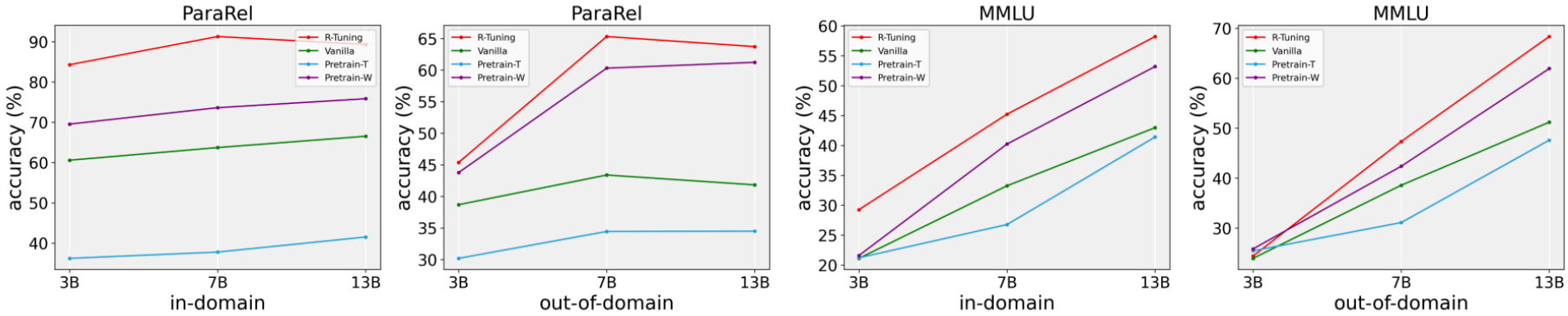
Experiments

$$AP = \sum_{k=0}^{n-1} (R(k+1) - R(k)) \times P(k), \quad (4)$$

$$P(k) = \frac{\text{\# of correct answers above k-threshold}}{\text{\# of answers above k-threshold}}, \quad (5)$$

$$R(k) = \frac{\text{\# of correct answers above k-threshold}}{\text{\# of correct answers}}. \quad (6)$$

Experiments



Dataset	Domain	Models	R-Tuning	Vanilla
ParaRel	ID	OpenLLaMA-3B	**93.23**	92.89
		LLaMA-7B	**93.64**	93.32
		LLaMA-13B	**94.44**	94.00
	OOD	OpenLLaMA-3B	**69.41**	68.42
		LLaMA-7B	74.61	**78.08**
		LLaMA-13B	**77.30**	64.12
MMLU	ID	OpenLLaMA-3B	**24.96**	24.19
		LLaMA-7B	**59.05**	58.16
		LLaMA-13B	**68.87**	51.93
	OOD	OpenLLaMA-3B	24.75	**26.08**
		LLaMA-7B	**68.69**	66.38
		LLaMA-13B	**77.41**	67.38

Table 1: Single-task experiments of R-Tuning and Vanilla on ParaRel and MMLU datasets with AP scores (%). ID and OOD denote in-domain and out-of-domain settings, respectively.

Dataset	Model	D_1	D_0
ParaRel	OpenLLaMA-3B	0.426	0.709
	LLaMA-7B	0.475	0.694
	LLaMA-13B	0.436	0.744
MMLU	OpenLLaMA-3B	0.347	0.389
	LLaMA-7B	0.330	0.400
	LLaMA-13B	0.239	0.457
WiCE	OpenLLaMA-3B	0.250	0.280
	LLaMA-7B	0.254	0.270
	LLaMA-13B	0.265	0.252
HotpotQA	OpenLLaMA-3B	0.534	0.747
	LLaMA-7B	0.605	0.719
	LLaMA-13B	0.528	0.797
FEVER	OpenLLaMA-3B	0.413	0.219
	LLaMA-7B	0.279	0.286
	LLaMA-13B	0.189	0.350

Table 7: Entropy of the training datasets. It is calculated from the frequency of every predicted answer among all predictions. A larger entropy denotes greater uncertainty of the system.

Uncertainty Learning

$$u = -\sum_{j=1}^{k} p(a_j|q) \ln p(a_j|q),$$

Domain	Model	R-Tuning	R-Tuning-U
ID	OpenLLaMA-3B	93.23	93.33
	LLaMA-7B	93.64	94.39
	LLaMA-13B	94.44	95.39
OOD	OpenLLaMA-3B	69.41	71.98
	LLaMA-7B	74.61	76.44
	LLaMA-13B	77.30	80.87

Table 4: Performance of R-Tuning-U with AP scores (%) compared with R-Tuning on the ParaRel dataset. ID and OOD denote in-domain and out-of-domain, respectively.

Thanks