Why not transform chat large language models to non-English?

Xiang GENG, Ming ZHU, Jiahuan LI, Zhejian LAI, Wei ZOU, Shuaijie SHE, Jiaxin GUO, Xiaofeng ZHAO, Yinglu LI, Yuang LI, Chang SU, Yanqing ZHAO, Xinglin LYU, Min ZHANG, Jiajun CHEN, Hao YANG, Shujian HUANG

Frontiers of Computer Science, DOI: 10.1007/s11704-025-50646-z

Problem & Method

- Problem: LLM Capability Imbalance Across Different Languages
- Method:
  - Leverage continual pre-training (CPT) to improve basic abilities needed for transfer
  - Use translation chain-of-thought (TCoT) to transfer English capability
  - Propose recovery knowledge distillation (RKD) to mitigate catastrophic forgetting

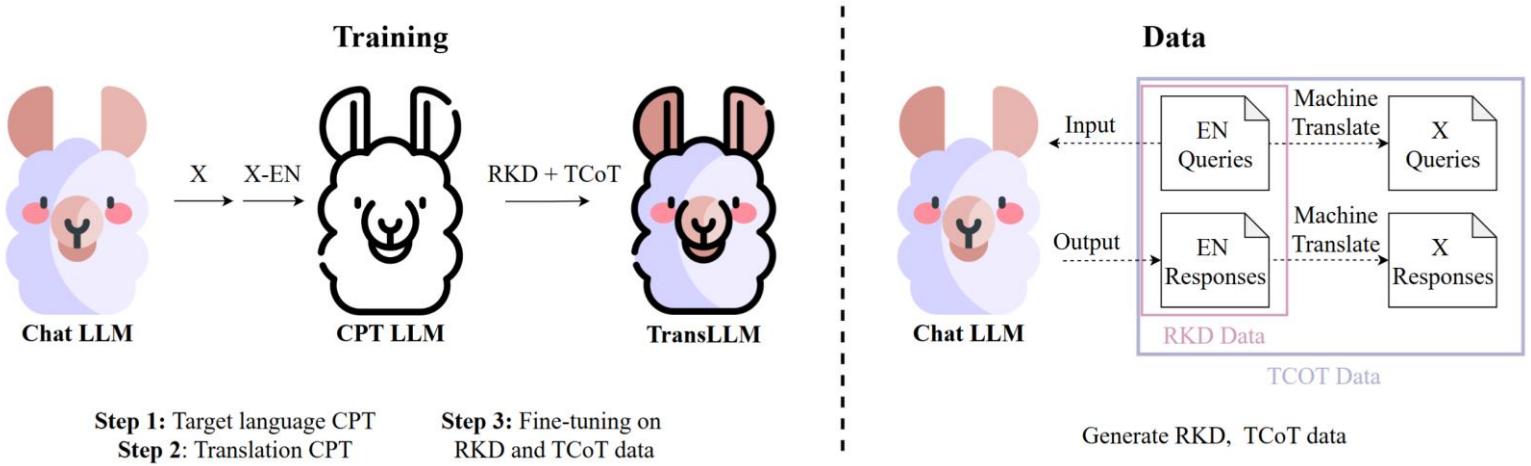

Fig. 1 The pipeline of TransLLM.

Results

- Surpass strong multilingual baselines

Comparison	MT-Bench 1st Turn		MT-Bench 2nd Turn		Aya-Test		
	TH	AR	TH	AR	PT	TE	TR
TransLLM vs. Aya-101 [7]	65.00	62.50	76.25	78.00	**60.40**	**42.80**	**58.00**
TransLLM vs. Llama-3.1† [2]	**31.25**	**32.50**	**43.75**	**35.00**	**29.60**	**36.40**	**49.60**
TransLLM vs. ChatGPT [8]	**37.50**	**32.50**	**36.25**	**30.00**	**14.80**	**40.00**	**16.00**
TransLLM vs. GPT-4 [9]	**32.50**	**-31.25**	17.50	**-43.75**	**-29.20**	**-30.80**	**-34.00**

Table 1 Comparison of TransLLM with various methods. The score is calculated as the win rate minus the loss rate.

- Knowledge is forgotten and recovered

	Model	$P(y\|x)$	Difference
1	Chat LLM	0.2363	-
2	CPT LLM	0.1666	0.0697
3	Ours w/o RKD	0.1972	0.0391
4	Ours w/o LoRA	0.1772	0.0592
5	Ours	0.2352	0.0055

Table 2 The difference of generation probabilities.