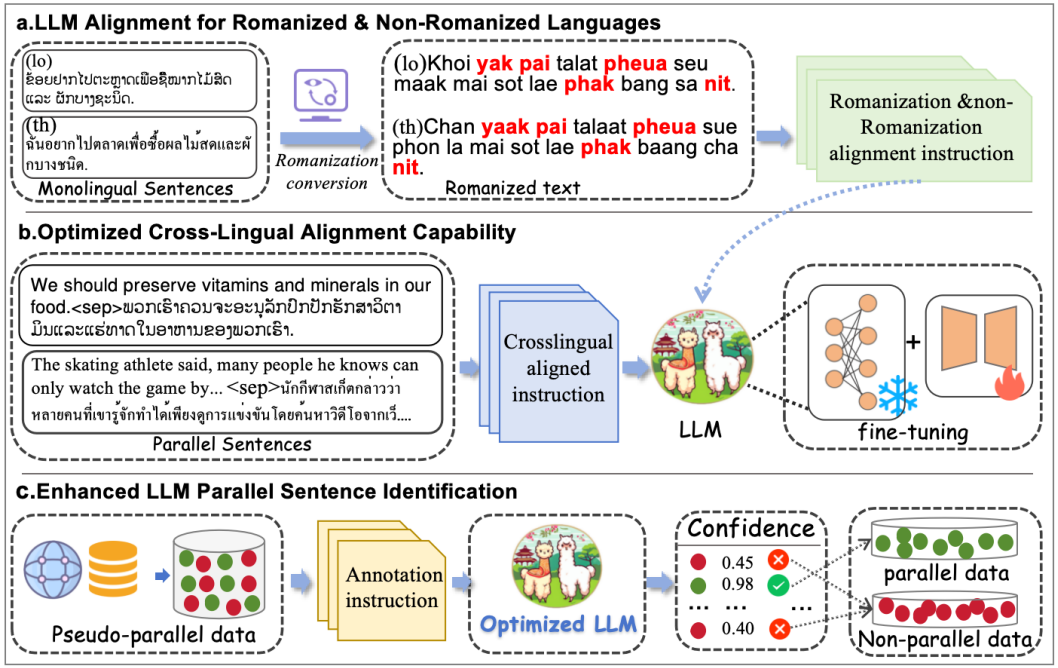
Romanization-Enhanced Large Language Models for Parallel Corpus Annotation

Siqi ZHANG, Kairong LIU, Ran SONG, Yuxin HUANG, Cunli MAO, Zhengtao YU

Frontiers of Computer Science, DOI: 10.1007/s11704-025-50704-6

Problems & Ideas

- Problems of conventional parallel data annotation methods:
  - rely on shallow statistical features and hand-crafted priors perform poorly in low-resource settings.
  - depend on high-quality bilingual embeddings also struggle to achieve good performance in low-resource scenariosn.
- Ideas: a Romanization-enhanced framework that integrates linguistic priors with multi-task joint incremental fine-tuning, improving representation unification and annotation robustness.



Proposed romanization-enhanced low-resource language parallel data annotation framework

Main Contributions

- Contributions:
  - The impact of romanization on improving Thai and Lao representation in LLMs under low-resource conditions is examined;
  - A multi-task parallel data annotation framework is introduced, integrating Romanized input standardization, English-mediated semantic bridging, and instruction-guided training to reduce cross-lingual semantic gaps and enhance alignment robustness;
  - The proposed method surpasses strong baselines in sentence annotation and yields substantial BLEU improvements for Thai–Chinese and Lao–Chinese translation.

Table 1 Results on zh-tha and zh-lao annotation task

Models	zh-tha			zh-lao		
	P/%	R/%	F1/%	P/%	R/%	F1/%
mBERT	77.48	75.46	76.45	56.43	54.37	55.38
LaBSE	93.74	86.90	90.19	94.87	91.55	93.18
E5	93.46	88.00	90.65	91.37	92.20	91.78
Llama-3-chinese-8b-instruct	100.00	81.30	89.69	83.77	47.50	60.63
Ours(Llama-3-Chinese-8B-sft)	**96.51**	**94.00**	**95.24**	**99.50**	**98.60**	**99.05**
Ours(Llama-3-Chinese-8B-Instruct-sft)	**98.42**	**93.50**	**95.90**	**99.50**	**99.00**	**99.25**

Table 2 Ablation Study on the Effects of Instruction Types

Models	zh-tha			zh-lao		
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
Ours	98.42	93.50	95.90	99.50	99.00	99.25
w/o Romanization-aligned	98.04	93.30	95.09	98.40	98.50	98.75
w/o Crosslingual-aligned	99.04	93.30	96.09	98.41	92.70	95.47
w/o Annotation	99.76	84.40	91.44	98.35	29.80	45.74
w/o Romanization-aligned + Crosslingual-aligned	98.84	93.00	95.36	99.30	98.80	99.05
w/o Romanization-aligned + Annotation	93.13	86.80	89.86	86.00	55.30	67.32
w/o Crosslingual-aligned + Annotation	64.95	75.80	69.69	50.83	89.00	64.70