EMNLP 2025 Suzhou, China | 中国苏州

From Unaligned to Aligned: Scaling Multilingual LLMs with Multi-Way Parallel Corpora

Yingli Shen[1*], Wen Lai[2,3*], Shuo Wang[1], Ge Gao[4]
Kangyang Luo[1], Alexander Fraser[2,3], Maosong Sun[1,5†]

[1] Department of Computer Science and Technology, Tsinghua University
[2] Technical University of Munich [3] Munich Center for Machine Learning
[4] Minzu University of China [5] Institute for AI, Tsinghua University
syl@mail.tsinghua.edu.cn, wen.lai@tum.de

Munich Center for Machine Learning

Background

- **Task Definition:** ***Multi-way parallel data***, where identical content is aligned across multiple languages, provides stronger cross-lingual consistency and offers greater potential for improving multilingual performance.
- **Motivation:**
  - Existing multi-way parallel datasets typically cover only a limited number of languages, domains and levels of parallelism.
  - Scaling multilingual LLMs using multi-way parallel data remains under explored.

Research Question

- **RQ1:** How does fine-tuning on multi-way parallel data compare to training on unaligned multilingual text in terms of zero-shot cross-lingual transfer and representation alignment?
- **RQ2:** Which strategies for selecting parallelism in multi-way parallel data (e.g., degree of parallelism and language subsets) lead to the greatest improvements in multilingual LLM performance?
- **RQ3:** Which instruction-tuning objectives can most effectively leverage the advantages of multi- way parallel data?

Contributions

(i) We construct **TED2025**, a 50-way parallel corpus derived from recent TED talk transcripts, covering 113 languages and 352 domains.

(ii) We present a systematic comparison of multilingual LLM fine-tuning using multi-way versus unaligned data, analyzing their effects on zero-shot transfer and cross-lingual representation alignment.

(iii) We explore instruction-tuning objectives specifically designed for multi-way parallel data and provide practical recommendations for optimizing multilingual LLM performance.

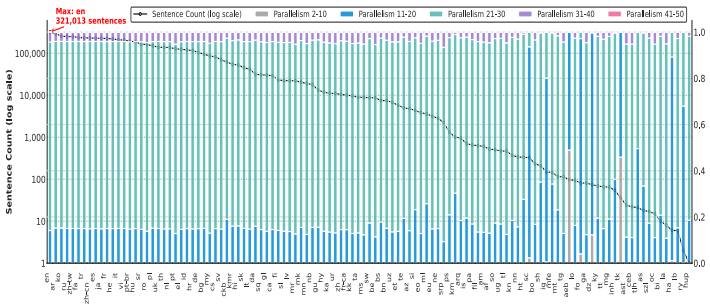

Fig. Distribution of sentence counts and parallelism spans across languages in the **TED2025**

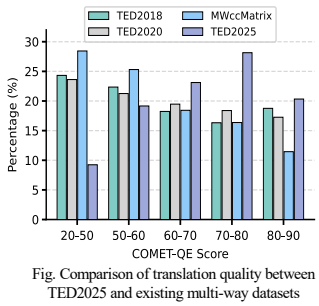

Fig. Comparison of translation quality between TED2025 and existing multi-way datasets

Impact Factors

Degree of Parallelism

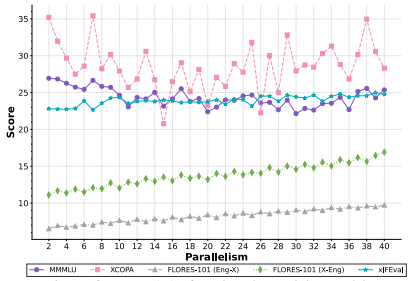

Fig. Performance (%) of continued pretraining models on downstream tasks with varying degrees of parallelism

English as Pivot

		MMMLU	XCOPA	FLORES (Eng-X)	FLORES (X-Eng)	xIFEval
Group 1	with	**23.17**	**30.93**	5.80	13.75	23.39
	w/o	20.84	30.07	**6.63**	**14.56**	**24.14**
Group 2	with	**22.19**	**35.33**	6.74	13.88	**22.19**
	w/o	18.51	31.60	**7.13**	**14.18**	**22.19**
Group 3	with	**26.47**	**36.73**	**6.40**	13.48	22.69
	w/o	23.83	35.78	6.38	**14.00**	**22.91**
Group 4	with	**23.42**	**33.84**	6.20	12.11	23.66
	w/o	20.26	30.84	**6.67**	**14.76**	**23.67**
Group 5	with	**23.89**	**39.64**	6.96	13.07	23.19
	w/o	20.39	34.15	**7.83**	**14.79**	**23.85**

Tab. Performance(%) comparison of models with and without (w/o) English across five different language groupings

Language Combinations

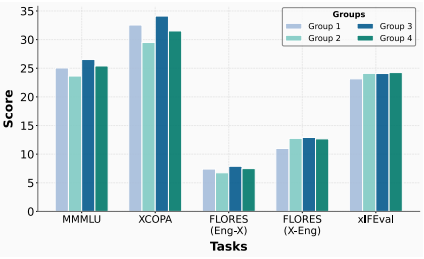

Fig. Impact of language family composition on model performance

Effectiveness of Multi-Way Corpora

Downstream Performance

	MMMLU		XCOPA		FLORES-101 (Eng-X)				FLORES-101 (X-Eng)				xIFEval	
					BLEU		COMET		BLEU		COMET			
	low	high	low	high	low	high	low	high	low	high	low	high	low	high
Baseline	18.27	33.72	23.46	34.29	6.03	11.67	57.15	61.03	13.37	22.49	75.24	82.32	17.14	24.43
Unaligned	19.64	36.26	24.62	34.76	6.12	11.78	57.51	62.11	13.84	22.74	75.82	82.58	17.28	24.44
Multi-Way	**22.48**	**41.38**	**27.58**	**57.22**	**6.32**	**12.08**	**58.06**	**67.44**	**14.45**	**25.03**	**76.25**	**86.43**	**18.79**	**27.41**

(a) LLaMA-3-8B

	MMMLU		XCOPA		FLORES-101 (Eng-X)				FLORES-101 (X-Eng)				xIFEval	
					BLEU		COMET		BLEU		COMET			
	low	high	low	high	low	high	low	high	low	high	low	high	low	high
Baseline	35.24	49.55	62.25	72.00	7.45	11.05	57.22	67.16	16.54	20.24	67.23	74.29	27.63	32.40
Unaligned	35.61	51.32	62.59	74.06	7.62	11.60	57.85	70.85	16.86	21.02	67.61	75.97	27.92	35.54
Multi-Way	**36.64**	**55.81**	**63.24**	**79.52**	**8.07**	**13.11**	**58.94**	**80.56**	**17.36**	**23.26**	**68.59**	**81.33**	**28.64**	**40.95**

(b) Qwen-2.5-14B

Tab. Performance (%) comparison of different models across multilingual benchmarks.

Cross-Lingual Representation Alignment

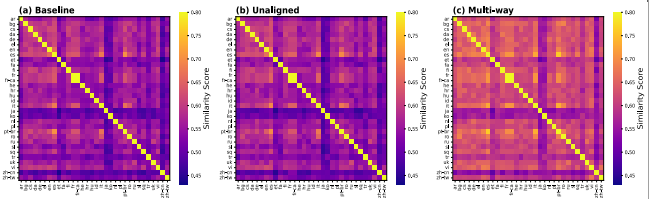

Fig. SVCCA alignment comparison between the Multi-Way, Unaligned and Baseline models across32-way language pairs

Cross-Domain Generalization

Downstream Performance

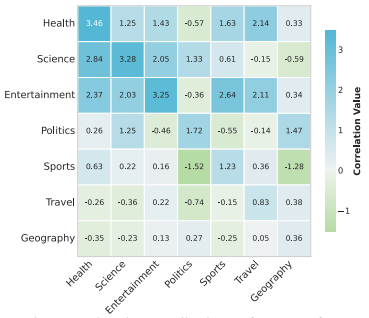

Fig. Cross-domain generalization performance of instruction-tuned models using multi-way parallel data

Training Data Size

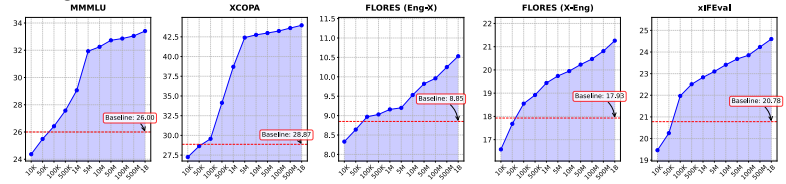

Fig. Impact of training data size on model performance across different token amounts