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

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

[1]Tsinghua University, [2]Technical University of Munich, [3]Minzu University of China

Background

- LLMs excel in high-resource languages but lag in low-resource languages.
- Approach to improve performance on low-resource languages:
  - Continue Pretraining / Instruction Tuning using more multilingual data.
- Two types of multilingual data:
  - Unaligned multilingual data
  - Aligned (parallel / multi-way parallel) data

Multi-way parallel data is underexplored for scaling multilingual LLMs.

Existing Multi-Way Datasets

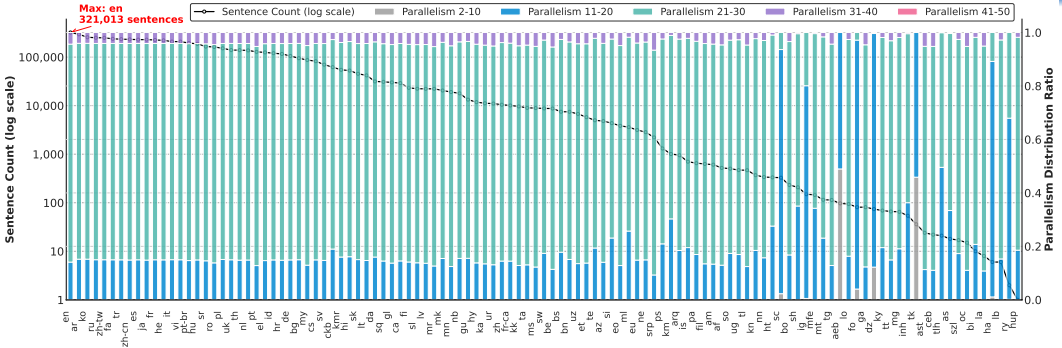
Dataset	Source	#Langs	#Domains	Max Parallelism	Type	Collection Method	Date Range	Open-source?
Bible	Bible	100	1	55	Training	Human Translator	2015	✓
UN Corpus	UN	6	1	6	Training	Human Translator	2016	✓
MMCR4NLP	Mix	59	5	13	Training	Mix Collection	2017	✗
GCP Corpus	Speech	10	4	10	Training	Machine Translation	2018	✗
FLORES-101	Wiki	101	10	101	Evaluation	Human Translator	2022	✓
FLORES-200	Wiki	204	10	204	Evaluation	Human Translator	2022	✓
BPCC	Many	22	13	22	Training	Mix Collection	2023	✓
XDailyDialog	DailyDialog	4	/	4	Training	Human Translator	2023	✓
MWccMatrix	Common Crawl	90	/	/	Training	Crawl	2024	✓
TED2018	TED	58	/	58	Training	Human Translator	2018	✓
TED2020	TED	108	/	/	Training	Human Translator	2020	✓
Ours (TED2025)	TED	113	352	50	Training	Human Translator	2025	✓

- Existing datasets: limited languages, domains, and parallelism.
- The largest TED-based datasets have not been updated since 2020.

Research Goal & Questions

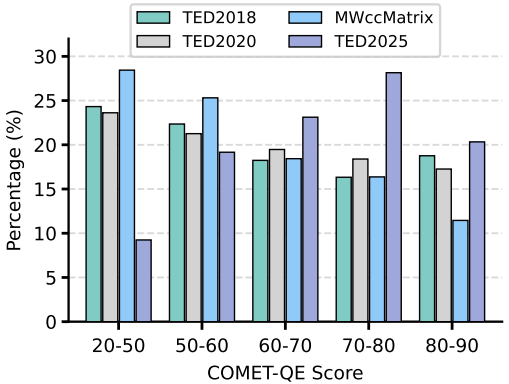
- **Research Goal:**
  - Construct TED2025 (large-scale multi-way corpus) and explore best practices.
- **Research Question:**
  - **RQ1:** Multi-way vs. unaligned → downstream task & cross-lingual transfer & representation alignment.
  - **RQ2:** Best practice on continued pretraining → parallelism & English as Pivot & Language Combinations & Training Data Size.
  - **RQ3:** Best practice on instruction tuning → different objectives.

Parallelism Distribution



- Contains 113 languages with up to 50-way parallelism.
- Most common parallelism: 21–30 languages.
- Low-Resource languages also achieve high parallelism.

Data Quality



- **Existing Dataset:**
  - TED2018 (Qi et al., 2018)
  - TED2020 (Reimers and Gurevych, 2020)
  - MWccMatrix (Thompson et al., 2024)
- **Evaluation Metric:**
  - COMET-QE (Rei et al., 2020)

- TED2025 has significantly more high-quality translations (COMET-QE > 60).
- COMET-QE score for all language pairs can be found in Appendix.

Setup

- **Datasets:**
  - Multi-Way aligned dataset: TED2025
  - Unaligned dataset: DCAD-2000 (Shen et al., 2025[a])
- **Models:**
  - LLaMA-3.1-8B and Qwen-2.5-14B
- **Training:**
  - Continued pretraining (RQ1 & RQ2) / Instruction tuning (RQ3)
  - LoRA instead Full Parameter Training

[a]DCAD-2000: A Multilingual Dataset across 2000+ Languages with Data Cleaning as Anomaly Detection (Shen et al., NeurIPS 2025)

- **Evaluation Benchmarks**

Benchmark	Task	#Langs	Metric
MMMLU	Understanding	14	Acc
XCOPA	Reasoning	11	Acc
FLORES-101	Generation	101	BLEU/COMET
FLORES-200	Generation	204	BLEU/COMET
xIFEval	Instruction Following	17	Acc
SIB-200	Text Classification	204	Acc

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**

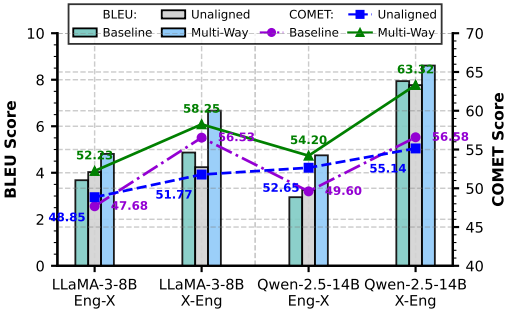
(a) LLaMA-3-8B

	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

- Multi-Way > Unaligned > Baseline (low- & high-resource).
- non-English-centric tasks are provided in Appendix C.

Zero-Shot Transfer



- Evaluation:
  - Machine Translation on FLORES-200.
  - We exclude all languages in the evaluation subset from training and assess the English-X translation quality.

- Multi-Way achieves stronger zero-shot transfer on FLORES-200.

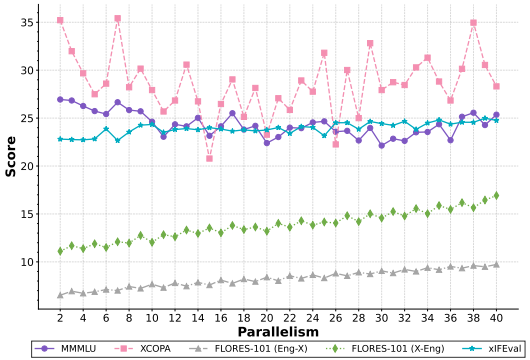
Representation Alignment

- We randomly select 32 aligned languages (with 100 sentences each) from TED2025
- Four similarity evaluation metric on embedding space:
  - Cosine Similarity
  - Centered Kernel Alignment (CKA; Kornblith et al., 2019)
  - Cross-Lingual Sentence Retrieval (Retrieval; Conneau et al., 2017)
  - Singular Vector Canonical Correlation Analysis (SVCCA; Raghu et al., 2017)

		LLaMA-3-8B			Qwen-2.5-14B		
		Baseline	Unaligned	Multi-Way	Baseline	Unaligned	Multi-Way
Cosine (↑)		0.27	$0.27_{-0.00}$	$\mathbf{0.30}_{+0.03}$	0.29	$0.27_{-0.02}$	$\mathbf{0.32}_{+0.03}$
CKA (↑)		0.54	$0.54_{-0.00}$	$\mathbf{0.60}_{+0.06}$	0.56	$0.57_{+0.01}$	$\mathbf{0.63}_{+0.07}$
Retrieval (↑)	P@1	0.09	$0.07_{-0.02}$	$\mathbf{0.13}_{+0.04}$	0.12	$0.14_{+0.02}$	$\mathbf{0.19}_{+0.07}$
	P@5	0.24	$0.23_{-0.01}$	$\mathbf{0.27}_{+0.03}$	0.26	$0.28_{+0.02}$	$\mathbf{0.33}_{+0.07}$
	P@10	0.35	$0.33_{-0.02}$	$\mathbf{0.39}_{+0.04}$	0.36	$0.38_{+0.02}$	$\mathbf{0.42}_{+0.06}$
SVCCA (↑)		0.55	$0.55_{-0.00}$	$\mathbf{0.61}_{+0.06}$	0.57	$0.58_{+0.01}$	$\mathbf{0.63}_{+0.06}$

- Multi-Way shows denser, language-agnostic embeddings.

Degree of Parallelism



- Experiment Setups
  - We fix the tokens to 5M in each parallelism.
- Evaluation

Benchmark	Task	#Langs	Metric
MMMLU	Understanding	14	Acc
XCOPA	Reasoning	11	Acc
FLORES-101	Generation	101	BLEU/COMET
FLORES-200	Generation	204	BLEU/COMET
xIFEval	Instruction Following	17	Acc
SIB-200	Text Classification	204	Acc

- MT benefits from higher parallelism.
- Reasoning tasks peak at 6–10 languages.

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**

- Five groups, each with both English included and English-excluded variants across ten languages.

English Pivoting Combinations	
Group	**Language List**
group 1	en,vi,ar,bg,de,es,fr,he,it,ja,ko
group 2	en,nl,pl,pt-br,zh-cn,ar,bg,de,es,fr,he
group 3	en,el,vi,ar,bg,de,es,fr,he,it,ja
group 4	en,el,ar,bg,de,es,fr,he,it,ja,ko
group 5	en,fa,hu,ar,bg,de,es,fr,he,it,ja

- English helps reasoning/understanding.
- English slightly hurts direct MT transfer.

Other Factors

- **Language Combinations**
  - Investigating an alternative sampling strategy: whether the selected languages belong to the same language family.
- **Training Data Size**
  - We conduct experiments by randomly sampling varying amounts of tokens—10K, 50K, 100K, 500K, 1M, 5M, 10M, 50M, 100M, 500M, and 1B—from the constructed TED2025 dataset.

- More details can be found in Section 5.3 and 5.4.

Objectives

- Four training objectives

Task	Prompt
Machine Translation (MT)	Translate the following {src_lang_1}, {src_lang_2}, ... ,{src_lang_m} sentence to {tgt_lang_1}, {tgt_lang_2}, ..., {tgt_lang_n}.\n {src_lang_1} Sentence: {src_txt_1}.\n {src_lang_2} Sentence: {src_txt_2}.\n ... {src_lang_m} Sentence: {src_txt_m}.\n Translation:\n {tgt_lang_1} Sentence: {tgt_txt_1}.\n {tgt_lang_2} Sentence: {tgt_txt_2}.\n ... {tgt_lang_n} Sentence: {tgt_txt_n}.\n
Cross-Lingual Text Similarity (CLTS)	Given the sentences below in different languages, rate how similar their meanings are on a scale of 0 to 1, where 0 means completely dissimilar and 1 means identical meanings.\n {lang_1} Sentence: {txt_1}.\n {lang_2} Sentence: {txt_2}.\n ... {lang_m} Sentence: {txt_m}.\n Similarity: {sim_score}.
Multilingual Text Classification (MTC)	Classify the following sentence in {lang_1}, {lang_2}, ..., {lang_m} into one of the following categories: {domain_list}.\n {lang_1} Sentence: {txt_1}.\n {lang_2} Sentence: {txt_2}.\n ... {lang_m} Sentence: {txt_m}.\n Categories: {target_domain}.
Cross-Lingual Paraphrasing (CLP)	Paraphrase the following {src_lang} sentence in {tgt_lang}.\n {src_lang} Sentence: {src_txt}.\n Paraphrasing:\n {tgt_lang} Sentence: {tgt_txt}.

Instruction Tuning Objectives

	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	41.96	46.68	64.59	66.79	5.51	10.50	56.95	60.02	14.98	18.41	68.29	73.51	35.99	38.56
MT	**45.28**	**51.06**	**68.17**	**69.38**	**11.26**	**13.25**	**63.03**	**65.61**	**22.25**	**21.60**	**70.76**	**75.41**	**41.72**	**44.52**
CLTS	39.99	45.44	62.87	66.47	3.63	9.90	56.48	59.28	13.25	16.60	66.89	73.15	34.91	38.38
MTC	40.68	44.49	64.19	65.16	5.02	9.08	56.04	57.84	14.38	18.16	67.86	73.12	34.03	38.01
CLP	42.49	47.01	65.41	68.42	6.38	11.46	57.23	61.28	15.76	19.27	69.87	73.91	36.17	40.07
MT + CLTS	42.23	47.94	65.00	**68.77**	6.19	10.51	**58.53**	60.33	16.83	18.98	68.70	**74.91**	36.47	**40.30**
MT + MTC	42.69	47.54	65.12	66.83	6.35	10.73	57.53	60.24	15.13	18.45	**69.28**	74.00	36.10	39.03
MT + CLP	43.20	**49.07**	**67.29**	68.47	**7.31**	10.51	58.40	**62.64**	**17.75**	**20.77**	69.18	74.47	**38.49**	39.67
CLTS + MTC	41.78	45.17	63.21	65.62	5.38	9.77	55.19	58.21	14.67	16.59	67.45	72.35	34.57	36.63
CLTS + CLP	**42.82**	46.74	65.30	67.12	5.58	**10.88**	57.84	60.41	15.41	19.33	68.84	73.58	36.99	38.83
MTC + CLP	42.62	46.70	65.16	67.56	6.48	10.84	57.74	60.98	15.29	19.22	68.76	73.77	36.87	38.96
MT + CLTS + MTC	41.03	46.59	64.15	66.05	5.14	10.35	56.14	59.71	14.64	18.06	67.64	73.39	35.24	38.14
MT + CLTS + CLP	**42.72**	**47.67**	64.83	67.57	**6.17**	10.94	57.78	60.11	**15.96**	**19.16**	68.67	**74.03**	**36.54**	38.77
MT + MTC + CLP	42.33	46.72	**65.26**	**67.58**	5.92	**10.98**	**57.93**	**60.76**	15.38	18.81	**68.98**	73.55	36.22	**39.15**
CLTS + MTC + CLP	41.56	46.65	64.16	66.45	4.67	10.38	56.20	59.63	14.95	17.62	68.00	73.03	35.27	38.21
MT + CLTS + MTC + CLP	43.07	47.67	66.64	67.38	7.59	12.42	57.75	60.79	17.62	19.62	71.13	75.24	36.95	40.52

- The improvements in MT are the largest and most stable in both high-resource and low-resource languages.

Instruction Tuning Objectives

	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	41.96	46.68	64.59	66.79	5.51	10.50	56.95	60.02	14.98	18.41	68.29	73.51	35.99	38.56
MT	**45.28**	**51.06**	**68.17**	**69.38**	**11.26**	**13.25**	**63.03**	**65.61**	**22.25**	**21.60**	**70.76**	**75.41**	**41.72**	**44.52**
CLTS	39.99	45.44	62.87	66.47	3.63	9.90	56.48	59.28	13.25	16.60	66.89	73.15	34.91	38.38
MTC	40.68	44.49	64.19	65.16	5.02	9.08	56.04	57.84	14.38	18.16	67.86	73.12	34.03	38.01
CLP	42.49	47.01	65.41	68.42	6.38	11.46	57.23	61.28	15.76	19.27	69.87	73.91	36.17	40.07
MT + CLTS	42.23	47.94	65.00	**68.77**	6.19	10.51	**58.53**	60.33	16.83	18.98	68.70	**74.91**	36.47	**40.30**
MT + MTC	42.69	47.54	65.12	66.83	6.35	10.73	57.53	60.24	15.13	18.45	**69.28**	74.00	36.10	39.03
MT + CLP	43.20	**49.07**	**67.29**	68.47	**7.31**	10.51	58.40	**62.64**	**17.75**	**20.77**	69.18	74.47	**38.49**	39.67
CLTS + MTC	41.78	45.17	63.21	65.62	5.38	9.77	55.19	58.21	14.67	16.59	67.45	72.35	34.57	36.63
CLTS + CLP	**42.82**	46.74	65.30	67.12	5.58	**10.88**	57.84	60.41	15.41	19.33	68.84	73.58	36.99	38.83
MTC + CLP	42.62	46.70	65.16	67.56	6.48	10.84	57.74	60.98	15.29	19.22	68.76	73.77	36.87	38.96
MT + CLTS + MTC	41.03	46.59	64.15	66.05	5.14	10.35	56.14	59.71	14.64	18.06	67.64	73.39	35.24	38.14
MT + CLTS + CLP	**42.72**	**47.67**	64.83	67.57	**6.17**	10.94	57.78	60.11	**15.96**	**19.16**	68.67	**74.03**	**36.54**	38.77
MT + MTC + CLP	42.33	46.72	**65.26**	**67.58**	5.92	**10.98**	**57.93**	**60.76**	15.38	18.81	**68.98**	73.55	36.22	**39.15**
CLTS + MTC + CLP	41.56	46.65	64.16	66.45	4.67	10.38	56.20	59.63	14.95	17.62	68.00	73.03	35.27	38.21
MT + CLTS + MTC + CLP	43.07	47.67	66.64	67.38	7.59	12.42	57.75	60.79	17.62	19.62	71.13	75.24	36.95	40.52

- Interestingly, the combination of tasks did not significantly improve performance.

Conclusion

- We introduce TED2025, a large-scale, high-quality multi-way parallel dataset covering 113 languages, with a maximum parallel degree of 50.
- Using TED2025, we investigate:
  - The effectiveness of using multi-way parallel data to improve the multilingual capabilities.
  - The best practice (impact factors) on using multi-way parallel data.
  - The best practice on using multi-way parallel data on instruction tuning.

Thank you!
Questions & Comments?

Paper

Code

Contact: wen.lai@tum.de syl@mail.tsinghua.edu.cn