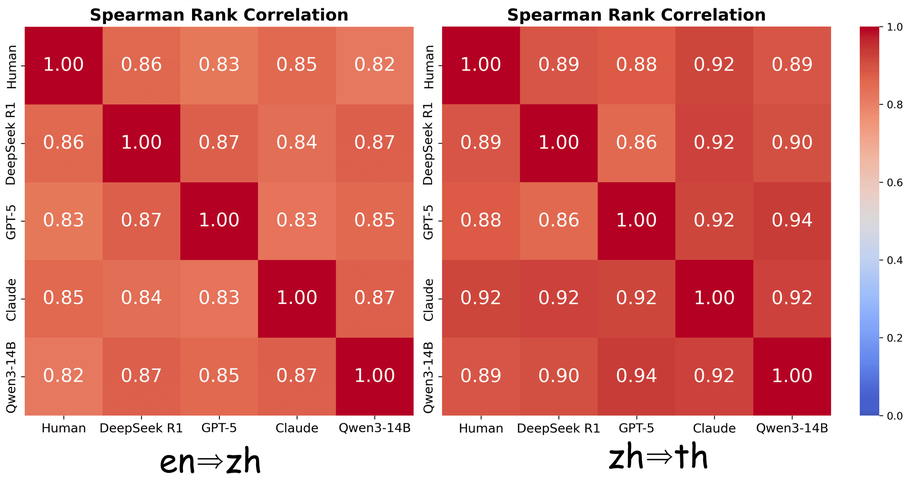
From Utterance to Vividity: Training Expressive Subtitle Translation LLM via Adaptive Local Preference Optimization

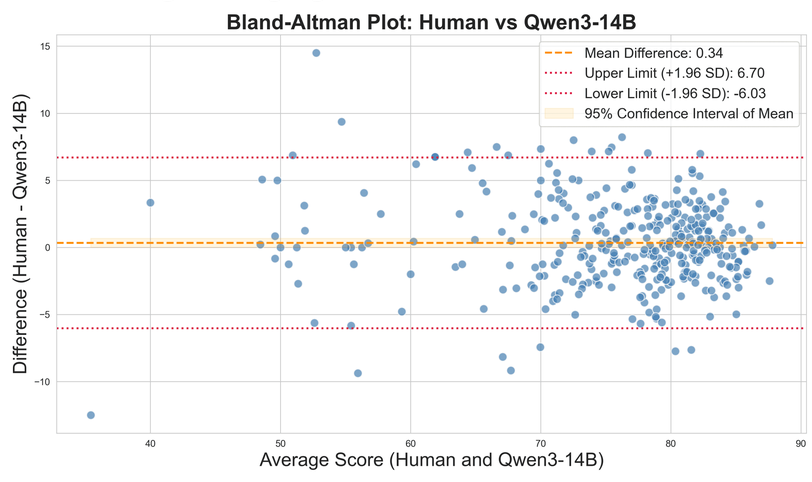
Chaoqun Cui, Shijing Wang, Liangbin Huang, Qingqing Gu,
Zhaolong Huang, Xiao Zeng, Wenji Mao

Empirical Investigation

1. LLM Is Excellent Translation Evaluator

We validate that LLMs can serve as reliable translation evaluators by comparing their vividness judgments with human ratings across multiple subtitle translation directions and showing strong agreement.





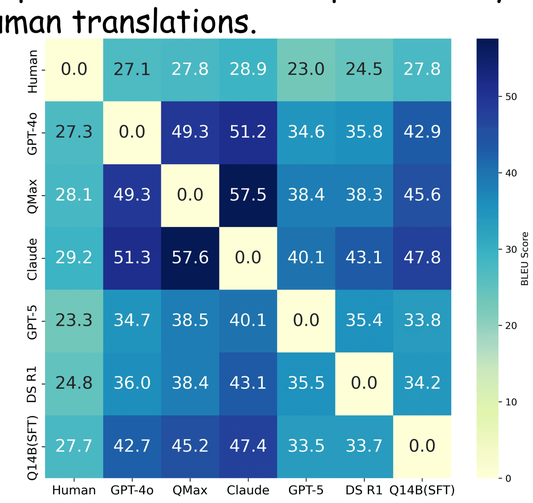
2. Parallel Corpora Are Actually Not Parallel

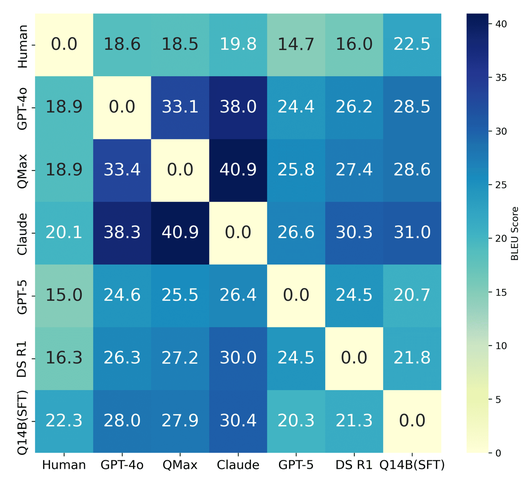
We quantitatively show that so-called parallel corpora across domains differ in how literally they mirror the source, with subtitle exhibiting much stronger liberal translation and thus being "not actually parallel" in a strict sense.

Dataset	Domain	en⇒de		en⇒fr		en⇒es	
		BLEU	ChrF++	BLEU	ChrF++	BLEU	ChrF++
OpenSubtitles	Visual Media	**15.00**	**37.08**	**17.84**	**37.80**	21.60	43.78
Books	Literature	17.22	43.20	21.51	47.14	**18.60**	**43.00**
bible-uedin	Religion	22.55	49.57	22.07	48.24	22.31	51.03
DGT	Legislation	19.85	50.05	22.44	48.10	21.88	52.24
JRC-Acquis		**27.55**	**61.95**	**28.83**	**69.65**	25.16	65.30
News-Commentary	News	25.00	55.23	24.87	54.37	**35.90**	65.84
ECDC	Medicine	23.18	60.77	23.86	62.97	27.20	65.39
EMEA		25.20	59.36	24.73	61.65	31.46	**69.15**

3. Chat LLMs Favor Literal, Reason LLMs Excel in Liberal

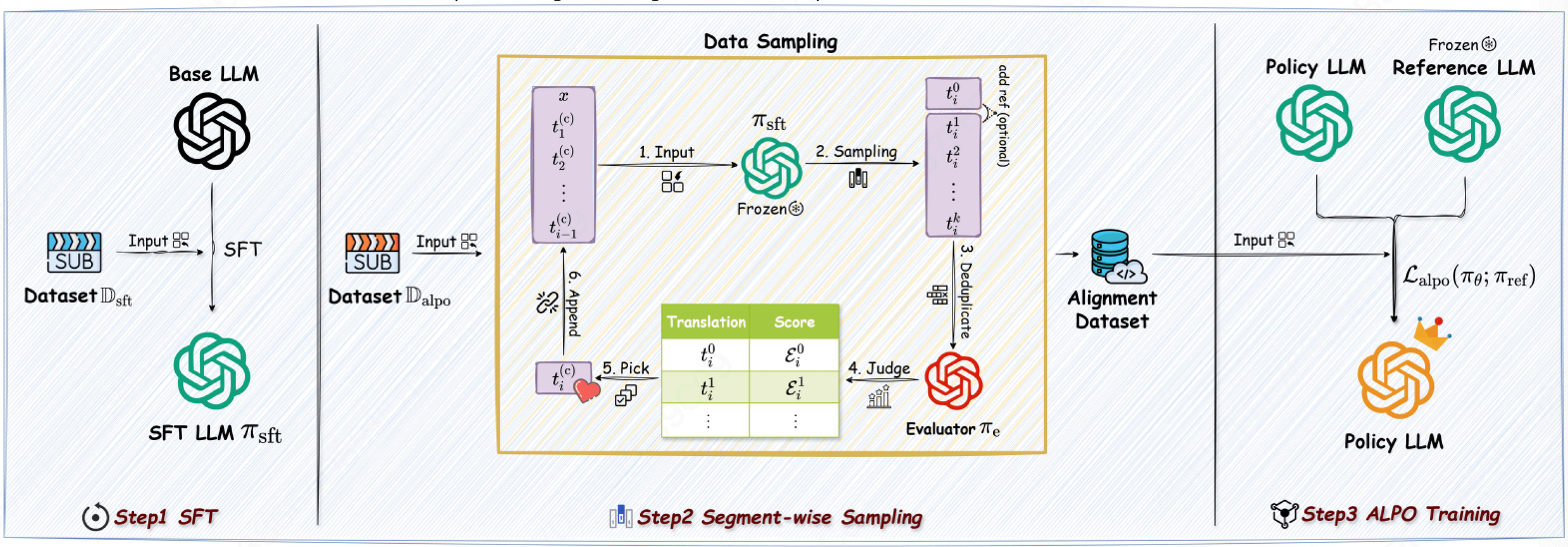
We showed that chat LLMs tend to produce more literal subtitle translations, whereas reasoning LLMs better generate outputs closer to the expressive style of human translations.





Adaptive Local Preference Optimization

We proposed ALPO, a process-supervised local preference optimization framework that samples segment-level subtitle translation candidates, scores them with an LLM evaluator, and trains the model with an adaptive fine-grained alignment loss to improve translation vividness.



Experiments

Models	Training	en⇒de			en⇒fr			en⇒zh		
		Accuracy	Naturalness	Vividness	Accuracy	Naturalness	Vividness	Accuracy	Naturalness	Vividness
Gold Reference	Human	84.8	83.8	**73.1**	83.5	85.0	74.8	83.6	82.6	**71.5**
VideoDubber	ST	41.5	35.5	41.0	48.2	47.3	48.5	46.9	51.9	49.7
NLLB-3.3B		75.8	70.1	59.2	76.9	74.6	61.8	61.4	54.0	43.7
MADLAD-10B		73.2	65.6	54.5	73.6	67.8	57.4	59.7	55.5	46.3
Google Translate		89.9	80.6	62.6	91.9	84.8	64.3	84.2	79.7	54.4
GPT-4o	ICL (C)	94.1	86.7	66.9	93.2	88.8	69.5	89.3	82.3	59.8
Qwen-Max		**95.5**	**89.2**	68.9	**94.1**	89.9	71.6	**91.9**	84.4	61.3
DeepSeek-V3.1		94.8	**89.2**	67.4	**94.9**	**90.7**	72.3	91.2	85.3	63.5
DeepSeek-R1	ICL (R)	92.8	87.4	70.0	93.6	90.0	73.4	90.5	**85.7**	70.8
GPT-5		93.6	88.6	72.7	92.2	**90.4**	**75.8**	**92.4**	**87.0**	71.1
Qwen2.5-14B	SFT	87.5	83.6	64.4	86.2	85.4	67.7	86.4	82.0	59.1
Qwen2.5-14B	**ALPO**	**95.4**	88.4	**74.8**	94.1	89.2	**78.8**	90.6	84.3	**76.6**