Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation

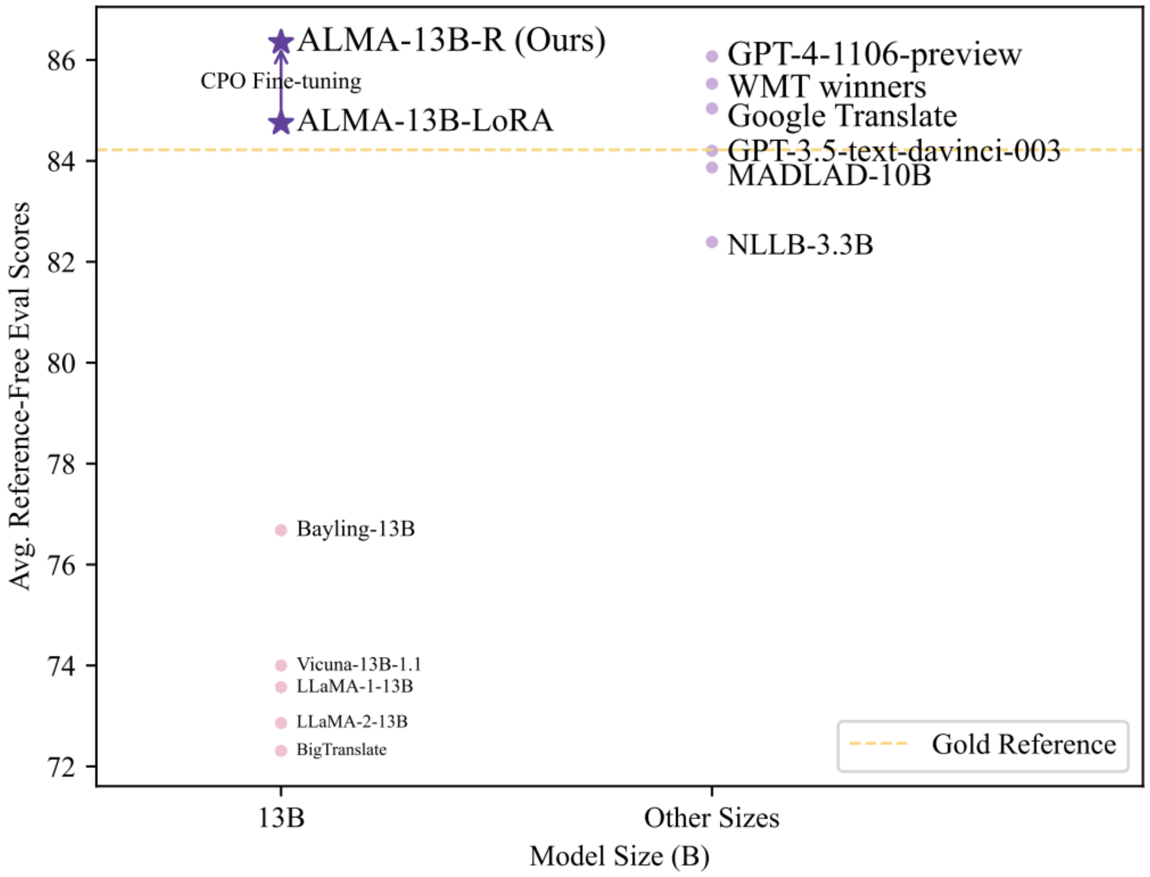
Haoran Xu

May 2024

To Appear @ICML 2024

ALMA Overview

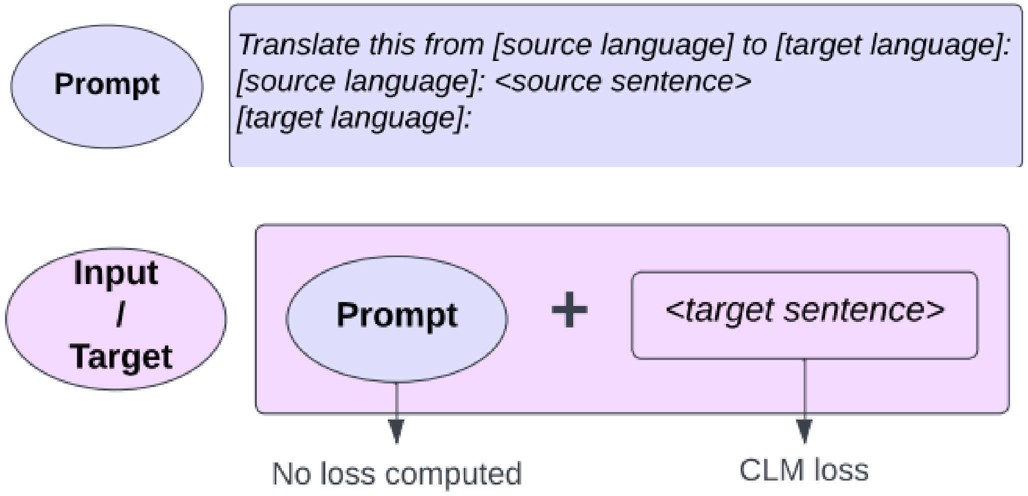
What is ALMA *(**A**dvanced **L**anguage **M**odel-Based Transl**A**tors)* ?
The first open-source LLM-based translation models which can beat GPT4



Better Instruction Tuning?

Fine-tune LLM on the translation task?

Reconsider the SFT objective, which is mimicking the gold reference. The performance can be capped by the quality of gold reference.



$$\mathcal{L}_{\text{NLL}} = -\mathbb{E}_{(x,y)\sim\mathcal{D}}[\log \pi_\theta(y|x)].$$

Beyond Gold References

Even human-written translations may not be perfect.

We compare the quality between the gold references and translation outputs from ALMA-13B and GPT-4.

Source: 这是马特利 (Martelly) 四年来第五次入选海地临时选举委员会 (CEP)。

Reference: It is Martelly's fifth CEP in four years.

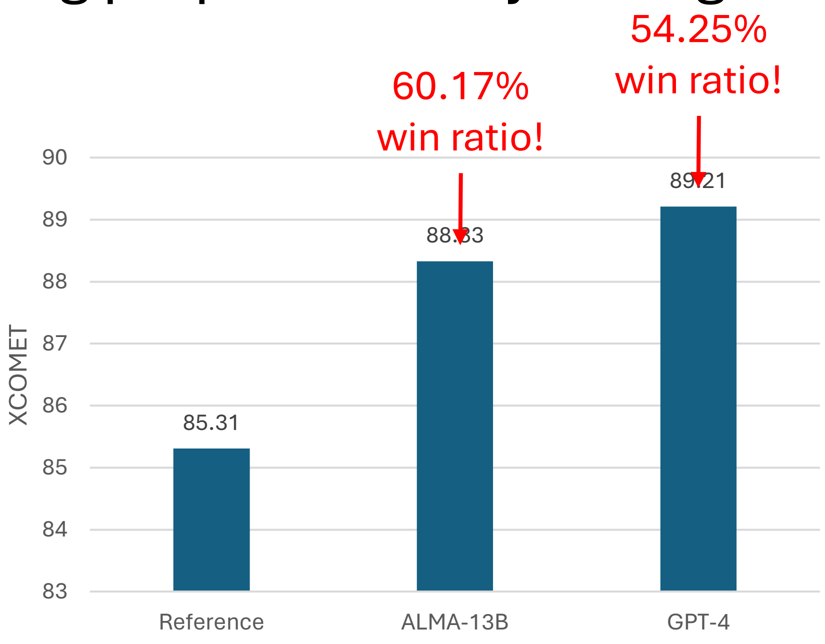
ALMA-13B-LoRA: This is Martelly's fifth time **being selected by the Provisional Electoral Council** (CEP) in four years.

GPT-4: This is the fifth time Martelly has been **selected for Haiti's Provisional Electoral Council** (CEP) in four years.

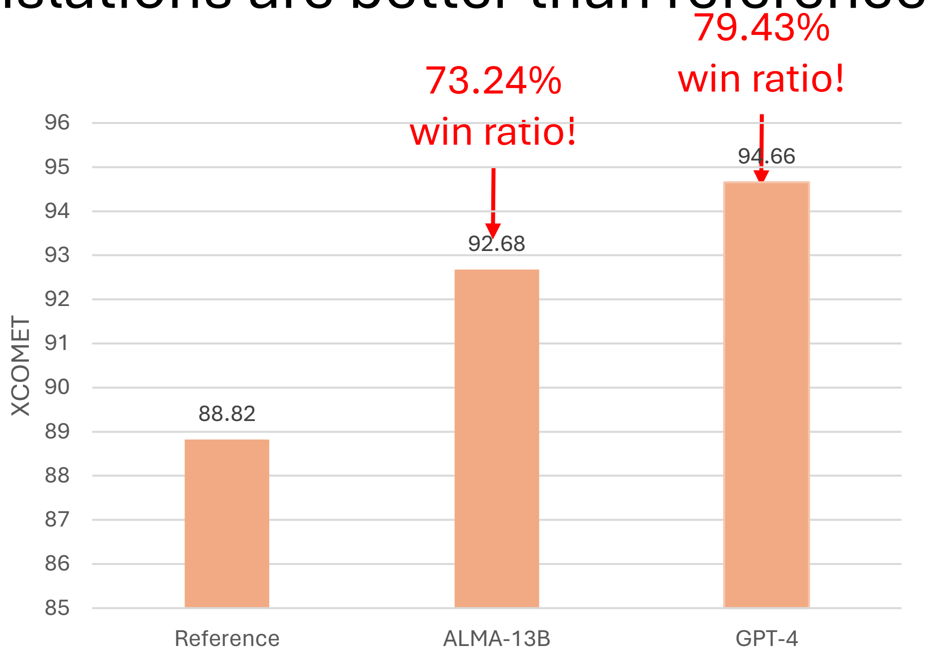
Beyond Gold References

Gold or Gilded? Scrutinizing Gold Reference Quality

A big proportion of system-generated translations are better than references.



XCOMET: Unbabel/XCOMET-XXL

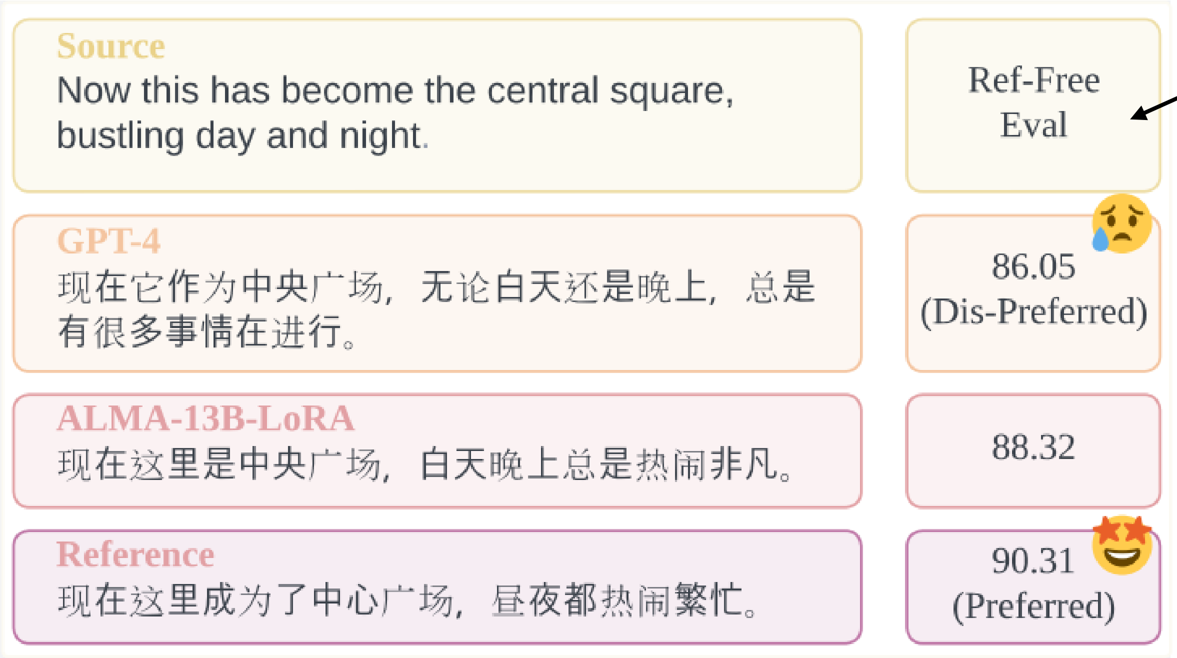


KIWI-XXL: Unbabel/wmt23-cometkiwi-da-xxl

Beyond Gold References

Motivation: Help The Model Learn Rejection

What is the best way to utilized these high-quality system-generated data? We believe the model need to learn how to reject “good but not perfect” translation.



We use avg. score of KIWI-XXL and XCOMET-XXL

Contrastive Preference Optimization

Building Preference Learning for MT. A popular way is to use DPO[1] ,a direct optimization in RLHF:

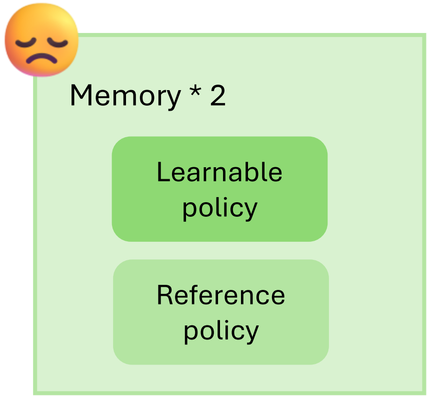
Learnable Policy

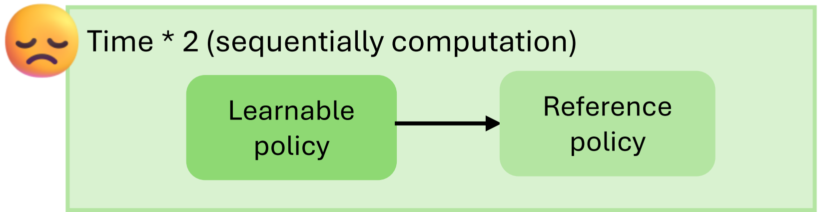
Preferred Translation

Dis-Preferred Translation

$$\mathcal{L}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right) \right]$$

Reference Policy





[1] Direct Preference Optimization: Your Language Model is Secretly a Reward Model

Johns Hopkins University

Microsoft

Contrastive Preference Optimization

Building Preference Learning for MT. A popular way is to use DPO[1] ,a direct optimization in RLHF:

Learnable Policy

Preferred Translation

Dis-Preferred Translation

$$\mathcal{L}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\log\sigma\Big(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\Big)\Big]$$

Reference Policy

Remove reference policy to approximate the optimization?

$$-\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\log\sigma\Big(\beta\log\pi_\theta(y_w|x) - \beta\log\pi_\theta(y_l|x)\Big)\Big]$$

[1] Direct Preference Optimization: Your Language Model is Secretly a Reward Model

Contrastive Preference Optimization

The answer is Yes! But why?

We only need to prove that $\mathcal{L}(\pi_\theta; \pi_{ref})$ is upper bounded by $\mathcal{L}(\pi_\theta; U)$

$$\mathcal{L}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}}\left[\log \sigma\left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right]$$

When reference policy is uniformly distributed

$$\mathcal{L}(\pi_\theta, U) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}}\left[\log \sigma\left(\beta \log \pi_\theta(y_w|x) - \beta \log \pi_\theta(y_l|x)\right)\right]$$

Appendix: Contrastive Preference Optimization

Behavior Cloning Constraint: a straightforward and strong signal to prevent the model from deviating the preferred data distribution:

$$\min_{\theta} \mathcal{L}(\pi_\theta, U) \text{ s.t. } \mathbb{E}_{(x,y_w)\sim\mathcal{D}}\Big[\mathbb{KL}(\pi_{\text{ref}}(y_w|x)||\pi_\theta(y_w|x))\Big] < \epsilon$$

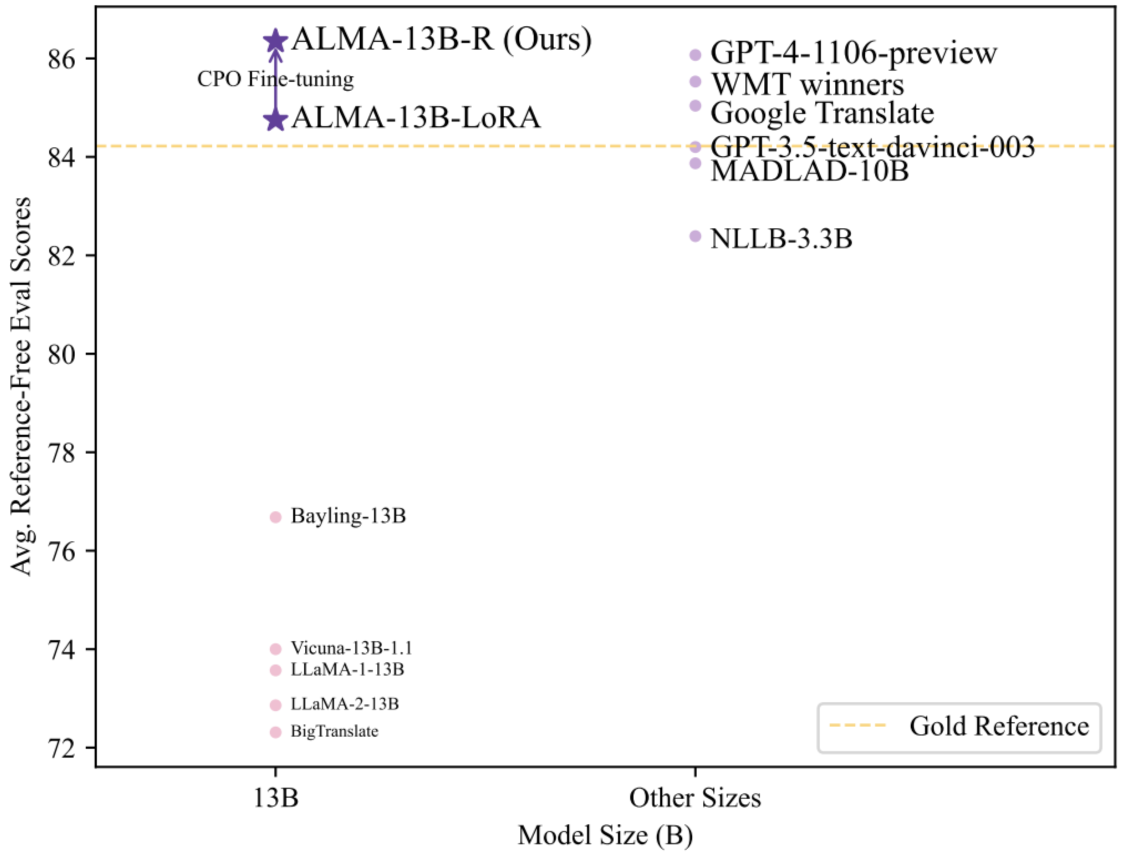
Equivalent to

$$\min_{\theta} \underbrace{\mathcal{L}(\pi_\theta, U)}_{\mathcal{L}_{prefer}} \underbrace{- \mathbb{E}_{(x,y_w)\sim\mathcal{D}}[\log \pi_\theta(y_w|x)]}_{\mathcal{L}_{nll}}$$

$$\mathcal{L}_{prefer} + \mathcal{L}_{nll}$$

Experiments

Performance?



Evaluation tools:

- wmt22-cometkiwi-da
- KIWI-XXL
- XCOMET-XXL

[1] Contrastive Preference Optimization: Pushing the Boundaries of LLM Performance in Machine Translation

Analyses

Analysis 1: CPO vs. DPO

Loss Objective	KIWI-22	KIWI-XXL	XCOMET	Memory Cost	FLOPs/tok
	Translating to English (xx→en)				
$\mathcal{L}_{\text{DPO}}$	80.51	81.36	86.58	2×	2×
$\mathcal{L}_{\text{DPO}} + \mathcal{L}_{\text{NLL}}$	81.28	82.42	89.05	2×	2×
$\mathcal{L}_{\text{prefer}} + \mathcal{L}_{\text{NLL}}$ (CPO)	**81.33**	**82.43**	**89.11**	1×	1×

Analyses

Analysis 2: Does The Quality of Dis-preferred Data Matter?

We consider a baseline where dis-preferred data is **manually created** by noising preferred data:

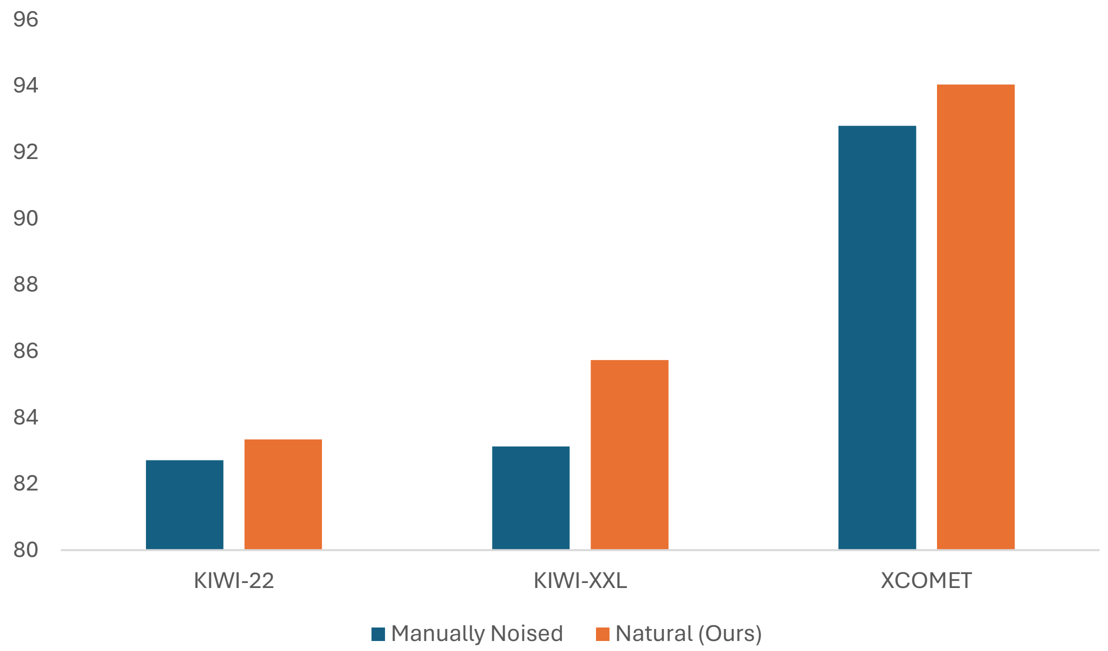
We applied random **deletions** of words with a probability of 0.15 and word **swaps** within a range of 1 with a probability of 0.3.

I like eating apples → like **apples eating**

Analyses

Analysis 2: Does The Quality of Dis-preferred Data Matter?

The quality of dis-preferred data does **matter**!



Conclusion

- Data Quality (even small!)is important.
- Maybe do not blindly trust the gold reference.
- Find a better alignment method:
    - SFT
    - DPO
    - CPO (CPO now is merged into huggingface now!)
    - ....

Many Thanks to My Collaborators!

Young Jin Kim

Kenton Murray

Hany Hassan Awadalla

Benjamin Van Durme

Amr Sharaf

Yunmo Chen

Weiting Tan

Lingfeng Shen

Questions?

Analyses

Analysis: Are Translations Really Better or Just Metric-Preferred?

Preferred data is selected by reference-free models and the same models are used for evaluation. Any "cheating" here?

In the method-level: Training on preferred data does not lead better performance on these metrics.

	KIWI-XXL
ALMA-13B	82.66
SFT on preferred data	82.42
DPO on preferred data	82.42
CPO on preferred data	**85.74**

Analyses

Analysis: Are Translations Really Better or Just Metric-Preferred?

Human Eval! 400 examples sampled from zh->en

	Avg. Score	Avg. Rank	Avg. Win Ratio (%)
ALMA-13B-LoRA	4.86	1.60	62.5
ALMA-13B-R	5.16	1.40	77.8

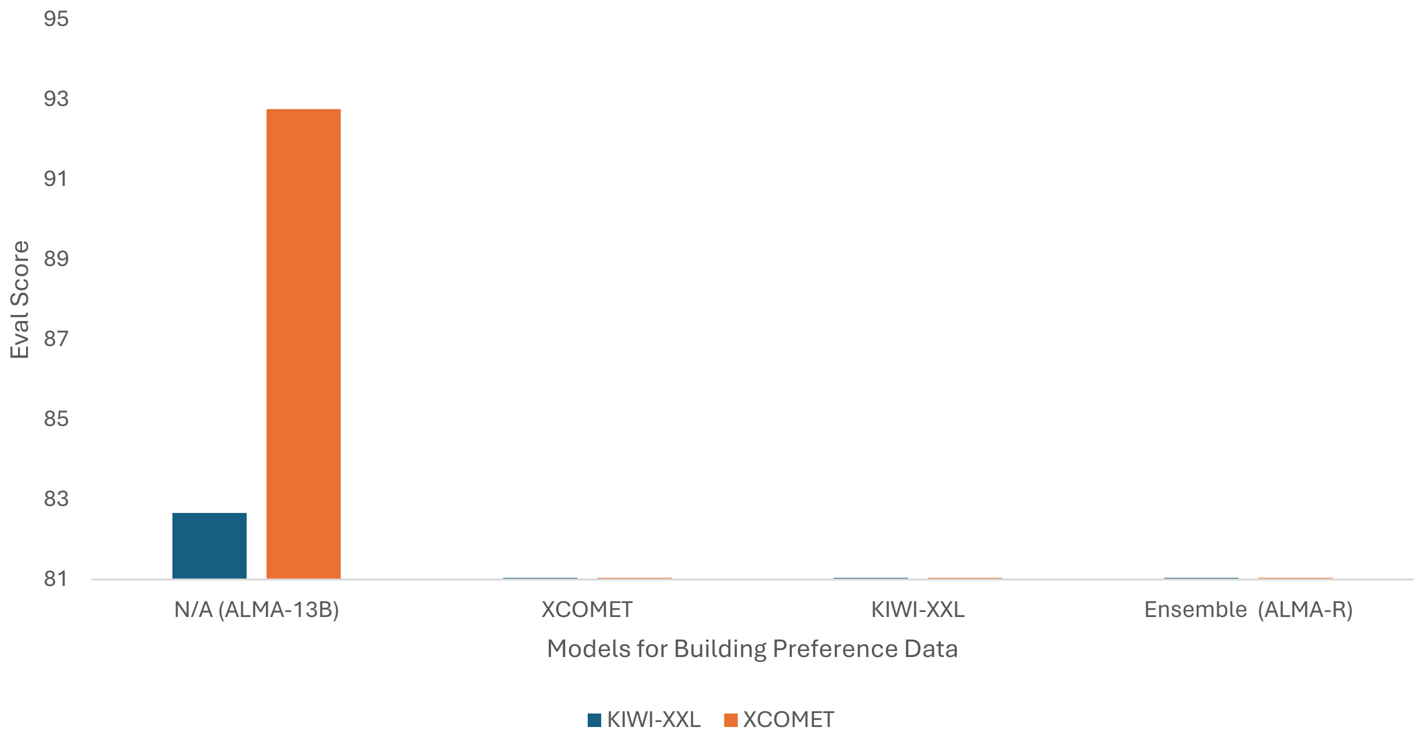
Appendix: Results on BLEURT

BLEURT-20	de	cs	is	zh	ru	Avg.
Translating to English (xx→en)						
ALMA-13B-LoRA	73.20	76.65	75.87	67.37	76.7	73.96
ALMA-13B-R	**73.62**	**76.94**	**76.98**	**69.48**	**76.91**	**74.79**
Translating from English (en→xx)						
ALMA-13B-LoRA	75.51	80.93	73.19	70.54	74.94	75.02
ALMA-13B-R	**77.20**	**81.87**	**73.43**	**71.51**	**76.19**	**76.04**

Analyses

Analysis: Are Translations Really Better or Just Metric-Preferred?

In the metric-level: No significant bias towards the metric used for selecting preferred data:



Johns Hopkins University

Microsoft

Appendix: ALMA-R Results for xx→en

Models	de			cs			is		
	KIWI-22	KIWI-XXL	XCOMET	KIWI-22	KIWI-XXL	XCOMET	KIWI-22	KIWI-XXL	XCOMET
Gold Reference	78.74	78.56	88.82	82.08	83.11	84.60	80.88	85.04	76.16
WMT Winners	81.38	83.59	93.74	82.47	82.53	85.65	81.39	85.60	78.14
GPT-4	**81.50**	**84.58**	**94.47**	82.52	83.55	**88.48**	81.49	**85.90**	**81.11**
ALMA-13B-LoRA	81.14	83.57	93.30	81.96	82.97	83.95	80.90	85.49	76.68
+ SFT on preferred data	81.36	83.98	93.84	82.36	83.15	86.67	81.32	85.61	80.20
+ DPO	81.13	83.52	93.25	81.82	82.69	83.84	80.89	85.22	76.09
+ CPO (Ours, ALMA-13B-R)	**81.50**	83.97	94.20	**82.63**	**83.75**	88.03	**81.57**	85.73	80.49

Models	zh			ru			Avg.		
	KIWI-22	KIWI-XXL	XCOMET	KIWI-22	KIWI-XXL	XCOMET	KIWI-22	KIWI-XXL	XCOMET
Gold Reference	77.09	74.19	90.70	80.74	79.59	88.56	79.91	80.10	85.77
WMT Winners	77.66	73.28	87.2	81.71	80.97	90.91	80.92	81.19	87.13
GPT-4	**79.33**	**77.65**	**92.06**	81.57	81.34	90.95	81.28	**82.60**	**89.41**
ALMA-13B-LoRA	77.32	74.41	89.88	81.31	81.05	89.89	80.53	81.50	86.74
+ SFT on preferred data	78.32	76.03	90.65	81.46	81.17	90.65	80.96	81.99	88.40
+ DPO	77.50	74.50	89.94	81.19	80.88	89.76	80.51	81.36	86.58
+ CPO (Ours, ALMA-13B-R)	79.24	77.17	91.65	**81.72**	**81.54**	**91.18**	**81.33**	82.43	89.11

Appendix: ALMA-R Results on WMT'23

	de→en			zh→en			ru→en		
	KIWI-22	KIWI-XXL	XCOMET	KIWI-22	KIWI-XXL	XCOMET	KIWI-22	KIWI-XXL	XCOMET
Gold Reference	78.93	75.96	84.23	74.46	68.80	83.51	79.46	77.84	83.60
WMT Winners	79.37	76.18	84.35	**80.17**	**79.53**	92.25	80.88	79.21	86.22
TowerInstruct	79.67	77.60	86.28	79.84	78.13	91.75	80.85	80.03	87.76
MADLAD-10B	78.52	75.50	83.85	77.68	73.72	88.07	79.65	77.58	85.15
ALMA-13B-LoRA	79.36	76.79	85.07	78.83	76.71	90.73	80.79	80.14	86.94
+ CPO (Ours, ALMA-13B-R)	**79.87**	**77.69**	**86.62**	80.01	78.42	**92.36**	**81.11**	**80.95**	**88.75**

	en→de			en→zh			en→ru		
	KIWI-22	KIWI-XXL	XCOMET	KIWI-22	KIWI-XXL	XCOMET	KIWI-22	KIWI-XXL	XCOMET
Gold Reference	80.12	**77.93**	88.91	79.60	73.47	86.15	79.87	79.36	91.41
WMT Winners	**80.80**	77.26	87.94	79.70	74.20	87.24	**82.51**	79.95	91.41
TowerInstruct	80.13	75.34	86.55	80.03	74.85	86.74	81.33	77.14	89.59
MADLAD-10B	77.48	70.87	86.18	74.63	62.07	79.12	79.24	72.40	86.64
ALMA-13B-LoRA	78.79	73.40	85.61	78.92	72.95	85.13	80.21	76.02	89.48
+ CPO (Ours, ALMA-13B-R)	79.85	77.05	**89.79**	**80.48**	**78.17**	**88.34**	81.97	**81.52**	**92.56**

Appendix: Contrastive Preference Optimization

The answer is Yes! But why?

We only need to prove that $\mathcal{L}(\pi_\theta; \pi_{ref})$ is upper bounded by $\mathcal{L}(\pi_\theta; U)$

$$\mathcal{L}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x)}{\cancel{\pi_{\text{ref}}(y_w|x)}} - \beta\log\frac{\pi_\theta(y_l|x)}{\cancel{\pi_{\text{ref}}(y_l|x)}}\right)\right]$$

Remove reference policy to approximate the optimization?

$$\mathcal{L}(\pi_\theta, U) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\log\sigma\Big(\beta\log\pi_\theta(y_w|x) - \beta\log\pi_\theta(y_l|x)\Big)\Big]$$

Appendix: Contrastive Preference Optimization

Theorem 1. $\mathcal{L}(\pi_\theta; \pi_{ref})$ is upper bounded by $\mathcal{L}(\pi_\theta; U)$ if π_{ref} is an ideal policy that perfectly aligns the true data distribution of the preferred data.

$$\pi_{ref}(y_w|x) = 1$$

$$0 \leq \pi_{ref}(y_l|x) \leq 1$$

$$\mathcal{L}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right]$$

$\pi_{\text{ref}}(y_w|x)$=1

Appendix: Contrastive Preference Optimization

Theorem 1. $\mathcal{L}(\pi_\theta; \pi_{ref})$ is upper bounded by $\mathcal{L}(\pi_\theta; U)$ if π_{ref} is an ideal policy that perfectly aligns the true data distribution of the preferred data.

$$\begin{aligned}
\mathcal{L}(\pi_\theta; \pi_{\text{ref}}) &= -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\log \pi_\theta(y_w|x)^\beta - \log\Big(\pi_\theta(y_w|x)^\beta \cdot \boxed{\pi_{\text{ref}}(y_l|x)^\beta} + \pi_\theta(y_l|x)^\beta\Big)\Big] \\
&\leq -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\log \pi_\theta(y_w|x)^\beta - \log\Big(\pi_\theta(y_w|x)^\beta \cdot \boxed{1} + \pi_\theta(y_l|x)^\beta\Big)\Big] \\
&= -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\log \sigma\Big(\beta \log \pi_\theta(y_w|x) - \beta \log \pi_\theta(y_l|x)\Big)\Big]. \\
&= \mathcal{L}(\pi_\theta, U)
\end{aligned}$$

Appendix: Contrastive Preference Optimization

Additional Constraint: a straightforward and strong signal to prevent the model from deviating the preferred data distribution:

$$\min_{\theta} \mathcal{L}(\pi_\theta, U) \text{ s.t. } \mathbb{E}_{(x,y_w)\sim\mathcal{D}}\Big[\mathbb{KL}(\pi_{\text{ref}}(y_w|x)||\pi_\theta(y_w|x))\Big] < \epsilon$$

Equivalent to

$$\min_{\theta} \underbrace{\mathcal{L}(\pi_\theta, U)}_{\mathcal{L}_{prefer}} \underbrace{- \mathbb{E}_{(x,y_w)\sim\mathcal{D}}[\log \pi_\theta(y_w|x)]}_{\mathcal{L}_{nll}}$$

$$\mathcal{L}_{prefer} + \mathcal{L}_{nll}$$

Appendix: Contrastive Preference Optimization

$$\min_{\theta} \mathcal{L}(\pi_\theta, U) \text{ s.t. } \mathbb{E}_{(x,y_w)\sim\mathcal{D}}\Big[\mathbb{KL}(\pi_{\text{ref}}(y_w|x)||\pi_\theta(y_w|x))\Big] < \epsilon$$

This is equivalent to the following objective via Lagrangian duality:

$$\min_{\theta} \mathcal{L}(\pi_\theta, U) + \lambda \cdot \mathbb{E}_{(x,y_w)\sim\mathcal{D}}\Big[\mathbb{KL}(\pi_w(y_w|x)||\pi_\theta(y_w|x))\Big]$$

$$\begin{aligned}
\mathcal{L}_{\text{CPO}} &= \mathcal{L}(\pi_\theta, U) + \mathbb{E}_{(x,y_w)\sim\mathcal{D}}\Big[\mathbb{KL}(\pi_w(y_w|x)||\pi_\theta(y_w|x))\Big] \\
&= \mathcal{L}(\pi_\theta, U) + \mathbb{E}_{(x,y_w)\sim\mathcal{D}}\Big[\pi_w(y_w|x) \cdot \log\Big(\pi_w(y_w|x)\Big) - \pi_w(y_w|x) \cdot \log\Big(\pi_\theta(y_w|x)\Big)\Big] \\
&= \mathcal{L}(\pi_\theta, U) + \mathbb{E}_{(x,y_w)\sim\mathcal{D}}\Big[1 \cdot 0 - 1 \cdot \log\Big(\pi_\theta(y_w|x)\Big)\Big] \\
&= \mathcal{L}(\pi_\theta, U) - \mathbb{E}_{(x,y_w)\sim\mathcal{D}}\Big[\log\Big(\pi_\theta(y_w|x)\Big)\Big].
\end{aligned}$$