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

Haoran Xu♠ Amr Sharaf♡ Yunmo Chen♠ Weiting Tan♠ Lingfeng Shen♠ Benjamin Van Durme♠
Kenton Murray* ♠ Young Jin Kim* ♡

Zhiyi Chen, Jeongrok Yu

Georgia Tech

Recap

- DPO: $$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w \mid x)}{\boxed{\pi_{\text{ref}}(y_w \mid x)}} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\boxed{\pi_{\text{ref}}(y_l \mid x)}} \right) \right]$$
    - Need reference model (SFT model) + Ensure the policy model doesn't diverge too far

Recap

- DPO: $$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w \mid x)}{\boxed{\pi_{\text{ref}}(y_w \mid x)}} - \beta\log\frac{\pi_\theta(y_l \mid x)}{\boxed{\pi_{\text{ref}}(y_l \mid x)}}\right)\right]$$
  - Need reference model (SFT model) + Ensure the policy model doesn't diverge too far
- SimPO: $$\mathcal{L}_{\text{SimPO}}(\pi_\theta) = -\mathbb{E}_{(x, y_w, y_l)\sim\mathcal{D}}\left[\log\sigma\left(\frac{\beta}{|y_w|}\log\pi_\theta(y_w|x) - \frac{\beta}{|y_l|}\log\pi_\theta(y_l|x) - \gamma\right)\right]$$
  - No need of reference model
  - Intuition: The reward being optimized during DPO training and the generation metric used for inference is different

$$p_\theta(y \mid x) = \frac{1}{|y|}\log\pi_\theta(y \mid x) = \frac{1}{|y|}\sum_{i=1}^{|y|}\log\pi_\theta(y_i \mid x, y_{<i})$$

  - Solution: Employs an implicit reward formulation that directly aligns with the generation metric

Georgia Tech

This paper: Relation to SimPO

- Contrastive Preference Optimization (CPO):
  - Shares a similar **reference-free preference learning framework** with SimPO
  - Key differences
    - Objective: CPO focuses on machine translation (MT) tasks, while SimPO targets more general tasks
    - Intuition: In MT tasks, the authors of CPO found that **human-written reference** data is often inferior in quality compared to system-generated translations

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.

CPO-related work timeline

- Translation LLM: Advanced Language Model-based trAnslator (ALMA)
  - *Paradigm Shift in Machine Translation: Boosting Translation Performance of Large Language Models* (ICLR 2024)
- Contrastive Preference Optimization (CPO) & ALMA-R model
  - *Contrastive Preference Optimization: Pushing the Boundaries of LLM Performance in Machine Translation* (ICML 2024)
- CPO-SimPO
  - GitHub repo: A new training approach combining objectives of CPO and SimPO

Georgia Tech

Takeaway 1

- CPO shares the same reference-free idea of SimPO, and their objectives can be combined to a even better objective
    - They’ve published the source code

Gold or Gilded? Scrutinizing Gold Reference Quality

- Goal: Based on FLORES-200 dataset, evaluate its **gold references** and translation outputs from **ALMA13B-LoRA2** and **GPT-4**.
- Approach: Use reference-free evaluation frameworks to rank and compare the gold references and system-generated translations
  - Evaluate the quality of a MT system's output without using human-produced reference translations for comparison
  - Model-based frameworks: two latest and largest reference-free models, each with a 10B parameter size
    - KIWI-XXL, XCOMET

Gold or Gilded? Scrutinizing Gold Reference Quality

- Scope: 5 English-centric language pairs, covering both translations from and to English (German (de), Czech (cs), Icelandic (is), Chinese (zh), and Russian (ru))
- Prompt:

GPT-4 Prompt

System:
You are a helpful translator and only output the result.

User:
Translate this from <source language> to <target language>, <source language>:
<source sentence>
<target language>:

ALMA Prompt

Translate this from <source language> to <target language>:
<source language>: <source sentence>
<target language>:

Gold or Gilded? Scrutinizing Gold Reference Quality

- Metrics: Average evaluation scores + win ratio (model outputs surpass the gold standard references)
- Observations?

	KIWI-XXL	Win Ratio (%)	XCOMET	Win Ratio (%)
Translating to English (xx→en)				
Reference	85.31	-	88.82	-
ALMA-13B-LoRA	88.33	73.24	92.68	60.17
GPT-4	89.21	79.43	94.66	54.25
Translating from English (en→xx)				
Reference	87.85	-	94.42	-
ALMA-13B-LoRA	85.62	42.15	93.07	35.46
GPT-4	87.30	49.13	94.21	38.09

Gold or Gilded? Scrutinizing Gold Reference Quality

- For the average performance of translation models in xx→en, system-generated translations significantly exceeds the human-written references
- In the en→xx direction, while the overall performance between the translations from reference and two systems is comparable, approximately 40% are still deemed superior to the reference translations

	KIWI-XXL	Win Ratio (%)	XCOMET	Win Ratio (%)
	Translating to English (xx→en)			
Reference	85.31	-	88.82	-
ALMA-13B-LoRA	88.33	73.24	92.68	60.17
GPT-4	89.21	79.43	94.66	54.25
	Translating from English (en→xx)			
Reference	87.85	-	94.42	-
ALMA-13B-LoRA	85.62	42.15	93.07	35.46
GPT-4	87.30	49.13	94.21	38.09

Takeaway 2

- Human-written references are not good enough -> we do not want our model to merely mimic (be fine-tuned) the gold references

Contrastive Preference Optimization – Preference data construction

- Preference data construction
  - Use the same setting in the previous evaluation: based on FLORES-200, use the two reference-free eval frameworks to rank (1) reference, (2) GPT-4, (3) ALMA translations based on average performance scores
  - $y_w = \mathbf{y}_{\arg\max_i(\mathbf{s})}, y_l = \mathbf{y}_{\arg\min_i(\mathbf{s})}$

Contrastive Preference Optimization – Objective

- Idea 1: Starting from the DPO objective, get rid of the reference policy term
  - Consider **the weakest policy** and the ideal policy:
    - Weakest: A uniform prior **U** (for a given x, predict the same score for all y)

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

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

Contrastive Preference Optimization – Objective

- Idea 1: Starting from the DPO objective, get rid of the reference policy term
  - Consider the weakest policy and **the ideal policy**:
    - Weakest: A uniform prior **U** (give all x the same prediction)
    - **Ideal:** Predicts 1 for the preferred translation π_w
  - The objective of the weakest policy gets rid of references (what we want), while the ideal policy is what we target
  - How to build the connection between the two objectives?

$\mathcal{L}(\pi_\theta; \pi_w)$ **?** $\mathcal{L}(\pi_\theta; U)$

Contrastive Preference Optimization – Proof 1

- The DPO loss of the ideal policy $\mathcal{L}(\pi_\theta; \pi_w)$ is upper-bounded by $\mathcal{L}(\pi_\theta; U)$

$$
\begin{aligned}
\mathcal{L}(\pi_\theta; \pi_w) &= -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\log\sigma\Big(\beta\log\frac{\pi_\theta(y_w|x)}{\boxed{\pi_w(y_w|x)}} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_w(y_l|x)}\Big)\Big] \\
&= -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\log\sigma\Big(\beta\log\pi_\theta(y_w|x) - \beta\log\pi_\theta(y_l|x) + \beta\log\pi_w(y_l|x)\Big)\Big]
\end{aligned}
$$

The ideal policy predicts 1 for the preferred translation

Contrastive Preference Optimization – Proof 1

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

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

$$\mathcal{L}(\pi_\theta;\pi_w) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\left(\frac{1}{1+e^{-\beta\log\pi_\theta(y_w|x)+\beta\log\pi_\theta(y_l|x)-\beta\log\pi_w(y_l|x)}}\right)\right]$$

$$= -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\left(\frac{1}{1+\frac{\pi_\theta(y_l|x)^\beta}{\pi_\theta(y_w|x)^\beta\cdot\pi_w(y_l|x)^\beta}}\right)\right]$$

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

Expanding the sigmoid function

Contrastive Preference Optimization – Proof 1

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

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

$$\mathcal{L}(\pi_\theta;\pi_w) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\log\Big(\frac{1}{1+e^{-\beta\log\pi_\theta(y_w|x)+\beta\log\pi_\theta(y_l|x)-\beta\log\pi_w(y_l|x)}}\Big)\Big]$$

$$= -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\log\Big(\frac{1}{1+\frac{\pi_\theta(y_l|x)^\beta}{\pi_\theta(y_w|x)^\beta\cdot\pi_w(y_l|x)^\beta}}\Big)\Big]$$

$$= -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\log\pi_\theta(y_w|x)^\beta + \boxed{\log\pi_w(y_l|x)^\beta} - \log\Big(\pi_\theta(y_w|x)^\beta\cdot\pi_w(y_l|x)^\beta + \pi_\theta(y_l|x)^\beta\Big)\Big]$$

$$\mathcal{L}'(\pi_\theta;\pi_w) \triangleq \mathcal{L}(\pi_\theta;\pi_w) + \boxed{\underbrace{\mathbb{E}_{(x,y_l)\sim\mathcal{D}}\Big[\log\pi_w(y_l|x)^\beta\Big]}_{C\text{ in the Theorem}}}$$

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

Not related to the ideal policy
Constant C

Contrastive Preference Optimization – Proof 1

$$\begin{aligned}\mathcal{L}(\pi_\theta;\pi_w) &= -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\log\sigma\Big(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_w(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_w(y_l|x)}\Big)\Big]\\ &= -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\log\sigma\Big(\beta\log\pi_\theta(y_w|x) - \beta\log\pi_\theta(y_l|x) + \beta\log\pi_w(y_l|x)\Big)\Big]\end{aligned}$$

$$\begin{aligned}\mathcal{L}(\pi_\theta;\pi_w) &= -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\log\Big(\frac{1}{1+e^{-\beta\log\pi_\theta(y_w|x)+\beta\log\pi_\theta(y_l|x)-\beta\log\pi_w(y_l|x)}}\Big)\Big]\\ &= -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\log\Big(\frac{1}{1+\frac{\pi_\theta(y_l|x)^\beta}{\pi_\theta(y_w|x)^\beta\cdot\pi_w(y_l|x)^\beta}}\Big)\Big]\\ &= -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\log\pi_\theta(y_w|x)^\beta + \log\pi_w(y_l|x)^\beta - \log\Big(\pi_\theta(y_w|x)^\beta\cdot\pi_w(y_l|x)^\beta + \pi_\theta(y_l|x)^\beta\Big)\Big]\end{aligned}$$

$$\begin{aligned}\mathcal{L}'(\pi_\theta;\pi_w) &\triangleq \mathcal{L}(\pi_\theta;\pi_w) + \underbrace{\mathbb{E}_{(x,y_l)\sim\mathcal{D}}\Big[\log\pi_w(y_l|x)^\beta\Big]}_{C\text{ in the Theorem}}\\ &= -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\log\pi_\theta(y_w|x)^\beta - \log\Big(\pi_\theta(y_w|x)^\beta\boxed{\cdot\,\pi_w(y_l|x)^\beta} + \pi_\theta(y_l|x)^\beta\Big)\Big]\end{aligned}$$

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

Upper-bound

Contrastive Preference Optimization – Objective

- Idea 2: incorporate a behavior cloning (BC) regularizer to ensure that the policy model does not deviate from the preferred data distribution

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

    - How to incorporate the regularizer into the objective?

Contrastive Preference Optimization – Proof 2

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

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

Lagrangian duality

Contrastive Preference Optimization – Proof 2

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

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

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

Set lambda to 1, ideal policy predicts 1 for the preferred translation

Contrastive Preference Optimization – Proof 2

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

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

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

- CPO Loss: $\min_{\theta} \underbrace{\mathcal{L}(\pi_\theta, U)}_{\mathcal{L}_{\text{prefer}}} \underbrace{-\mathbb{E}_{(x,y_w)\sim\mathcal{D}}[\log \pi_\theta(y_w|x)]}_{\mathcal{L}_{\text{NLL}}}$
    - Preference optimization term (reference-free) + negative log-likelihood term

Experimental Setup

- English, Czech, Chinese, German, Russian, Icelandic (10 translation directions)

- Preference dataset: FLORES-200 + Human-labeled

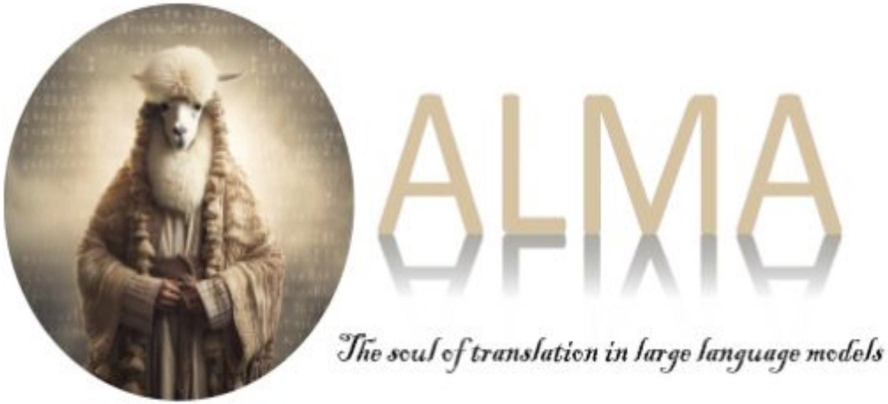
The statistic of how many translations win or tie by each system evaluated by human.

	Google Wins	GPT-4 Wins	Ties
en→de	418	435	203
en→zh	362	412	282

- ALMA-13B-LoRA vs WMT Winner vs GPT 4 vs Gold Reference vs SFT vs DPO vs CPO on WMT 21, WMT 22, assessed with reference-free evaluation models (KIWI-22, XXL, XCOMET…)

ALMA-13B-LoRA

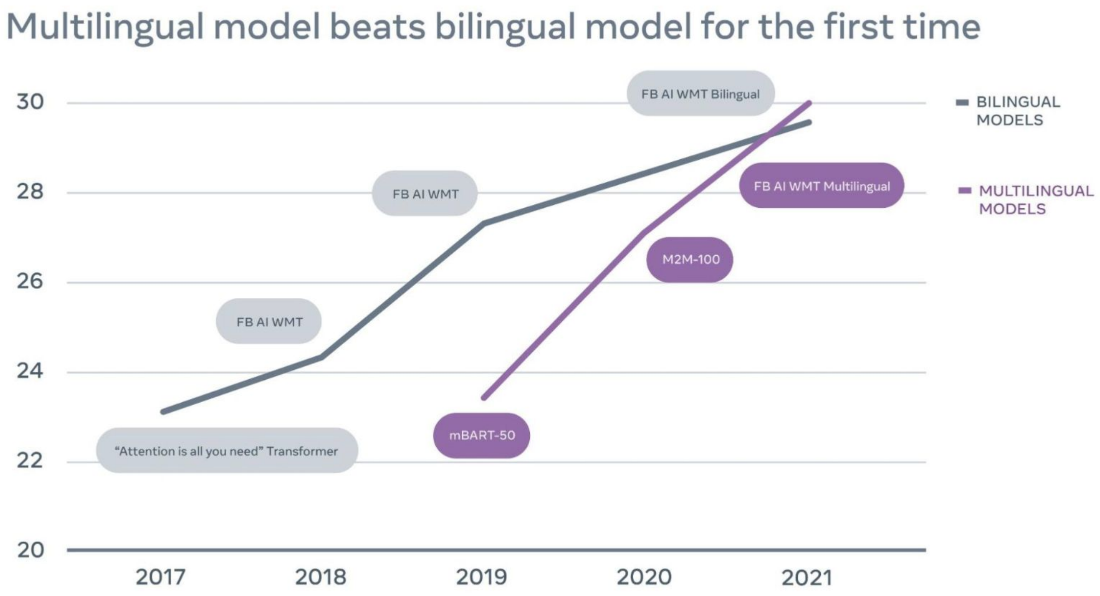
- Llama2-13B→ Full-weight training on monolingual data → LoRA on high quality parallel data



ALMA: Advanced Language Model-based translator

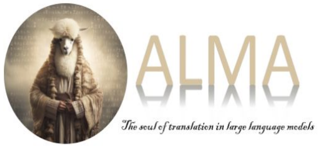
WMT

- ALMA-13B-LoRA vs WMT Winner vs GPT 4 vs Gold Reference vs SFT vs DPO vs CPO on WMT 21, WMT 22, assessed with reference-free evaluation models (KIWI-22, XXL, XCOMET…)



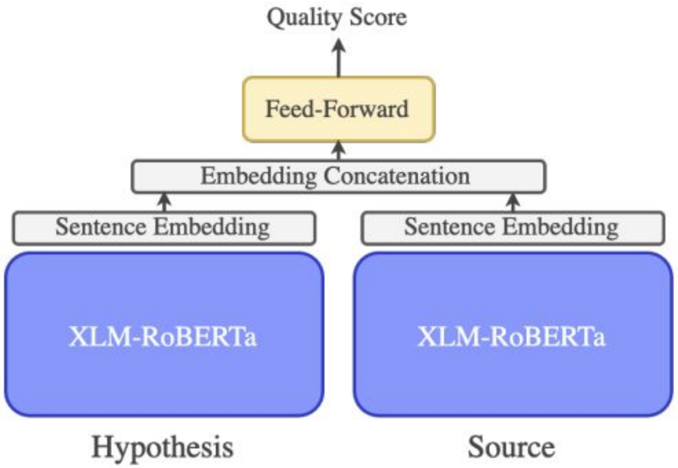
Reference-free Evaluation Models

- ALMA-13B-LoRA vs WMT Winner vs GPT 4 vs Gold Reference vs SFT vs DPO vs CPO on WMT 21,22 assessed with reference-free evaluation models (KIWI-22, XXL, XCOMET…)



ALMA: Advanced Language Model-based translator

I am a student.



Ich bin Student.

Overall Results for Multilingual Outputs (en → xx)

Models	de			cs			is		
	KIWI-22	KIWI-XXL	XCOMET	KIWI-22	KIWI-XXL	XCOMET	KIWI-22	KIWI-XXL	XCOMET
Gold Reference	82.67	84.01	**97.85**	83.19	81.83	90.27	80.51	85.20	91.52
WMT Winners	**83.56**	83.70	96.99	85.31	**87.27**	**94.38**	81.77	84.94	91.61
GPT-4	83.48	**84.91**	97.56	84.81	85.35	93.48	81.03	81.21	90.00
ALMA-13B-LoRA	82.62	81.64	96.49	84.14	84.24	92.38	81.71	83.31	91.20
+ SFT on preferred data	82.75	81.85	96.67	84.14	83.46	91.99	81.48	82.11	90.30
+ DPO	82.40	81.20	96.40	83.86	83.45	91.68	81.43	82.66	90.33
+ CPO (Ours, ALMA-13B-R)	83.28	84.25	97.48	**84.99**	87.06	93.61	**82.18**	**85.68**	**91.93**

Models	zh			ru			Avg.		
	KIWI-22	KIWI-XXL	XCOMET	KIWI-22	KIWI-XXL	XCOMET	KIWI-22	KIWI-XXL	XCOMET
Gold Reference	80.92	81.70	90.42	82.96	84.62	94.17	82.05	83.47	92.85
WMT Winners	82.04	81.13	91.14	**84.35**	87.01	94.79	**83.41**	84.81	93.78
GPT-4	81.73	81.53	90.79	83.64	86.15	94.3	82.94	83.83	93.23
ALMA-13B-LoRA	80.82	79.96	89.92	83.10	84.17	93.79	82.48	82.66	92.76
+ SFT on preferred data	81.25	80.51	90.18	83.23	84.15	93.54	82.57	82.42	92.54
+ DPO	80.74	79.64	89.58	82.94	83.40	93.25	82.27	82.07	92.25
+ CPO (Ours, ALMA-13B-R)	**82.25**	**84.32**	**92.03**	83.98	**87.37**	**95.22**	83.34	**85.74**	**94.05**

Georgia Tech

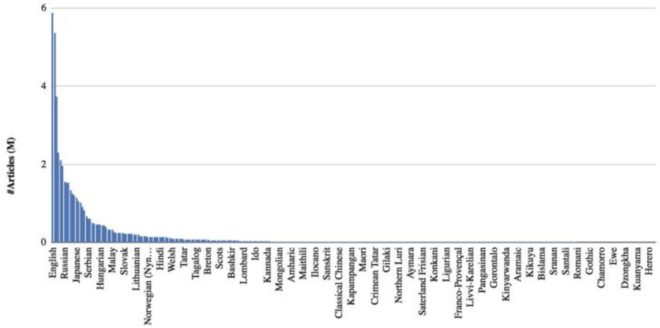
Overall Results for Multilingual Inputs (xx → en)

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

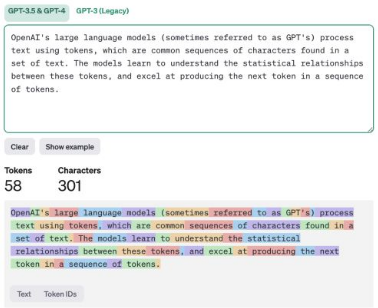
Discussion 1: For LLMs, what is more challenging? Translating multilingual input or multilingual output? Why?

Paucity of data

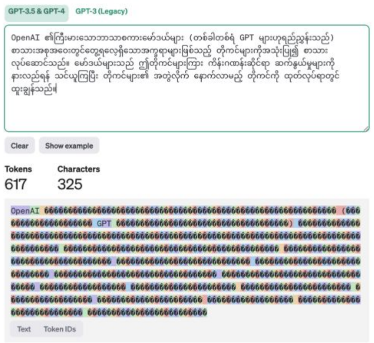


Tokenization Disparity

English



Burmese/Myanmar (Google Translated)



Similar content, 10.6x the tokens!

Discussion 2: Reliability of Reference-free Evaluation?

Model Evaluated on FLORES-200

	KIWI-XXL	Win Ratio (%)	XCOMET	Win Ratio (%)
Translating to English (xx→en)				
Reference	85.31	-	88.82	-
ALMA-13B-LoRA	88.33	73.24	92.68	60.17
GPT-4	89.21	79.43	94.66	54.25
Translating from English (en→xx)				
Reference	87.85	-	94.42	-
ALMA-13B-LoRA	85.62	42.15	93.07	35.46
GPT-4	87.30	49.13	94.21	38.09

	en → xx			xx → en		
	Avg.			Avg.		
	KIWI-22	KIWI-XXL	XCOMET	KIWI-22	KIWI-XXL	XCOMET
Gold Reference	82.05	83.47	92.85	79.91	80.10	85.77
WMT-Winners	**83.41**	84.81	93.78	80.92	81.19	87.13
GPT4	82.94	83.83	93.23	81.28	**82.60**	**89.41**
ALMA13B-LoRA	82.48	82.66	92.76	80.53	81.50	86.74
SFT	82.57	82.42	92.54	80.96	81.99	88.40
DPO	82.27	82.07	92.25	80.51	81.36	86.58
CPO	83.34	**85.74**	**94.05**	**81.33**	82.43	89.11

Georgia Tech

Are Translations Really Better of Just Metric-Preferred?

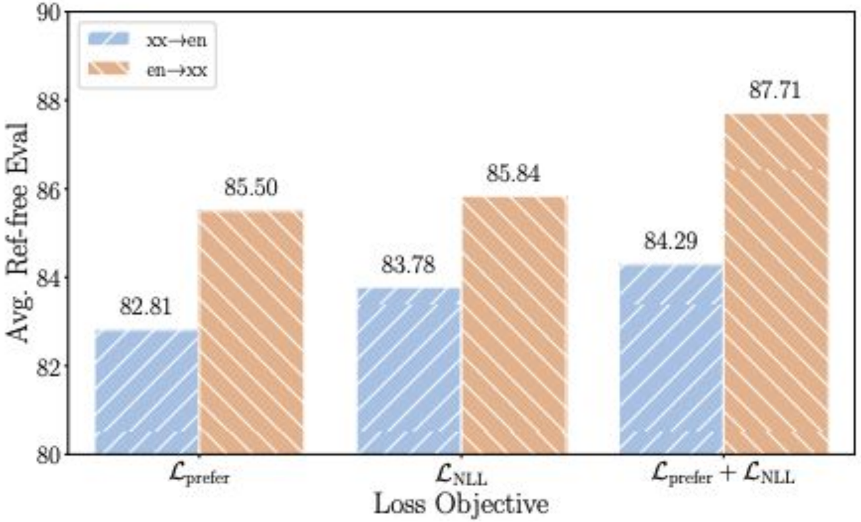
Models for Building Preference Data	KIWI-22	KIWI-XXL	XCOMET
Translating to English (xx→en)			
N/A (ALMA-13B-LoRA baseline)	80.53	81.50	86.74
KIWI-XXL	**81.33**	**82.59**	88.82
XCOMET	81.27	82.33	**89.17**
Ensemble of above (Original)	**81.33**	82.43	89.11
Translating from English (en→xx)			
N/A (ALMA-13B-LoRA baseline)	82.48	82.66	92.76
KIWI-XXL	83.31	**85.87**	93.97
XCOMET	83.09	85.43	**94.09**
Ensemble of above (Original)	**83.34**	85.74	94.05

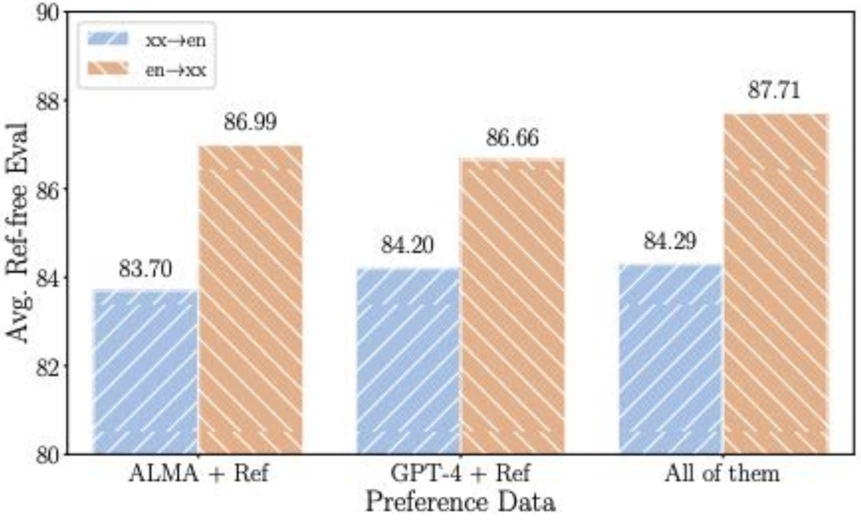
Human Evaluation on Sampled WMT 22 dataset

Human Evaluation on sampled zh → en

	Avg. score ↑	Avg. rank ↓	Avg. win ratio (%)	Ties (%)
ALMA-13B-LoRA	4.86	1.60	62.50	40.30
ALMA-13B-R	**5.16**	**1.40**	**77.80**	40.30

Ablation Study





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×
Translating from English (en→xx)					
$\mathcal{L}_{\text{DPO}}$	82.27	82.07	92.25	2×	2×
$\mathcal{L}_{\text{DPO}} + \mathcal{L}_{\text{NLL}}$	83.13	84.74	93.53	2×	2×
$\mathcal{L}_{\text{prefer}} + \mathcal{L}_{\text{NLL}}$ (CPO)	**83.34**	**85.74**	**94.05**	1×	1×

Impact of Human-Labeled Preference Data

Table 12. A comparison of translation performance when utilizing solely triplet data versus a combination of triplet data and human-labeled data (our original setup) in the en→xx direction. The **bold** number indicates superior performance. There is not obvious performance difference adding our human-labeled data.

Dataset	de			cs			is		
	KIWI-22	KIWI-XXL	XCOMET	KIWI-22	KIWI-XXL	XCOMET	KIWI-22	KIWI-XXL	XCOMET
Only Triplet Data	**83.43**	**84.63**	**97.56**	84.97	**87.24**	93.50	82.05	85.37	91.83
Triplet Data + Human-Labeled Data	83.28	84.25	97.48	**84.99**	87.06	**93.61**	**82.18**	**85.68**	**91.93**
Dataset	zh			ru			Avg.		
	KIWI-22	KIWI-XXL	XCOMET	KIWI-22	KIWI-XXL	XCOMET	KIWI-22	KIWI-XXL	XCOMET
Only Triplet Data	82.15	84.08	91.59	**84.05**	**87.43**	**95.26**	83.33	**85.75**	93.95
Triplet Data + Human-Labeled Data	**82.25**	**84.32**	**92.03**	83.98	87.37	95.22	**83.34**	85.74	**94.05**

Table 13. A comparison of translation performance when utilizing solely triplet data versus a combination of triplet data and human-labeled data (our original setup) in the en→xx direction. The **bold** number indicates superior performance. Interestingly, the inclusion of our human-labeled data results in a slight decrease in average performance.

Dataset	de			cs			is		
	KIWI-22	KIWI-XXL	XCOMET	KIWI-22	KIWI-XXL	XCOMET	KIWI-22	KIWI-XXL	XCOMET
Only Triplet Data	**81.57**	**84.25**	**94.32**	**82.68**	83.70	87.97	**81.63**	**85.87**	80.89
Triplet Data + Human-Labeled Data	81.50	83.97	94.20	82.63	**83.75**	**88.03**	81.57	85.73	**80.49**
Dataset	zh			ru			Avg.		
	KIWI-22	KIWI-XXL	XCOMET	KIWI-22	KIWI-XXL	XCOMET	KIWI-22	KIWI-XXL	XCOMET
Only Triplet Data	**79.34**	**77.31**	**91.76**	**81.76**	**81.63**	**91.34**	**81.40**	**82.55**	**89.26**
Triplet Data + Human-Labeled Data	79.24	77.17	91.65	81.72	81.54	91.18	81.33	82.43	89.11

Discussion 3: Limitation and Discussions?

Discussion 3: Limitations and Discussions

- Why did you pick only 5 out of 7 pairs of languages for WMT 21 and 22 challenge? 3 languages for WMT23→

	KIWI-22	KIWI-XXL	XCOMET
Gold Reference	78.74	75.56	86.30
WMT Winners	**80.57**	77.72	88.24
TowerInstruct	80.31	77.18	88.11
ALMA-13B-LoRA	79.48	76.00	87.16
+ CPO (Ours, ALMA-13B-R)	80.55	**78.97**	**89.74**

- How does ALMA-13B-R perform when compared with more sophisticated multilingual prompts?
  - Cross-Lingual Thought Prompting (Huang et. al., 2023) Cross-Lingual Consistent Prompting (Qin et. al, 2023), Cross-lingual Transfer Prompting (Kim et al., 2023), Prompts Augmented by Retrieval Cross Lingual (Nie et al., 2023), Chain-of-Dictionary (Lu et al., 2023) …

ALMA Prompt

Translate this from <source language> to <target language>:
<source language>: <source sentence>
<target language>: