Bridging the Inference Gap: Consistent Training for Low-Resource Cross-Lingual Intervention

Kshitij Ambilduke, Shubha Sanket Samantaray
ENS Paris-Saclay

Introduction

Language Models perform well overall but remain heavily English-centric

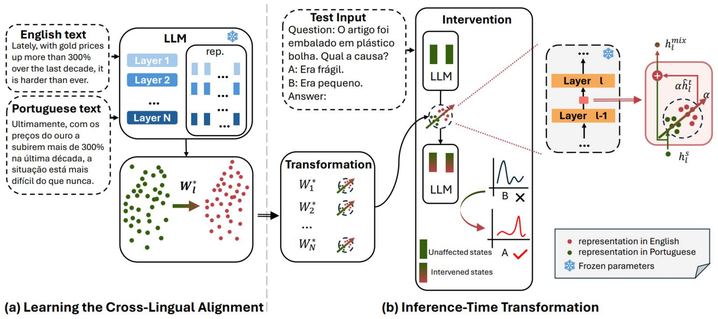
- Large performance gaps across low-resource languages
- Fine-tuning and retraining are costly

Question: *Can we reduce cross-lingual gaps without retraining?*

We propose INCLINE : Inference-Time Cross-Lingual Intervention

- Aligns internal representations across languages
- Transfers knowledge from high- to low-performing languages
- Applied only at inference time with no parameter updates

Methodology



(a) Learning the Cross-Lingual Alignment (b) Inference-Time Transformation

Training: *Learning Alignment*
We learn a matrix W_l^* to map source language representations (h^s) to the target language space (h^t) by minimizing the distance between them

$$W_l^* = argmin_{W_l} \sum_{i=1}^{N} ||W_l h_{i,l}^s - h_{i,l}^t||^2$$

Inference: *Transformation & Intervention*
During inference, we project the test input $(h_{q,l}^s)$ into the target space and blend it with the original representation using α, a strength hyperparameter

$$\text{Projection:} \quad \hat{h}_{q,l}^t = W_l^* h_{q,l}^s$$
$$\text{Intervention:} \quad h_{q,l}^{mix} = h_{q,l}^s + \alpha \hat{h}_{q,}^t$$

Where,
l is Layer index
h^s/h^t is Hidden states of source/target

Limitations

- **Over-reliance on XStoryCloze:** Most experiments focus on XStoryCloze (a discriminative task), while limited discussion is provided on generative tasks like Translation.
- **Focus on High-Resource (X) to English:** While the method has potential for low-resource languages, only high-resource (X) to (English) translation was considered.
- **Reliance on BLEU:** Evaluation is limited to BLEU scores even though LLMs tend to paraphrase significantly.

Improved methodology

Proposed Improvement: *Consistency*
INCLINE trains W for direct projection ($Wh^s \approx h^t$) but uses it for mixing ($h^s + \alpha W h^s$) during inference, creating an objective mismatch.

We modify the training objective to directly optimize the post-intervention vector, ensuring W learns the specific shift needed given the mixing factor α.
New Training Objective:

$$W_l^* = \arg\min_{W_l} \sum_{i=1}^{N} ||(h_{i,l}^s + \alpha W_l h_{i,l}^s) - h_{i,l}^t||^2$$

Implementation details

- We start with Llama-3-8B-Instruct as our base LLM.
- We focus exclusively on the low-resource translation task.
- All experiments use quantized model for viable compute.
- Sentence representations are extracted from the last token of both source and target sentences.
- We use L2 regularization since the system is highly over parameterized.
- Translation quality is evaluated using BLEU and ChrF++. For African languages, we additionally report AfriCOMET.

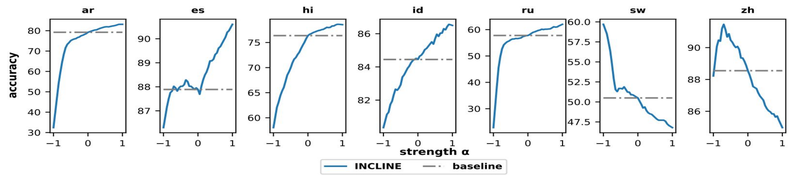
Evaluated Approaches

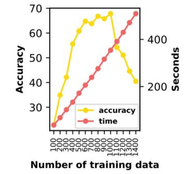
- **No intervention** uses the quantized base model for translation, with no alignment or model modification.
- **Intervention** applies the learned alignment matrices at inference time, modifying hidden states using a fixed intervention strength (referred to as INCLINE).
- **Consistent intervention** applies the proposed improvement in training the alignment matrices and applies these alignment matrices in the same way during inference (referred to as Improved INCLINE).
- **LoRA** trains low-rank adapters on the same FLORES translation data. Since both LoRA fine-tuning and INCLINE adds additional intermediate weights, by comparing both, we gauge into the effectiveness of the indirect INCLINE objective against the more apparent token prediction loss.

Takeaway

- INCLINE performs well even in severe low-resource settings.
- LoRA significantly outperforms INCLINE, but INCLINE offers a much quicker and more efficient way to improve base model.
- Improvement is consistent across scripts, though better for Latin scripts (e.g., Serbian Latin vs. Cyrillic).
- INCLINE shows improvements even for languages with very sparse tokenization (Lao & Burmese).
- Our proposed INCLINE improvement achieves better performance for African Languages while matching performance for others.
- The regularized objective is stable, unlike the non-regularized approach which failed on Sanskrit.

Relevant existing experiments





Flores	en	ar	el	es	fr	hi	ru	tr	vi	zh
BASELINE	-	66.59	15.30	48.52	67.86	71.97	35.66	12.38	40.40	56.11
INCLINE	-	**73.95**†	**15.79**†	**56.11**†	**75.84**†	**77.85**†	**39.33**†	**12.92**†	**48.62**†	**60.19**†

We use **500 parallel sentences** from the **FLORES** devtest set.

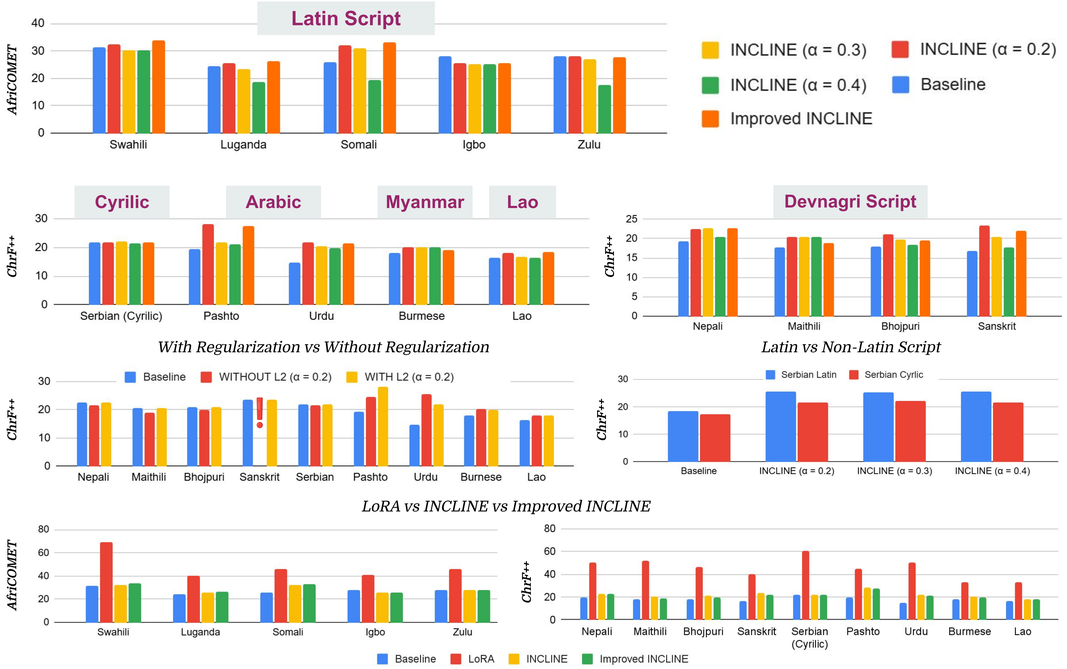
WMT23	en	de	ja	ru	uk	zh
BASELINE	-	18.26	10.17	14.73	11.36	14.39
INCLINE	-	**18.85**†	**10.30**†	**15.24**†	**11.71**†	**15.05**†

Notice that NONE of the languages considered for translation are low-resource

Our experiments

Questions we seek to answer:

- Whether INCLINE works well for low resource languages
- Whether performance of INCLINE is dependent on the script of the high resource language
- Whether tokenization density affects the performance.



References

- Wang, Jiayi et al. "AfriMTE and AfriCOMET: Enhancing COMET to Embrace Under-resourced African Languages." North American Chapter of the Association for Computational Linguistics (2023).
- Popović, Maja. "chrF++: words helping character n-grams." Proceedings of the second conference on machine translation. 2017.
- AI4Bharat et al. "IndicTrans2: Towards High-Quality and Accessible Machine Translation Models for all 22 Scheduled Indian Languages." ArXiv abs/2305.16307 (2023)
- Wang, Weixuan, et al. "Bridging the language gaps in large language models with inference-time cross-lingual intervention." Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025.
- team, Nllb et al. "No Language Left Behind: Scaling Human-Centered Machine Translation." ArXiv abs/2207.04672 (2022)
- Hu, J. Edward et al. "LoRA: Low-Rank Adaptation of Large Language Models." ArXiv abs/2106.09685 (2021)