Translation in LLMs

2026.01.29 (목)

홍성태

ghdchlwls123@korea.ac.kr

ICLR 2024

A PARADIGM SHIFT IN MACHINE TRANSLATION: BOOSTING TRANSLATION PERFORMANCE OF LARGE LANGUAGE MODELS

Haoran Xu[♠], **Young Jin Kim**[♡], **Amr Sharaf**[♡], **Hany Hassan Awadalla**[♡]

[♠]Johns Hopkins University, [♡]Microsoft

hxu64@jhu.edu
{youki,amrsharaf,hanyh}@microsoft.com

ICLR 2025

X-ALMA: PLUG & PLAY MODULES AND ADAPTIVE REJECTION FOR QUALITY TRANSLATION AT SCALE

Haoran Xu[°] **Kenton Murray**[◁] **Philipp Koehn**[◁] **Hieu Hoang**[°] **Akiko I. Eriguchi**[°] **Huda Khayrallah**[◇,‡]
[°]Microsoft, [◁]Johns Hopkins University, [◇]Amazon
haoranxu@microsoft.com, {kenton, phi}@jhu.edu,
{hihoan,akikoe}@microsoft.com, hudakh@amazon.com

A Paradigm Shift in Machine Translation: Boosting Translation Performance of Large Language Models

Haoran Xu[♠], **Young Jin Kim**[♡], **Amr Sharaf**[♡], **Hany Hassan Awadalla**[♡]

[♠]Johns Hopkins University, [♡]Microsoft

hxu64@jhu.edu
{youki,amrsharaf,hanyh}@microsoft.com

https://huggingface.co/collections/haoranxu/alma

https://arxiv.org/abs/2309.11674

Introduction

배경 및 문제 제기

- 번역 작업에서는 Decoder-only LLM들이 힘을 못씀, Closed Model 같은 초대형 모델만이 NLLB와 비등.
- 저자원 언어 번역에서는 여전히 한계를 보이며 특히 SLLM들과 기존 번역 모델 간의 격차는 더욱 뚜렷.
- 예를 들어, XGLM (7B)모델은 NLLB-1.3B 보다 BLEU가 30점이나 낮음.

기존 연구의 한계

1. 최근 7B나 13B 규모의 LLM을 활용해 번역 성능을 높이려는 연구들(BigTranslate, BayLing 등)이 있으나 성능 향상 폭이 적음
2. 이러한 미미한 성과가 부적절한 Training recipe 때문이라고 판단

Introduction

ALMA (Advanced Language Model-based trAnslator)를 제안

1단계: 일반적인 다국어 언어 지식 학습

➔ 대부분의 LLM이 영어 중심 데이터로 학습, 비영어 단일 언어 데이터로 미세 조정하여 해당 언어로 CPT.

2단계: 번역 학습

➔ 소량의 고품질 병렬 데이터를 사용하여 SFT

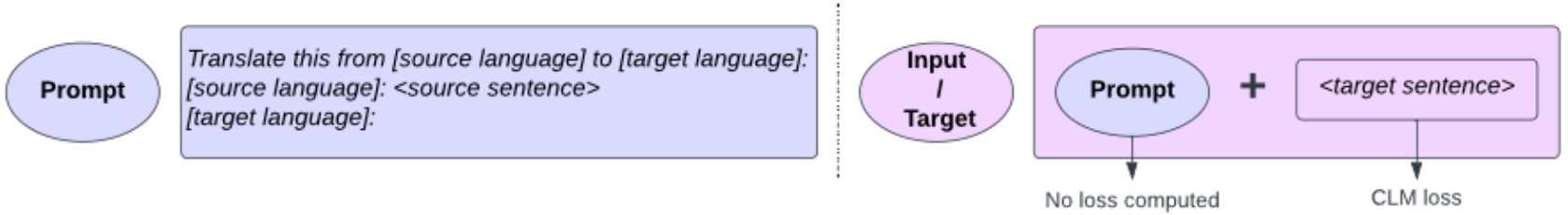
Contributions

1. 병렬 데이터 필요성 감소: 방대한 병렬 데이터가 모델의 고유 지식을 희석시켜 오히려 성능을 해칠 수 있음을 발견
2. ALMA 모델 공개: LLaMA-2 기반의 새로운 2-Stage FT를 통해, GPT-3.5 & NLLB-54B보다

KOREA UNIVERSITY

PRELIMINARY

TASK DEFINITION



- Decoder-only Transformer
- x: Source Sentence, y: Target Sentence, I: Fixed-Pr
- Hendy et al.(2023)과 동일한 프롬프트 사용

$$\mathcal{L}_{\text{NLL}}(\mathbf{x}, \mathbf{y}, \theta) = -\log P(\mathbf{y}|\mathbf{x}, \mathcal{I}; \theta)$$
$$= -\sum_{t=1}^{T} \log P(y_t|\mathbf{y}_{<t}, \mathbf{x}, \mathcal{I}; \theta),$$

A BACKBONE LLM FOR TRANSLATION

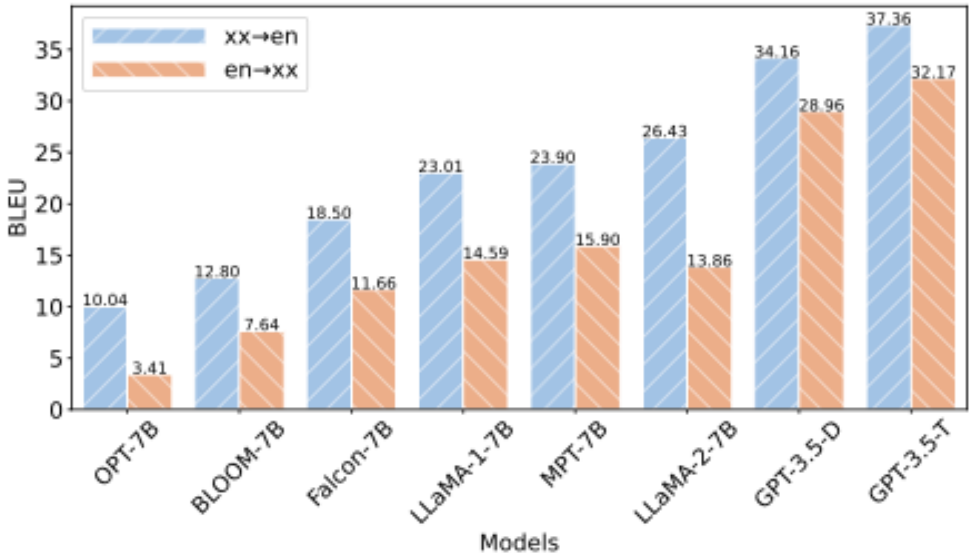
: Zero-shot 번역 성능을 통해 LLM 을 우선적으로 평가

- 비교모델: OPT-7B, Falcon-7B, BLOOM-7B, MPT-7B, LLaMA-1-7B, LLaMA-2-7B.
- 베이스라인: GPT-3.5-text-davinci-003 , GPT-3.5-turbo-0301의 결과도 함께 제시

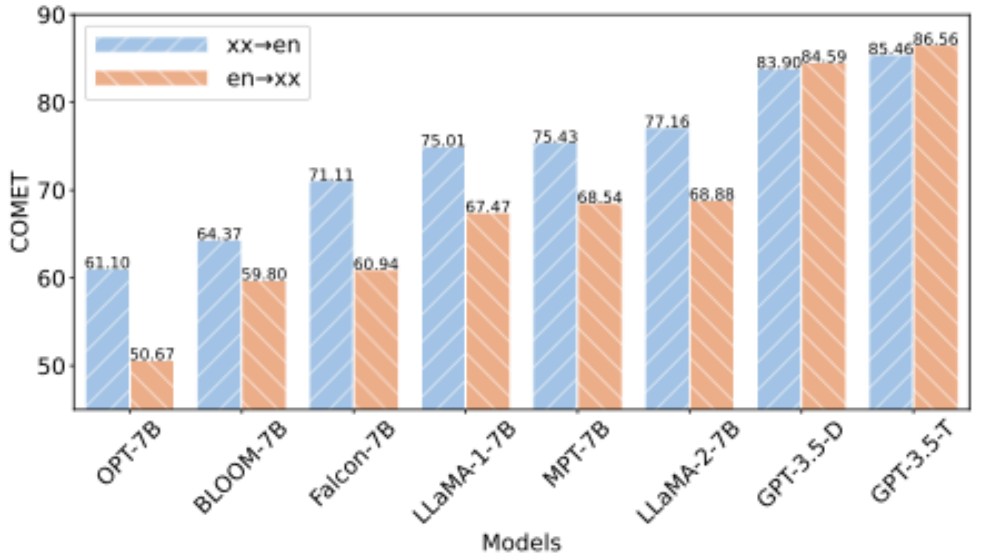
PRELIMINARY

A BACKBONE LLM FOR TRANSLATION (cont.)

- 5개 언어 (독일어, 체코어, 아이슬란드어, 중국어, 러시아어) ⇔ 영어, 양방향 번역을 평가
- WMT'21, WMT'22 / Beam: 5 / sacreBLEU / COMET



(a) BLEU



(b) COMET

➔ LLaMA-2-7B와 MPT-7B를 백본 모델로 선택

DO LLMS HAVE AN APPETITE FOR PARALLEL DATA?

A BACKBONE LLM FOR TRANSLATION

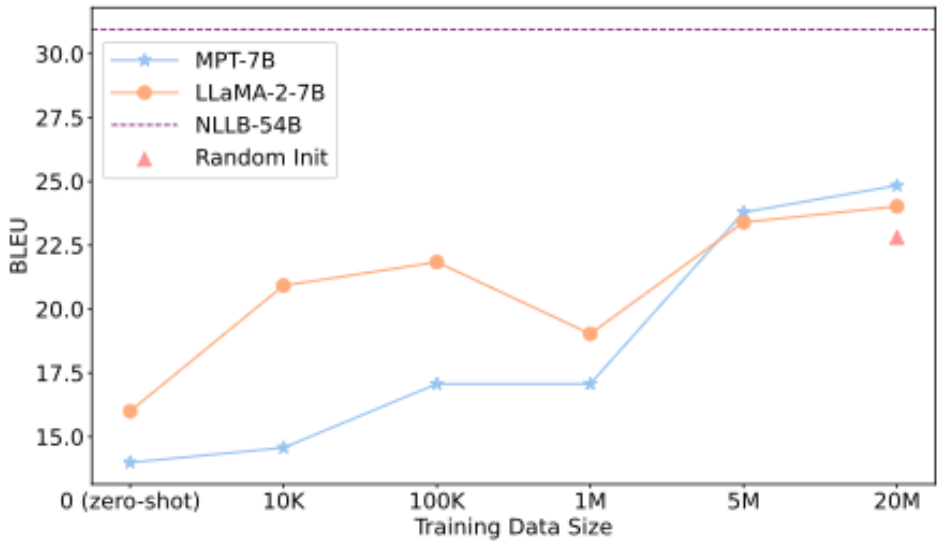
- 전통적인 기계 번역 훈련은 주로 인코더-디코더 프레임워크 내에서 방대한 양의 병렬 데이터셋을 활용
- Decoder-only LLM에서 여전히 상당한 양의 병렬 데이터에 의존하나? 그리고 전통적인 훈련 패러다임을 따르나?

Experimental Design

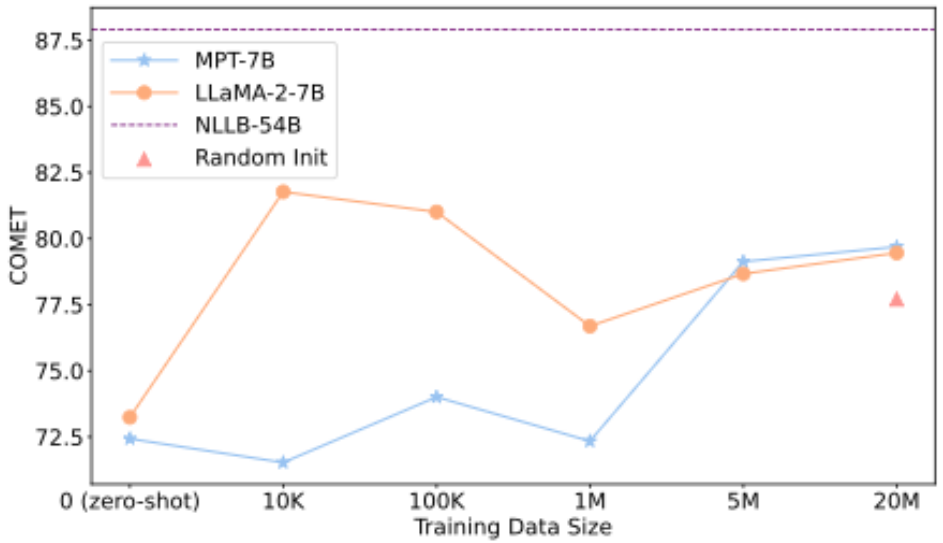
- LLaMA-2-7B와 MPT-7B의 SFT
- 심층적인 분석을 위해 En→Ru 언어 쌍에만 초점
- 7,500만 개의 Parallel set에서 필터링, 데이터 크기를 10K, 100K, 1M, 5M, 20M의 5단계로

DO LLMS HAVE AN APPETITE FOR PARALLEL DATA?

OBSERBATIONS



(a) BLEU



(b) COMET

- 데이터가 많다고 무조건 좋은 것은 아니더라!
- 적은 훈련 데이터로도 충분함: COMET 점수에 10~ 100K 데이터에서 정점을 찍은 후 성능이 하락
- MPT-7B는 데이터가 많아질수록 성능이 계속 향상
- LLaMA-2-7B는 유능한 번역을 수행하는 데 단 10K~100K 만으로도 충분

A NEW TRAINING RECIPE

① Monolingual data CPT & ② High-quality parallel data SFT ➔ ALMA

① Monolingual Data CPT

: 다국어 능력이 필수적인 번역 작업에서 다국어 성능 개선을 위해 번역 작업에 포함된 비영어권 언어의 단일 언어 데이터로 모델 학습

1) 영어 지식의 망각을 방지하기 위해 미세 조정 과정에 영어 단일 언어 데이터도 함께 추가

| 이전 연구들(Tan et al., 2023; BigTranslate; PolyLM; Li et al., 2023) 또한 단일 언어 데이터가 번역 성능 향상에 도움이 된다는 단서 제공

2) 단 10억 개의 토큰(6개 언어 혼합)만으로도 10개 번역 방향에서 상당한 성능 향상을 이룰 수 있음을 보여줌

A NEW TRAINING RECIPE

① Monolingual data CPT & ② High-quality parallel data SFT ➔ ALMA

② High-quality parallel data SFT

: 3.2절의 실험과 데이터 품질의 중요성을 강조한 기존 연구들에 기반하여, 작지만 고품질인 병렬 데이터셋으로 SFT

1) WMT test & Flores-200 dev/test set : Human-written Dataset 수집

2) Full SFT & LoRA 모두 탐구

| LoRA: FFN, Down Projection

Exp Setup

Dataset

1) CPT: OSCAR Corpus, 언어별 샘플링 비율은 De(20%), Cs(14%), Is(8%), Zh(19%), Ru(22%), En(17%)로 설정.
2) Train: WMT'17~'20 Test & Flores-200의 dev, test 수집. (모든 언어에 대해 총 58,000개)
3) Test 앞선 연구와 동일하게 10개 번역 방향(cs↔en, de↔en, is↔en, zh↔en, ru↔en)을 사용

Train

LLaMA-2-7B 13B에 대해 Full Tuning, Lora (Rank=16) 각각 학습

Baseline

BigTranslate, TIM, SWIE, BayLing, NLLB-54B / GPT-3.5-D, GPT-3.5-T, GPT-4

Exp Results

En => XX

Models	de		cs		is		zh		ru		Avg.	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
SoTA Models												
NLLB-54B	34.50	86.45	37.60	90.15	24.15	81.76	27.38	78.91	30.96	87.92	30.92	85.04
GPT-3.5-D, zero-shot	31.80	85.61	31.30	88.57	15.90	76.28	38.30	85.76	27.50	86.74	28.96	84.59
GPT-3.5-T, zero-shot	34.40	87.00	32.92	90.17	18.74	81.04	44.90	87.00	29.90	87.60	32.17	86.56
GPT-4, zero-shot	35.38	87.44	34.53	90.77	-	-	43.98	87.49	30.45	88.87	-	-
Prior Similar Studies												
TIM-BLOOMZ-7B	20.63	74.16	-	-	-	-	37.20	84.89	-	-	-	-
TIM-LLaMA-1-7B	25.59	82.56	-	-	-	-	19.33	75.46	-	-	-	-
SWIE-BLOOMZ-7B	21.83	75.17	-	-	-	-	36.88	84.53	-	-	-	-
SWIE-LLaMA-1-7B	27.21	82.36	-	-	-	-	31.24	80.63	-	-	-	-
BigTranslate-13B	21.48	78.81	20.67	80.65	2.28	35.56	28.56	81.31	17.66	78.21	18.13	70.91
Bayling-13B	25.62	82.69	16.43	78.22	-	-	37.92	84.62	12.77	71.01	-	-
Our Recipe with Backbone Model: LLaMA-2-7B												
LLaMA-2-7B, zero-shot	19.00	76.39	16.02	79.13	1.33	43.83	16.97	71.80	16.00	73.24	13.86	68.88
ALMA-7B (Ours)	30.31	85.59	29.88	89.10	25.71	85.52	36.48	85.05	27.09	87.17	29.89	86.49
ALMA-7B-LoRA (Ours)	30.16	85.45	30.17	89.05	25.19	85.44	36.47	84.87	26.93	87.05	29.78	86.37
Our Recipe with Backbone Model: LLaMA-2-13B												
LLaMA-2-13B, zero-shot	13.69	75.55	0.87	68.57	2.36	38.47	30.00	79.70	0.59	63.84	9.50	65.23
ALMA-13B (Ours)	31.37	85.45	31.12	89.42	26.67	85.85	39.05	85.76	28.76	87.50	31.39	86.80
ALMA-13B-LoRA (Ours)	**31.47**	**85.62**	**32.38**	**89.79**	**26.68**	**86.08**	**39.84**	**85.96**	**28.96**	**87.53**	**31.87**	**87.00**

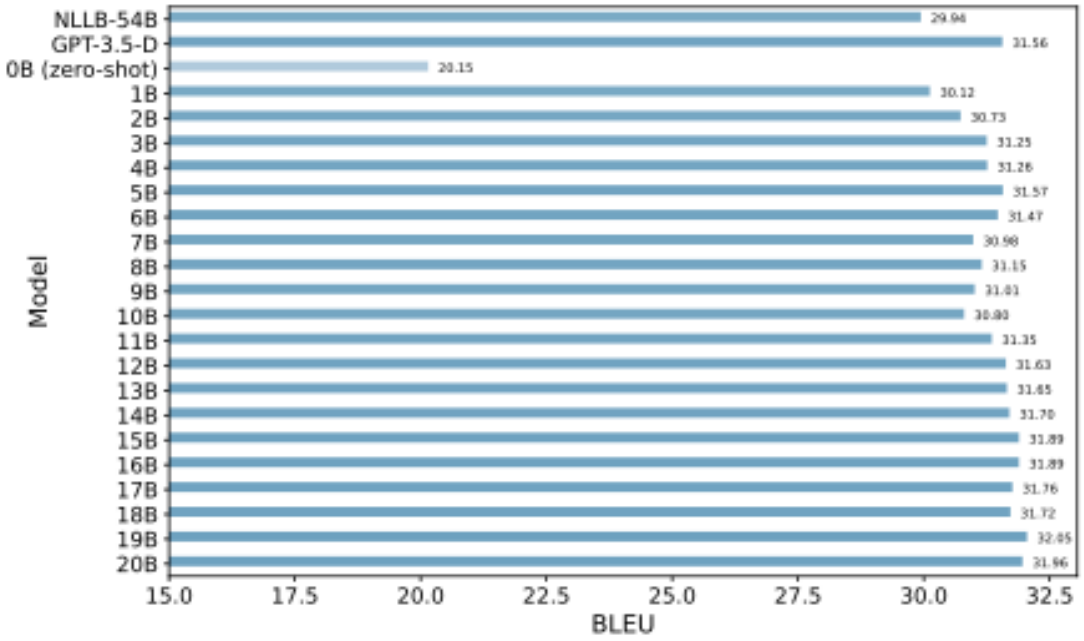
Exp Results

XX => En

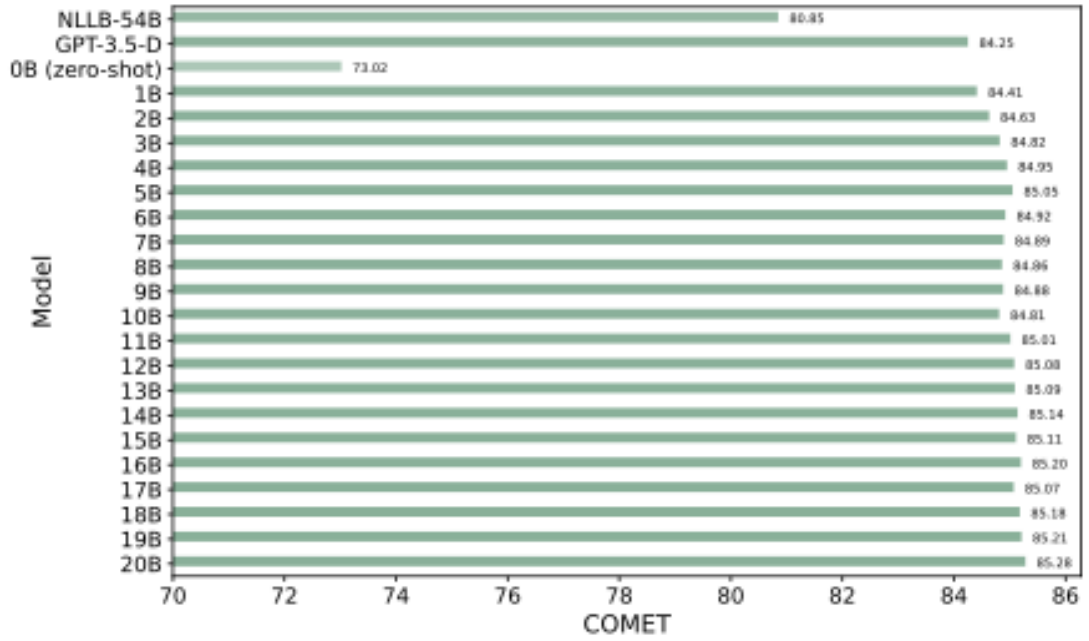
Models	de		cs		is		zh		ru		Avg.	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
SoTA Models												
NLLB-54B	26.89	78.94	39.11	80.13	23.09	71.66	16.56	70.70	39.11	81.88	28.95	76.66
GPT-3.5-D, zero-shot	30.90	84.79	44.50	86.16	31.90	82.13	25.00	81.62	38.50	84.80	34.16	83.90
GPT-3.5-T, zero-shot	33.10	85.50	47.20	87.30	37.50	85.50	26.60	82.90	42.40	86.10	37.36	85.46
GPT-4, zero-shot	33.87	85.62	48.67	87.43	-	-	27.20	82.79	43.51	86.18	-	-
Prior Similar Studies												
TIM-BLOOMZ-7B	24.31	77.65	-	-	-	-	23.42	79.50	-	-	-	-
TIM-LLaMA-1-7B	27.91	82.80	-	-	-	-	19.33	75.46	-	-	-	-
SWIE-BLOOMZ-7B	25.95	78.80	-	-	-	-	23.40	79.36	-	-	-	-
SWIE-LLaMA-1-7B	30.48	82.97	-	-	-	-	21.30	76.48	-	-	-	-
BigTranslate-13B	23.35	80.68	33.67	81.19	6.51	54.71	14.16	74.26	26.81	77.80	20.90	73.80
Bayling-13B	27.34	83.02	33.87	81.65	-	-	20.12	77.72	33.95	82.07	-	-
Our Recipe with Backbone Model: LLaMA-2-7B												
LLaMA-2-7B, zero-shot	30.42	82.74	36.56	82.42	10.98	62.79	18.19	75.00	36.02	82.84	26.43	77.16
ALMA-7B (Ours)	29.49	83.98	42.91	85.90	35.26	85.97	23.52	79.73	38.93	84.81	34.02	84.08
ALMA-7B-LoRA (Ours)	29.56	83.95	43.49	85.93	35.64	86.09	23.64	79.78	39.21	84.84	34.31	84.12
Our Recipe with Backbone Model: LLaMA-2-13B												
LLaMA-2-13B, zero-shot	31.06	83.01	40.02	83.27	15.77	66.35	21.81	78.10	36.50	82.91	29.03	78.73
ALMA-13B (Ours)	30.73	84.42	44.68	86.29	36.46	86.30	24.65	79.90	40.37	85.09	35.38	84.40
ALMA-13B-LoRA (Ours)	**31.14**	**84.56**	**45.28**	**86.47**	**36.95**	**86.42**	**25.46**	**80.21**	**40.27**	**85.27**	**35.82**	**84.59**

Analyses

HOW MUCH MONOLINGUAL DATA TO USE?



(a) BLEU



(b) COMET

Analyses

THE EFFECT OF MONOLINGUAL DATA AND PARALLEL DATA QUALITY

Use Mono.	Parallel Data Quality	Avg. xx→en		Avg. en→xx	
		BLEU	COMET	BLEU	COMET
✘	✘	26.43	77.16	13.86	68.88
✘	Random	28.24	78.69	19.68	73.89
✘	Filtered	28.39	78.94	19.56	74.35
✘	HW	29.39	80.00	22.17	76.52
✔	✘	28.49	80.32	26.35	84.73
✔	Random	32.47	83.02	26.98	83.15
✔	Filtered	32.32	83.03	27.38	83.98
✔	HW	**34.02**	**84.08**	**29.89**	**86.49**

X-ALMA: Plug & Play Modules and Adaptive Rejection for Quality Translation at Scale

Haoran Xu[°] **Kenton Murray**[◁] **Philipp Koehn**[◁] **Hieu Hoang**[°] **Akiko I. Eriguchi**[°] **Huda Khayrallah**[◇,‡]
[°]Microsoft, [◁]Johns Hopkins University, [◇]Amazon
haoranxu@microsoft.com, {kenton, phi}@jhu.edu,
{hihoan,akikoe}@microsoft.com, hudakh@amazon.com

https://huggingface.co/collections/haoranxu/x-alma

https://arxiv.org/abs/2410.03115

Introduction

배경 및 문제 제기

- Multilingual LLM이 수백 개의 언어를 지원한다고 주장
- 실제로는 mid- and low-resource lanugages에 대해 고품질의 응답을 제공하지 못함
- 성능이 고자원 언어에 크게 치우치는 불균형 문제를 지적

당시 흐름..

1. LLaMA 3 및 3.1은 어휘 집합을 32K에서 128K로 확장 및 다국어 사전 학습 수행,
2. Aya-101은 101개 언어를 지원하는 다국어 생성 모델로 도입
3. 또한 BigTranslate와 LLaMAX는 LLM 기반 다국어 번역 모델을 100개 이상의 언어로 확장

Introduction

배경 및 문제 제기 (cont.)

모델이 지원하는 언어 수가 늘어났음에도 두 가지 주요 약점이 존재

1. Mid-Low에서 성능이 고자원 만큼 기대 성능을 충족하지 못함.
2. 지원 언어가 늘어날수록, 고자원 언어의 성능이 더 적은 언어로 학습된 모델보다 떨어짐 (Curse of Multilinguality)

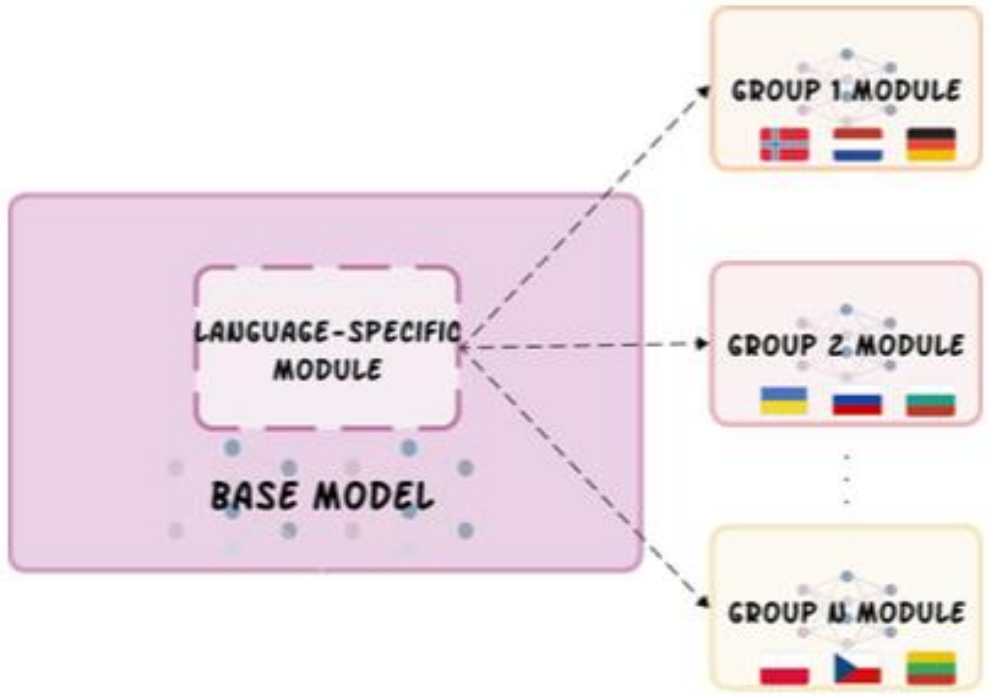
➔ 이러한 약점을 해결하고 자원 수준에 무관한 50개의 다양한 언어에서 높은 성능을 달성하는 X-ALMA 공개

Contributions

1. Plug-and-Play Architecture: 용량 문제를 해결을 위해 유사한 언어별로 서빙하는 여러 언어별(LS) 모듈을 설계

Methods

1. Model Architecture



- 1) Dense base model과 (2) language-specific 모듈로 구성
- MoE 방식과 유사하지만, LS 모듈에게 할당하기 위해 신경망 기반의 게이트를 사용하지 않음
- Hard-Gating 사용, 해당 언어에 지정된 모듈에만 독점적으로 할당

Methods

1. Model Architecture (cont.)

Details

Base Model: LLaMA-2 arch.

Language Specific: LoRA (Attn + MLP, 15%)

Why This Design?

- 선택적 로딩: 필요한 단일 LS 모듈만 로드하여 GPU 메모리 제약을 완화
- 모델 병합: LS 모듈을 베이스 모델과 병합하여 일반 모델과 동일한 파라미터 수를 가진 새로운 모델을 생성
- 결합 로딩: 베이스 모델과 모든 LS 모듈을 함께 로드하여 MoE처럼 운용 가능

Methods

2. LANGUAGE GROUPING

: 14개의 script와 18개의 language family를 포함하는 총 50개의 언어를 고려

- 두 가지 기준에 따라 8개 그룹으로 나눔 (언어학적 기준으로 나눔, 영어는 모든 그룹에 포함)

(1) 각 그룹은 언어적으로 최대한 유사해야 하며,

(2) 각 그룹에 포함된 언어 수의 균형이 맞아야 합니다.

Table 1: Language grouping based on linguistic features and balanced number of languages.

Group ID	Linguistic Feature	Languages
1	Germanic languages	af, da, de, is, nl, no, sv, (en)
2	Romance Languages	ca, es, gl, it, pt, ro, (en)
3	Eastern and Southern Slavic Languages	bg, mk, ru, sr, uk, (en)
4	Southeast Asian Languages	fr, id, mg, ms, th, vi, (en)
5	Central and Eastern European Languages	cs, el, hu, lt, lv, pl, (en)
6	Eurasian Language Mix	et, fi, ja, ka, ko, zh, (en)
7	Indo-Aryan Languages	gu, hi, mr, ne, ur, (en)
8	Turkic and Semitic Languages	ar, az, fa, he, kk, ky, tr, uz, (en)

Methods

3. Training Recipe

Pre-Training Stage 1: Monolingual Fine-Tuning Base Model

- 베이스모델에 대해 50개 언어의 20B 토큰을 사용한 CPT

Pre-Training Stage 2: Monolingual Fine-Tuning Language-Specific Modules

- 베이스 모델 Freeze, 각 LS 모듈은 해당 그룹 언어의 10B 토큰 CPT

Pre-Training Stage 3: Pseudo-Monolingual Fine-Tuning

- 병렬 데이를 무작위 순서로 연결, 유사 단일 언어 데이터를 사용한 CPT. 각 모듈당 1.25B 토큰 사용

Post-Training Stage 1: Supervised Fine-tuning

- 고품질인 병렬 데이터셋을 사용한 SFT

Post-Training Stage 2: Preference Optimization

- 번역 품질을 향상을 위한 Adaptive Rejection Preference Optimization(ARPO) 적용

ARPO

4.1. Limitations In Current Preference Learning

- DPO 데이터셋 구축 시, 유의미한 모델 개선을 보장하기 위해서는 선호되지 않는 번역문 또한 고품질이여야함
- 이로 인해 선호되는 번역과 선호되지 않는 번역이 종종 매우 유사하며 단 몇 개의 단어만 다른 시나리오가 존재
- 이는 일반적인 Open Domain QA에서 사용되는 선호도 데이터와는 상당히 다름
- 기존의 방식이 선호되는 번역과 유사한 선호되지 않는 번역 전체를 거부하는 경향이 존재. ➔ MT에 부적합
- 이러한 접근 방식은 선호되는 번역 대부분의 토큰까지 의도치 않게 거부하게 만듦, 이를

ARPO

4.2. Adaptive Rejection

: Over-Rejection 현상을 완화하기 위해, 손실 함수에서 dis-preferred term의 강도를 조절

$$\mathcal{L}_{ARPO} = \mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}[f(r_\theta(y_w|x) - \tau_\theta(y_w,y_l)\cdot r_\theta(y_l|x))]$$

$$\tau_\theta(y_w,y_l) = \min(e^{\eta\cdot z_\theta(y_w,y_l)} - 1, 1)$$

$$z_\theta(y_w,y_l) = \text{abs}\left(\frac{\log(\pi_\theta(y_w|x))}{|y_w|} - \frac{\log(\pi_\theta(y_l|x))}{|y_l|}\right)$$

- 선호되는 응답과 선호되지 않는 응답 사이의 거리를 average log-likelihoods로 수치화

Exp Setup

Dataset

1) CPT: 1, 2단계 OSCAR Corpus, 3단계 NLLB, OPUS parallel set (50개 언어, 98개 방향)
2) SFT: Flores-200 dev & NTREX Test set & WNT'15-22 Test set (방향당 평균 4k)
3) PO: SFT를 마친 X-ALMA로 번역본 생성한 뒤, gold를 'win', 생성 번역을 'lose'로 지정, 고자원의 GPT-4o로 수정

Evaluation

Set: Flores-200 test, WMT'23 / Metric: COMET-22, XCOMET-XL (No-Ref), BLEU

Baseline

NLLB-200, Aya-101, LLaMAX3-Alpaca, Aya-23-8B/35B, LLaMA-3.1-8B-Instruct

Training Setup

Backbone: LLaMA-2

Exp Results

Models	Group 1		Group 2		Group 3		Group 4		Group 5		Group 6		Group 7		Group 8	
	All	High	All	High	All	High	All	High	All	High	All	High	All	High	All	High
LLaMA-3.1-8B-Instruct	80.8	79.8	83.7	84.2	79.1	69.2	76.3	85.8	79.0	81.2	71.1	71.8	70.1	69.6	78.3	84.4
NLLB-3.3B	88.2	88.8	88.3	88.1	89.4	89.1	87.1	88.2	89.2	89.8	87.5	87.5	80.1	80.9	88.1	87.5
LLaMAX3-Alpaca-8B	86.4	86.9	86.8	86.6	85.7	82.0	81.7	86.2	86.6	87.1	86.0	87.3	76.5	76.6	82.6	83.6
Aya-101	85.0	85.7	86.8	86.2	87.7	85.6	85.8	85.5	88.4	88.7	87.5	87.3	76.2	75.5	86.8	86.3
Aya-23-8B	75.1	84.7	86.6	86.6	74.4	75.7	74.6	88.7	70.6	77.3	67.1	79.8	68.9	79.3	76.0	87.9
Aya-23-35B	79.6	86.5	87.1	87.0	77.6	78.5	76.7	88.6	82.1	86.0	73.9	84.4	61.9	79.1	68.8	87.8
X-ALMA (only SFT)	89.5	89.7	89.2	88.9	90.7	90.2	88.1	89.1	90.6	90.7	90.1	90.4	82.6	81.4	89.2	88.9
X-ALMA	**89.6**	**89.9**	**89.4**	**89.0**	**90.9**	**90.5**	**88.6**	**89.5**	**91.0**	**91.1**	**90.6**	**90.8**	**83.2**	**81.9**	**89.4**	**89.2**

Models	Group 1		Group 2		Group 3		Group 4		Group 5		Group 6		Group 7		Group 8	
	All	High	All	High	All	High	All	High	All	High	All	High	All	High	All	High
LLaMA-3.1-8B-Instruct	68.8	77.6	70.9	76.9	51.2	53.8	65.6	76.4	54.8	60.2	58.8	66.9	47.6	53.7	53.7	67.5
NLLB-3.3B	79.1	81.8	84.5	85.0	84.3	83.8	81.1	85.4	74.9	76.0	76.1	77.3	88.3	88.9	79.5	81.6
LLaMAX3-Alpaca-8B	88.3	88.5	88.1	87.9	87.0	86.8	86.2	87.9	87.0	87.2	81.7	87.7	83.6	88.9	84.7	87.7
Aya-101	87.2	88.2	87.6	87.6	85.4	85.5	86.2	87.6	86.3	86.5	86.5	86.7	84.8	87.5	85.9	87.1
Aya-23-8B	84.6	88.2	87.9	87.7	83.3	83.3	79.8	88.5	82.9	85.2	79.9	86.1	71.8	89.1	76.7	88.0
Aya-23-35B	87.4	88.9	88.8	88.6	86.3	86.2	82.3	88.7	86.4	87.2	85.9	88.0	79.4	89.6	82.7	88.6
X-ALMA (only SFT)	89.1	89.2	88.8	88.6	87.9	87.7	87.7	88.8	87.9	88.1	88.2	88.3	89.3	89.8	87.5	88.4
X-ALMA	**89.4**	**89.5**	**89.2**	**89.0**	**88.1**	**87.8**	**88.0**	**88.9**	**88.2**	**88.4**	**88.7**	**88.8**	**89.6**	**90.1**	**88.0**	**88.9**

Analysis

1. Preference Optimization Comparison

ARPO를 DPO, KTO, ORPO, SimPO, CPO와 비교. 공정한 비교를 위해, BC 정규화 항을 추가

Over-Rejection 문제: 다른 방식들은 어휘 일치도를 나타내는 BLEU 점수가 크게 하락,

➔ 이는 모델이 선호 데이터의 Writing style에서 너무 멀어지기 때문

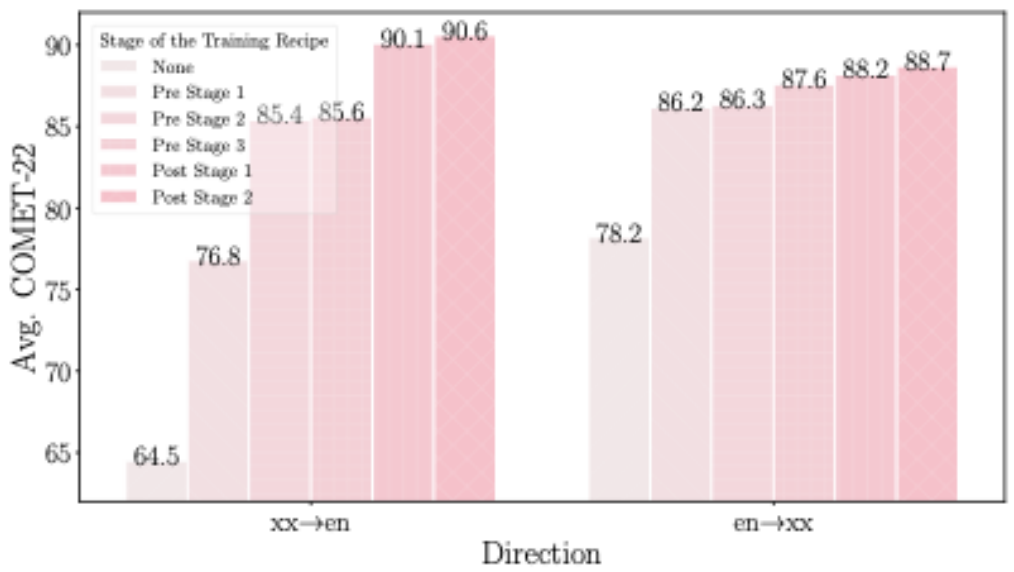
Table 5: Average performance comparison of various preference optimization methods for en→xx and xx→en on Group 6.

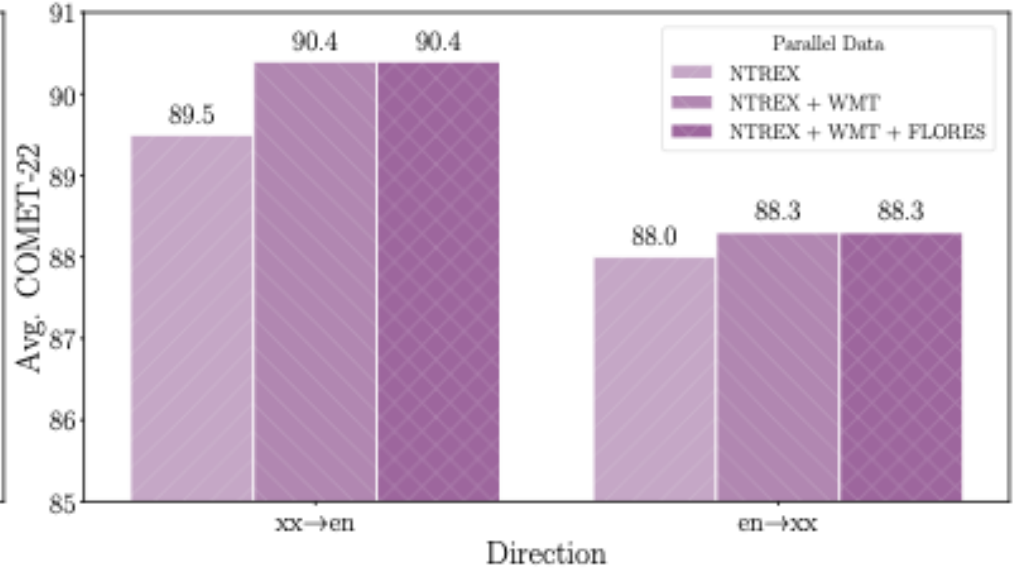
Models	Avg. en→xx			Avg. xx→en		
	BLEU	COMET-22	XCOMET-XL	BLEU	COMET-22	XCOMET-XL
XALMA (only SFT)	26.5	90.1	80.3	32.1	88.2	77.4
+ DPO	0.7	53.6	51.1	7.1	79.2	64.6
+ BC	23.5	90.2	80.0	27.8	87.5	77.4
+ SimPO	0.0	16.7	1.5	0.0	16.4	8.9
+ BC	23.3	89.7	78.7	26.6	87.1	76.5
+ KTO	22.1	89.8	79.2	26.4	87.1	76.5
+ BC	26.4	90.3	80.4	29.2	87.5	77.1
+ ORPO	23.0	85.8	75.9	22.7	81.8	70.8
+ CPO	22.2	90.2	79.9	26.5	87.8	77.0
+ ARPO (Final X-ALMA)	**27.8**	**90.6**	**81.3**	**32.2**	**88.7**	**78.4**

ARPO의 해결책: CPO 같은 방식은 의미는 통하지만 참조 문장과 스타일이 다른 번역을 남발하여 오류를 만듦

Analysis

2. Ablatin Stu





Recipe

- 사전 학습부터 사후 학습까지 각 단계를 거칠 때마다 성능이 일관되게 향상

SFT를 위한 병렬 데이터 구성

- NTREX 데이터만으로도 이미 인상적인 성능을 보여줌
- WMT 추가 시 영어로의 번역 성능 향상, 이는 영어 데이터의 다양성이 과적합되는 것을 막아줌
- 반면 Flores-200 dev 데이터를 추가는 도움 X

감사합니다