Rethinking Multilinguality: Balancing Cross-Lingual Uniformity and Locality in Multilingual Models

HyoJung Han

DEPARTMENT OF COMPUTER SCIENCE

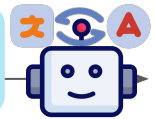
Guest Lecture at CMSC848T, April 2026

What Makes a Good “Multilingual” Model?

What Makes a Good “Multilingual” Model?

Uniformity

60% of the body is water.



体の約60%は水でできている。

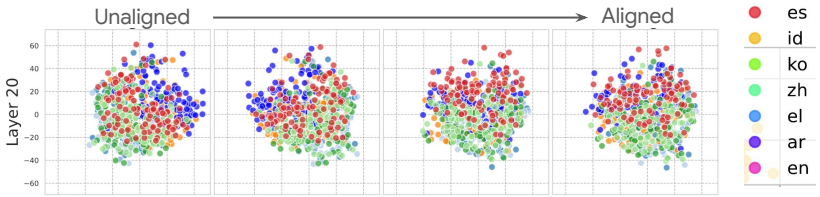
What % of the body is water?

우리 몸은 몇 퍼센트가 물인가요?

Πόσο % νερό είναι το σώμα;

体の何パーセントが水ですか？ → 60% ✓

Πόσο % νερό είναι το σώμα; → 60% ✓



* Example from CulturalBench: A Robust, Diverse and Challenging Benchmark for Measuring LMs' Cultural Knowledge Through Human-AI Red-Teaming (Chiu et al., ACL 2025)

What Makes a Good "Multilingual" Model?

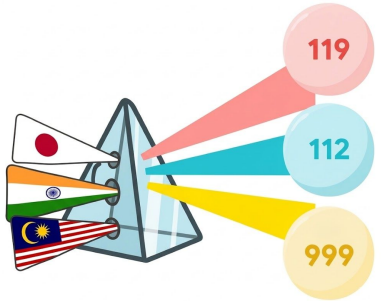
Locality / Cultural Awareness

In Singapore culture, what item is commonly used to reserve seats at a hawker center?*

Tissue papers

Source …frère de Dominique de Villepin…
Literal Translation …brother of Dominique de Villepin …
Translation with Explicitation …brother of the former French Prime Minister Dominique de Villepin …

🇺🇸😀 "Ah ha"



What's the ambulance number?

Both perspectives have been addressed independently rather than jointly

Uniformity

60% of the body is water.

体の約60%は水でできている。

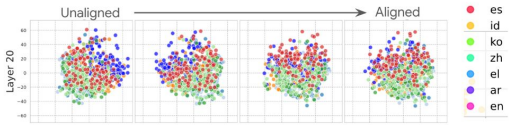
体の何パーセントが水ですか？ → 60% ✓

Πόσο % νερό είναι το σώμα; → 60% ✓

What % of the body is water?

우리 몸은 몇 퍼센트가 물인가요?

Πόσο % νερό είναι το σώμα;



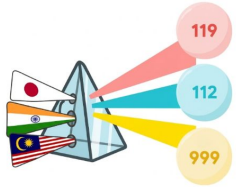
Locality

In Singapore culture, what item is commonly used to reserve seats at a hawker center?*

Tissue papers

Source …frère de Dominique de Villepin…
Literal Translation …brother of Dominique de Villepin …
Translation with Explicitation …brother of the former French Prime Minister Dominique de Villepin …

"Ah ha"

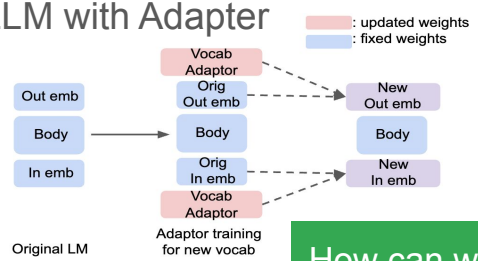


What's the ambulance number?

In this talk…

Uniformity

Vocabulary Transfer of Multilingual LLM with Adapter



Adapters for Altering LLM Vocab… Benefit the Most? (Han et al., ICL…

Locality

Explanatory Translation for Bridging **Cultural Background** Knowledge **Gap**

Source …frère de Dominique de Villepin…

Literal Translation …brother of Dominique de Villepin …

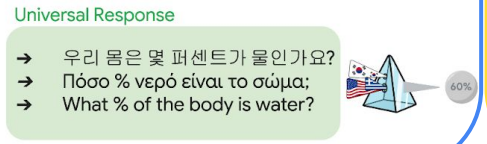
…brother of the former French Prime …ique de Villepin …

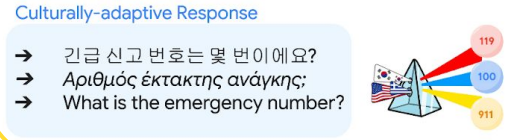
…ge Gaps in Translation …Han et al., EMNLP 2023)

How can we think about both sides together, and how can models better balance the two?

How can we construct new vocabulary representations … existing models to promote uniformity at the surface lev…

How can we frame and evaluate language generation tasks that adapt to users culture and background knowledge?



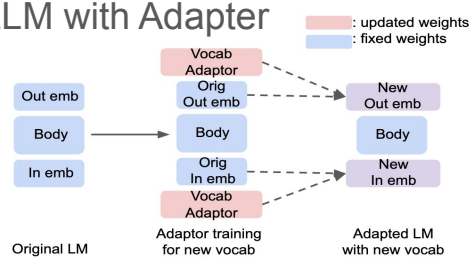


Rethinking Cross-lingual Alignment: Balancing Transfer and Cultural Erasure in Multilingual LLMs (Han et al., arXiv 2025)

Promoting Uniformity at the Surface Level

Uniformity

Vocabulary Transfer of Multilingual LLM with Adapter



Adapters for Altering LLM Vocabularies: What Languages Benefit the Most? (Han et al., ICLR 2025)

How can we construct new vocabulary representations using existing models to promote uniformity at the surface level?

Issue of over-fragmentation (words are excessively split by the tokenizer)

This mountain range, which is one of the largest in Europe, stretches across the territory of eight countries.
(**22** tokens by Mistral tokenizer)

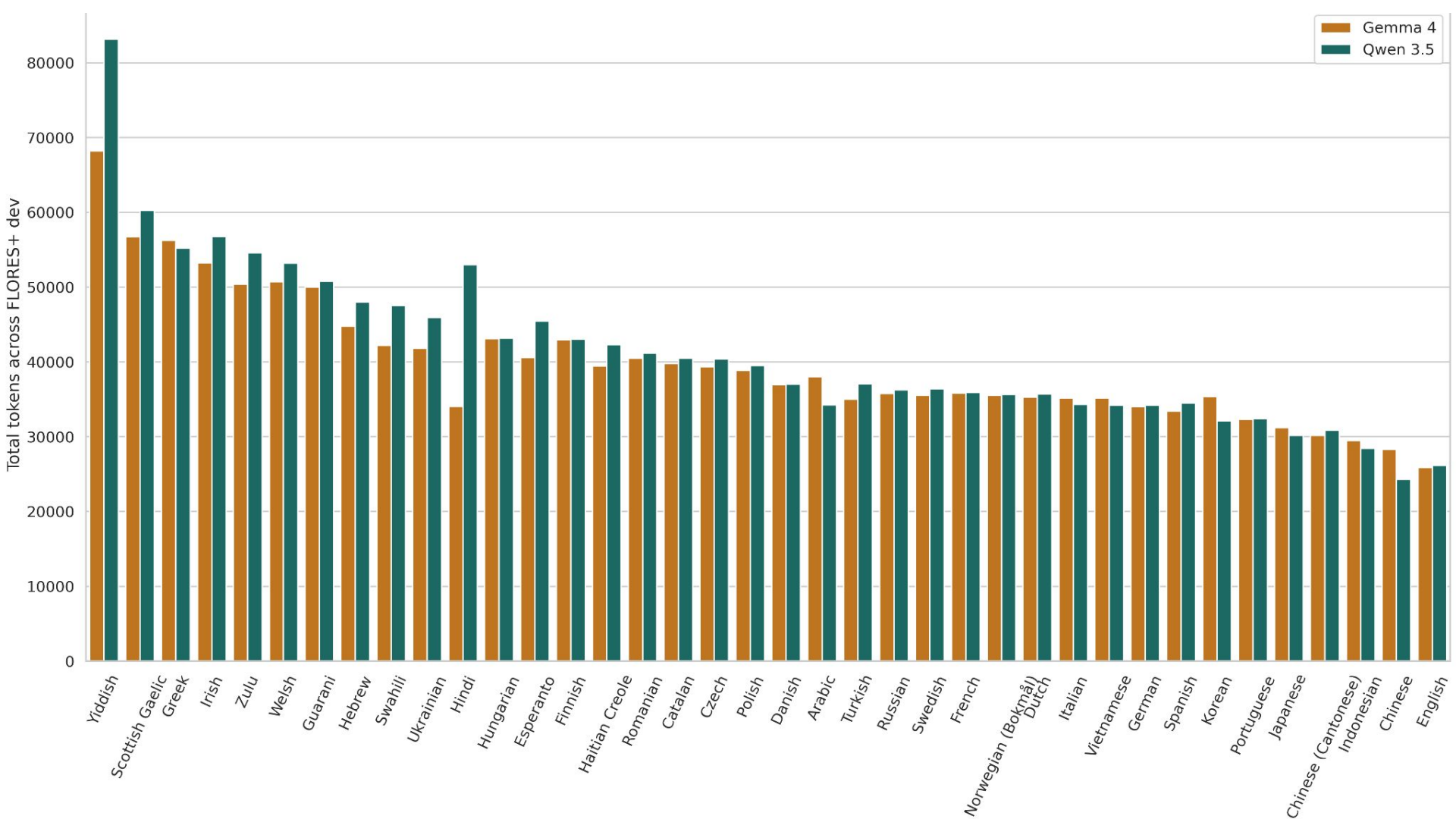
Ten łańcuch górski, będący jednym z największych w Europie, ciągnie się przez terytorium ośmiu krajów.
(**43** tok by Mistral tokenizer)

"This induces unfair treatment for some language communities in regard to the **cost** of accessing commercial language services, the **processing time** and **latency**,..."

"We show evidence that speakers of a large number of the supported languages are **overcharged** while obtaining **poorer results**."

Quoted from Language Model Tokenizers Introduce Unfairness Between Languages (Petrov et al., NeurIPS 2023)

Quoted from Do All Languages Cost the Same? Tokenization in the Era of Commercial Language Models (Ahia et al., EMNLP 2023)


Gemma 4
Qwen 3.5
Total tokens across FLORES+ dev
0
10000
20000
30000
40000
50000
60000
70000
80000
Yiddish
Scottish Gaelic
Greek
Irish
Zulu
Welsh
Guarani
Hebrew
Swahili
Ukrainian
Hindi
Hungarian
Esperanto
Finnish
Haitian Creole
Romanian
Catalan
Czech
Polish
Danish
Arabic
Turkish
Russian
Swedish
French
Norwegian (Bokmål)
Dutch
Italian
Vietnamese
German
Spanish
Korean
Portuguese
Japanese
Chinese (Cantonese)
Indonesian
Chinese
English

Adapters for Altering LLM Vocabularies: What Languages Benefit the Most? (Han et al., ICLR 2025)

Vocabulary Adaptation

Modifying pre-trained LMs to leverage a new vocabulary

01. Mitigate Fragmentation

Reduces excessive splitting of words by the tokenizer

02. Multilingual Expansion

Seamlessly introduces new languages into an existing model

03. Boost Performance

Improves downstream results in the specific target language

Limitations of Existing Vocabulary Adaptation Approaches

Heuristic Initialization

Lacks adaptability; requires **LAPT** (language adaptive pretraining) for new embeddings.

External Dependencies

Reliance on dictionaries increases complexity and limits scalability.

Approach	Resources
RAMEN (2020)	FastAlign, fastText
OFA (2024)	ColexNet+
FOCUS (2023)	fastText
WECHSEL (2022)	fastText, Biling Dicts
CW2V (2024)	Bilingual dictionaries

Limitations of Existing Vocabulary Adaptation Approaches

Heuristic Initialization

Lacks adaptability; requires **LAPT** (language adaptive pretraining) for new embeddings.

External Dependencies

Reliance on dictionaries increases complexity and limits scalability.

Language-Specific Restrictions

Existing methods are often restricted to a small subset of languages, failing to scale across diverse linguistic settings.

Vocabulary Adaptation	Grouping
ZeTT (Minixhofer et al., 2024)	lang-specific
RAMEN (Tran, 2020)	lang-specific
FOCUS (Dobler & de Melo, 2023)	lang-specific
MAD-X (Pfeiffer et al., 2020)	lang-specific
WECHSEL (Minixhofer et al., 2022)	lang-specific
CLP (Ostendorff & Rehm, 2023)	lang-specific
CLP+ (Yamaguchi et al., 2024)	lang-specific

Adapters for Altering LLM Vocabularies: What Languages Benefit the Most? (Han et al., ICLR 2025)

Understudied Impact of Vocabulary Adaptation

Limited Language Scope

Most prior work investigates only a few languages. Research involving many languages often lacks the necessary detailed analysis.

Gap in Generative Tasks

Impact on cross-lingual and generative tasks like MT is understudied. Evaluations typically focus on discriminative tasks (NLI, QA).

Vocabulary Adaptation	# Langs	Generative Task
ZeTT (Minixhofer et al., 2024)	26	x
RAMEN (Tran, 2020)	6	x
FVT (Gee et al., 2022)	1 (en)	x
VIPI (Mosin et al., 2023)	1 (en)	x
OFA (Liu et al., 2024)	min 369	x
CLP (Ostendorff & Rehm, 2023)	1 (de)	x
CLP+ (Yamaguchi et al., 2024)	4	summarization

ADAPTERS FOR ALTERING LLM VOCABULARIES: WHAT LANGUAGES BENEFIT THE MOST?

HyoJung Han[‡]
University of Maryland
hjhan@cs.umd.edu

Akiko I. Eriguchi
Microsoft
akikoe@microsoft.com

Haoran Xu
Microsoft
haoranxu@microsoft.com

Hieu Hoang
Microsoft
hihoan@microsoft.com

Marine Carpuat
University of Maryland
marine@cs.umd.edu

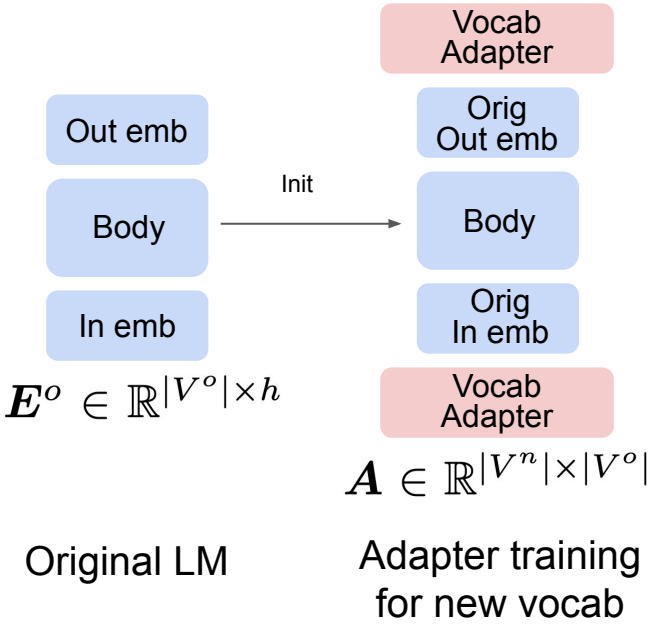
Huda Khayrallah[‡]
Amazon
hudakh@amazon.com

DEPARTMENT OF COMPUTER SCIENCE

ICLR

Microsoft

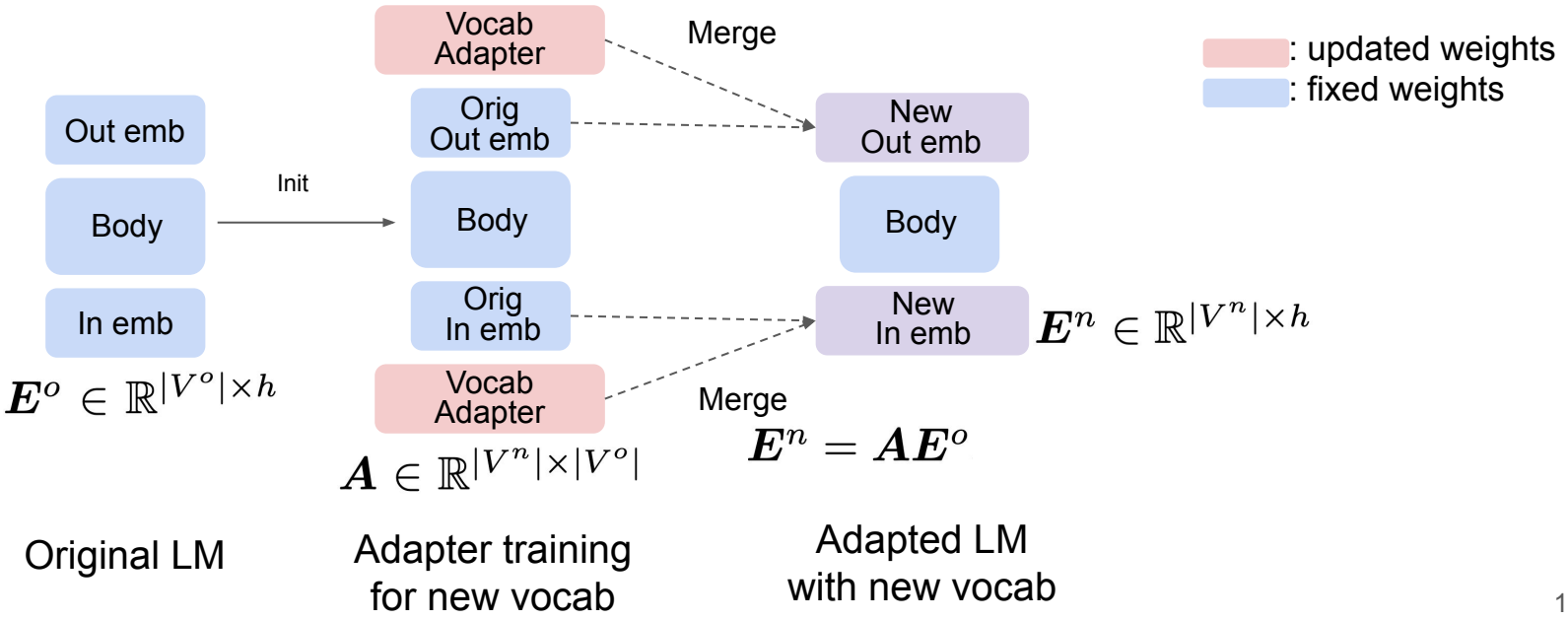
VocADT: Multilingual Vocabulary Adaptation with Adapters



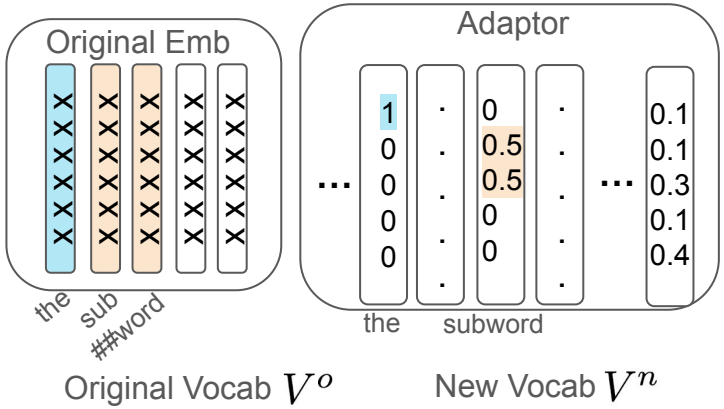
: updated weights
: fixed weights

V^o and V^n
original and **n**ew vocabulary.
h is an embedding dimension

VocADT: Multilingual Vocabulary Adaptation with Adapters



Initialization Scheme for the Vocabulary Adapter



a tokenizer associated with a vocabulary V^x

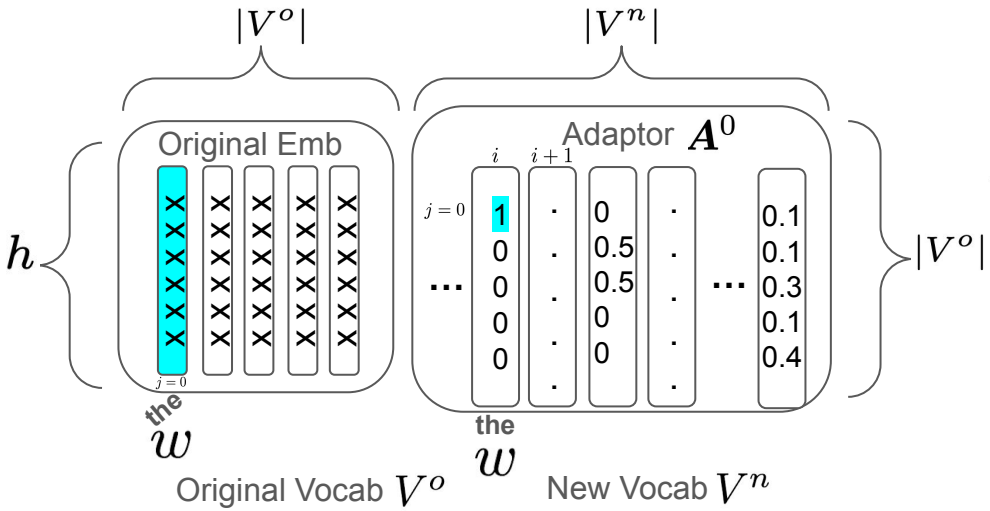
$$\mathcal{T}^x : w \rightarrow (t_1, t_2, \ldots, t_k)$$

$i = \mathcal{I}^n(w)$ be the index of a token w in V^n

$\mathcal{I}^x : V^x \rightarrow \mathbb{Z}$ be the mapping function of a token to an index in a vocabulary V^x

Initialization Scheme for the Vocabulary Adapter (1)

1. Copying the original embeddings of overlapping tokens



$$A^0_{i,\mathcal{I}^o(w)} = 1, \quad A_{i,j} = 0 \quad \forall j \neq \mathcal{I}^o(w)$$

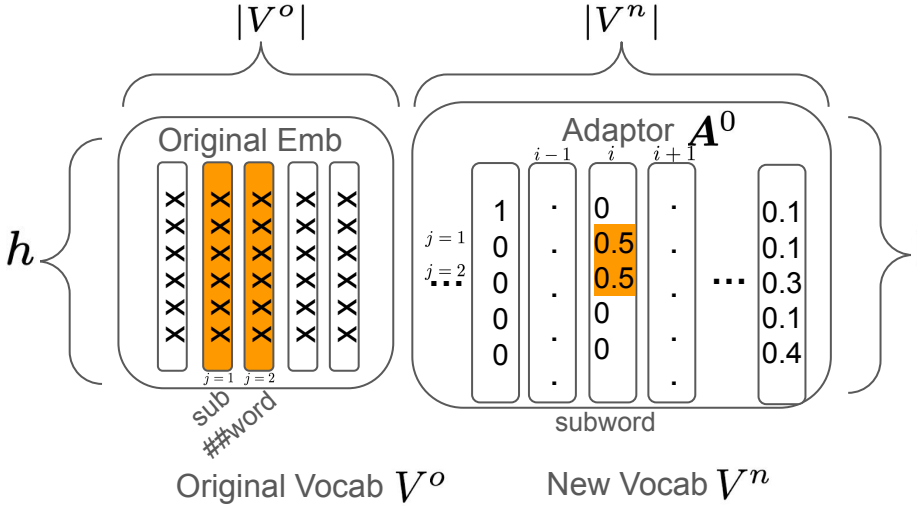
where $w \in V^o \cap V^n$

$i = \mathcal{I}^n(w)$ be the index of a token w in V^n

$\mathcal{I}^x : V^x \rightarrow \mathbb{Z}$ be the mapping function of a token to an index in a vocabulary V^x

Initialization Scheme for the Vocabulary Adapter (2)

2. Initializing the row of a token whose partitioned tokens by the original tokenizer are subset of the original vocabulary



subword sub ##word

$$\mathcal{T}^o(w) = \{t_1, \ldots, t_m\} \subset V^o, m > 1$$

$$A^0_{i,j} = \begin{cases} \frac{1}{m} & \text{if } j \in \{\overset{j=1}{\mathcal{I}^o(t_1)}, \ldots, \overset{j=2}{\mathcal{I}^o(t_m)}\} \\ 0 & \text{otherwise} \end{cases}$$

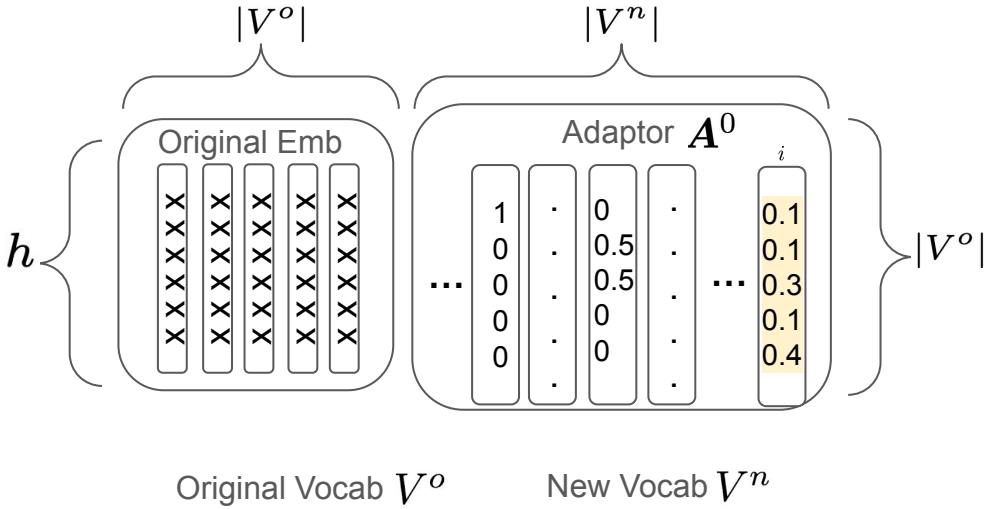
where $w \in V^n \backslash (V^o \cap V^n)$ and $w \in S = \{w \mid \mathcal{T}^o(w) = \{t_{1:m}\} \subset V^o\}$

$i = \mathcal{I}^n(w)$ be the index of a token w in V^n

$\mathcal{I}^x : V^x \rightarrow \mathbb{Z}$ be the mapping function of a token to an index in a vocabulary V^x

Initialization Scheme for the Vocabulary Adapter (3)

3. Otherwise, we randomly initialize a row vector of the adapter with the uniform distribution whose sum of each element is one.



$$A_i^0 = \frac{\mathbf{u}}{\sum_{j=1}^{|V^o|} u_j}$$

$$u_j \sim \text{Uniform}(0,1),\ j = 1, \ldots, |V^o|$$

$$\text{where } w \in V^n \backslash (V^o \cap V^n) \backslash S$$

$$S = \{w \mid \mathcal{T}^o(w) = \{t_{1:m}\} \subset V^o\}$$

Adapters for Altering LLM Vocabularies: What Languages Benefit the Most? (Han et al., ICLR 2025)

When and how should we perform vocabulary adaptation?

Language Benefit

Which languages benefit the most from vocabulary adaptation techniques?

Creation Strategies

What are the best strategies for creating new vocabularies? Is script consistency a necessary factor?

Translation Impact

How does vocabulary adaptation impact performance in machine translation tasks?

Experiment Design

idx	Full Name	Short	Script	Resource	FLORES	XNLI	XCOPA	Belebele
1	English	en	Latin	High	✓	✓		✓
2	Swahili	sw	Latin	Low	✓	✓	✓	✓
3	Indonesian	id	Latin	Mid	✓		✓	✓
4	Estonian	et	Latin	Mid	✓		✓	✓
5	Haitian Creole	ht	Latin	Low	✓		✓	✓
6	Korean	ko	Hangul	High	✓			✓
7	Greek	el	Greek	Mid	✓	✓		✓
8	Russian	ru	Cyrillic	High	✓	✓		✓
9	Bulgarian	bg	Cyrillic	Mid	✓	✓		✓
10	Ukrainian	uk	Cyrillic	Mid	✓			✓
11	Kazakh	kk	Cyrillic	Mid	✓			✓

Latin **group**

Mixed **group**

Cyrillic **group**

Experimental Setting

Basemodel: Mistral-7B (32k tokens)

Baselines: ZeTT (Minixhofer et al., 2024), FOCUS (Dobler & de Melo, 2023), and

OFA (Liu et al., 2024).

Vocabulary: SentencePiece. 50k tokens. For each language groups plus English

Adapter Training: MADLAD-400, train 0.5B monolingual tokens per language, totaling 2.5B mixed by 5 languages (English + 4 non-English from each corresponding group)

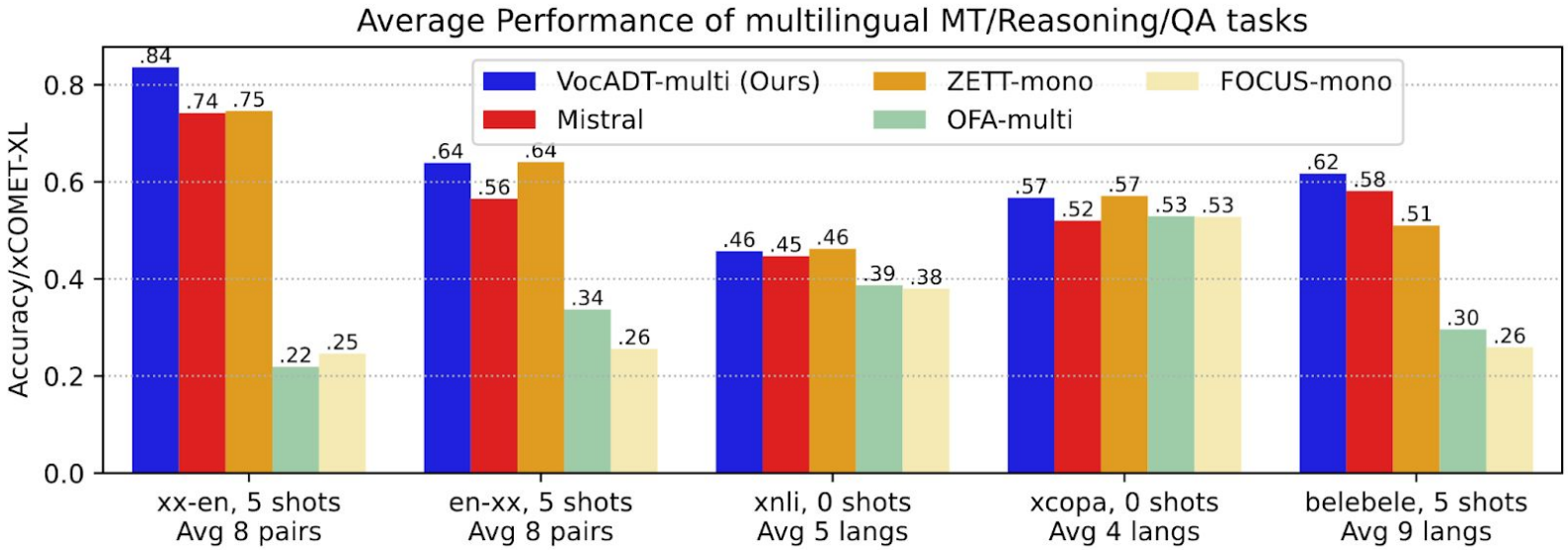
Evaluation: FLORES (xCOMET-XL, 5-shot) for xx-en & en-xx MT, Belebele (Accuracy, 5-shot), XNLI and XCOPA (Accuracy, 0-shot)

Zero-Shot Tokenizer Transfer (Minixhofer et al., arXiv 2024)
FOCUS: Effective Embedding Initialization for Monolingual Specialization of Multilingual Models (Dobler & de Melo, EMNLP 2023)
OFA: A Framework of Initializing Unseen Subword Embeddings for Efficient Large-scale Multilingual Continued Pretraining (Liu et al., Findings 2024)

Adapters for Altering LLM Vocabularies: What Languages Benefit the Most? (Han et al., ICLR 2025)

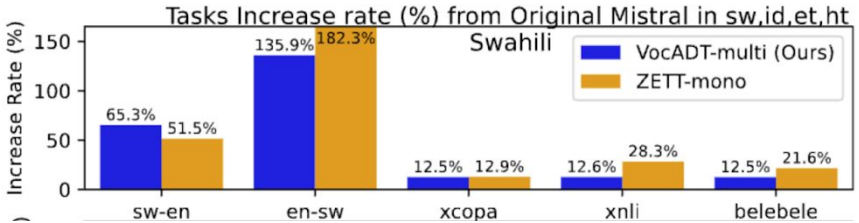
Overall Task Performance

Adapting the vocabulary using VocADT generally leads to better performance compared to the original Mistral model, and either surpasses or performs on par with competitive baselines.



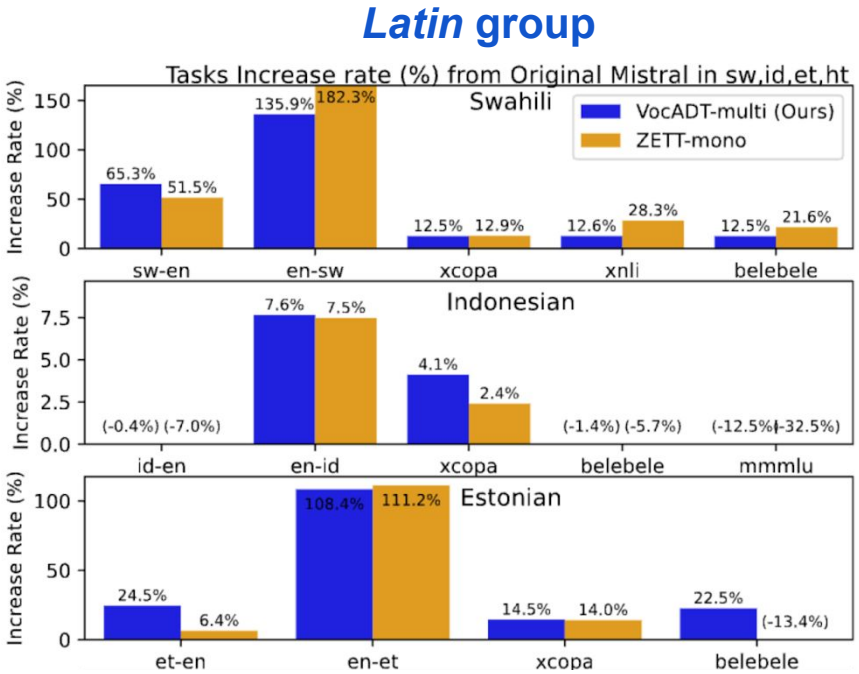
Adapters for Altering LLM Vocabularies: What Languages Benefit the Most? (Han et al., ICLR 2025)

Which Languages Benefit the Most from Vocab Adaptation?



Adapters for Altering LLM Vocabularies: What Languages Benefit the Most? (Han et al., ICLR 2025)

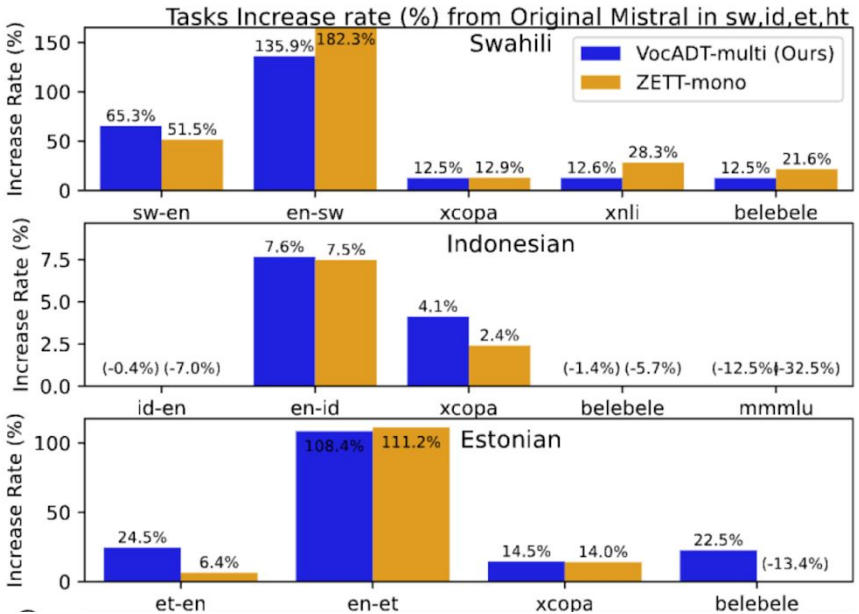
Which Languages Benefit the Most from Vocab Adaptation?



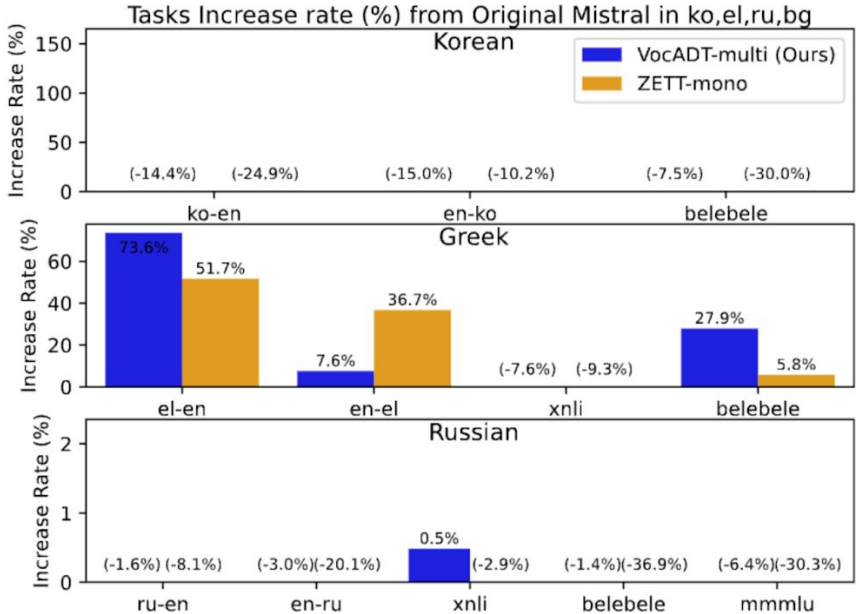
Adapters for Altering LLM Vocabularies: What Languages Benefit the Most? (Han et al., ICLR 2025)

Which Languages Benefit the Most from Vocab Adaptation?

Latin group



Mixed group

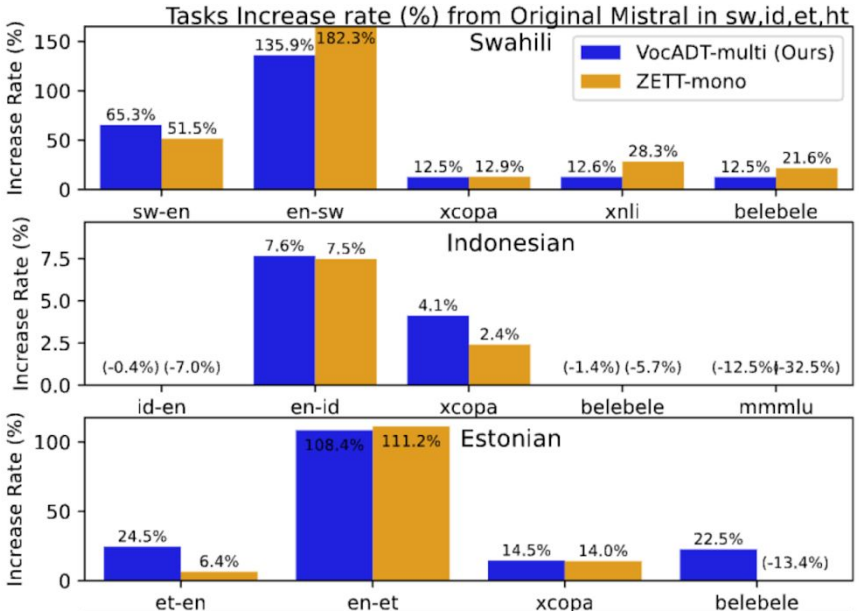


Adapters for Altering LLM Vocabularies: What Languages Benefit the Most? (Han et al., ICLR 2025)

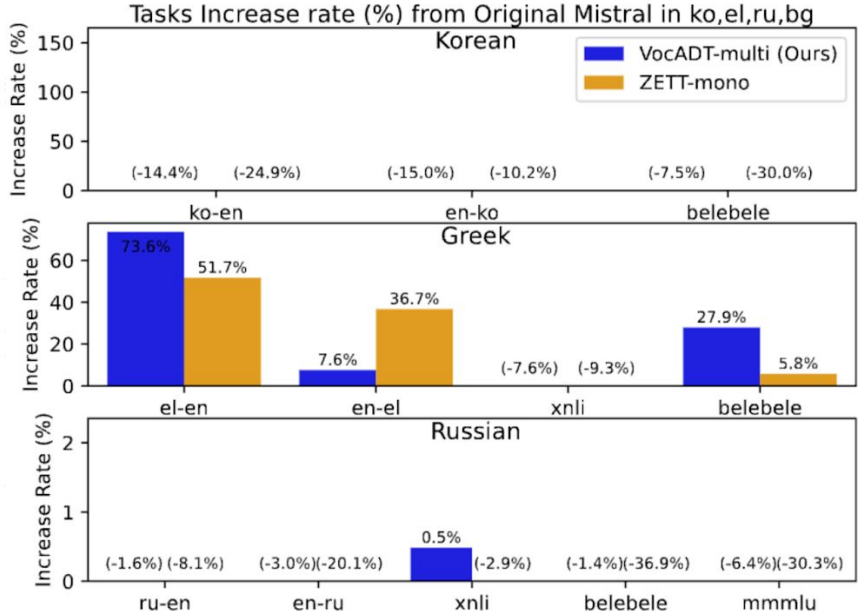
Which Languages Benefit the Most from Vocab Adaptation?

Increased rate of task performance after Vocabulary Adaptation: Languages with **Latin Scripts** or Severe Fragmentation Benefit the Most

***Latin* group**



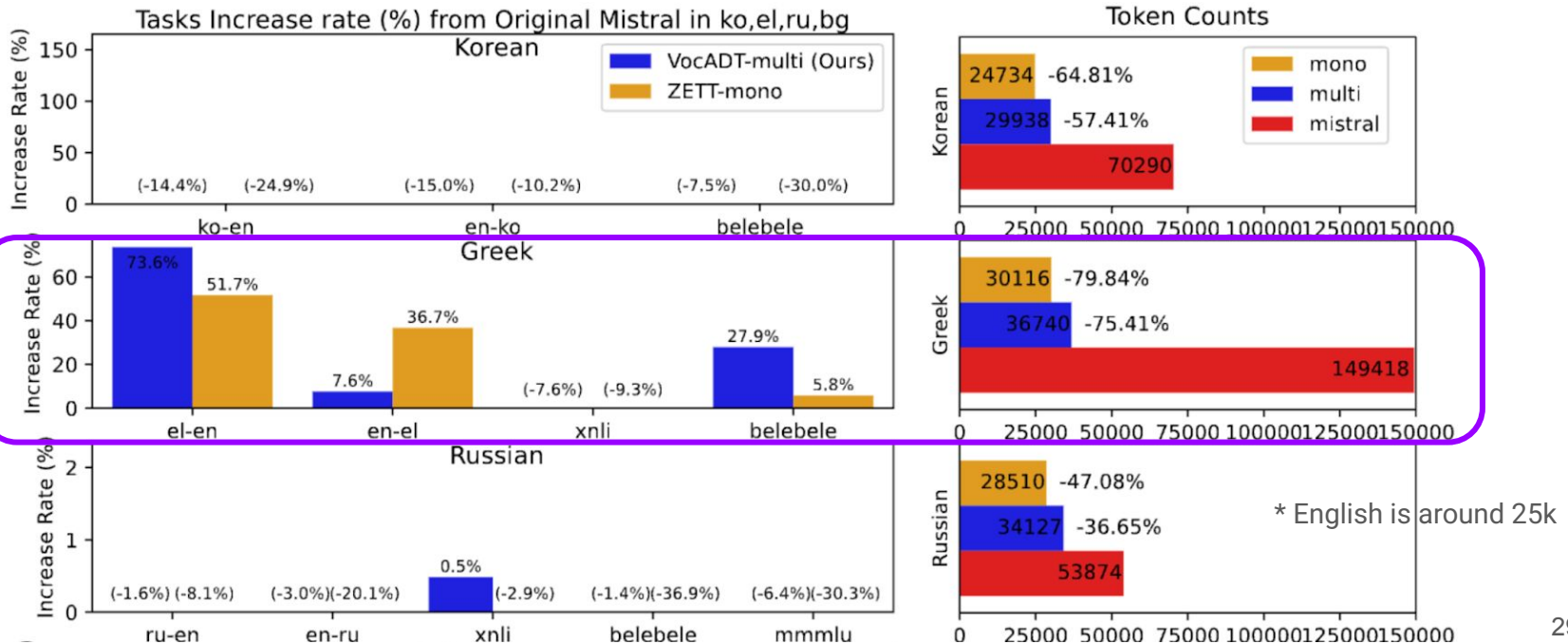
***Mixed* group**



Adapters for Altering LLM Vocabularies: What Languages Benefit the Most? (Han et al., ICLR 2025)

Which Languages Benefit the Most from Vocab Adaptation?

Languages with Latin Scripts or **Severe Fragmentation** Benefit the Most



* English is around 25k

Adapters for Altering LLM Vocabularies: What Languages Benefit the Most? (Han et al., ICLR 2025)

ADAPTERS FOR ALTERING LLM VOCABULARIES: WHAT LANGUAGES BENEFIT THE MOST?

HyoJung Han Akiko I. Eriguchi Haoran Xu

Hieu Hoang Marine Carpuat Huda Khayrallah

ICLR

github.com/h-j-han/VocADT

hjhan@cs.umd.edu

- **VocADT**, a simple and effective solution for vocabulary adaptation using **adapters** that addresses key limitations in prior work

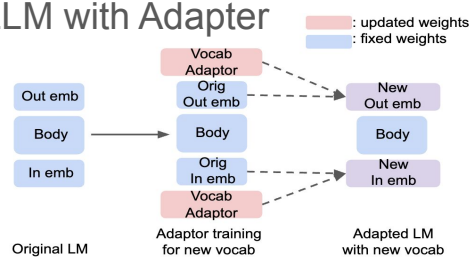
- Finding that languages with Latin scripts or severe fragmentation **benefit the most** and that having a consistent grouping of scripts for multilingual vocabulary is helpful

- VocADT consistently outperforms the original language model and is more effective than, or on par with, strong vocabulary adaptation baselines

Talk Overview

Uniformity

Vocabulary Transfer of Multilingual LLM with Adapter



Adapters for Altering LLM Vocabularies: What Languages Benefit the Most? (Han et al., ICLR 2025)

Locality

Explanatory Translation for Bridging **Cultural Background** Knowledge **Gap**

Source …frère de Dominique de Villepin…

Literal Translation …brother of Dominique de Villepin …

with Explicitation …brother of the former French Prime Minister Dominique de Villepin …

Bridging Background Knowledge Gaps in Translation with Automatic Explicitation (Han et al., EMNLP 2023)

The vocabulary adaptation addresses uniformity at the very-surface level by mitigating the over-fragmentation of mid/low resources languages.

…ure

Universal Response

→ 우리 몸은 몇 퍼센트가 물인가요?
→ Πόσο % νερό είναι το σώμα;
→ What % of the body is water?

Culturally-adaptive Response

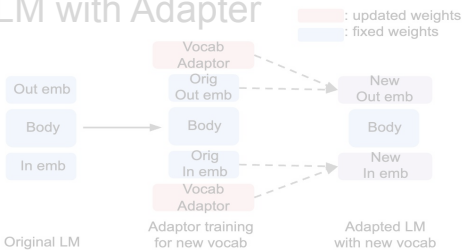
→ 긴급 신고 번호는 몇 번이에요?
→ *Αριθμός έκτακτης ανάγκης;*
→ What is the emergency number?

Rethinking Cross-lingual Alignment: Balancing Transfer and Cultural Erasure in Multilingual LLMs (Han et al., arXiv 2025)

Talk Overview

Uniformity

Vocabulary Transfer of Multilingual LLM with Adapter



Adapters for Altering LLM Vocabularies: What Languages Benefit the Most? (Han et al., ICLR 2025)

Locality

Explanatory Translation for Bridging **Cultural Background** Knowledge **Gap**

🇫🇷🙂 Source: …frère de Dominique de Villepin…

🇺🇸🤔 Literal Translation: …brother of Dominique de Villepin …

🇺🇸😃 with Explicitation: …brother of the former French Prime Minister Dominique de Villepin …

Bridging Background Knowledge Gaps in Translation with Automatic Explicitation (Han et al., EMNLP 2023)

Trade-offs of **Cross-lin**

between **Universal Tra**

Universal Response

→ 우리 몸은 몇 퍼센트가 물인가요?
→ Πόσο % νερό είναι το σώμα;
→ What % of the body is water?

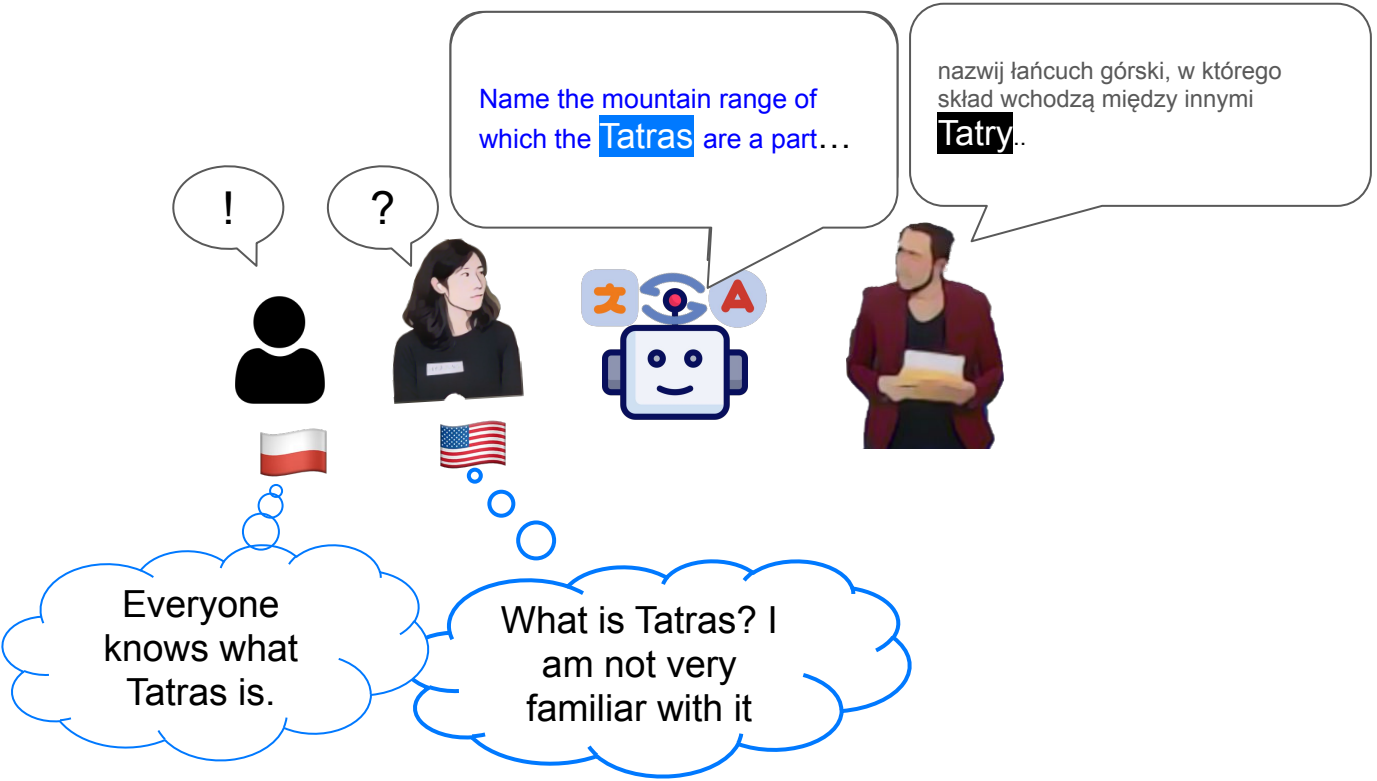
Culturally-adaptive Response

→ 긴급 신고 번호는 몇 번이에요?
→ *Αριθμός έκτακτης ανάγκης;*
→ What is the emergency number?

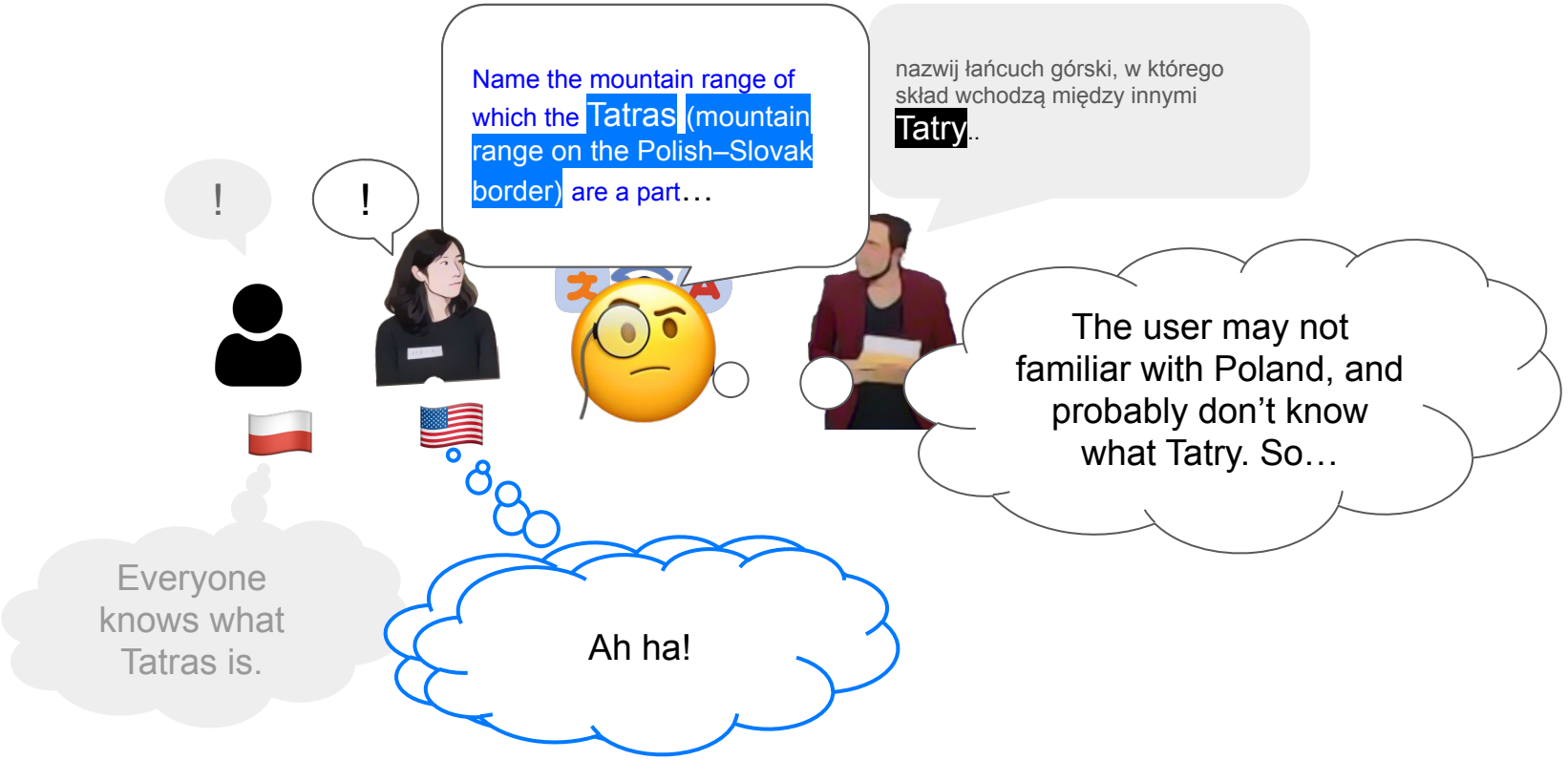
Rethinking Cross-lingual Alignment: Balancing Transfer and Cultural Erasure in Multilingual LLMs (Han et al., arXiv 2025)

How can we frame and evaluate language generation tasks that adapt to users culture and background knowledge?

Literal Multilingual Task



Explicitation: explicit realization of implicit BG knowledge



What is **Explicitation**?

Kinga Klaudy

Vinay and Darbelnet

*"A procedure of **explicitly introducing explicitly explains implicit background knowledge** which remain **implicit to the source speakers**."*

Pragmatic Explicitation, (Klaudy, 1993)

How can we make models adaptive to cultural background knowledge?

How can we test these capabilities rigorously and automatically at scale?

Images from : https://into23.com/what-are-the-main-translation-methods-and-techniques/
Images from : https://scholar.google.com/citations?user=mXSFEcoAAAAJ

DEPARTMENT OF COMPUTER SCIENCE

EMNLP 2023

Bridging Cultural Background Knowledge Gaps in Translation with Automatic Explicitation

How can we make models adaptive to cultural background knowledge?

How can we test these capabilities rigorously and automatically at scale?

HyoJung Han Jordan Boyd-Graber Marine Carpuat

How to generate explicitations automatically?

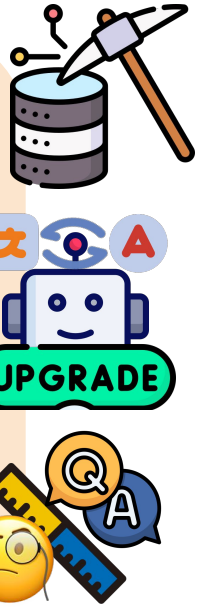
Challenges

- No labeled data

- No evaluation metrics

Our Approach

- Collect samples from Wikipedia semi-automatically (WIKIEXPL)
- Design automatic methods to mimic collected data
- Evaluate intrinsically and extrinsically: How useful are explicitations in multilingual QA?

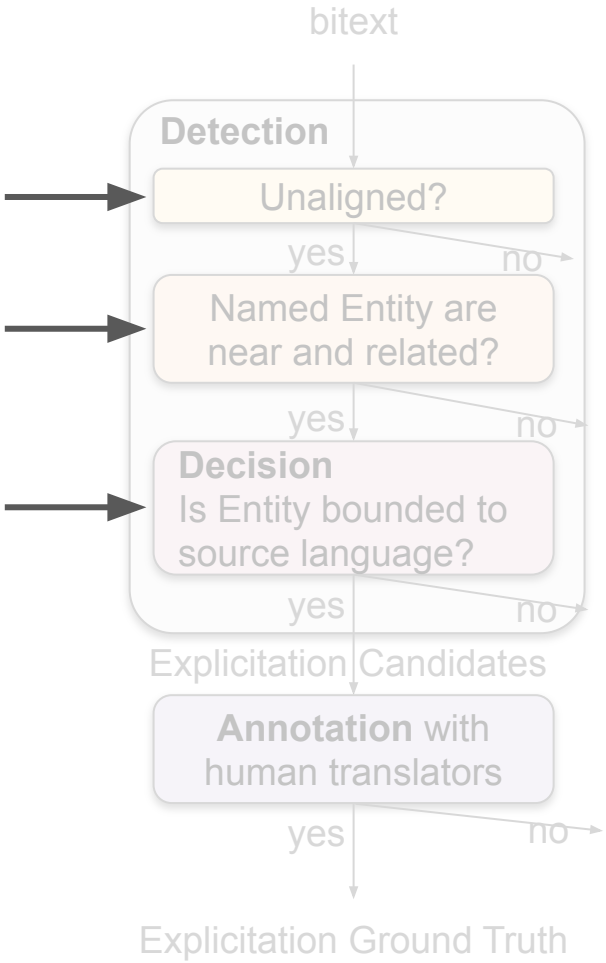


Building WIKIEXPL Dataset

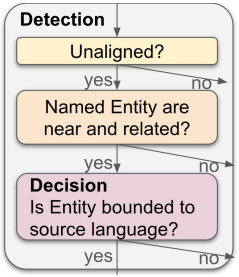
1. **Extracting** Explicitation Candidates

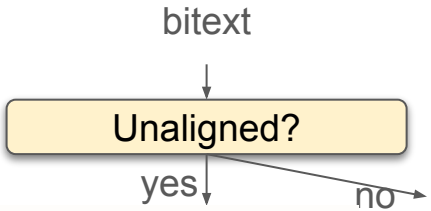
   - **Detecting** the Explicitation in Bitext
   - **Deciding** If Explicitation is Needed

2. **Annotating** final explicitation with human annotators



Detecting the Explicitation in Bitext

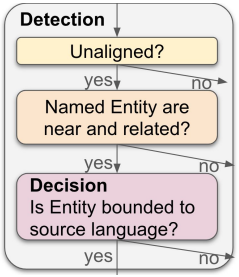


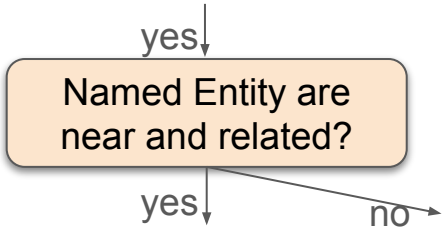


Assumption 1. Explicitations are part of **unaligned** token sequences:

		Explicitation?	Why?
Source	**…avec des vols sans escale vers Antigua,**		
Gloss	…with non-stop flights to Antigua,	X	No unaligned segments
Target	**…with direct service to Antigua,**		
Source	**…l'attentat contre Charlie Hebdo…**		
Gloss	…the attack against Charlie Hebdo…		
Target (unaligned)	**…the attack against the French satirical newspaper Charlie Hebdo…**	?	

Detecting the Explicitation in Bitext

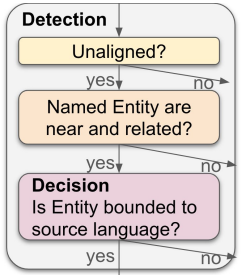


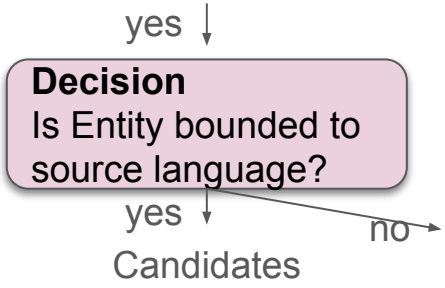


Assumption 2. Explicitations are **adjacent** and **relevant** to named entities (NE):

		Explicitation?	Why?
Source	…comme les persécutions à Metz en 888.		
Gloss	…like the persecution in Metz in 888.	X	Unaligned segment is **not related** to NE
Target (unaligned)	…such as the one in Metz in 888.		
Source	…l'attentat contre Charlie Hebdo…		
Gloss	…the attack against Charlie Hebdo…		
Target (unaligned)	…the attack against the French satirical newspaper Charlie Hebdo…	?	
Source	À l' adolescence , il écoute de la musique latino-américaine,…		
Gloss	In adolescence, he listened to Latin American music,…		
Target (unaligned)	In his teenage years in Bogotá , he acquired a taste for Latin American music,…	X	Unaligned segment **far from** NE

Deciding If Explicitation is Needed





Assumption 3. Explicitations are more likely for ***culturally distant*** entities from target language:

Explicitation? Why?

Source	**…l'attentat contre Charlie Hebdo…**
Gloss	…the attack against Charlie Hebdo…
Target (unaligned)	**…the attack against the French satirical newspaper Charlie Hebdo…**

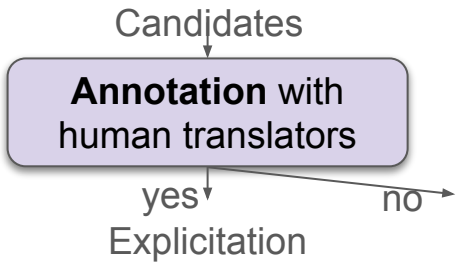
?

Source	**…et au sud du Texas.**
Gloss	…and in southern Texas,
Target (unaligned)	**…and into southern Texas in the United States.**

X

Texas is **not culturally bounded** to French community, and also not culturally distant from English speaking country

Annotating Explicitation with translators



Present the candidates to **human** annotators and label ground truth examples.

Explicitation? Why?

Source	**…l'attentat contre Charlie Hebdo…**
Gloss	…the attack against Charlie Hebdo…
Target (unaligned)	**…the attack against the French satirical newspaper Charlie Hebdo…**

O

This specifies a socio-cultural function to help the non-French reader identify Charlie Hebdo, a newspaper that is well-known in France.

Source	**…nommé d' après Pierre André de Suffren,**
Gloss	…appointed after Pierre André de Suffren,
Target (unaligned)	**…named after the 18th century admiral Pierre André de Suffren,**

X

The century cannot be expected to be known by either the French or target readers

Resulting WIKIEXPL Dataset

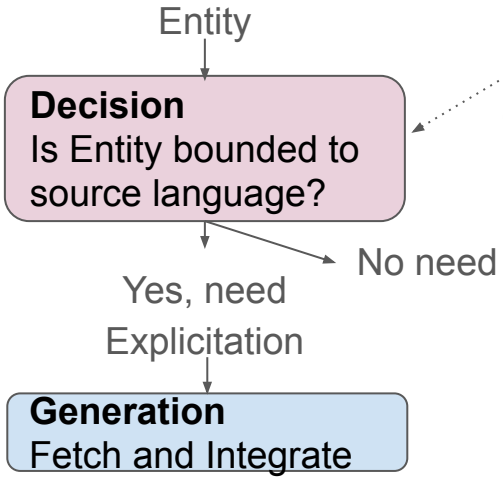
Source Language	French	Polish	Spanish
WikiMatrix	29826	21392	28900
Candidates	791	245	307
Top 1 country			
Annotated	460	244	220
Explicitation	116	67	44

Resulting WIKIEXPL Dataset

Type	Source	Target
Hypernym (h)	la Sambre	the Sambre ***river***
Occupation/Title (o)	Javier Gurruchaga	***showman*** Javier Gurruchaga
Acronym Expansion (a)	PP	***People 's Party*** (PP)
Full names (f)	Cervantes	***Miguel*** de Cervantes
Nationality (n)	Felipe II	Philip II ***of Spain***
Integrated (i)	Dominique de Villepin	***former French Prime Minister*** Dominique...

Automating Explicitation

1. Deciding if Explicitation is Needed



2. Generating the Explanation

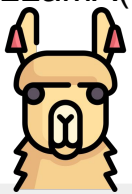
Fetch Information from Wikipedia and Wikidata, and merge it to target text.

Type	Length	Form	Example of "Sambre,"
SHORT	1–2 words	Appositive	Sambre river,
MID	3 words–a phrase	+ Parenthetical	Sambre, river in France and Belgium,
LONG	1–3 sentences	Footnote	Sambre (a river in northern France and in Wallonia, Belgium. It is a left-bank tributary of the Meuse, which it joins in the Wallonian capital Namur.)

Evaluating Explicitation

Same Human Annotators in building WIKIEXPL

LLaMA(1)



Multilingual Question Answering System

Polish Question	Ten łańcuch górski, będący jednym z największych w Europie, ciągnie się przez terytorium ośmiu krajów. Ten obszar stanowi dział wodny między zlewiskiem Morza Bałtyckiego i Morza Czarnego oraz wypływa z niego wiele rzek, w tym Wisła. Aby otrzymać punkt, nazwij łańcuch górski, w którego skład wchodzą między innymi Tatry, a najwyższym szczytem jest Gerlach.
English Question	This mountain range, which is one of the largest in Europe, continent, stretches across the territory of eight countries. This area is the watershed between the catchment areas of the Baltic Sea and the Black Sea and there are many rivers flowing out of it, including the Vistula. To get a point, name the mountain range of which the Tatras are a part, and the highest peak is Gerlach.
Answer	Karpaty, Carpathian Mountain
Short Explicitation	... the Tatras, Slovakia, are ...
Mid Explicitation	... the Tatras (mountain range on the Polish–Slovak border) are ...
Long Explicitation	... the Tatras* are ...
Footnotes (For Long)	*Tatras : The Tatra Mountains, Tatras, or Tatra are a series of mountains within the Western Carpathians that form a natural border between Slovakia and Poland. They are the highest mountains in the Carpathians.

Figure 6: Generation example from our experiment in extrinsic evaluation in XQB-*pl*.

Evaluating Explicitation - Direct Human Evaluation

- Rated with a three-step Likert scale mapping high as 1, mid as 0.5, and low as

Is our automatic explicitation is as useful as natural ones?

How do you assess additive information?

How well do you think this extra information is integrated into the English translation?

Decision	Type	Generation	Integration
0.71	SHORT	0.63	0.79
	MID	0.82	0.92
	LONG	0.95	_

About 70% of automated explicitation are marked as valid decision

The quality of generation decreases with shorter explanations.

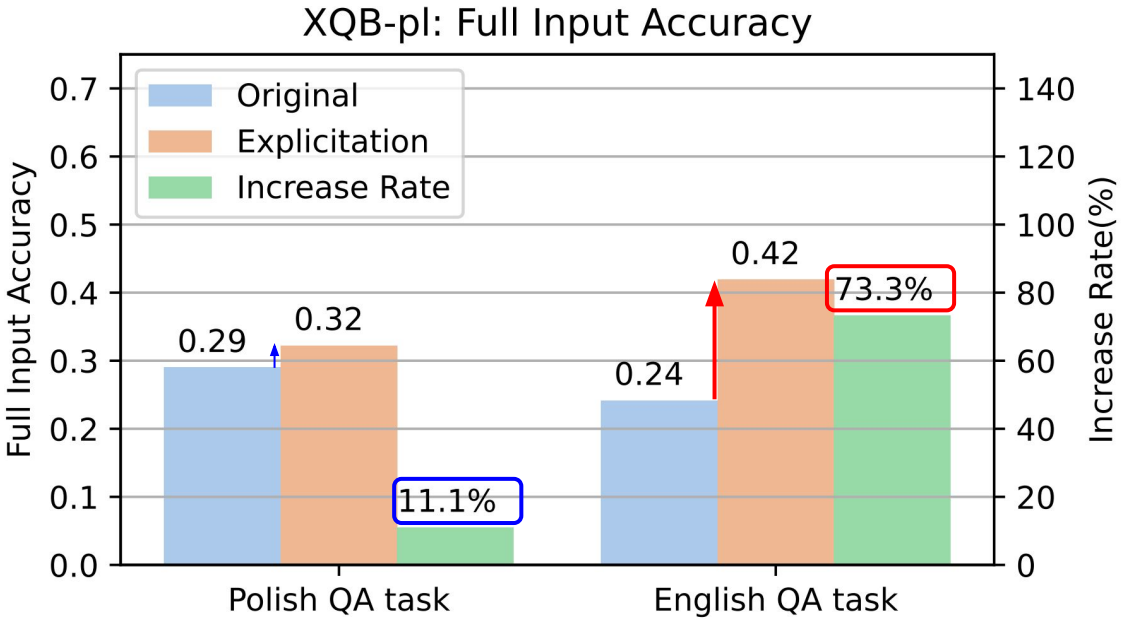
Evaluating Explicitation - Extrinsic Evaluation with QA

	Useful in target QA	Not Useful in target QA
Useful in source QA	Subject we don't know commonly	Subject is not popular in source community
Not Useful in source QA	**Subject is well-known only in source community**	Subject is well-known globally

Question Type	Input text	System Answer
Polish Question	**...Aby otrzymać punkt, nazwij łańcuch górski, w którego skład wchodzą między innymi Tatry,...**	Karpaty (O)

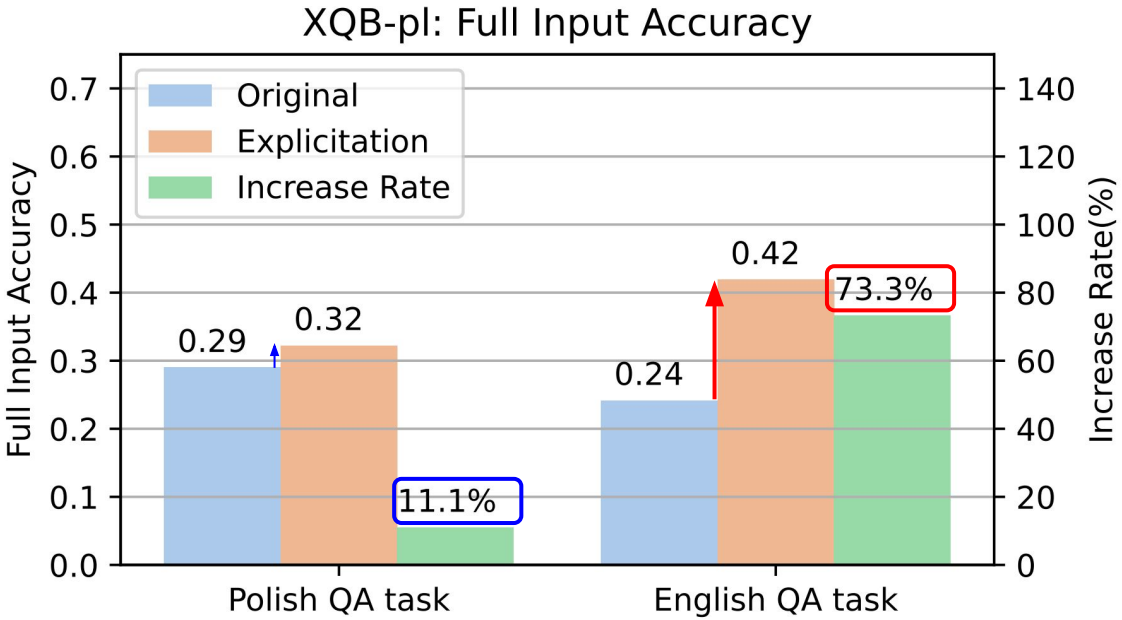
Evaluating Explicitation - Extrinsic Evaluation

A higher increase rate in the English QA task than that in Polish QA task indicates the effectiveness of our automatic explicitation methods.



Evaluating Explicitation - Extrinsic Evaluation

	Useful in target QA	Not Useful in target QA
Useful in source QA	Subject we don't know commonly	Subject is not popular in source community
Not Useful in source QA	Subject is well-known only in source community → **Need Explicitation!**	Subject is well-known globally



Bridging Background Knowledge Gaps in Translation with Automatic Explicitation

EMNLP 2023

HyoJung Han
Computer Science
University of Maryland
hjhan@cs.umd.edu

Jordan Boyd-Graber
CS, UMIACS, iSchool, LCS
University of Maryland
jbg@umiacs.umd.edu

Marine Carpuat
Computer Science, UMIACS
University of Maryland
marine@cs.umd.edu

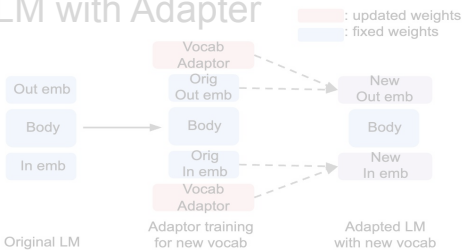
To bridge the gap of background knowledge between source and target audiences in MT, we contribute:

- **WIKIEXPL** : semi-automatically collected from Wikipedia
- **Automatic Explicitation** : techniques to **automatically decide** what to explain and **how**
- **Extensive Evaluation** : explicitations are valid based on human evaluation and useful based on extrinsic multilingual QA.

Talk Overview

Uniformity

Vocabulary Transfer of Multilingual LLM with Adapter



Adapters for Altering LLM Vocabularies: What Languages Benefit the Most? (Han et al., ICLR 2025)

Locality

Explanatory Translation for Bridging **Cultural Background** Knowledge **Gap**

🇫🇷🙂 Source: …frère de Dominique de Villepin…

🇺🇸🤔 Literal Translation: …brother of Dominique de Villepin …

🇺🇸😃 with Explicitation: …brother of the former French Prime Minister Dominique de Villepin …

Bridging Background Knowledge Gaps in Translation with Automatic Explicitation (Han et al., EMNLP 2023)

Trade-offs of **Cross-**
between **Universal Tr**

Universal Response

→ 우리 몸은 몇 퍼센트가 물인가요?
→ Πόσο % νερό είναι το σώμα;
→ What % of the body is water?

→ 긴급 신고 번호는 몇 번이에요?
→ *Αριθμός έκτακτης ανάγκης;*
→ What is the emergency number?

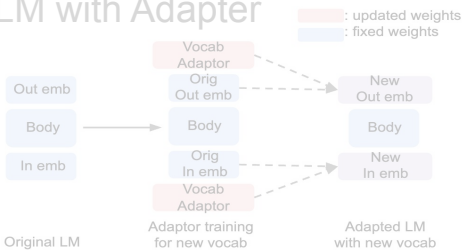
Rethinking Cross-lingual Alignment: Balancing Transfer and Cultural Erasure in Multilingual LLMs (Han et al., arXiv 2025)

Multilingual models may need to behave differently at the output level to support targeted users who speaks different languages and have different groundings.

Talk Overview

Uniformity

Vocabulary Transfer of Multilingual LLM with Adapter



Adapters for Altering LLM Vocabularies: What Languages Benefit the Most? (Han et al., ICLR 2025)

Locality

Explanatory Translation for Bridging **Cultural Background** Knowledge **Gap**

Source …frère de Dominique de Villepin…

Literal Translation …brother of Dominique de Villepin …

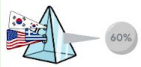
with Explicitation …brother of the former French Prime Minister Dominique de Villepin …

Bridging Background Knowledge Gaps in Translation with Automatic Explicitation (Han et al., EMNLP 2023)

Trade-offs of **Cross-lingual Alignment** between **Universal Transfer** vs **Cultural Erasure**

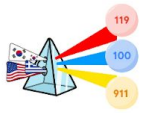
Universal Response

- 우리 몸은 몇 퍼센트가 물인가요?
- Πόσο % νερό είναι το σώμα;
- What % of the body is water?



Culturally-adaptive Response

- 긴급 신고 번호는 몇 번이에요?
- *Αριθμός έκτακτης ανάγκης;*
- What is the emergency number?



Rethinking Cross-lingual Alignment: Balancing Transfer and Cultural Erasure in Multilingual LLMs (Han et al., arXiv 2025)

Why cross-lingual alignment?

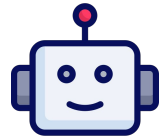
Which vitamin is required for vision in dim light?

A. Vitamin A
B. Vitamin D
C. Vitamin E
D. Vitamin K

Ποια βιταμίνη είναι απαραίτητη για την όραση σε αμυδρό φως;

A. Βιταμίνη A
B. Βιταμίνη D
C. Βιταμίνη E
D. Βιταμίνη K

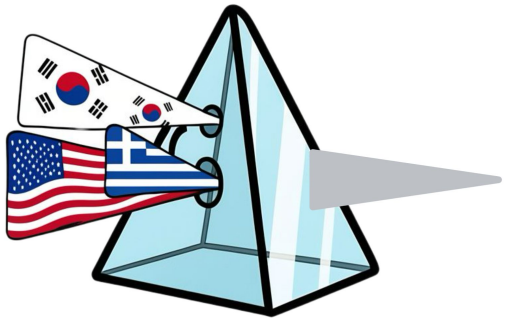
Vitamin A



Βιταμίνη E

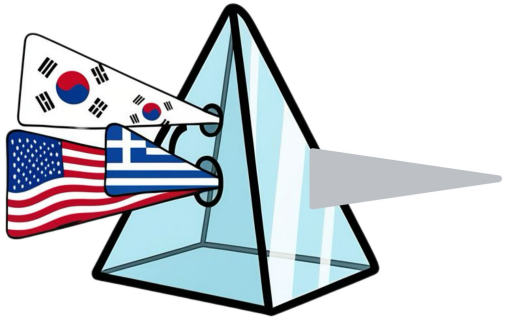
What is the goal of cross-lingual alignment (**CLA**)?

Cross-lingual alignment encourages explicitly or implicitly **convergence** in model's responses across languages.

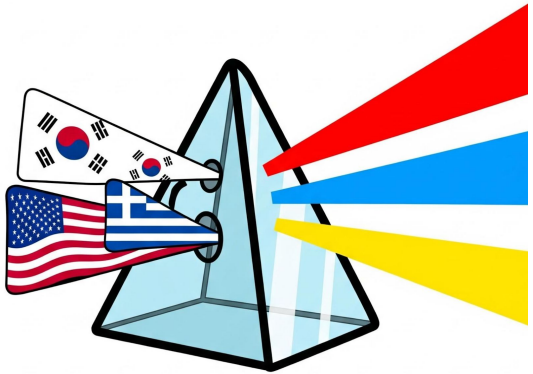


Is there a hidden cost of cross-lingual alignment (CLA)?

Cross-lingual alignment encourages explicitly or implicitly **convergence** in model's responses across languages.



Yes, but… knowledge can be culturally situated, requiring model's responses to **divergence** across languages.



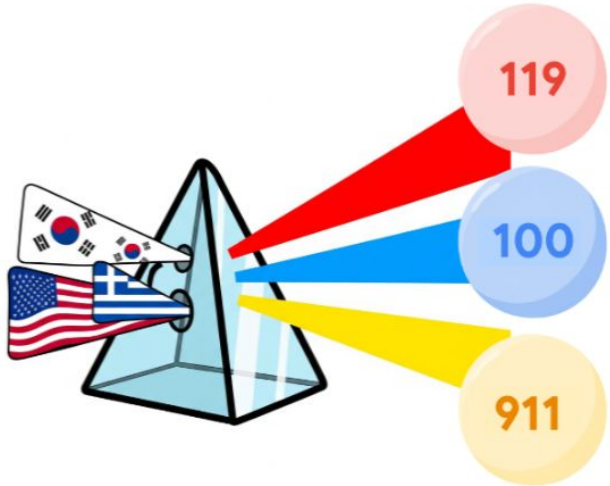
What gets lost in alignment? → **context-dependent** outputs

Each language is associated with a different cultural context → The **answer could be different by the language**

Yes, but… knowledge can be culturally situated, requiring model's responses to **divergence** across languages.

Culturally-adaptive Response

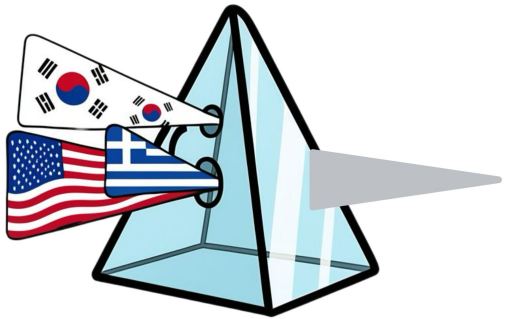
➔ 긴급 신고 번호는 몇 번이에요?
➔ *Αριθμός έκτακτης ανάγκης;*
➔ What is the emergency number?



Research Gap

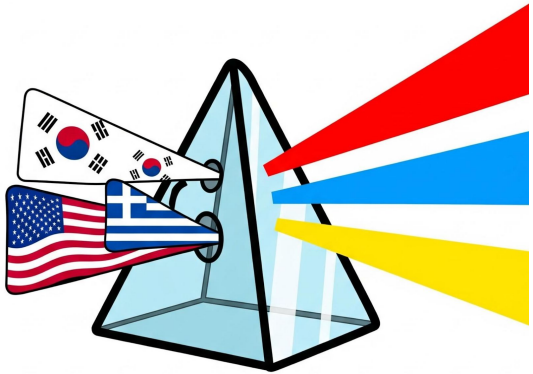
Focus of prior CLA work!

Cross-lingual alignment encourages explicitly or implicitly **convergence** in model's responses across languages.



Still Unaddressed!

Yes, but… knowledge can be culturally situated, requiring model's responses to **divergence** across languages.



RETHINKING CROSS-LINGUAL ALIGNMENT: BALANCING TRANSFER AND CULTURAL ERASURE IN MULTILINGUAL LLMS

HyoJung Han[‡]
University of Maryland
hjhan@cs.umd.edu

Sweta Agrawal
Google
swetaagrawal@google.com

Eleftheria Briakou
Google
ebriakou@google.com

DEPARTMENT OF COMPUTER SCIENCE

Google

Rethinking Cross-lingual Alignment

How can we evaluate both the gains and losses of alignment?

A holistic evaluation framework built on a two-dimensional transfer-localization plane

Rethinking Cross-lingual Alignment

How can we evaluate both the gains and losses of alignment?

A holistic evaluation framework built on a two-dimensional transfer-localization plane

What hidden cultural costs accompany cross-lingual alignment?

Re-evaluate Cross-lingual Alignment Methods on transfer-localization plane

Rethinking Cross-lingual Alignment

How can we evaluate both the gains and losses of alignment?

A holistic evaluation framework built on a two-dimensional transfer-localization plane

What hidden cultural costs accompany cross-lingual alignment?

Re-evaluate Cross-lingual Alignment Methods on transfer-localization plane

How can we design culturally-aware alignment techniques?

Identify a key distinction in how knowledge is encoded
→ Layer-specific Steering Intervention

Rethinking Cross-lingual Alignment

How can we evaluate both the gains and losses of alignment?

A holistic evaluation framework built on a two-dimensional transfer-localization plane

What hidden cultural costs accompany cross-lingual alignment?

Re-evaluate Cross-lingual Alignment Methods on transfer-localization plane

How can we design culturally-aware alignment techniques?

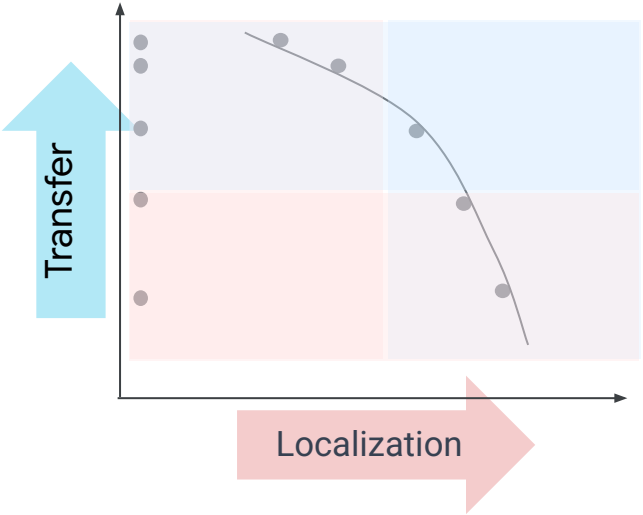
Identify a key distinction in how knowledge is encoded
→ Layer-specific Steering Intervention

Measuring the Transfer-Localization Trade-off

Δ Acc = Acc after CLA - Acc before CLA

Universal Response

➔ 우리 몸은 몇 퍼센트가 물인가요?
➔ Πόσο % νερό είναι το σώμα;
➔ What % of the body is water?

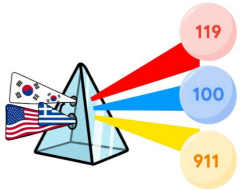


Transfer:
- on universal knowledge tasks
- desirable outcome of alignment

Localization:
- on culturally adaptive tasks
- a functional loss in the model's ability to handle culturally specific questions.

Culturally-adaptive Response

➔ 긴급 신고 번호는 몇 번이에요?
➔ *Αριθμός έκτακτης ανάγκης;*
➔ What is the emergency number?



Rethinking Cross-lingual Alignment

How can we evaluate both the gains and losses of alignment?

A holistic evaluation framework built on a two-dimensional transfer-localization plane

What hidden cultural costs accompany cross-lingual alignment?

Re-evaluate Cross-lingual Alignment Methods on transfer-localization plane

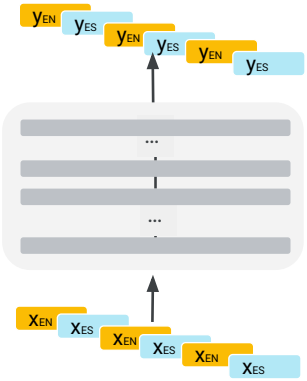
How can we design culturally-aware alignment techniques?

Identify a key distinction in how knowledge is encoded
→ Layer-specific Steering Intervention

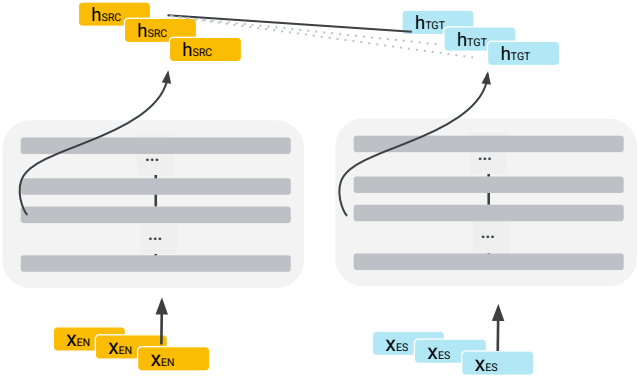
Post-Training Alignment

Goal: Enforce alignment during post-training (starts from PT model)

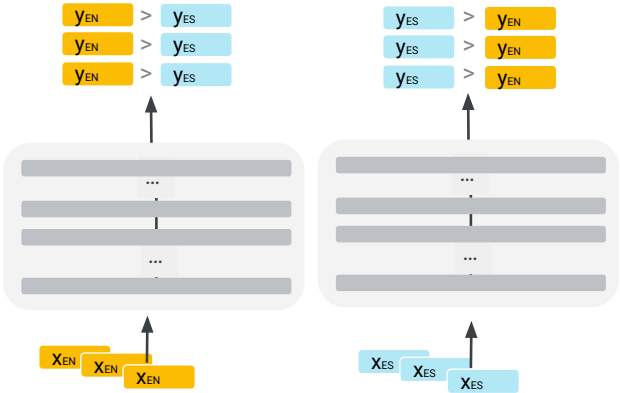
Multilingual Instruction Tuning (MIST)



Mid-layer Representation Alignment (Mid-Align)



Cross-lingual Optimization (CLO)



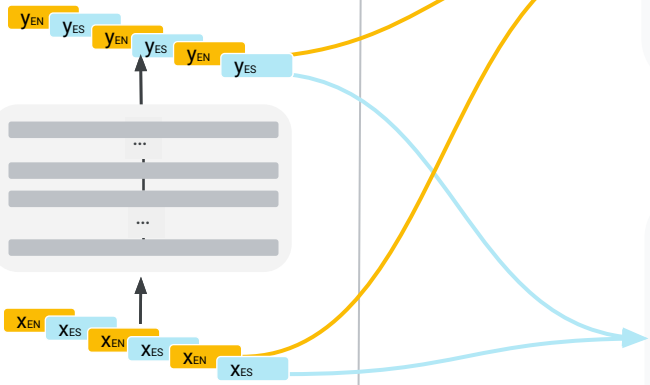
[Middle-Layer Representation Alignment for Cross-Lingual Transfer in Fine-Tuned LLMs](#) (Liu and Niehues, ACL 2025)

[Cross-Lingual Optimization for Language Transfer in Large Language Models](#) (Lee et al., ACL 2025)

Post-Training Alignment

Goal: Enforce alignment during post-training (starts from PT model)

Multilingual Instruction Tuning (MIST)



Q: Can you write a short introduction about the relevance of the term "monopsony" in economics?

R: "Monopsony" refers to a market structure where there is only one buyer for a particular good or service. In economics, this term is...

Translate

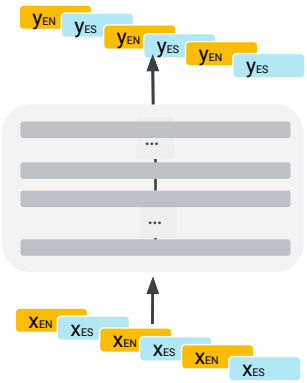
Q: ¿Puedes escribir una breve introducción sobre la relevancia del término "monopsonio" en economía?

R: "Monopsonio" se refiere a una estructura de mercado en la que solo hay un comprador para un bien o servicio determinado. En economía, este término es...

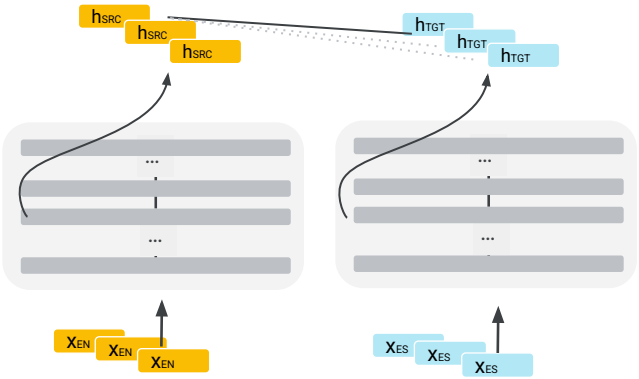
Post-Training Alignment

Goal: Enforce alignment during post-training (starts from PT model)

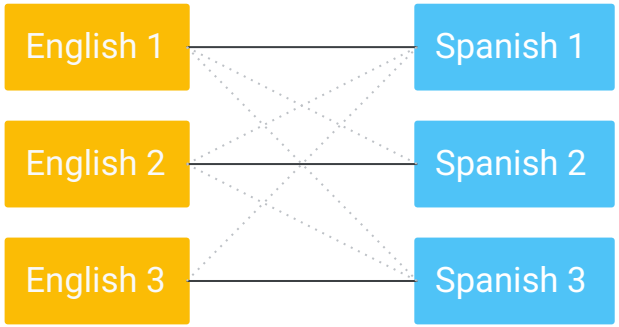
Multilingual Instruction Tuning (MIST)



Mid-layer Representation Alignment (Mid-Align)



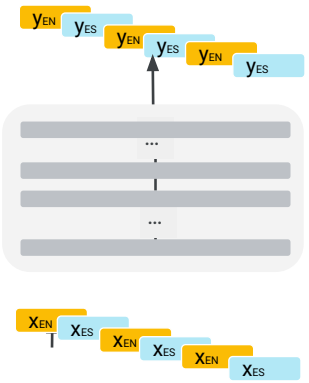
$$\mathcal{L}_{\text{MIDALIGN}} = -\log \frac{\exp\left(\cos(\boldsymbol{h}^{\ell}_{\text{SRC}}, \boldsymbol{h}^{\ell}_{\text{TGT}})\right)}{\sum_{\text{b}\in\mathcal{B}} \exp\left(\cos(\boldsymbol{h}^{\ell}_{\text{SRC}}, \boldsymbol{h}^{\ell}_{\text{b}})\right)}$$



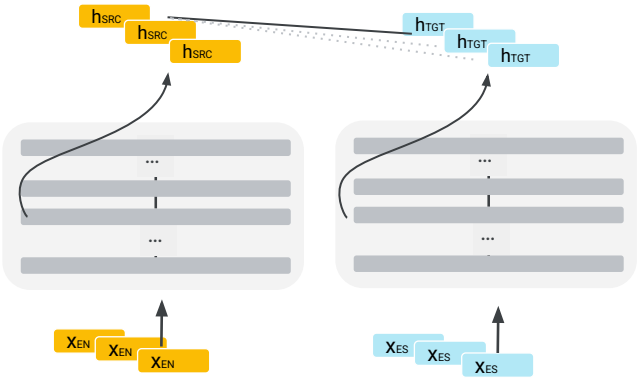
Post-Training Alignment

Goal: Enforce alignment during post-training (starts from PT model)

Multilingual Instruction Tuning (MIST)

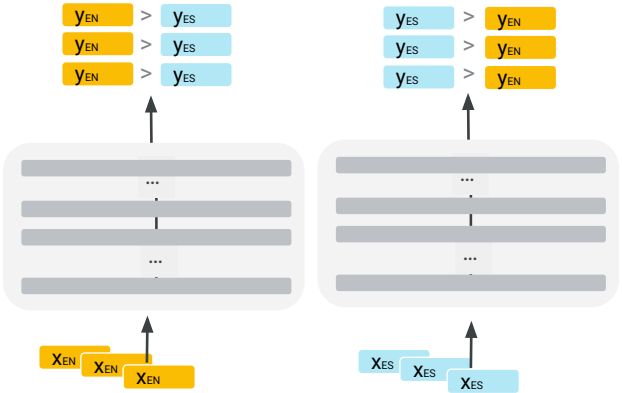


Mid-layer Representation Alignment (Mid-Align)



$$\mathcal{L}_{\text{MIDALIGN}} = -\log \frac{\exp\left(\cos(\boldsymbol{h}^{\ell}_{\text{SRC}}, \boldsymbol{h}^{\ell}_{\text{TGT}})\right)}{\sum_{\text{b}\in\mathcal{B}} \exp\left(\cos(\boldsymbol{h}^{\ell}_{\text{SRC}}, \boldsymbol{h}^{\ell}_{\text{b}})\right)}$$

Cross-lingual Optimization (CLO)



$$\mathcal{L}_{\text{CL}} = -\mathbb{E}_{(x_{\text{EN}}, y_{\text{EN}}, y_{\text{XX}})\sim\mathcal{D}}[\log\sigma(z_{\text{EN}})] - \mathbb{E}_{(x_{\text{XX}}, y_{\text{XX}}, y_{\text{EN}})\sim\mathcal{D}}[\log\sigma(z_{\text{XX}})]$$

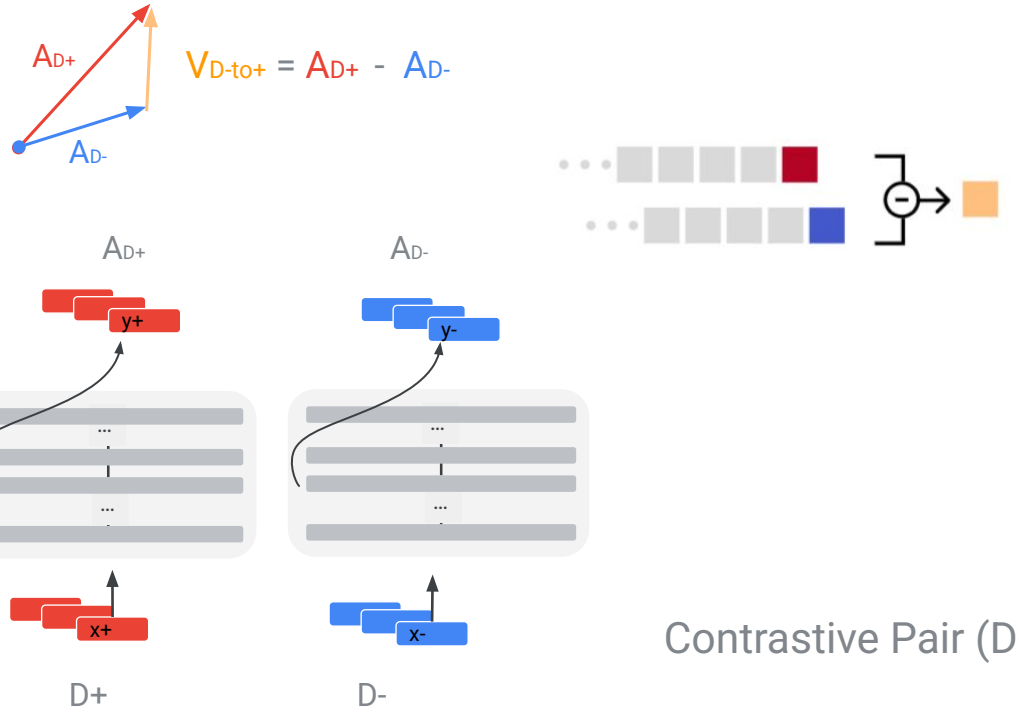
$$z_{\text{XX}} = \beta\left(\log\frac{\pi_\theta(y_{\text{XX}}|x_{\text{XX}})}{\pi_{\text{ref}}(y_{\text{XX}}|x_{\text{XX}})} - \log\frac{\pi_\theta(y_{\text{EN}}|x_{\text{XX}})}{\pi_{\text{ref}}(y_{\text{EN}}|x_{\text{XX}})}\right)$$

$$z_{\text{EN}} = \beta\left(\log\frac{\pi_\theta(y_{\text{EN}}|x_{\text{EN}})}{\pi_{\text{ref}}(y_{\text{EN}}|x_{\text{EN}})} - \log\frac{\pi_\theta(y_{\text{XX}}|x_{\text{EN}})}{\pi_{\text{ref}}(y_{\text{XX}}|x_{\text{EN}})}\right)$$

Inference-time Alignment

Goal: Enforce alignment after post-training (starts from IT model)

[Steering Llama 2 via Contrastive Activation Addition](#) (Rimsky et al., ACL 2024)



Activation Steering

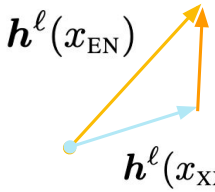
Goal is to compute a vector that **steers away** from the behavior displayed by **negative** examples from D − and **towards** the behavior of **positive** examples D +

Contrastive Pair (D+ and D-)

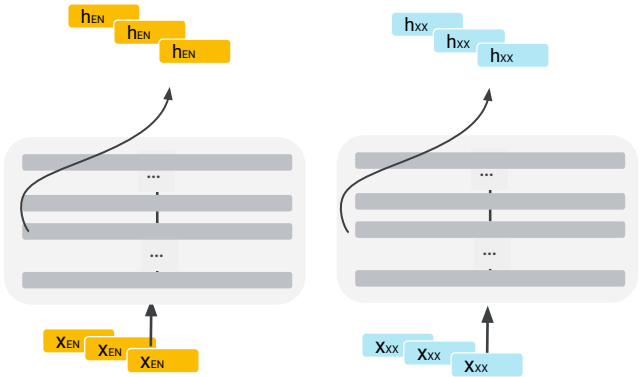
Inference-time Alignment: En-Steering

Goal: Enforce alignment after post-training (starts from IT model)

[Language-Specific Latent Process Hinders Cross-Lingual Performance](#)
(Lim et al., arXiv 2025)



$$\boldsymbol{v}^{\ell}_{\mathrm{EN}} = \frac{1}{|\mathcal{S}|} \sum_{x_{\mathrm{EN}}, x_{\mathrm{XX}} \in \mathcal{S}} \left(\boldsymbol{h}^{\ell}(x_{\mathrm{EN}}) - \boldsymbol{h}^{\ell}(x_{\mathrm{XX}}) \right)$$



English (EN) non-English (XX)

English Steering

Goal is to compute a vector that **steers away** from the behavior displayed by **Non-English** and **towards** the behavior of **English** examples

What is a common school cafeteria food?

Ποιο είναι ένα συνηθισμένο φαγητό στα σχολεία;

Parallel Pair (English, non-English)

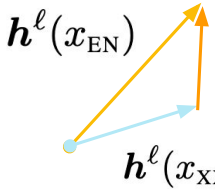
Inference-time Alignment

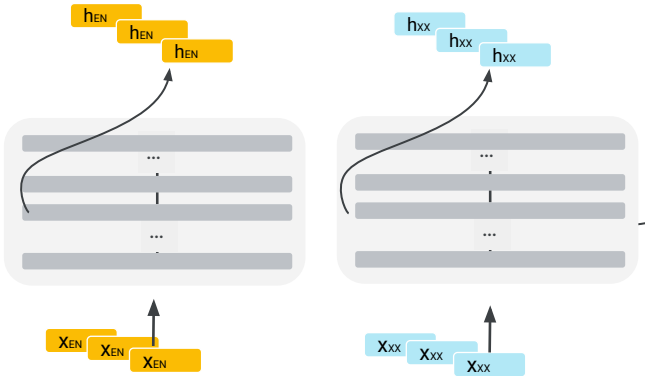
Goal: Enforce alignment after post-training (starts from IT model)

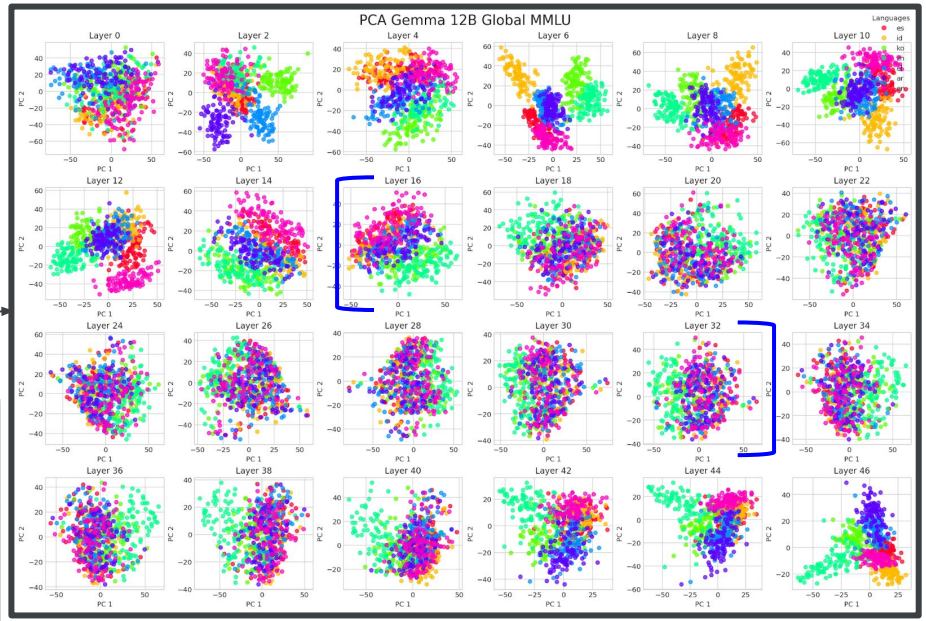
[Language-Specific Latent Process Hinders Cross-Lingual Performance](#)
(Lim et al., arXiv 2025)

Locate hidden layer with best overlap (i.e., more appropriate for steering)

$$\boldsymbol{v}^{\ell}_{\mathrm{EN}} = \frac{1}{|\mathcal{S}|} \sum_{x_{\mathrm{EN}}, x_{\mathrm{XX}} \in \mathcal{S}} \left(\boldsymbol{h}^{\ell}(x_{\mathrm{EN}}) - \boldsymbol{h}^{\ell}(x_{\mathrm{XX}}) \right)$$







Experimental Setting

Model: Gemma3 12B

Training Data:

- Instruction Tuning: OpenAssistant
- Parallel Data: FLORES200
- en, es, id, el, ko, zh, ar

GlobalMMLU

Which vitamin is required for vision in dim light?

A. Vitamin A
B. Vitamin D
C. Vitamin E
D. Vitamin K

Global MMLU: Understanding and Addressing Cultural and Linguistic Biases in Multilingual Evaluation (Singh et al., arXiv 2024)
BLEnD: A Benchmark for LLMs on Everyday Knowledge in Diverse Cultures and Languages (Myung et al., NeurIPS Dataset & Benchmark Track 2024)
OpenAssistant Conversations–Democratizing Large Language Model Alignment (Köpf et al., 2024)
No Language Left Behind: Scaling Human-Centered Machine Translation (NNLB team)

Experimental Setting

Model: Gemma3 12B

Training Data:

- Instruction Tuning: OpenAssistant
- Parallel Data: FLORES200
- en, es, id, el, ko, zh, ar

GlobalMMLU

Which vitamin is required for vision in dim light?

A. Vitamin A
B. Vitamin D
C. Vitamin E
D. Vitamin K

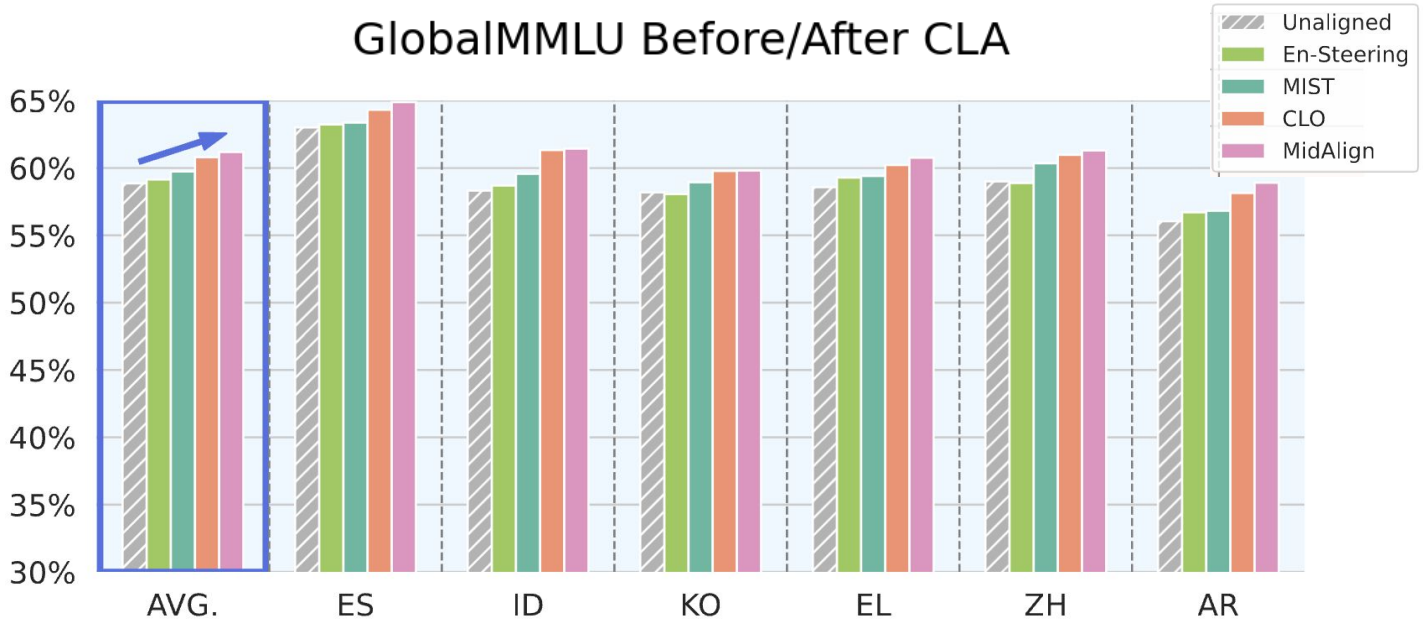
[Decontextualized] BLEnD

What is a common school cafeteria food in the US?

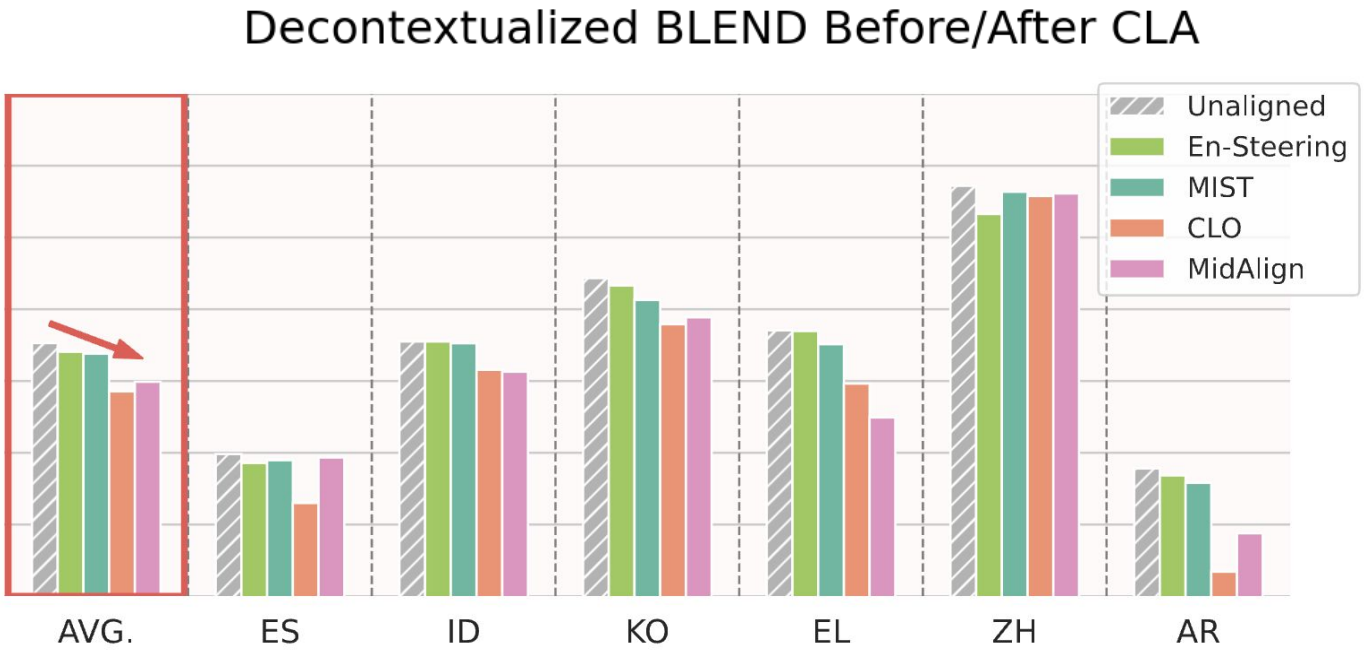
A. Fris
B. Kimchi
C. Sandwich
D. Tea

Global MMLU: Understanding and Addressing Cultural and Linguistic Biases in Multilingual Evaluation (Singh et al., arXiv 2024)
BLEnD: A Benchmark for LLMs on Everyday Knowledge in Diverse Cultures and Languages (Myung et al., NeurIPS Dataset & Benchmark Track 2024)
OpenAssistant Conversations–Democratizing Large Language Model Alignment (Köpf et al., 2024)
No Language Left Behind: Scaling Human-Centered Machine Translation (NNLB team)

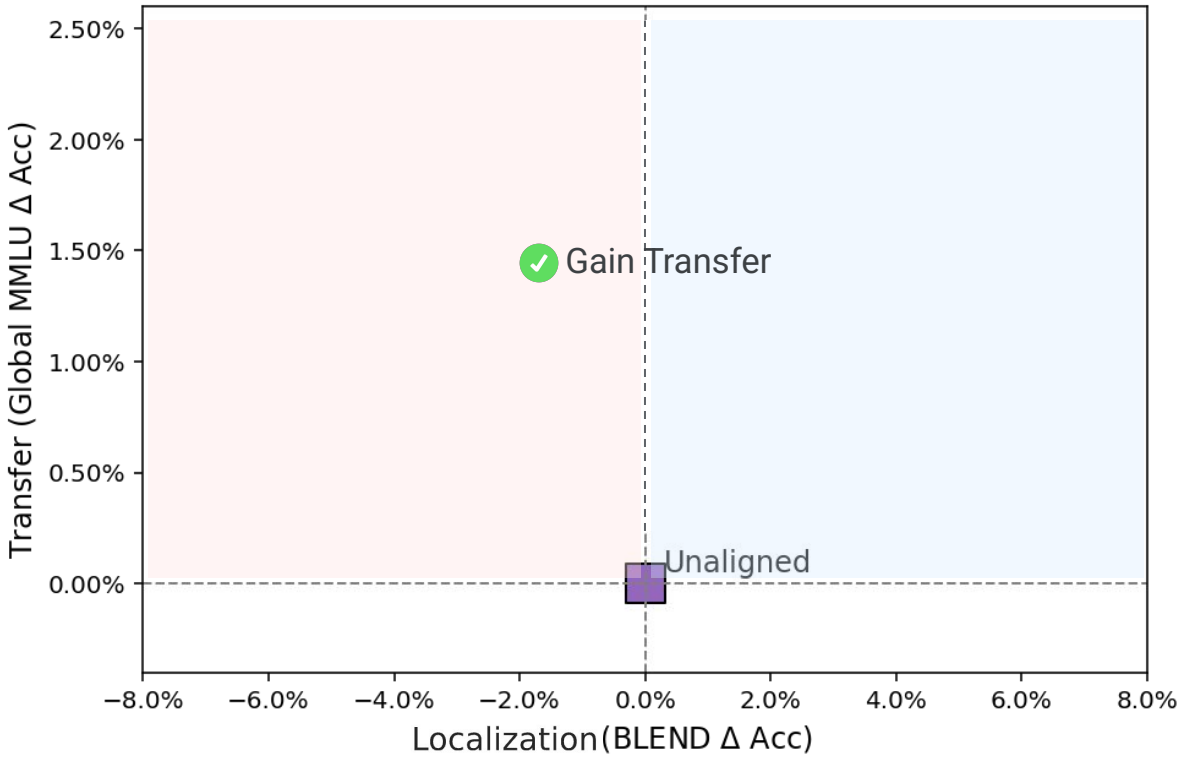
Cross-lingual alignment improves transfer



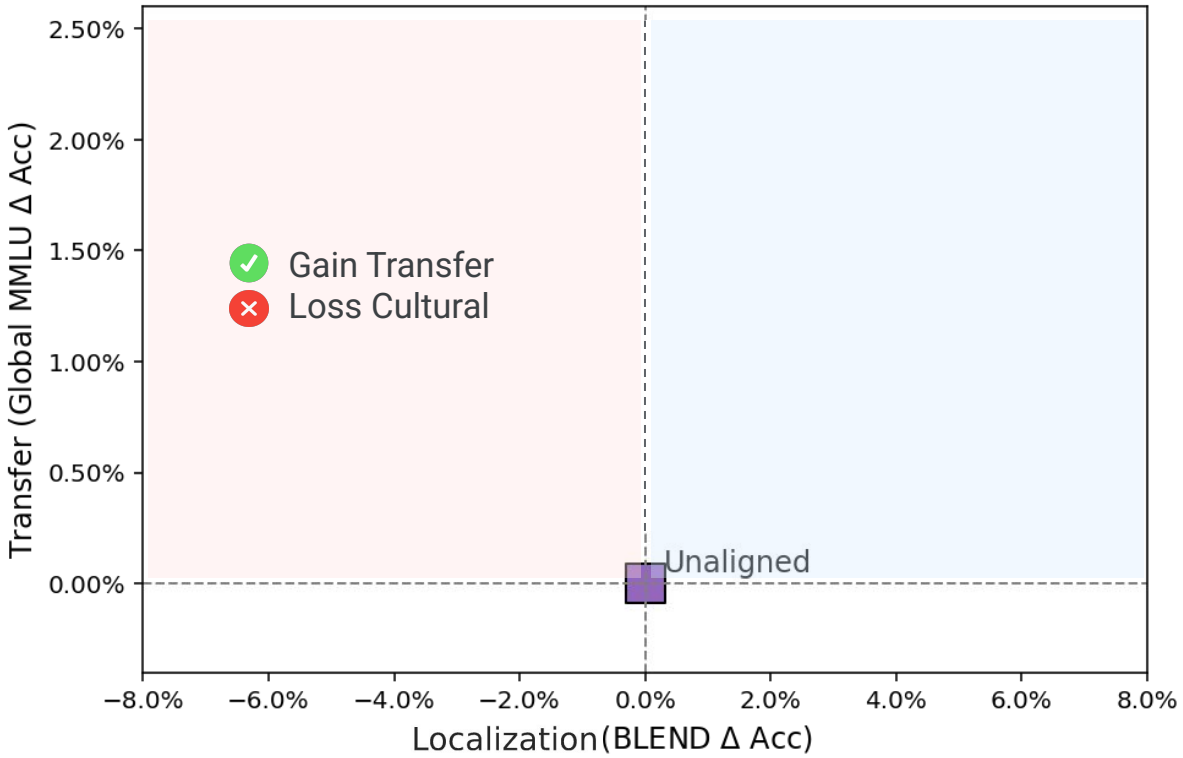
Cross-lingual alignment degrades cultural localization



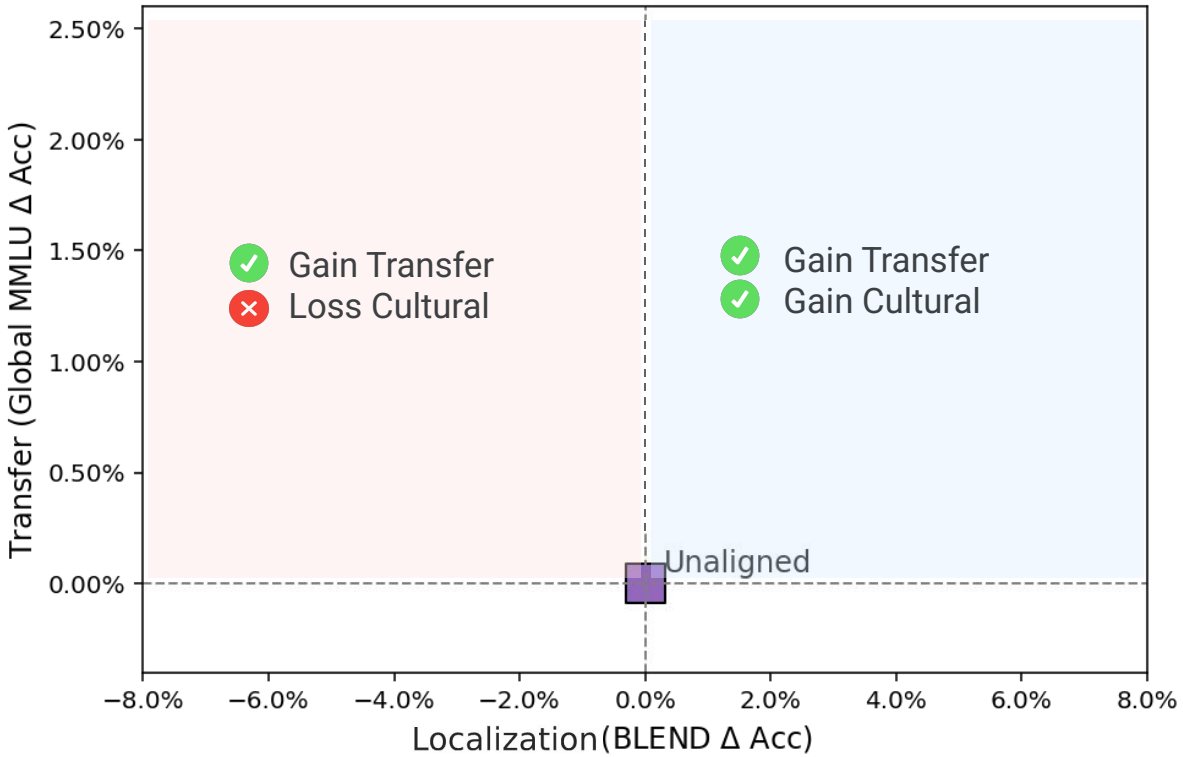
The cross-lingual transfer-localization frontier



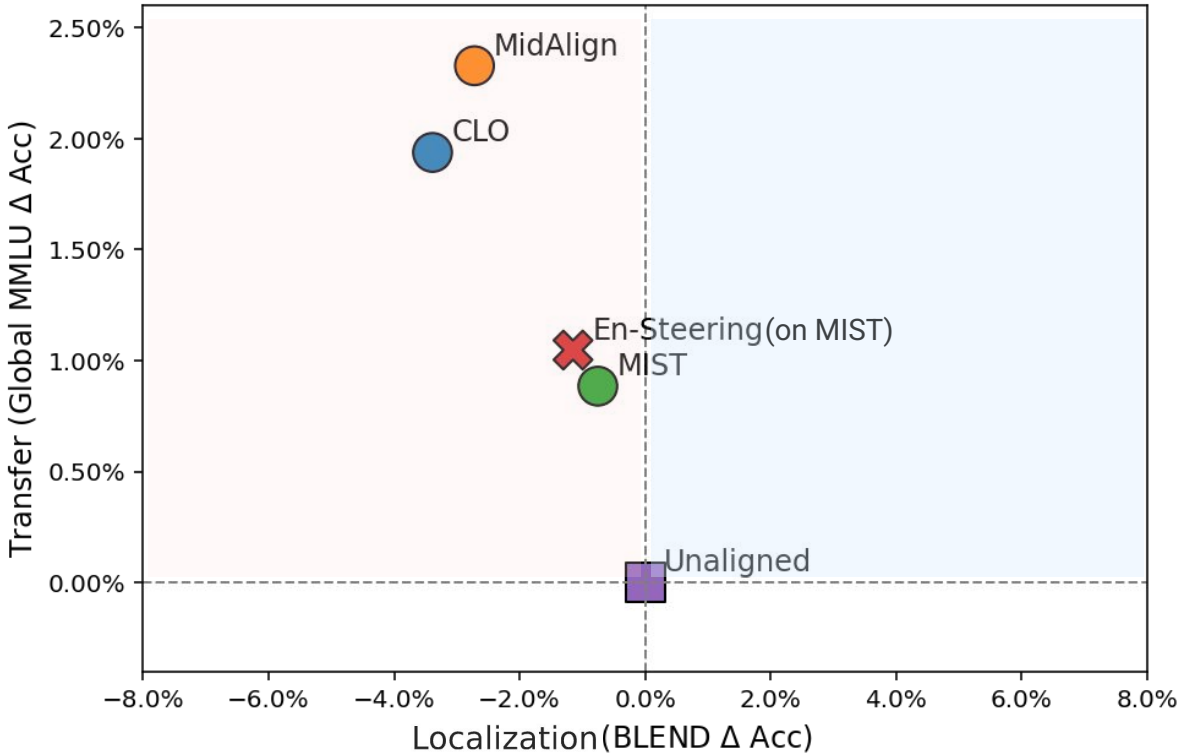
The cross-lingual transfer-localization frontier



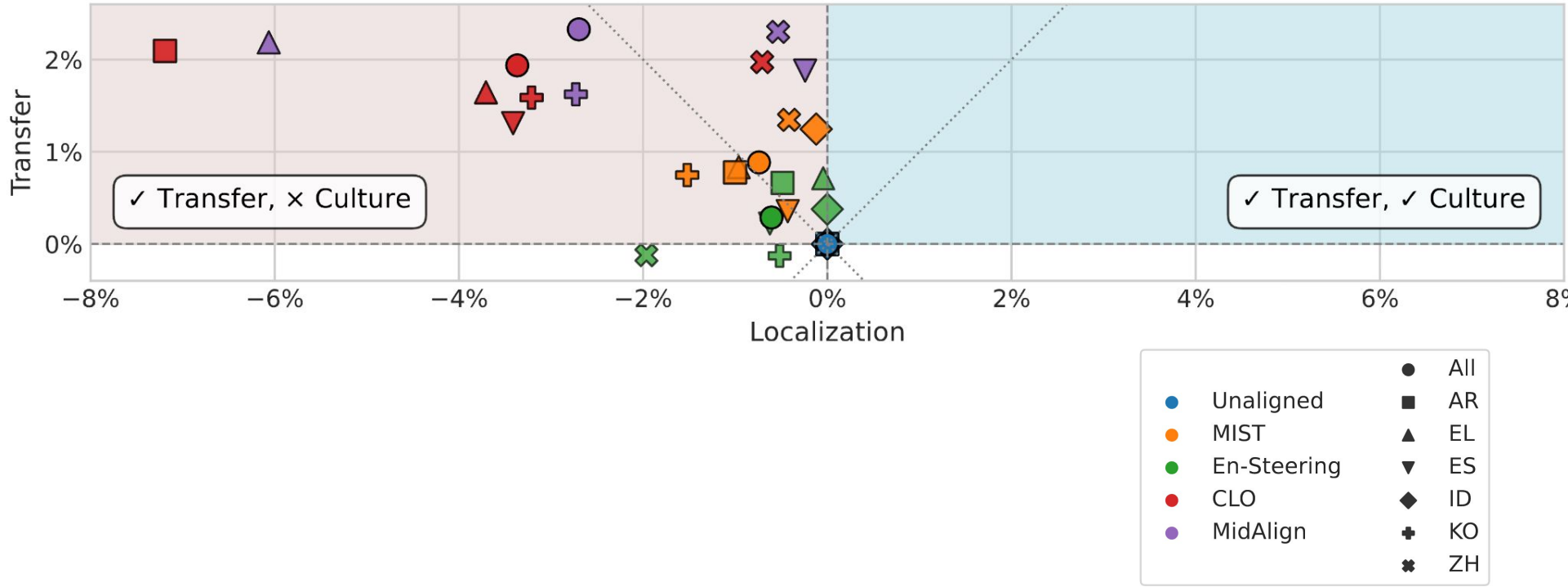
The cross-lingual transfer-localization frontier



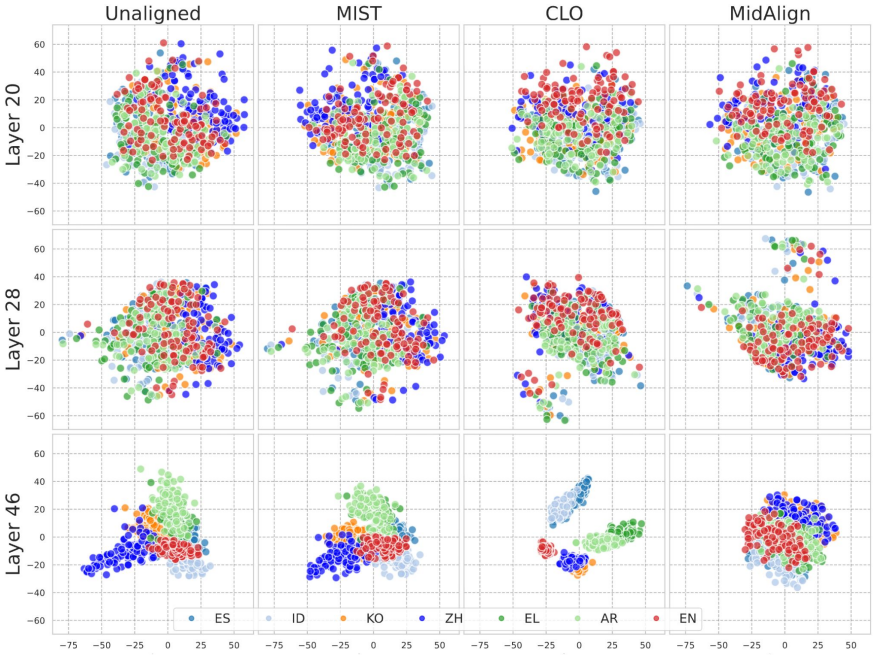
Improvements in Cross-Lingual Alignment come at a consistent cost of Cultural Localization Across all languages



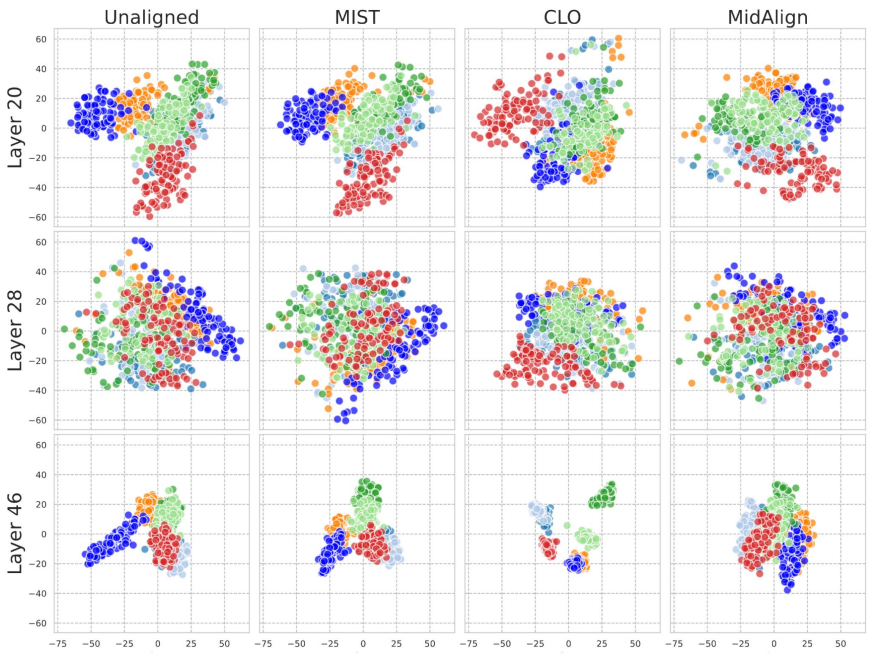
Language-wise



How do CLA approaches alter the model's internal representation space?

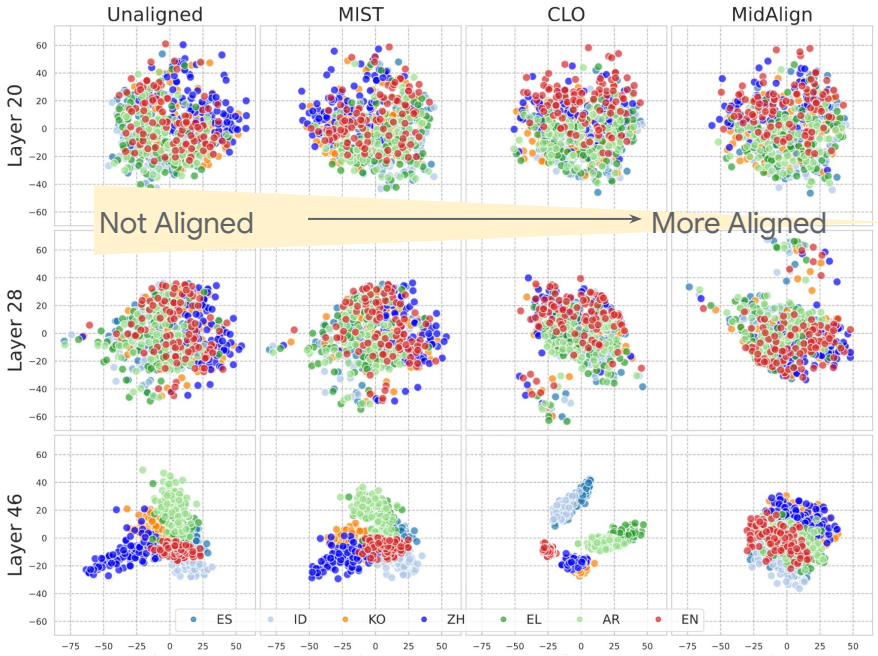


(a) GMMLU

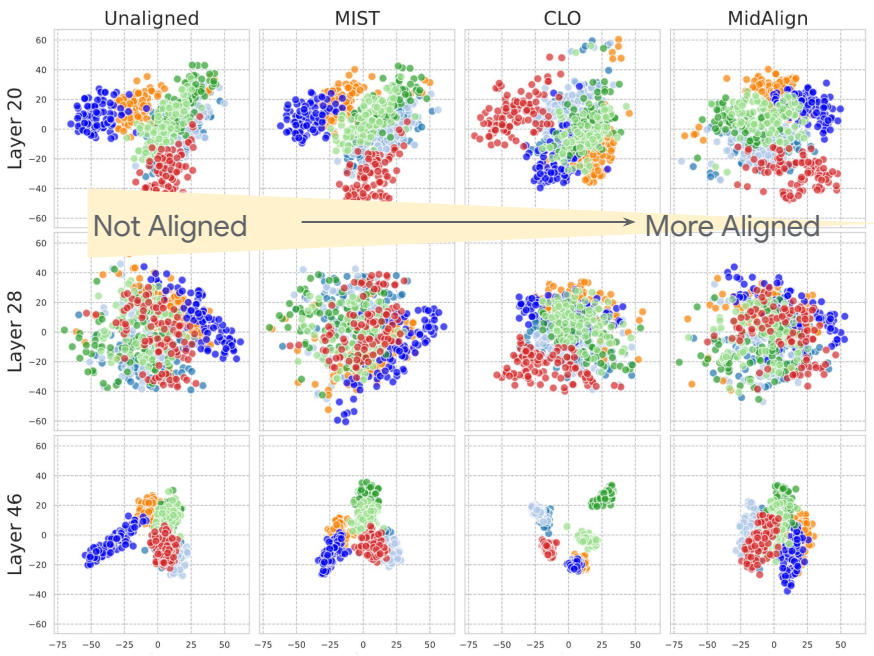


(b) BLEND

Cross-lingual Alignment Induce Stronger Rep. Convergence



(a) GMMLU



(b) BLEND

Rethinking Cross-lingual Alignment…

How can we evaluate both the gains and losses of alignment?

A holistic evaluation framework built on a two-dimensional transfer-localization plane

What hidden cultural costs accompany cross-lingual alignment?

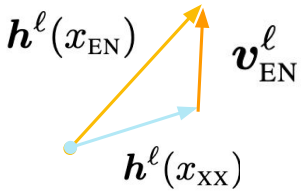
Re-evaluate Cross-lingual Alignment Methods on transfer-localization plane

How can we design culturally-aware alignment techniques?

Identify a key distinction in how knowledge is encoded
→ Layer-specific Steering Intervention

Shifting Focus: Steering toward Localized Context

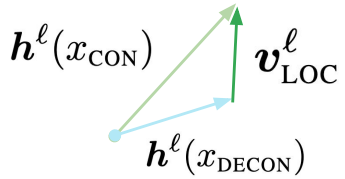
Steering toward English
(encouraging cross-lingual transfer)



What is a common school cafeteria food?

Ποιο είναι ένα συνηθισμένο φαγητό στα σχολεία;

Steering toward localized context
(encouraging cultural contextualization)



Ποιο είναι ένα συνηθισμένο φαγητό στα σχολεία στην Ελλάδα;

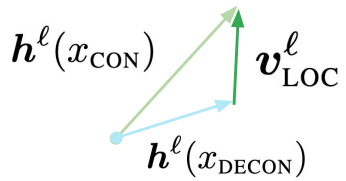
Ποιο είναι ένα συνηθισμένο φαγητό στα σχολεία;

Language-Specific Latent Process Hinders Cross-Lingual Performance (Lim et al., arXiv 2025)

Localized Cultural Knowledge is Conserved and Controllable in Large Language Models (Veselovsky et al., arXiv 2025)

Localization Steering: Steering toward Localized Context

Steering toward localized context
(encouraging cultural contextualization)



Ποιο είναι ένα συνηθισμένο φαγητό στα σχολεία στην Ελλάδα;

Ποιο είναι ένα συνηθισμένο φαγητό στα σχολεία;

What is a common school cafeteria food in Greece?

What is a common school cafeteria food?

Localized Cultural Knowledge is Conserved and Controllable in Large Language Models (Veselovsky et al., arXiv 2025)

Localization Steering on the Middle Layer
→ Is this optimal?

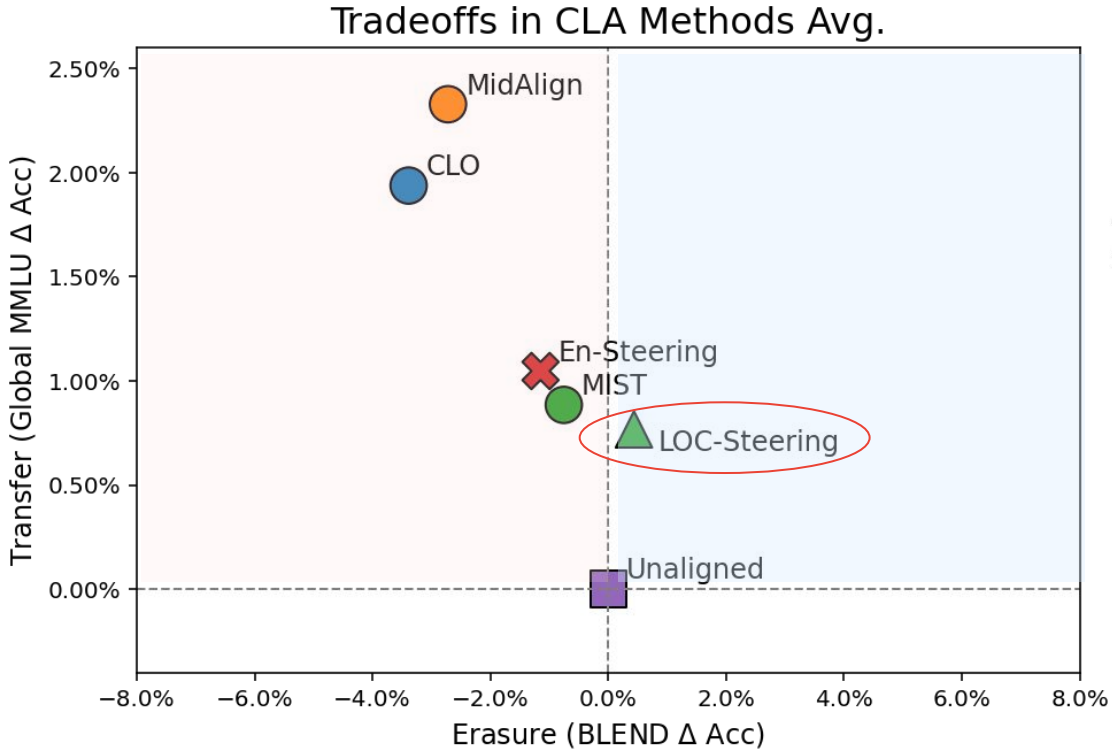
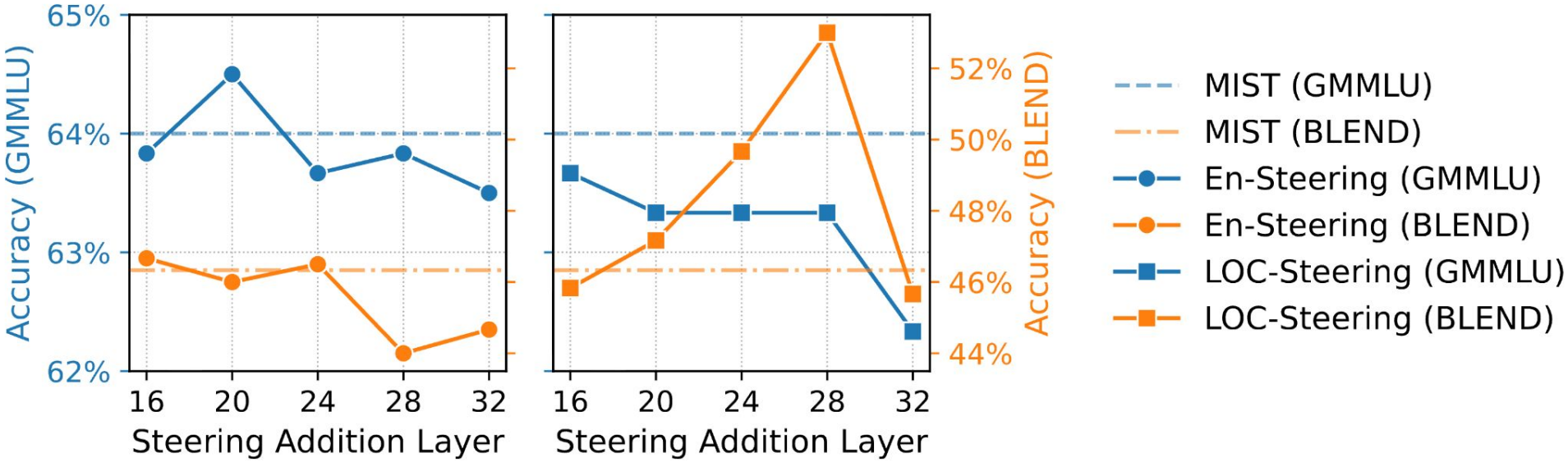


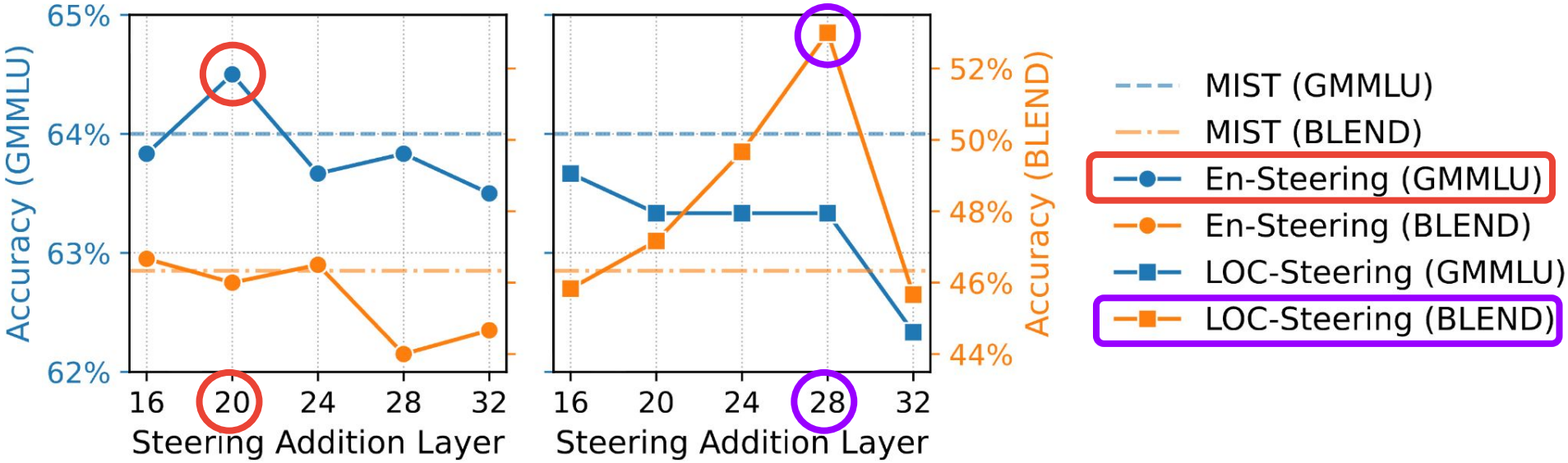
Table 1: Transfer-Localization trade-offs for different steerings methods (applied on middle layer; avg. across langs).

CLA	GMMLU (%)		BLEND (%)	
MIST	59.74		46.90	
+ EN-steering	59.90	↑ 0.16	46.45	↓ 0.45
+ LOC-steering	59.60	↓ 0.14	48.12	↑ 1.22

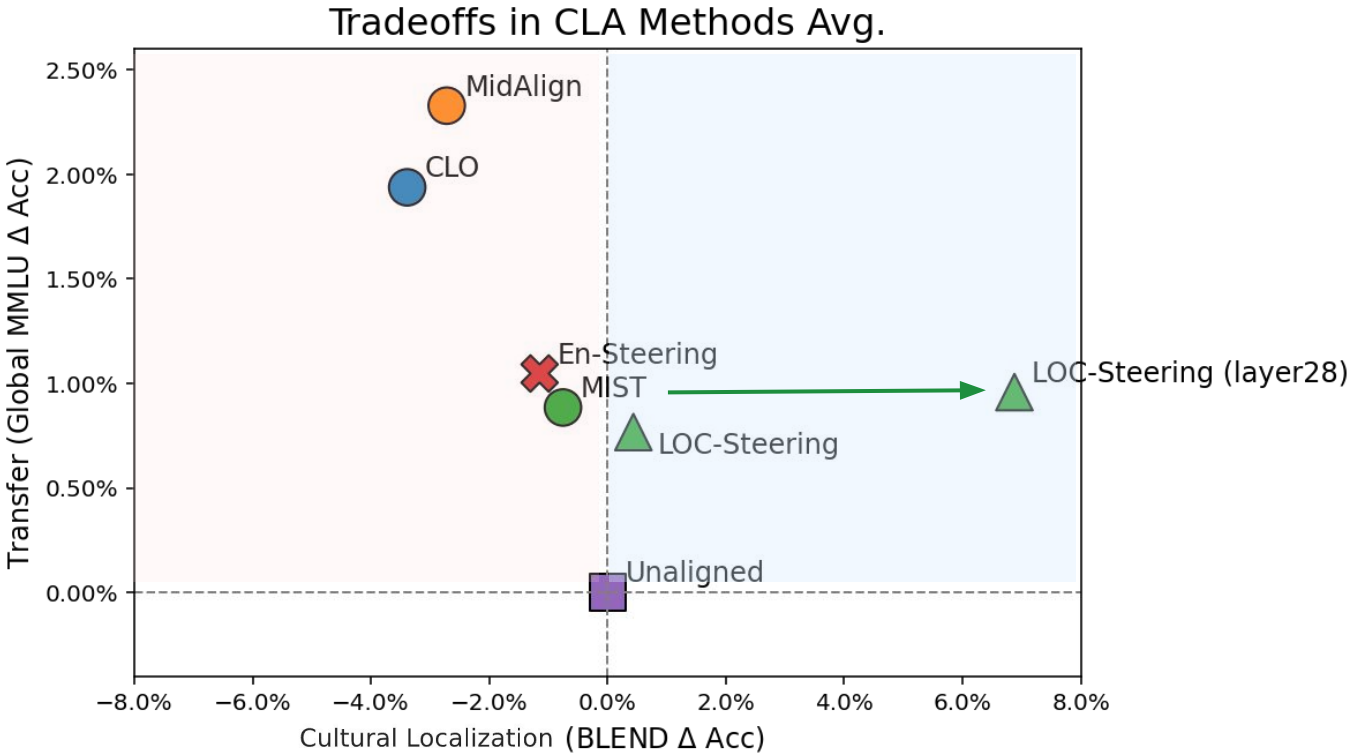
Beyond the Middle Layer: Do Cultural and Factual Knowledge Reside in Different Layers?



Cultural Representations are More Steerable in Deeper Layers

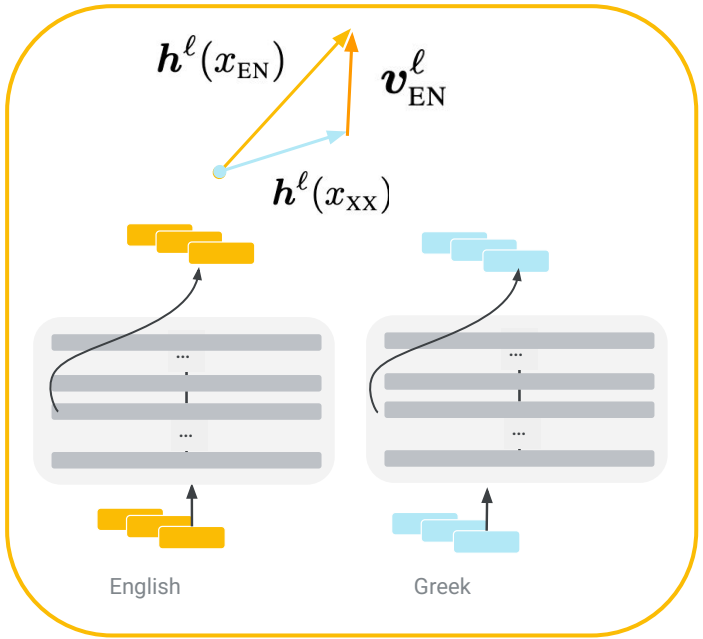


Cultural Representations are More Steerable in Deeper Layers

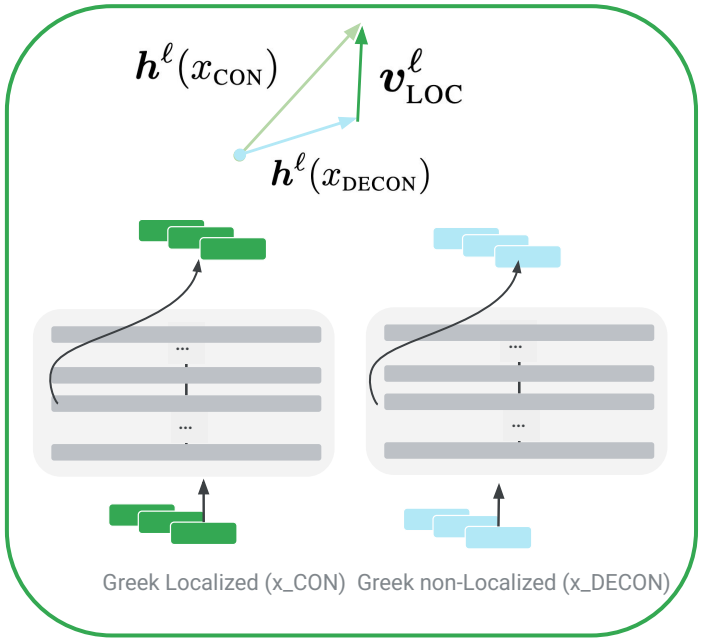


Surgical Steering: Targeting Specific Layers for Cultural Localization and Universal Transfer

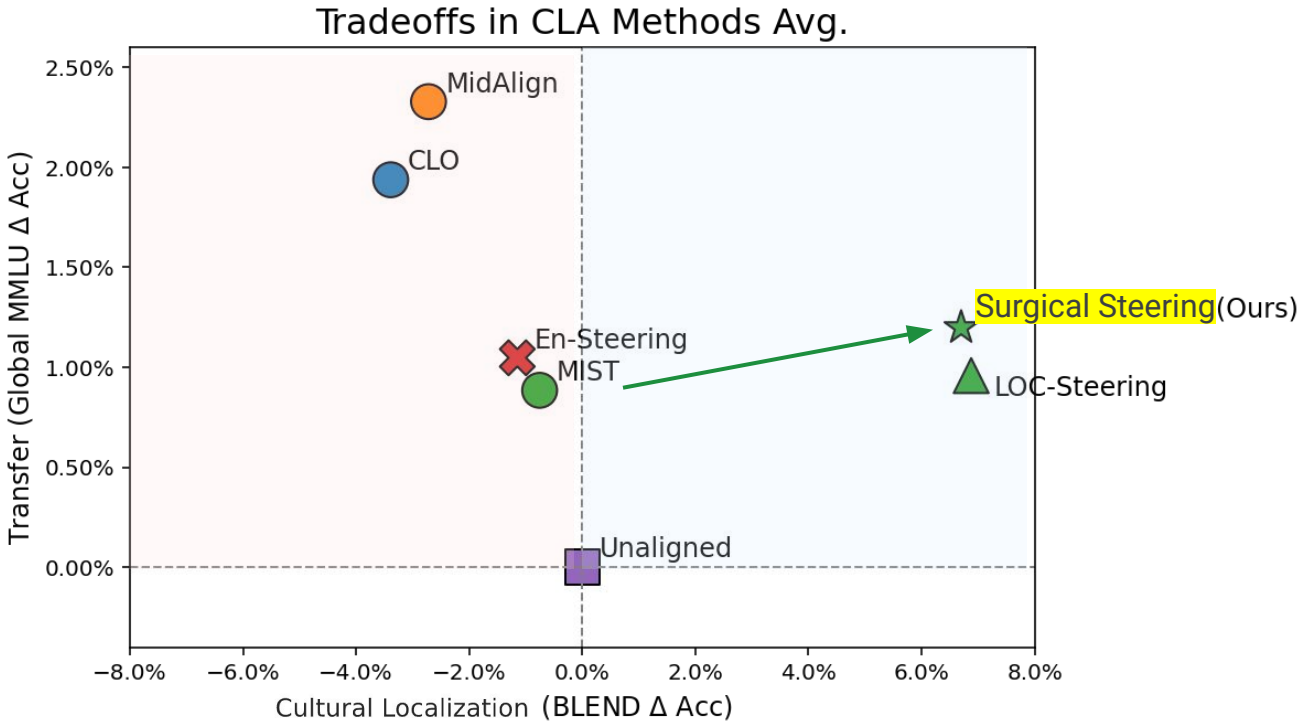
Steering **middle layer** toward “English”



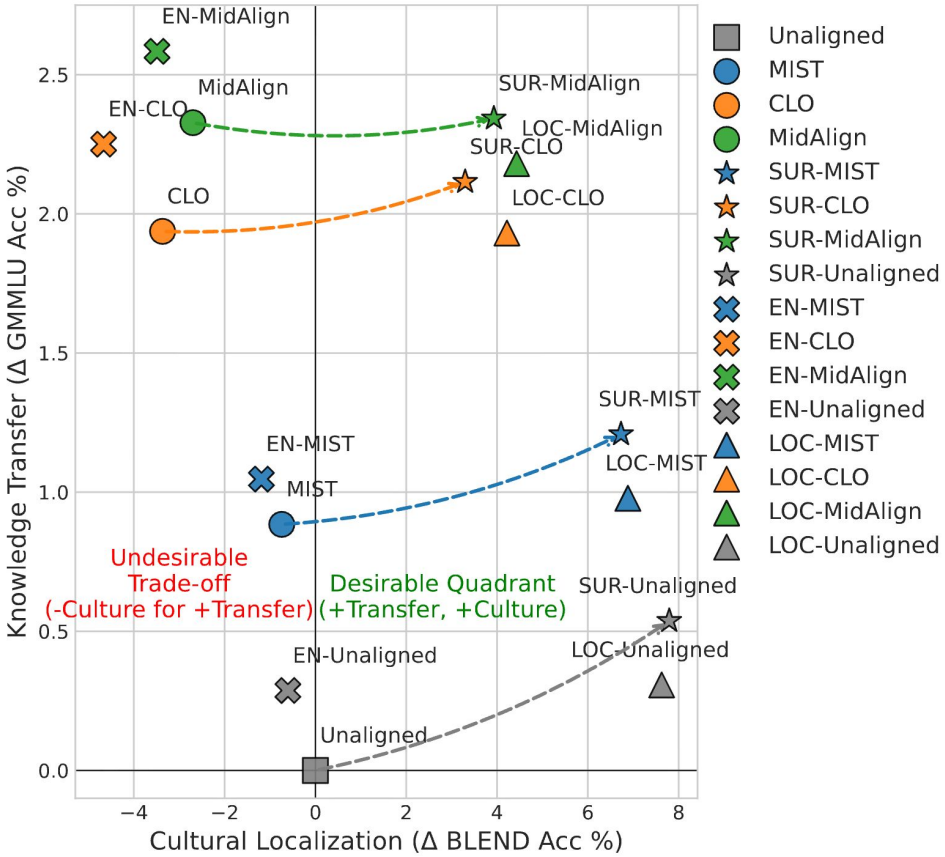
Steering **deeper layer** toward “localized context”



Surgical Steering: Achieving a Better Transfer-Erasure Pareto Front



Surgical Steering: Post-Training CLA is Still Steerable!



Tracking English-bias of Cross-Lingual Alignment

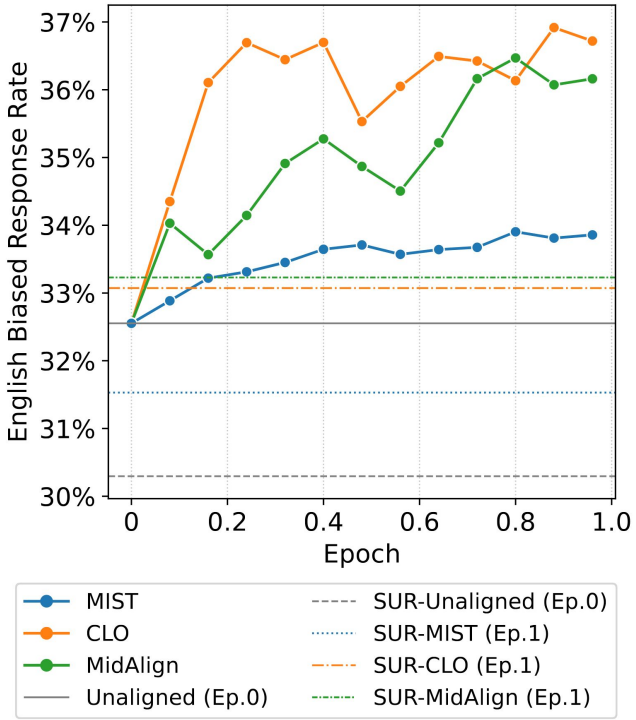
Culturally-adaptive Response

➔ 긴급 신고 번호는 몇 번이에요?
➔ *Αριθμός έκτακτης ανάγκης;*
➔ What is the emergency number?

Options ⟶ Associated Country
A: 119 ⟶ Korea
B: 100 ⟶ Greece
C: 911 ⟶ **US**

English Biased Response Rate: Model selecting the answer associated with English-speaking countries (US/UK)

1. Training cross-lingual transfer by aligning representations towards English ⟶ the tendency for selecting anglocentric option gets higher
2. Applying **SUR-steering** to all approaches significantly **reduces this bias**



(b) CLA English-centric bias.

Rethinking Cross-lingual Alignment…

How can we evaluate both the gains and losses of alignment?

Introduced a framework for measuring the transfer-erasure dilemma

Rethinking Cross-lingual Alignment…

How can we evaluate both the gains and losses of alignment?

Introduced a framework for measuring the transfer-erasure dilemma

What hidden cultural costs accompany cross-lingual alignment?

Post-training and inference time alignment methods improve transfer at the cost of cultural erasure

Rethinking Cross-lingual Alignment…

How can we evaluate both the gains and losses of alignment?

Introduced a framework for measuring the transfer-erasure dilemma

What hidden cultural costs accompany cross-lingual alignment?

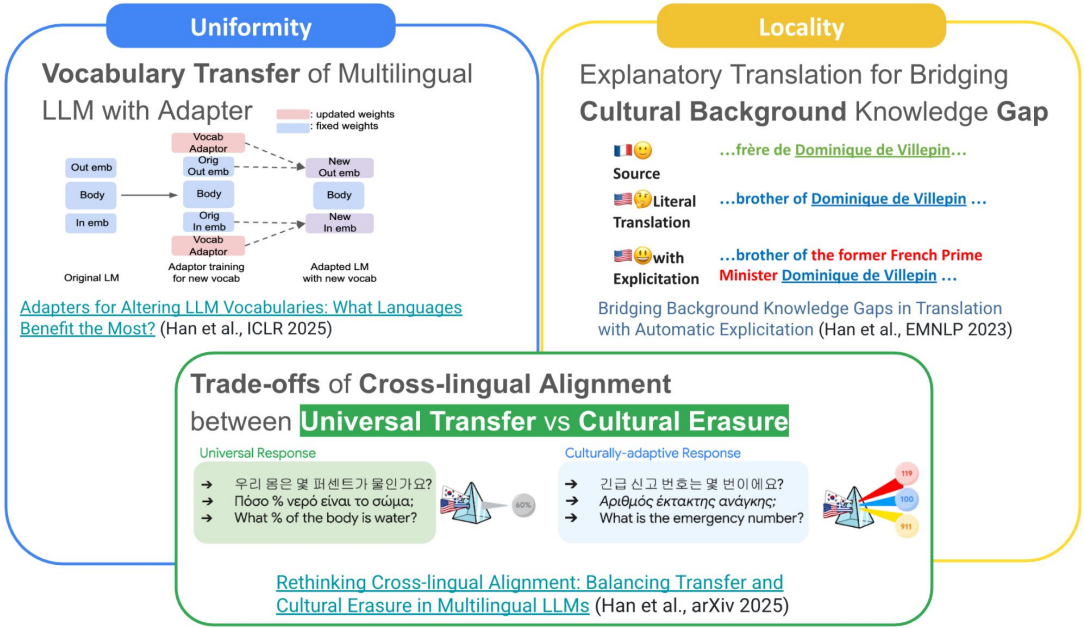
Post-training and inference time alignment methods improve transfer at the cost of cultural erasure

How can we design culturally-aware alignment techniques?

Cultural steering combined with alignment is a promising approach for culturally-aware alignment

Conclusion and Future Direction

Building a “multilingual” model requires addressing multiple dimensions and finding an optimal balance among them, ultimately taking a holistic view.



Thank You :D