Adapters for Altering LLM Vocabularies: What Languages Benefit the Most?

HyoJung Han, Akiko Eriguchi, Haoran Xu, Hieu Hoang, Marine Carpuat, Huda Khayrallah

hjhan@cs.umd.edu

Vocabulary Adaptation

A process of modifying a pre-trained language model (LM) to use a new vocabulary

Advantages:

1. Introduction of new languages into a model
2. Improving downstream performance in target language
3. Mitigating **over-fragmentation** (words are excessively split by the tokenizer)

Vocabulary Adaptation

A process of modifying a pre-trained language model (LM) to vocabulary

Advantages:

(Ahia et al., EMNLP 2023)

1. Introduction of new languages into a model
2. Improving downstream performance in target language
3. Mitigating **over-fragmentation** (words are excessively split by the tokenizer)

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

Language Model Tokenizers Introduce Unfairness Between Languages (Petrov et al., NeurIPS 2023)
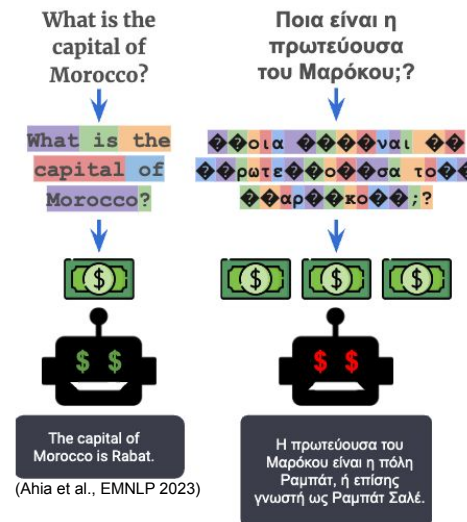
"We show evidence that speakers of a large number of the supported languages are **overcharged** while obtaining **poorer results**."

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

Limitations of Existing Vocabulary Adaptation Approaches

- **Heuristics based initialization** for new embeddings from existing ones

  → Lack adaptability, not fully integrated, requires additional training

- **Dependency on external embeddings** or dictionaries

  → Increase complexity and limit scalability

Vocabulary Adaptation	External Resources
RAMEN (Tran, 2020)	FastAlign, fastText
OFA (Liu et al., 2024)	ColexNet+
FOCUS (Dobler & de Melo, 2023)	fastText
WECHSEL (Minixhofer et al., 2022)	fastText, bilingual dictionaries
CW2V (Mundra et al., 2024)	bilingual dictionaries

Limitations of Existing Vocabulary Adaptation Approaches

- **Heuristics based initialization** for new embeddings from existing ones

  → Lack adaptability, not fully integrated, requires additional training

- **Dependency on external embeddings** or dictionaries

  → Increase complexity and limit scalability

- **Language-specific** approach or
  restrictions on the number of languages

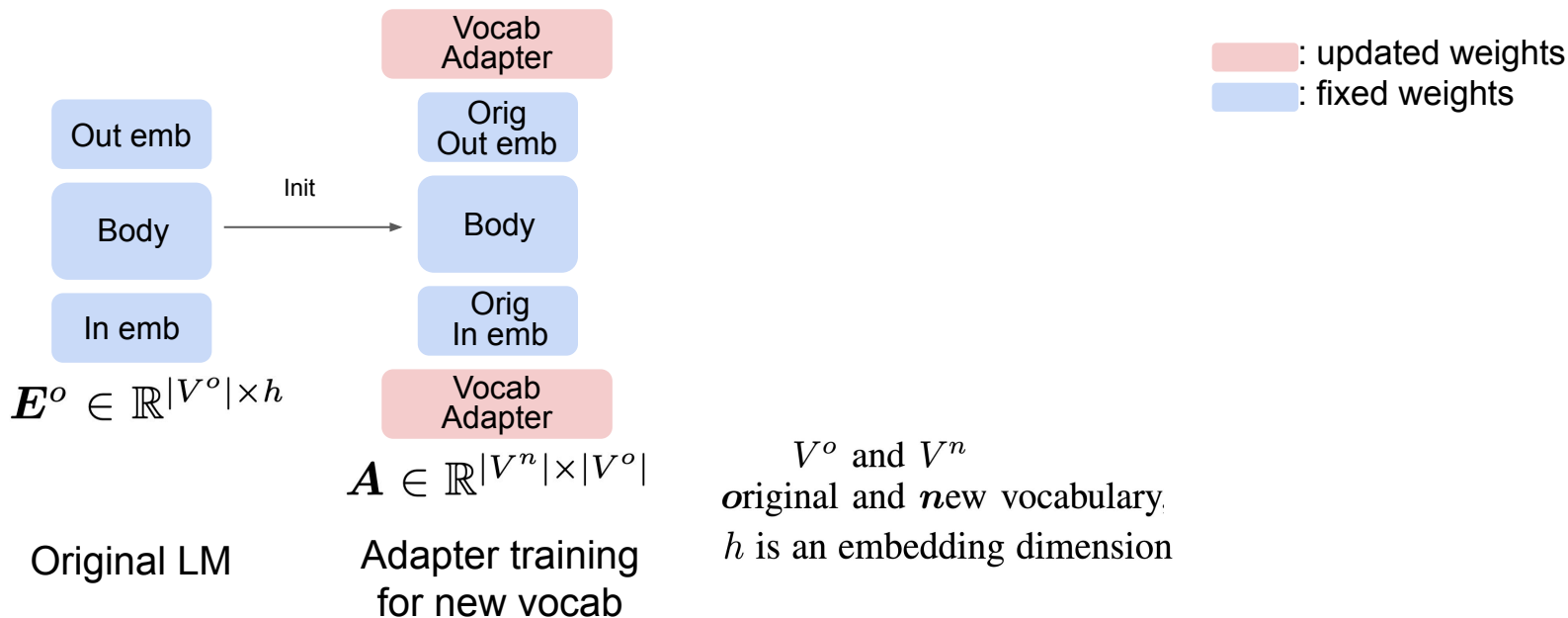| Vocabulary Adaptation	Grouping
ZeTT (Minixhofer et al., 2024)	lang-specific
RAMEN (Tran, 2020)	lang-specific
FOCUS (Dobler & de Melo, 2023)	lang-specific
MAD-X (Pfeiffer et al., 2020)	lang-specific
WECHSEL (Minixhofer et al., 2022)	lang-specific
CLP (Ostendorff & Rehm, 2023)	lang-specific
CLP+ (Yamaguchi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

Understudied Impact of Vocabulary Adaptation Across Diverse Linguistic and Task Settings

- Most prior work investigates few languages
  - Works with many languages lacks detailed analysis

- The impact of vocabulary adaptation on cross-lingual and generative tasks like machine translation (MT) is understudied.
  - Evaluated instead on non-cross-lingual and discriminative tasks (NLI, multiple-choice QA)

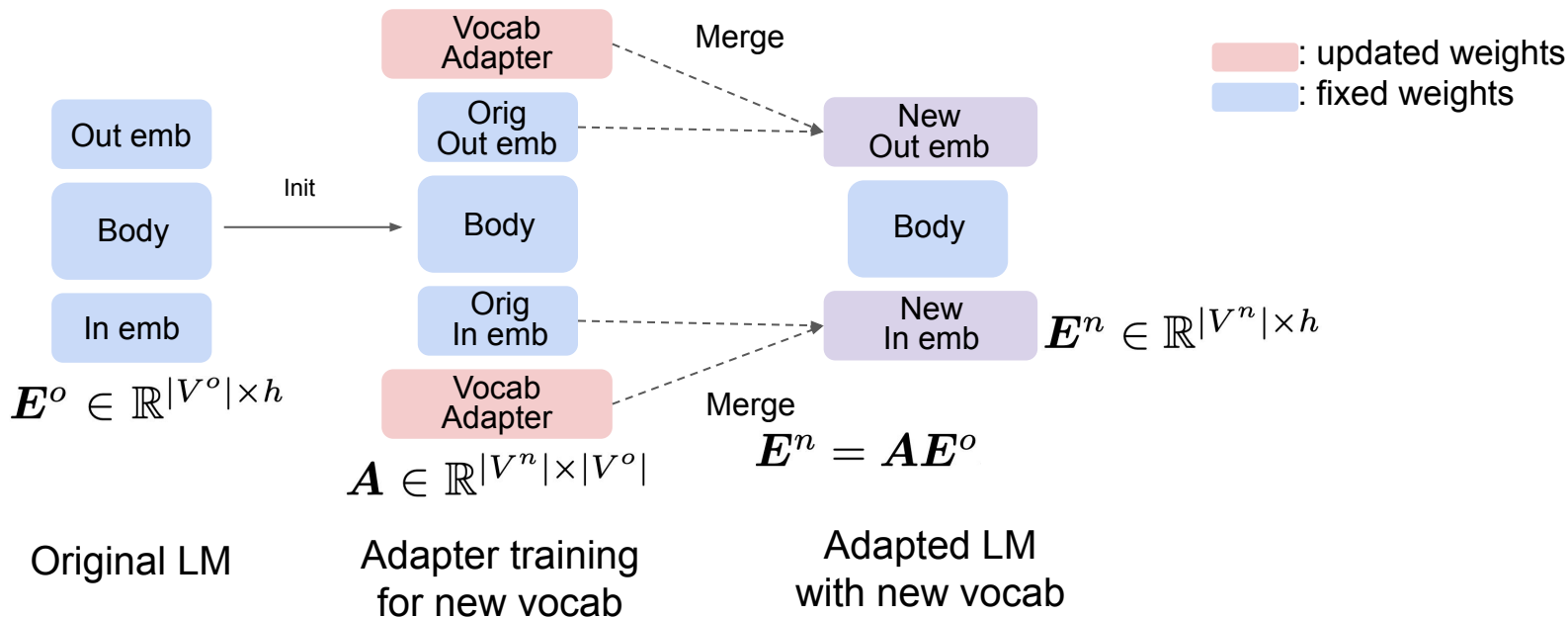| Vocabulary Adaptation	# Langs	Generative Task
ZeTT (Minixhofer et al., 2024)	26	x
RAMEN (Tran, 2020)	6	x
FVT (Gee et al., 2022)	1 (en)	x
VIPI (Mosin et al., 2023)	1 (en)	x
OFA (Liu et al., 2024)	min 369	x
CLP (Ostendorff & Rehm, 2023)	1 (de)	x
CLP+ (Yamaguchi et al., 2024)	4	summarization

VocADT: Multilingual Vocabulary Adaptation with Adapters

Adapter modules to learn the best combination of the original embeddings without relying on heuristics, external embeddings, or dictionaries.

: updated weights
: fixed weights

$\boldsymbol{E}^o \in \mathbb{R}^{|V^o| \times h}$

Original LM

$\boldsymbol{A} \in \mathbb{R}^{|V^n| \times |V^o|}$

Adapter training
for new vocab

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

7

# VocADT: Multilingual Vocabulary Adaptation with Adapters

Better adaptability with only its embeddings replaced and more flexibility in the number of languages while removing the necessity of external pre-trained resources.

$$E^o \in \mathbb{R}^{|V^o| \times h}$$

$$A \in \mathbb{R}^{|V^n| \times |V^o|}$$

$$E^n \in \mathbb{R}^{|V^n| \times h}$$

$$E^n = AE^o$$

Original LM — Adapter training for new vocab — Adapted LM with new vocab

: updated weights
: fixed weights

Initialization Scheme for the Vocabulary Adapter

Effective initialization of the new embedding is crucial in adapting to a new vocabulary.
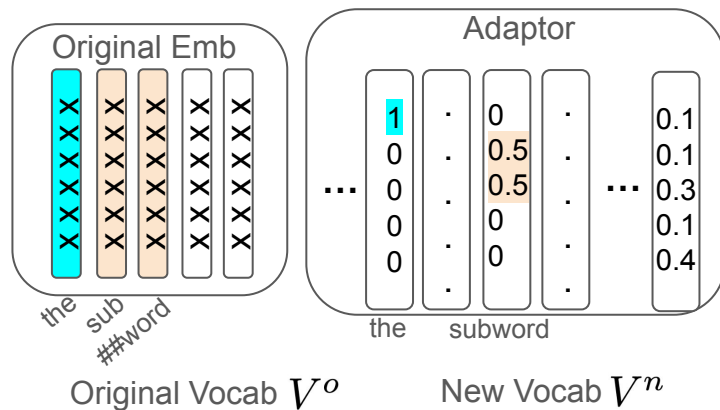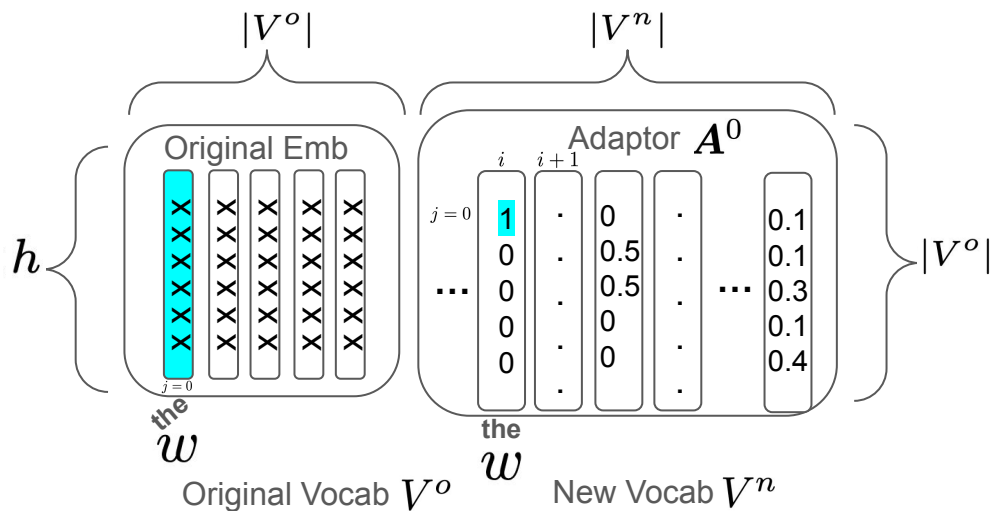

a tokenizer associated with a vocabulary V^x

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

$i = \mathcal{I}^n(w)$ 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

9

Initialization Scheme for the Vocabulary Adapter (1)

1. Copying the original embeddings of overlapping tokens

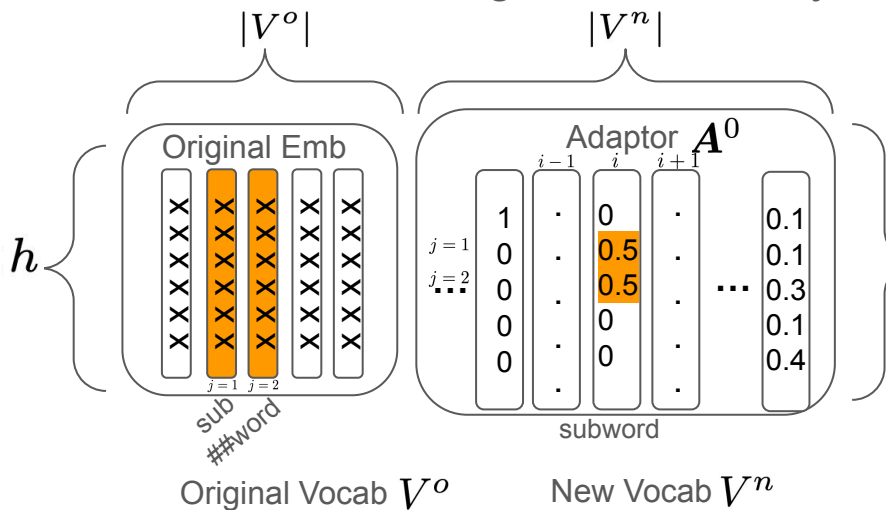$$A^0_{i,\mathcal{I}^o(w)} = 1, \quad A_{i,j} = 0 \quad \forall j \neq \mathcal{I}^o(w)$$

$$\text{where } w \in V^o \cap V^n$$

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

$\mathcal{I}^x : V^x \to \mathbb{Z}$ 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

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

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

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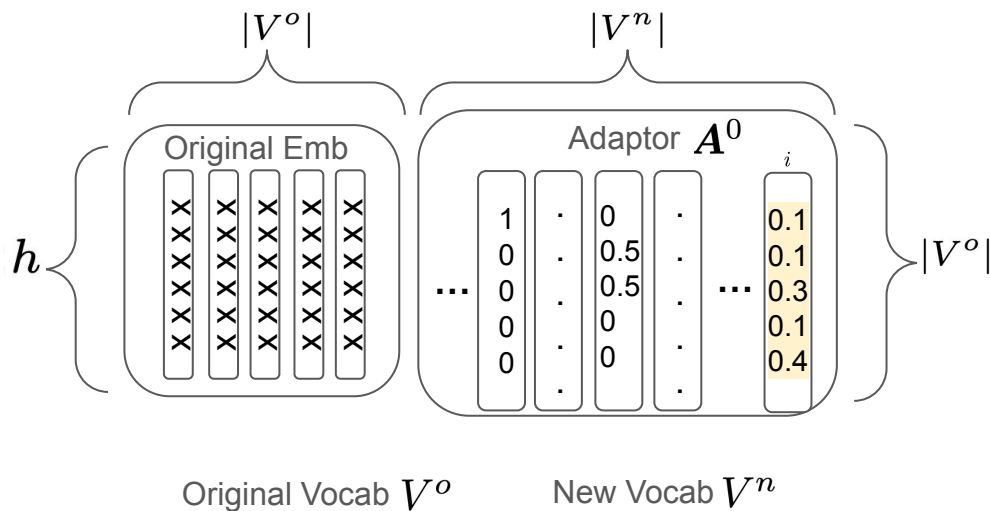$$w \in V^n \backslash (V^o \cap V^n) \quad \text{and}$$
$$w \in S = \{w \mid \mathcal{T}^o(w) = \{t_{1:m}\} \subset V^o\}$$

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

$\mathcal{I}^x : V^x \to \mathbb{Z}$ 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.

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

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

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

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

When and how should we perform vocabulary adaptation?

Key questions to understand the effectiveness and behavior of vocabulary adaptation:

1. Which languages benefit the most from vocabulary adaptation?

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

3. How does vocabulary adaptation impact machine translation?

Experiment Design

Covering 11 languages—with diverse scripts, resource availability, and fragmentation

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 (idx 2–5)

Mixed group (idx 6–9)

Cyrillic group (idx 8–11)

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)

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.

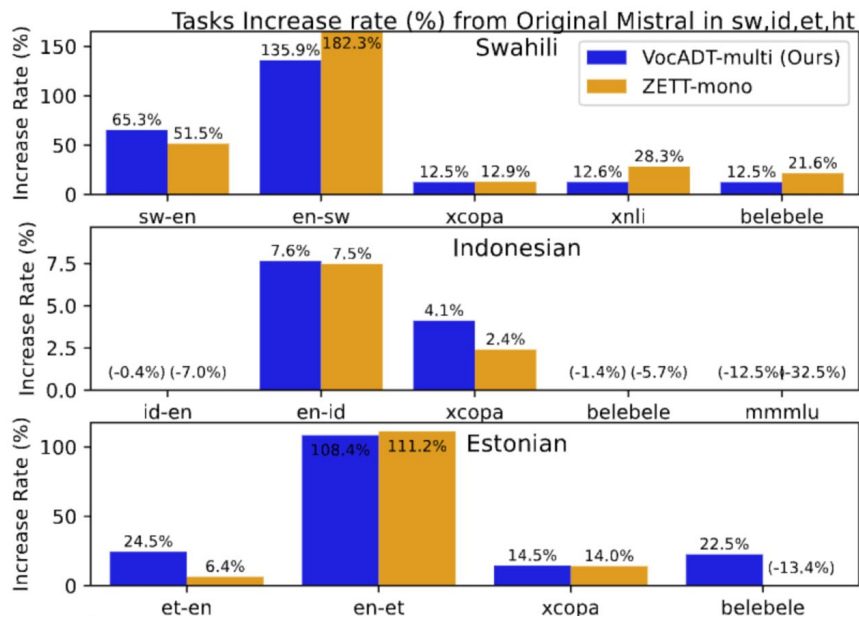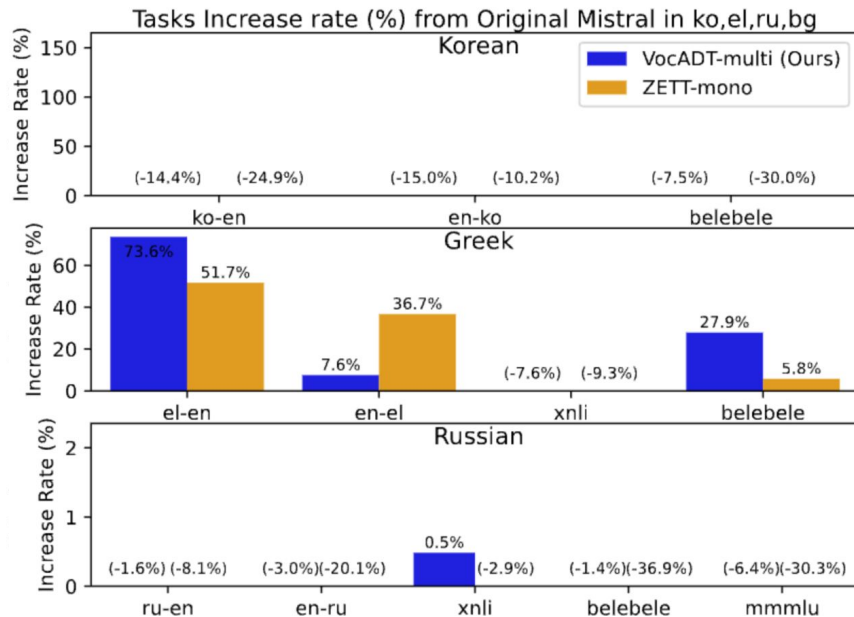Average Performance of multilingual MT/Reasoning/QA tasks
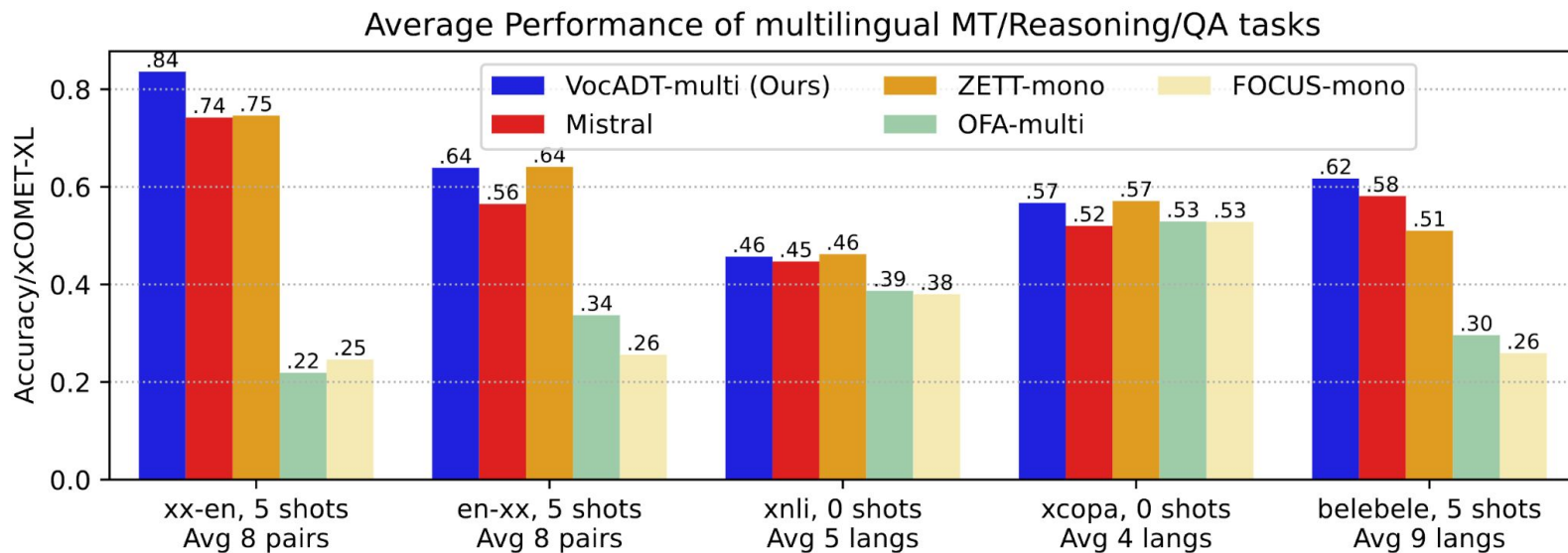
Which Languages Benefit the Most from Vocab Adaptation?

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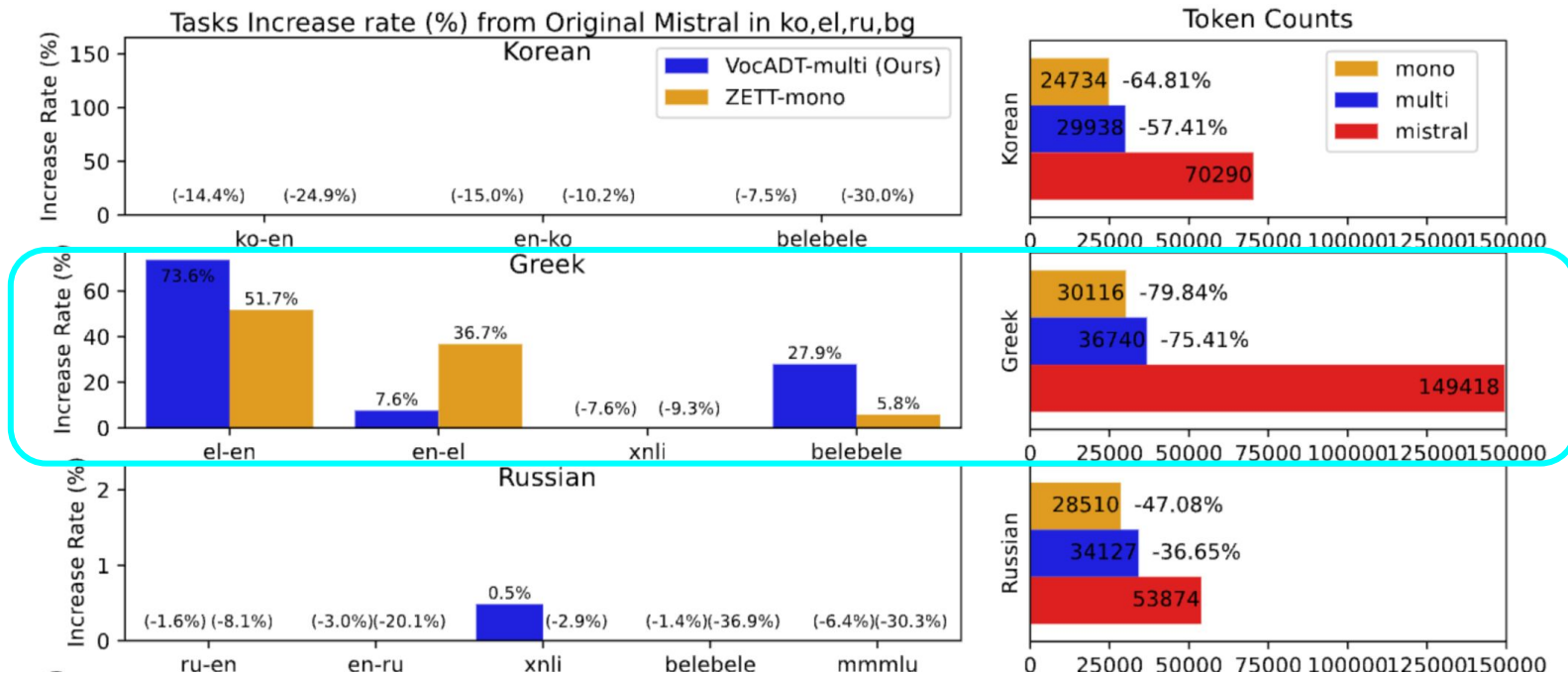

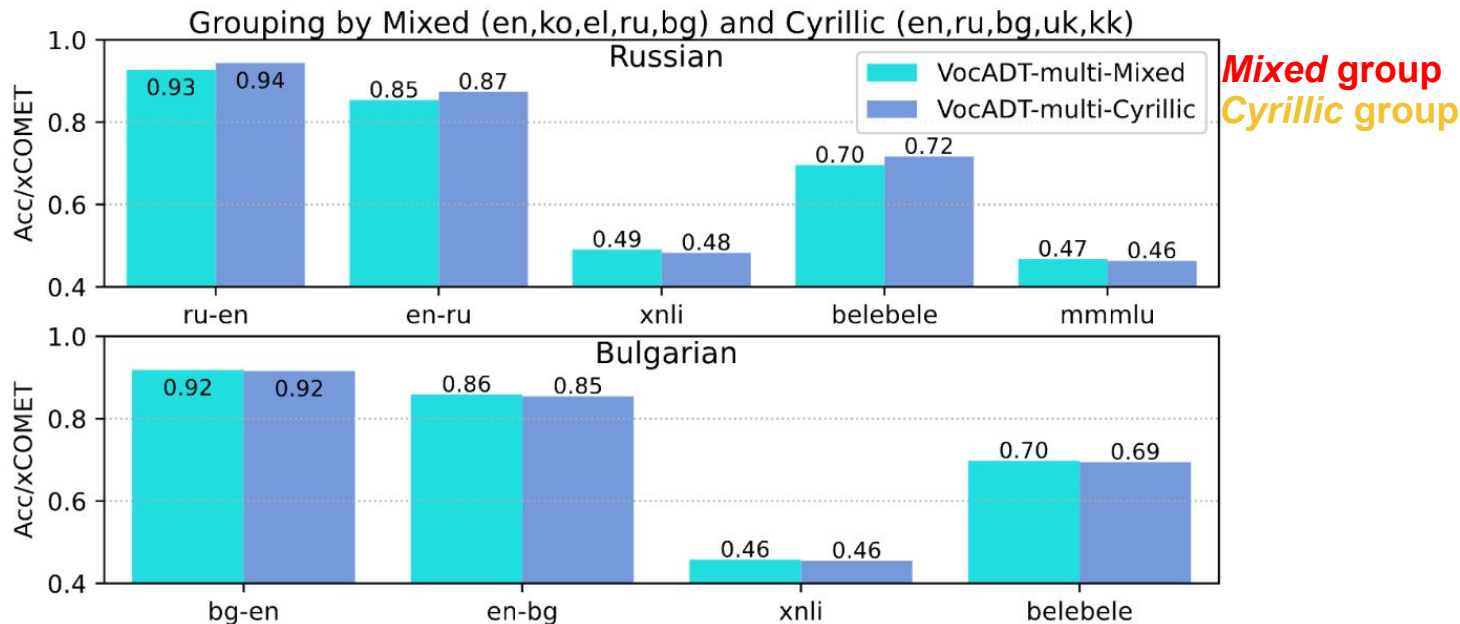**Latin group**

Mixed group

Which Languages Benefit the Most from Vocab Adaptation?

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

Does Script Matter for Language grouping?

Maintaining a **consistent script** within a group enhances performance, though outcomes tend to be influenced more by the language type itself than by the grouping strategy.

Grouping by Mixed (en,ko,el,ru,bg) and Cyrillic (en,ru,bg,uk,kk)

Mixed group
Cyrillic group

DEPARTMENT OF COMPUTER SCIENCE

Microsoft

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

ICLR

HyoJung Han   **Akiko I. Eriguchi**   **Haoran Xu**

Hieu Hoang   **Marine Carpuat**   **Huda Khayrallah**

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