University of Sheffield

Cross-lingual Vocabulary Adaptation for Large Language Models

April 14th, 2025 @ Glasgow

Atsuki Yamaguchi

In this talk…

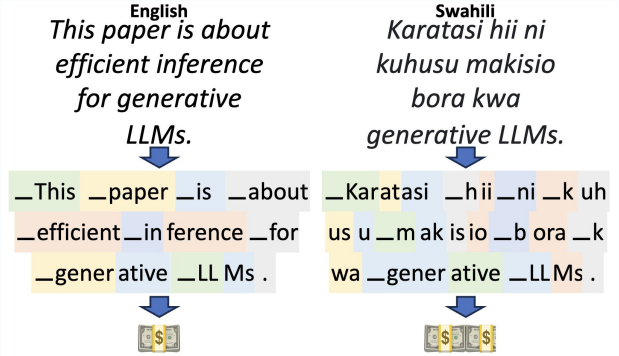
1. Give overview of cross-lingual vocabulary adaptation (CVA) (15 min.)
2. Introduce our work on CVA (10 min.)
3. Discuss challenges in CVA (5 min.)

What is cross-lingual vocabulary adaptation?

- Cross-lingual vocabulary adaptation (**CVA**) is efficient method for cross-lingual transfer
  - Vocabulary of source model is firstly updated (or replaced) with tokens from target language
  - Typically followed by continual pre-training on target language data

Why do we need CVA in the LLM era?

- LLMs show degraded performance in non-English languages
  (Ahia et al., EMNLP 2023; Petrov et al., NeurIPS 2023)
  - Key reason: Not included or underrepresented in training data
- Non-English languages require more inference steps than English – known as *overfragmentation*
  - LLM tokenisers do not have enough multilingual vocabularies



13 tokens ⟪ 23 tokens

CVA can help!

🏅 Benefits of CVA

- **Original motivation** –
  More resource-efficient than training from scratch (Tran, 2019)
- **Other known benefits**
  - Better downstream performance than baselines
    (i.e. source & adapted models w/o CVA) (Dobler & de Melo, EMNLP 2023)
  - Faster inference (Ours: EMNLP Findings 2024)

(Tran, 2019) From English To Foreign Languages: Transferring Pre-trained Language Models
(Dobler & de Melo, EMNLP 2023) FOCUS: Effective Embedding Initialization for Monolingual Specialization of Multilingual Models
(Yamaguchi et al., EMNLP Findings 2024) An Empirical Study on Cross-lingual Vocabulary Adaptation for Efficient Language Model Inference

Different flavours of CVA

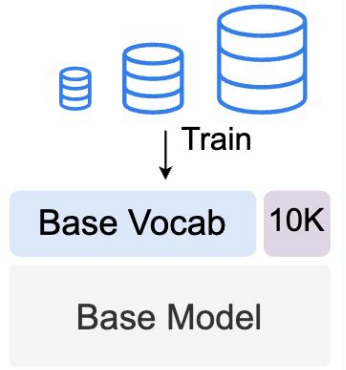
Many variations in considering new target vocabulary:

- **Vocabulary expansion** – Expand source vocabulary
  - Most popular approach for decoder-only LLMs
- **Vocabulary replacement** – Replace source vocabulary
  - Used to be popular for encoder-based models.
- **Other recent flavours**:
  - Hypernetwork (Minixhofer et al., NeurIPS 2024)
  - Embedding/LM head adapters (Han et al., ICLR 2025)

(Minixhofer et al., NeurIPS 2024) Zero-Shot Tokenizer Transfer
(Han et al., ICLR 2025) Adapters for Altering LLM Vocabularies: What Languages Benefit the Most?

– Vocabulary expansion

- Add fixed number of new target language tokens to source vocabulary
- New embeddings are typically randomly initialised in earlier work (Wang et al., EMNLP Findings 2020; Chau et al., EMNLP Findings 2020)
  - Recent studies employ more sophisticated initialisation methods for efficient adaptation
    (Fujii et al., COLM 2024; Mundra et al., CoNLL 2024; Tejaswi et al., EMNLP Findings 2024)
- Followed by continual pre-training
- Quite popular for LLM adaptation
  – Chinese (Cui et al., 2023), Japanese (Fujii et al., COLM 2024), Korean (Choi et al., LREC-COLING 2024), Portuguese (Larcher et al., 2023), etc.



Taken from Tejaswi et al. (2024)

(Wang et al., EMNLP Findings 2020) Extending Multilingual BERT to Low-Resource Languages
(Chau et al., EMNLP Findings 2020) Parsing with Multilingual BERT, a Small Corpus, and a Small Treebank
(Mundra et al., CoNLL 2024) An Empirical Comparison of Vocabulary Expansion and Initialization Approaches For Language Models
(Tejaswi et al., EMNLP Findings 2024) Exploring Design Choices for Building Language-Specific LLMs
(Cui et al., 2023) Efficient and Effective Text Encoding for Chinese LLaMA and Alpaca
(Choi et al., LREC-COLING 2024) Optimizing Language Augmentation for Multilingual Large Language Models: A Case Study on Korean
(Larcher et al., 2023) Cabrita: closing the gap for foreign languages

– Vocabulary replacement

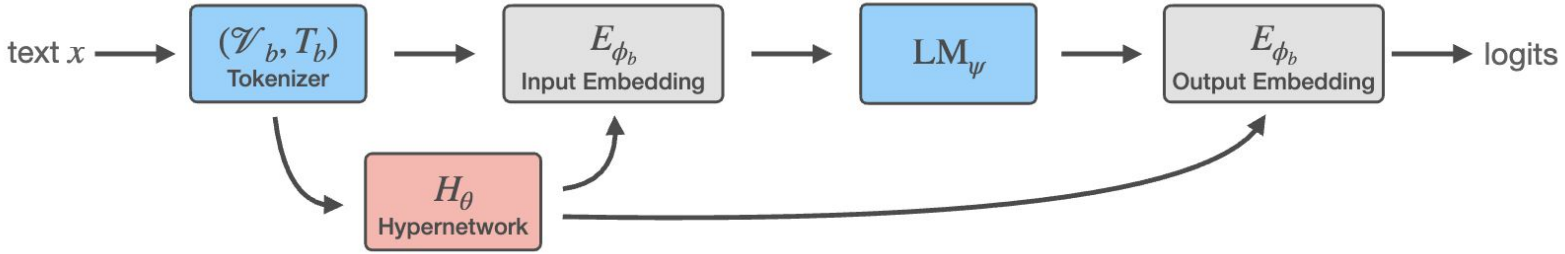
- Replace the entire (or partial) source vocabulary with target one
- Target embeddings are **NOT** usually randomly initialised
  - Semantic similarity-based initialisation: WECHSEL (Minixhofer et al., NAACL 2022), FOCUS (Dobler & de Melo, EMNLP 2023), etc.
  - Heuristic-based initialisation: Downey et al. (MRL 2023)
- Followed by continual pre-training
- Often used for encoder-based models but not so popular for decoder-based models
  - Dagan et al. (ICML 2024) – More than 50B tokens are necessary to swap tokeniser for successful adaptation 🤔

(Minixhofer et al., NAACL 2022) WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models
(Downey et al., MRL 2023) Embedding Structure Matters: Comparing Methods to Adapt Multilingual Vocabularies to New Languages
(Dagan et al., ICML 2024) Getting the most out of your tokenizer for pre-training and domain adaptation

– **Hypernetwork** (Minixhofer et al., NeurIPS 2024)

1. Training mapping function (*hypernetwork*), while keeping source embeddings/LM head unchanged
   - Hypernetwork can generate embeddings/LM head weights corresponding to new target token
2. Followed by continual pre-training

- Reportedly perform better than vocabulary replacement



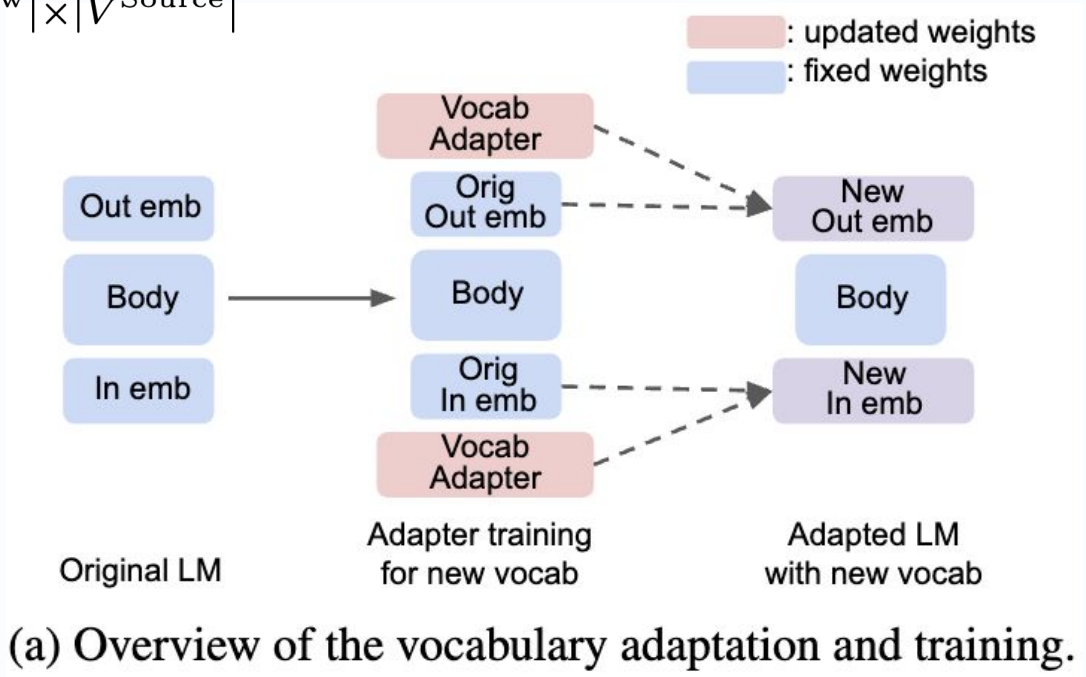
Taken from Minixhofer et al. (2024)

– **Adapters** (Han et al., ICLR 2025)

1. Train an adapter $\mathbf{A} \in \mathbb{R}^{|V^{\text{New}}| \times |V^{\text{Source}}|}$
2. Get new embeddings

$$\mathbf{E}^{\text{New}} = \mathbf{A}\mathbf{E}^{\text{Source}}$$

- Adapter is initialised using mean initialisation (i.e. assign the average of the corresponding source embeddings)
- Performs better than hypernetwork



(a) Overview of the vocabulary adaptation and training.

Taken from Han et al. (2025)

Data requirement in CVA

Question: How much target language data do we need for CVA?

Answer: *It varies. But people usually start with 500M tokens.*

Approach	Suggested/used data size	Target language(s)
Vocab expansion	Order of hundreds of millions of tokens (e.g. 500M tokens) – Tejaswi et al. (2024)	Typically a single language
Vocab replacement	50B tokens? – Dagan et al. (2024)	Typically a single language
Hypernetwork	800M tokens	Up to 26 languages
Adapter	500M tokens per language	en + 4 other languages

(Tejaswi et al., EMNLP Findings 2024) Exploring Design Choices for Building Language-Specific LLMs
(Dagan et al., ICML 2024) Getting the most out of your tokenizer for pre-training and domain adaptation

🚀 Other recent advances in CVA

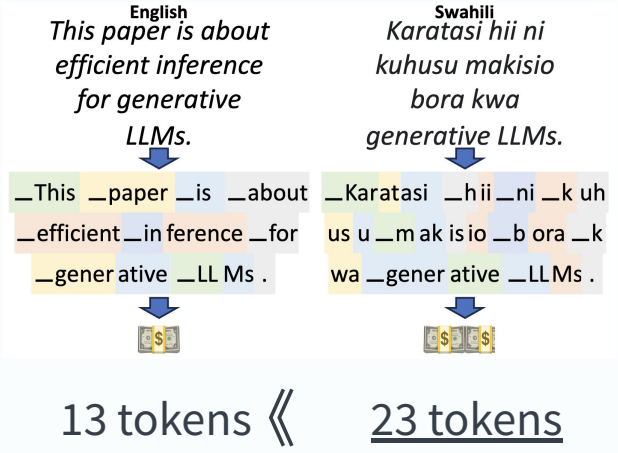
- Multiple language adaptation (> 500 languages)
  (Liu et al., NAACL Findings 2024)
  - Most studies target single or a few languages
- **Ours**:
  - CVA for inference speedups [1]
  - CVA for extremely low-resource settings (< 10M tokens) [2]
  - CVA for chat models w/o chat data [3]

(Liu et al., NAACL Findings 2024) OFA: A Framework of Initializing Unseen Subword Embeddings for Efficient Large-scale Multilingual Continued Pretraining
[1] An Empirical Study on Cross-lingual Vocabulary Adaptation for Efficient Language Model Inference (2024)
[2] How Can We Effectively Expand the Vocabulary of LLMs with 0.01GB of Target Language Text? (2024)
[3] ElChat: Adapting Chat Language Models Using Only Target Unlabeled Language Data (2024)

CVA for inference speedups (EMNLP Findings'24)

2
Our work

- **Problem**: LLMs' over-fragmentation in non-English text
- **Hypothesis**: CVA (i.e. *vocabulary replacement*) can improve LLM inference efficiency in a target language
- **Questions**:
  1. How much does CVA improve inference efficiency?
  2. Does CVA cause task performance degradation?



13 tokens ≪ 23 tokens

CVA for inference speedups (EMNLP Findings'24)

2
Our work

1. How much does CVA improve inference efficiency?

Answer: Up to 272% average speedups 🚀

Key factors on speedups:

- **Task type**: Discriminative tasks see less speedup benefits
- **Pre-training data/vocabulary**:
  How much language-specific data is included in pre-training data

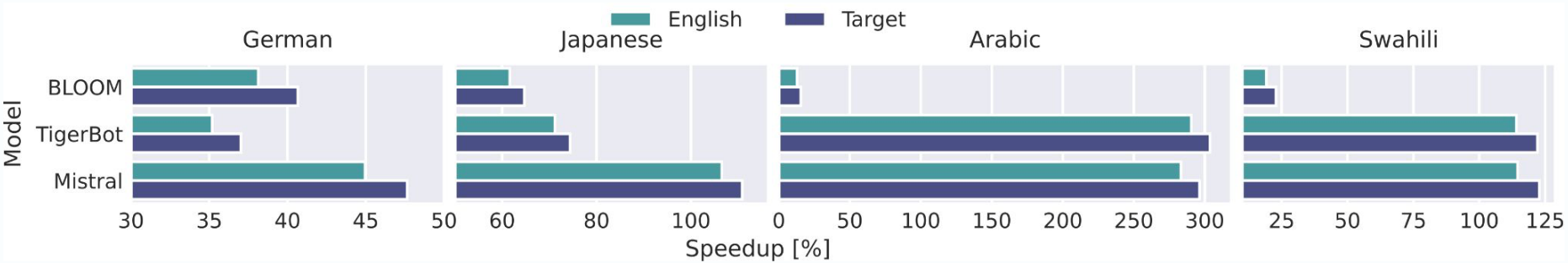


Figure: Inference speedups in summarisation tasks.

CVA for inference speedups (EMNLP Findings'24)

2
Our work

1. How much does CVA improve inference efficiency?

Answer: Up to 272% average speedups 🚀

Key factors on speedups:

- **Task type**: Discriminative tasks see less speedup benefits
- **Pre-training data/vocabulary**:
  How much language-specific data is included in pre-training data

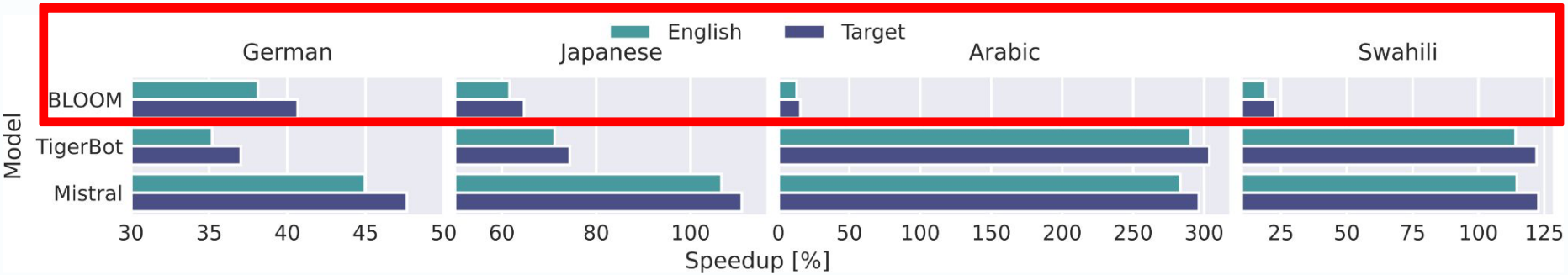


Figure: Inference speedups in summarisation tasks.

CVA for inference speedups (EMNLP Findings'24)

2
Our work

2. Does CVA cause task performance degradation?

Answer: CVA's task performance effect differs by model.

- Multilingual models show more stable performance
- English-centric models more often suffer from degradation

* **Higher is better**

Model	**German**		**Japanese**		**Arabic**		**Swahili**	
	SUM	SPAN	SUM	SPAN	SUM	SPAN	SUM	SPAN
BLOOM 7B	23.1	15	19.0	33	11.5	25	14.3	18
CPT only	19.7	14	21.6	36	11.5	21	13.0	14
CVA (Heuristics)	18.7	21	19.5	38	10.7	21	11.6	16
Mistral	24.1	35	23.7	60	11.2	21	15.4	7
CPT only	24.2	28	23.4	60	10.8	14	16.2	12
CVA (Heuristics)	21.2	22	19.7	43	10.7	13	10.6	14

CVA for inference speedups (EMNLP Findings'24)

2. *Does CVA cause task performance degradation?*

Answer: CVA's task performance effect differs by model.

- Multilingual models show more stable performance
- English-centric models more often suffer from degradation

* **Higher is better**

Model	**German**		**Japanese**		**Arabic**		**Swahili**	
	SUM	SPAN	SUM	SPAN	SUM	SPAN	SUM	SPAN
BLOOM 7B	23.1	15	19.0	33	11.5	25	14.3	18
CPT only	19.7	14	21.6	36	11.5	21	13.0	14
CVA (Heuristics)	18.7	21	19.5	38	10.7	21	11.6	16
Mistral	24.1	35	23.7	60	11.2	21	15.4	7
CPT only	24.2	28	23.4	60	10.8	14	16.2	12
CVA (Heuristics)	21.2	22	19.7	43	10.7	13	10.6	14

CVA for inference speedups (EMNLP Findings'24)

2
Our work

2. *Does CVA cause task performance degradation?*

Answer: CVA's task performance effect differs by model.

- Multilingual models show more stable performance
- English-centric models more often suffer from degradation

* **Higher is better**

Model	**German**		**Japanese**		**Arabic**		**Swahili**	
	SUM	SPAN	SUM	SPAN	SUM	SPAN	SUM	SPAN
BLOOM 7B	23.1	15	19.0	33	11.5	25	14.3	18
CPT only	19.7	14	21.6	36	11.5	21	13.0	14
CVA (Heuristics)	18.7	21	19.5	38	10.7	21	11.6	16
Mistral	24.1	35	23.7	60	11.2	21	15.4	7
CPT only	24.2	28	23.4	60	10.8	14	16.2	12
CVA (Heuristics)	21.2	22	19.7	43	10.7	13	10.6	14

CVA for inference speedups (EMNLP Findings'24)

2. *Does CVA cause task performance degradation?*

Answer: CVA's task performance effect differs by model.

- Multilingual models show more stable performance
- English-centric models more often suffer from degradation

* **Higher is better**

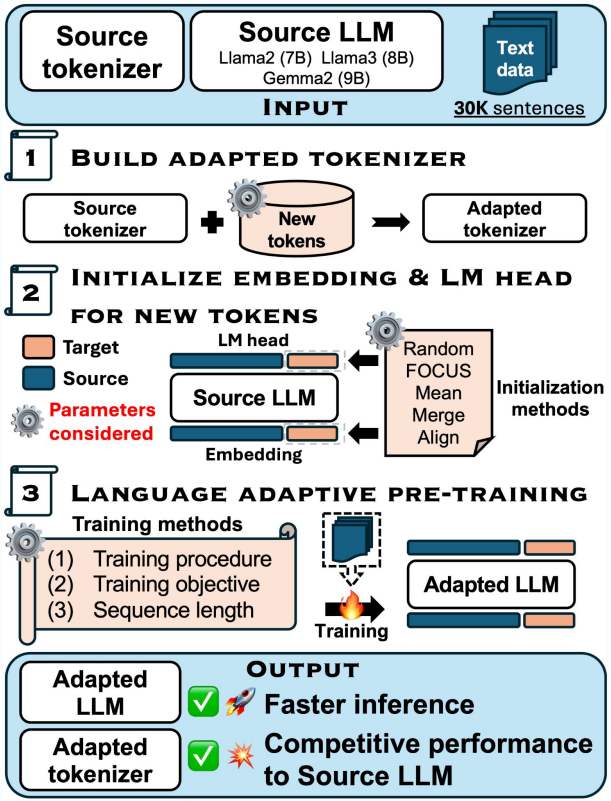
Model	German		Japanese		Arabic		Swahili	
CPT only	24.2	28	23.4	60	10.8	14	16.2	12
CVA (Heuristics)	21.2	22	19.7	43	10.7	13	10.6	14

⚠️ **Given recent models are multilingual,**
CVA will be less likely to hurt target language performance.

CVA for low-resource settings (Under review)

2
Our work

- **Problem**: Vocabulary expansion (VE) mainly tested under high-resource settings (e.g. 10+GB text)
- **Questions**:
  1. Does commonly used VE approach work under low-resource settings?
  2. What is the best possible training strategy for low-resource settings?



CVA for low-resource settings (Under review)

2
Our work

1. **Does conventional VE approach work in low-resource?**

   Answer: Yes. But not optimal.
   Reason: Underfitting

Conventional approaches

Model	ar	my	de	el	hi	ja	si	sw	te	th
Source	8.3	5.0	35.7	4.8	6.4	20.4	3.5	47.2	2.4	9.4
LAPT	4.2	2.7	10.9	2.9	3.2	5.4	2.3	12.7	1.8	4.3
Random	11.9	11.9	12.0	7.1	8.3	15.0	8.7	13.9	7.1	8.8
FOCUS	11.5	13.8	12.1	6.6	8.9	14.8	9.3	13.7	7.9	9.4
Mean	9.4	11.6	**11.7**	6.0	6.4	**14.7**	8.1	**13.5**	6.9	7.8
Merge	9.8	12.6	11.8	6.1	6.7	15.1	8.7	13.7	7.3	8.0
Align	**9.3**	**11.2**	**11.7**	**5.9**	**6.3**	15.1	**8.0**	**13.5**	**6.7**	**7.6**

Table: Perplexity on held-out target language data

CVA for low-resource settings (Under review)

2. What is the best possible training strategy for low-resource settings?

Best recipe

Procedure: Top and bottom two layers (Remy et al., COLM 2024)
Learning objective: Multi-token prediction (Gloeckle et al., ICML 2024)
Sequence length: 512 (shorter)

Up to 12-point gain!

Higher is better

Model	Burmese		Sinhala		Telugu	
	MT	SUM	MT	SUM	MT	SUM
Source	3	25	5	26	6	21
CPT only	11	24	7	27	9	29
LoRA +CLM +2048	4	15	8	30	6	27
2x2LS +MTP +512	**16**	**26**	**13**	**33**	**12**	**29**

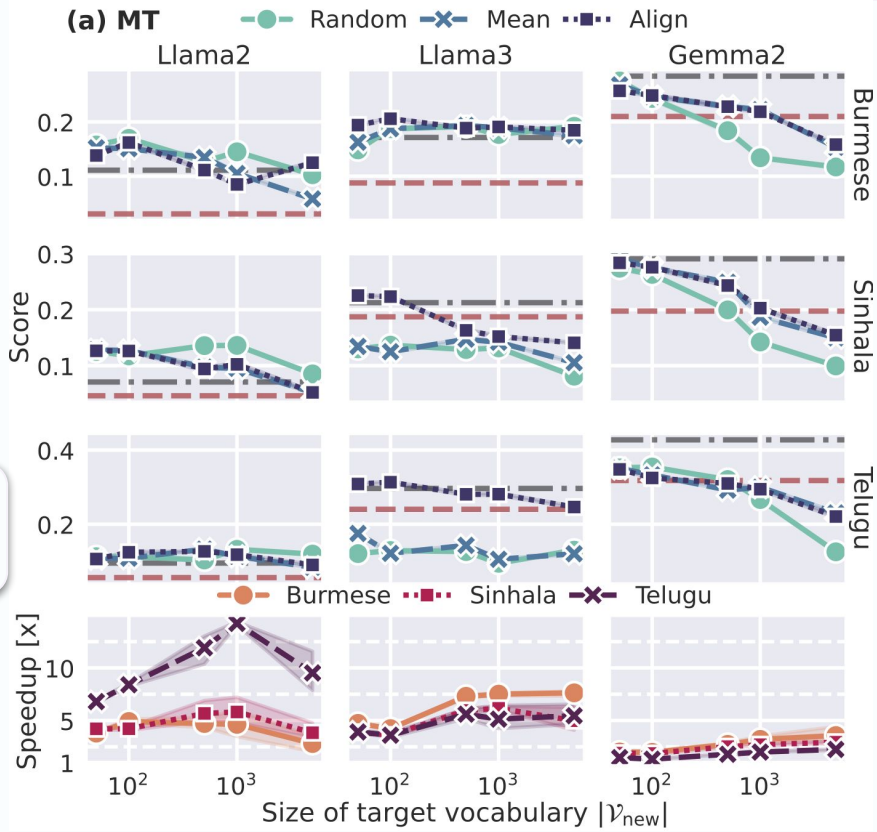
Table: Task performance of Llama 2-based models

CVA for low-resource settings (Under review)

3. **How about vocabulary size?**

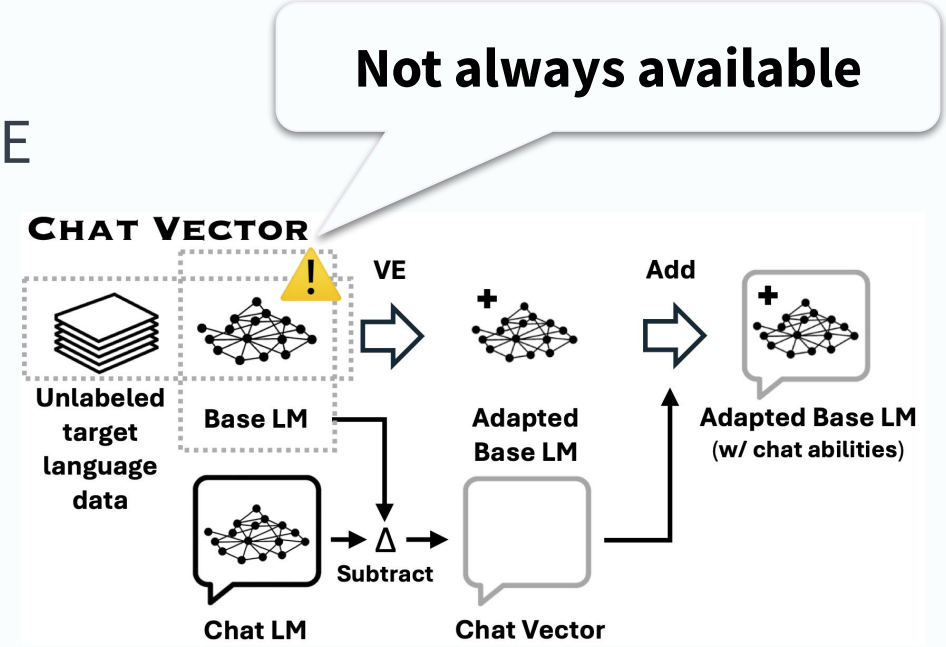
   If too large (e.g. 10K)**:**
   Underfitting (also no substantial speedup gain)

500~1K is good starting point!



CVA for chat models (Under review)

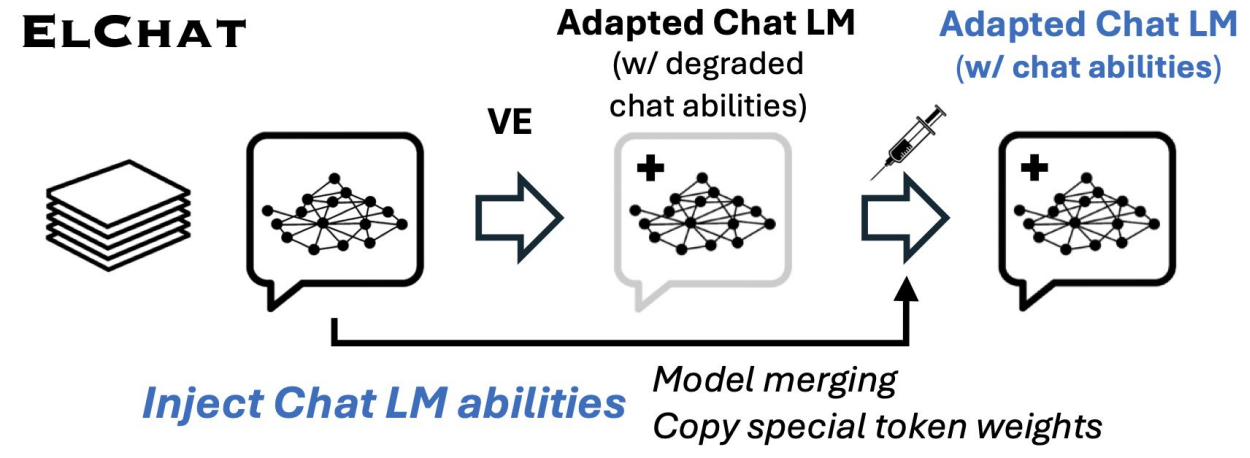
- **Background**: Few has tried to adapt chat models using VE
- **Challenges**:
  - Limited amount of target language chat data
  - Chat Vector is useful but not applicable to models without its base variant (Phi models etc.)



(Huang et al., ACL 2024) Chat Vector: A Simple Approach to Equip LLMs with Instruction Following and Model Alignment in New Languages

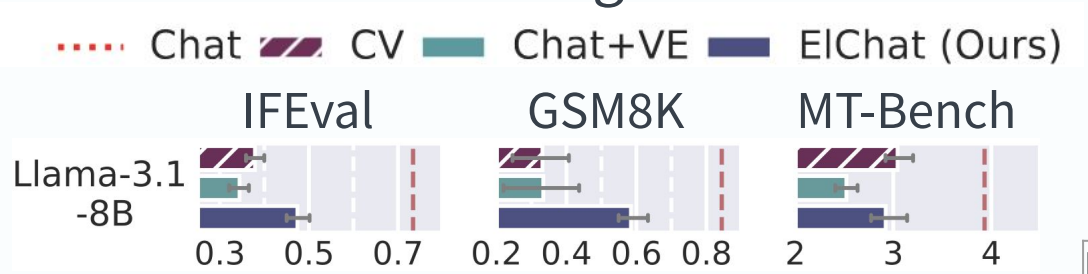
CVA for chat models (Under review)

- **Our proposal: ElChat** (:Eliciting chat capabilities)
  1. Directly adapt Chat model on unlabeled data
  2. Elicit chat capabilities by model merging and weight copy



CVA for chat models (Under review)

- **Chat Vector vs. ElChat** 🏆(Ours)
  Chat and reasoning tasks



Model	MGSM (Multilingual GSM8K)	
	Bengali	Telugu
Chat	30	12
CV	31	24
Chat+VE	26	28
ElChat (Ours)	**51**	**41**

- **Component analysis**
  Merge and copy complement each other

Model	IFEval	GSM8K	MGSM	MT-Bench
ElChat	47	58	46	2.92
No Merge	-13	-18	-9	-0.39
No Copy	-8	-21	-14	-0.27

Remaining key challenges in CVA

3
Challenges

1. Which is the best flavour – ***vocabulary expansion***, ***vocabulary replacement***, ***hypernetwork***, ***vocabulary adapter***, *etc*?
   ⚠️ No conclusive and comparative studies!
2. CVA for fine-tuned (e.g. multi-agent specialised etc.) models
   ⚠️ Limited availability of language-specific fine-tuning datasets.
   ⚠️ What side-effects can arise when applying CVA to fine-tuned models?
3. Necessity of continual pre-training and its drawbacks (e.g. underfitting, catastrophic forgetting, resource-intensive)
   ⚠️ Need to align new tokens well to make CVA work – Unavoidable?

Remaining key challenges in CVA

3
Challenges

1. Which is the best flavour – ***vocabulary expansion***, ***vocabulary replacement***, ***hypernetwork***, ***vocabulary adapter***, *etc*?
   ⚠️ No conclusive and comparative studies!
2. CVA for fine-tuned (e.g. multi-agent specialised etc.) models
   ⚠️ Limited availability of language-specific fine-tuning datasets.
   ⚠️ What side-effects can arise when applying CVA to fine-tuned models?
3. Necessity of continual pre-training and its drawbacks (e.g. underfitting, catastrophic forgetting, resource-intensive)
   ⚠️ Need to align new tokens well to make CVA work – Unavoidable?

1. CVA Arena Project (Work in progress 🚧)

3
Challenges

We are validating the efficacy of different CVA approaches across varying data requirements, language varieties (below), and tasks (both generative & discriminative tasks)!

Language	Language Code	Language Family	Language Script	Region	Joshi's Class
German	de	Indo-European	Latin	Europe 1	5
Japanese	ja	Japonic	Kanji / Hiragana / Katakana	Asia 3	5
Portuguese	pt	Indo-European	Latin	Europe 1	4
Turkish	tr	Turkic	Latin	Asia 1	4
Greek	el	Indo-European	Greek	Europe 1	3
Bengali	bn	Indo-European	Bengali	Asia 2	3
Amharic	am	Afro-Asiatic	Ethiopic	Africa	2
Yoruba	yo	Atlantic-Congo	Latin	Africa	2
Igbo	ig	Atlantic-Congo	Latin	Africa	1
Sinhala	si	Indo-European	Sinhala	Asia 2	0

Remaining key challenges in CVA

3
Challenges

1. Which is the best flavour – ***vocabulary expansion***, ***vocabulary replacement***, ***hypernetwork***, ***vocabulary adapter***, *etc*?
   ⚠️ No conclusive and comparative studies!
2. CVA for fine-tuned (e.g. multi-agent specialised etc.) models
   ⚠️ Limited availability of language-specific fine-tuning datasets.
   ⚠️ What side-effects can arise when applying CVA to fine-tuned models?
3. Necessity of continual pre-training and its drawbacks (e.g. underfitting, catastrophic forgetting, resource-intensive)
   ⚠️ Need to align new tokens well to make CVA work – Unavoidable?

Remaining key challenges in CVA

3
Challenges

1. Which is the best flavour – ***vocabulary expansion***, ***vocabulary replacement***, ***hypernetwork***, ***vocabulary adapter***, *etc*?
   ⚠️ No conclusive and comparative studies!
2. CVA for fine-tuned (e.g. multi-agent specialised etc.) models
   ⚠️ Limited availability of language-specific fine-tuning datasets.
   ⚠️ What side-effects can arise when applying CVA to fine-tuned models?
3. Necessity of continual pre-training and its drawbacks (e.g. underfitting, catastrophic forgetting, resource-intensive)
   ⚠️ Need to align new tokens well to make CVA work – Unavoidable?

3. Beyond continual pre-training

3
Challenges

- Latest work (Feher et al., 2024) has tried to adapt LM in zero-shot (using off-the-shelf hypernetwork)
  - One step forward to CVA without continual pre-training
- Pros: No training required
- Cons:
  - Only applicable to pre-filling stage (i.e. prompt)
  - Well-trained hypernetwork is required

Taken from Feher et al. (2024)

Language	Original Subword Tokenization	#tokens
English	A sub/stantial im/prove/ment fosters further im/prove/ment/s	12
Swahili	U/bor/esh/aj/i mk/ub/wa una/ku/za u/bor/esh/aj/i za/idi	18
#merges	**Dynamic Tokenization**	**#tokens**
1	A sub/stantial *improve*/ment fosters further *improve*/ment/s	10 (83%)
1	U/*boresh*/aj/i m /wa una/ku/za u/*boresh*/aj/i za/idi	16 (89%)

Find the most frequent pair and merge

Merged token weights can be obtained with hypernetwork

(Feher et al., 2024) Retrofitting Large Language Models with Dynamic Tokenization

Summary

- CVA can enable (1) efficient transfer; (2) faster inference in target language(s)
- Increasing number of CVA papers published for the last two years - To improve inference efficiency?
- Many remaining challenges to address!

Thank you!

𝕏 (Twitter): @_gucciiiii
✉️: ayamaguchi1@sheffield.ac.uk