Adapting Chat Language Models Using Only Target Unlabeled Language Data

Anonymous authors
Paper under double-blind review

Abstract

Vocabulary expansion (VE) is the de-facto approach to language adaptation of large language models (LLMs) by adding new tokens and continuing pre-training on target data. While this is effective for base models trained on unlabeled data, it poses challenges for *chat* models trained to follow instructions through labeled conversation data. Directly adapting the latter with VE on target unlabeled data may result in forgetting chat abilities. While ideal, target chat data is often unavailable or costly to create for low-resource languages, and machine-translated alternatives are not always effective. To address this issue, previous work proposed using a base and chat model from the same family. This method first adapts the base LLM with VE on target unlabeled data and then converts it to a chat model by adding a chat vector (CV) derived from the weight difference between the source base and chat models. We propose ElChat, a new language adaptation method for chat LLMs that adapts a chat model directly on target unlabeled data, *without a base model*. It elicits chat abilities by injecting information from the source chat model. ElChat offers more robust and competitive target language and safety performance while achieving superior English, chat, and instruction-following abilities compared to CV.[1]

1 Introduction

Vocabulary expansion (VE) is the de-facto approach to adapting large language models (LLMs) in a target language (Cui et al., 2024; Fujii et al., 2024; Choi et al., 2024). It typically consists of two main steps: (i) new tokens are added to the model vocabulary by expanding the input embedding and output head matrices; and (ii) continual pre-training on target data to learn the input and output embeddings of the new tokens (Cui et al., 2024; Fujii et al., 2024; Choi et al., 2024; Tejaswi et al., 2024; Mundra et al., 2024, *inter alia*). VE is important because LLMs including chat models often perform poorly in languages underrepresented in the training data (Geng et al., 2025; Huang et al., 2024). Moreover, target language tokenization suffers from overfragmentation due to the heavy reliance on data and vocabulary from particular languages (e.g. English), resulting to more inference steps especially in low-resource languages (Ahia et al., 2023; Petrov et al., 2023; Ali et al., 2024). While the vocabularies of frontier LLMs are often large, e.g. 152K for Qwen2.5 (Yang et al., 2024) and 128K for Llama 3.1 (Dubey et al., 2024), they still suffer from this overfragmentation in underrepresented languages. This means such languages often require substantially more inference steps than their high-resource counterparts, as seen with Amharic, which requires 3.48x more inference steps using the default Qwen2.5 without VE.

While VE is effective for *base* models trained on unlabeled data, its application poses significant challenges when the LLM at hand is a *chat* model trained to follow instructions through labeled conversation data. Ideally, we need access to target chat data to effectively adapt chat models. However, this is often unavailable or costly to create for low-resource languages, including the acquisition of human feedback (Huang et al., 2024). Alternatively, machine-translated chat data are not consistently effective (Tao et al., 2024).

To address this issue, Huang et al. (2024) proposed chat vector (CV), a method to obtain a chat model in the target language with access to target unlabeled data only. CV first adapts the base LLM with VE

[1]Our anonymous code is available as supplementary material.

on target unlabeled data and then converts it to a chat model by adding a chat vector derived from the weight difference between the source base and chat models. However, this requires access to base and chat models from the same family that might not always be available, hindering its applicability. For example, the Phi-3 (Abdin et al., 2024a) and Phi-4 (Abdin et al., 2024b) do not provide base models due to safety reasons.[2] Similarly, Velvet[3], EXAONE-3.5[4], and Trillion (Han et al., 2025) models are available only as a chat model. Crucially, it is completely to the discretion of developers to decide whether they publish both base and chat variants.

In this paper, we propose ElChat, a new language adaptation method for chat LLMs that adapts a chat model directly on target unlabeled data, eliminating the need for a base model (Figure 1). We hypothesize that direct adaptation of a source chat model with VE on target unlabeled data negatively impacts its chat and instruction-following abilities by altering its parametric knowledge. However, we posit that these can still be recovered. For this purpose, ElChat leverages information from the source chat model to elicit chat abilities through two key mechanisms. First, we employ model merging to integrate distinct parametric knowledge from the source and target models (Wortsman et al., 2022; Yadav et al., 2023; Yu et al., 2024; Goddard et al., 2024). We hypothesize that model merging helps restore the chat and instruction-following abilities of the source model while preserving the target language performance achieved by the target model. Second, we reuse the weights of special tokens from the source model. For example, tokens that mark the start of a conversation turn should be crucial for activating the instruction-following ability as they are used to structure raw input into chat format. However, direct adaptation on target unlabeled data may degrade their functionality as they are modified during VE. To mitigate this, we copy these token weights directly from the source model to the target model.

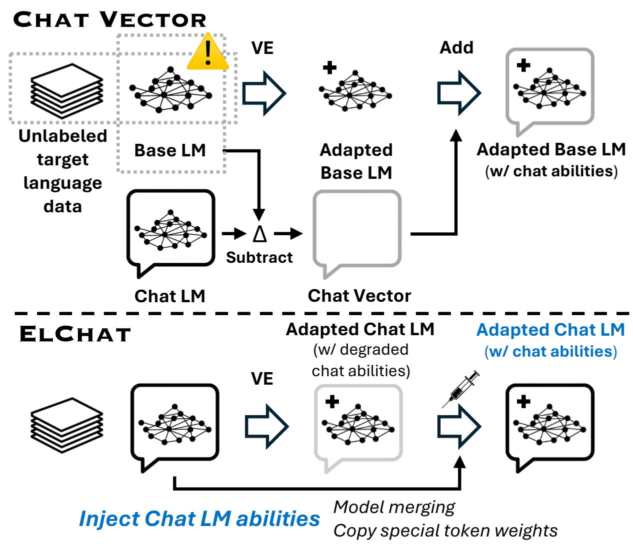


Figure 1: Chat LLM language adaptation with Chat Vector (Huang et al., 2024) and ElChat (ours). Note that a base LM in this paper refers to an LM pre-trained on unlabeled data without any further post-training. A chat LM, on the other hand, is a base model that has been further supervised fine-tuned on labeled conversational data, enabling it to follow instructions.

We investigate the efficacy of ElChat by experimenting with two popular chat models across seven typologically diverse languages. Our evaluation includes safety, chat, and instruction-following performance. Additionally, we also assess target and source language task performance and target language inference speed. Our key contributions are as follows:

- We propose ElChat that adapts a chat model directly on target unlabeled data, eliminating the need for (i) a base model and (ii) target chat data.
- ElChat achieves better chat and instruction-following abilities and source language performance than CV. It is also competitive and more robust (i.e. consistently outperforming the source chat model) in the target language and safety tasks compared to CV (§5.2, §5.3, §5.1).
- Despite model modifications, ElChat achieves comparable target inference speedups across models and tasks, matching the performance of the adapted VE and CV models (§6).

[2]https://huggingface.co/microsoft/phi-4/discussions/4
[3]https://huggingface.co/Almawave/Velvet-14B
[4]https://huggingface.co/collections/LGAI-EXAONE/exaone-35-674d0e1bb3dcd2ab6f39dbb4

2 Related Work

2.1 Cross-lingual Vocabulary Adaptation

The most popular approach to adapting LLMs to a target language is by expanding their vocabulary (VE) with tokens from the target language (Balachandran, 2023; Larcher et al., 2023; Pipatanakul et al., 2023; Lin et al., 2024; Cui et al., 2024; Kim et al., 2024; Fujii et al., 2024; Choi et al., 2024; Nguyen et al., 2024; Tejaswi et al., 2024; Mundra et al., 2024).

Other methods to language adaptation include full or partial vocabulary replacement with a new target vocabulary (Ostendorff & Rehm, 2023; Csaki et al., 2023; Da Dalt et al., 2024; Remy et al., 2024; Yamaguchi et al., 2024a; Dobler & de Melo, 2024; Cahyawijaya et al., 2024), hypernetwork for tokenizer transfer (Minixhofer et al., 2024), and adapters for vocabulary alignment (Han et al., 2024). Our work focuses on VE as it has been widely used recently for mostly base LLM adaptation in languages such as Chinese, Japanese, Korean, and Persian (Cui et al., 2024; Fujii et al., 2024; Choi et al., 2024; Mahdizadeh Sani et al., 2025, *inter alia.*).

2.2 Language Adaptation of Chat Models

Recent work has proposed developing chat models in a target language from source base models. For example, Toraman (2024) and Zhao et al. (2024a) apply VE to *base* models using target language chat data, consisting of 52k samples and 500M tokens, respectively. Bandarkar et al. (2024b) also adapt base models using 30-40k target language chat data. Their approach also adapts a task-specific (i.e. math) model on 200k English math samples, followed by merging the two models to enhance math performance in the target language. Alexandrov et al. (2024) iteratively merge models trained on subsets of available target language data to effectively mitigate catastrophic forgetting. Their method first adapts base models with continual pre-training (CPT) on target unlabeled language data, followed by instruction tuning on target language chat data samples. However, this approach requires substantial target language data (at least 50B tokens of unlabeled data for CPT, and 78K samples of target language chat data) to ensure that each subset contains sufficient information for effective adaptation. A different approach, proposed by Tao et al. (2024), involves merging two base models: (1) one supervised fine-tuned on 162k English data samples, and (2) another trained on at least eight billion tokens of target unlabeled language data. However, it still relies on the availability of a base model. Geng et al. (2025) propose adapting source chat models directly through a multi-stage training approach. This method involves target unlabeled data and transfer fine-tuning (i.e. supervised fine-tuning tasks using translated target chat data.)

The main limitation with this line of work is that it requires access to target chat data (real or translated), typically in large volumes. Chat data is often unavailable or costly to produce for low-resource languages, while machine-translated chat data is not always effective for adaptation (Tao et al., 2024). For example, Burmese, one of our experimental languages, consists only of 472 manually annotated instruction samples in the Aya Dataset (Singh et al., 2024b). This is insufficient for direct application of VE, as its typical data requirements are in the order of millions of tokens (Tejaswi et al., 2024), making these methods not applicable in such settings.

Huang et al. (2024) assumes a language adaptation setting where there is no access to target chat data (real or translated). This is a more realistic scenario for low-resource languages. For example, Burmese has only 172k target unlabeled language data in MADLAD-400 (Kudugunta et al., 2023) available for adaptation. Their proposed CV method obtains a target chat model using a source base and a source chat model. This approach to chat LLM adaptation is the closest to our work.

3 ElChat: Eliciting Chat and Instruction-following Abilities

Similar to Huang et al. (2024), we aim to adapt a chat LLM to a target language, assuming that we only have access to target unlabeled data and no target chat data (real or translated) at all. Unlike Huang et al.

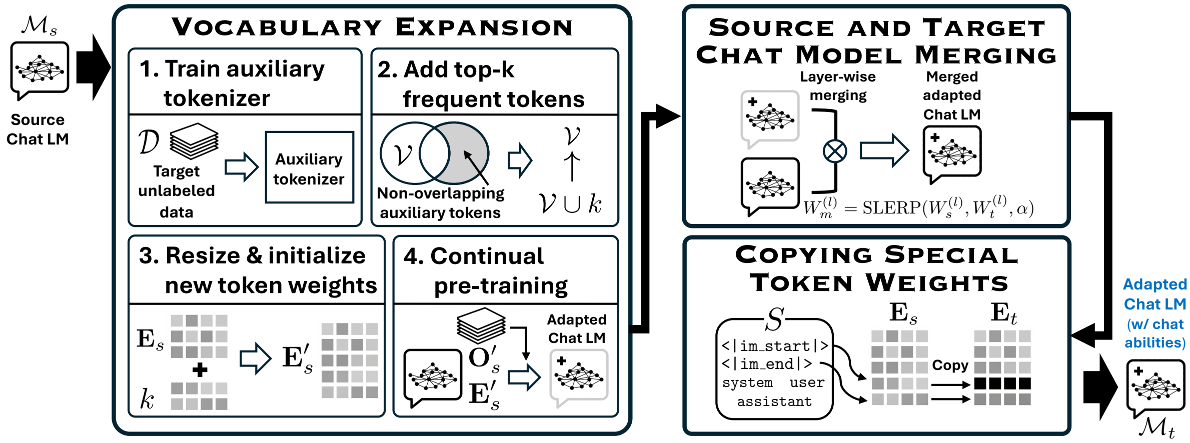


Figure 2: ElChat: A three-step adaptation process overview.

(2024), our goal is to remove the dependence on a base model that might not always be available and adapt a chat model directly on target unlabeled data with VE.

To achieve this, we introduce ElChat that consists of three steps: (i) VE on the source chat model using target unlabeled data to obtain an initial adapted target chat model; (ii) merging the source and target models; (iii) injecting information from the source to the target by copying tokens that are specific to chat and instruction-following capabilities. Figure 2 visualizes these steps. While the individual components of these are established techniques, the key novelty of ElChat resides in its strategic combination to overcome the dependence on the base model.

We expect that modifying the parametric knowledge of the source model by training on target unlabeled data in step (i) will improve its target language skills by updating specific areas of the network but negatively impact its chat and instruction-following abilities due to catastrophic forgetting. To remedy this, we hypothesize that we can elicit the latter through (ii) model merging (Wortsman et al., 2022; Yadav et al., 2023; Yu et al., 2024; Goddard et al., 2024). It allows the integration of distinct parametric knowledge from the source and target models without incurring any additional training costs, i.e. chat abilities from the source and target language knowledge from the target. Finally, in step (3), we restore the representation of special tokens used in the chat template. We assume this will guide the target model in effectively responding to user instructions by transferring this information from the source.

Vocabulary Expansion (VE). Given the sole availability of target unlabeled data, we apply VE on the source chat LLM by expanding its input and output head matrices with new tokens, followed by CPT on the target data. This process follows the standard protocol adopted in Tejaswi et al. (2024) and Mundra et al. (2024) for resizing and initializing these matrices.

More specifically, given target unlabeled data $\mathcal{D}$ and a source chat model $\mathcal{M}_s$ with initial vocabulary $\mathcal{V}$:

1. An auxiliary target language tokenizer is first trained on $\mathcal{D}$. This tokenizer utilizes the same underlying tokenization function (e.g. byte-level BPE in this paper) as $\mathcal{M}_s$.

2. From this auxiliary tokenizer, we then select the top-k most frequent tokens (e.g. $k = 10\text{K}$ by default, as in Tejaswi et al. (2024)) that do not overlap with $\mathcal{V}$ but are present in the auxiliary tokenizer's vocabulary. These selected k tokens are then added to $\mathcal{V}$.

3. To accommodate these k new tokens in $\mathcal{M}_s$, its input $\mathbf{E}_s \in \mathbb{R}^{|\mathcal{V}| \times H}$ and output head $\mathbf{O}_s \in \mathbb{R}^{H \times |\mathcal{V}|}$ matrices are resized. They become $\mathbf{E}'_s \in \mathbb{R}^{(|\mathcal{V}|+k) \times H}$ and $\mathbf{O}'_s \in \mathbb{R}^{H \times (|\mathcal{V}|+k)}$ respectively, where H denotes the hidden dimensionality of $\mathcal{M}_s$. The weights corresponding to the newly added tokens are then initialized using mean initialization (Yao et al., 2021), a popular and simple, yet effective method in VE (Fujii et al., 2024; Tejaswi et al., 2024; Mundra et al., 2024). The weight of each

newly added token is initialized as the average embedding or language modeling head weight of their corresponding source tokens, obtained using the corresponding source tokenizer.

After the above procedure, $\mathcal{M}_s$ undergoes continual pre-training on $\mathcal{D}$ using a causal language modeling objective. A key difference between chat and base models is that the former uses a chat template. This includes specific roles (e.g. user, system, or assistant) and has a placeholder for message text in the prompt (see Appendix A for details). For example, a chat template might structure input as follows:

```
<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
What is the capital of the US?<|im_end|>
<|im_start|>assistant
```

During CPT, we remove the default chat template of the model to support unlabeled data because the unlabeled data typically lacks these explicit role annotations. During inference, we append it to task-specific prompt templates (see Table 3 for task-specific prompts).

Source and Target Chat Model Merging (Merge). After VE, we merge the source and target chat models. We employ a popular merging method: spherical linear interpolation (SLERP) (Goddard et al., 2024) to merge each layer of the source and adapted models.[5] This process excludes the embedding and language modeling head from merging because the source and target models use different vocabularies.

Given the source chat model $\mathcal{M}_s$ and the target-adapted chat model $\mathcal{M}_t$ (which has undergone VE), the weights of each corresponding layer l (e.g. self-attention and feed-forward layers) are merged. For each such layer l, let $\mathbf{W}_s^{(l)}$ denote the weight matrix from $\mathcal{M}_s$, and $\mathbf{W}_t^{(l)}$ denote the corresponding weight matrix from $\mathcal{M}_t$. The merged weight matrix, denoted $\mathbf{W}_m^{(l)}$, is then computed using SLERP:

$$\mathbf{W}_m^{(l)} = \mathrm{SLERP}(\mathbf{W}_s^{(l)}, \mathbf{W}_t^{(l)}, \alpha)$$

where $\alpha \in [0, 1]$ is an interpolation coefficient. This α serves as a hyperparameter controlling the balance between $\mathcal{M}_s$ and $\mathcal{M}_t$ learned representations, where $\alpha = 0$ corresponds to using only the weights of $\mathcal{M}_s$ and $\alpha = 1$ to using only those of $\mathcal{M}_t$.

Copying Special Token and Language Modeling Head Weights (Copy). Special tokens used in a chat template (e.g. <im_start> in Qwen2.5 to represent the start of a turn) should be critical in supporting chat and instruction-following abilities of a model (see Appendix A for a full list of special tokens). Although the embedding and language modeling heads are excluded from merging due to vocabulary differences, leaving them unchanged may not be optimal for eliciting the chat and instruction-following abilities of the adapted chat model. Hence, we copy all the special token weights from the source model to the adapted model.

Specifically, let S be the set of token IDs corresponding to the special tokens defined by the chat template of $\mathcal{M}_s$. The input embedding matrix and the language modeling head of $\mathcal{M}_t$ are updated as follows:

For each special token ID $x \in S$, the embedding vector for token x in $\mathcal{M}_t$, $\mathbf{E}_t \in \mathbb{R}^{(|\mathcal{V}|+k)\times H}$, is updated by copying the corresponding vector from $\mathcal{M}_s$'s input embedding matrix ($\mathbf{E}_s$): $\mathbf{E}_t[x, :] \leftarrow \mathbf{E}_s[x, :]$, where $\mathbf{E}[i, :]$ denotes the embedding vector for token ID i.

Similarly, the language modeling head weights corresponding to token x in $\mathcal{M}_t$, $\mathbf{O}_t \in \mathbb{R}^{H\times(|\mathcal{V}|+k)}$, are updated by copying the corresponding vector from $\mathcal{M}_s$: $\mathbf{O}_t[:, x] \leftarrow \mathbf{O}_s[:, x]$, where $\mathbf{O}[:, j]$ denotes the column of weights for token ID j in the language modeling head.

[5] We also test linear merging (Wortsman et al., 2022) yielding similar results (see Appendix B). This suggests the robustness of our merging approach regardless of the specific interpolation method. Nonetheless, we primarily utilize SLERP to streamline our experiments due to its superior fine-grained control for merging, even though it may not be empirically superior across all tasks (Goddard et al., 2024).

4 Experimental Setup

This section describes the experimental setup in this paper. More details are listed in Appendix A.

4.1 Source Models

We use two popular chat models as source: Qwen2.5 7B (Yang et al., 2024); and Llama 3.1 8B (Dubey et al., 2024), across experiments. Additionally, to ensure consistency in our analysis, we also incorporate the state-of-the-art chat model Qwen3 14B (Yang et al., 2025). The corresponding results and analysis for Qwen3 are presented in Appendix C.3.

4.2 Target Languages and Adaptation Data

We experiment with the following seven typologically diverse languages, assuming that they are likely to be underrepresented compared to English in the pre-training data of the source models, or entirely absent: Amharic (Afroasiatic), Bengali (Indo-European), Burmese (Sino-Tibetan), Gujarati (Indo-European), Sinhala (Indo-European), Tamil (Dravidian), and Telugu (Dravidian). The ratio of training data in each model for each source base and chat model has not been explicitly disclosed (Appendix C.2). Note that we do not consider Latin script target languages as they are less likely to suffer from overfragmentation and usually benefit less from VE (Yamaguchi et al., 2024b; Tejaswi et al., 2024). Due to computational constraints, we use Qwen3 for Amharic, Bengali, and Telugu only.[6]

For the CPT part of VE, we use MADLAD-400 (Kudugunta et al., 2023), which consists of highly-filtered document-level samples sourced from CommonCrawl, and randomly sample 250K language-specific documents for each language as the target unlabeled data.[7]

4.3 Continual Pre-training

Following Remy et al. (2024), we train the embedding, LM head, and the top and bottom two layers of a source model. This approach aims to calibrate only the parts closely related to the encoding and decoding of the target language (Wendler et al., 2024; Tang et al., 2024; Zhao et al., 2024b), minimizing changes to the source model while allowing cost-effective tuning.

4.4 Baselines

We compare ElChat against the following baselines:

- Off-the-shelf base (**Base**) and chat (**Chat**) models without target language adaptation.
- Base and Chat models adapted using standard VE, denoted by **Base+VE** and **Chat+VE** respectively. Note that the latter uses a chat template in inference (see §3).
- **CV** proposed by Huang et al. (2024), augmenting Base+VE with chat vector using Base and Chat.

For reference, we also experiment with adapting Chat and Base using only CPT on the same target language data without VE (i.e. no inference speedup in a target language). We provide the results and analysis of these CPT-only models in Appendix C.1.

4.5 Evaluation Tasks

We evaluate the efficacy of ElChat in safety, chat, and instruction-following performance, and target and source language performance.

[6] Amharic is selected because it shows the most significant speedup gains in target language generative tasks with Qwen2.5 and Llama 3.1 (Tables 17 and 18). Bengali and Telugu are chosen as they are the only languages covered by MGSM.

[7] For languages with less than 250K documents (i.e. Amharic and Burmese), we use the full articles.

Safety, Chat, and Instruction-Following. Following Cahyawijaya et al. (2024), we conduct safety evaluation on target language translated data including TruthfulQA (Lin et al., 2022), ToxicGen (Hartvigsen et al., 2022), and ImplicitHate (ElSherief et al., 2021). We also measure chat and instruction-following abilities in the source language (English) using IFEval (Zhou et al., 2023), GSM8K (Cobbe et al., 2021) as multi-turn few-shot, and MT-Bench (Zheng et al., 2023). Furthermore, we measure the performance on English AlpacaEval v2.0 (Li et al., 2023; Dubois et al., 2024) for additional analysis.

Target language evaluation is challenging for instruction-following and chat tasks due to the limited data availability. LLM-as-a-Judge (Zheng et al., 2023) is also unstable according to Azime et al. (2024) in low-resource languages. Hence, we use multi-turn MGSM (Shi et al., 2023) for target language evaluation as it consists of manually translated high-quality data.

Target Language. We use both generative and discriminative target language tasks. For generative tasks, we use summarization (sum) using XL-SUM (Hasan et al., 2021) and English-to-target machine translation (mt) using FLORES-200 (NLLB Team et al., 2022). For a discriminative task, we employ multiple-choice text classification (mc) using Belebele (Bandarkar et al., 2024a) and Global MMLU (gmmlu) (Singh et al., 2024a) as general target language understanding benchmarks.

Source Language (English). We assess the extent to which the adapted models retain their general task-solving abilities in English sum, target-to-English mt, and English mc using the same datasets as those employed for target languages. We also use mmlu (Hendrycks et al., 2021) as an English language understanding benchmark and English bbh (Srivastava et al., 2023; Suzgun et al., 2023) as a stress-test benchmark.

Following Ahia et al. (2023), we use 500 random samples for generative tasks: sum and mt. The rest use the full test sets for evaluation.

4.6 Evaluation Metrics

Task Performance. We report the standard metrics for each task: accuracy for mc, gmmlu, mmlu, bbh, TruthfulQA, and IFEval (strict prompt), and exact match for GSM8K and MGSM. For MT-Bench, we use the mean score over two turns across all questions. Adhering to the standard protocol in LightEval (Fourrier et al., 2023), each score is determined using `Flow-Judge-v0.1` and follows a Likert-5 scale. For AlpacaEval, we use a win-rate over GPT-4 (1106 Preview) measured by GPT-4.1 nano (2025-04-14).[8] For sum and mt, we primarily use chrF (Popović, 2015).[9] For ToxicGen and ImplicitHate, we use safety score, which is the percentage of likeliness of the model producing benign over harmful sentences, following Cahyawijaya et al. (2024).

We report average zero- and three-shot performance across three different runs for sum and mt, respectively. For the remaining tasks, we report single-run zero-shot performance for IFEval, MT-Bench, ToxicGen and ImplicitHate, three-shot performance for mc, TruthfulQA, five-shot performance for gmmlu, mmlu, GSM8K, and MGSM as these tasks are deterministically evaluated with temperature set to zero.

Inference Efficiency. VE offers inference speedups in a target language compared to source models (Tejaswi et al., 2024; Mundra et al., 2024; Yamaguchi et al., 2024b). To quantify this, we measure the number of tokens generated per second (tokens/s) (Hong et al., 2024).

[8] `https://openai.com/index/gpt-4-1/`

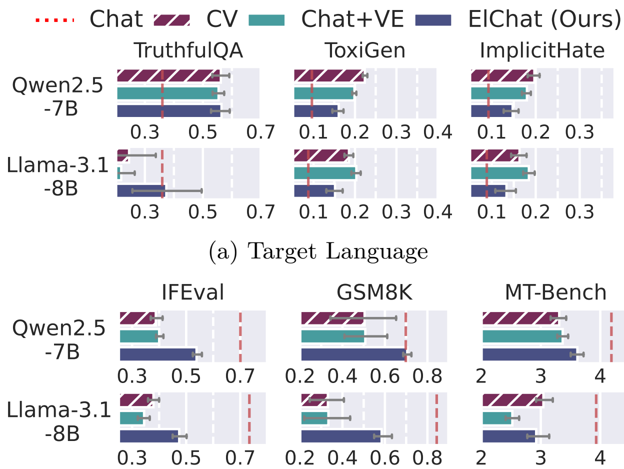
[9] Although chrF has been a widely used metric for sum and mt (Ebrahimi et al., 2023; Remy et al., 2024, *inter alia*), we also show ROUGE-L (Lin, 2004) for sum and chrF++ (Popović, 2017) for mt in Appendix B.

5 Task Performance

5.1 Safety, Chat, and Instruction-following

Safety. Figure 3 (a) shows the aggregated mean performance in safety tasks across the seven target languages. We first observe that ElChat outperforms CV in TruthfulQA for both Qwen2.5 and Llama 3.1. In particular, it substantially helps Llama 3.1, achieving 13-point gains over CV on average. We speculate that CV may be less effective for classification tasks as reflected in its performance on other discriminative target language tasks (§5.2).

In contrast, CV often surpasses ElChat in ToxiGen and ImplicitHate for both models, with differences of up to 6 points, e.g. in ToxiGen with Qwen2.5. This is primarily due to the use of Merge instead of Copy in ElChat (see Table 7). Specifically, ElChat without Copy follows similar trends to ElChat, while ElChat without Merge exhibits similar trends to Chat+VE and CV. This is intuitive, as merging a model with the lowest target safety performance (i.e. Chat) with Chat+VE can degrade safety performance.



(a) Target Language

(b) English (Source)

Figure 3: Aggregated mean performance (↑) across seven target languages for each model on safety, chat, and instruction-following tasks. Full results are in the Appendix (Tables 6 and 7).

We finally find that ElChat, CV, and Chat+VE outperform the Chat baseline across tasks and models, with gains ranging from 1.4 (TruthfulQA with Llama 3.1 using ElChat) to 20 points (TruthfulQA with Qwen2.5 using ElChat). The only exceptions are CV and Chat+VE in TruthfulQA, where they underperform Chat by 11 and 14 points, respectively. This instability highlights the advantage of ElChat, as it consistently enhances safety performance over Chat across tasks.

Chat and Instruction-following. Figure 3 (b) shows the aggregated mean performance across chat and instruction-following tasks in English (source). We first analyze the extent to which ElChat impacts performance on chat and instruction-following tasks compared to the Chat baseline. Note that some performance degradation is inevitable, as adapting Chat to *target* unlabeled data inherently affects the *source* chat and instruction-following abilities (§3).

As anticipated, we find that ElChat exhibits performance degradation across tasks and models compared to the Chat baseline. However, the extent of this degradation varies depending on the model, task, and adaptation approach. For example, ElChat achieves comparable performance on GSM8K but experiences reductions of 16 and 0.57 points on IFEval and MT-Bench, respectively. Despite these drops, ElChat successfully improves instruction-following performance compared to the respective adapted model, Chat+VE. It demonstrates improvements of 14 and 13 points over Chat+VE for Qwen2.5 and Llama 3.1, respectively. These results indicate that ElChat can inject instruction-following capabilities into the adapted model.

We next observe that ElChat generally outperforms CV across tasks and models in five out of six cases, with performance differences ranging from 9.5 points (IFEval with Llama 3.1) to 25 points (GSM8K with Llama

Table 1: MGSM performance in Bengali (bn) and Telugu (te) by model.

(a) Qwen2.5 7B

Model	EM	
	bn	te
Chat	.23	.06
CV	**.60**	.27
Chat+VE	.39	.27
ElChat (Ours)	.46	**.35**

(b) Llama 3.1 8B

Model	EM	
	bn	te
Chat	.30	.12
CV	.31	.24
Chat+VE	.26	.28
ElChat (Ours)	**.51**	**.41**

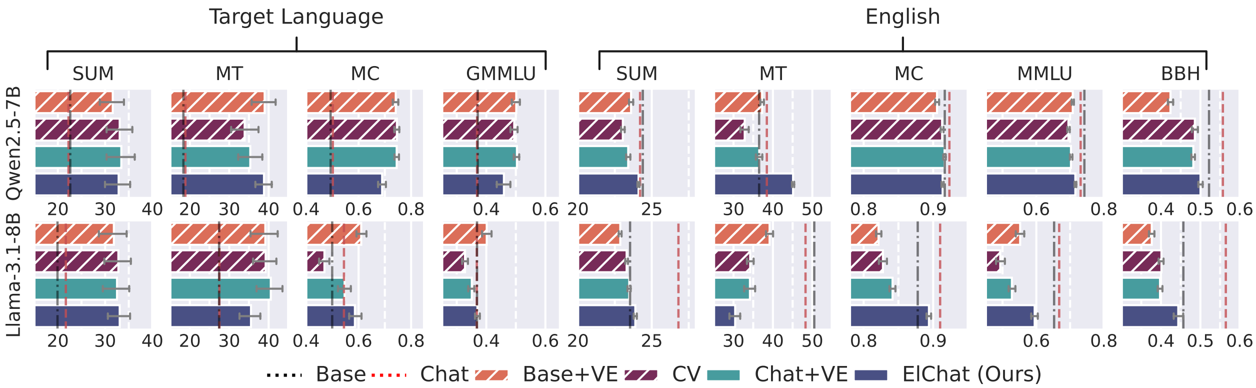


Figure 4: Aggregated mean performance across seven target languages for each model (error bars indicate 95% confidence interval). Full results are in Tables 11 and 12.

3.1). A similar trend is observed on AlpacaEval (Table 10 in the Appendix), where ElChat substantially outperforms CV with up to a 16-point gain in Qwen2.5, while still performing competitively with only a 0.95-point drop in Llama 3.1. These results suggest that ElChat is more effective than CV in enhancing both chat and instruction-following abilities.

Finally, ElChat's performance advantage extends to target language tasks. Table 1 shows model performance in MGSM (covering Bengali and Telugu). Overall, ElChat surpasses both CV and the original chat model, Chat, in three and four out of four cases, respectively. Notably, ElChat significantly improves Telugu performance by 29 points with both Qwen2.5 and Llama 3.1. Additionally, while CV enhances Bengali performance by just 1 point over Chat in Llama 3.1, ElChat achieves a substantial 21-point gain. These results further support the superiority of ElChat over CV.

5.2 Target Language

Figure 4 (left) shows the aggregated mean performance across seven languages for all source and adapted target models in target language.

We note that ElChat consistently outperforms its source chat model (Chat) across all models and tasks. This improvement is particularly notable in generative tasks (i.e. SUM and MT), with gains ranging from 8 points (Llama 3.1 on MT) to 20 points (Qwen2.5 on MT). While ElChat generally maintains competitive performance (within 3 points) compared to Chat+VE, it exhibits slightly reduced performance (up to 5.9 points with Qwen2.5 on MC) in discriminative tasks (i.e. MC and GMMLU) and MT with Llama 3.1. These results suggest that ElChat can overall preserve the target language performance, while the model modifications do not substantially degrade performance.

We further observe that ElChat demonstrates competitive performance with CV, with each method outperforming the other in half of the evaluated cases. Specifically, CV generally outperforms ElChat in Qwen2.5, except for the MT task, whereas ElChat typically achieves better performance than CV in Llama 3.1, excluding MT. However, CV notably underperforms both the source base and chat models in the two discriminative tasks (i.e. MC and GMMLU) with Llama 3.1. Thus, although ElChat and CV achieve similar overall performance, our method is more likely to yield improvements in target language tasks.

5.3 Source Language (English)

Figure 4 (right) shows the aggregated mean performance across seven languages for all source and adapted target models in English tasks. ElChat generally underperforms the source Chat baseline. However, the degree of performance degradation varies considerably depending on the task and model similar to §5.1. For instance, with Llama 3.1, ElChat exhibits substantial performance drops of 7, 12, and 18 points on MMLU, BBH and MT, respectively. In contrast, the performance degradation observed with Qwen2.5 is less

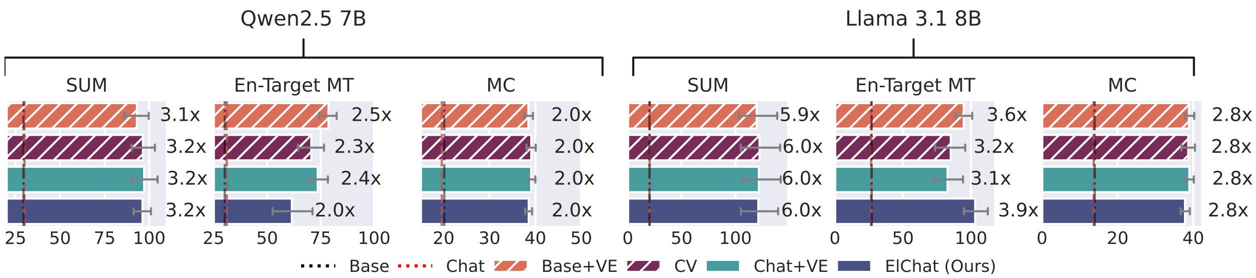


Figure 5: Aggregated mean inference speedup (tokens/s) across seven target languages. The value next to each bar represents the speedup ratio over Chat. Full results are in the Appendix (Tables 17 and 18).

pronounced, with a maximum decrease of 5.7 points on BBH. Interestingly, ElChat even demonstrates a 6.7-point improvement on MT with Qwen2.5. This improvement likely stems from two key factors: (1) ElChat's effective utilization of source tokens (95%) during generation, compared to other approaches (i.e. Base+VE, CV, and Chat+VE) that achieve at most 71% source token utilization; and (2) its successful early stopping, generating a similar number of tokens (33) as Chat (see Figures 7 and 8 in the Appendix).

Comparing ElChat and Chat+VE, we find that ElChat generally yields better or comparable performance across models and tasks, with the exception of MT with Llama 3.1. This improvement suggests that ElChat can generally alleviate catastrophic forgetting not only in chat and instruction-following tasks (§5.1) but also in source language tasks by injecting the source chat information into Chat+VE.

Conversely, CV exhibits poor performance on English tasks, typically falling short of ElChat in eight out of ten cases. Moreover, the benefits of catastrophic forgetting mitigation are not consistently observed with CV, as it only outperforms its adapted base model, Base+VE, in five out of ten cases. This result somewhat contrasts with Huang et al. (2024), who highlight the benefits of using CV to mitigate catastrophic forgetting and improve knowledge retention and language ability. We speculate that modifying models through simple arithmetic operations, as in CV, may be less robust than our method. These results suggest that ElChat more effectively integrates target language abilities while mitigating performance degradation across chat, instruction-following, and source English tasks compared to CV.

6 Inference Efficiency

VE offers inference speedups on target language tasks by reducing text overfragmentation (§4.6). Since ElChat and CV involve modifications to model weights, they may increase the generation of source tokens, potentially slowing inference compared to Chat+VE and Base+VE. Furthermore, the inference efficiency of CV with VE has not been previously studied. Therefore, we investigate the inference efficiency of both methods. Figure 5 shows the aggregated mean inference speedups across seven languages and three target tasks.

Overall, both ElChat and CV achieve comparable speedups to their respective adapted models, Chat+VE and Base+VE, demonstrating 2.0x to 6.0x speedups over the source chat model, Chat. The largest drop is observed between ElChat and Chat+VE on the MT task with Qwen2.5 (0.4x). This likely happens because ElChat frequently terminates generation early (see Table 8), resulting in fewer generated tokens (and thus a smaller numerator in the tokens/second calculation). In summary, these results suggest that the model modifications introduced by ElChat and CV do not negatively impact the inference speedups provided by VE.

7 Analysis

To better understand the behavior of ElChat, we perform an ablation in chat and instruction-following tasks using Llama 3.1. We also conduct a qualitative analysis using MGSM to gain further insights.

[User] టౌలా ఒక బేకరీకి వెళ్లి, వివిధ రకాలైన పేస్ట్రీలను కొనుగోలు చేసింది. ఆమె డజను $68 ఖరీదుతో 3 డజన్ల డోనట్‌లను, డజను $80 డాలర్లు ఖరీదుతో 2 డజన్ల మినీ కప్ కేకులను, డజను $55 ఖరీదుతో 6 డజన్ల మినీ చీజ్‌కేక్‌లను కొనుగోలు చేసింది. మొత్తం ఖర్చు ఎంత?

Toula went to a bakery and bought a variety of pastries. She bought 3 dozen donuts for $68 a dozen, 2 dozen mini cupcakes for $80 a dozen, and 6 dozen mini cheesecakes for $55 a dozen. What was the total cost?

→[ElChat] మినీ చీజ్‌కేక్‌లు 6*55=330 డాలర్లు ఖరీదుతో 6 డజన్ల కొనుగోలు చేయబడ్డాయి. మినీ కప్ కేకులు 2*80=160 డాలర్లు ఖరీదుతో 2 డజన్ల కొనుగోలు చేయబడ్డాయి. డోనట్‌లు 3*68=204 డాలర్లు ఖరీదుతో 3 డజన్ల కొనుగోలు చేయబడ్డాయి. మొత్తం ఖర్చు 330+160+204=694 డాలర్లు. సమాధానం 694.

*Mini cheesecakes were purchased at a cost of 6*55=330 dollars, 6 dozen. Mini cupcakes were purchased at a cost of 2*80=160 dollars, 2 dozen. Donuts were purchased at a cost of 3*68=204 dollars, 3 dozen. Total cost was 330+160+204=694 dollars. Answer is 694.*

(a) Correct answer - Correct reasoning

[User] রায়ান তাঁর বাগানে একদিনে 2টি ফুল লাগিয়েছেন। 15 দিন পরে, যদি 5টি না বেড়ে ওঠে তবে মোট কটি ফুল থাকবে?

Ryan planted 2 flowers in his garden in one day. After 15 days, if 5 do not grow, how many flowers will be left in total?

→[ElChat] রায়ান তাঁর বাগানে 1 দিনে 2টি ফুল লাগিয়েছেন। 15 দিন পরে তিনি 2 * 15 = 30টি ফুল লাগিয়েছেন। তিনি 5টি ফুল লাগিয়েছেন না। তাই তাঁর বাগানে মোট 30 - 5 = 25টি ফুল থাকবে। উত্তর হল 25।

*Ryan planted 2 flowers in his garden in 1 day. After 15 days he planted 2 * 15 = 30 flowers. He did not plant 5 flowers. So there will be a total of 30 - 5 = 25 flowers in his garden. The answer is 25.*

(b) Correct answer - Wrong reasoning

Figure 6: Output examples on MGSM. We use Google Translate for translation. Few-shot demonstrations (conversation history) are omitted due to space constraints.

Ablation. Table 2 shows the importance of each ElChat component. Removing either of the two components (§3) significantly reduces performance across chat and instruction-following tasks, demonstrating their complementarity in eliciting ElChat's abilities. For instance, ablating the weight copying (Copy) component, leaving only VE and model merging, causes performance to drop by an average of 8 points on IFEval, 21 points on GSM8K, 14 points on MGSM, and 0.27 points on MT-Bench relative to ElChat. This substantially undermines the advantage of ElChat over CV as it greatly lags behind CV in MT-Bench and shows marginally better performance than CV with only a few points (up to 4 points on GSM8K) differences in other tasks, unlike at least 9 points difference observed between ElChat and CV. These results clearly demonstrate that while the individual components are established techniques, their synergistic effect within a chat model context is non-trivial, and their combination is essential for ElChat's strong performance.

Table 2: Mean performance across languages for chat and instruction-following tasks using Llama 3.1. Ablation results in other tasks are available in Appendix B.

Model	**IFEval**	**GSM8K**	**MGSM**	**MT-Bench**
CV	$.38_{.03}$	$.33_{.11}$	$.27_{.03}$	$3.04_{.20}$
ElChat	$.47_{.04}$	$.58_{.06}$	$.46_{.05}$	$2.92_{.27}$
– Merge	$.34_{.05}$	$.40_{.18}$	$.37_{.06}$	$2.53_{.22}$
– Copy	$.39_{.02}$	$.37_{.17}$	$.32_{.04}$	$2.65_{.23}$

Qualitative Analysis. Figure 6 presents examples of ElChat's output on MGSM, highlighting both successes and challenges. Case (a) showcases correct reasoning and answer generation in Telugu. However, case (b) demonstrates that even when providing a correct answer in Bengali, ElChat can exhibit wrong reasoning. The misinterpretation of "five flowers do not grow" as "Ryan did not plant five flowers" suggests a potential limitation in understanding nuanced language. Notably, in the same case, it correctly reasons in English that "5 did not grow". Focusing on improving target language abilities further during VE while mitigating catastrophic forgetting of chat and instruction-following could address this issue. For instance, making iterative model merging (Alexandrov et al., 2024) applicable to low-resource settings is a potential avenue for future investigation.

8 Conclusion

We introduced ElChat, a method for directly adapting *chat* models with VE using unlabeled data, eliminating the need for a *base* model and target chat data. To mitigate potential catastrophic forgetting in the adapted chat models after VE, ElChat elicits chat abilities by injecting information from the source chat model without requiring further training. Extensive experiments across safety, chat, and instruction-following, target language, and source language tasks demonstrated that ElChat outperforms the previous state-of-the-art CV approach in chat and instruction-following, and English tasks while being competitive and more robust in the target language and safety tasks. These results highlight ElChat's superior abilities.

Limitations

Continual Pre-training Methods. This paper uses a continual pre-training method proposed by Remy et al. (2024), which tunes the top and bottom two layers of a model and its embedding and language modeling head, for efficient and effective target language adaptation. Nonetheless, other continual pre-training methods exist, including adapter-based training (e.g. LoRA (Hu et al., 2022)) and full fine-tuning. It would be interesting to extensively investigate the effect of different training methods for future work, but this is beyond the scope of this paper.

Model Merging Methods. We experiment with linear and SLERP merging as representative model merging methods for simplicity. More recent methods like TIES (Yadav et al., 2023) and DARE-TIES (Yu et al., 2024) might perform even better in ElChat. Given the resource constraints, we leave this investigation for future work.

Languages. This paper covers seven typologically diverse languages, following previous work on language adaptation that has also tested a similar number of languages. For instance, Minixhofer et al. (2022) tested eight languages. Note that Huang et al. (2024) used three languages (i.e. Chinese, Traditional Chinese, and Korean) to verify the effectiveness of CV. Experimenting with more languages is an interesting avenue for future work but is out of the scope of this paper, given our limited computing capacity.

Chat and Instruction-following Evaluation. Our chat and instruction-following evaluation is mainly on English data except for MGSM due to the limited availability of manually curated language-specific evaluation resources. Azime et al. (2024) has also noted the instability of using LLM-as-a-Judge in Amharic, which is also one of our experimental languages. It would be an interesting avenue to explore more chat and instruction-following evaluation in target languages for future work. We hope our work inspires the development of extensive evaluation benchmarks in low-resource languages.

Ethical Considerations

Although we conducted extensive experiments across diverse public datasets to validate the effectiveness of ElChat, these datasets do not fully represent all real-world scenarios. Therefore, any model derived from or based on this work should be used with caution.

While this work does not appear to raise immediate concerns, the deployment of adapted chat LMs, especially in under-monitored, low-resource language regions, warrants further analysis. These models could inadvertently perpetuate harmful biases, compromise safety in ways not captured by current benchmarks, or be exploited for misinformation and other malicious purposes. Further research and responsible deployment strategies are crucial to address these potential risks.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro

Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone, 2024a. URL `https://arxiv.org/abs/2404.14219`.

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 technical report, 2024b. URL `https://arxiv.org/abs/2412.08905`.

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David Mortensen, Noah Smith, and Yulia Tsvetkov. Do all languages cost the same? tokenization in the era of commercial language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9904–9923, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.614. URL `https://aclanthology.org/2023.emnlp-main.614/`.

Anton Alexandrov, Veselin Raychev, Mark Niklas Müller, Ce Zhang, Martin Vechev, and Kristina Toutanova. Mitigating catastrophic forgetting in language transfer via model merging. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 17167–17186, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.1000. URL `https://aclanthology.org/2024.findings-emnlp.1000/`.

Mehdi Ali, Michael Fromm, Klaudia Thellmann, Richard Rutmann, Max Lübbering, Johannes Leveling, Katrin Klug, Jan Ebert, Niclas Doll, Jasper Buschhoff, Charvi Jain, Alexander Weber, Lena Jurkschat, Hammam Abdelwahab, Chelsea John, Pedro Ortiz Suarez, Malte Ostendorff, Samuel Weinbach, Rafet Sifa, Stefan Kesselheim, and Nicolas Flores-Herr. Tokenizer choice for LLM training: Negligible or crucial? In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 3907–3924, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.247. URL `https://aclanthology.org/2024.findings-naacl.247/`.

Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, Geeta Chauhan, Anjali Chourdia, Will Constable, Alban Desmaison, Zachary DeVito, Elias Ellison, Will Feng, Jiong Gong, Michael Gschwind, Brian Hirsh, Sherlock Huang, Kshiteej Kalambarkar, Laurent Kirsch, Michael Lazos, Mario Lezcano, Yanbo Liang, Jason Liang, Yinghai Lu, C. K. Luk, Bert Maher, Yunjie Pan, Christian Puhrsch, Matthias Reso, Mark Saroufim, Marcos Yukio Siraichi, Helen Suk, Shunting Zhang, Michael Suo, Phil Tillet, Xu Zhao, Eikan Wang, Keren Zhou, Richard Zou, Xiaodong Wang, Ajit Mathews, William Wen, Gregory Chanan, Peng Wu, and Soumith Chintala. Pytorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, ASPLOS '24, pp. 929–947, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703850. doi: 10.1145/3620665.3640366. URL `https://doi.org/10.1145/3620665.3640366`.

Israel Abebe Azime, Atnafu Lambebo Tonja, Tadesse Destaw Belay, Mitiku Yohannes Fuge, Aman Kassahun Wassie, Eyasu Shiferaw Jada, Yonas Chanie, Walelign Tewabe Sewunetie, and Seid Muhie Yimam. Walia-LLM: Enhancing Amharic-LLaMA by integrating task-specific and generative datasets. In

Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 432–444, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.25. URL https://aclanthology.org/2024.findings-emnlp.25/.

Abhinand Balachandran. Tamil-Llama: A new Tamil language model based on Llama 2, 2023. URL https://arxiv.org/abs/2311.05845.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 749–775, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.44. URL https://aclanthology.org/2024.acl-long.44/.

Lucas Bandarkar, Benjamin Muller, Pritish Yuvraj, Rui Hou, Nayan Singhal, Hongjiang Lv, and Bing Liu. Layer swapping for zero-shot cross-lingual transfer in large language models, 2024b. URL https://arxiv.org/abs/2410.01335.

Samuel Cahyawijaya, Holy Lovenia, Fajri Koto, Rifki Putri, Wawan Cenggoro, Jhonson Lee, Salsabil Akbar, Emmanuel Dave, Nuurshadieq Nuurshadieq, Muhammad Mahendra, Rr Putri, Bryan Wilie, Genta Winata, Alham Aji, Ayu Purwarianti, and Pascale Fung. Cendol: Open instruction-tuned generative large language models for Indonesian languages. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14899–14914, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.796. URL https://aclanthology.org/2024.acl-long.796/.

ChangSu Choi, Yongbin Jeong, Seoyoon Park, Inho Won, HyeonSeok Lim, SangMin Kim, Yejee Kang, Chanhyuk Yoon, Jaewan Park, Yiseul Lee, HyeJin Lee, Younggyun Hahm, Hansaem Kim, and KyungTae Lim. Optimizing language augmentation for multilingual large language models: A case study on Korean. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 12514–12526, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.1095/.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Zoltan Csaki, Pian Pawakapan, Urmish Thakker, and Qiantong Xu. Efficiently adapting pretrained language models to new languages, 2023. URL https://arxiv.org/abs/2311.05741.

Yiming Cui, Ziqing Yang, and Xin Yao. Efficient and effective text encoding for Chinese LLaMA and Alpaca, 2024. URL https://arxiv.org/abs/2304.08177.

Severino Da Dalt, Joan Llop, Irene Baucells, Marc Pamies, Yishi Xu, Aitor Gonzalez-Agirre, and Marta Villegas. FLOR: On the effectiveness of language adaptation. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 7377–7388, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.650/.

Konstantin Dobler and Gerard de Melo. Language adaptation on a tight academic compute budget: Tokenizer swapping works and pure bfloat16 is enough. In *2nd Workshop on Advancing Neural Network Training: Computational Efficiency, Scalability, and Resource Optimization (WANT@ICML 2024)*, 2024. URL https://openreview.net/forum?id=VYfJaHeVod.

C.m. Downey, Terra Blevins, Nora Goldfine, and Shane Steinert-Threlkeld. Embedding structure matters: Comparing methods to adapt multilingual vocabularies to new languages. In Duygu Ataman (ed.), *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pp. 268–281, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.mrl-1.20. URL https://aclanthology.org/2023.mrl-1.20/.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The Llama 3 herd of models, 2024. URL `https://arxiv.org/abs/2407.21783`.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

Abteen Ebrahimi, Manuel Mager, Shruti Rijhwani, Enora Rice, Arturo Oncevay, Claudia Baltazar, María Cortés, Cynthia Montaño, John E. Ortega, Rolando Coto-solano, Hilaria Cruz, Alexis Palmer, and Katharina Kann. Findings of the AmericasNLP 2023 shared task on machine translation into indigenous languages. In Manuel Mager, Abteen Ebrahimi, Arturo Oncevay, Enora Rice, Shruti Rijhwani, Alexis Palmer, and Katharina Kann (eds.), *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pp. 206–219, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.americasnlp-1.23. URL `https://aclanthology.org/2023.americasnlp-1.23/`.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. Latent hatred: A benchmark for understanding implicit hate speech. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 345–363, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.29. URL `https://aclanthology.org/2021.emnlp-main.29/`.

Clémentine Fourrier, Nathan Habib, Thomas Wolf, and Lewis Tunstall. Lighteval: A lightweight framework for llm evaluation, 2023. URL `https://github.com/huggingface/lighteval`.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. Continual pre-training for cross-lingual LLM adaptation: Enhancing japanese language capabilities. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=TQdd1VhWbe.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 12 2023. URL https://zenodo.org/records/10256836.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

Xiang Geng, Ming Zhu, Jiahuan Li, Zhejian Lai, Wei Zou, Shuaijie She, Jiaxin GUO, Xiaofeng Zhao, Yinglu Li, Yuang Li, Chang Su, Yanqing Zhao, Xinglin Lyu, Min Zhang, Jiajun Chen, Hao Yang, and Shujian Huang. Why not transform chat large language models to non-english?, 2025. URL https://openreview.net/forum?id=US2UCMvzvP.

Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vladimir Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. Arcee's MergeKit: A toolkit for merging large language models. In Franck Dernoncourt, Daniel Preoţiuc-Pietro, and Anastasia Shimorina (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 477–485, Miami, Florida, US, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-industry.36. URL https://aclanthology.org/2024.emnlp-industry.36/.

HyoJung Han, Akiko Eriguchi, Haoran Xu, Hieu Hoang, Marine Carpuat, and Huda Khayrallah. Adapters for altering LLM vocabularies: What languages benefit the most?, 2024. URL https://arxiv.org/abs/2410.09644.

Sungjun Han, Juyoung Suk, Suyeong An, Hyungguk Kim, Kyuseok Kim, Wonsuk Yang, Seungtaek Choi, and Jamin Shin. Trillion 7b technical report, 2025. URL https://arxiv.org/abs/2504.15431.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3309–3326, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.234. URL https://aclanthology.org/2022.acl-long.234/.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 4693–4703, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.413. URL https://aclanthology.org/2021.findings-acl.413/.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Jimin Hong, Gibbeum Lee, and Jaewoong Cho. Accelerating multilingual language model for excessively tokenized languages. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 11095–11111, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.660. URL https://aclanthology.org/2024.findings-acl.660/.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Shih-Cheng Huang, Pin-Zu Li, Yu-chi Hsu, Kuang-Ming Chen, Yu Tung Lin, Shih-Kai Hsiao, Richard Tsai, and Hung-yi Lee. Chat vector: A simple approach to equip LLMs with instruction following and model alignment in new languages. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10943–10959, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.590. URL https://aclanthology.org/2024.acl-long.590/.

Seungduk Kim, Seungtaek Choi, and Myeongho Jeong. Efficient and effective vocabulary expansion towards multilingual large language models, 2024. URL https://arxiv.org/abs/2402.14714.

Sneha Kudugunta, Isaac Rayburn Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. MADLAD-400: A multilingual and document-level large audited dataset. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=Y45ZCxslFx.

Celio Larcher, Marcos Piau, Paulo Finardi, Pedro Gengo, Piero Esposito, and Vinicius Caridá. Cabrita: closing the gap for foreign languages, 2023. URL https://arxiv.org/abs/2308.11878.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. Datasets: A community library for natural language processing. In Heike Adel and Shuming Shi (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 175–184, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-demo.21. URL https://aclanthology.org/2021.emnlp-demo.21/.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. AlpacaEval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval, 5 2023.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Peiqin Lin, Shaoxiong Ji, Jörg Tiedemann, André F. T. Martins, and Hinrich Schütze. Mala-500: Massive language adaptation of large language models, 2024. URL https://arxiv.org/abs/2401.13303.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229/.

Samin Mahdizadeh Sani, Pouya Sadeghi, Thuy-Trang Vu, Yadollah Yaghoobzadeh, and Gholamreza Haffari. Extending LLMs to new languages: A case study of llama and Persian adaptation. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 8868–8884, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.594/.

Joshua Maynez, Priyanka Agrawal, and Sebastian Gehrmann. Benchmarking large language model capabilities for conditional generation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9194–9213, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.511. URL https://aclanthology.org/2023.acl-long.511/.

Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3992–4006, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.293. URL https://aclanthology.org/2022.naacl-main.293/.

Benjamin Minixhofer, Edoardo Ponti, and Ivan Vulić. Zero-shot tokenizer transfer. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=RwBObRsIzC.

Nandini Mundra, Aditya Nanda Kishore Khandavally, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, and Mitesh M Khapra. An empirical comparison of vocabulary expansion and initialization approaches for language models. In Libby Barak and Malihe Alikhani (eds.), *Proceedings of the 28th Conference on Computational Natural Language Learning*, pp. 84–104, Miami, FL, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.conll-1.8. URL https://aclanthology.org/2024.conll-1.8/.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Zhiqiang Hu, Chenhui Shen, Yew Ken Chia, Xingxuan Li, Jianyu Wang, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. SeaLLMs - large language models for Southeast Asia. In Yixin Cao, Yang Feng, and Deyi Xiong (eds.), *Proceedings of the 62nd Annual Meeting of the Association* *for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 294–304, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-demos.28. URL `https://aclanthology.org/2024.acl-demos.28/`.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind: Scaling human-centered machine translation. *ArXiv*, abs/2207.04672, 2022. URL `https://arxiv.org/abs/2207.04672`.

Malte Ostendorff and Georg Rehm. Efficient language model training through cross-lingual and progressive transfer learning, 2023. URL `https://arxiv.org/abs/2301.09626`.

Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Bibi. Language model tokenizers introduce unfairness between languages. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=78yDLKi95p`.

Kunat Pipatanakul, Phatrasek Jirabovonvisut, Potsawee Manakul, Sittipong Sripaisarnmongkol, Ruangsak Patomwong, Pathomporn Chokchainant, and Kasima Tharnpipitchai. Typhoon: Thai large language models, 2023. URL `https://arxiv.org/abs/2312.13951`.

Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina (eds.), *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pp. 392–395, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3049. URL `https://aclanthology.org/W15-3049/`.

Maja Popović. chrF++: words helping character n-grams. In Ondřej Bojar, Christian Buck, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, and Julia Kreutzer (eds.), *Proceedings of the Second Conference on Machine Translation*, pp. 612–618, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4770. URL `https://aclanthology.org/W17-4770/`.

Matt Post. A call for clarity in reporting BLEU scores. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor (eds.), *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 186–191, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6319. URL `https://aclanthology.org/W18-6319/`.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL `https://qwenlm.github.io/blog/qwen2.5/`.

François Remy, Pieter Delobelle, Hayastan Avetisyan, Alfiya Khabibullina, Miryam de Lhoneux, and Thomas Demeester. Trans-tokenization and cross-lingual vocabulary transfers: Language adaptation of LLMs for low-resource NLP. In *First Conference on Language Modeling*, 2024. URL `https://openreview.net/forum?id=sBxvoDhvao`.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=fR3wGCk-IXp`.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I. Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne

Longpre, Wei-Yin Ko, Madeline Smith, Antoine Bosselut, Alice Oh, Andre F. T. Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation, 2024a. URL `https://arxiv.org/abs/2412.03304`.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Chien, Sebastian Ruder, Surya Guthikonda, Emad Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. Aya dataset: An open-access collection for multilingual instruction tuning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11521–11567, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.620. URL `https://aclanthology.org/2024.acl-long.620/`.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Johan Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew Kyle Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, Cesar Ferri, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Christopher Waites, Christian Voigt, Christopher D Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, C. Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodolà, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germàn Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Xinyue Wang, Gonzalo Jaimovitch-Lopez, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Francis Anthony Shevlin, Hinrich Schuetze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B Simon, James Koppel, James Zheng, James Zou, Jan Kocon, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh Dhole, Kevin Gimpel, Kevin Omondi, Kory Wallace Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros-Colón, Luke Metz, Lütfi Kerem Senel, Maarten

Bosma, Maarten Sap, Maartje Ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramirez-Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael Andrew Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Andrew Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter W Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan Le Bras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Russ Salakhutdinov, Ryan Andrew Chi, Seungjae Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel Stern Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima Shammie Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven Piantadosi, Stuart Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsunori Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Venkatesh Ramasesh, vinay uday prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Sophie Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL `https://openreview.net/forum?id=uyTL5Bvosj`. Featured Certification.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13003–13051, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.824. URL `https://aclanthology.org/2023.findings-acl.824/`.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. Language-specific neurons: The key to multilingual capabilities in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5701–5715, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.309. URL `https://aclanthology.org/2024.acl-long.309/`.

Mingxu Tao, Chen Zhang, Quzhe Huang, Tianyao Ma, Songfang Huang, Dongyan Zhao, and Yansong Feng. Unlocking the potential of model merging for low-resource languages. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 8705–8720, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.508. URL https://aclanthology.org/2024.findings-emnlp.508/.

Atula Tejaswi, Nilesh Gupta, and Eunsol Choi. Exploring design choices for building language-specific LLMs. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10485–10500, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.614. URL https://aclanthology.org/2024.findings-emnlp.614/.

Cagri Toraman. LlamaTurk: Adapting open-source generative large language models for low-resource language, 2024. URL https://arxiv.org/abs/2405.07745.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Do llamas work in English? on the latent language of multilingual transformers. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15366–15394, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.820. URL https://aclanthology.org/2024.acl-long.820/.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6/.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23965–23998. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/wortsman22a.html.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 483–498, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41. URL https://aclanthology.org/2021.naacl-main.41/.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-merging: Resolving interference when merging models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=xtaX3WyCj1.

Atsuki Yamaguchi, Aline Villavicencio, and Nikolaos Aletras. An empirical study on cross-lingual vocabulary adaptation for efficient language model inference. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 6760–6785, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.396. URL https://aclanthology.org/2024.findings-emnlp.396/.

Atsuki Yamaguchi, Aline Villavicencio, and Nikolaos Aletras. How can we effectively expand the vocabulary of LLMs with 0.01GB of target language text?, 2024b. URL https://arxiv.org/abs/2406.11477.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024. URL `https://arxiv.org/abs/2407.10671`.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL `https://arxiv.org/abs/2505.09388`.

Yunzhi Yao, Shaohan Huang, Wenhui Wang, Li Dong, and Furu Wei. Adapt-and-distill: Developing small, fast and effective pretrained language models for domains. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 460–470, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.40. URL `https://aclanthology.org/2021.findings-acl.40/`.

Zheng Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, Genta Winata, Stella Biderman, Edward Raff, Dragomir Radev, and Vassilina Nikoulina. BLOOM+1: Adding language support to BLOOM for zero-shot prompting. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11682–11703, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.653. URL `https://aclanthology.org/2023.acl-long.653/`.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *Forty-first International Conference on Machine Learning*, 2024. URL `https://openreview.net/forum?id=fq0NaiU8Ex`.

Jun Zhao, Zhihao Zhang, Luhui Gao, Qi Zhang, Tao Gui, and Xuanjing Huang. LLaMA beyond English: An empirical study on language capability transfer, 2024a. URL `https://arxiv.org/abs/2401.01055`.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. How do large language models handle multilingualism? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024b. URL `https://openreview.net/forum?id=ctXYOoAgRy`.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL `https://openreview.net/forum?id=uccHPGDlao`.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023. URL `https://arxiv.org/abs/2311.07911`.

A Experimental Setup

A.1 Chat Template and Special Tokens

Model-specific chat templates and special tokens are accessible via the following links:

- **Qwen2.5**: `https://huggingface.co/Qwen/Qwen2.5-7B-Instruct/blob/main/tokenizer_config.json`
- **Llama 3.1**: `https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct/blob/main/tokenizer_config.json`
- **Qwen3**: `https://huggingface.co/Qwen/Qwen3-14B/blob/main/tokenizer_config.json`

Below are excerpts from the chat templates of each model with placeholders for a prompt and output:

Qwen2.5.

```
<|im_start|>system
You are Qwen, created by Alibaba Cloud.
You are a helpful assistant.
<|im_end|>
<|im_start|>user
{prompt}
<|im_end|>
<|im_start|>assistant
{output}
```

Llama 3.1.

```
<|begin_of_text|>
<|start_header_id|>system
<|end_header_id|>
Cutting Knowledge Date: December 2023
Today Date: 26 Jul 2024
<|eot_id|>
<|start_header_id|>user
<|end_header_id|>
{prompt}
<|eot_id|>
<|start_header_id|>assistant
<|end_header_id|>
{output}
```

Qwen3.

```
<|im_start|>user
{prompt}
<|im_end|>
<|im_start|>assistant
<think>
</think>
{output}
```

A.2 Prompt Template

We translate the English prompt templates provided by Ahia et al. (2023) for SUM with a machine translation API, following Yong et al. (2023). For MT and MC, we formulate a task-specific English prompt, followed by machine translation for each language. For the remaining tasks, except for MT-BENCH, we use the default templates provided in `lm-evaluation-harness` (Gao et al., 2023). For MT-BENCH, we use the default template provided in LightEval (Fourrier et al., 2023). Table 3 shows the prompt templates used in our evaluation. Note that we do not make any changes to the task-specific prompt to allow for a fair comparison between models with and without a chat template (i.e. Base and Chat, respectively).

Table 4: Hyperparameters for continual pre-training.

Hyperparameters	**Values**
Batch size	32
Number of training steps	30,517
Adam ϵ	1e-8
Adam β_1	0.9
Adam β_2	0.999
Sequence length	512
Learning rate	5e-5
Learning rate scheduler	cosine
Warmup steps	First 5% of steps
Weight decay	0.01
Attention dropout	0.0
Training precision	BF16

Table 5: Parameters for non-greedy generative tasks: MT and SUM.

Parameters	**Values**
Temperature	0.8
Repetition penalty	1.1
Top k	40
Top p	0.9
Beam width	5
Sampling	True
Early stopping	True
Maximum number of generated tokens	128

A.3 Implementation Details

Our experimental design is based on the findings from Tejaswi et al. (2024). They report that (1) *there are no significant gains when adding more than 10K tokens to the source vocabulary*, and (2) *additional CPT in the order of millions of tokens is sufficient for model adaptation*. Given this, we set the vocabulary size of the auxiliary target language tokenizer $|\mathcal{V}_{\text{aux}}|$ to 50K across languages and the number of new target tokens k to 10K. We train each model for 500M tokens with a batch size of 32, a maximum learning rate of 5e-5, and a sequence length of 512.

Tokenizer Training. We train tokenizers using Hugging Face Tokenizers.

Preprocessing. We preprocess datasets with Hugging Face Datasets (Lhoest et al., 2021).

Continual Pre-training. We implement our models using PyTorch (Ansel et al., 2024) and Hugging Face Transformers (Wolf et al., 2020). Table 4 lists the hyperparameters in CPT.

Model Merging. To ensure a smooth transition between layers, we use a 0.3 : 0.7 = Chat : Chat+VE mixing ratio for the top and bottom layers of all merged models, favoring Chat+VE as these layers are adjacent to the embeddings and language modeling head of Chat+VE. For the second top and bottom layers, we use 0.5 : 0.5 = Chat : Chat+VE, balancing the contributions of Chat and Chat+VE.

Evaluation. We use Hugging Face LightEval[10] for evaluation on all tasks except for AlpacaEval, IFEval, GSM8K and MGSM. For AlpacaEval, we use the official implementation available on GitHub[11] (v0.6.6). For IFEval, GSM8K, and MGSM, we use `lm-evaluation-harness` (Gao et al., 2023). To compute ROUGE-L, we split sentences with an mT5 (Xue et al., 2021) tokenizer as preprocessing following Maynez et al. (2023) and subsequently call `rouge_scorer`[12] to compute the metric. To compute chrF and chrF++, we use SacreBLEU (Post, 2018). For safety evaluation, we follow Cahyawijaya et al. (2024) and use their implementation available on GitHub: `https://github.com/IndoNLP/cendol`.

Table 5 lists the parameters used during evaluation for generative tasks: mt and sum. To make a fair comparison, we do not conduct any generation parameter tuning and use the same ones across all approaches.

Hardware. We use either a single NVIDIA A100 (80GB), NVIDIA H100 (80GB), or NVIDIA GH200 (96GB) for CPT. For CPT with Qwen3 14B, we use a single AMD MI300X GPU. For evaluation, we use a single NVIDIA A100 (80GB) for all Llama 3.1 models, a single NVIDIA H100 (80GB) for all Qwen2.5 models, and a single AMD MI300X GPU for all Qwen3 models to ensure accurate measurement of inference efficiency.

Model Links. We list all the source model URLs in the following:

- **Qwen2.5** (Chat): `https://huggingface.co/Qwen/Qwen2.5-7B-Instruct`
- **Qwen2.5** (Base): `https://huggingface.co/Qwen/Qwen2.5-7B`
- **Llama 3.1** (Chat): `https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct`
- **Llama 3.1** (Base): `https://huggingface.co/meta-llama/Llama-3.1-8B`
- **Qwen3** (Base): `https://huggingface.co/Qwen/Qwen3-14B-Base`
- **Qwen3** (Chat): `https://huggingface.co/Qwen/Qwen3-14B`

B Results

B.1 Task Performance

- **Safety, Chat, and Instruction-following**: Tables 6, 7, and 8 provide a detailed breakdown of the task performance results for Qwen2.5, Llama 3.1, and Qwen3 across safety, chat, and instruction-following tasks. Table 9 shows the results of Chat+CPT and ElChat-related ablation models in mgsm. Table 10 shows the results of Chat-related models on AlpacaEval.
- **Target Language and Source (English)**: Tables 11, 12, and 13 provide a detailed breakdown of the task performance results for Qwen2.5, Llama 3.1, and Qwen3 across target language and source (English) language tasks.
- **sum and mt Results with Auxiliary Metrics**: Tables 14, 15, and 16 provide a detailed breakdown of sum and mt performance results, measured by ROUGE-L for sum and chrF++ for mt, for Qwen2.5, Llama 3.1, and Qwen3, respectively.

[10] `https://github.com/huggingface/lighteval`
[11] `https://github.com/tatsu-lab/alpaca_eval`
[12] `https://github.com/google-research/google-research/tree/master/rouge`

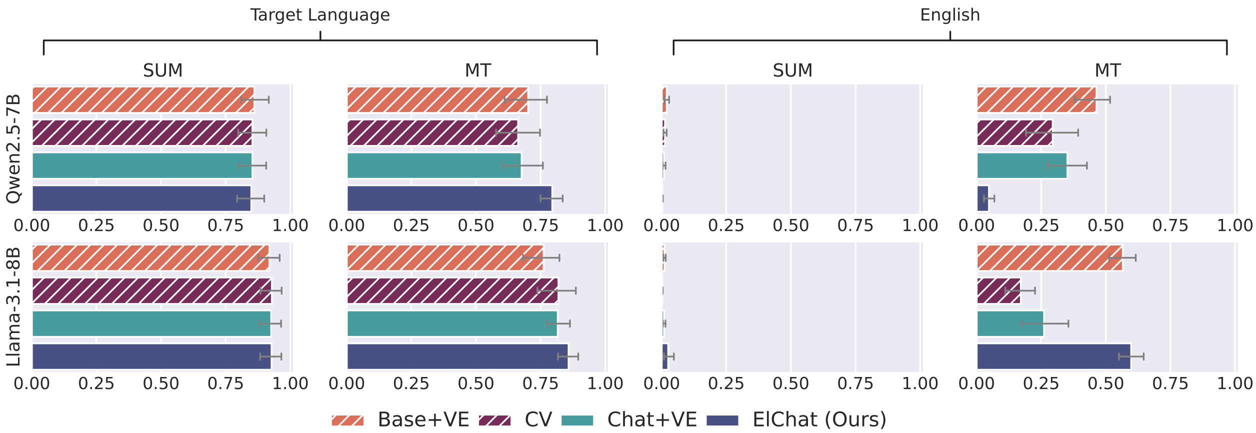


Figure 7: Aggregated average ratio of target tokens in output per sample across seven target languages for each model (error bars indicate 95% confidence interval).

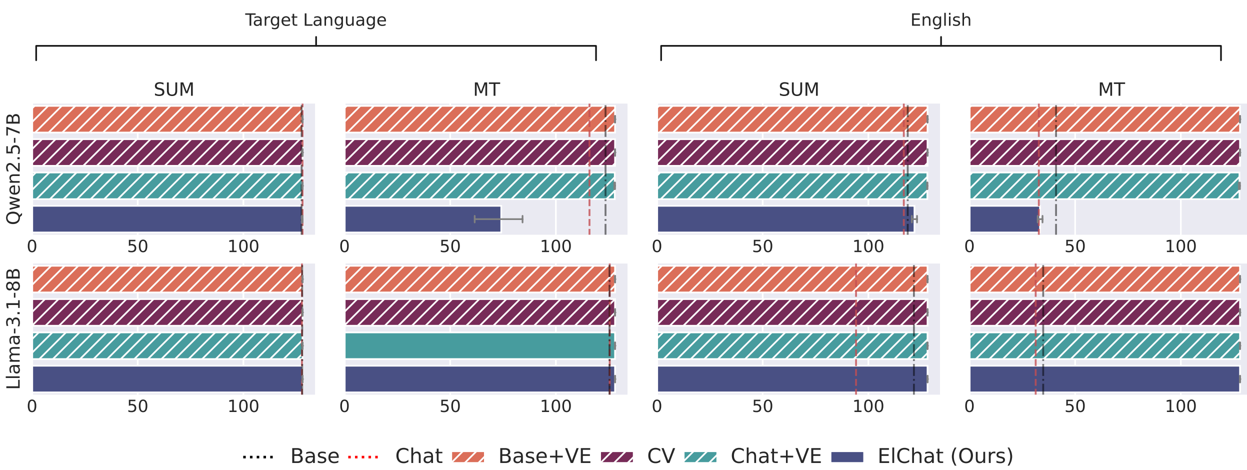


Figure 8: Aggregated average number of generated tokens per sample across seven target languages for each model (error bars indicate 95% confidence interval).

B.2 Inference Efficiency

Tables 17, 18, and 19 provide a detailed breakdown of the inference efficiency results for Qwen2.5, Llama 3.1, and Qwen3, respectively.

B.3 Ratio of Target Language Tokens

Figure 7 shows the aggregated mean ratio of target new tokens in output per sample across seven target languages for each model.

B.4 Number of Generated Tokens

Figure 8 shows the aggregated average number of generated tokens per sample across seven target languages for each model.

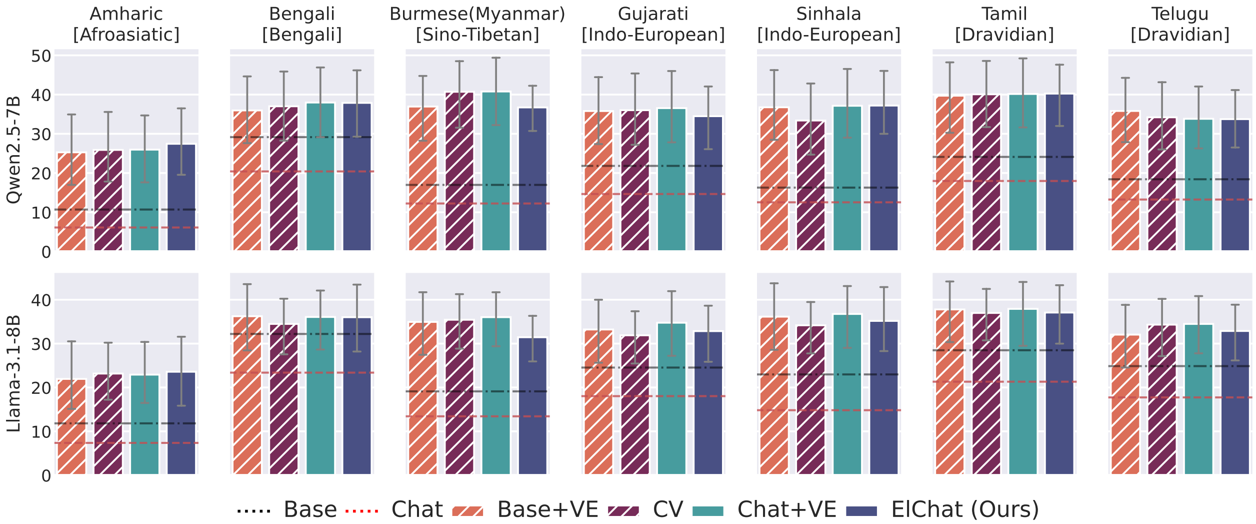


Figure 9: Aggregated mean performance across target language tasks for each model by language (error bars indicate 95% confidence interval).

C Analysis and Discussion

C.1 CPT-only vs. VE

Comparing the adapted chat models (Chat+VE and ElChat) with the CPT-only model (Chat+CPT) in Tables 11 and 12, we observe three key trends: (1) Chat+VE generally performs better than Chat+CPT on target language tasks across both models; (2) Chat+CPT often outperforms Chat+VE on MT tasks; and (3) ElChat either matches or surpasses Chat+CPT on nearly all tasks and models, except for Target MC and GMMLU with Qwen2.5 and target-to-English MT with Llama 3.1. This performance advantage of ElChat is also confirmed for chat and instruction-following tasks (Tables 6, 7, and 9), where ElChat substantially outperforms Chat+CPT in almost all cases across languages, tasks, and models.

We also observe a similar trend between Base+VE and the CPT-only adapted base model (Base+CPT) in Tables 11 and 12: Base+VE often outperforms Base+CPT in most of the tasks, while Base+CPT excels in target-to-English MT across models.

These results somewhat contradict the recent observations (Downey et al., 2023; Yamaguchi et al., 2024a;b) that CPT-only models often perform better than vocabulary adapted models, possibly due to the robustly aligned original embeddings. However, Fujii et al. (2024) reported that "*the overall impact of vocabulary expansion on performance is minor.*" Further, Dobler & de Melo (2024) also claimed that "*we do not see a clear trend of better performance with or without tokenizer swapping*" for their vocabulary adaptation experiments. We hypothesize that the superiority of CPT is greatly affected by the amount of CPT data, and it can be more apparent in low-resource settings as in Yamaguchi et al. (2024b), where new embeddings are likely to be not well aligned.

It is important to note that the CPT-only models (i.e. Base+CPT and Chat+CPT) have no speedups at all (Tables 17 and 18) as they use the same vocabulary as the source models (i.e. Base and Chat).

C.2 Additional Analysis by Language

We conduct additional analysis of the target language task performance of models by language. Figure 9 presents the aggregated mean performance across three target language tasks (i.e. Target SUM, English-to-target MT, and Target MC). Figure 10 shows the aggregated mean performance across *generative* target language tasks (i.e. Target SUM and English-to-target MT).

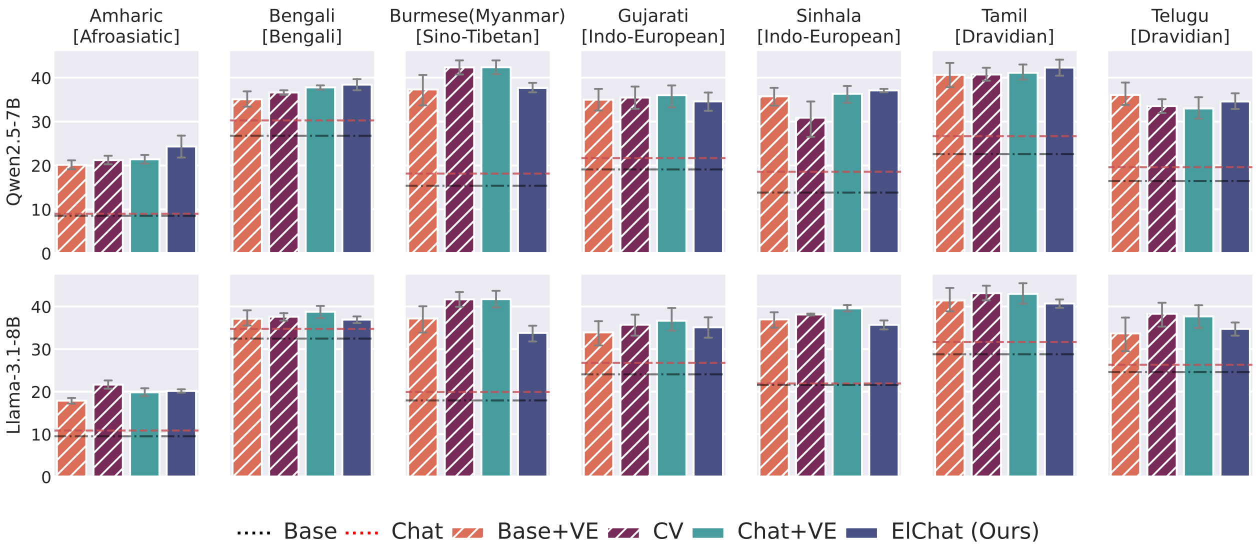


Figure 10: Aggregated mean performance across target language generative tasks (SUM and MT) for each model by language (error bars indicate 95% confidence interval).

Performance improvements achieved with VE are evident across languages and models. Overall, we observe from Figure 9 that ElChat consistently outperforms Base and Chat in all the target languages across models.

The extent to which adapted chat models improve target language task performance substantially varies by language and model. We observe from Figures 9 and 10 that ElChat substantially improves their target language performance over Base and Chat across models in Amharic, Burmese (Myanmar), and Sinhala. However, its improvement in Bengali remains minimal for both Qwen2.5 and Llama 3.1, especially in generative tasks (Figure 10). Similarly, for Llama 3.1, the performance gains in Tamil, Telugu, and Gujarati are less pronounced, particularly in generative tasks (Figure 10). *We hypothesize that this disparity strongly correlates with the amount of language-specific data used to train the source chat model.* Note that similar findings have already been reported by Yamaguchi et al. (2024a) and Tejaswi et al. (2024), *inter alia.*

How many languages do source chat LLMs support? The exact language coverage of the source LLMs remains unclear, as none explicitly list their supported languages.

Qwen2.5 (Qwen Team, 2024; Yang et al., 2024) reportedly supports over 29 languages, including Chinese, English, French, Spanish, Portuguese, German, Italian, Russian, Japanese, Korean, Vietnamese, Thai, and Arabic. However, none of our target languages are explicitly included in this list.

Llama 3.1 (Dubey et al., 2024) officially supports English, German, French, Italian, Portuguese, Hindi, Spanish, and Thai. However, it employs a FastText-based language identification model to categorize documents into 176 languages during pre-processing, suggesting that some of our target languages may be included in its pre-training data.

Additionally, Llama 3.1 reportedly utilizes 15T multilingual tokens (Dubey et al., 2024) for pre-training, while Qwen2.5 uses 18T tokens *in total* (not exclusively multilingual) (Qwen Team, 2024; Yang et al., 2024). This suggests that Llama 3.1 likely benefits from exposure to a broader set of languages. The relatively modest performance improvements observed for Gujarati, Tamil, and Telugu in Llama 3.1 could indicate that these languages were already present in its training data.

Approximating the language coverage. We can roughly estimate the language coverage of the source LLMs using MADLAD-400. The rationale is that these open-weight models are generally trained mainly on a mixture of publicly available data (Gemma Team et al., 2024; Dubey et al., 2024). Given that MADLAD-400 is sourced from CommonCrawl as mentioned in §4, its data distribution can potentially approximate the relative coverage of our target languages.

Table 20 presents the data statistics of MADLAD-400 for our target languages. We observe that Burmese, Amharic, and Sinhala have the fewest articles and total characters in MADLAD-400. Notably, these languages also exhibit the largest performance gains in our experiments. This suggests a strong correlation between the size of language-specific data in MADLAD-400 and the effectiveness of VE in task performance. On the basis of these results, we hypothesize that the two source LLMs used in our experiments might have been trained on very limited language-specific data, or possibly not at all, for Burmese, Amharic, and Sinhala. In contrast, Tamil, Telugu, Bengali, and Gujarati each have over 1 million articles, making them at least 7.4 times larger than the Burmese dataset. This further suggests a higher likelihood of their inclusion in the pre-training data of the source LLMs.

Table 20: Data size of MADLAD-400 Kudugunta et al. (2023) for each language.

Language	**Number of articles**	**Number of characters**
Tamil	5.6M	10.6B
Telugu	2.5M	5.1B
Bengali	4.3M	4.3B
Gujarati	1.3M	2.1B
Sinhala	788K	1.9B
Amharic	245.2K	509M
Burmese	176.5K	1.3B

C.3 Additional Analysis with Newer and Larger Model

We examine the efficacy of ElChat against a newer and larger model compared to Qwen2.5 7B and Llama 3.1 8B. To this end, we employ a state-of-the-art Qwen3 14B model, applying ElChat for adaptation. Due to computational limitations, our evaluation focuses on Amharic, Bengali, and Telugu. Amharic is selected because it shows the most significant speedup gains in target language generative tasks with Qwen2.5 and Llama 3.1 (Tables 17 and 18). Bengali and Telugu are chosen as they are the only languages covered by MGSM. Our analysis will focus on two key questions: (i) *Does Qwen3 exhibit a similar trend in performance with ElChat?* (ii) *If not, what are the potential reasons for the divergence?*

Safety. Table 8 shows that the results of Qwen3 align with those of Qwen2.5 in Figure 3. Specifically, CV outperforms ElChat on ToxiGen and ImplicitHate by up to 5 points, while ElChat performs slightly better on TruthfulQA (except for Telugu). Notably, both ElChat and CV generally surpass the Chat baseline across tasks, indicating that they improve safety performance even in this newer, larger model.

Chat and Instruction-following. Table 8 demonstrates that ElChat significantly better recovers chat and instruction-following abilities compared to CV across all tasks and languages. For example, ElChat shows drops of up to 6, 6, and 0.41 points from the Chat baseline in IFEval, GSM8K, and MT-Bench, respectively. In contrast, CV experiences considerably larger drops of up to 41, 56, and 0.66 points for the same tasks. A similar trend is observed in AlpacaEval (Table 10). On MGSM (Table 9), both ElChat and CV outperform the Chat baseline; they are competitive in Bengali with a 1-point difference, though CV outperforms ElChat in Telugu by 6 points. Overall, these trends are consistent with our previous observation in §5.1 that ElChat is more effective than CV in enhancing both chat and instruction-following capabilities.

Target Language. Contrary to our observations with Qwen2.5 and Llama 3.1 (§5.2; Figure 4, left), ElChat does not consistently outperform Chat in Qwen3 for target language tasks, as shown in Table 13. While we see consistent gains for Amharic, Bengali and Telugu show performance drops in MT, MC, and GMMLU with decreases up to 18 points (e.g. Bengali GMMLU). Similar drops are also observed in Chat+VE and CV, whereas Chat+CPT generally maintains performance comparable to Chat.

We hypothesize that Bengali and Telugu may already be well-represented within Qwen3 due to its significantly enhanced multilingual capabilities. Indeed, Qwen3 was pre-trained on 36T tokens, covering up to 119 languages and dialects. This represents a substantial increase from Qwen2.5, specifically, 29 supported languages and 18T tokens, suggesting a much broader and deeper understanding of various languages. This hypothesis is further supported by Table 36 in Yang et al. (2025), which lists both Bengali and Telugu as officially supported by Qwen3, while Amharic is not.

Finally, when comparing ElChat and CV, we observe they perform competitively across tasks and languages, with a maximum difference of 5.3 points in Amharic MT. This generally aligns with our observations in Qwen2.5 and Llama 3.1.

Source Language (English). Consistent with observations in §5.3 for Qwen2.5 and Llama 3.1, we observe from Table 13 that ElChat generally improves source language (English) performance compared to Chat+VE across tasks and languages. However, these improvements are marginal or modest, with a maximum gain of 6.3 points (Telugu MT). This smaller improvement in English indicates less performance degradation during the VE-process (i.e. less catastrophic forgetting of general English capabilities), suggesting enhanced robustness in Qwen3.

Comparing ElChat with CV, ElChat consistently demonstrates superior or equivalent performance in source language tasks across all languages. This aligns with our observation in §5.3. Notably, the gains tend to be more pronounced in the two generative tasks, SUM and MT, with improvements of up to 19.3 points (Bengali MT).

Inference Efficiency. Finally, Table 19 shows that both ElChat and CV consistently provide speedup gains of up to 3.8x (Amharic SUM with ElChat) in generative tasks within Qwen3. While CV exhibits some reductions in speedup ratios compared to Base+VE (e.g., in Amharic and Bengali), ElChat maintains inference speedups similar to Chat+VE across all tasks and languages. These findings are generally consistent with the results presented in §6 for Qwen2.5 and Llama 3.1.

D License

This study uses various publicly available models and datasets with different licenses, as detailed below, all of which permit their use for academic research.

D.1 Models

Qwen2.5 and Qwen3 are distributed under Apache License 2.0. Llama 3.1 is licensed under the Llama 3 Community License Agreement.[13]

D.2 Datasets

XL-Sum is licensed under CC BY-NC-SA 4.0. Belebele and FLORES-200 are licensed under CC BY-SA 4.0. BBH, MMLU, GSM8K, and ImplicitHate are distributed under the MIT License. MGSM is distributed under CC BY 4.0. ToxiGen is licensed under Community Data License Agreement - Permissive - Version 2.0. AlpacaEval, IFEval, MT-Bench, and TruthfulQA are distributed under Apache License 2.0.

[13] `https://llama.meta.com/llama3/license/`

Table 3: Prompt template for each task and language.

Task	Language	Template
En-Target MT	English	Translate English to {X: a target language}: {sentence} =
	Amharic	**እንግሊዝኛን ወደ አማርኛ ተርጉም**: {sentence} =
	Bengali	ইংরেজি থেকে বাংলায় অনুবাদ করুন: {sentence} =
	Burmese	အင်္ဂလိပ်မှ မြန်မာသို့ ဘာသာပြန်ပါ။: {sentence} =
	Gujarati	અંગ્રેજીમાંથી ગુજરાતીમાં અનુવાદ કરો: {sentence} =
	Sinhala	ඉංග්‍රීසි සිංහලයට පරිවර්තනය කරන්න: {sentence} =
	Tamil	ஆங்கிலத்திலிருந்து தமிழுக்கு மொழிபெயர்க்கவும்: {sentence} =
	Telugu	ఆంగ్లం నుండి తెలుగుకు అనువదించండి: {sentence} =
Target-En MT	English	Translate {X: a target language} to English: {sentence} =
	Amharic	**አማርኛን ወደ እንግሊዝኛ ተርጉም**: {sentence} =
	Bengali	বাংলা থেকে ইংরেজিতে অনুবাদ করুন: {sentence} =
	Burmese	မြန်မာမှ အင်္ဂလိပ်သို့ ဘာသာပြန်ပါ။: {sentence} =
	Gujarati	ગુજરાતીમાંથી અંગ્રેજીમાં અનુવાદ કરો: {sentence} =
	Sinhala	සිංහලයෙන් ඉංග්‍රීසියට පරිවර්තනය කරන්න: {sentence} =
	Tamil	தமிழிலிருந்து ஆங்கிலத்திற்கு மொழிபெயர்க்கவும்: {sentence} =
	Telugu	తెలుగు నుండి ఆంగ్లంకు అనువదించండి: {sentence} =
SUM	English	Write a short summary of the following text in {language}. Article: {text} Summary:
	Amharic	**የታችኛው ጽሁፍን በአማርኛ አጭር በማድረግ አሳትረኝ።**: {text} **አጭር መግለጫ**:
	Bengali	নিম্নলিখিত লেখাটি বাংলায় সংক্ষেপে লিখুন।: {text} সংক্ষিপ্তসার:
	Burmese	အောက်ပါစာသားကို မြန်မာဘာသာဖြင့် အကျဉ်းချုပ်ရေးပါ။ ဆောင်းပါး: {text} အကျဉ်းချုပ်:
	Gujarati	નીચે આપેલા લખાણને ગુજરાતીમાં સંક્ષિપ્ત લખો.: {text} સંક્ષેપ:
	Sinhala	පහත පාඨයේ සාරාංශය සිංහලෙන් ලියන්න. ලිපිය: {text} සාරාංශය:
	Tamil	கீழே உள்ள உரையை தமிழில் சுருக்கமாக எழுதுங்கள்: {text} சுருக்கம்:
	Telugu	క్రింది వచనం యొక్క సారాంశం తెలుగులో రాయండి. వ్యాసం: {text} సారాంశం:
MC	English	{passage} \n Question: {question} \n A. {answer 1}\n B. {answer 2}\n C. {answer 3}\n D. {answer 4}\n Answer:
	Amharic	{passage} \n **ጥያቄ**: {question} \n A. {answer 1}\n B. {answer 2}\n C. {answer 3}\n D. {answer 4}\n **መልስ**:
	Bengali	{passage} \n প্রশ্ন: {question} \n A. {answer 1}\n B. {answer 2}\n C. {answer 3}\n D. {answer 4}\n উত্তর:
	Burmese	{passage} \n မေးခွန်း: {question} \n A. {answer 1}\n B. {answer 2}\n C. {answer 3}\n D. {answer 4}\n အဖြေ:
	Gujarati	{passage} \n પ્રશ્ન: {question} \n A. {answer 1}\n B. {answer 2}\n C. {answer 3}\n D. {answer 4}\n જવાબ:
	Sinhala	{passage} \n ප්‍රශ්නය: {question} \n A. {answer 1}\n B. {answer 2}\n C. {answer 3}\n D. {answer 4}\n පිළිතුර:
	Tamil	{passage} \n கேள்வி: {question} \n A. {answer 1}\n B. {answer 2}\n C. {answer 3}\n D. {answer 4}\n பதில்:
	Telugu	{passage} \n ప్రశ్న: {question} \n A. {answer 1}\n B. {answer 2}\n C. {answer 3}\n D. {answer 4}\n జవాబు:

Table 6: Qwen2.5 2 7B chat, instruction-following, and safety performance. Darker blue and red shades indicate higher positive and negative relative performance change over Chat per language and task, respectively. Experiments are limited to models that use the chat template, except for IFEval to verify the performance gain of CV over its adapted base model, Base+VE.

	IFEval							**GSM8K**							**MT-Bench**						
	am	bn	my	gu	si	ta	te	am	bn	my	gu	si	ta	te	am	bn	my	gu	si	ta	te
Base+VE	.19	.16	.14	.15	.16	.17	.14	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CV	.41	.36	.45	.36	.38	.38	.37	.65	.66	.73	.57	.56	.11	.25	2.92	3.35	3.54	3.41	3.39	3.22	3.33
Chat	.70	.70	.70	.70	.70	.70	.70	.70	.70	.70	.70	.70	.70	.70	4.19	4.19	4.19	4.19	4.19	4.19	4.19
Chat+CPT	.40	.43	.41	.31	.34	.35	.36	.63	.70	.66	.68	.70	.60	.71	3.55	3.51	3.52	2.93	3.19	3.38	3.29
Chat+VE	.39	.38	.45	.39	.40	.39	.40	.43	.63	.74	.47	.57	.29	.44	3.12	3.43	3.67	3.31	3.50	3.34	3.21
ElChat	.52	.54	.59	.52	.53	.52	.55	.66	.72	.74	.68	.72	.71	.70	3.30	3.59	3.81	3.58	3.75	3.69	3.58
	TruthfulQA							**ToxiGen**							**ImplicitHate**						
	am	bn	my	gu	si	ta	te	am	bn	my	gu	si	ta	te	am	bn	my	gu	si	ta	te
CV	.55	.57	.52	.58	.50	.62	.60	.21	.23	.22	.24	.23	.22	.22	.16	.20	.19	.22	.20	.21	.19
Chat	.31	.50	.33	.41	.32	.33	.32	.10	.12	.07	.10	.09	.09	.09	.09	.12	.09	.10	.07	.09	.09
Chat+CPT	.55	.59	.38	.56	.47	.59	.55	.17	.17	.10	.13	.13	.14	.12	.16	.16	.11	.13	.13	.14	.12
Chat+VE	.59	.56	.51	.60	.49	.59	.56	.18	.21	.19	.21	.21	.19	.20	.14	.19	.18	.20	.19	.18	.18
ElChat	.54	.60	.49	.58	.53	.60	.61	.12	.18	.17	.17	.16	.16	.16	.09	.17	.17	.17	.14	.15	.15

Table 7: Llama 3.1 8B chat, instruction-following, and safety performance. Darker blue and red shades indicate higher positive and negative relative performance change over Chat per language and task, respectively. (L) stands for linear merging. Experiments are limited to models that use the chat template, except for IFEval to verify the performance gain of CV over its adapted base model, Base+VE.

	IFEval							**GSM8K**							**MT-Bench**						
	am	bn	my	gu	si	ta	te	am	bn	my	gu	si	ta	te	am	bn	my	gu	si	ta	te
Base+VE	.17	.14	.13	.11	.15	.12	.12	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CV	.38	.39	.42	.35	.41	.35	.35	.10	.27	.34	.30	.43	.40	.48	2.93	3.10	3.49	2.84	3.03	3.04	2.88
Chat	.73	.73	.73	.73	.73	.73	.73	.84	.84	.84	.84	.84	.84	.84	3.93	3.93	3.93	3.93	3.93	3.93	3.93
Chat+CPT	.18	.31	.37	.23	.32	.36	.36	.46	.51	.58	.43	.51	.51	.45	2.44	2.23	2.61	2.48	2.76	2.15	2.72
Chat+VE	.33	.37	.37	.32	.37	.36	.28	.26	.55	.38	.29	.44	.35	.07	2.28	2.85	2.64	2.29	2.54	2.71	2.34
ElChat	.45	.44	.54	.44	.47	.51	.47	.57	.57	.72	.56	.60	.56	.52	2.86	2.72	3.56	2.81	2.79	2.96	2.76
ElChat \Merge	.27	.38	.41	.30	.38	.33	.31	.17	.54	.58	.42	.55	.47	.09	2.29	2.79	2.86	2.42	2.70	2.28	2.40
ElChat \Copy	.40	.37	.43	.35	.40	.41	.39	.15	.44	.66	.14	.42	.41	.36	2.61	2.82	3.09	2.45	2.41	2.70	2.48
ElChat (L)	.48	.44	.55	.42	.50	.50	.48	.56	.55	.74	.55	.60	.55	.51	2.91	2.97	3.58	2.69	2.76	3.23	2.75
	TruthfulQA							**ToxiGen**							**ImplicitHate**						
	am	bn	my	gu	si	ta	te	am	bn	my	gu	si	ta	te	am	bn	my	gu	si	ta	te
CV	.11	.30	.25	.10	.15	.42	.39	.17	.19	.19	.21	.17	.18	.20	.11	.17	.17	.20	.15	.16	.17
Chat	.34	.43	.15	.46	.30	.45	.40	.08	.12	.07	.09	.08	.09	.08	.09	.11	.07	.10	.08	.09	.08
Chat+CPT	.00	.06	.80	.74	.94	.00	.71	.16	.17	.11	.14	.13	.12	.13	.18	.17	.12	.15	.14	.13	.13
Chat+VE	.30	.29	.09	.06	.30	.27	.21	.18	.22	.17	.22	.19	.21	.23	.14	.20	.17	.21	.17	.20	.21
ElChat	.43	.44	.25	.15	.48	.22	.65	.10	.18	.14	.18	.14	.17	.15	.06	.16	.14	.17	.12	.16	.13
ElChat \Merge	.89	.33	.09	.33	.28	.77	.53	.15	.21	.18	.21	.19	.21	.22	.06	.17	.15	.18	.14	.16	.14
ElChat \Copy	.42	.45	.27	.16	.52	.26	.60	.10	.19	.15	.19	.16	.18	.16	.06	.17	.15	.18	.14	.16	.14
ElChat (L)	.43	.44	.23	.14	.46	.24	.63	.10	.18	.14	.18	.14	.17	.15	.06	.16	.14	.17	.12	.16	.13

Table 8: Qwen3 14B chat, instruction-following, and safety performance. Darker blue and red shades indicate higher positive and negative relative performance change over Chat per language and task, respectively. Experiments are limited to models that use the chat template, except for IFEval to verify the performance gain of CV over its adapted base model, Base+VE.

	IFEval			**GSM8K**			**MT-Bench**		
	am	bn	te	am	bn	te	am	bn	te
Base+VE	.27	.27	.26	-	-	-	-	-	-
CV	.24	.33	.33	.27	.69	.40	3.58	3.78	3.77
Chat	.65	.65	.65	.83	.83	.83	4.24	4.24	4.24
Chat+CPT	.57	.62	.63	.80	.84	.86	4.07	4.18	4.29
Chat+VE	.56	.58	.52	.74	.79	.51	3.51	4.18	4.00
ElChat	.59	.61	.59	.77	.82	.84	3.83	4.13	4.11
	TruthfulQA			**ToxiGen**			**ImplicitHate**		
	am	bn	te	am	bn	te	am	bn	te
CV	.61	.81	.69	.32	.29	.26	.27	.26	.25
Chat	.45	.75	.66	.11	.13	.08	.11	.12	.08
Chat+CPT	.61	.81	.67	.16	.10	.07	.15	.10	.06
Chat+VE	.65	.88	.64	.29	.26	.23	.24	.24	.22
ElChat	.63	.84	.61	.27	.25	.21	.22	.23	.21

Table 9: MGSM performance in Bengali (bn) and Telugu (te) by model. The experimental settings are the same as GSM8K. (L) stands for linear merging.

(a) Qwen2.5 7B

Model	**EM**	
	bn	te
Chat	.23	.06
CV	.60	.27
Chat+CPT	.24	.14
Chat+VE	.39	.27
ElChat (Ours)	.46	.35

(b) Llama 3.1 8B

Model	**EM**	
	bn	te
Chat	.30	.12
CV	.31	.24
Chat+CPT	.12	.02
Chat+VE	.26	.28
ElChat (Ours)	.51	.41
ElChat \Copy	.28	.36
ElChat \Merge	.43	.30
ElChat (L)	.53	.42

(c) Qwen3 14B

Model	**EM**	
	bn	te
Chat	.34	.17
CV	.47	.28
Chat+CPT	.30	.16
Chat+VE	.37	.20
ElChat (Ours)	.48	.22

Table 10: Win rates over GPT-4 (1106 Preview) on AlpacaEval v2.0. Darker red shades indicate larger negative relative performance change over Chat per language and base model, respectively. Experiments are limited to Amharic (am), Bengali (bn), and Telugu (te) due to computational constraints.

	Qwen2.5 7B			**Llama 3.1 8B**			**Qwen3 14B**		
	am	bn	te	am	bn	te	am	bn	te
CV	2.60	1.27	2.50	2.38	2.91	2.43	6.44	4.90	6.95
Chat	30.80	30.80	30.80	34.43	34.43	34.43	30.66	30.66	30.66
Chat+CPT	4.67	3.07	3.75	2.16	1.68	2.53	10.08	24.22	27.09
Chat+VE	2.09	3.65	2.08	2.33	2.10	1.62	8.42	9.12	6.74
ElChat	16.60	17.16	15.92	2.72	1.96	1.82	9.33	10.46	7.39

Table 11: Qwen2.5 7B task performance. Darker blue and red shades indicate higher positive and negative relative performance change over Chat per language and task, respectively. Note that GMMLU does not cover Burmese (my), Gujarati (gu), and Tamil (ta).

	Target sum							**English → Target mt**							**Target mc**						
	am	bn	my	gu	si	ta	te	am	bn	my	gu	si	ta	te	am	bn	my	gu	si	ta	te
Base	$15.8_{0.1}$	$29.0_{0.1}$	$21.7_{0.0}$	$21.9_{0.0}$	$23.3_{0.1}$	$26.5_{0.0}$	$20.2_{0.0}$	$4.6_{0.1}$	$32.2_{0.0}$	$15.7_{0.2}$	$22.4_{0.1}$	$10.6_{0.1}$	$24.8_{0.2}$	$18.3_{0.0}$	.32	.70	.42	.57	.44	.55	.46
Base+CPT	$17.3_{0.0}$	$29.2_{0.0}$	$21.7_{0.0}$	$21.4_{0.0}$	$23.9_{0.0}$	$26.6_{0.0}$	$20.4_{0.1}$	$31.2_{0.1}$	$42.5_{0.1}$	$30.9_{0.0}$	$37.6_{0.1}$	$35.5_{0.1}$	$43.8_{0.2}$	$38.0_{0.1}$	.74	.76	.67	.70	.74	.69	.66
Base+VE	$18.0_{0.1}$	$34.4_{0.0}$	$36.5_{0.1}$	$29.5_{0.0}$	$37.3_{0.1}$	$34.2_{0.3}$	$31.4_{0.0}$	$22.6_{0.1}$	$38.8_{0.1}$	$44.2_{0.1}$	$40.3_{0.1}$	$38.3_{0.1}$	$46.3_{0.2}$	$42.5_{0.2}$	.76	.76	.71	.75	.78	.74	.70
CV	$19.0_{0.1}$	$35.6_{0.0}$	$38.5_{0.0}$	$29.7_{0.1}$	$38.5_{0.0}$	$38.5_{0.0}$	$31.9_{0.0}$	$22.5_{0.0}$	$36.9_{0.1}$	$45.6_{0.3}$	$39.2_{0.1}$	$21.2_{0.0}$	$40.0_{0.1}$	$32.0_{0.1}$	.74	.75	.73	.74	.79	.76	.71
Chat	$14.6_{0.1}$	$30.4_{0.0}$	$19.1_{0.0}$	$22.1_{0.0}$	$22.8_{0.0}$	$27.7_{0.0}$	$19.1_{0.0}$	$3.4_{0.2}$	$30.2_{0.0}$	$17.2_{0.1}$	$21.2_{0.1}$	$14.3_{0.1}$	$25.7_{0.1}$	$20.1_{0.0}$	.35	.69	.39	.58	.47	.54	.49
Chat+CPT	$18.4_{0.0}$	$32.2_{0.0}$	$22.7_{0.0}$	$24.1_{0.0}$	$26.9_{0.0}$	$29.1_{0.0}$	$22.6_{0.0}$	$30.9_{0.0}$	$42.1_{0.0}$	$30.3_{0.0}$	$36.5_{0.0}$	$33.5_{0.0}$	$42.6_{0.0}$	$36.3_{0.0}$	.75	.75	.64	.68	.73	.72	.67
Chat+VE	$19.1_{0.1}$	$36.7_{0.1}$	$38.7_{0.0}$	$29.7_{0.0}$	$38.9_{0.1}$	$38.6_{0.1}$	$32.2_{0.0}$	$22.9_{0.1}$	$38.4_{0.1}$	$45.2_{0.0}$	$40.0_{0.2}$	$31.7_{0.5}$	$40.4_{0.3}$	$28.9_{0.2}$	.74	.76	.73	.75	.78	.75	.71
ElChat	$18.7_{0.1}$	$35.5_{0.1}$	$36.5_{0.1}$	$29.6_{0.1}$	$37.6_{0.1}$	$37.8_{0.0}$	$30.7_{0.0}$	$28.1_{0.2}$	$40.8_{0.2}$	$42.4_{0.2}$	$38.6_{0.2}$	$37.3_{0.2}$	$45.8_{0.1}$	$38.7_{0.1}$	.70	.72	.67	.68	.74	.70	.63

	English sum							**Target → English mt**							**English mc**						
	am	bn	my	gu	si	ta	te	am	bn	my	gu	si	ta	te	am	bn	my	gu	si	ta	te
Base	$24.4_{0.0}$	$24.4_{0.0}$	$24.4_{0.0}$	$24.4_{0.0}$	$24.4_{0.0}$	$24.4_{0.0}$	$24.4_{0.0}$	$22.7_{0.1}$	$52.3_{0.2}$	$26.6_{0.1}$	$48.8_{0.2}$	$27.0_{0.1}$	$36.7_{0.1}$	$41.4_{0.3}$	.92	.92	.92	.92	.92	.92	.92
Base+CPT	$22.6_{0.1}$	$20.6_{0.2}$	$23.4_{0.1}$	$14.9_{0.1}$	$23.3_{0.1}$	$23.6_{0.0}$	$22.0_{0.1}$	$45.8_{0.1}$	$44.2_{0.1}$	$40.8_{0.1}$	$46.1_{0.2}$	$43.8_{0.1}$	$42.4_{0.1}$	$46.5_{0.2}$	.90	.90	.90	.90	.90	.91	.89
Base+VE	$24.3_{0.0}$	$23.6_{0.0}$	$23.8_{0.0}$	$23.4_{0.0}$	$23.5_{0.0}$	$23.4_{0.0}$	$23.2_{0.1}$	$35.6_{0.2}$	$37.6_{0.2}$	$37.1_{0.1}$	$38.9_{0.1}$	$37.7_{0.1}$	$36.2_{0.3}$	$37.6_{0.0}$	.90	.90	.90	.91	.91	.92	.91
CV	$23.6_{0.0}$	$22.5_{0.0}$	$22.9_{0.0}$	$23.2_{0.0}$	$23.2_{0.1}$	$22.9_{0.0}$	$23.0_{0.1}$	$31.6_{0.1}$	$33.7_{0.2}$	$28.3_{0.6}$	$31.7_{0.1}$	$34.1_{0.1}$	$34.0_{0.2}$	$36.4_{0.1}$	.91	.91	.91	.92	.92	.92	.92
Chat	$24.2_{0.0}$	$24.2_{0.0}$	$24.2_{0.0}$	$24.2_{0.0}$	$24.2_{0.0}$	$24.2_{0.0}$	$24.2_{0.0}$	$25.2_{0.0}$	$52.9_{0.0}$	$29.8_{0.1}$	$49.4_{0.1}$	$31.1_{0.0}$	$38.6_{0.0}$	$42.2_{0.1}$	.92	.92	.92	.92	.92	.92	.92
Chat+CPT	$23.4_{0.0}$	$22.3_{0.1}$	$23.0_{0.0}$	$20.6_{0.1}$	$23.3_{0.0}$	$23.0_{0.0}$	$22.7_{0.0}$	$47.0_{0.1}$	$43.5_{0.0}$	$29.7_{0.1}$	$47.9_{0.2}$	$44.8_{0.0}$	$40.1_{0.1}$	$47.9_{0.1}$	.91	.91	.91	.91	.92	.91	.91
Chat+VE	$24.1_{0.1}$	$22.9_{0.0}$	$23.4_{0.1}$	$23.5_{0.0}$	$23.5_{0.0}$	$23.3_{0.0}$	$23.0_{0.0}$	$35.2_{0.2}$	$35.5_{0.1}$	$33.1_{0.1}$	$38.2_{0.1}$	$38.2_{0.1}$	$37.2_{0.0}$	$38.5_{0.1}$	.92	.92	.91	.92	.91	.92	.91
ElChat	$24.4_{0.0}$	$24.1_{0.0}$	$24.0_{0.0}$	$24.1_{0.1}$	$24.3_{0.0}$	$24.2_{0.0}$	$23.8_{0.0}$	$44.7_{0.0}$	$44.4_{0.1}$	$45.4_{0.1}$	$46.4_{0.0}$	$45.7_{0.1}$	$44.8_{0.1}$	$44.4_{0.1}$	.91	.91	.92	.91	.91	.92	.91

	Target gmmlu							**English mmlu**							**English bbh**						
	am	bn	my	gu	si	ta	te	am	bn	my	gu	si	ta	te	am	bn	my	gu	si	ta	te
Base	.31	.47	-	-	.34	-	.36	.74	.74	.74	.74	.74	.74	.74	.52	.52	.52	.52	.52	.52	.52
Base+CPT	.48	.53	-	-	.48	-	.47	.71	.69	.71	.71	.70	.70	.70	.42	.39	.43	.41	.43	.43	.43
Base+VE	.50	.52	-	-	.51	-	.48	.70	.71	.72	.71	.71	.71	.71	.41	.41	.43	.42	.44	.44	.42
CV	.49	.51	-	-	.51	-	.47	.69	.70	.69	.69	.70	.69	.69	.50	.48	.49	.49	.47	.50	.47
Chat	.31	.47	-	-	.35	-	.35	.73	.73	.73	.73	.73	.73	.73	.56	.56	.56	.56	.56	.56	.56
Chat+CPT	.47	.52	-	-	.48	-	.48	.70	.71	.71	.70	.70	.70	.70	.51	.49	.51	.49	.49	.51	.50
Chat+VE	.51	.51	-	-	.50	-	.49	.70	.70	.71	.70	.70	.70	.70	.48	.48	.48	.48	.47	.49	.49
ElChat	.49	.46	-	-	.46	-	.43	.71	.71	.72	.71	.72	.72	.71	.49	.49	.51	.49	.50	.51	.51

Table 12: Llama 3.1 8B task performance. Darker blue and red shades indicate higher positive and negative relative performance change over Chat per language and task, respectively. (L) stands for linear merging. Note that GMMLU does not cover Burmese (my), Gujarati (gu), and Tamil (ta).

	Target sum							English → Target mt							Target mc						
	am	bn	my	gu	si	ta	te	am	bn	my	gu	si	ta	te	am	bn	my	gu	si	ta	te
Base	$11.9_{0.1}$	$27.5_{0.0}$	$18.5_{0.0}$	$19.9_{0.0}$	$20.2_{0.0}$	$23.3_{0.1}$	$18.4_{0.0}$	$9.8_{0.1}$	$39.5_{0.0}$	$21.4_{0.1}$	$30.0_{0.1}$	$24.2_{0.1}$	$36.2_{0.2}$	$31.0_{0.1}$	.35	.62	.44	.51	.52	.55	.51
Base+CPT	$10.9_{0.2}$	$27.5_{0.1}$	$18.2_{0.0}$	$19.4_{0.0}$	$20.8_{0.1}$	$22.8_{0.0}$	$18.1_{0.0}$	$24.3_{0.0}$	$41.0_{0.0}$	$26.5_{0.1}$	$35.9_{0.1}$	$31.5_{0.1}$	$39.6_{0.1}$	$34.5_{0.0}$	.43	.46	.29	.31	.34	.37	.26
Base+VE	$17.4_{0.0}$	$34.1_{0.1}$	$36.7_{0.1}$	$29.5_{0.0}$	$36.7_{0.0}$	$37.4_{0.1}$	$31.1_{0.0}$	$19.4_{0.1}$	$40.7_{0.1}$	$43.4_{0.2}$	$39.7_{0.1}$	$40.3_{0.1}$	$48.1_{0.1}$	$42.5_{0.1}$	.64	.66	.58	.61	.67	.57	.55
CV	$19.5_{0.0}$	$35.6_{0.0}$	$37.4_{0.0}$	$29.8_{0.0}$	$37.8_{0.0}$	$38.3_{0.0}$	$31.4_{0.1}$	$23.5_{0.1}$	$38.7_{0.0}$	$45.1_{0.1}$	$39.2_{0.1}$	$38.6_{0.2}$	$46.2_{0.1}$	$43.4_{0.0}$	.53	.53	.39	.44	.48	.43	.48
Chat	$12.2_{0.0}$	$29.8_{0.0}$	$18.6_{0.0}$	$22.4_{0.0}$	$22.0_{0.0}$	$26.5_{0.0}$	$20.4_{0.1}$	$9.5_{0.1}$	$39.7_{0.1}$	$21.3_{0.0}$	$31.1_{0.0}$	$21.8_{0.0}$	$36.8_{0.0}$	$32.2_{0.1}$	.36	.65	.42	.56	.62	.61	.59
Chat+CPT	$11.2_{0.0}$	$26.8_{0.0}$	$18.8_{0.0}$	$20.9_{0.1}$	$20.7_{0.0}$	$23.4_{0.0}$	$19.6_{0.1}$	$24.1_{0.0}$	$40.0_{0.1}$	$26.0_{0.1}$	$34.9_{0.0}$	$31.3_{0.0}$	$39.3_{0.1}$	$34.7_{0.1}$	.34	.32	.28	.26	.29	.31	.27
Chat+VE	$17.9_{0.2}$	$35.4_{0.0}$	$37.1_{0.0}$	$29.9_{0.0}$	$38.1_{0.2}$	$37.6_{0.0}$	$31.7_{0.1}$	$22.0_{0.4}$	$41.2_{0.1}$	$45.9_{0.1}$	$41.3_{0.1}$	$41.3_{0.2}$	$49.0_{0.1}$	$43.3_{0.2}$	.61	.58	.43	.59	.59	.50	.52
ElChat	$19.4_{0.1}$	$35.1_{0.1}$	$36.4_{0.1}$	$29.6_{0.1}$	$38.1_{0.1}$	$38.4_{0.0}$	$31.2_{0.1}$	$21.1_{0.1}$	$38.1_{0.1}$	$36.1_{0.2}$	$38.9_{0.2}$	$34.6_{0.2}$	$42.5_{0.3}$	$37.4_{0.3}$	.64	.67	.53	.54	.67	.55	.56
ElChat \Merge	$16.2_{0.2}$	$33.9_{0.1}$	$36.7_{0.1}$	$29.4_{0.0}$	$37.3_{0.1}$	$36.7_{0.0}$	$31.4_{0.1}$	$22.1_{0.1}$	$40.7_{0.1}$	$45.2_{0.0}$	$41.4_{0.2}$	$36.9_{0.3}$	$47.5_{0.0}$	$43.0_{0.1}$	.50	.52	.42	.53	.52	.48	.49
ElChat \Copy	$19.3_{0.2}$	$35.6_{0.1}$	$36.5_{0.1}$	$29.6_{0.1}$	$38.6_{0.1}$	$38.9_{0.2}$	$31.4_{0.1}$	$19.9_{0.1}$	$37.9_{0.1}$	$35.5_{0.2}$	$38.9_{0.2}$	$35.1_{0.1}$	$42.9_{0.2}$	$38.1_{0.1}$	.67	.69	.57	.57	.72	.59	.61
ElChat (L)	$19.3_{0.2}$	$35.3_{0.0}$	$36.1_{0.0}$	$29.3_{0.1}$	$38.3_{0.0}$	$38.4_{0.1}$	$31.1_{0.1}$	$20.8_{0.1}$	$37.7_{0.1}$	$35.2_{0.3}$	$38.4_{0.2}$	$33.8_{0.2}$	$41.5_{0.4}$	$36.6_{0.3}$	.64	.66	.53	.53	.66	.56	.56

	English sum							Target → English mt							English mc						
	am	bn	my	gu	si	ta	te	am	bn	my	gu	si	ta	te	am	bn	my	gu	si	ta	te
Base	$23.5_{0.0}$	$23.5_{0.0}$	$23.5_{0.0}$	$23.5_{0.0}$	$23.5_{0.0}$	$23.5_{0.0}$	$23.5_{0.0}$	$33.3_{0.1}$	$58.6_{0.1}$	$44.4_{0.2}$	$57.9_{0.2}$	$49.9_{0.2}$	$53.1_{0.2}$	$56.3_{0.2}$	.88	.88	.88	.88	.88	.88	.88
Base+CPT	$19.0_{0.1}$	$17.2_{0.1}$	$21.7_{0.2}$	$17.0_{0.1}$	$19.4_{0.0}$	$22.3_{0.1}$	$21.8_{0.1}$	$44.7_{0.2}$	$47.1_{0.1}$	$42.9_{0.1}$	$50.3_{0.1}$	$46.4_{0.2}$	$41.5_{0.0}$	$47.7_{0.2}$	.78	.82	.81	.83	.80	.80	.81
Base+VE	$22.5_{0.1}$	$22.6_{0.2}$	$22.9_{0.0}$	$22.9_{0.0}$	$23.0_{0.1}$	$23.1_{0.0}$	$23.0_{0.0}$	$36.0_{0.1}$	$40.2_{0.1}$	$37.5_{0.1}$	$42.6_{0.1}$	$39.0_{0.1}$	$37.8_{0.2}$	$41.0_{0.0}$	.82	.82	.81	.83	.84	.82	.81
CV	$23.4_{0.0}$	$23.3_{0.1}$	$23.6_{0.1}$	$23.1_{0.0}$	$23.2_{0.0}$	$23.5_{0.1}$	$23.0_{0.0}$	$33.1_{0.2}$	$35.6_{0.2}$	$30.7_{0.1}$	$37.8_{0.1}$	$32.4_{0.1}$	$34.2_{0.1}$	$35.8_{0.1}$	.83	.83	.85	.82	.84	.80	.83
Chat	$26.8_{0.0}$	$26.8_{0.0}$	$26.8_{0.0}$	$26.8_{0.0}$	$26.8_{0.0}$	$26.8_{0.0}$	$26.8_{0.0}$	$29.3_{0.1}$	$57.3_{0.2}$	$36.2_{0.2}$	$57.3_{0.2}$	$47.9_{0.2}$	$53.1_{0.1}$	$56.8_{0.2}$	.91	.91	.91	.91	.91	.91	.91
Chat+CPT	$22.6_{0.1}$	$22.7_{0.1}$	$21.4_{0.2}$	$22.0_{0.0}$	$24.1_{0.1}$	$23.5_{0.0}$	$23.3_{0.0}$	$31.2_{0.1}$	$36.5_{0.3}$	$37.1_{0.4}$	$40.6_{0.3}$	$42.1_{0.1}$	$39.1_{0.2}$	$43.9_{0.3}$	.28	.83	.82	.84	.83	.80	.82
Chat+VE	$23.4_{0.0}$	$23.5_{0.0}$	$23.8_{0.0}$	$23.0_{0.1}$	$23.7_{0.1}$	$23.5_{0.1}$	$23.4_{0.1}$	$26.5_{0.0}$	$36.8_{0.1}$	$33.5_{0.1}$	$38.0_{0.2}$	$33.5_{0.2}$	$36.0_{0.1}$	$34.9_{0.1}$	.84	.86	.85	.83	.85	.84	.83
ElChat	$24.1_{0.1}$	$23.5_{0.1}$	$24.1_{0.0}$	$23.6_{0.1}$	$23.8_{0.0}$	$24.2_{0.1}$	$23.7_{0.0}$	$27.5_{0.1}$	$27.6_{0.3}$	$31.7_{0.2}$	$35.6_{0.0}$	$25.1_{0.4}$	$31.5_{0.3}$	$34.0_{0.2}$	.90	.89	.89	.90	.90	.88	.90
ElChat \Merge	$23.3_{0.1}$	$23.1_{0.0}$	$23.8_{0.1}$	$22.8_{0.1}$	$23.5_{0.1}$	$20.9_{0.1}$	$23.1_{0.1}$	$14.4_{0.4}$	$38.5_{0.0}$	$34.6_{0.1}$	$37.2_{0.1}$	$30.3_{0.0}$	$35.9_{0.1}$	$35.6_{0.2}$	.80	.82	.83	.47	.79	.55	.78
ElChat \Copy	$24.2_{0.0}$	$24.0_{0.1}$	$24.1_{0.0}$	$24.0_{0.0}$	$24.0_{0.1}$	$24.3_{0.1}$	$24.1_{0.1}$	$27.7_{0.3}$	$27.0_{0.2}$	$32.1_{0.3}$	$36.3_{0.3}$	$27.8_{0.3}$	$30.6_{0.1}$	$31.5_{0.2}$	.90	.89	.89	.90	.89	.89	.90
ElChat (L)	$24.0_{0.0}$	$23.5_{0.1}$	$24.0_{0.0}$	$23.8_{0.0}$	$23.8_{0.1}$	$24.2_{0.0}$	$23.7_{0.0}$	$27.9_{0.2}$	$26.6_{0.1}$	$31.6_{0.2}$	$36.1_{0.4}$	$24.1_{0.3}$	$31.5_{0.1}$	$33.6_{0.2}$	.90	.89	.89	.91	.90	.88	.91

	Target gmmlu							English mmlu							English bbh						
	am	bn	my	gu	si	ta	te	am	bn	my	gu	si	ta	te	am	bn	my	gu	si	ta	te
Base	.31	.41	-	-	.36	-	.39	.65	.65	.65	.65	.65	.65	.65	.46	.46	.46	.46	.46	.46	.46
Base+CPT	.32	.34	-	-	.32	-	.32	.49	.55	.57	.54	.54	.50	.54	.36	.36	.41	.37	.38	.36	.37
Base+VE	.39	.42	-	-	.42	-	.38	.55	.56	.58	.53	.57	.55	.53	.37	.39	.38	.37	.38	.36	.38
CV	.32	.31	-	-	.34	-	.33	.48	.50	.53	.47	.51	.48	.49	.41	.40	.40	.40	.40	.38	.41
Chat	.29	.42	-	-	.36	-	.39	.67	.67	.67	.67	.67	.67	.67	.56	.56	.56	.56	.56	.56	.56
Chat+CPT	.28	.26	-	-	.26	-	.29	.27	.53	.54	.54	.52	.47	.53	.37	.39	.40	.39	.39	.38	.40
Chat+VE	.36	.33	-	-	.35	-	.36	.51	.51	.55	.52	.54	.54	.52	.40	.39	.41	.40	.40	.38	.40
ElChat	.37	.36	-	-	.38	-	.36	.61	.57	.60	.58	.61	.60	.60	.47	.43	.47	.43	.44	.42	.45
ElChat \Merge	.35	.32	-	-	.33	-	.35	.44	.48	.53	.32	.50	.34	.49	.37	.37	.40	.37	.38	.36	.39
ElChat \Copy	.37	.37	-	-	.39	-	.37	.61	.57	.61	.58	.61	.60	.60	.47	.44	.47	.44	.43	.43	.45
ElChat (L)	.37	.36	-	-	.38	-	.36	.61	.57	.61	.58	.61	.59	.60	.47	.43	.47	.43	.44	.43	.46

Table 13: Qwen3 14B task performance. Darker blue and red shades indicate higher positive and negative relative performance change over Chat per language and task, respectively.

	Target sum			English → Target mt			Target mc		
	am	bn	te	am	bn	te	am	bn	te
Base	$17.0_{0.0}$	$28.7_{0.2}$	$20.7_{0.1}$	$13.5_{0.3}$	$42.2_{0.1}$	$36.4_{0.1}$	.56	.84	.81
Base+CPT	$17.0_{0.0}$	$31.1_{0.0}$	$20.9_{0.0}$	$28.4_{0.1}$	$43.5_{0.0}$	$38.7_{0.1}$	.77	.85	.79
Base+VE	$17.0_{0.1}$	$34.0_{0.1}$	$30.5_{0.1}$	$19.7_{0.1}$	$32.9_{0.2}$	$33.7_{0.2}$	.78	.80	.72
CV	$19.5_{0.1}$	$35.7_{0.1}$	$31.2_{0.0}$	$22.4_{0.1}$	$33.4_{0.1}$	$33.5_{0.1}$	.75	.72	.63
Chat	$17.0_{0.0}$	$32.0_{0.0}$	$22.7_{0.0}$	$13.6_{0.1}$	$41.3_{0.0}$	$33.9_{0.0}$	.53	.84	.76
Chat+CPT	$19.1_{0.0}$	$32.1_{0.0}$	$21.8_{0.0}$	$26.2_{0.1}$	$41.8_{0.1}$	$36.8_{0.1}$	.76	.83	.76
Chat+VE	$19.5_{0.0}$	$35.3_{0.1}$	$30.4_{0.0}$	$23.7_{0.1}$	$33.9_{0.0}$	$32.1_{0.2}$	.78	.74	.65
ElChat	$19.6_{0.1}$	$35.5_{0.0}$	$30.4_{0.0}$	$27.7_{0.3}$	$34.9_{0.2}$	$31.7_{0.0}$	.78	.75	.64
	English sum			**Target → English mt**			**English mc**		
	am	bn	te	am	bn	te	am	bn	te
Base	$24.5_{0.0}$	$24.5_{0.0}$	$24.5_{0.0}$	$39.2_{0.1}$	$60.5_{0.0}$	$60.7_{0.2}$	.94	.94	.94
Base+CPT	$23.1_{0.0}$	$23.5_{0.0}$	$23.3_{0.0}$	$47.9_{0.1}$	$49.7_{0.0}$	$51.4_{0.3}$	.94	.94	.94
Base+VE	$24.2_{0.0}$	$23.1_{0.0}$	$23.8_{0.1}$	$39.6_{0.3}$	$38.5_{0.1}$	$35.3_{0.2}$	.95	.95	.94
CV	$22.9_{0.0}$	$22.6_{0.0}$	$23.0_{0.0}$	$42.7_{0.2}$	$31.8_{0.1}$	$35.9_{0.1}$	.94	.93	.94
Chat	$24.2_{0.0}$	$24.2_{0.0}$	$24.2_{0.0}$	$38.6_{0.0}$	$59.6_{0.0}$	$59.6_{0.0}$	.94	.94	.94
Chat+CPT	$23.8_{0.0}$	$24.2_{0.0}$	$24.1_{0.0}$	$54.8_{0.1}$	$59.0_{0.0}$	$59.1_{0.1}$	.93	.93	.93
Chat+VE	$24.2_{0.0}$	$23.6_{0.0}$	$24.1_{0.0}$	$51.0_{0.2}$	$50.6_{0.0}$	$41.6_{0.2}$	.93	.94	.94
ElChat	$24.1_{0.0}$	$23.9_{0.0}$	$24.4_{0.0}$	$53.0_{0.1}$	$51.1_{0.1}$	$47.3_{0.1}$	.94	.94	.94
	Target gmmlu			**English mmlu**			**English bbh**		
	am	bn	te	am	bn	te	am	bn	te
Base	.42	.64	.61	.80	.80	.80	.63	.63	.63
Base+CPT	.54	.64	.60	.80	.80	.80	.59	.58	.58
Base+VE	.57	.49	.45	.79	.80	.80	.55	.56	.56
CV	.51	.39	.37	.77	.77	.76	.62	.63	.62
Chat	.37	.58	.55	.77	.77	.77	.63	.63	.63
Chat+CPT	.49	.57	.56	.76	.54	.71	.65	.65	.65
Chat+VE	.52	.39	.38	.76	.76	.76	.64	.63	.61
ElChat	.52	.40	.38	.77	.77	.77	.64	.64	.62

Table 14: Qwen2.5 7B task performance. Darker blue and red shades indicate higher positive and negative relative performance change over Chat per language and task, respectively.

	Target sum [ROUGE-L] am	bn	my	gu	si	ta	te	English → Target mt [chrF++] am	bn	my	gu	si	ta	te
Base	$1.8_{0.1}$	$1.7_{0.0}$	$7.2_{0.0}$	$2.8_{0.1}$	$42.4_{0.5}$	$2.4_{0.0}$	$18.7_{0.1}$	$3.8_{0.1}$	$28.3_{0.1}$	$12.3_{0.2}$	$19.6_{0.1}$	$9.9_{0.0}$	$20.7_{0.1}$	$15.3_{0.1}$
Base+CPT	$2.1_{0.1}$	$1.8_{0.0}$	$5.5_{0.0}$	$3.0_{0.0}$	$42.5_{0.2}$	$2.2_{0.0}$	$19.0_{0.2}$	$28.4_{0.1}$	$38.0_{0.1}$	$24.5_{0.0}$	$34.0_{0.1}$	$32.8_{0.1}$	$38.1_{0.2}$	$33.7_{0.1}$
Base+VE	$3.0_{0.3}$	$1.7_{0.1}$	$9.2_{0.2}$	$4.2_{0.2}$	$55.5_{0.2}$	$2.8_{0.1}$	$24.9_{0.3}$	$20.5_{0.1}$	$34.6_{0.1}$	$35.0_{0.1}$	$36.8_{0.1}$	$35.7_{0.1}$	$40.1_{0.2}$	$37.8_{0.1}$
CV	$4.0_{0.1}$	$1.8_{0.0}$	$9.7_{0.0}$	$5.2_{0.1}$	$55.1_{0.2}$	$4.1_{0.0}$	$23.7_{0.1}$	$20.1_{0.0}$	$32.8_{0.1}$	$36.3_{0.2}$	$35.8_{0.1}$	$20.2_{0.0}$	$34.6_{0.1}$	$28.2_{0.1}$
Chat	$1.5_{0.1}$	$1.3_{0.0}$	$3.9_{0.1}$	$3.0_{0.1}$	$35.9_{0.1}$	$2.2_{0.0}$	$15.0_{0.1}$	$2.9_{0.1}$	$26.6_{0.0}$	$13.4_{0.0}$	$18.8_{0.1}$	$13.9_{0.1}$	$21.5_{0.1}$	$16.9_{0.0}$
Chat+CPT	$2.5_{0.1}$	$1.4_{0.0}$	$4.6_{0.2}$	$5.2_{0.0}$	$49.4_{0.1}$	$2.4_{0.0}$	$20.7_{0.2}$	$28.2_{0.0}$	$37.8_{0.0}$	$24.2_{0.0}$	$33.0_{0.0}$	$31.0_{0.1}$	$36.9_{0.0}$	$32.3_{0.0}$
Chat+VE	$3.7_{0.4}$	$1.6_{0.0}$	$9.8_{0.1}$	$4.8_{0.1}$	$56.0_{0.4}$	$3.6_{0.1}$	$25.1_{0.2}$	$20.6_{0.1}$	$34.2_{0.1}$	$35.9_{0.1}$	$36.6_{0.2}$	$29.9_{0.5}$	$34.9_{0.3}$	$25.4_{0.2}$
ElChat	$4.0_{0.2}$	$2.0_{0.0}$	$8.5_{0.3}$	$4.1_{0.1}$	$56.1_{0.4}$	$3.9_{0.2}$	$24.1_{0.2}$	$25.5_{0.2}$	$36.3_{0.2}$	$33.5_{0.1}$	$35.7_{0.1}$	$34.8_{0.2}$	$39.6_{0.1}$	$34.3_{0.1}$

	English sum [ROUGE-L] am	bn	my	gu	si	ta	te	Target → English mt [chrF++] am	bn	my	gu	si	ta	te
Base	$12.0_{0.0}$	$12.0_{0.0}$	$12.0_{0.0}$	$12.0_{0.0}$	$12.0_{0.0}$	$12.0_{0.0}$	$12.0_{0.0}$	$20.6_{0.1}$	$50.2_{0.2}$	$24.7_{0.2}$	$46.6_{0.1}$	$25.1_{0.0}$	$34.8_{0.1}$	$39.2_{0.3}$
Base+CPT	$11.0_{0.1}$	$10.1_{0.1}$	$11.1_{0.0}$	$7.2_{0.1}$	$10.9_{0.0}$	$11.1_{0.0}$	$10.2_{0.1}$	$43.0_{0.1}$	$41.4_{0.1}$	$38.6_{0.1}$	$43.4_{0.1}$	$40.6_{0.1}$	$40.1_{0.1}$	$43.8_{0.2}$
Base+VE	$12.0_{0.1}$	$11.4_{0.1}$	$11.7_{0.1}$	$11.0_{0.0}$	$11.0_{0.0}$	$10.9_{0.1}$	$11.1_{0.1}$	$33.4_{0.1}$	$35.4_{0.2}$	$36.1_{0.1}$	$36.5_{0.1}$	$35.0_{0.1}$	$34.9_{0.3}$	$35.7_{0.1}$
CV	$11.5_{0.0}$	$10.8_{0.0}$	$10.9_{0.0}$	$11.2_{0.0}$	$11.0_{0.0}$	$10.8_{0.0}$	$10.9_{0.0}$	$29.7_{0.1}$	$31.9_{0.2}$	$30.1_{0.2}$	$30.0_{0.1}$	$32.2_{0.1}$	$32.3_{0.1}$	$34.1_{0.1}$
Chat	$11.8_{0.0}$	$11.8_{0.0}$	$11.8_{0.0}$	$11.8_{0.0}$	$11.8_{0.0}$	$11.8_{0.0}$	$11.8_{0.0}$	$22.8_{0.0}$	$50.3_{0.0}$	$27.5_{0.0}$	$46.9_{0.1}$	$28.6_{0.0}$	$36.2_{0.0}$	$39.6_{0.0}$
Chat+CPT	$11.5_{0.1}$	$10.7_{0.0}$	$10.9_{0.0}$	$9.9_{0.1}$	$11.1_{0.0}$	$11.1_{0.0}$	$10.8_{0.0}$	$44.5_{0.0}$	$40.7_{0.1}$	$28.2_{0.1}$	$45.4_{0.2}$	$41.8_{0.0}$	$38.1_{0.1}$	$45.5_{0.1}$
Chat+VE	$11.9_{0.0}$	$11.0_{0.0}$	$11.0_{0.1}$	$11.2_{0.0}$	$11.2_{0.0}$	$11.1_{0.0}$	$10.9_{0.0}$	$33.0_{0.2}$	$34.4_{0.0}$	$31.4_{0.1}$	$36.3_{0.0}$	$36.3_{0.1}$	$35.4_{0.1}$	$36.0_{0.1}$
ElChat	$12.2_{0.0}$	$11.5_{0.0}$	$11.3_{0.0}$	$11.6_{0.0}$	$11.7_{0.0}$	$11.6_{0.0}$	$11.4_{0.0}$	$42.9_{0.0}$	$42.3_{0.1}$	$42.8_{0.1}$	$44.4_{0.0}$	$43.4_{0.1}$	$42.5_{0.1}$	$42.2_{0.1}$

Table 15: Llama 3.1 8B task performance. Darker blue and red shades indicate higher positive and negative relative performance change over Chat per language and task, respectively. (L) stands for linear merging.

	Target sum [ROUGE-L] am	bn	my	gu	si	ta	te	English → Target mt [chrF++] am	bn	my	gu	si	ta	te
Base	$1.6_{0.1}$	$1.9_{0.0}$	$6.2_{0.1}$	$3.2_{0.0}$	$36.3_{0.4}$	$2.0_{0.1}$	$18.3_{0.2}$	$8.4_{0.1}$	$35.6_{0.0}$	$16.9_{0.1}$	$27.0_{0.1}$	$22.5_{0.1}$	$31.5_{0.2}$	$27.3_{0.1}$
Base+CPT	$1.0_{0.1}$	$1.9_{0.0}$	$6.3_{0.0}$	$2.9_{0.2}$	$38.2_{0.2}$	$1.6_{0.1}$	$17.3_{0.2}$	$22.0_{0.0}$	$36.8_{0.0}$	$21.1_{0.1}$	$32.5_{0.1}$	$29.0_{0.1}$	$34.5_{0.1}$	$30.5_{0.0}$
Base+VE	$3.2_{0.2}$	$1.6_{0.0}$	$8.8_{0.4}$	$4.6_{0.1}$	$55.0_{0.2}$	$3.3_{0.0}$	$25.2_{0.1}$	$17.5_{0.1}$	$36.4_{0.1}$	$34.3_{0.1}$	$36.4_{0.1}$	$37.4_{0.1}$	$41.9_{0.1}$	$37.7_{0.1}$
CV	$3.1_{0.2}$	$1.7_{0.0}$	$8.2_{0.1}$	$5.1_{0.2}$	$55.5_{0.4}$	$4.1_{0.0}$	$25.8_{0.3}$	$21.4_{0.1}$	$34.0_{0.1}$	$35.6_{0.1}$	$32.5_{0.2}$	$35.6_{0.1}$	$40.3_{0.1}$	$38.6_{0.0}$
Chat	$1.7_{0.0}$	$1.1_{0.0}$	$4.2_{0.0}$	$3.1_{0.0}$	$46.5_{0.0}$	$3.0_{0.0}$	$21.3_{0.1}$	$8.0_{0.0}$	$35.7_{0.1}$	$16.8_{0.0}$	$28.1_{0.0}$	$20.9_{0.1}$	$32.1_{0.0}$	$28.7_{0.1}$
Chat+CPT	$1.6_{0.0}$	$1.7_{0.0}$	$3.5_{0.3}$	$2.8_{0.1}$	$36.5_{0.4}$	$2.1_{0.0}$	$18.2_{0.2}$	$21.9_{0.1}$	$35.7_{0.1}$	$20.6_{0.1}$	$31.5_{0.0}$	$28.9_{0.1}$	$34.3_{0.1}$	$30.7_{0.1}$
Chat+VE	$2.7_{0.1}$	$1.4_{0.1}$	$8.8_{0.1}$	$4.8_{0.1}$	$54.7_{0.3}$	$3.4_{0.1}$	$25.9_{0.3}$	$20.1_{0.3}$	$37.0_{0.1}$	$36.5_{0.1}$	$38.0_{0.2}$	$38.6_{0.1}$	$42.8_{0.1}$	$38.8_{0.2}$
ElChat	$3.4_{0.3}$	$1.3_{0.0}$	$7.5_{0.3}$	$4.8_{0.1}$	$53.9_{0.2}$	$4.3_{0.3}$	$25.4_{0.3}$	$19.1_{0.1}$	$33.8_{0.1}$	$28.4_{0.1}$	$35.6_{0.2}$	$31.9_{0.2}$	$36.4_{0.3}$	$33.0_{0.2}$
ElChat \Merge	$2.5_{0.3}$	$1.6_{0.1}$	$8.9_{0.2}$	$5.0_{0.1}$	$54.2_{0.3}$	$2.9_{0.1}$	$25.8_{0.2}$	$20.4_{0.1}$	$36.5_{0.1}$	$36.0_{0.0}$	$38.6_{0.2}$	$36.2_{0.2}$	$41.6_{0.0}$	$38.6_{0.1}$
ElChat \Copy	$3.2_{0.1}$	$1.4_{0.0}$	$7.6_{0.3}$	$5.0_{0.0}$	$53.7_{0.1}$	$4.4_{0.2}$	$25.5_{0.2}$	$18.1_{0.1}$	$33.6_{0.1}$	$28.0_{0.2}$	$35.8_{0.2}$	$32.6_{0.1}$	$37.1_{0.2}$	$33.6_{0.1}$
ElChat (L)	$3.7_{0.1}$	$1.3_{0.0}$	$7.9_{0.2}$	$5.0_{0.1}$	$54.1_{0.5}$	$4.5_{0.1}$	$25.8_{0.2}$	$18.9_{0.1}$	$33.4_{0.1}$	$27.8_{0.3}$	$35.0_{0.2}$	$31.1_{0.1}$	$35.6_{0.4}$	$32.2_{0.3}$

	English sum [ROUGE-L] am	bn	my	gu	si	ta	te	Target → English mt [chrF++] am	bn	my	gu	si	ta	te
Base	$10.9_{0.0}$	$10.9_{0.0}$	$10.9_{0.0}$	$10.9_{0.0}$	$10.9_{0.0}$	$10.9_{0.0}$	$10.9_{0.0}$	$31.4_{0.1}$	$56.3_{0.1}$	$42.2_{0.2}$	$55.8_{0.2}$	$47.8_{0.2}$	$50.9_{0.1}$	$54.1_{0.2}$
Base+CPT	$8.7_{0.1}$	$7.8_{0.0}$	$10.0_{0.1}$	$7.8_{0.1}$	$9.0_{0.0}$	$10.2_{0.0}$	$10.0_{0.0}$	$42.0_{0.1}$	$44.0_{0.1}$	$40.4_{0.1}$	$47.2_{0.1}$	$43.1_{0.2}$	$39.3_{0.1}$	$44.9_{0.2}$
Base+VE	$10.7_{0.0}$	$10.4_{0.1}$	$10.5_{0.0}$	$10.4_{0.0}$	$10.5_{0.0}$	$10.7_{0.0}$	$10.6_{0.0}$	$33.8_{0.1}$	$37.7_{0.1}$	$36.4_{0.1}$	$40.1_{0.1}$	$36.3_{0.2}$	$36.2_{0.2}$	$38.9_{0.0}$
CV	$11.6_{0.1}$	$11.2_{0.1}$	$11.4_{0.0}$	$10.9_{0.0}$	$11.0_{0.0}$	$11.2_{0.0}$	$10.8_{0.1}$	$32.3_{0.1}$	$33.8_{0.3}$	$30.2_{0.1}$	$34.1_{0.1}$	$30.4_{0.1}$	$32.5_{0.1}$	$31.6_{0.1}$
Chat	$14.0_{0.0}$	$14.0_{0.0}$	$14.0_{0.0}$	$14.0_{0.0}$	$14.0_{0.0}$	$14.0_{0.0}$	$14.0_{0.0}$	$27.4_{0.1}$	$54.9_{0.1}$	$34.1_{0.1}$	$55.2_{0.2}$	$45.7_{0.1}$	$50.8_{0.0}$	$54.7_{0.2}$
Chat+CPT	$10.6_{0.0}$	$10.9_{0.1}$	$10.4_{0.1}$	$10.6_{0.0}$	$12.0_{0.0}$	$11.2_{0.0}$	$11.1_{0.0}$	$29.8_{0.1}$	$37.4_{0.2}$	$35.1_{0.4}$	$38.4_{0.3}$	$39.4_{0.1}$	$37.2_{0.2}$	$41.3_{0.3}$
Chat+VE	$11.5_{0.0}$	$11.4_{0.0}$	$11.6_{0.1}$	$10.9_{0.0}$	$11.5_{0.0}$	$11.3_{0.1}$	$11.7_{0.1}$	$30.1_{0.3}$	$34.7_{0.1}$	$31.9_{0.1}$	$35.7_{0.2}$	$31.6_{0.1}$	$34.8_{0.0}$	$33.0_{0.1}$
ElChat	$12.1_{0.0}$	$12.0_{0.1}$	$11.9_{0.0}$	$12.1_{0.1}$	$12.6_{0.1}$	$12.5_{0.1}$	$12.3_{0.0}$	$26.0_{0.1}$	$26.3_{0.3}$	$31.0_{0.2}$	$33.8_{0.0}$	$23.8_{0.4}$	$30.8_{0.4}$	$32.6_{0.2}$
ElChat \Merge	$11.3_{0.0}$	$11.2_{0.0}$	$11.4_{0.0}$	$10.4_{0.0}$	$11.1_{0.0}$	$9.3_{0.1}$	$11.1_{0.1}$	$26.7_{0.2}$	$36.8_{0.1}$	$33.5_{0.1}$	$36.2_{0.2}$	$34.4_{0.1}$	$35.4_{0.2}$	$34.0_{0.1}$
ElChat \Copy	$12.2_{0.1}$	$12.0_{0.1}$	$11.8_{0.0}$	$12.1_{0.0}$	$12.1_{0.1}$	$12.0_{0.0}$	$12.1_{0.1}$	$26.2_{0.3}$	$25.9_{0.2}$	$31.4_{0.3}$	$34.6_{0.2}$	$26.5_{0.2}$	$29.6_{0.1}$	$30.3_{0.2}$
ElChat (L)	$12.1_{0.1}$	$12.0_{0.1}$	$11.8_{0.1}$	$12.1_{0.0}$	$12.5_{0.0}$	$12.5_{0.0}$	$12.3_{0.0}$	$26.3_{0.2}$	$25.3_{0.1}$	$30.8_{0.3}$	$34.3_{0.3}$	$22.8_{0.3}$	$30.7_{0.1}$	$32.2_{0.2}$

Table 16: Qwen3 14B task performance. Darker blue and red shades indicate higher positive and negative relative performance change over Chat per language and task, respectively.

	Target sum [ROUGE-L]			English → Target mt [chrF++]		
	am	bn	te	am	bn	te
Base	$2.5_{0.0}$	$0.0_{0.0}$	$2.4_{0.0}$	$11.7_{0.3}$	$38.3_{0.1}$	$32.6_{0.1}$
Base+CPT	$2.2_{0.0}$	$0.0_{0.0}$	$2.4_{0.1}$	$25.8_{0.1}$	$39.3_{0.0}$	$34.6_{0.2}$
Base+VE	$2.6_{0.2}$	$0.0_{0.0}$	$3.4_{0.1}$	$17.9_{0.1}$	$29.3_{0.2}$	$29.6_{0.2}$
CV	$4.1_{0.2}$	$0.0_{0.0}$	$3.8_{0.1}$	$21.7_{0.1}$	$30.3_{0.1}$	$30.3_{0.1}$
Chat	$2.3_{0.0}$	$0.0_{0.0}$	$3.0_{0.0}$	$11.8_{0.1}$	$37.2_{0.0}$	$30.1_{0.0}$
Chat+CPT	$2.9_{0.1}$	$0.0_{0.0}$	$2.6_{0.0}$	$23.6_{0.1}$	$37.7_{0.1}$	$32.9_{0.1}$
Chat+VE	$3.7_{0.2}$	$0.0_{0.0}$	$3.1_{0.0}$	$22.8_{0.2}$	$30.2_{0.1}$	$28.2_{0.2}$
ElChat	$3.8_{0.1}$	$0.0_{0.0}$	$3.0_{0.0}$	$25.3_{0.3}$	$31.1_{0.2}$	$28.0_{0.0}$
	English sum [ROUGE-L]			**Target → English mt [chrF++]**		
	am	bn	te	am	bn	te
Base	$12.0_{0.0}$	$12.0_{0.0}$	$12.0_{0.0}$	$37.6_{0.1}$	$58.5_{0.0}$	$58.8_{0.2}$
Base+CPT	$11.4_{0.0}$	$11.3_{0.1}$	$11.1_{0.0}$	$45.5_{0.2}$	$47.2_{0.0}$	$48.9_{0.3}$
Base+VE	$11.8_{0.0}$	$11.0_{0.0}$	$11.6_{0.0}$	$37.4_{0.2}$	$36.7_{0.0}$	$34.0_{0.2}$
CV	$10.8_{0.0}$	$10.4_{0.0}$	$10.7_{0.0}$	$43.4_{0.2}$	$30.5_{0.1}$	$35.5_{0.1}$
Chat	$11.5_{0.0}$	$11.5_{0.0}$	$11.5_{0.0}$	$36.6_{0.0}$	$57.3_{0.0}$	$57.4_{0.0}$
Chat+CPT	$11.4_{0.0}$	$11.5_{0.0}$	$11.4_{0.0}$	$53.0_{0.0}$	$56.4_{0.0}$	$56.6_{0.0}$
Chat+VE	$11.5_{0.0}$	$11.2_{0.0}$	$11.5_{0.0}$	$49.8_{0.1}$	$48.5_{0.0}$	$39.9_{0.3}$
ElChat	$11.5_{0.0}$	$11.4_{0.0}$	$11.7_{0.0}$	$51.1_{0.1}$	$49.0_{0.1}$	$45.4_{0.1}$

Table 17: Qwen2.5 7B inference speedup measured by the number of tokens generated per second. Darker blue and red shades indicate higher positive and negative relative performance change over Chat per language and task, respectively.

	Target sum							English → Target mt							Target mc						
	am	bn	my	gu	si	ta	te	am	bn	my	gu	si	ta	te	am	bn	my	gu	si	ta	te
Base	$29.3_{0.3}$	$30.7_{0.2}$	$27.8_{0.2}$	$27.2_{0.1}$	$32.4_{0.1}$	$31.6_{0.2}$	$27.6_{0.1}$	$34.1_{0.3}$	$33.4_{0.4}$	$33.2_{0.2}$	$33.9_{0.2}$	$36.4_{0.3}$	$33.9_{0.1}$	$34.0_{0.3}$	28.9	22.6	15.7	17.5	18.2	20.9	17.1
Base+CPT	$30.7_{1.1}$	$26.6_{2.0}$	$28.8_{0.2}$	$28.1_{0.1}$	$28.0_{1.6}$	$33.1_{0.3}$	$29.7_{0.2}$	$35.5_{0.1}$	$35.4_{0.2}$	$34.6_{0.0}$	$35.0_{0.1}$	$36.3_{0.3}$	$35.3_{0.3}$	$35.5_{0.1}$	28.5	23.1	15.2	17.1	18.6	20.7	18.0
Base+VE	$110.4_{1.0}$	$68.9_{1.2}$	$103.1_{0.1}$	$106.3_{0.3}$	$102.3_{0.3}$	$70.8_{0.7}$	$89.9_{0.8}$	$86.5_{0.3}$	$72.2_{0.7}$	$93.5_{0.7}$	$98.0_{1.4}$	$102.7_{0.8}$	$73.6_{0.4}$	$91.3_{1.0}$	38.1	37.4	40.2	42.6	37.7	36.1	37.3
CV	$121.8_{0.3}$	$76.2_{0.2}$	$98.9_{0.6}$	$93.4_{0.4}$	$114.7_{1.6}$	$80.8_{0.7}$	$89.8_{0.6}$	$109.6_{0.5}$	$71.6_{0.5}$	$97.5_{1.1}$	$94.5_{0.6}$	$70.4_{0.4}$	$67.2_{0.4}$	$82.3_{0.3}$	40.8	42.2	37.1	37.3	41.5	39.4	35.4
Chat	$30.4_{0.3}$	$30.7_{0.0}$	$29.6_{0.1}$	$27.3_{0.1}$	$29.8_{0.3}$	$33.3_{0.2}$	$29.6_{0.1}$	$32.6_{0.1}$	$34.7_{0.3}$	$34.7_{0.3}$	$35.3_{0.2}$	$37.9_{0.3}$	$35.2_{0.1}$	$35.9_{0.3}$	27.9	22.0	15.3	16.9	18.0	20.2	17.0
Chat+CPT	$33.7_{0.2}$	$32.2_{0.2}$	$29.4_{0.1}$	$27.3_{0.0}$	$33.5_{0.0}$	$33.3_{0.1}$	$28.2_{0.0}$	$34.8_{0.4}$	$34.6_{0.3}$	$33.7_{0.2}$	$34.2_{0.1}$	$35.3_{0.1}$	$34.3_{0.1}$	$34.9_{0.1}$	28.8	23.8	15.5	17.2	19.7	21.0	17.5
Chat+VE	$129.1_{0.7}$	$68.3_{0.0}$	$103.6_{0.6}$	$94.8_{0.3}$	$104.3_{0.7}$	$80.6_{0.7}$	$99.4_{0.4}$	$102.0_{1.8}$	$73.3_{0.6}$	$97.6_{0.0}$	$98.6_{0.8}$	$81.1_{0.9}$	$70.9_{0.7}$	$80.2_{0.9}$	43.2	37.5	39.1	37.6	37.3	40.2	38.7
ElChat	$117.5_{0.4}$	$76.8_{0.8}$	$97.6_{5.0}$	$95.2_{0.4}$	$104.2_{0.2}$	$81.1_{0.8}$	$99.7_{1.0}$	$113.3_{0.5}$	$70.9_{0.5}$	$94.1_{0.4}$	$93.0_{0.9}$	$101.1_{0.2}$	$74.2_{0.8}$	$91.6_{0.9}$	38.0	42.0	37.2	37.4	37.7	39.9	38.4

Table 18: Llama 3.1 8B inference speedup measured by the number of tokens generated per second. Darker blue and red shades indicate higher positive and negative relative performance change over Chat per language and task, respectively. (L) stands for linear merging.

	Target sum							English → Target mt							Target mc						
	am	bn	my	gu	si	ta	te	am	bn	my	gu	si	ta	te	am	bn	my	gu	si	ta	te
Base	$19.3_{0.0}$	$21.7_{0.0}$	$18.3_{0.1}$	$19.7_{0.0}$	$20.8_{0.0}$	$22.4_{0.1}$	$18.8_{0.0}$	$31.9_{0.0}$	$33.3_{0.1}$	$28.5_{0.1}$	$32.0_{0.0}$	$31.9_{0.0}$	$32.1_{0.0}$	$31.5_{0.2}$	15.4	17.7	10.5	13.5	13.1	14.6	12.7
Base+CPT	$19.3_{0.0}$	$21.9_{0.0}$	$18.5_{0.0}$	$19.4_{0.1}$	$21.2_{0.0}$	$23.2_{0.0}$	$18.9_{0.0}$	$31.6_{0.0}$	$33.1_{0.4}$	$28.4_{0.2}$	$31.6_{0.1}$	$31.6_{0.1}$	$31.8_{0.1}$	$31.4_{0.2}$	15.3	17.9	10.5	13.6	13.1	14.7	12.7
Base+VE	$206.2_{3.3}$	$78.5_{0.5}$	$114.7_{0.3}$	$100.6_{0.5}$	$121.0_{0.4}$	$91.4_{0.4}$	$105.4_{0.2}$	$141.5_{0.8}$	$78.9_{0.5}$	$111.6_{0.4}$	$98.2_{0.8}$	$122.5_{1.2}$	$84.9_{8.8}$	$98.3_{0.8}$	44.7	39.7	36.9	36.9	37.0	37.8	38.7
CV	$224.3_{2.9}$	$78.2_{0.1}$	$120.2_{0.2}$	$107.6_{0.6}$	$132.7_{0.1}$	$90.1_{1.1}$	$103.5_{0.2}$	$185.1_{0.8}$	$84.8_{0.4}$	$123.5_{1.0}$	$135.6_{1.4}$	$123.2_{1.0}$	$91.8_{1.1}$	$109.5_{1.2}$	44.5	33.0	28.6	33.4	34.7	31.0	31.4
Chat	$19.2_{0.0}$	$22.0_{0.0}$	$18.3_{0.0}$	$19.6_{0.0}$	$21.1_{0.1}$	$22.9_{0.0}$	$18.9_{0.0}$	$31.6_{0.1}$	$33.0_{0.2}$	$28.2_{0.0}$	$31.5_{0.1}$	$31.7_{0.0}$	$31.6_{0.1}$	$31.0_{0.1}$	15.0	17.3	10.3	13.2	12.9	14.4	12.5
Chat+CPT	$18.9_{0.0}$	$21.8_{0.0}$	$18.4_{0.0}$	$19.5_{0.0}$	$21.1_{0.0}$	$23.2_{0.1}$	$18.9_{0.0}$	$31.8_{0.1}$	$33.4_{0.1}$	$28.3_{0.0}$	$31.7_{0.2}$	$31.3_{0.3}$	$31.9_{0.1}$	$31.3_{0.1}$	14.9	17.2	10.4	13.3	13.0	14.4	12.7
Chat+VE	$218.9_{3.1}$	$78.6_{0.3}$	$120.2_{0.7}$	$107.2_{0.5}$	$135.1_{0.4}$	$93.2_{0.6}$	$99.3_{0.5}$	$168.8_{3.7}$	$79.7_{0.2}$	$115.5_{1.2}$	$100.4_{1.5}$	$124.4_{0.9}$	$89.0_{0.5}$	$96.9_{0.4}$	44.3	39.1	34.7	39.5	40.6	37.6	36.9
ElChat	$233.9_{1.9}$	$75.8_{0.5}$	$118.2_{2.6}$	$105.5_{0.5}$	$126.3_{0.6}$	$90.3_{1.4}$	$101.7_{0.4}$	$212.0_{3.4}$	$81.3_{0.8}$	$113.0_{1.2}$	$107.7_{0.4}$	$124.1_{0.7}$	$89.8_{0.8}$	$107.2_{0.5}$	44.5	37.9	34.1	38.9	38.0	37.6	37.3
ElChat \Merge	$201.3_{2.6}$	$79.3_{0.0}$	$121.6_{0.4}$	$109.1_{0.2}$	$132.1_{1.1}$	$88.9_{0.4}$	$103.8_{0.2}$	$164.1_{2.2}$	$76.3_{0.8}$	$111.4_{1.5}$	$84.2_{1.1}$	$102.0_{2.7}$	$84.3_{0.4}$	$94.5_{0.3}$	42.2	39.1	34.4	40.2	39.9	37.3	37.6
ElChat \Copy	$235.0_{4.3}$	$79.2_{0.1}$	$118.6_{0.2}$	$107.3_{0.2}$	$123.7_{0.7}$	$85.6_{1.5}$	$105.0_{0.6}$	$207.3_{1.6}$	$83.7_{1.4}$	$113.2_{0.7}$	$112.1_{0.3}$	$129.2_{0.5}$	$92.4_{0.8}$	$110.9_{2.0}$	46.7	39.6	34.3	39.5	37.5	36.4	37.8
ElChat (L)	$230.7_{2.4}$	$78.7_{0.2}$	$119.6_{0.6}$	$106.3_{0.8}$	$135.9_{0.2}$	$92.6_{0.9}$	$103.9_{0.8}$	$211.4_{3.1}$	$78.9_{0.7}$	$111.2_{1.1}$	$106.0_{0.7}$	$122.2_{0.5}$	$88.4_{1.9}$	$105.8_{0.5}$	44.9	39.2	34.3	39.4	40.8	37.7	37.8

Table 19: Qwen3 14B inference speedup measured by the number of tokens generated per second. Darker blue and red shades indicate higher positive and negative relative performance change over Chat per language and task, respectively.

	Target sum			**English → Target mt**			**Target mc**		
	am	bn	te	am	bn	te	am	bn	te
Base	$14.6_{0.1}$	$13.5_{0.1}$	$13.2_{0.1}$	$15.4_{0.1}$	$15.6_{0.2}$	$15.6_{0.1}$	17.7	16.3	12.6
Base+CPT	$14.6_{0.0}$	$14.6_{0.1}$	$13.8_{0.1}$	$15.5_{0.2}$	$16.1_{0.2}$	$15.4_{0.2}$	17.5	16.6	12.5
Base+VE	$52.3_{0.2}$	$33.7_{0.5}$	$43.5_{0.2}$	$36.3_{0.3}$	$29.6_{0.5}$	$38.6_{0.3}$	17.8	17.5	17.8
CV	$56.8_{0.2}$	$34.3_{0.1}$	$44.0_{0.4}$	$27.9_{0.1}$	$26.3_{0.2}$	$35.2_{0.4}$	17.0	17.6	17.5
Chat	$15.2_{0.2}$	$14.6_{0.1}$	$14.1_{0.1}$	$15.7_{0.1}$	$15.9_{0.2}$	$15.8_{0.2}$	17.9	16.3	12.3
Chat+CPT	$14.9_{0.1}$	$14.5_{0.1}$	$14.1_{0.0}$	$16.0_{0.0}$	$15.9_{0.2}$	$15.8_{0.1}$	17.4	15.9	12.3
Chat+VE	$55.8_{0.2}$	$34.3_{0.3}$	$44.9_{0.1}$	$28.1_{0.3}$	$32.3_{0.3}$	$42.8_{0.7}$	17.4	17.6	17.6
ElChat	$58.0_{0.5}$	$34.5_{0.1}$	$44.1_{0.2}$	$43.5_{0.7}$	$30.9_{0.5}$	$40.3_{0.5}$	17.4	17.6	17.8