

# 000 001 002 003 004 005 006 007 008 009 010 RETHINKING ALIGNMENT IN CROSS-LINGUAL KNOWL- EDGE TRANSFER

005 **Anonymous authors**

006 Paper under double-blind review

## 009 ABSTRACT

011 Despite LLMs' advanced performance in multilingual tasks, they usually have  
 012 performance gap on the same task in dominant languages (e.g., English) and  
 013 non-dominant languages (e.g., Turkish). In this paper, we analyze LLMs' ca-  
 014 pacity to transfer the task knowledge learned in the dominant language to the  
 015 non-dominant language. We first formulate the cross-lingual transfer problem  
 016 into a gradient alignment problem and then connect it to the representation align-  
 017 ment problem. We show that pre-trained LLMs with decent representation align-  
 018 ment ability can easily transfer knowledge from the dominant language by simple  
 019 fine-tuning, while others need carefully designed training strategies. For the  
 020 latter, we propose a cross-lingual in-context prompt tuning (CL-ICP) model to  
 021 enhance gradient alignment, which utilizes in-context attentions to generalize to  
 022 unseen data. In addition, we apply a representation shift to enhance represen-  
 023 tation alignment between demonstration and target samples. Experiments show  
 024 that CL-ICP improves cross-lingual transfer in both high and low resource sce-  
 025 narios. The code is available in <https://anonymous.4open.science/r/Cross-Lingual-Alignment-94C2>.

## 027 1 INTRODUCTION

030 State of the art LLMs are pre-trained to have strong capacities in solving multilingual downstream  
 031 tasks (Shi et al., 2023; Zhao et al., 2025a). However, their performance varies across different  
 032 languages even for the same task (Huang et al., 2023). Specifically, models tend to perform better  
 033 in dominant languages like English than in non-dominant languages like Turkish (Asai et al., 2024;  
 034 Gurgurov et al., 2024a). Assuming the major knowledge for solving a task is language agnostic,  
 035 models performing well in the dominant language should have the knowledge to solve the task and are  
 036 supposed to also perform well in non-dominant languages. In this paper, we study the cross-lingual  
 037 transfer problem which utilizing the knowledge learned from a dominant language to enhance the  
 038 model's task performance in a non-dominant language.

039 Previous works have identified that models' alignment between languages is the key to cross-lingual  
 040 transfer (Stap et al., 2023; Tanwar et al., 2023a; Qin et al., 2024; Zeng et al., 2025). The multi-  
 041 lingual alignment includes: (1) language-level alignment, which utilizes LLMs' ability to encode the  
 042 knowledge from different languages to the dominant language's space (Wang et al., 2024; Zhao et al.,  
 043 2024b; Wendler et al., 2024; Schut et al., 2025); (2) representation-level alignment, where models  
 044 are trained to connect data representations from different languages in supervised and unsupervised  
 045 manners (Bornea et al., 2021; Li et al., 2024; Zhang et al., 2025a). However, what kind of alignment  
 046 is necessary for knowledge transfer remains underexplored (Hämmerl et al., 2024). Such lack of  
 047 understanding presents a challenge in optimizing cross-lingual transfer strategies. For example, in  
 048 post-training, some works claim that simple fine-tuning on the dominant language is already strong  
 049 for cross-lingual transfer (Chirkova & Nikoulina, 2024); while other works emphasize the need for  
 additional alignment between language representations (Li et al., 2024; Zhang et al., 2025a).

050 To better understand the effect of alignment in cross-lingual transfer, we first formulate cross-lingual  
 051 transfer as a gradient alignment problem that aims to maximize the inner product between gradients  
 052 on the dominant and non-dominant languages' data during training (Riemer et al., 2019). Therefore,  
 053 the decrease of loss on the dominant language's data will also lead to a loss decrease on the non-  
 dominant language's data and cause knowledge transfer. Then we show that gradient alignment

depends on the alignment between representations of dominant and non-dominant languages’ data. For pre-trained LLMs with decent representation alignment ability, fine-tuning on the dominant language may directly transfer the knowledge to the non-dominant language.

In the case that needs further tuning with the non-dominant language’s data, we show that simple fine-tuning and multi-task learning tend to maximize gradient alignment within and across languages and therefore have strong performance with sufficient training data. However, data in non-dominant languages are usually low-resourced (Huang et al., 2023). In the low-resource scenario, existing models which enforce alignment between accessible data may not be generalized enough. To address this problem, we propose a cross lingual in-context prompt tuning (CL-ICP) model which utilizes the target sample’s attention on different demonstration samples to infer its relation to the unseen data. When training with in-context samples, it enforces gradient alignment based on the contribution of each dominant language’s data (demonstration) to the prediction of non-dominant language’s data (target). To improve the relevance between demonstration and target samples for better transfer, we shift the target sample’s representation to make it closer to the demonstration’s representation. Experiments show that CL-ICP outperforms existing models in both high and low resource scenarios.

In summary, our work makes the following contributions:

- We provide a new and unified perspective for studying the alignment problem in cross-lingual transfer. Specifically, we formulate the problem with a gradient alignment objective; analyze the connection between gradient and representation alignment; and show the alignment effects in existing models.
- We show that adding in-context samples may further improve the gradient alignment during training; and propose a CL-ICP model with representation shift to improve both gradient and representation alignment.
- Experiments in different cross-lingual transfer scenarios show the effectiveness of our CL-ICP model, which further supports our alignment claims.

## 2 RELATED WORK

**Cross-lingual Transfer** Despite strong multilingual capacity (Shi et al., 2023; Zhao et al., 2025a), current LLMs usually have limited proficiency in low-resource languages (Huang et al., 2023; Gurgurov et al., 2024a). This calls for cross-lingual transfer models that leverage knowledge from high-resource (dominant) languages to improve performance in low-resource (non-dominant) languages (Ruder et al., 2019). Existing works improve cross-lingual transfer from the pre-training to post-training stages. Conneau & Lample (2019); Ouyang et al. (2021) use translation based pre-training to enhance LLMs’ multilingual transfer ability. Gupta et al. (2023); Zhao et al. (2024a); Fujii et al. (2024) continually pre-train the LLMs to better understand non-dominant languages. Pan et al. (2020); Wang et al. (2022); Gurgurov et al. (2024b); Cassano et al. (2024) use adaptation techniques to transfer pre-trained knowledge to non-dominant languages’ tasks.

Recently, In-Context Learning (ICL) has emerged as a powerful paradigm for LLM post training (Brown et al., 2020). Previous works have also explored the benefit of ICL in multilingual task learning (Li et al., 2024), but only few of them explored the transfer ability of using cross-lingual samples in ICL (Tanwar et al., 2023b; Zhang et al., 2024). Even for the above works, they aimed to improve ICL in the multilingual settings without training, and thus not exploring the cross-lingual knowledge transfer ability of in-context training compared to fine-tuning or multi-task learning. Our work analyzes the benefits of training with in-context samples in cross-lingual transfer, showing that it may achieve better alignment to unseen data and thus benefit in the low-resource scenario.

**Cross-Lingual Alignment** Alignment has been recognized as a crucial factor in effective cross-lingual transfer (Stap et al., 2023; Tanwar et al., 2023a; Qin et al., 2024; Zeng et al., 2025). Previous works show that LLMs tend to map different languages’ representations to one language’s representation space (Wang et al., 2024; Zhao et al., 2024b; Wendler et al., 2024; Schut et al., 2025), which indicates their language alignment ability before post-training. In post-training, existing works add extra representation alignment to better align cross-lingual representations via translation-based objectives (Bornea et al., 2021), contrastive learning (Li et al., 2024) and representation shift (Zhang et al., 2025a; Sundar et al., 2025). However, to what extent the extra alignment is necessary in post-

108 training remains unexplored (Hämmerl et al., 2024). In this work, we first analyze the alignment in  
 109 the gradient perspective and connect it to the representation alignment. Our analysis bridges existing  
 110 training strategies and representation alignment methods, indicating the cases when representation  
 111 alignment can compensate existing learning strategies for better cross-lingual transfer.  
 112

### 113 3 PROBLEM STATEMENT

115 In this paper, we study the cross-lingual transfer problem which utilizes the task knowledge learned  
 116 in the dominant language to enhance the same task’s performance in the non-dominant language.  
 117

#### 118 3.1 CROSS-LINGUAL TRANSFER SETTINGS

120 We consider two main cross-lingual transfer scenarios: (1) **Dominant language direct transfer** where  
 121 the model learns the task in the dominant language and then is directly applied to the non-dominant  
 122 language; (2) **Mixed language transfer** where the model learns the task in both the dominant and  
 123 non-dominant languages to improve performance in the non-dominant language.

124 Based on different data accessibility in the above scenarios, we divide the data in cross-lingual  
 125 transfer into the following categories:  
 126

- 127 • The accessible data  $\mathcal{D}_s$  in the dominant language  $s$ ;
- 128 • The accessible data  $\mathcal{D}_t$  in the non-dominant language  $t$ ;
- 129 • The inaccessible data  $\mathcal{D}_{unseen}$ , which is a minimal complementary set that enables the LLM  
 130 to sufficiently learn the task in language  $t$  when added to the accessible data.  
 131

132 In the dominant language direct transfer only  $\mathcal{D}_s$  is used in training, while in the mixed language  
 133 transfer both  $\mathcal{D}_s$  and  $\mathcal{D}_t$  are used in training. When sufficiently learned the task in the non-dominant  
 134 language  $t$ , the learned model should achieve good performance on the language  $t$ ’s data  $\mathcal{D}_t$  and the  
 135 inaccessible data  $\mathcal{D}_{unseen}$ .  
 136

#### 137 3.2 PROBLEM FORMULATION

139 We study the cross-lingual transfer problem from a gradient perspective. Since the goal of cross-  
 140 lingual transfer is to use a dominant language to improve the accuracy of a non-dominant language  
 141 in the downstream task (Lin et al., 2019), the training that decreases the loss on the dominant or  
 142 mixed languages’ data should also decrease the loss (i.e., improve the accuracy) on the target data  $\mathcal{D}_t$   
 143 and  $\mathcal{D}_{unseen}$ . Training by gradient descent, this can be achieved by maximizing the inner product  
 144 between the model’s gradients on the losses (Riemer et al., 2019), which is our **gradient alignment**  
 145 objective. By maximizing the gradient inner product, the gradient on the training data tends to be in  
 146 the same direction as that on the target data and thus decreases the losses simultaneously.  
 147

148 Denote the training data as  $\mathcal{D}_{train}$  and the model’s loss on it as  $\mathcal{L}_w(\mathcal{D}_{train})$ , where  $w$  represents  
 149 the model parameters. The gradient alignment objective is to maximize the inner product between  
 150 gradients on the training and non-dominant language’s data:  
 151

$$150 \max_w \underbrace{\nabla \mathcal{L}_w(\mathcal{D}_{train}) \cdot \nabla \mathcal{L}_w(\mathcal{D}_t)}_{\text{Gradient alignment to } \mathcal{D}_t} + \underbrace{\nabla \mathcal{L}_w(\mathcal{D}_{train}) \cdot \nabla \mathcal{L}_w(\mathcal{D}_{unseen})}_{\text{Gradient alignment to } \mathcal{D}_{unseen}}, \quad (1)$$

153 which includes the alignment with gradients on the accessible language  $t$ ’s data  $\mathcal{D}_t$  and the inaccessible  
 154 data  $\mathcal{D}_{unseen}$ . In the scenario of dominant language direct transfer,  $\mathcal{D}_{train} = \mathcal{D}_s$ ; in the scenario of  
 155 mixed language transfer,  $\mathcal{D}_{train} = [\mathcal{D}_s, \mathcal{D}_t]$ .  
 156

**157 Connection to Representation Alignment** Gradient alignment depends on representation align-  
 158 ment, which is shown to be important for cross-lingual transfer (Li et al., 2024; Zhang et al., 2025a).  
 159 We show the connection by theoretical analysis of a toy example.

160 In the example, we have a linear regression model with parameter  $w \in \mathbb{R}^d$ . For any single sample  
 161 input  $x \in \mathbb{R}^{1 \times d}$ , the model outputs  $\hat{y} = xw$ . Assume there is an optimal parameter  $w^*$  that fits  
 162 samples from the same task in both language  $s$  and  $t$ , the loss on each sample is  $\|\hat{y} - xw^*\|_2^2$ .  
 163

162 Then the inner product of gradients on  $\mathcal{D}_{train}$  and  $\mathcal{D}_t$  is:  
 163

$$\nabla \mathcal{L}_{\mathbf{w}}(\mathcal{D}_{train}) \cdot \nabla \mathcal{L}_{\mathbf{w}}(\mathcal{D}_t) = \gamma \sum_{i,j} \underbrace{\mathbf{x}_t^i (\mathbf{x}_{train}^j)^T}_{RepAlign 1} \sum_m \underbrace{\lambda_m (\mathbf{x}_t^i \mathbf{q}_m) (\mathbf{x}_{train}^i \mathbf{q}_m)^T}_{RepAlign 2}, \quad (2)$$

167 where  $\gamma$  is a positive scalar,  $\mathbf{x}_t^i \in \mathbb{R}^{1 \times d}$  is the  $i$ -th sample feature in  $\mathcal{D}_t$ , and  $\mathbf{x}_{train}^j$  is the  $j$ -th sample  
 168 feature in  $\mathcal{D}_{train}$ .  $\mathbf{q}_m$  is an orthonormal basis of the solution space  $(\mathbf{w} - \mathbf{w}^*)(\mathbf{w} - \mathbf{w}^*)^T$ , with the  
 169 corresponding eigenvalue  $\lambda_m \geq 0$ . The derivation is in Appendix A.  
 170

171 In Eq. (2), the inner product of gradients  
 172 depends on the correlations *RepAlign1* and  
 173 *RepAlign2* between data representations (fea-  
 174 tures). *RepAlign1* shows the correlation between  
 175 input data representations, while *RepAlign2*  
 176 shows the correlation between data representa-  
 177 tions projected to the solution space. An exam-  
 178 ple of gradient alignments influenced by repre-  
 179 sentation alignments is shown in Fig. 1. In Fig.  
 180 1(a), correlations between input representa-  
 181 tions  $\mathbf{x}_{train}$ ,  $\mathbf{x}_t$  and their projections to the solution  
 182 space are in the same direction, which causes positive  
 183 gradient alignment. In Fig. 1(b), although input  
 184 representations are positively correlated, their pro-  
 185 jections to the solution space  $\mathbf{p}_1$  are nega-  
 186 tively correlated, which causes negative align-  
 187 ment. In Fig. 1(c),  $\mathbf{x}_{train} \perp \mathbf{x}_t$  and there is no  
 188 gradient alignment effect.  
 189

190 We discuss ways to improve gradient alignments in different scenarios in the following sections.  
 191

## 4 DOMINANT LANGUAGE DIRECT TRANSFER

192 In the dominant language direct transfer, we train the model only on data from the dominant language  
 193 (i.e.  $\mathcal{D}_{train} = \mathcal{D}_s$ ) and then directly apply the model on the non-dominant language's data. Therefore,  
 194 it is infeasible to explicitly maximize the gradient alignment between  $\nabla \mathcal{L}_{\mathbf{w}}(\mathcal{D}_{train})$  and  $\nabla \mathcal{L}_{\mathbf{w}}(\mathcal{D}_t)$   
 195 in Eq. (1) during training. However, considering the connection between gradient and representation  
 196 alignment, LLMs with aligned pre-trained representations in language  $s$  and  $t$  and can still activate  
 197 the gradient alignment effect. For example, if pre-trained representations of samples in  $\mathcal{D}_s$  and  $\mathcal{D}_t$   
 198 have large positive correlations (*RepAlign1*) when they rely on the same knowledge to solve the task  
 199 (*RepAlign2*), they can still achieve positive gradient alignment even only tuning on  $\mathcal{D}_s$  for the task.  
 200

201 We hypothesize that an LLM has aligned pre-trained representations if it has highly correlated pre-  
 202 trained representations of translated data pairs, which rely on similar knowledge to solve the task. We  
 203 quantify LLMs' pre-trained representation alignment ability by the deviation of cosine similarities  
 204 between pre-trained token representations in and across languages (Appendix B). The smaller cosine  
 205 deviations indicate better representation alignment ability. Then we evaluate the correlation between  
 206 different LLMs' representation alignment ability and their direct transfer performance in Fig. 2.  
 207

208 In Fig. 2(c), the models' direct transfer performance tend to decrease when the cosine deviation  
 209 gets large. This suggests that better representation alignment may lead to better direct transfer  
 210 performance. The visualization of cosine similarities between tokens are shown in Fig. 2(a). Qwen  
 211 2.5 7B has evenly distributed cosine similarity within and across English (En) and Chinese (Zh). On  
 212 the other hand, Llama 3.1 8B's cosine similarity deviation between En and Zh is larger and its direct  
 213 transfer performance on Zh is also worse than Qwen. We also visualize representations of the En and  
 214 Zh sentences by average pooling their token representations. Better direct transfer models (Qwen  
 215 2.5) may have close sentence representations in middle layers as well. However, different pooling  
 216 strategies may influence the representation distance (Zhang et al., 2025a).

217 For models with good representation alignment ability, direct transfer from the dominant language  
 218 may enable decent performance on non-dominant languages. This suggests improving models  
 219 pre-trained representation alignment across different languages and tasks for cross-lingual transfer  
 220 (Muennighoff et al., 2023; Chua et al., 2025). However, we can further improve models' alignment  
 221 ability by post-training with task data in non-dominant language, discussed in the following section.  
 222

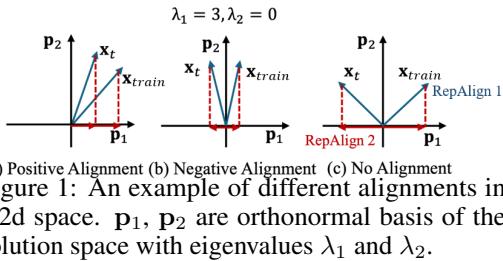


Figure 1: An example of different alignments in a 2d space.  $\mathbf{p}_1, \mathbf{p}_2$  are orthonormal basis of the solution space with eigenvalues  $\lambda_1$  and  $\lambda_2$ .

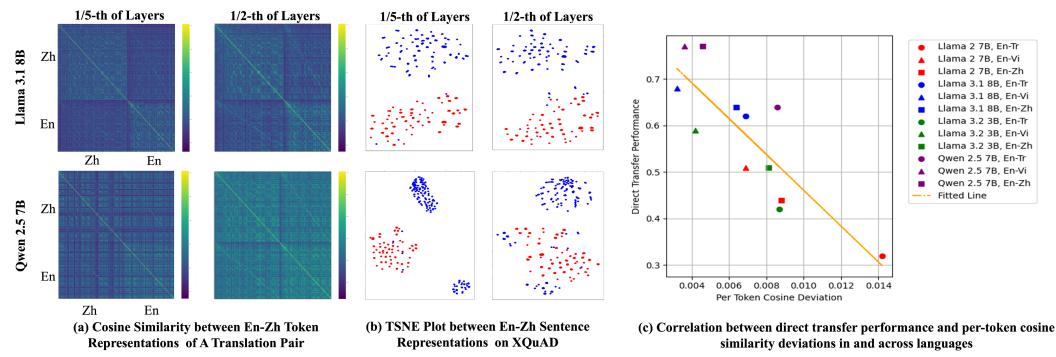


Figure 2: (a) The cosine similarity between *token* representations in a En-Zh translated sentence on XQuAD; (b) The scatter of En-Zh *sentence* representations, which are computed by average pooling the token representations in sentences; (c) Correlation between LLMs’ direct transfer performance (XQuAD F1) and their cosine similarity deviations between En-En and En-Zh token representations.

## 5 MIXED LANGUAGE TRANSFER

When mixed language data  $\mathcal{D}_s$  and  $\mathcal{D}_t$  is available in training, we can explicitly train the model to maximize the gradient alignment between language  $s$  and  $t$ ; and further enhance the representation alignment. In this section, we show that existing training strategies including target language fine-tuning and multi-task learning already improve the gradient alignment between  $\mathcal{D}_{train}$  and the non-dominant data  $\mathcal{D}_t$ . However, they may ignore potential gradient alignment to inaccessible data  $\mathcal{D}_{unseen}$  especially in the low-resource scenario. To address this issue, we use in-context prompt tuning (CL-ICP) model which utilizes the target sample’s attention on mixed languages’ samples to infer its relation to the unseen data  $\mathcal{D}_{unseen}$ . In addition, we shift the representation of the target sample to better align with the representation of the demonstration samples.

### 5.1 HOW DO EXISTING METHODS IMPROVE GRADIENT AND REPRESENTATION ALIGNMENT

**Gradient Alignment** Many works show that simply using fine-tuning (FT) and multi-task learning (MTL) already perform well in many cross-lingual transfer scenarios (Chirkova & Nikoulina, 2024; Chua et al., 2025; Gaschi et al., 2023; Wu et al., 2023; M’hamdi et al., 2021; Mousi et al., 2024). By calculating the second-order Taylor expansion of gradients (Nichol et al., 2018) in FT and MTL, we show that this is because FT and MTL improve the gradient alignment:

*Target Language Fine-tuning (FT):* In FT, we have  $\mathcal{D}_{train} = \mathcal{D}_t$  and the gradient  $g_w(\mathcal{D}_{train})$  is:

$$g_w(\mathcal{D}_{train}) = \nabla \mathcal{L}_w(\mathcal{D}_{train}) - \frac{\beta}{2} \underbrace{\nabla (\nabla \mathcal{L}_w(\mathcal{D}_t) \cdot \nabla \mathcal{L}_w(\mathcal{D}_t))}_{\text{Gradient alignment to } \mathcal{D}_t} \quad (3)$$

where  $\beta$  is the learning rate times the number of gradient descent steps.

*Multi-Task Learning (MTL):* In MTL, we have  $\mathcal{D}_{train} = [\mathcal{D}_s, \mathcal{D}_t]$  and the gradient  $g_w(\mathcal{D}_{train})$  is:

$$g_w(\mathcal{D}_{train}) = \nabla \mathcal{L}_w(\mathcal{D}_{train}) - \frac{\beta}{2} \nabla (\nabla \mathcal{L}_w(\mathcal{D}_s) \cdot \nabla \mathcal{L}_w(\mathcal{D}_s)) - \frac{\beta}{2} \underbrace{\nabla ((2 \nabla \mathcal{L}_w(\mathcal{D}_s) + \nabla \mathcal{L}_w(\mathcal{D}_t)) \cdot \nabla \mathcal{L}_w(\mathcal{D}_t))}_{\text{Gradient alignment to } \mathcal{D}_t}.$$

The derivation is in Appendix A. As shown above, gradient descent of FT and MTL tends to maximize the inner products between gradient on the training data and the data  $\mathcal{D}_t$ . This is a part of our gradient alignment objective in Eq. (1).

As discussed in Section 4, when sample representations correlate well after pre-training, simply using FT or MTL may achieve good gradient alignment and have decent cross-lingual performance.

Otherwise, the gradient alignment may be limited (e.g., due to small absolute value of *RepAlign 1*). This may require explicit alignment between representations.

**Representation Alignment** Previous works explicitly strengthen representation alignment to achieve better cross-lingual transfer performance (Li et al., 2024; Zhang et al., 2025b; Tang et al., 2024; Zhang et al., 2025b), which directly improves *RepAlign 1* and may make models more capable for gradient alignment. However, accurate representation alignment (e.g. via contrastive learning) usually requires rich data of translation pairs, and thus may be infeasible for non-dominant language  $t$  which usually has low-resource data  $\mathcal{D}_t$  (Zhang et al., 2025a). On the other hand, coarse representation alignment may cause negative gradient alignment (Fig. 1(b)) and not improve performance.

As analyzed above, existing gradient and representation alignment methods focus on the alignment between training data and accessible data  $\mathcal{D}_t$ , which may omit the alignment with the unseen data  $\mathcal{D}_{unseen}$  especially in the low-resource scenarios (Fig. 5).

## 5.2 CROSS-LINGUAL IN CONTEXT PROMPT TUNING (CL-ICP)

In this section, we show that one can utilize LLMs’ in-context learning (ICL) ability to improve the model’s gradient alignment with unseen data  $\mathcal{D}_{unseen}$ .

Traditional ICL uses some data samples as demonstration to enhance a target sample’s prediction. As shown in Table 1, we find that pre-trained LLMs allocate higher attention on demonstration samples that contribute more to the target sample’s prediction. The target sample’s various attention on demonstration samples may infer its relation to unseen samples in  $\mathcal{D}_{unseen}$ . For example, assume that  $\mathcal{D}_{train}$  does not include translation pairs in language  $s$  and  $t$ . Then for a target sample in  $\mathcal{D}_t$ , demonstration samples in  $\mathcal{D}_s$  with high in-context attention may be close to its translation in language  $s$ . In training, the target samples are supposed to have more alignment with more related demonstration samples.

**In-Context Prompt Tuning (ICP)** To incorporate ICL’s alignment ability to  $\mathcal{D}_{unseen}$  during training, we train for the target task by learning soft-prompts  $\mathbf{p}$  with in-context samples. We learn soft prompts instead of full fine-tuning to better preserve the model’s pre-trained knowledge in arranging high attention to related demonstrations.

In ICP, our model input is  $\{\mathbf{p}, \mathbf{x}_d, y_d, \mathbf{x}_t\}$ , where  $\{\mathbf{x}_d, y_d\} = \{\mathbf{x}_{d1}, y_{d2}, \dots, \mathbf{x}_{dk}, y_{dk}\}$  includes  $k$  demonstration samples;  $\{\mathbf{x}_t, y_t\}$  is the target sample from  $\mathcal{D}_t$ . We train soft prompts  $\mathbf{p}$  to predict the target value  $y_t$  and the demonstration value  $y_{dk}$  to learn the knowledge in both languages. The objective is  $\max_{\mathbf{p}} p(y_t|\{\mathbf{p}, \mathbf{x}_d, y_d, \mathbf{x}_t\}) + p(y_{dk}|\{\mathbf{p}, \mathbf{x}_{d1}, y_{d2}, \dots, \mathbf{x}_{dk}\})$ .

**Alignment Effect of ICP** We analyze the gradient alignment effect of ICP under the attention mechanism (Vaswani et al., 2017). For simplicity, we assume each sample representation in  $\mathbf{x}_d$  has the size  $\mathbb{R}^{1 \times d}$ . Then the representation  $\mathbf{x}$  for predicting  $y_t$  is the weighted sum of in-context sample representations:

$$\mathbf{x} = \text{attn}_t(\mathbf{p})\mathbf{p} + \text{attn}_t(\mathbf{x}_d)\mathbf{x}_d + \text{attn}_t(\mathbf{x}_t)\mathbf{x}_t + \text{attn}_t(y_d)y_d,$$

where  $\text{attn}_t(\cdot)$  is the self attention with the query of  $\cdot$  for  $y_t$  prediction. Based on Table 1,  $\text{attn}_t(\cdot)$  is higher when  $\cdot$  contributes more to  $y_t$ ’s prediction.

Based on Eq. (3), by tuning for a target task with data  $\mathbf{x}^i$  and  $\mathbf{x}^j$ , gradient descent tends to maximize the inner product between gradients  $\nabla \mathcal{L}_w(\mathbf{x}^i) \cdot \nabla \mathcal{L}_w(\mathbf{x}^j)$ . Based on the expansion of gradients in Eq. (2),  $\nabla \mathcal{L}_w(\mathbf{x}^i) \cdot \nabla \mathcal{L}_w(\mathbf{x}^j)$  is related to the inner product between  $\mathbf{x}^i$  and  $\mathbf{x}^j$ :

$$\mathbf{x}^i(\mathbf{x}^j)^T = \underbrace{\phi_1 \mathbf{x}_d^i(\mathbf{x}_d^j)^T + \phi_2 \mathbf{x}_d^i(\mathbf{x}_t^j)^T + \phi_3 \mathbf{x}_t^i(\mathbf{x}_t^j)^T}_{\text{Pre-trained correlation between sample representations}} + f(\{\mathbf{x}_d^{i,j}, \mathbf{x}_t^{i,j}, y_d^{i,j}\})\mathbf{p}^T, \quad (4)$$

where  $f(\{\mathbf{x}_d^{i,j}, \mathbf{x}_t^{i,j}, y_d^{i,j}\})\mathbf{p}^T$  is the correlation between soft prompts  $\mathbf{p}$  and in-context representations;  $\phi_1 = \text{attn}_t(\mathbf{x}_d^i)\text{attn}_t(\mathbf{x}_d^j)$ ,  $\phi_2 = 2\text{attn}_t(\mathbf{x}_d^i)\text{attn}_t(\mathbf{x}_t^j)$ ,  $\phi_3 = \text{attn}_t(\mathbf{x}_t^i)\text{attn}_t(\mathbf{x}_t^j)$  are pre-trained attention correlations. We omit the non-correlation terms in the equation.

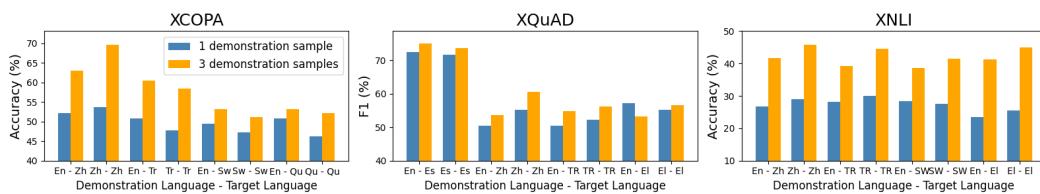


Figure 3: Llama 3.1 8B’s ICL performance with different demonstration samples.

To maximize the inner product between gradients, the model may learn prompts  $\mathbf{p}$  to increase the representation inner product  $\mathbf{x}^i(\mathbf{x}^j)^T$ . With in-context samples, such process is guided by the pre-trained correlation  $\phi$  between sample representations (Eq. (4)). Without in-context samples, we have the representation  $\mathbf{x} = \mathbf{x}_t$  and the inner product  $\mathbf{x}^i(\mathbf{x}^j)^T = \mathbf{x}_t^i(\mathbf{x}_t^j)^T + f(\mathbf{x}_t^{i,j} \mathbf{p}^T)$ , which do not include the pre-trained correlation between samples and may require sufficient data to learn.

**Mixed Language Demonstrations** According to Tanwar et al. (2023b); Zhang et al. (2024), LLMs’ ICL ability highly depends on the selection of demonstration samples. In Fig. 3, we have two observations about cross-lingual ICL: (1). For a target sample in the non-dominant language  $t$ , using demonstration samples in language  $t$  sometimes achieves better ICL performance than using English (dominant language) demonstrations. (2). Using more demonstration samples ( $k = 3$ ) achieves better ICL performance than only using one.

Based on these observations, our CL-ICP model uses 3 demonstration samples in most tasks. Instead of choosing demonstrations from one language, we randomly select demonstration samples in  $\{\mathbf{x}_d, y_d\}$  from both  $\mathcal{D}_s$  and  $\mathcal{D}_t$  to balance the relatedness between samples and the knowledge from the dominant language.

**Representation Shift** In Eq. (4), when pre-trained correlation between sample representations is low, CL-ICP may not well connect the demonstration and target samples and thus downgrade to single-sample prompt tuning. Therefore, we shift representations of  $\mathbf{x}_t$  to further enhance its representation alignment to the demonstrations  $\mathbf{x}_d$ . An overview of the shifting operation is in Fig. 4.

In each layer of transformer models, the sentence inputs  $\mathbf{x}_d$  and  $\mathbf{x}_t$  are sequences of token representations. We denote their layer-wise token representations as  $\mathbf{H}_d^L = [\mathbf{h}_{d,1}^L, \dots, \mathbf{h}_{d,n_d}^L]$  and  $\mathbf{H}_t^L = [\mathbf{h}_{t,1}^L, \dots, \mathbf{h}_{t,n_t}^L]$  at the layer  $L$ . Inspired by Xu et al. (2023); Zhang et al. (2025b), we add representation deviations on each token representation in the target sample at the  $L_{\text{to}}$  and  $L_{\text{back}}$  layers where representations from different languages are close to each other. The representation deviations are calculated by

$$\Delta \mathbf{h}^{L_{\text{to}}} = \text{sentpool}(\mathbf{H}_d^{L_{\text{to}}}) - \text{sentpool}(\mathbf{H}_t^{L_{\text{to}}}); \quad \Delta \mathbf{h}^{L_{\text{back}}} = \text{sentpool}(\mathbf{H}_t^{L_{\text{back}}}) - \text{sentpool}(\mathbf{H}_d^{L_{\text{back}}})$$

where sentpool calculates the average token representations as the sentence representation.

The representation shift first projects the target sample representations close to the demonstration’s representation by  $\Delta \mathbf{h}^{L_{\text{to}}}$  and then projects them back by  $\Delta \mathbf{h}^{L_{\text{back}}}$ . By making target sample representations close to demonstration representations, the model may easier find the connection between the demonstration and target samples, and thus benefit in-context prompt tuning.

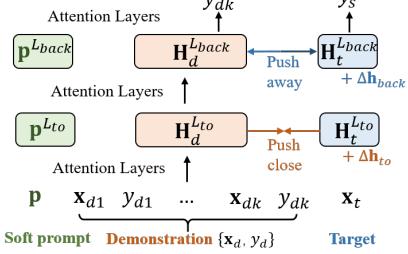


Figure 4: CL-ICP with representation shift.

## 6 EXPERIMENTS

### 6.1 SETUP

**Datasets** Our experiments are performed on three cross-lingual datasets: (1) *XQuAD* (Artetxe et al., 2019) for multilingual question answering, with Chinese (Zh), Spanish (Es), Turkish (Tr), Greek (El) as target non-dominant languages. We report exact-match and F1 score on this dataset; (2) *XCOPA* (Ponti et al., 2020) for multilingual causal reasoning, with Chinese (Zh), Swahili (Sw), Turkish (Tr), Quechua (Qu) as target non-dominant languages. We report accuracy on the dataset. (3). *XNLI* (Conneau et al., 2018) for text classification with Chinese (Zh), Swahili (Sw), Turkish (Tr), Greek

378 Table 2: Evaluation Results. The **bold** and underline scores are the first and second best scores.  
379

380 Language $t$	XQuAD (EM/F1 %)					XCOPA (Acc %)					XNLI (Acc %)				
	Zh	Es	Tr	El	Avg	Zh	Tr	Sw	Qu	Avg	Zh	Tr	Sw	El	Avg
<b>Llama3.1</b>	58.0/63.7	38.2/62.5	<b>33.2/62.4</b>	32.8/59.7	40.5/62.1	63.2	54.6	52.6	<u>52.2</u>	55.7	61.6	58.3	51.8	58.7	57.6
TL	54.3/62.3	40.9/67.2	17.6/49.1	35.7/61.7	37.1/60.1	81.7	62.3	52.5	<b>53.6</b>	62.5	67.8	65.4	53.2	68.3	63.7
MTL	51.3/57.5	42.3/68.6	18.1/51.3	38.2/62.3	37.5/59.9	83.5	65.4	53.1	48.4	62.6	67.0	64.2	54.1	67.5	63.2
Shifcon	57.4/64.1	42.4/68.8	23.4/54.3	<b>42.2/66.3</b>	41.3/63.4	84.2	65.7	54.3	50.9	63.8	68.6	65.8	53.6	67.6	63.9
CL-ICP	53.4/65.5	<b>46.4/72.3</b>	25.6/58.7	39.1/66.1	41.1/65.7	87.6	<u>77.9</u>	<b>58.8</b>	49.1	<b>68.4</b>	70.3	<u>66.1</u>	<b>59.6</b>	72.0	<u>67.0</u>
CL-ICP (S)	<b>61.6/69.1</b>	44.8/71.9	23.1/56.3	<b>41.2/67.0</b>	<b>42.7/66.1</b>	<b>87.7</b>	<b>78.5</b>	<u>57.7</u>	49.0	<u>68.2</u>	<b>70.4</b>	<b>68.3</b>	<u>57.4</u>	<b>73.6</b>	<u>67.4</u>
<b>Qwen2.5</b>	73.9/77.3	51.7/75.5	42.0/64.5	39.4/65.2	51.8/70.6	91.0	75.2	52.6	51.4	67.5	77.0	69.1	44.3	70.9	65.3
TL	<b>78.6/82.2</b>	53.5/75.4	35.3/65.1	42.0/66.6	52.3/72.3	92.1	77.2	53.5	52.0	68.7	81.7	74.1	51.4	76.3	70.9
MTL	76.9/80.8	<b>54.1/75.6</b>	38.1/67.2	<b>43.6/67.6</b>	<u>53.2/72.8</u>	93.8	77.2	54.1	51.8	69.2	81.4	76.7	52.0	<b>78.4</b>	<u>72.1</u>
Shifcon	76.8/80.9	<b>54.3/76.1</b>	37.8/67.0	42.3/67.4	52.8/72.8	93.3	76.7	53.2	50.9	68.5	81.0	<u>77.1</u>	51.4	<b>78.9</b>	<u>72.1</u>
CL-ICP	<b>78.7/81.4</b>	51.5/75.7	<u>43.7/68.2</u>	40.2/66.2	<u>53.5/72.9</u>	93.9	79.6	<u>55.6</u>	<b>54.0</b>	70.8	81.5	<b>77.2</b>	52.6	77.8	<b>72.3</b>
CL-ICP (S)	77.9/80.6	51.5/75.2	<b>44.1/68.0</b>	39.4/66.3	53.2/72.5	<b>94.0</b>	<b>82.2</b>	<b>55.9</b>	<u>52.6</u>	<u>71.2</u>	<b>81.9</b>	76.2	<b>54.1</b>	76.3	<u>72.1</u>

390  
391  
392 (El) as target non-dominant languages. We randomly sample 1000 training and testing samples for  
393 XNLI (Tu et al., 2022; Zhao et al., 2025b). For all datasets, we set English as the dominant language.  
394 The non-dominant languages are selected based on their dominance and the richness of linguistic  
395 resources (Li et al., 2024). For each dataset, we select at least one language from each dominance  
396 level (high  $\rightarrow$  low).

397 **Models** We use Qwen 2.5 7B (Qwen et al., 2025) and Llama 3.1 8B (Dubey et al., 2024) as base  
398 models of testing methods. The methods we compare are: (1) *Dominant Language Direct Transfer*  
399 (*default model performance*): only fine-tuning the pre-trained LLM on the dominant language and  
400 evaluating its performance on non-dominant language. (2) *Target Language Fine-Tuning (FT)*: fine-  
401 tuning the pre-trained LLM on the non-dominant language. (3) *Transfer Learning (TL)*: first fine-  
402 tuning the LLM on the dominant language and then fine-tuning on the non-dominant language. (4)  
403 *Multi-Task Learning (MTL)*: Fine-tuning the model on both dominant and non-dominant languages’  
404 data. (5) *ShifCon* (Zhang et al., 2025a): MTL model with representation projection and contrastive  
405 learning to align sample representations. (6) *CL-ICP*: our cross-lingual in-context prompt tuning  
406 without representation shift. (7) *CL-ICP (S)*: our CL-ICP model with representation shift. This is a  
407 short term of CL-ICP (Shift).

408 For ICP models, we use demonstration samples in the training set at test time for a fair comparison  
409 with other methods. Due to the input length constraint, we use 1 demonstration samples for XQuAD  
410 and 3 demonstration samples for other datasets. For every method, we compute the scores averaged  
411 from three random seeds. Detailed training settings are in the appendix.

## 412 6.2 EXPERIMENTAL RESULTS

413 **Results of the Gradient Alignment** We show results of different methods in Table 2. CL-ICP (with  
414 or without shift) achieves the best average performance across all datasets. The effectiveness of  
415 CL-ICP varies between tasks, base models and languages.

416 For the question answering task XQuAD which may have sequence-to-sequence distribution similar  
417 to pre-trained data, the direct transfer already achieves decent performance and the improvement  
418 of CL-ICP is marginal. For other tasks XCOPA and XNLI whose outputs are choices instead of  
419 sentences, the direct transfer is not effective enough and CL-ICP achieves much better performance  
420 compared to other baselines.

421 Comparing performance on different base models, CL-ICP outperforms baselines by larger margins  
422 on Llama 3.1 than Qwen 2.5. This may be because Qwen 2.5 has better pre-trained representation  
423 alignment than Llama 3.1 (as shown in Fig. 2(c)), which makes simple fine-tuning on two languages  
424 (e.g. MTL) strong baselines.

425 Relying on pre-trained LLMs’ ICL ability, CL-ICP’s performance also depends on LLMs pre-trained  
426 capacity on different languages. For the rare language that are poorly learned in pre-training (e.g., Qu  
427 in Llama 3.1, El in Qwen 2.5), CL-ICP does not outperform baselines with full fine-tuning. For El in  
428 Qwen 2.5, the model may need full fine-tuning instead of prompt tuning to learn the knowledge in the  
429 non-dominant language. For Qu in Llama 3.1, the model may have difficulty learning the relationship  
430 between En and Qu, which makes the two-stage training (direct transfer and TL) performs best.

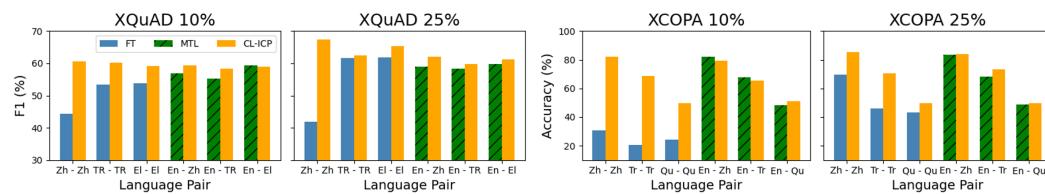


Figure 5: Llama 3.1 8B’s performance with different ratios of non-dominant language data. Under the language pair  $s - t$ , we compare CL-ICP with mixed language demonstrations and MTL; under the language pair  $t - t$ , we compare CL-ICP with non-dominant language demonstrations and FT.

**Results of the Representation Alignment** In Table 2, we compare the models with additional representation alignment, ShifCon and CL-ICP (Shift). With contrastive learning on shifted representations, ShifCon outperforms MTL in most cases on Llama 3.1. With shifted representations, CL-ICP (Shift) also outperforms CL-ICP in 2/3 datasets on Llama 3.1.

However, the effect of representation alignment is critical in different scenarios. On Qwen 2.5, Shifcon does not outperform MTL in some datasets and languages; CL-ICP (Shift)’s performance also differs in different base models and datasets. That may be because representation alignment needs carefully designed strategies and sufficient data for training (Zhang et al., 2025a). By only learning on the target tasks’ data, ShifCon and CL-ICP (Shift) may not learn the true relationship between shifted representations of data in different languages. As shown in Fig 1(b), this may cause negative gradient alignment and even decrease the performance.

**Results in Low Resource Scenarios** We show results of training with different ratios of non-dominant language data in Fig. 5. We use full dominant language (English) data in this experiment. Results show that with low-resource non-dominant language data, FT suffers from the generalization issue which makes it underperform MTL by a large margin (with less non-dominant language data, MTL is closer to the direct transfer). However, using the same data, CL-ICP significantly outperform FT, which shows its generalization ability in aligning with unseen data.

However, when the demonstration samples are in mixed languages, CL-ICP needs sufficient data to address the relationship between demonstration and target samples. With 10% data in XQuAD and XCOPA, CL-ICP with mixed language demonstrations performs worse than MTL in some cases and CL-ICP with non-dominant language demonstrations achieves better performance. With more than 25% non-dominant data, CL-ICP consistently outperform MTL.

### 6.3 ABLATION STUDY

**Influence of Demonstration** The experiments of ablation studies use Llama 3.1 8B as the base model. We show the influence of the number and mixture of demonstration samples in Table 3. Using 3 demonstration samples in mixed languages achieves best performance. Using 3 demonstration samples in the target non-dominant language outperforms that in the dominant language (English). In addition, using mixed language demonstration with only 1 sample sometimes outperforms using English demonstration with 3 samples. These indicate the importance of close distribution between demonstration and target samples in in-context training.

Table 3: CL-ICP with different demonstrations on XCOPA.

Lang. #	Zh	Tr	Sw	Qu	Avg
English, 3	86.0	74.4	55.9	50.1	66.6
Target, 3	87.3	77.1	56.7	<b>50.9</b>	68.0
Mixed, 1	83.7	75.7	57.9	49.7	66.8
Mixed, 3	<b>87.6</b>	<b>77.9</b>	<b>58.8</b>	49.1	<b>68.4</b>

Table 4: Models with in-context samples at inference time on XCOPA.

Model	Zh	Tr	Sw	Qu	Avg
TL	81.7	62.3	52.5	<b>53.6</b>	62.5
+ ICL	78.3	67.2	53.3	50.5	62.3
MTL	83.5	65.4	53.1	48.4	62.6
+ ICL	80.5	70.0	55.0	52.0	64.4
CL-ICP	<b>87.6</b>	<b>77.9</b>	<b>58.8</b>	49.1	<b>68.4</b>

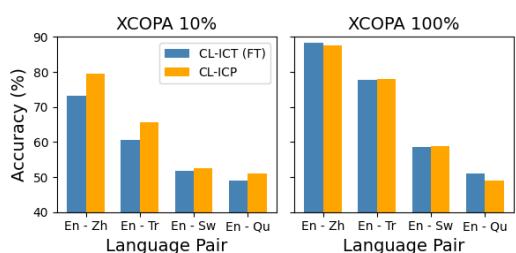


Figure 6: Comparison between CL-ICP and in-context fine-tuning (CL-ICT).

Table 5: Multi-task learning (MTL) and cross-lingual in-context tuning (CL-ICT) by only training prompts (PT) or LoRA blocks.

		Zh	Tr	Sw	Qu	Avg
PT	MTL	56.4	55.8	52.6	50.6	53.8
	CL-ICT	87.6	77.9	58.8	49.1	68.4
LoRA	MTL	87.4	75.6	59.7	50.0	68.2
	CL-ICT	88.2	81.5	63.3	53.1	71.5

**Influence of Prompt Tuning** We compare the influence of prompt tuning and full parameter tuning in our cross-lingual in-context training in Fig. 6. Although in-context fine-tuning achieves slightly better performance with 100% non-dominant language’s data, it underperforms CL-ICP in the low-resource scenario. This may be because in the low-resource scenario, full parameter tuning may be easier to lose alignment with unseen data due to the distortion pre-trained sample representations.

We also compare the in-context training using LoRA and prompt tuning with MTL in Table 5. In practice, prompt tuning only may be hard to train and the model may have low capacity in solving the target tasks. Using LoRA can improve the model’s performance compared to full finetuning. However, such a benefit is also applicable to our in-context training model. Using LoRA, our model still achieves better performance than MTL. This suggests that our improvements come not only from the parameter-efficient training, but also from the inclusion of in-context samples during training.

## 7 CONCLUSION

In this paper, we presented a comprehensive study on the cross-lingual transfer in large language models, with a particular focus on the interplay between representation and gradient alignment. To improve cross-lingual transfer in post-training, we propose a cross-lingual in-context prompt tuning model to improve gradient alignment and add a representation shift to better align representations between demonstration and target samples. Experiments show that our models have improved performance in both low and high resource scenarios. The limitation of our model is that it depends on LLMs’ ICL capacity. When adding the demonstration in the input, the input sequence length will increase and is inefficient for long samples. We leave these questions for future study.

540 REFERENCES  
541

542 Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolin-  
543 gual representations. *arXiv preprint arXiv:1910.11856*, 2019.

544 Akari Asai, Sneha Kudugunta, Xinyan Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov,  
545 Sebastian Ruder, and Hannaneh Hajishirzi. BUFFET: Benchmarking large language models for few-  
546 shot cross-lingual transfer. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings  
547 of the 2024 Conference of the North American Chapter of the Association for Computational  
548 Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1771–1800, Mexico  
549 City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.  
550 naacl-long.100. URL <https://aclanthology.org/2024.naacl-long.100/>.

551 Mihaela Bornea, Lin Pan, Sara Rosenthal, Radu Florian, and Avirup Sil. Multilingual transfer  
552 learning for qa using translation as data augmentation. In *Proceedings of the AAAI conference on  
553 artificial intelligence*, volume 35, pp. 12583–12591, 2021.

554 Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal,  
555 Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are  
556 few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

558 Federico Cassano, John Gouwar, Francesca Lucchetti, Claire Schlesinger, Anders Freeman, Car-  
559 ollyn Jane Anderson, Molly Q Feldman, Michael Greenberg, Abhinav Jangda, and Arjun Guha.  
560 Knowledge transfer from high-resource to low-resource programming languages for code llms.  
561 *Proceedings of the ACM on Programming Languages*, 8(OOPSLA2):677–708, 2024.

562 Nadezhda Chirkova and Vassilina Nikoulina. Key ingredients for effective zero-shot cross-lingual  
563 knowledge transfer in generative tasks. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.),  
564 *Proceedings of the 2024 Conference of the North American Chapter of the Association for Compu-  
565 tational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7222–7238,  
566 Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/  
567 2024.naacl-long.401. URL <https://aclanthology.org/2024.naacl-long.401/>.

568 Lynn Chua, Badih Ghazi, Yangsibo Huang, Pritish Kamath, Ravi Kumar, Pasin Manurangsi, Amer  
569 Sinha, Chulin Xie, and Chiyuan Zhang. Crosslingual capabilities and knowledge barriers in  
570 multilingual large language models, 2025. URL <https://openreview.net/forum?id=BCyAlMoyx5>.

573 Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. *Advances in  
574 neural information processing systems*, 32, 2019.

575 Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk,  
576 and Veselin Stoyanov. XNLI: Evaluating cross-lingual sentence representations. In Ellen Riloff,  
577 David Chiang, Julia Hockenmaier, and Jun’ichi Tsujii (eds.), *Proceedings of the 2018 Conference  
578 on Empirical Methods in Natural Language Processing*, pp. 2475–2485, Brussels, Belgium,  
579 October–November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1269.  
580 URL <https://aclanthology.org/D18-1269/>.

581 Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha  
582 Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models.  
583 *arXiv e-prints*, pp. arXiv–2407, 2024.

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

589 Felix Gaschi, Patricio Cerdá, Parisa Rastin, and Yannick Toussaint. Exploring the relationship  
590 between alignment and cross-lingual transfer in multilingual transformers. In Anna Rogers,  
591 Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational  
592 Linguistics: ACL 2023*, pp. 3020–3042, Toronto, Canada, July 2023. Association for Computational  
593 Linguistics. doi: 10.18653/v1/2023.findings-acl.189. URL <https://aclanthology.org/2023.findings-acl.189/>.

594 Kshitij Gupta, Benjamin Thérien, Adam Ibrahim, Mats L Richter, Quentin Anthony, Eugene  
 595 Belilovsky, Irina Rish, and Timothée Lesort. Continual pre-training of large language models:  
 596 How to (re) warm your model? *arXiv preprint arXiv:2308.04014*, 2023.

597 Daniil Gurgurov, Tanja Bäumel, and Tatiana Anikina. Multilingual large language models and curse  
 598 of multilinguality. *arXiv preprint arXiv:2406.10602*, 2024a.

600 Daniil Gurgurov, Mareike Hartmann, and Simon Ostermann. Adapting multilingual llms to low-  
 601 resource languages with knowledge graphs via adapters. *arXiv preprint arXiv:2407.01406*, 2024b.

603 Katharina Hä默rl, Jindřich Libovický, and Alexander Fraser. Understanding cross-lingual  
 604 Alignment—A survey. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of*  
 605 *the Association for Computational Linguistics: ACL 2024*, pp. 10922–10943, Bangkok, Thailand,  
 606 August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.649.  
 607 URL <https://aclanthology.org/2024.findings-acl.649/>.

608 Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. Not all  
 609 languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought  
 610 prompting. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association*  
 611 *for Computational Linguistics: EMNLP 2023*, pp. 12365–12394, Singapore, December 2023.  
 612 Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.826. URL  
 613 <https://aclanthology.org/2023.findings-emnlp.826/>.

614 Chong Li, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. Improving in-context learning of  
 615 multilingual generative language models with cross-lingual alignment. In Kevin Duh, Helena  
 616 Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American*  
 617 *Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume*  
 618 *I: Long Papers)*, pp. 8058–8076, Mexico City, Mexico, June 2024. Association for Computational  
 619 Linguistics. doi: 10.18653/v1/2024.naacl-long.445. URL <https://aclanthology.org/2024.naacl-long.445/>.

621 Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani,  
 622 Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham  
 623 Neubig. Choosing transfer languages for cross-lingual learning. In Anna Korhonen, David Traum,  
 624 and Lluís Márquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Compu-*  
 625 *tational Linguistics*, pp. 3125–3135, Florence, Italy, July 2019. Association for Computational Lin-  
 626 *guistics*. doi: 10.18653/v1/P19-1301. URL <https://aclanthology.org/P19-1301/>.

627 Meryem M'hamdi, Doo Soon Kim, Franck Dernoncourt, Trung Bui, Xiang Ren, and Jonathan  
 628 May. X-METRA-ADA: Cross-lingual meta-transfer learning adaptation to natural language  
 629 understanding and question answering. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer,  
 630 Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao  
 631 Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association*  
 632 *for Computational Linguistics: Human Language Technologies*, pp. 3617–3632, Online, June  
 633 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.283. URL  
 634 <https://aclanthology.org/2021.naacl-main.283/>.

635 Basel Mousi, Nadir Durrani, Fahim Dalvi, Majd Hawasly, and Ahmed Abdelali. Exploring alignment  
 636 in shared cross-lingual spaces. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.),  
 637 *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume*  
 638 *I: Long Papers)*, pp. 6326–6348, Bangkok, Thailand, August 2024. Association for Computational  
 639 Linguistics. doi: 10.18653/v1/2024.acl-long.344. URL <https://aclanthology.org/2024.acl-long.344/>.

641 Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven  
 642 Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir  
 643 Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson,  
 644 Edward Raff, and Colin Raffel. Crosslingual generalization through multitask finetuning. In Anna  
 645 Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting*  
 646 *of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15991–16111,  
 647 Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.  
 648 acl-long.891. URL <https://aclanthology.org/2023.acl-long.891/>.

648 Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *CoRR*,  
 649 abs/1803.02999, 2018. URL <http://arxiv.org/abs/1803.02999>.

650

651 Xuan Ouyang, Shuohuan Wang, Chao Pang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. ERNIE-  
 652 M: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual  
 653 corpora. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.),  
 654 *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 27–  
 655 38, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational  
 656 Linguistics. doi: 10.18653/v1/2021.emnlp-main.3. URL <https://aclanthology.org/2021.emnlp-main.3/>.

657

658 Lin Pan, Chung-Wei Hang, Haode Qi, Abhishek Shah, Saloni Potdar, and Mo Yu. Multilingual bert  
 659 post-pretraining alignment. *arXiv preprint arXiv:2010.12547*, 2020.

660

661 Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen.  
 662 XCOPA: A multilingual dataset for causal commonsense reasoning. In Bonnie Webber,  
 663 Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical  
 664 Methods in Natural Language Processing (EMNLP)*, pp. 2362–2376, Online, November  
 665 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.185. URL  
 666 <https://aclanthology.org/2020.emnlp-main.185/>.

667

668 Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and  
 669 Philip S. Yu. Multilingual large language model: A survey of resources, taxonomy and frontiers,  
 670 2024. URL <https://arxiv.org/abs/2404.04925>.

671

672 Qwen, ;, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan  
 673 Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang,  
 674 Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin  
 675 Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi  
 676 Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan,  
 677 Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL  
 678 <https://arxiv.org/abs/2412.15115>.

679

680 Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, , and Gerald  
 681 Tesauro. Learning to learn without forgetting by maximizing transfer and minimizing interference.  
 682 In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=B1gTShAct7>.

683

684 Sebastian Ruder, Ivan Vulić, and Anders Søgaard. A survey of cross-lingual word embedding models.  
 685 *J. Artif. Int. Res.*, 65(1):569–630, May 2019. ISSN 1076-9757. doi: 10.1613/jair.1.11640. URL  
 686 <https://doi.org/10.1613/jair.1.11640>.

687

688 Lisa Schut, Yarin Gal, and Sebastian Farquhar. Do multilingual llms think in english? *arXiv preprint  
 689 arXiv:2502.15603*, 2025.

690

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

695

696 David Stap, Vlad Niculae, and Christof Monz. Viewing knowledge transfer in multilingual ma-  
 697 chine translation through a representational lens. In Houda Bouamor, Juan Pino, and Ka-  
 698 lika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*,  
 699 pp. 14973–14987, Singapore, December 2023. Association for Computational Linguistics.  
 700 doi: 10.18653/v1/2023.findings-emnlp.998. URL <https://aclanthology.org/2023.findings-emnlp.998/>.

701

702 Anirudh Sundar, Sinead Williamson, Katherine Metcalf, Barry-John Theobald, Skyler Seto, and  
 703 Masha Fedzechkina. Steering into new embedding spaces: Analyzing cross-lingual alignment  
 704 induced by model interventions in multilingual language models. *arXiv preprint arXiv:2502.15639*,  
 705 2025.

702 Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu  
 703 Wei, and Ji-Rong Wen. Language-specific neurons: The key to multilingual capabilities in large  
 704 language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the*  
 705 *62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,  
 706 pp. 5701–5715, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi:  
 707 10.18653/v1/2024.acl-long.309. URL [https://aclanthology.org/2024.acl-long.](https://aclanthology.org/2024.acl-long.309/)  
 708 309/.

709 Eshaan Tanwar, Subhabrata Dutta, Manish Borthakur, and Tanmoy Chakraborty. Multilingual  
 710 LLMs are better cross-lingual in-context learners with alignment. In Anna Rogers, Jordan Boyd-  
 711 Gruber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association*  
 712 *for Computational Linguistics (Volume 1: Long Papers)*, pp. 6292–6307, Toronto, Canada, July  
 713 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.346. URL  
 714 <https://aclanthology.org/2023.acl-long.346/>.

715 Eshaan Tanwar, Subhabrata Dutta, Manish Borthakur, and Tanmoy Chakraborty. Multilingual llms  
 716 are better cross-lingual in-context learners with alignment. *arXiv preprint arXiv:2305.05940*,  
 717 2023b.

718 Lifu Tu, Caiming Xiong, and Yingbo Zhou. Prompt-tuning can be much better than fine-tuning  
 719 on cross-lingual understanding with multilingual language models. In Yoav Goldberg, Zornitsa  
 720 Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics:*  
 721 *EMNLP 2022*, pp. 5478–5485, Abu Dhabi, United Arab Emirates, December 2022. Association  
 722 for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.401. URL <https://aclanthology.org/2022.findings-emnlp.401/>.

723 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez,  
 724 Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von  
 725 Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Ad-*  
 726 *vances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.,  
 727 2017. URL [https://proceedings.neurips.cc/paper\\_files/paper/2017/file/3f5ee243547dee91fdb053c1c4a845aa-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fdb053c1c4a845aa-Paper.pdf).

728 Hetong Wang, Pasquale Minervini, and Edoardo Ponti. Probing the emergence of cross-lingual  
 729 alignment during LLM training. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.),  
 730 *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12159–12173, Bangkok,  
 731 Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.  
 732 findings-acl.724. URL <https://aclanthology.org/2024.findings-acl.724/>.

733 Xinyi Wang, Sebastian Ruder, and Graham Neubig. Expanding pretrained models to thousands more  
 734 languages via lexicon-based adaptation. *arXiv preprint arXiv:2203.09435*, 2022.

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

741 Linjuan Wu, Zongyi Guo, Baoliang Cui, Haihong Tang, and Weiming Lu. Good meta-tasks make a  
 742 better cross-lingual meta-transfer learning for low-resource languages. In *The 2023 Conference*  
 743 *on Empirical Methods in Natural Language Processing*, 2023. URL <https://openreview.net/forum?id=XhR6ebeEXo>.

744 Shaoyang Xu, Junzhuo Li, and Deyi Xiong. Language representation projection: Can we trans-  
 745 fer factual knowledge across languages in multilingual language models? *arXiv preprint*  
 746 *arXiv:2311.03788*, 2023.

747 Hongchuan Zeng, Senyu Han, Lu Chen, and Kai Yu. Converging to a lingua franca: Evolution  
 748 of linguistic regions and semantics alignment in multilingual large language models. In Owen  
 749 Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven  
 750 Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*,

751 752 753 754 755

756 pp. 10602–10617, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics.  
 757 URL <https://aclanthology.org/2025.coling-main.707/>.  
 758

759 Hengyuan Zhang, Chenming Shang, Sizhe Wang, Dongdong Zhang, Yiyao Yu, Feng Yao, Renliang  
 760 Sun, Yujiu Yang, and Furu Wei. ShifCon: Enhancing non-dominant language capabilities with  
 761 a shift-based multilingual contrastive framework. In Wanxiang Che, Joyce Nabende, Ekaterina  
 762 Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the  
 763 Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4818–4841, Vienna,  
 764 Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-251-0. URL  
 765 <https://aclanthology.org/2025.acl-long.239/>.  
 766

766 Hengyuan Zhang, Chenming Shang, Sizhe Wang, Dongdong Zhang, Yiyao Yu, Feng Yao, Renliang  
 767 Sun, Yujiu Yang, and Furu Wei. Shifcon: Enhancing non-dominant language capabilities with  
 768 a shift-based multilingual contrastive framework, 2025b. URL <https://arxiv.org/abs/2410.19453>.  
 769

770 Miaoan Zhang, Vagrant Gautam, Mingyang Wang, Jesujoba O Alabi, Xiaoyu Shen, Dietrich Klakow,  
 771 and Marius Mosbach. The impact of demonstrations on multilingual in-context learning: A  
 772 multidimensional analysis. *arXiv preprint arXiv:2402.12976*, 2024.  
 773

774 Jun Zhao, Zhihao Zhang, Luhui Gao, Qi Zhang, Tao Gui, and Xuanjing Huang. Llama beyond  
 775 english: An empirical study on language capability transfer, 2024a. URL <https://arxiv.org/abs/2401.01055>.  
 776

777 Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min,  
 778 Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen,  
 779 Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and  
 780 Ji-Rong Wen. A survey of large language models, 2025a. URL <https://arxiv.org/abs/2303.18223>.  
 781

782 Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. How do large  
 783 language models handle multilingualism? In *The Thirty-eighth Annual Conference on Neural  
 784 Information Processing Systems*, 2024b. URL <https://openreview.net/forum?id=ctXYOoAgRy>.  
 785

786 Yiran Zhao, Wenxuan Zhang, Huiming Wang, Kenji Kawaguchi, and Lidong Bing. AdaMergeX:  
 787 Cross-lingual transfer with large language models via adaptive adapter merging. In Luis Chiruzzo,  
 788 Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas  
 789 Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume  
 790 1: Long Papers)*, pp. 9785–9800, Albuquerque, New Mexico, April 2025b. Association for  
 791 Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.nacl-long.493.  
 792 URL <https://aclanthology.org/2025.nacl-long.493/>.  
 793

794

795

796

797

798

799

800

801

802

803

804

805

806

807

808

809

810 A DERIVATION  
811812 A.1 CONNECTION BETWEEN GRADIENT ALIGNMENT AND REPRESENTATION ALIGNMENT  
813814 This is the detailed derivation of Eq. (2).  
815816 In the example, we have a linear regression model with parameter  $\mathbf{w} \in \mathbb{R}^d$ . For the task in language  $s$   
817 and  $t$ , we have data  $\mathcal{D}_{train} = \{\mathbf{X}_{train} \in \mathbb{R}^{n_{train} \times d}, \mathbf{y}_{train} \in \mathbb{R}^{n_s}\}$  and  $\mathcal{D}_t = \{\mathbf{X}_t \in \mathbb{R}^{n_t \times d}, \mathbf{y}_t \in \mathbb{R}^{n_s}\}$  respectively, containing data features and the true outputs. In different transfer scenarios,  
818  $\mathcal{D}_{train}$  contains different mixes of  $\mathcal{D}_s$  and  $\mathcal{D}_t$ . Since samples in different languages are for the  
819 same task, we assume that there is an optimal parameter  $\mathbf{w}^*$  that fits both  $\mathcal{D}_{train}$  and  $\mathcal{D}_t$ . Then for  
820 any single sample  $\{\mathbf{x} \in \mathbb{R}^{1 \times d}, y \in \mathbb{R}\}$ , the model outputs  $\hat{y} = \mathbf{x}\mathbf{w}$  and the loss on the sample is  
821  $\|\hat{y} - \mathbf{x}\mathbf{w}^*\|_2^2$ .  
822823 The inner product of gradients on  $\mathcal{D}_s$  and  $\mathcal{D}_t$  is:  
824

825 
$$\begin{aligned} & \nabla \mathcal{L}_{\mathbf{w}}(\mathcal{D}_t) \cdot \nabla \mathcal{L}_{\mathbf{w}}(\mathcal{D}_{train}) \\ &= \gamma(\mathbf{w} - \mathbf{w}^*)^T \mathbf{X}_t^T \mathbf{X}_t \mathbf{X}_{train}^T \mathbf{X}_{train} (\mathbf{w} - \mathbf{w}_j^*) \\ &= \gamma \sum_{i,j} \mathbf{x}_t^i (\mathbf{x}_{train}^j)^T [\mathbf{x}_t^i (\mathbf{w} - \mathbf{w}^*) (\mathbf{w} - \mathbf{w}^*)^T (\mathbf{x}_{train}^j)^T] \end{aligned} \quad (5)$$
  
826

827 where  $\gamma = \frac{1}{n_t n_{train}}$ ,  $\mathbf{x}_t^i \in \mathbb{R}^{1 \times d}$  is the  $i$ -th row of the matrix  $\mathbf{X}_t$  (i.e., the input feature of the  $i$ -th  
828 sample in language  $t$ ), and  $\mathbf{x}_s^j$  is the  $j$ -th row of the matrix  $\mathbf{X}_s$ .  
829830 We further decompose  $(\mathbf{w} - \mathbf{w}^*)(\mathbf{w} - \mathbf{w}^*)^T = \mathbf{Q}\Sigma\mathbf{Q}^T$  where each column  $m$  of  $\mathbf{Q}$  is an orthonormal  
831 vector  $\mathbf{q}_m$  with the corresponding eigenvalue  $\lambda_m \geq 0$  as the  $m$ -th diagonal element in  $\Sigma$ . Then Eq.  
832 (5) can be written as  
833

834 
$$\begin{aligned} \text{Eq. (5)} &= \gamma \sum_{i,j} \mathbf{x}_t^i (\mathbf{x}_{train}^j)^T [\mathbf{x}_t^i \mathbf{Q} \Sigma \mathbf{Q}^T (\mathbf{x}_{train}^j)^T] \\ &= \gamma \sum_{i,j} \underbrace{\mathbf{x}_t^i (\mathbf{x}_{train}^j)^T}_{\text{RepAlign 1}} \underbrace{\sum_m \lambda_m \mathbf{x}_t^i \mathbf{q}_m \mathbf{q}_m^T (\mathbf{x}_{train}^j)^T}_{\text{RepAlign 2}}, \end{aligned}$$
  
835

836 which is Eq. (2) in the main paper.  
837838 A.2 FT AND MTL FOR GRADIENT ALIGNMENT  
839840 This is the detailed derivation of gradient alignment effect in the FT and MTL.  
841842 A.2.1 TARGET LANGUAGE FINE-TUNING  
843844 For a model with parameter  $\mathbf{w}$ , and data  $\mathcal{D}_{train} = \mathcal{D}_t$  from languages  $t$ , the objective of Target  
845 Language Fine-Tuning is to minimize the loss  $L_{\mathbf{w}}(\mathcal{D}_t)$ . When updating the model by gradient  
846 descent, the parameter  $\mathbf{w}$  is updated as  
847

848 
$$\mathbf{w}_k = \mathbf{w}_0 - \beta \sum_k L'_{\mathbf{w}_{k-1}}(\mathcal{D}_t),$$
  
849

850 where  $\beta$  is the learning rate,  $\mathbf{w}_0$  is the initialized weight of  $\mathbf{w}$ ,  $k$  is the updating step.  
851852 The gradient of target language fine-tuning at the step  $k$  is:  
853

854 
$$g_{\mathbf{w}_k}(\mathcal{D}_t) = L'_{\mathbf{w}_k}(\mathcal{D}_t) \quad (6)$$
  
855

856 We now take the second-order Taylor expansion of the gradient:  
857

858 
$$\begin{aligned} L'_{\mathbf{w}_k}(\mathcal{D}_t) &= L'_{\mathbf{w}_0}(\mathcal{D}_t) + L''_{\mathbf{w}_0}(\mathcal{D}_t)(\mathbf{w}_k - \mathbf{w}_0) + \mathcal{O}(\|\mathbf{w}_k - \mathbf{w}_0\|^2) \\ &= L'_{\mathbf{w}_0}(\mathcal{D}_t) + L''_{\mathbf{w}_0}(\mathcal{D}_t)(\mathbf{w}_k - \mathbf{w}_0) + \mathcal{O}(\beta^2) \\ &= L'_{\mathbf{w}_0}(\mathcal{D}_t) - \beta L''_{\mathbf{w}_0}(\mathcal{D}_t) \sum_k L'_{\mathbf{w}_{k-1}}(\mathcal{D}_t). \end{aligned}$$
  
859

We omit the term  $\mathcal{O}(\beta^2)$  here since the learning rate is usually small for LLMs (e.g., 1e-5 for fine-tuning). Using the Taylor expansion of  $L'_{\mathbf{w}_{k-1}}(\mathcal{D}_t)$ , we have

$$\beta L'_{\mathbf{w}_{k-1}}(\mathcal{D}_t) = \beta L'_{\mathbf{w}_0}(\mathcal{D}_t) + \underbrace{\beta \mathcal{O}(\|\mathbf{w}_k - \mathbf{w}_0\|)}_{\mathcal{O}(\beta^2)}$$

The latter  $\mathcal{O}(\beta^2)$  is omitted. Then the gradient in Eq. (6) is:

$$g_{\mathbf{w}_k}(\mathcal{D}_t) = L'_{\mathbf{w}_0}(\mathcal{D}_t) - \beta k L''_{\mathbf{w}_0}(\mathcal{D}_t) L'_{\mathbf{w}_0}(\mathcal{D}_t)$$

Based on the product rule of derivatives, we have

$$\begin{aligned} g_{\mathbf{w}_k}(\mathcal{D}_t) &= L'_{\mathbf{w}_0}(\mathcal{D}_t) - \frac{\beta k}{2} \nabla_{\mathbf{w}_0} (L'_{\mathbf{w}_0}(\mathcal{D}_t) L'_{\mathbf{w}_0}(\mathcal{D}_t)) \\ &= L'_{\mathbf{w}_0}(\mathcal{D}_{train}) - \frac{\beta k}{2} \nabla_{\mathbf{w}_0} (L'_{\mathbf{w}_0}(\mathcal{D}_t) L'_{\mathbf{w}_0}(\mathcal{D}_t)) \quad (\mathcal{D}_{train} = \mathcal{D}_t) \end{aligned}$$

This suggests that the gradient of multi-task learning encourages to maximize the inner product of gradient on data within samples of  $\mathcal{D}_t$ .

### A.2.2 MULTI-TASK LEARNING

For a model with parameter  $\mathbf{w}$ , and data  $\mathcal{D}_{train} = [\mathcal{D}_s, \mathcal{D}_t]$  from languages  $s$  and  $t$ , we have the training loss  $L_{\mathbf{w}}(\mathcal{D}_s)$  and  $L_{\mathbf{w}}(\mathcal{D}_t)$ , respectively. The objective of multi-task learning is to minimize the loss:

$$L_{\mathbf{w}}(\mathcal{D}_s + \mathcal{D}_t) = L_{\mathbf{w}}(\mathcal{D}_s) + L_{\mathbf{w}}(\mathcal{D}_t).$$

Following the toy example, we assume the parameter  $\mathbf{w}$  is a  $d$ -dimensional vector. When updating the model by gradient descent, the parameter  $\mathbf{w}$  is updated as:

$$\mathbf{w}_k = \mathbf{w}_0 - \beta \sum_k (L'_{\mathbf{w}_{k-1}}(\mathcal{D}_s) + L'_{\mathbf{w}_{k-1}}(\mathcal{D}_t))$$

where  $\mathbf{w}_0$  is the initialized weight of  $\mathbf{w}$ ,  $k$  is the updating step,  $\beta$  is the learning rate. The gradient of multi-task learning at the step  $k$  is:

$$g_{\mathbf{w}_k}(\mathcal{D}_s + \mathcal{D}_t) = L'_{\mathbf{w}_k}(\mathcal{D}_s) + L'_{\mathbf{w}_k}(\mathcal{D}_t) \quad (7)$$

We now take the second-order Taylor expansion of each gradient. For the gradient on  $\mathcal{D}_s$  we have

$$\begin{aligned} L'_{\mathbf{w}_k}(\mathcal{D}_s) &= L'_{\mathbf{w}_0}(\mathcal{D}_s) + L''_{\mathbf{w}_0}(\mathcal{D}_s)(\mathbf{w}_k - \mathbf{w}_0) + \mathcal{O}(\|\mathbf{w}_k - \mathbf{w}_0\|^2) \\ &= L'_{\mathbf{w}_0}(\mathcal{D}_s) + L''_{\mathbf{w}_0}(\mathcal{D}_s)(\mathbf{w}_k - \mathbf{w}_0) + \mathcal{O}(\beta^2) \\ &= L'_{\mathbf{w}_0}(\mathcal{D}_s) - \beta L''_{\mathbf{w}_0}(\mathcal{D}_s) \sum_k (L'_{\mathbf{w}_{k-1}}(\mathcal{D}_s) + L'_{\mathbf{w}_{k-1}}(\mathcal{D}_t)). \end{aligned}$$

We omit the term  $\mathcal{O}(\beta^2)$  here since the learning rate is usually small for LLMs (e.g., 1e-5 for fine-tuning). Then since  $L'_{\mathbf{w}_{k-1}}(\mathcal{D}_s) = L'_{\mathbf{w}_0}(\mathcal{D}_s) + \mathcal{O}(\beta)$ , we further approximate  $L'_{\mathbf{w}_k}(\mathcal{D}_s)$  as:

$$L'_{\mathbf{w}_k}(\mathcal{D}_s) = L'_{\mathbf{w}_0}(\mathcal{D}_s) - \beta k L''_{\mathbf{w}_0}(\mathcal{D}_s) (L'_{\mathbf{w}_0}(\mathcal{D}_s) + L'_{\mathbf{w}_0}(\mathcal{D}_t)).$$

Similarly, we have

$$L'_{\mathbf{w}_k}(\mathcal{D}_t) = L'_{\mathbf{w}_0}(\mathcal{D}_t) - \beta k L''_{\mathbf{w}_0}(\mathcal{D}_t) (L'_{\mathbf{w}_0}(\mathcal{D}_s) + L'_{\mathbf{w}_0}(\mathcal{D}_t)).$$

Then we expand Eq. (7) as:

$$\begin{aligned} g_{\mathbf{w}_k}(\mathcal{D}_s + \mathcal{D}_t) &= L'_{\mathbf{w}_0}(\mathcal{D}_s) + L'_{\mathbf{w}_0}(\mathcal{D}_t) \\ &\quad - \beta k L''_{\mathbf{w}_0}(\mathcal{D}_s) L'_{\mathbf{w}_0}(\mathcal{D}_s) \\ &\quad - \beta k L''_{\mathbf{w}_0}(\mathcal{D}_t) L'_{\mathbf{w}_0}(\mathcal{D}_t) \\ &\quad - \beta k (L''_{\mathbf{w}_0}(\mathcal{D}_s) L'_{\mathbf{w}_0}(\mathcal{D}_t) + L''_{\mathbf{w}_0}(\mathcal{D}_t) L'_{\mathbf{w}_0}(\mathcal{D}_s)) \end{aligned}$$

918 Based on the product rule of derivatives, we integrate the terms in the above gradients as below.  
919

$$\begin{aligned}
920 \quad L''_{\mathbf{w}_0}(\mathcal{D}_s)L'_{\mathbf{w}_0}(\mathcal{D}_s) &= \frac{1}{2}L''_{\mathbf{w}_0}(\mathcal{D}_s)L'_{\mathbf{w}_0}(\mathcal{D}_s) + \frac{1}{2}L''_{\mathbf{w}_0}(\mathcal{D}_s)L'_{\mathbf{w}_0}(\mathcal{D}_s) \\
921 \quad &= \frac{1}{2}\nabla_{\mathbf{w}_0}(L'_{\mathbf{w}_0}(\mathcal{D}_s)L'_{\mathbf{w}_0}(\mathcal{D}_s)) \\
922 \quad L''_{\mathbf{w}_0}(\mathcal{D}_t)L'_{\mathbf{w}_0}(\mathcal{D}_t) &= \frac{1}{2}\nabla_{\mathbf{w}_0}(L'_{\mathbf{w}_0}(\mathcal{D}_t)L'_{\mathbf{w}_0}(\mathcal{D}_t)) \\
923 \quad L''_{\mathbf{w}_0}(\mathcal{D}_s)L'_{\mathbf{w}_0}(\mathcal{D}_t) + L''_{\mathbf{w}_0}(\mathcal{D}_t)L'_{\mathbf{w}_0}(\mathcal{D}_s) &= \nabla_{\mathbf{w}_0}(L'_{\mathbf{w}_0}(\mathcal{D}_s)L'_{\mathbf{w}_0}(\mathcal{D}_t))
\end{aligned}$$

924 Then the gradient of multi-task learning is finally expanded as:  
925

$$\begin{aligned}
926 \quad g_{\mathbf{w}_k}(\mathcal{D}_s + \mathcal{D}_t) &= L'_{\mathbf{w}_0}(\mathcal{D}_{train}) \quad (L'_{\mathbf{w}_0}(\mathcal{D}_{train}) = L'_{\mathbf{w}_0}(\mathcal{D}_s) + L'_{\mathbf{w}_0}(\mathcal{D}_t)) \\
927 \quad &\quad - \frac{\beta k}{2}\nabla_{\mathbf{w}_0}(L'_{\mathbf{w}_0}(\mathcal{D}_s)L'_{\mathbf{w}_0}(\mathcal{D}_s)) \\
928 \quad &\quad - \frac{\beta k}{2}\nabla_{\mathbf{w}_0}(L'_{\mathbf{w}_0}(\mathcal{D}_t)L'_{\mathbf{w}_0}(\mathcal{D}_t)) \\
929 \quad &\quad - \beta k\nabla_{\mathbf{w}_0}(L'_{\mathbf{w}_0}(\mathcal{D}_s)L'_{\mathbf{w}_0}(\mathcal{D}_t))
\end{aligned}$$

930 This suggests that the gradient of multi-task learning encourages to maximize the inner product of  
931 gradient on data within and across languages.  
932

## 933 B EXPERIMENTAL SETTINGS

### 934 B.1 METRIC FOR COSINE DEVIATION

935 We hypothesize that an LLM has aligned representations if it has highly correlated representations of  
936 translated data pairs, which rely on similar knowledge to solve the task. Therefore, we compare the  
937 cosine similarity of representations between translated data pairs and shuffled data in the dominant  
938 language. Since translated pairs rely on similar knowledge to solve the task, they are expected to  
939 have higher cosine similarity than shuffled data in the dominant language.

940 We use the cosine deviation below to quantify LLMs' pre-trained representation alignment ability.  
941 For the  $L$ -th transformer layer, we have pre-trained data representations of each sentence pair in  
942 language  $s$  and  $t$  as  $\mathbf{H}_s^L = [\mathbf{h}_{s,1}^L, \dots, \mathbf{h}_{s,n_s}^L]$  and  $\mathbf{H}_t^L = [\mathbf{h}_{t,1}^L, \dots, \mathbf{h}_{t,n_t}^L]$  where  $n_s$  and  $n_t$  are numbers  
943 of tokens in each sentence;  $\mathbf{h}_{t,n_t}^L$  and  $\mathbf{h}_{s,n_s}^L$  are token representations in the corresponding sentence.  
944 The cosine deviation of all sentence pairs is computed as:

$$\begin{aligned}
945 \quad \text{cos\_dev}(\mathbf{H}_s^L, \mathbf{H}_t^L) &= \frac{1}{n_s} \sum_i \cos(\mathbf{h}_{s,i}^L, \mathbf{H}_r^L; \text{top-10}) - \cos(\mathbf{h}_{s,i}^L, \mathbf{H}_t^L; \text{top-10}) \\
946 \quad \text{cos\_dev}(\mathbf{H}_s, \mathbf{H}_t) &= \mathbb{E}\left[\frac{1}{L} \sum_L \text{cos\_dev}(\mathbf{H}_s^L, \mathbf{H}_t^L)\right]
\end{aligned}$$

947 where  $\mathbf{H}_r^L$  is the random sentence selected from language  $s$ .  $\cos(\mathbf{h}_{s,i}^L, \mathbf{H}_t^L; \text{top-10})$  is the average of  
948 the top-10 highest cosine similarities between the token representation  $\mathbf{h}_{s,i}^L$  and all token representations in  
949  $\mathbf{H}_t^L$ . Since the translated sentences in different languages may have varied token numbers  
950 due to different tokenization, we pick top-10 highest cosine similarities and select language pairs  
951 with comparable token numbers in our experiment.

### 952 B.2 TRAINING DETAILS IN SECTION 6

953 We train our models on 4 Nvidia L40S machines. The detailed training settings for each model is:  
954

955 For each baseline model, we select training epochs from the set  $\{1, 2, 3, 4, 5, 8\}$  and pick the epoch  
956 with stable top performance.

957 For parameter efficient test in Table 5, we use LoRA with learning rate 2e-4, training batch 8 with 3  
958 epochs training on 1 Nvidia A100 machine. For MTL prompt tuning, we select best results from the

Model	LR	Batch Size	XQuAD	XCOPA				XNLI			
TL	2e-6	4 per-device	Epoch 1 on each single language	Epoch 3 on each single language				Epoch 3 on each single language			
MTL	2e-6	4 per-device	Epoch 2	Epoch 3				Epoch 3			
ShifCon	2e-6	4 per-device	Epoch 2	Epoch 4				Epoch 4			
CL-ICP	1e-3	4 per-device	Epoch 5, prompt 1, demosample 1	Epoch 20, prompt 10, demosample 3				Epoch 20, prompt 10, demosample 3			

976

977

978 following settings: the same prompt setting as our CL-ICP; increasing the prompt number (20, 30,  
 979 50) and training epochs (10, 20, 30); adjusting the learning rate from (2e-4, 1e-3, 5e-3, 1e-2). The  
 980 prompt embeddings are initialized the same as CL-ICP.

981

982

## C INFLUENCE OF DEMONSTRATION TRAINING

983

984

Table 6: CL-ICL models without training to predict demonstration samples.

985

Dataset	XQuAD (EM/F1 %)					XCOPA (Acc %)					XNLI (Acc %)					
	Language $t$	Zh	Es	Tr	El	Avg	Zh	Tr	Sw	Qu	Avg	Zh	Tr	Sw	El	Avg
CL-ICP		53.4/65.5	46.4/72.3	25.6/58.7	39.1/66.1	41.1/65.7	87.6	77.9	<b>58.8</b>	49.1	<b>68.4</b>	70.3	66.1	<b>59.6</b>	72.0	<b>67.0</b>
- Demotrain		64.4/68.9	<b>47.3/72.6</b>	28.4/58.5	40.8/66.5	<b>45.2/66.6</b>	<b>88.3</b>	77.9	55.2	<b>50.7</b>	68.0	69.7	65.8	55.7	71.0	65.5
CL-ICP (S)		61.6/69.1	44.8/71.9	23.1/56.3	41.2/67.0	42.7/66.1	87.7	78.5	57.7	49.0	<b>68.2</b>	<b>70.4</b>	<b>68.3</b>	57.4	<b>73.6</b>	<b>67.4</b>
- Demotrain		<b>65.4/69.3</b>	45.5/71.2	<b>31.4/60.1</b>	<b>42.0/67.0</b>	<b>46.1/66.9</b>	<b>88.3</b>	<b>80.1</b>	54.2	50.1	<b>68.2</b>	70.0	64.9	53.4	71.0	64.8

991

992 We show the influence of training with the demonstration prediction loss in CL-ICP in Table 6. For the  
 993 question answering task XQuAD which has sequence to sequence distribution close to pre-training  
 994 data, adding the demonstration loss does not increase the performance. However, for task XCOPA  
 995 and XNLI whose outputs are choices instead of sequences, adding the demonstration loss achieves  
 996 better performance. We hypothesize that this is because the model can easily acquire knowledge from  
 997 demonstration samples in XQuAD, while it needs further training to acquire knowledge of the task in  
 998 XCOPA and XNLI. So for tasks which have dissimilar distribution to pre-trained data, training with  
 999 the demonstration loss is beneficial.

1000

1001

1002

1003

1004

1005

1006

1007

1008

1009

1010

1011

1012

1013

1014

1015

1016

1017

1018

1019

1020

1021

1022

1023

1024

1025