Multilingual Large Language Model: A Survey of Resources, Taxonomy and Frontiers

Libo Qin[♣*] **Qiguang Chen**[♠*] **Yuhang Zhou**[♠] **Zhi Chen**[♢] **Yinghui Li**[♮]
Lizi Liao[♯] **Min Li**[♣] **Wanxiang Che**[♠] **Philip S. Yu**[♡]
[♣] Central South University [♠] Harbin Institute of Technology [♢] Shanghai AI Laboratory
[♮] Tsinghua University [♯] Singapore Management University [♡] University of Illinons at Chicago
lbqin@csu.edu.cn, {qgchen,car}@ir.hit.edu.cn

Abstract

Multilingual Large Language Models are capable of using powerful Large Language Models to handle and respond to queries in multiple languages, which achieves remarkable success in multilingual natural language processing tasks. Despite these breakthroughs, there still remains a lack of a comprehensive survey to summarize existing approaches and recent developments in this field. To this end, in this paper, we present a thorough review and provide a unified perspective to summarize the recent progress as well as emerging trends in multilingual large language models (MLLMs) literature. The contributions of this paper can be summarized: (1) ***First survey***: to our knowledge, we take the first step and present a thorough review in MLLMs research field according to multi-lingual alignment; (2) ***New taxonomy***: we offer a new and unified perspective to summarize the current progress of MLLMs; (3) ***New frontiers***: we highlight several emerging frontiers and discuss the corresponding challenges; (4) ***Abundant resources***: we collect abundant open-source resources, including relevant papers, data corpora, and leaderboards. We hope our work can provide the community with quick access and spur breakthrough research in MLLMs.

1 Introduction

In recent years, remarkable progress has been witnessed in large language models (LLMs) (Brown et al., 2020; Touvron et al., 2023a; Bang et al., 2023; Zhao et al., 2023b), which have achieved excellent performance on various natural language processing tasks (Pan et al., 2023; Nguyen et al., 2023a; Trivedi et al., 2023). In addition, LLMs raise surprising emergent capabilities, including in-context learning (Min et al., 2022; Dong et al., 2022), chain-of-thought reasoning (Wei et al., 2022; Huang et al., 2023a; Qin et al., 2023a), and even planning (Driess et al., 2023; Hu et al., 2023b).

[*] Equal Contribution

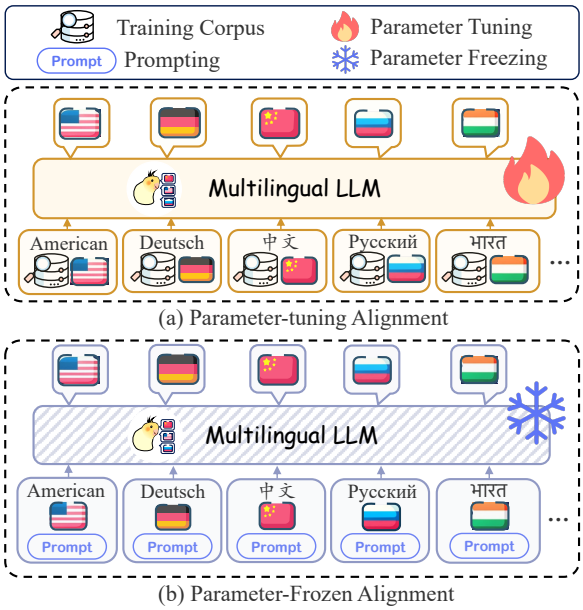


Figure 1: Parameter-Tuning Alignment (§4.1) v.s. Parameter-Frozen Alignment (§4.2). The former requires the model to fine-tune the MLLM parameters for cross-lingual alignment, while the latter directly uses prompts for alignment without parameter tuning.

Nevertheless, the majority of LLMs are English-centric, primarily focusing on English tasks (Held et al., 2023; Zhang et al., 2023i), which makes them somewhat weak for multilingual settings, especially in low-resource scenarios.

Actually, there are over 7,000 languages in the world. With the acceleration of globalization, the success of large language models should be considered to serve diverse countries and languages. To this end, multilingual large language models (MLLMs) possess the advantage of comprehensively handling multiple languages, gaining increasing attention. Specifically, the existing MLLMs can be broadly divided into two groups based on different stages. The first series of works (Xue et al., 2020; Workshop et al., 2022; Zhang et al., 2023g; Muennighoff et al., 2022) leverage multilingual

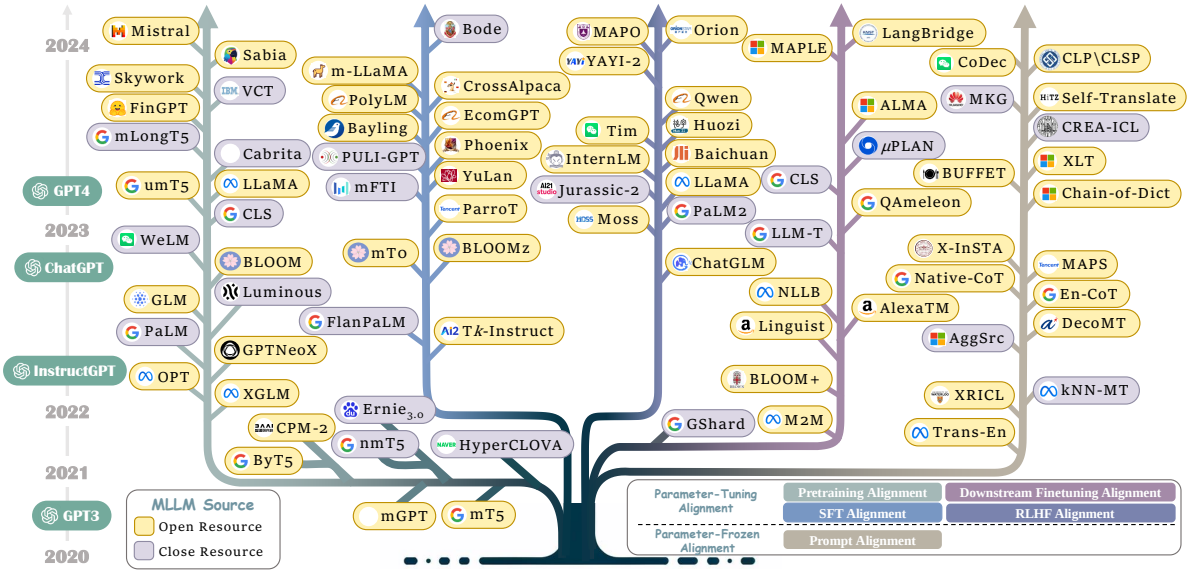


Figure 2: Evolution of selected MLLMs over the past five years, where colored branches indicate different alignment stages. For models with multiple alignment stages, the final stage is represented.

data to tuning the parameters to boost the overall multilingual performance. The second series of work (Shi et al., 2022a; Qin et al., 2023b; Huang et al., 2023a) also adapt the advanced prompting strategies to unlock deeper multilingual potential of MLLMs during parameter-frozen inference stage.

While remarkable success has been achieved in the MLLMs, there still remains a lack of a comprehensive review and analysis of recent efforts in the literature, which hinders the development of MLLMs. To bridge this gap, we make the first attempt to conduct a comprehensive and detailed analysis of MLLMs. Concretely, we first introduce the widely used data resource (§3). Furthermore, due to the key challenge of alignment across languages, we introduce a novel taxonomy according to alignment strategies (§4), aiming to provide a unified perspective in the literature, which includes: ***parameter-tuning alignment*** and ***parameter-frozen alignment*** (as shown in Figure 1). Specifically, *parameter-tuning alignment* requires the fine-tuning of model parameters to enhance alignment between English and target languages during pre-training, supervised fine-tuning, reinforcement learning from human feedback and downstream fine-tuning. *parameter-frozen alignment* refers to the alignment achieved by prompting across languages that can be achieved without the need for parameter tuning. Finally, we point out some potential frontier areas as well as the corresponding challenges for MLLMs, hoping to inspire the follow-up research (§5).

The contributions of this work can be summarized as follows: (1) ***First survey***: To the best of our knowledge, we are the first to present a comprehensive survey in the MLLMs literature according to multi-lingual alignment; (2) ***New taxonomy***: We introduce a novel taxonomy categorizing MLLMs into two alignment types: *parameter-frozen* and *parameter-tuning*, offering a unified view for understanding the MLLMs literature; (3) ***New frontiers***: We discuss some emerging frontiers and highlight their challenges as well as opportunities, hoping to pave the way for future research developments; (4) ***Exhaustive resources***: We make the first attempt to organize MLLMs resources including open-source software, diverse corpora, and a curated list of relevant publications, accessible at https://multilingual-llm.net.

We hope that this work can serve as a valuable resource for researchers and inspire more breakthroughs in future research[1].

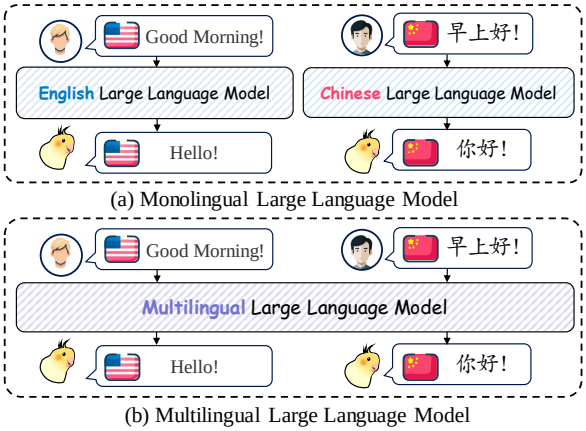
2 Preliminary

In this section, we will formally describe the definitions of monolingual large language model (§2.1) and multilingual large language model (§2.2).

2.1 Monolingual Large Language Model

Monolingual large language models (LLM) can only process one language at a time. For example, as illustrated in Figure 3 (a), English and Chinese

[1]Figure 2 illustrates the evolution of selected MLLMs over the past five years.



(a) Monolingual Large Language Model

(b) Multilingual Large Language Model

Figure 3: Monolingual Large Language Model v.s. Multilingual Large Language Model.

LLM can separately handle English and Chinese language, respectively. Formally, considering a set of languages $\mathcal{L} = \{\mathcal{L}_i\}_{i=0}^{|\mathcal{L}|}$, given input utterance $\mathcal{X}_i \in \mathcal{L}_i$ in languages $\mathcal{L}_i$, the process of monolingual LLM ($\mathcal{M}_{\texttt{mono}}$) generating the output $\mathcal{Y}_i$ can be defined as:

$$\mathcal{Y}_i = \begin{cases} \mathcal{M}_{\texttt{mono}}(\mathcal{X}_i, \mathcal{L}_i), & \texttt{mono} = \mathcal{L}_i; \\ \texttt{Unexpect}, & \texttt{mono} \neq \mathcal{L}_i, \end{cases} \quad (1)$$

where `Unexpect` indicates that the LLM generates output in an unintended language; `mono` denotes the single language.

2.2 Multilingual Large Language Model

As shown in Figure 3 (b), unlike monolingual LLM, a multilingual LLM is capable of handling and producing content in various languages simultaneously, like English and Chinese. Formally, for MLLM $\mathcal{M}_{\texttt{multi}}$, where $\texttt{multi} \subseteq \mathcal{L}$ and $|\texttt{multi}| \geq 2$, the model's response is given by:

$$\mathcal{Y} = \mathcal{M}_{\texttt{multi}}(\mathcal{X}), \quad (2)$$

where $\mathcal{X}$ and $\mathcal{Y}$ belong to multiple languages, `multi`.

3 Data Resource

In this section, we describe the widely used data resources in pre-training (§3.1), supervised fine-tuning (SFT) (§3.2) and reinforcement learning from human feedback (RLHF) (§3.3) stage (Zhao et al., 2023b) for multilingual large language model. Detailed statistics can be found in Table 1 and Table 2 in the Appendix.

3.1 Multilingual Pretraining Data

The widely used multilingual corpora for pre-training in MLLMs can be divided into 3 categories: (1) ***Manual Creation***: obtains high-quality pre-training corpora through manual creation and proofreading, which consists of the Bible Corpus (Mayer and Cysouw, 2014) and MultiUN (Ziemski et al., 2016). (2) ***Web Crawling***: involves crawling extensive multilingual data from the internet, which includes OSCAR (Suárez et al., 2019), CC-100 (Conneau et al., 2020), mC4 (Xue et al., 2021) and Redpajama-v2 (Computer, 2023). Another series of data are extracted from Wikipedia to enhance the knowledge of MLLMs. Common datasets include Wikipedia (Foundation) , WikiMatrix (Schwenk et al., 2021) and WikiExpl (Han et al., 2023). (3) ***Benchmark Adaptation***: means re-cleaning or integrating existing benchmarks to enhance data quality which includes OPUS-100 (Zhang et al., 2020), Culturax (Nguyen et al., 2023c), OPUS (Tiedemann, 2012), WMT (Kocmi et al., 2023) and ROOTS (Laurençon et al., 2022).

3.2 Multilingual SFT Data

Similarly, we categorize the existing multilingual SFT data into 4 classes: (1) ***Manual Creation***: acquires SFT corpora through manual creation and proofreading, which includes Sup-NatInst (Wang et al., 2022b), OpenAssist (Köpf et al., 2023) and COIG-PC$_{\text{lite}}$ (Team, 2023a). (2) ***Machine Translation***: translates the existing monolingual datasets into multilingual instruction datasets, which comprises xP3-MT (Muennighoff et al., 2022), MGSM8K$_{\text{Instruct}}$ (Chen et al., 2023b), CrossAlpaca (Ranaldi et al., 2023b; Cui et al., 2023), MultilingualSIFT (Chen et al., 2023i) and Bactrain-X (Li et al., 2023b). (3) ***Benchmark Adaptation***: involves transformation from existing benchmarks to instruction format. Widely used datasets include xP3 (Muennighoff et al., 2022), PolyglotPrompt (Fu et al., 2022), and BUFFET (Asai et al., 2023). (4) ***MLLMs Aided Generation***: means that the data are automatically synthesized by the MLLMs, containing Vicuna (Chiang et al., 2023), OverMiss (Chen et al., 2023g), ShareGPT (ShareGPT, 2023), BELLE (Yunjie Ji, 2023), MultiAlpaca (Wei et al., 2023c), Guanaco (Dettmers et al., 2023) and Alpaca-4 (Peng et al., 2023).

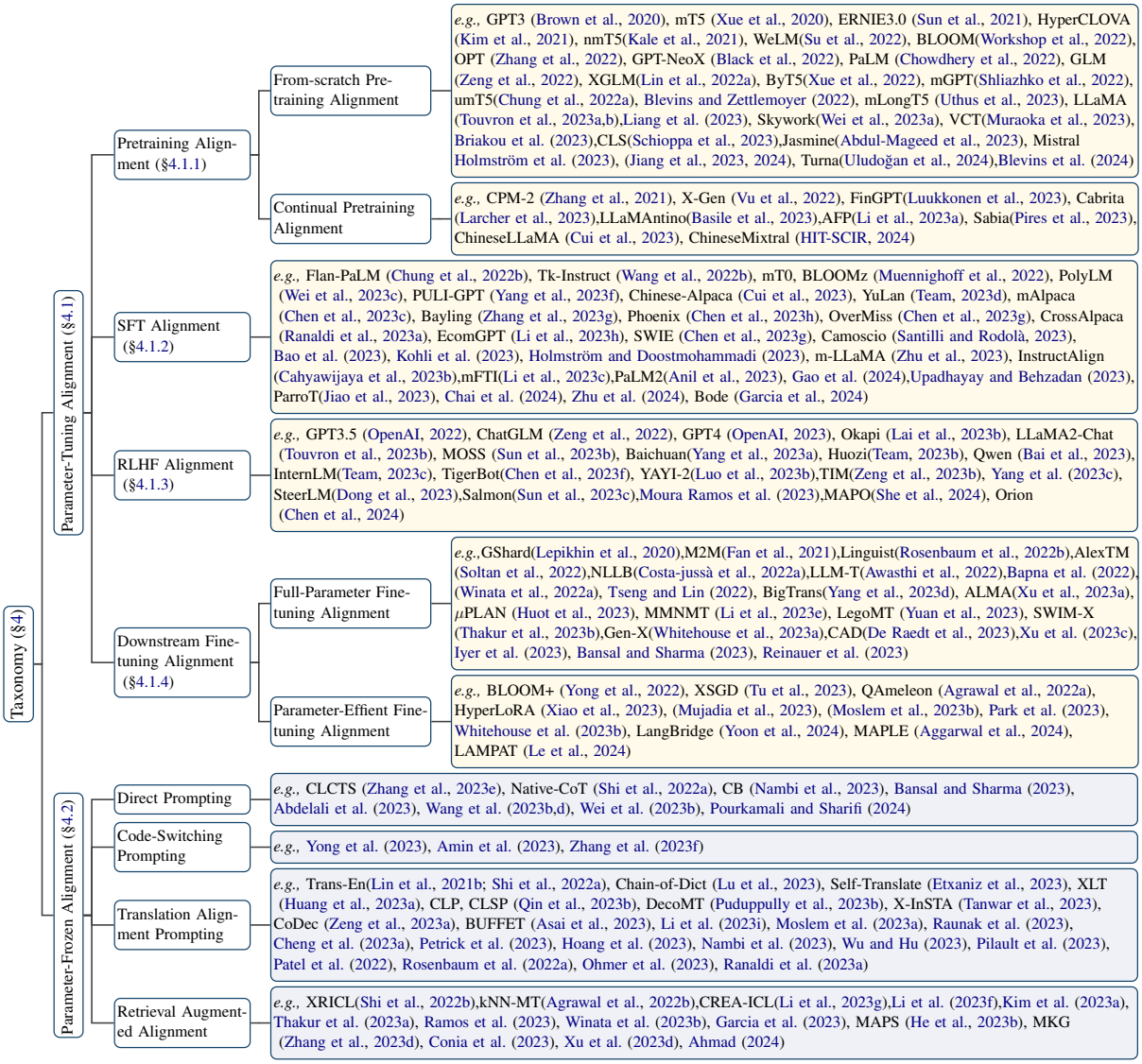


Figure 4: Taxonomy of MLLMs which includes *Parameter-Tuning Alignment Methodology* and *Parameter-Frozen Alignment Methodology*.

3.3 Multilingual RLHF Data

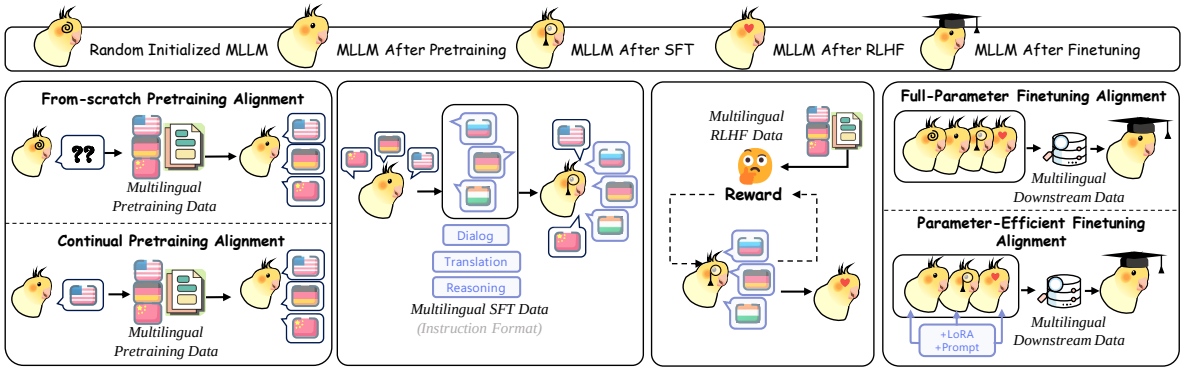
Some work leveraged the multilingual RLHF data to improve alignment. Specifically, Lai et al. (2023b) leverages multilingual ranking data for training a reward model using RLHF. Zeng et al. (2023b) introduce the TIM dataset to train a more effective reward model in multilingual contexts.

4 Taxonomy

As shown in Figure 4, we introduce a novel taxonomy including *parameter-tuning alignment* (§4.1) and *parameter-frozen alignment* (§4.2), which aims to provide a unified view for researchers to understand the MLLMs literature. Specifically, parameter tuning alignment (PTA) comprises a series of progressively advanced training and alignment strategies, including Pretraining Alignment, Supervised Fine-Tuning (SFT) Alignment, Reinforcement Learning from Human Feedback (RLHF) Alignment, and, ultimately, Downstream Fine-Tuning Alignment. These stages collectively aim to refine model parameters to align the multilingual performance systematically. Conversely, the parameter frozen alignment (PFA) focuses on four prompting strategies based on PTA: Direct Prompting, Code-Switching Prompting, Translation Alignment Prompting, and Retrieval-Augmented Alignment. This method maintains the original model parameters to achieve desired outcomes.

4.1 Parameter-Tuning Alignment

Parameter-tuning alignment indicates that MLLMs should tune their parameters for better cross-lingual alignment (Wen-Yi and Mimno, 2023). As shown in Figure 5, we discuss the four categories of



(a) PTA in Pretraining Stage (b) PTA in SFT Stage (c) PTA in RLHF Stage (d) PTA in Downstream Finetuning Stage

Figure 5: Overview of Parameter-Tuning Alignment (§ 4.1) Methods, which including *PTA in Pretraining Stage* (§ 4.1.1), *PTA in SFT stage* (§ 4.1.2), *PTA in RLHF stage* (§ 4.1.3) and *PTA in Downstream Finetuning stage* (§ 4.1.4).

parameter-tuning alignment (PTA), including PTA in pretraining stage (§4.1.1), PTA in SFT stage (§4.1.2), PTA in RLHF stage (§4.1.3) and PTA in Finetuning stage (§4.1.4).

4.1.1 PTA in Pretraining Stage

From-scratch Pretraining Alignment. A series of approaches have achieved to alignment across languages by tuning the initially random parameters of MLLMs during pretraining (see Figure 5 (a)). Specifically, Blevins and Zettlemoyer (2022); Briakou et al. (2023); Holmström et al. (2023) observed that adding a few multilingual data during the from-scratch pretraining alignment, even unintentionally, can significantly boost the multilingual performance. Inspired by this, Zeng et al. (2022); Su et al. (2022) used bilingual data in their from-scratch pretraining for alignment. mT5 (Xue et al., 2020), Ernie3.0 (Sun et al., 2021), ByT5 (Xue et al., 2022), BLOOM (Workshop et al., 2022), LLaMA (Touvron et al., 2023b,a), PaLM (Chowdhery et al., 2022), Mistral (Jiang et al., 2023), Mixtral (Jiang et al., 2024), PolyLM (Wei et al., 2023c), Kale et al. (2021); Kim et al. (2021); Shliazhko et al. (2022); Chai et al. (2022); Schioppa et al. (2023); Abdul-Mageed et al. (2023); Uthus et al. (2023); Wei et al. (2023a); Uludoğan et al. (2024) incorporated multilingual data in pretraining stage for better alignment. Blevins et al. (2024) utilizes Mixture-of-Experts (MoE) to independently train language models on subsets of multilingual corpora to alleviate the problem of multilingual parameter competition. Furthermore, to enhance the performance of low-resource languages, umT5 (Chung et al., 2022a) and XGLM (Lin et al., 2022a) adopted equitable data sampling methods during from-scratch pretraining. Muraoka et al. (2023) introduced VCT to leverage vision for indirect cross-lingual alignment in from-scratch pretraining.

Continual Pretraining Alignment. To address the high computational cost of from-scratch pretraining, continual pretraining alignment builds the pretraining process upon pretrained MLLMs (as shown in Figure 5 (a)). Specifically, CPM-2 (Zhang et al., 2021), Sabia (Pires et al., 2023), FinGPT (Luukkonen et al., 2023), X-Gen (Vu et al., 2022), AFP (Li et al., 2023a), Cabrita (Larcher et al., 2023), LLaMAntino (Basile et al., 2023) focused on adding more target language data during continual pretraining for general performance. Further, Cui et al. (2023); HIT-SCIR (2024) emphasized extending the MLLMs' vocabularies to adapt to new languages.

4.1.2 PTA in SFT Stage

As illustrated in Figure 5 (b), PTA in SFT stage means leveraging multiple multilingual task data with instruction format for tuning parameters (Fu et al., 2022; Yang et al., 2023f; Team, 2023d; Chen et al., 2023c,g; Ranaldi et al., 2023a; Li et al., 2023h; Chen et al., 2023g; Santilli and Rodolà, 2023; Bao et al., 2023; Kohli et al., 2023; Holmström and Doostmohammadi, 2023; Garcia et al., 2024). In particular, models like Flan-PaLM (Chung et al., 2022b), mT0, BLOOMz (Muennighoff et al., 2022), PolyLM (Wei et al., 2023c), Tk-Instruct (Wang et al., 2022b), Chinese-Alpaca (Cui et al., 2023), Bayling (Zhang et al., 2023g) and Phoenix (Chen et al., 2023h), directly incorporated multilingual data in the SFT stage to achieve implicit multilingual alignment across languages. Besides, to